Phenotyping wheat structural traits from millimetric resolution RGB imagery in field conditions

Simon Madec

THESE

Présentée par

Simon Madec

Pour obtenir le titre de

DOCTEUR DE L'UNIVERSITE D'AVIGNON ET DES PAYS DE VAUCLUSE

Spécialité : Télédétection de la Végétation

Phenotyping wheat structural traits from millimetric resolution RGB imagery in field conditions

Date de soutenance : 25 Juin 2019

Composition du jury :

Thierry Artières : Rapporteur

Gilles Charmet : Rapporteur

Scott Chapman : Examinateur

Ryad Bendoula : Examinateur

Frédéric Baret : Directeur de thèse

Remerciements

Je remercie avant tout mon directeur de thèse Fred pour la confiance qu'il m'a accordée. Mais aussi ses bonnes idées, sa sympathie et les expériences partagées (en missions, à l'étranger, sur le terrain…). Petit lapinou content.

Je remercie aussi toutes les personnes qui m'ont accompagné durant ces 3 années. Il y en a beaucoup, ils sont un peu partout: mes parents, grand-parents et ma famille, à CAPTE, chez Arvalis, chez HiPhen, à l'INRA, en Chine, à Montpellier, en Australie (merci Scott pour ton accueil !), en Bretagne, à Paris, au Japon …

Merci à Alexis – Benoit – Fred de faire vivre ce jolie écosystème : l'UMT CAPTE.

Je tiens aussi à remercier mes rapporteurs et examinateurs de consacrer une partie de leur précieux temps.

List of abbreviations and acronyms

AGB	Above Ground Biomass
CNN	Convolutional neural network
DaS	Date after sowing (in GDD
DC	Dense Cloud
DSM	Digital surface model
DTM	Digital terrain model
FCover	Fraction of Vegetation Cover
FN	False negative
FP	False positive
GAN	Generative adversarial network
GDD	Growth degree day
GPU	Graphics Processing Unit
GSD	Ground sample distance
GxExM	Genetics by Environment by Management interactions
H^2	Broad sense heritability
IOT	Internet Of Things
IoU	Intersection over Union
LiDAR	Light detection and ranging
MAE	Mean Absolute Error
PH	Plant Height
QTL	Quantitative Trait Loci
RCNN	Region-based Convolutional neural networks
Rp	Rank percentile
RPN	Region Proposal Networks
SfM	Structure from Motion
TP	True Positive
TRL	Trait readiness Level
UAV	Unnamed Aerial Vehicule
WS	Water Stress
WW	Well Wattered

Content

Introduction

1.1 The main challenges for agriculture: population growth, climate change, environmental footprint & natural resources

The last century has been marked by the green revolution. This change, characterized by industrialized agriculture, area expansion and chemical-intensive technologies with their high responsiveness varieties have tripled the agricultural yields. This transformation has led to a doubling of the world population.

This trend of population growth is predicted to continue in the developing countries and the world's population is expected to reach 9.2 billion by 2050. Hence, the future farmers will then need to produce food for 2.3 billion more people. In addition to that, the consumption per capita is expected to change: a higher purchasing power and a greater demand for meat, fish and processed food is expected. These changes are and will continue to be especially true in developing countries as predicted by (Bongaarts and Watkins, 1996). This adds pressure to the food supply system. According to some studies, the world will need 50 to 100% more food by 2050 (Baulcombe et al., 2009; Porter and Semenov, 2005).

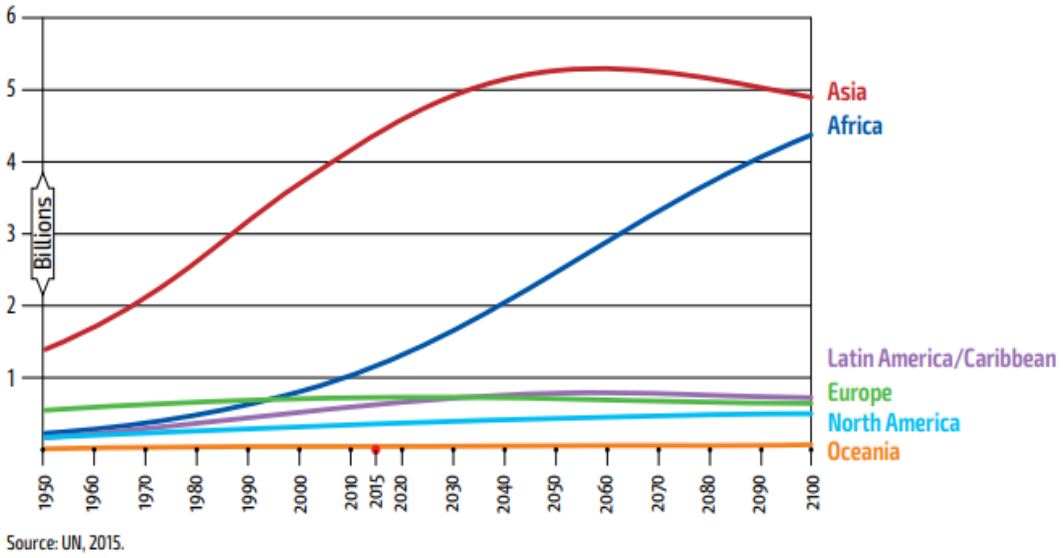


Figure 1: Population growth to 2100, by region (medium variant) Source: UN, 2015

On the other hand, it is likely that this general trend of population growth along with the expansion of agriculture will continue to impact the climate. Future atmospheric conditions will be characterized by increased variability in temperature, precipitation and elevated atmospheric carbon dioxide (CO2) concentration (Guillaume, 2018). These changes, in particular regarding the temperature, will impact the food production. A stronger year to year variability is expected and short-term variability in supply might put in danger the whole food supply system (Schlenker and Roberts, 2009). For instance, increased temperatures can lead to greater evaporative demand

and thus water stress on crops. Crops are more likely to be subject to abiotic and biotic stress (Zandalinas et al., 2018). Most of the simulations of crop yield under a climate change predict that yields of most crop species will be reduced. Under pessimistic high CO2 emissions scenario (RCP8.5) wheat and maize yields are projected to decrease respectively by 22% and 27% by 2050 (Porfirio et al., 2018). The global wheat production is supposed to fall by 6% for each increase of 1.0 °C (Asseng et al., 2015). It has to be noticed that crop yields are more negatively affected across most tropical areas than at higher latitudes (Wheeler and Von Braun, 2013). However, the ranges of the uncertainty of these simulations under the diversity of environmental conditions are high. We still need more knowledge regarding the parameterization of crop models for the CO2 fertilization (Leakey et al., 2009; Ziska and Bunce, 2007), the temperature response functions to extremely hot days and the degree of the climatic changes. The lack of knowledge on cultivars needed for modeling also contributes to the uncertainties (Wang et al., 2017).

This challenge for food security as mentioned above gives importance to the choices of the agricultural practices (Foley et al., 2011; Jarecki et al., 2018), the food distribution and consumption, crop management and the appropriate use of genetic variability (He et al., 2018; Parent et al., 2018). We will focus on the next sections on the genetic adaptation of crops.

1.2 Genetic improvement

1.2.1 A short history about plant domestication

Sedentary agriculture and the domestication of plants are estimated to have come into practice 10000 years ago (Frumin, 2017; Zohary et al., 2012). Farmers selected plants with desirable characteristics and used their seeds for subsequent generations. Plants with less desirable characteristics were then culled. It has to be noted that domestication was also enforced with natural reason and unintentional genetic mutation. This has resulted in an accumulation of characteristics over time.

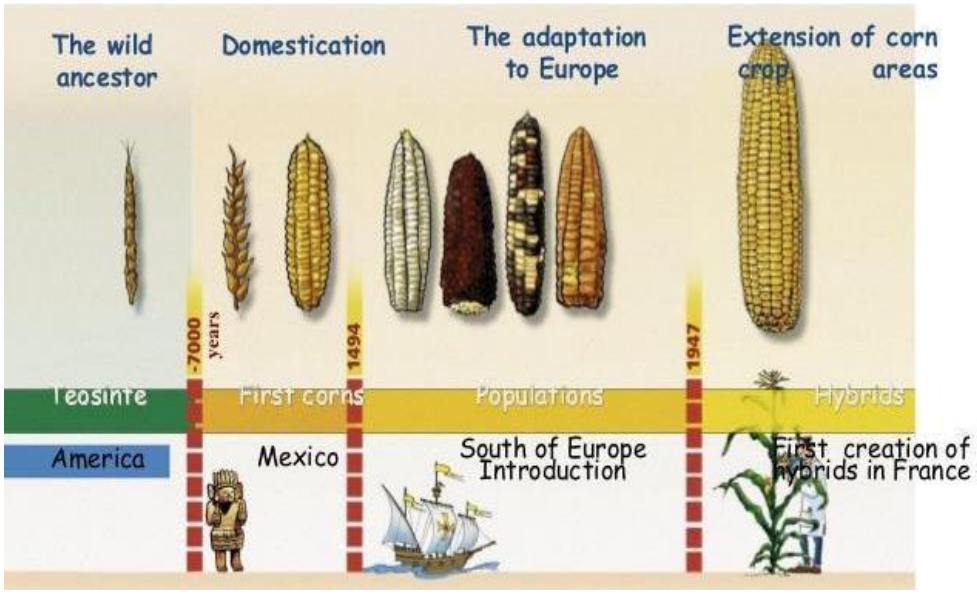


Figure 2: Illustration of the evolution of maize. Source: ScienceMag

By the end of the 19th century, the first genetic cross of chosen parents was realized. For instance, the first hybrid maize was created in 1908 (Crow, 1998). The next improvement in crop production was achieved through plant breeding and crossing. The technology of plant breeding has indeed advanced greatly in the last century.

Plant breeding was realized through cycles of crossing and phenotypic selection. This has induced a large increase in yield. After World War II, this breeding technology was developed in many developing countries. The subsequent yield increase was then more homogeneously observed. It was of the main component responsible for the green revolution.

The continuous observation, migration, selection, and improvements in cross-breeding have resulted in homogenous cultivar with well-defined criterions and fixed in the time (reproducible).

The discovery of DNA and the next generation sequencing (Metzker, 2010) then brought plant breeding into a new era (Evenson and Gollin, 2003). Now, it is not only possible to identify all the genes and genetic variants contributing to agronomic traits but also assess these changes made during the breeding process.

1.2.2 The need for new genotypes

The main goal of plant breeders is to create genotypes that are well-adapted to the future growing conditions defined by the meteorological conditions, soil properties and management practices in the production basin of interest.

A recent study suggested that the productivity of major food crops is either stagnating or not increasing at the rate needed to ensure food security (Ousseini et al., 2017). In addition to that: climate change, population growth, and environmentally limited resources place a three-fold pressure on agriculture crop production and food security.

To meet the future food demand, breeding efforts have to target more resource-efficient and stress-tolerant crops in different managing systems. This places a big demand on new cultivar with more resource efficiency and well-targeted to their environments (Licker et al., 2010; Neumann et al., 2010).

1.2.3 Genomic selection

Advances in sequencing technologies and bioinformatics tools have allowed rapid progress since the sequencing and assembly of the rice genome in 2005 (Sasaki, 2005). During the last decades, the tremendous decrease in the cost of DNA sequencing has led to a rapid rise in the size of crop genomic data. Today, there are over 260 land plant nuclear genomes publicly available in GenBank (Benson et al., 2012), including that of major crops. This enables a new way of genomic improvement (Heffner et al., 2009; Jannink et al., 2010).

These new tools and techniques represent a paradigm shift by facilitating the direct study of all the genotypes and their relationships with the phenotype (Tester and Langridge, 2010). This can be very useful for the study of complex traits that have a multi-genic nature and a complex interaction with the environment such as yield and abiotic stresses. By genomic selection, the efficiency and speed of the selection process have also been improved by avoiding multiple cycles of phenotyping selection.

Hence, genomic-based breeding has been applied to produce many crops (Dwivedi et al., 2017; M Perez-de-Castro et al., 2012; Mousavi-Derazmahalleh et al., 2019).

1.2.4 The phenotyping bottleneck

Currently, we have a lot of genomic information for several crops and the tools to relate them to their phenotype. This applies even for complex species such as wheat (Appels et al., 2018). To harness this wealth of genomic information, it has to be comprehensively linked to the phenotype under a given environment.

Because of the limitations of marker-assisted selections for assessing complex traits, the genome-enabled prediction is becoming increasingly important in plant breeding. Genome selection (Hayes and Goddard, 2001) includes all markers' information in prediction models, thus, avoiding biased marker affect estimates and capturing the variation due to small-effect QTL (Quantitative Trait Loci) better (Heffner et al., 2009).

Genomic selection needs a 'training population' of the genotype considered that has been both genotyped and phenotyped to develop a model that takes genotypic data from a 'candidate population' of untested individuals and produces genomic estimated breeding values (GEBVs). This breeding technology has been remarkably accurate (Habier et al., 2007; Hayes and Goddard, 2001; Zhong et al., 2009) and has brought a revolution in plant breeding (Heffner et al., 2009).

This paradigm shift changes the role of phenotyping, which would now then serve to both update prediction models or simply to train a prediction model.

Typical plant breeding programs grow thousands of cultivar in a target environment. The varieties (cultivar) are identified through a genetic recombination pipeline (crossing of lines). This is most often carried out in complex private programs (Messina et al., 2018) which explains the lack of information and why a clear explanation of this program is not possible here.

Cultivars are typically aggregated for plots of plants grown together in mini canopies of 1 to 20 m^2 depending on the purpose and crop type, we call them microplots. These microplots are tested for adaptation to abiotic and biotic conditions. For instance, a part of the trial could be subjected to water stress and thus this part would not be irrigated. The presence of genotype by environment interaction is critical for the analysis of complex traits and the analysis is thus performed over multiple sites and several years.

Because of the labor-intensive and the expensive nature of field phenotyping, many crop breeding programs make a single measurement of the final yield on replicated plots in contrasting environments over multiple seasons (Millet et al., 2016). However, yield is not a heritable trait and hence it is complicated to be linked with the genotype.

The presence of Genotype-by-Environment-by–Management interactions (GxExM) is a significant challenge for the development of predictive approaches. Improvement in the speed and accuracy of phenotyping along with the integration of estimated traits into models representing the functioning of the plant will greatly enhance the value of the genomic selection.

Affordable phenotyping associated with genomic selection may give us new insight into the stress resilience of crops; this will be relevant the food security challenge.

This is the reason for the urgent need to develop accurate, cheap, robust and easy to integrate phenotyping pipeline. Heritability of a measurement methodology is a measure of the

degree to which it can be used as a selection or trait discovery tool (Piepho et al., 2008). The missing heritability (Maher, 2008) problem has to be taken into account, large relative measurement errors could lead to low heritability. This explains why the objective of our study is to deliver heritable traits to the breeders.

We will present in the next section a brief overview of high throughput phenotyping in the field and also highlight the work done with images.

1.3 State of the art of the field high-throughput crop phenotyping

Plant phenotyping under controlled conditions have been well developed (Bai et al., 2016; Cabrera Bosquet et al., 2015; Pereyra-Irujo et al., 2012). However, progress in the field has been relatively slow. We will cover the advances and remaining challenges for field phenotyping in this section.

1.3.1 Some definitions

To get unambiguous use of the terms 'trait', 'variable' and 'measurement' we will propose the generally accepted definitions in the phenotyping community:

- **Trait:** it is an intrinsic characteristic of the organ, plant or canopy (e. g. height, length, surface, intercepted radiation, stomatal conductance of the leaf...).
- **Variable**: it is a discrete (class, order) or continuous quantity that quantifies the trait and can be used to compare different individuals/genotypes/treatments. A variable will be associated with a unit. It may be:
  - the estimated value of the variable derived from the measurement of a parent variable and using a given processing method. This will provide the "apparent" or "estimated" value of the variable: estimates of the same trait from several methods may show differences.
  - the value of a parameter of a model that is adjusted on a set of (parent) measurements. This value will depend on the model used, the estimation method and the distribution of the parent variables.

  A variable results from the combination of four elements:
  1. The trait to be estimated
  2. The specie or group of species on which the trait should be quantified
  3. The method used to estimate the trait
  4. The unit in which the trait is quantified.
  5.

1.3.2 Traditional field phenotyping approaches

Traditional breeders select cultivar based on the physical appearance in their environment. This could be, for example, the measurement of height with a ruler, times series measurement of the appearance of the phenological stages or comparative numerical scores or indices (leaf rolling, disease infection). Agronomic performances such as yield and biomass sampling are also typically conducted. However, these traits although derived from relatively reliable and simple measurements are often lacking heritability since they are strongly controlled by the GxExM interactions (Hallauer et al., 2010). In addition, these measurements are generally

labor intensive, low throughput and prone to errors in the sampling. This poses a limitation when breeders want to screen large phenotyping platforms.

1.3.3 Field phenotyping platforms

A standard phenotyping platform is constituted of hundreds of microplots. The data then have to be gathered over a large area (typically between 1 and 10 thousand meter square) where thousands of microplots can be sowed. Each microplot corresponds to a genotype, conducted under a given management practice. There might be several management practices investigated within one platform, including sowing density and date, nitrogen fertilization, irrigation or protection from biotic stresses. Further, each genotype conducted under a given management practice could be repeated several times to control possible variation of the environmental effects (mostly induced by soil heterogeneity) and ensures to quantify the repeatability of the measurements performed.

Figure 3: Aerial view of the field phenotyping platform of Arvalis in Greoux-les-Bains France.

These phenotyping systems are often divided in two categories: the ones that operate at the ground level and those operating aerially. The choice of the platform depends on the cost, scalability, spatial and temporal resolution desired.

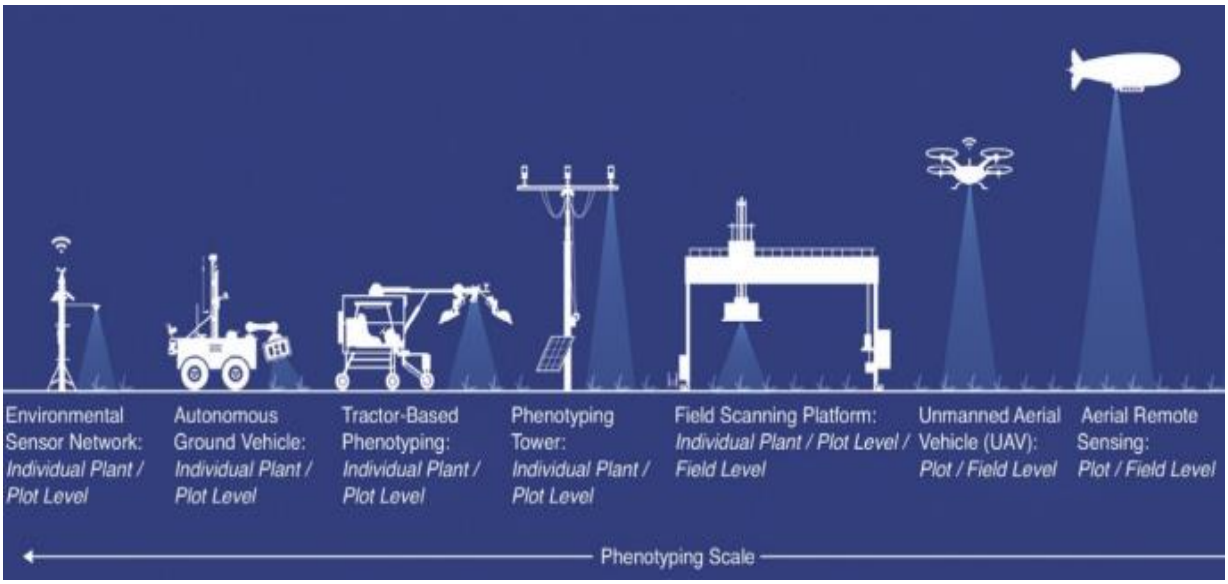


Figure 4: Illustration of the different field phenotyping systems. Source : (Shakoor et al., 2017)

1.3.3.1 Ground phenotyping systems

Fixed Ground field phenotyping systems are well suited to develop phenotyping methodologies (Sadeghi-Tehran et al., 2017a, 2017b) by facilitating the data collection (e.g avoiding geo-referencing errors or vehicle mobility issues) and advantaged the spatial resolution while allowing the use of heavier sensors. This was the case for the zip-line (Kirchgessner et al., 2017) and gantry crane (Virlet et al., 2016) installations. The drawback of these systems was their limited study area, restraining the multi-site experiment. Mobile ground systems have also been developed. The simplest versions of these are the wheeled buggies (White and Conley, 2013). These carts can also be motorized (White and Conley, 2013) and made fully autonomous, thus allowing high throughput phenotyping. A lot of research has been conducted to improve the robustness of these autonomous ground vehicles in the recent years (Burud et al., 2017; Grimstad and From, 2017; Underwood et al., 2017). They are relatively expensive systems with moderate throughput ranging from 50 to 250 microplots per hour.

1.3.3.2 Aerial phenotyping systems

The availability of aerial systems owing to the commercialization and affordable UAV (Unmanned Aerial Vehicles) combined with the accessibility of photogrammetric software have gathered the attention of many researchers in the last ten years (Yang et al., 2017). UAV or Drones have the ability to fly at an altitude between 10 and 100 meters which allows millimetric resolution images (from few millimeters to 1m) (Aasen, 2017; Jay et al., 2018). Their ability to cover larger areas within a short period of time has made them suitable for field trials. Their main limitation is the size of the payload (<20 kg and much lower in most models) and weather conditions as they ideally perform best under stable weather conditions with a light wind. See (Yang et al., 2017) for an extended review of the use of UAV in field crop phenotyping. These aerial systems could cover a platform within few minutes to few hours depending on the area of the platform and the ground spatial resolution required.

1.3.4 Sensors

Images are now playing an important role in plant phenotyping. The advances in sensors (smaller, lighter and cheaper, increased resolution and sensitivity) have contributed to their wide usage.

The most affordable sensors are visible light (400-700 nm) cameras (RGB images). RGB cameras have a high versatility (Remondino and El-Hakim, 2006) and are easy to maintain. They can be easily integrated in many systems and provide (2D) color images. RGB images have a lot of direct applications such as Canopy Cover estimation and LAI estimation. Indeed, every pixel of the image can be classified as vegetation or background, thus providing an estimation of the gap fraction in a given direction. A limitation of this method is the occlusion of the leaves. The illumination condition is another limitation that may affect the results to a great extent. However, these sensors offer an unrivaled spatial resolution, for instance, the best smartphone on the market now offers a spatial resolution between 12 and 20 megapixels.

This high spatial resolution has made the computation of some features say, HOG, SIFT, and SURF (Dalal and Triggs, 2005). Associated with a classifier, these features have been used for phenological stages detection or segmentation (Guo et al., 2015; Sadeghi-Tehran et al., 2017a, 2017b). Some of these features are also scale-invariant and have made 3D reconstruction of the scene possible. This technique, famously known as structure from motion, is based on the knowledge of the camera intrinsic parameters and requires a series of images acquired from various angles and positions. Some interesting phenotype information can be extracted from this 3D structure of the scene. For example, the traits such as plant height, leaf angle and cases of lodging can be extracted from the reconstructed 3D canopy structure. Advances in computer vision now allow plant and organ counting in field conditions which will be more developed in the following sections. It has to be noted that several of the traits developed have been computed in laboratory conditions or with hand-held sensors which ease the acquisition procedure. The main challenges in adopting them to field conditions are the various conditions of data acquisition and to increase the throughput which results in a spatial resolution trade-off.

Leaf and plant reflectance are governed by leaf surface properties and internal structure, as well as by the concentration and distribution of biochemical components. Remote sensing analysis in the infrared domains adds valuable information to assess the physiological status of a plant. In the near infrared domain, no strong absorption feature is present because of the specific structural discontinuities encountered in the leaf (Campbell and Wynne, 2011; Gao, 1996; Lillesaeter, 1982). Meanwhile, the pigment in plant leaves including chlorophyll strongly absorbs visible light. At the canopy level, spectral reflectance is a combination of soil and vegetation reflectance. The use of spectral vegetation indices calculated as a simple ratio or normalized difference from infrared and visible bands have been commonly used as indicators of the vegetation status (green area index (Verger et al., 2014), stay-green, vegetation cover …). Development of low weight and low-cost multispectral cameras (Berni et al., 2009) also referred to as 2D snapshot multi-cameras has facilitated their integration in field phenotyping platform and the computation of many vegetation indices.

While multispectral cameras take images at a specific wavelength, hyperspectral cameras capture spectral information within every exposure. Hyperspectral information may access specific traits, but most of the times, dedicated wavelength are selected to compute vegetation indices (Aasen and Bolten, 2018). A general trade-off is spatial versus spectral resolution. Both of them are crucial for plant phenotyping and indeed, the accuracy of the phenotype measured is crucial to get good repeatability and enable genetic analysis. The high throughput natures of these

measurements have the potential to reduce the cost associated with the screening of large phenotyping platform.

A common drawback of these spectral imagers is their calibration procedure. Radiometric calibration is often performed with a reference panel placed in the field. Stable illumination condition between the acquisition time of the radiometric target and the microplots is mandatory. Indeed, the presence of cloud affects the incident illumination that will be absorbed and reflected by the crops. Anisotropy effects should also be minimized by restricting the observation to a limited range of viewing angle. A spatial co-registration also has to be performed, which is a limitation for close range phenotyping applications. State-of-the-art methods in UAV spectral remote sensing were evaluated and well discussed in (Aasen and Bolten, 2018; Adão et al., 2017).

The table below lists the traits now accessible in field conditions, robustness and limitation of the methods are evaluated by their technological readiness level (TRL)(Mankins, 1995).

Table 1 : A summary of crop traits and their associated sensors, spatial resolution and methodology Technological Readiness Level

	TRAIT	METHOD	RGB	Multispectral	LiDAR	TIR	Other sensors	Diurnal cycle	View direction	Resolution		Drone	Ground system	Technological Readiness Level (TRL)								
										Small (wheat)	Big (maize)			1	2	3	4	5	6	7	8	9
CANOPY STRUCTURE	Vegetation Index	Band combination							0°	20 cm	20 cm											
	Canopy height	Structure from motion							0°	1 cm	1 cm											
		Distribution of height							0°	5 mm	5 mm											
		Photogrammetry							0°	1 mm	1 mm											
	Green Fraction	Green classification (SVM)							0°-45°	1 mm	5 mm											
		Green segmentation (deep)							0°-45°	1 mm	5 mm											
		Empirical transfer functions							0°-45°	20 cm	20 cm											
		RT model inversion							0°	20 cm	20 cm											
		Height segmentation							0°	5 mm	5 mm											
	GAI	Empirical transfer							0°	20 cm	20 cm											
		1D RT inversion reflectance							0°	20 cm	20 cm											
		1D RT inversion Green Fraction							0°-45°	1 mm	5 mm											
		3D RT inversion reflectance							0°	20 cm	20 cm											
		3D RT inversion Green Fraction							0°-45°	1 mm	5 mm											
		3D LiDAR inversion							0°	20 cm	20 cm											
	FIPAR	Empirical transfer							0°-45°	20 cm	20 cm											
		RT inversion							0°-45°	20 cm	20 cm											
		3D inversion							0°	5 mm	5 mm											
	Leaf orientation	RT inversion							0°-45°	20 cm	20 cm											
		3D inversion							0°-45°	5 mm	5 mm											
	row structure	row cross section							0°	5 mm	5 mm											
	Lodging	Height (LiDAR)							0°	5 mm	5 mm											
		height (SfM/photogrammetry)							0°	1 cm	1 cm											
	Leaf Rolling	Empirical transfer function						D	0°	20 cm	20 cm											
	Canopy temperature	with cover fraction		&				D	45°	20 cm	20 cm											
	Canopy fluorescence	SIF (with cover fraction)						?	0°-45°	20 cm	20 cm											
		LIFT						?	0°													
PLANT/ORGAN CHARACTERISTICS	Plant density @emergence	Small plants							45°	0.4 mm	-											
		Big plants							0°	-	5 mm											
	Stem density	After harvest							0°	0.2 mm	1 mm											
	Stem diameter	After harvest							0°	0.2 mm	2 mm											
	Organ density	Machine learning							0°	0.5 mm	0.5 mm											
	Leaf size & orientation	Stereo	S	S					?	0.5 mm	5 mm											
		Fusion			&					5 mm	5 mm											
	Leaf Glaucousness	Glaucousness							0°	20 cm	20 cm											
	weed infestation	shape and position							0°		1-5 mm											
	Diseases	Cercosporia							0°		0.2-0.5 mm											
		Jaunisse																				
	Chlorophyll Content	RT model inversion							0°		1-5 cm											
		VI empirical (average microplot)							0°		0.5-5 cm											
		VI empirical (high spatial resolution)							0°		0.5-1 mm											
	Leaf temperature	Distribution of T°						D	45°		1-5 mm											
	Stomatal conductance	Co-distribution T° & APAR						D	45°		1-5 mm											
	Leaf fluorescence	LIFT									1-5 mm											

This table was compiled by Frederic Baret, inspired from the literature and the work mainly developed within the CAPTE unit.

We can see from this table that a large number of traits are currently estimated with 2D images.

Images generally have a throughput advantages against other sensor-derived methodology. The high levels of technology of imaging devices can be explained by their overriding place in the public sector, this induce a reduction of the cost for phenotyping (Reynolds et al., 2018).

The variables that estimates traits differs from the used sensors and methodologies. A large number of traits are not at a full readiness level. The reasons behind will be highlighted in the next section.

1.3.5 From measurements to traits

The extraction of phenotype-related traits from plants goes through many processing and calibration steps that can impact the accuracy of the estimates. For instance, during the acquisition of images from sensors mounted on UAVs, the images could be blurred if an inappropriate flying height and shutter speed are used. 3D mosaicking and microplots extraction might fail if protocols were not followed during the collection of ground control points. For example, a ground control point covered by leaves. Radiometric calibration could be compromised by the passage of a cloud. In another acquisition scenario, the collection of information from various angle and frequency may suffer from the BRDF effects. One more example is the vigneting correction that has to be performed for every multispectral image acquired since we have already experienced commercial cameras where no vigneting correction had been made.

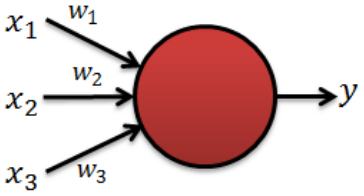
This highlights the importance to trace the metadata such as the flight parameters and calibration procedure. This meta-information should be accessible and there must be no black box in the processing pipeline. It is also important to standardize the processing chains and sensors used.

The recent review on vectors and sensors showed that this is a currently matures aspect of the phenotyping activities, although significant and continuous improvements are expected. The phenotyping bottleneck is now shifted from the data acquisition to the data processing issues. In the next section we will present a brief state of the art of the machine learning methods currently used in field phenotyping based on images.

1.4 Machine and Deep learning for trait estimation from images

1.4.1 Introduction to deep learning

Machine learning is the science of training systems to automatically learn and improve from experiences. They are trained to perform an assigned task and they learn to discriminate between the labeled input features using statistics. Originally inspired by the human brain and the biological neuron, Franck Rosenblatt in 1958 proposed the perceptron (Rosenblatt, 1958). The perceptron is an artificial neuron which gives a boolean output based on a linear operation on boolean inputs x_i and weights w_i. This is followed by an activation function. The perceptron can be represented as below:



$$if \sum x_i.w_i > 0:$$
$$y = 1$$
$$else:$$
$$y = 0$$

Figure 5: A visual representation of the perceptron

A single neuron performs only elementary operations. However, when associated in a network, neurons perform complex tasks. Neural networks take this principle up by interconnecting multiple elementary units to solve a large range of problems. They are organized in layers where each layer responds to a level of abstraction of a problem and then passing the extracted information to the others.

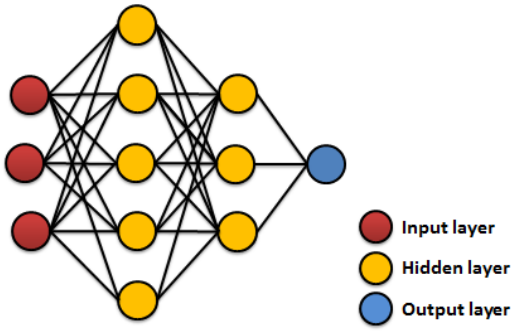


Figure 6: A visual representation of a neural network

The difference between machine learning and deep learning is that deep learning automatically learns to extract the relevant features that will help to take the decision. To perform this task, deep learning models typically have more than three layers.

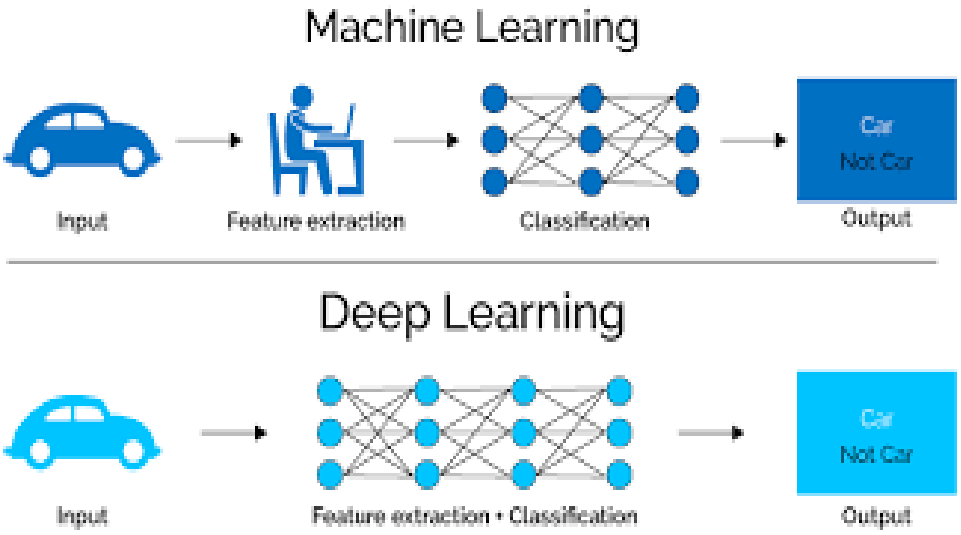


Figure 7: Illustration of the difference between machine learning and deep learning

Deep learning has made incredible progress in the last decade and is currently booming. It is a revolution in the era of artificial intelligence and the following reasons explain its success :

- The creation of large annotated databases (such as ImageNet (Deng et al., 2009)) has allowed having enough supervised learning data to achieve good results. Indeed, the performance of deep learning approaches increased with the amount of training data whereas with the traditional machine learning approaches their performance stabilizes.
- The continuous development of the GPU technology allows the computation of millions of parameters at the same iteration. This technology has enabled to train deep network within a reasonable time.

The first popular deep network for 2D image classification was made of 62.3 millions of parameters. AlexNet (Krizhevsky et al., 2012) famously won the 2012 ImageNet LSVRC-2012 competition by a large margin (15.3% VS 26.25% (second place) error rates). The model had been trained for 5 to 6 days on the ImageNet dataset which at that moment contained over 1.3 million annotated images from a total of over 1000 categories. The ImageNet dataset now contains 15 million images for 22000 categories.

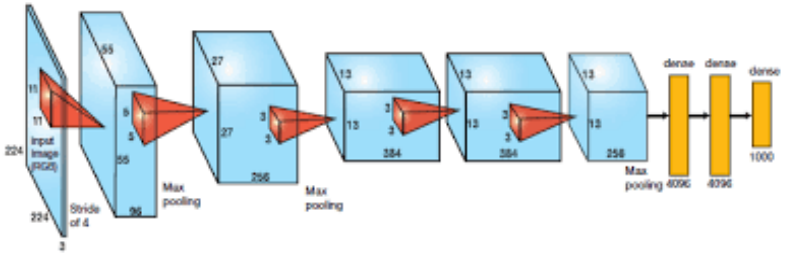


Figure 8: A visual represation of the Alexnet model

There are different research topics in the field of deep learning. In this chapter, we will focus on the models used to process images: the convolutional neural networks (CNN) which are definitely the leading research topic in deep learning.

A convolution is a small filter operation that is passed over the image. The processed image after a convolution contains weighted sums of small sub-areas of the original image. The figure below shows an example of a simple convolution operation (horizontal edge detection filter in this case).

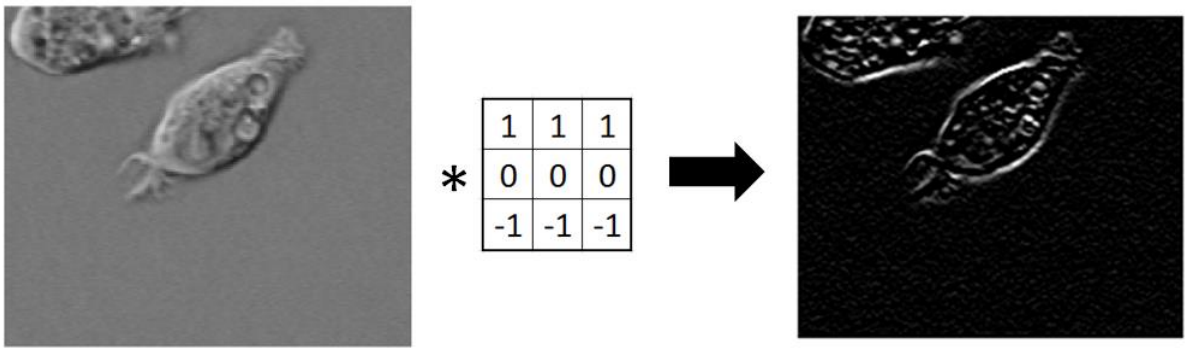


Figure 9: Example of a convolution operation (Edge detection)

A CNN replaces the normal neurons by filtering operations (convolution). A convolution is followed by an activation where the activation adds some non-linearity in the models. Standard activation like ReLu has the advantage of vanishing large gradients while increasing the training efficiency. After few layers of convolution and pooling operation, the image is transformed into a set of features. Pooling operation helps to reduce the spatial dimensionality of a layer and also has a gainful local spatial invariance property and do not need any parameter to be tuned.

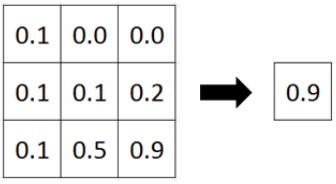


Figure 10: Example of a pooling operation (in this case, max-pooling)

In most cases, the network is built for a classification purpose and hence the final layer is often a classifier. In order to have a probability for each class, a softmax layer is often placed before the classifier. Fully connected (or Dense) layers connect every neuron of the previous layers to neurons of the next layer, and thus destroy the spatial organization of the input. This layer is used to classify the images using all the computed features.

Many different network architectures have been proposed in the last decade. Indeed the more layers you use, the more difficult it will be to train them. The challenge in network architecture design is to avoid overfitting issues while training those millions of parameters. Some networks have achieved good accuracy by using 152 layers! They have used a residual connection for this purpose. We refer readers to the literature for a more detailed review of the recent network architecture (Alom et al., 2018; Canziani et al., 2016; Real et al., 2018).

A key ingredient to reach high accuracy of the traits extracted is the training process. The parameters of the model are initialized randomly and are then updated with stochastic gradient descent. Gradient descent is an optimization algorithm which only requires a cost function to be defined. This process is repeated for mini-batches and stops when a criterion that can be computed on a validation dataset is reached. An optimizer needs a learning rate schedule and is also implemented with a Nesterov momentum. The selection of optimizers and optimization strategies is an active area of research (Martens and Grosse, 2015).

Training a network also involves some regularization. For example, in Dropout (Srivastava et al., 2014) a randomly selected subset of activations are set to zero within a layer which reduces the network dependency to individual neurons and gives a more redundant and robust representation of the data.

The features contained in the bottleneck of a trained CNN can be used for another application which is known as transfer learning. Training a model from scratch on large scale dataset (ImageNet, Coco) takes days or weeks on a standard GPU. CNNs are therefore often already trained (initialized) on a publicly available dataset. It is then possible to replace the last layer by another classifier and then train the model on a smaller dataset. This strategy which is known as fine-tuning is really useful when a small annotated dataset is available. This is currently the case in the agriculture field, especially in phenotyping. This is a challenging context that will need to focus on.

Today CNN in computer vision are applied with impressive accuracy and over a range of applications: classification, localization, segmentation, regression. The transfer of these technologies can add a lot of value to plant breeding.

1.4.2 Deep learning applied to phenotyping under field conditions

Deep learning for phenotyping has already been demonstrated under controlled conditions. Thanks to the initiative of the IPPN imaging group (Tsaftaris and Scharr, 2018), the leaf counting and segmentation challenges and the associated dataset have gathered the attention of many people in recent years. Not surprisingly, deep learning and state of the art algorithms have outperformed more classical methods (Ubbens and Stavness, 2017). It is interesting to note that the best approaches now use synthetic images generated as a training data augmentation (Giuffrida et al., 2017) strategy. This may be difficult to reproduce elsewhere because of the difficulty to simulate images of crops with realistic rendering and texture. However adding images

with variability of the shape of the object targeted is expected to increase the scalability and robustness of the CNN models (Geirhos et al., 2018).

No study using deep learning for images acquired under field conditions were found before 2016 (Kamilaris and Prenafeta-Boldú, 2018). Deep learning methods were first used in the field for weeding activities (McCool et al., 2017). (Mortensen et al., 2016) use a RGB (Red-Green-Blue) + NIR (Near infrared) sensor, and these four channel images were used to train their models that detect vegetation. The vegetation was then classified between crop and weeds with impressive accuracy. CNN models have tackled the problem of image classification for several years and have been recently applied to plants (Lim and Sugita, 2018), leaves (Ghosal et al., 2017), species (Botella et al., 2018; Joly et al., 2018) classification as well. The PlantNet experience and results suggest that state of the art deep learning is now close to top-level human expertise(Joly et al., 2018). The challenging problems of fruits counting have also been addressed by (Chen et al., 2017; Sa et al., 2016).

There is an exponential use of deep learning in crop production. An important component of this adoption is the development of robotic and autonomous vehicle and the internet of things technologies. It allows collecting millions of images under field conditions. The CNN models used in crop production and plant phenotyping were originally developed within the computer vision community where they were benchmarked on large datasets. The transfer of the technology has become easy with the general open source initiative. This is the case with the open-source libraries Tensorflow and Keras developed within Google and PyTorch releases by Facebook. In most cases, new research work in this domain has its own repository and is in theory reproducible. The same CNN architecture can be used for a large range of applications, and indeed the need for the design of specific handcrafted features targeted for a precise application is eliminated. A new trick concerning layer design or optimization strategy could then have a huge impact on a wide range of applications. For example, this was instrumental in the cases of the inception layer, residual network, deconvolution layer and optimizers (Adam). With these initiatives and an open-source environment, the deep learning era is developing fast. Recently (Zoph et al., 2018; Zoph and Le, 2016) designed the architecture of a neural network by automatically exploring several architectures. Their model has been tested on another dataset and has achieved a better accuracy than the state of the art, while preserving a reasonable computation cost. This change of paradigm emphasizes the importance and value to the dataset and the computer power as illustrated in Figure 11.

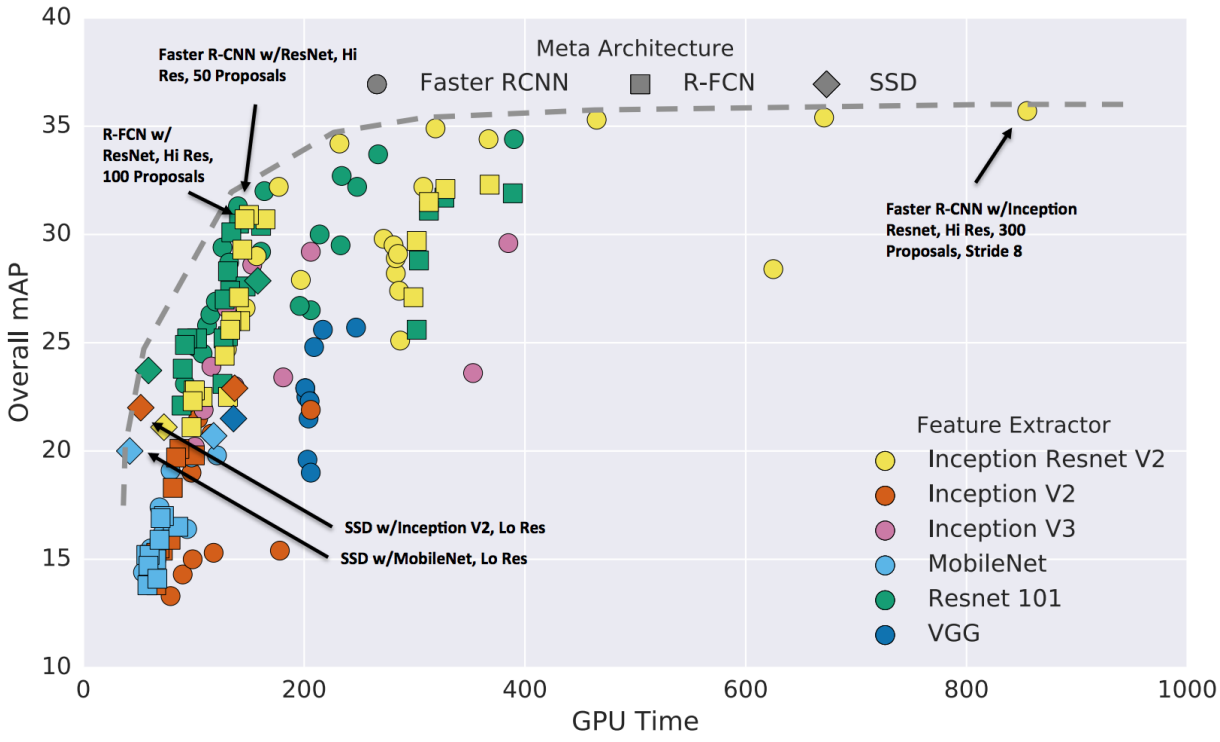


Figure 11. Evolution of the performances of the deep learning models as a function of the GPU requirements. Source : (Huang et al., 2016)

In the plant phenotyping vision community, it is not yet possible to have a large annotated dataset under field conditions. The labor-intensive task of image annotation and the lack of crowd-funded solutions explain this problem. The operational agriculture sector, most of the activities are conducted by the private sector which makes the development of a shared accessible field phenotyping dataset difficult owing to privacy concerns. In phenotyping, we specially require a large dataset as the models are supposed to be able to generalize and to be accurate over a wide range of cultivars and conditions.

1.4.3 Main challenges to address for efficient use of deep learning for trait estimation from images acquired under field conditions

To be utilized by breeders, the images have to be transformed into a functional trait. In this section, we attempt to describe the main challenges that have to be solved and the prerequisites for the deployment of deep learning in a plant breeding trial.

A typical trait which is of interest to the breeders is the number of plants or organs per unit area. From the deep learning point of view, different strategies could be used. One method could be to train a CNN to identify each plant in an image by an object detection algorithm. Another method could be to design an algorithm which would directly derive the number of plants in an image using a regression layer (continuous output) with a square loss cost function. To get an estimate of the green area index or fCover, the algorithms could directly classify each pixel into corresponding classes using some deconvolution layers. An alternative algorithm would be to first detect the plants in the image and then generate a segmentation mask for each instance of the plants. It is important to know which is the best strategy for the phenotyping applications we considered. The choice of the algorithms should be made depending on the type of crop

(wheat, maize) and phenological stages (early stage measurement, earing stage …), on their performances, robustness as well as the amount of efforts required for image labelling.

In the field phenotyping community, deep learning is just emerging and it will certainly be adopted widely and rapidly. Given this context, it is important to standardize the methodology and the sensors used. Indeed, a deep learning model that was trained to detect a disease with high-resolution RGB images will not be able to do the same with images from a multispectral camera with a degraded spatial resolution and spectral bands that are different: transfer learning will not be directly possible, and the model will have to be trained over a new dataset. A similar problem will be encountered when using RGB images acquired with different spatial resolutions or even under a large range of variation in the illumination conditions and for variability regarding in the phenological stages. CNNs methods are known to be sensitive to the spatial resolution of the scene. Hence, a model trained to detect objects of size 50*50 only will fail to detect smaller objects. A bias introduced by one of these changes can have a direct impact on the heritability and subsequently on the genetic analysis and selection. It is thus important to test the models on a different dataset and inform users of their respective limitation and robustness.

Because the data is the key to good performances, training the models with data augmentation strategy has been usually associated with an increase of the accuracy. Basic data augmentation strategies include flipping, rotating, zooming and histogram-based method. Domain style transfer algorithms with the use of Generative Adversarial Networks (GAN) have also been recently used for this purpose with promising results (Sankaranarayanan et al., 2018). The idea behind GAN is to transfer the image from the source domain to the target domain and to train the model with the source ground truth with the source and target data. More details on GAN and style transfer are available in this paper (Zhu et al., 2017). The use of plant models and 3D canopy structure models could also potentially help to solve the data diversity problems and lack of annotated images. The work that needs to be done here is really similar to the one in the autonomous driving field with the Kitti (Gaidon et al., 2016) or game engines dataset. The challenge here is to reduce or fill the gap between synthetic and real images. Many authors have already attempted to solve this problem using GANs (Alhaija et al., 2017; Mahmood et al., 2017; Nogue et al., 2018; Shrivastava et al., 2017; Zhu et al., 2017).

The fine-tuning and transfer learning techniques have the potential to take advantage of large-scale datasets. It is interesting to study if the off-the-shelf pre-trained CNN features will achieve better results than transfer learning. This will probably depend on the size and diversity of the datasets considered. It may also be important to freeze some layers during the transfer learning process. Further, key questions are still pending: is it better to build a generic model that can be applied over several species or develop specific models for each crop? How to deal with the uneven class balance? Should we perform domain adaptation for each acquisition?

1.5 Objectives and organization of the study

The aim of this study is to estimate key structural traits from high-resolution images acquired with high throughput phenotyping systems which are now operational. It will contribute towards unlocking the bottleneck of phenotyping data interpretation which is now widely recognized.

We propose to focus on a few structural traits that were up to now difficult to measure at high throughput. The spatial resolution of the acquired images varied from fraction of millimeter to 1 cm. The study was developed and evaluated on wheat, and can be easily transferred to others

crops. For each of these traits, we evaluate the performance of the high-throughput methods with a reference method. The acquisitions were performed using phenotyping platforms. It was thus also possible to compute the broad-sense heritability of the estimated traits. The traits are also tentatively used to derive biophysical parameters such as the biomass. The chapters are therefore organized as listed below, following the several traits considered:

- ***Plant height*** derived from high resolution images taken from UAVs: With the development of unmanned aerial vehicle platforms, it is now possible to access this information from a high throughput platform. In our first chapter, we evaluate this technology for the estimation of wheat plant height from the 3D point could derived from the Structure from Motion algorithm. Frequent LiDAR measurements were used as reference. The derived traits and workflow were evaluated in the context of plant breeding. This chapter corresponds to a paper that is currently published:

    - **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.* 8. doi:10.3389/fpls.2017.02002.

- ***Wheat ear density estimation*** is currently labor intensive and prone to error due to sampling problems while using the low throughput invasive techniques. The feasibility of deep learning approaches for this application was evaluated in the second chapter. Robustness and limitation of the method was evaluated in the context of high throughput phenotyping. This chapter corresponds to a paper that is currently published:

    - **Madec, S.**, Jin, X., Lu, H., De Solan, B., Liu, S., Duyme, F., et al. (2019). Ear density estimation from high resolution RGB imagery using deep learning technique. *Agricultural and Forest Meteorology* 264, 225–234. doi:10.1016/j.agrformet.2018.10.013.

- The methodology developed in the second chapter is tentatively transferred to the ***estimation of stem density*** after harvest. The stem diameter was also estimated. These two traits combined with the height were finally used to access the biomass. This chapter corresponds to a paper that has been recently accepted.

    - Jin, X., **Madec, S.**, Dutartre, D., Solan, B. de, Comar, A., and press, F. B. High-throughput measurements of stem characteristics to estimate ear density and above ground biomass. Available at: https://spj.sciencemag.org/plantphenomics/aip/4820305/

Finally, a conclusion is proposed to highlight the main findings and propose few improvements and application to a range of other traits.

References

Aasen, H. (2017). State-of-the-art in UAV remote sensing survey-First insights into applications of UAV sensing systems. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* 42, 1–4.

Aasen, H., and Bolten, A. (2018). Multi-temporal high-resolution imaging spectroscopy with hyperspectral 2D imagers – From theory to application. *Remote Sens. Environ.* 205, 374–389. doi:10.1016/j.rse.2017.10.043.

Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., et al. (2017). Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry. *Remote Sens.* 9, 1110. doi:10.3390/rs9111110.

Alhaija, H. A., Mustikovela, S. K., Mescheder, L., Geiger, A., and Rother, C. (2017). Augmented reality meets deep learning for car instance segmentation in urban scenes. in *British Machine Vision Conference*, 2.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Hasan, M., Van Esesn, B. C., et al. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. *ArXiv Prepr. ArXiv180301164*.

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, eaar7191.

Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., et al. (2015). Rising temperatures reduce global wheat production. *Nat. Clim. Change* 5, 143.

Bai, G., Ge, Y., Hussain, W., Baenziger, P. S., and Graef, G. (2016). A multi-sensor system for high throughput field phenotyping in soybean and wheat breeding. *Comput. Electron. Agric.* 128, 181–192.

Baulcombe, D., Crute, I., Davies, B., Dunwell, J., Gale, M., Jones, J., et al. (2009). Reaping the benefits: science and the sustainable intensification of global agriculture. The Royal Society Available at: http://royalsociety.org/policy/publications/2009/reaping-benefits/ [Accessed October 18, 2018].

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2012). GenBank. *Nucleic Acids Res.* 41, D36–D42.

Berni, J. A. J., Zarco-Tejada, P. J., Suarez, L., and Fereres, E. (2009). Thermal and Narrowband Multispectral Remote Sensing for Vegetation Monitoring From an Unmanned Aerial Vehicle. *IEEE Trans. Geosci. Remote Sens.* 47, 722–738. doi:10.1109/TGRS.2008.2010457.

Bongaarts, J., and Watkins, S. C. (1996). Social interactions and contemporary fertility transitions. *Popul. Dev. Rev.* Available at: http://link.galegroup.com/apps/doc/A19140226/AONE?sid=googlescholar [Accessed April 17, 2019].

Botella, C., Joly, A., Bonnet, P., Monestiez, P., and Munoz, F. (2018). Species distribution modeling based on the automated identification of citizen observations. *Appl. Plant Sci.* 6, e1029.

Burud, I., Lange, G., Lillemo, M., Bleken, E., Grimstad, L., and From, P. al J. (2017). Exploring robots and uavs as phenotyping tools in plant breeding. *IFAC-Pap.* 50, 11479–11484.

Cabrera Bosquet, L., Brichet, N., Fournier, C., Grau, A., Mineau, J., Negre, V., et al. (2015). PHENOARCH, a multiscale phenotyping platform for plant architecture, growth rate, water use efficiency and radiation use efficiency. *2015 Recent Prog. Drought Toler. Genet. Model. Montp. FRA 2015-06-08-2015-06-09 81*. Available at: http://agris.fao.org/agris-search/search.do?recordID=LV2016028780 [Accessed January 7, 2019].

Campbell, J. B., and Wynne, R. H. (2011). *Introduction to remote sensing*. Guilford Press.

Canziani, A., Paszke, A., and Culurciello, E. (2016). An analysis of deep neural network models for practical applications. *ArXiv Prepr. ArXiv160507678*.

Chen, S. W., Shivakumar, S. S., Dcunha, S., Das, J., Okon, E., Qu, C., et al. (2017). Counting Apples and Oranges With Deep Learning: A Data-Driven Approach. *IEEE Robot. Autom. Lett.* 2, 781–788.

Crow, J. F. (1998). 90 years ago: the beginning of hybrid maize. *Genetics* 148, 923–928.

Dalal, N., and Triggs, B. (2005). Histograms of oriented gradients for human detection. in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 886–893 vol. 1. doi:10.1109/CVPR.2005.177.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. in *2009 IEEE conference on computer vision and pattern recognition* (Ieee), 248–255.

Dwivedi, S. L., Scheben, A., Edwards, D., Spillane, C., and Ortiz, R. (2017). Assessing and exploiting functional diversity in germplasm pools to enhance abiotic stress adaptation and yield in cereals and food legumes. *Front. Plant Sci.* 8, 1461.

Evenson, R. E., and Gollin, D. (2003). Assessing the impact of the Green Revolution, 1960 to 2000. *science* 300, 758–762.

Foley, J. A., Ramankutty, N., Brauman, K. A., Cassidy, E. S., Gerber, J. S., Johnston, M., et al. (2011). Solutions for a cultivated planet. *Nature* 478, 337–342. doi:10.1038/nature10452.

Frumin, S. (2017). Invasion Biology Analysis in Archaeobotany: Philistine Culture at Tell eṣ-Ṣâfi/Gath as a Case Study. *PhD Diss Bar-Ilan Univ.*

Gaidon, A., Wang, Q., Cabon, Y., and Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4340–4349.

Gao, B.-C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* 58, 257–266.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv Prepr. ArXiv181112231*.

Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., and Sarkar, S. (2017). Interpretable Deep Learning applied to Plant Stress Phenotyping. *ArXiv Prepr. ArXiv171008619*.

Giuffrida, M. V., Scharr, H., and Tsaftaris, S. A. (2017). ARIGAN: Synthetic arabidopsis plants using generative adversarial network. in *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), Venice, Italy*, 22–29.

Grimstad, L., and From, P. J. (2017). The Thorvald II Agricultural Robotic System. *Robotics* 6, 24. doi:10.3390/robotics6040024.

Guillaume, A. (2018). Sortie du rapport du Giec sur le réchauffement planétaire de 1, 5° C. *La Météorologie*.

Guo, W., Fukatsu, T., and Ninomiya, S. (2015). Automated characterization of flowering dynamics in rice using field-acquired time-series RGB images. *Plant Methods* 11, 7.

Habier, D., Fernando, R. L., and Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397.

Hallauer, A. R., Miranda Filho, J. de, and Carena, M. J. (2010). "Breeding plans," in *Quantitative genetics in maize breeding* (Springer), 577–653.

Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

He, W., Yang, J. Y., Drury, C. F., Smith, W. N., Grant, B. B., He, P., et al. (2018). Estimating the impacts of climate change on crop yields and N2O emissions for conventional and no-tillage in Southwestern Ontario, Canada. *Agric. Syst.* 159, 187–198.

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., et al. (2016). Speed/accuracy trade-offs for modern convolutional object detectors. *ArXiv161110012 Cs*. Available at: http://arxiv.org/abs/1611.10012.

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Brief. Funct. Genomics* 9, 166–177.

Jarecki, M., Grant, B., Smith, W., Deen, B., Drury, C., VanderZaag, A., et al. (2018). Long-term Trends in Corn Yields and Soil Carbon under Diversified Crop Rotations. *J. Environ. Qual.*

Jay, S., Baret, F., Dutartre, D., Malatesta, G., Héno, S., Comar, A., et al. (2018). Exploiting the centimeter resolution of UAV multispectral imagery to improve remote-sensing estimates of canopy structure and biochemistry in sugar beet crops. *Remote Sens. Environ.* doi:10.1016/j.rse.2018.09.011.

Joly, A., Goëau, H., Botella, C., Glotin, H., Bonnet, P., Vellinga, W.-P., et al. (2018). Overview of lifeclef 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of ai. in *International Conference of the Cross-Language Evaluation Forum for European Languages* (Springer), 247–266.

Kamilaris, A., and Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Comput. Electron. Agric.* 147, 70–90.

Kirchgessner, N., Liebisch, F., Yu, K., Pfeifer, J., Friedli, M., Hund, A., et al. (2017). The ETH field phenotyping platform FIP: a cable-suspended multi-sensor system. *Funct. Plant Biol.* 44, 154–168.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. in *Advances in neural information processing systems*, 1097–1105. Available at: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

Leakey, A. D., Ainsworth, E. A., Bernacchi, C. J., Rogers, A., Long, S. P., and Ort, D. R. (2009). Elevated CO2 effects on plant carbon, nitrogen, and water relations: six important lessons from FACE. *J. Exp. Bot.* 60, 2859–2876.

Licker, R., Johnston, M., Foley, J. A., Barford, C., Kucharik, C. J., Monfreda, C., et al. (2010). Mind the gap: how do climate and agricultural management explain the 'yield gap'of croplands around the world? *Glob. Ecol. Biogeogr.* 19, 769–782.

Lillesaeter, O. (1982). Spectral reflectance of partly transmitting leaves: laboratory measurements and mathematical modeling. *Remote Sens. Environ.* 12, 247–254.

Lim, L., and Sugita, G. (2018). What's That Plant? WTPlant is a Deep Learning System to Identify Plants in Natural Images.

M Perez-de-Castro, A., Vilanova, S., Cañizares, J., Pascual, L., M Blanca, J., J Diez, M., et al. (2012). Application of genomic tools in plant breeding. *Curr. Genomics* 13, 179–195.

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.* 8. doi:10.3389/fpls.2017.02002.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nat. News* 456, 18–21.

Mahmood, F., Chen, R., and Durr, N. J. (2017). Unsupervised reverse domain adaption for synthetic medical images via adversarial training. *ArXiv Prepr. ArXiv171106606*.

Mankins, J. C. (1995). TECHNOLOGY READINESS LEVELS.

Martens, J., and Grosse, R. (2015). Optimizing Neural Networks with Kronecker-factored Approximate Curvature. *ArXiv150305671 Cs Stat*. Available at: http://arxiv.org/abs/1503.05671 [Accessed January 25, 2019].

McCool, C., Perez, T., and Upcroft, B. (2017). Mixtures of lightweight deep convolutional neural networks: applied to agricultural robotics. *IEEE Robot. Autom. Lett.* 2, 1344–1351.

Messina, C. D., Technow, F., Tang, T., Totir, R., Gho, C., and Cooper, M. (2018). Leveraging biological insight and environmental variation to improve phenotypic prediction: Integrating crop growth models (CGM) with whole genome prediction (WGP). *Eur. J. Agron.*

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nat. Rev. Genet.* 11, 31.

Millet, E., Welcker, C., Kruijer, W., Negro, S., Nicolas, S., Praud, S., et al. (2016). Genome-wide analysis of yield in Europe: allelic effects as functions of drought and heat scenarios. *Plant Physiol.*, pp–00621.

Mortensen, A. K., Dyrmann, M., Karstoft, H., Jørgensen, R. N., and Gislum, R. (2016). Semantic segmentation of mixed crops using deep convolutional neural network. in *CIGR-AgEng Conference, 26-29 June 2016, Aarhus, Denmark. Abstracts and Full papers* (Organising Committee, CIGR 2016), 1–6.

Mousavi-Derazmahalleh, M., Bayer, P. E., Hane, J. K., Valliyodan, B., Nguyen, H. T., Nelson, M. N., et al. (2019). Adapting legume crops to climate change using genomic approaches. *Plant Cell Environ.* 42, 6–19.

Neumann, K., Verburg, P. H., Stehfest, E., and Müller, C. (2010). The yield gap of global grain production: A spatial analysis. *Agric. Syst.* 103, 316–326.

Nogue, F. C., Huie, A., and Dasgupta, S. (2018). Object Detection using Domain Randomization and Generative Adversarial Refinement of Synthetic Images. *ArXiv Prepr. ArXiv180511778*.

Ousseini, I. S., Bakasso, Y., Kane, N. A., Couderc, M., Zekraoui, L., Mariac, C., et al. (2017). Myosin XI is associated with fitness and adaptation to aridity in wild pearl millet. *Heredity* 119, 88.

Parent, B., Leclere, M., Lacube, S., Semenov, M. A., Welcker, C., Martre, P., et al. (2018). Maize yields over Europe may increase in spite of climate change, with an appropriate use of the genetic variability of flowering time. *Proc. Natl. Acad. Sci.* 115, 10642–10647. doi:10.1073/pnas.1720716115.

Pereyra-Irujo, G. A., Gasco, E. D., Peirone, L. S., and Aguirrezábal, L. A. (2012). GlyPh: a low-cost platform for phenotyping plant growth and water use. *Funct. Plant Biol.* 39, 905–913.

Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161, 209–228.

Porfirio, L. L., Newth, D., Finnigan, J. J., and Cai, Y. (2018). Economic shifts in agricultural production and trade due to climate change. *Palgrave Commun.* 4, 111.

Porter, J. R., and Semenov, M. A. (2005). Crop responses to climatic variation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 2021–2035.

Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. (2018). Regularized evolution for image classifier architecture search. *ArXiv Prepr. ArXiv180201548*.

Remondino, F., and El-Hakim, S. (2006). Image-based 3D modeling: a review. Available at: http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=8913373 [Accessed March 22, 2016].

Reynolds, D., Baret, F., Welcker, C., Bostrom, A., Ball, J., Cellini, F., et al. (2018). What is cost-efficient phenotyping? Optimizing costs for different scenarios. *Plant Sci.*

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., and McCool, C. (2016). Deepfruits: A fruit detection system using deep neural networks. *Sensors* 16, 1222.

Sadeghi-Tehran, P., Sabermanesh, K., Virlet, N., and Hawkesford, M. J. (2017a). Automated Method to Determine Two Critical Growth Stages of Wheat: Heading and Flowering. *Front. Plant Sci.* 8. doi:10.3389/fpls.2017.00252.

Sadeghi-Tehran, P., Virlet, N., Sabermanesh, K., and Hawkesford, M. J. (2017b). Multi-feature machine learning model for automatic segmentation of green fractional vegetation cover for high-throughput field phenotyping. *Plant Methods* 13, 103. doi:10.1186/s13007-017-0253-8.

Sankaranarayanan, S., Balaji, Y., Castillo, C. D., and Chellappa, R. (2018). Generate to Adapt: Aligning Domains Using Generative Adversarial Networks. in, 8503–8512. Available at: http://openaccess.thecvf.com/content_cvpr_2018/html/Sankaranarayanan_Generate_to_Adapt_CVPR_2018_paper.html [Accessed April 17, 2019].

Sasaki, T. (2005). The map-based sequence of the rice genome. *Nature* 436, 793.

Schlenker, W., and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proc. Natl. Acad. Sci.* 106, 15594–15598.

Shakoor, N., Lee, S., and Mockler, T. C. (2017). High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Curr. Opin. Plant Biol.* 38, 184–192.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2017). Learning from Simulated and Unsupervised Images through Adversarial Training. in *CVPR*, 5.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Tester, M., and Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. *Science* 327, 818–822.

Tsaftaris, S. A., and Scharr, H. (2018). Sharing the Right Data Right: A Symbiosis with Machine Learning. *Trends Plant Sci.*

Ubbens, J. R., and Stavness, I. (2017). Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. *Front. Plant Sci.* 8. doi:10.3389/fpls.2017.01190.

Underwood, J., Wendel, A., Schofield, B., McMurray, L., and Kimber, R. (2017). Efficient in-field plant phenomics for row-crops with an autonomous ground vehicle. *J. Field Robot.* 34, 1061–1083. doi:10.1002/rob.21728.

Verger, A., Vigneau, N., Chéron, C., Gilliot, J.-M., Comar, A., and Baret, F. (2014). Green area index from an unmanned aerial system over wheat and rapeseed crops. *Remote Sens. Environ.* 152, 654–664.

Virlet, N., Sabermanesh, K., Sadeghi-Tehran, P., and Hawkesford, M. J. (2016). Field Scanalyzer: An automated robotic field phenotyping platform for detailed crop monitoring. *Funct. Plant Biol.*, NULL. doi:10.1071/FP16163.

Wang, E., Martre, P., Zhao, Z., Ewert, F., Maiorano, A., Rötter, R. P., et al. (2017). The uncertainty of crop yield projections is reduced by improved temperature response functions. *Nat. Plants* 3, 17102.

Wheeler, T., and Von Braun, J. (2013). Climate change impacts on global food security. *Science* 341, 508–513.

White, J. W., and Conley, M. M. (2013). A flexible, low-cost cart for proximal sensing. *Crop Sci.* 53, 1646–1649.

Yang, G., Liu, J., Zhao, C., Li, Z., Huang, Y., Yu, H., et al. (2017). Unmanned Aerial Vehicle Remote Sensing for Field-Based Crop Phenotyping: Current Status and Perspectives. *Front. Plant Sci.* 8. doi:10.3389/fpls.2017.01111.

Zandalinas, S. I., Mittler, R., Balfagón, D., Arbona, V., and Gómez-Cadenas, A. (2018). Plant adaptations to the combination of drought and high temperatures. *Physiol. Plant.* 162, 2–12.

Zhong, S., Dekkers, J. C., Fernando, R. L., and Jannink, J.-L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. in *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Ziska, L. H., and Bunce, J. A. (2007). Predicting the impact of changing CO2 on crop yields: some thoughts on food. *New Phytol.* 175, 607–618.

Zohary, D., Hopf, M., and Weiss, E. (2012). *Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin*. Oxford University Press on Demand.

Zoph, B., and Le, Q. V. (2016). Neural architecture search with reinforcement learning. *ArXiv Prepr. ArXiv161101578*.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8697–8710.

2 High-Throughput Phenotyping of Plant Height: Comparing Unmanned Aerial Vehicles and Ground LiDAR Estimates

Plant height is an attractive trait for breeders who currently measure using rulers. This approach is low throughput and prone to errors. With the development of UAV and ground vehicle platforms, it is now possible to access this information at high throughput. The recent adoption of UAV makes the collection of high-resolution images from various angles affordable. By applying the structure from motion method, it is possible to derive a dense 3D cloud. In this chapter we explore the potential of this technology for plant height phenotyping. The acquisition campaigns were realized in 2015. Plant height was estimated for different dates during the growing season. We have investigated many critical parameters for the computation of height from both RGB images and LiDAR point cloud. Wheat plant height is known to have a high heritability, and this was further validated by the high repeatability computed.

frontiers in Plant Science

METHODS
published: 27 November 2017
doi: 10.3389/fpls.2017.02002

High-Throughput Phenotyping of Plant Height: Comparing Unmanned Aerial Vehicles and Ground LiDAR Estimates

Simon Madec[1], Fred Baret[1], Benoît de Solan[2], Samuel Thomas[2], Dan Dutartre[3], Stéphane Jezequel[2], Matthieu Hemmerlé[3], Gallian Colombeau[1] and Alexis Comar[3]*

[1] INRA, UMR EMMAH, Avignon, France, [2] ARVALIS – Institut du végétal, Avignon, France, [3] HIPHEN, Avignon, France

The capacity of LiDAR and Unmanned Aerial Vehicles (UAVs) to provide plant height estimates as a high-throughput plant phenotyping trait was explored. An experiment over wheat genotypes conducted under well watered and water stress modalities was conducted. Frequent LiDAR measurements were performed along the growth cycle using a phénomobile unmanned ground vehicle. UAV equipped with a high resolution RGB camera was flying the experiment several times to retrieve the digital surface model from structure from motion techniques. Both techniques provide a 3D dense point cloud from which the plant height can be estimated. Plant height first defined as the *z*-value for which 99.5% of the points of the dense cloud are below. This provides good consistency with manual measurements of plant height (RMSE = 3.5 cm) while minimizing the variability along each microplot. Results show that LiDAR and structure from motion plant height values are always consistent. However, a slight underestimation is observed for structure from motion techniques, in relation with the coarser spatial resolution of UAV imagery and the limited penetration capacity of structure from motion as compared to LiDAR. Very high heritability values ($H^2 > 0.90$) were found for both techniques when lodging was not present. The dynamics of plant height shows that it carries pertinent information regarding the period and magnitude of the plant stress. Further, the date when the maximum plant height is reached was found to be very heritable ($H^2 > 0.88$) and a good proxy of the flowering stage. Finally, the capacity of plant height as a proxy for total above ground biomass and yield is discussed.

Keywords: plant height, high throughput, unmanned aerial vehicles, dense point cloud, LiDAR, phenotyping, broad-sense heritability

OPEN ACCESS

Edited by:
Yann Guédon, Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), France

Reviewed by:
Barbara George-Jaeggli, The University of Queensland, Australia
Andreas Bolten, University of Cologne, Germany

****Correspondence:***
Simon Madec
simon.madec@inra.fr

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *22 August 2017*
Accepted: *09 November 2017*
Published: *27 November 2017*

Citation:
Madec S, Baret F, de Solan B, Thomas S, Dutartre D, Jezequel S, Hemmerlé M, Colombeau G and Comar A (2017) High-Throughput Phenotyping of Plant Height: Comparing Unmanned Aerial Vehicles and Ground LiDAR Estimates. Front. Plant Sci. 8:2002. doi: 10.3389/fpls.2017.02002

INTRODUCTION

Plant height is recognized as a good proxy of biomass (Yin et al., 2011; Bendig et al., 2014; Ota et al., 2015; Tilly et al., 2015). Stem height that defines plant height appears to be sensitive to the stresses subjected by the crop (Rawson and Evans, 1971). It is also one of the input of models used to evaluate water stress (Blonquist et al., 2009). Plant height is known to make the crop

Abbreviations: DaS, day after sowing; $D_{\text{flowering}}$, date of flowering; $D_{\max(\text{PH})}$, date of maximum height; GDD, growth degree day; GSD, ground sampling distance; H^2, broad sense heritability; LiDAR, light detection and ranging; RMSE, root mean square error; R_p, rank percentile; UAV, unmanned aerial vehicle; WS, water stress modality; WW, well watered modality.

more sensitive to lodging (Berry et al., 2003). Plant height appears thus a highly appealing trait for plant breeders within phenotyping experiments, particularly under natural field conditions. Current methods based on manual evaluation with a ruler on a limited sample size for each microplot are labor intensive, low throughput and prone to errors in the sampling, ruler adjustment, reading and recording the data. Alternative methods have been developed either from LiDAR (Light Detection And Range) often called laser scanning (Hoffmeister et al., 2015), ultrasonic sensors also called sonar (Turner et al., 2007), or from depth camera also called time of flight camera (Chéné et al., 2012; Schima et al., 2016), and finally from RGB high resolution imagery associated with structure from motion algorithms. Depth cameras are limited to close range applications (Schima et al., 2016). Ultrasonic systems are considered as a relatively low-cost solution and user friendly. However, LiDAR measurements have been generally preferred for their increased spatial resolution, higher throughput and independency from air temperature and wind (Tumbo et al., 2002; Escolà et al., 2011; Llorens et al., 2011). LiDAR scanning can be performed from the ground with terrestrial laser scanner). However, terrestrial laser scanners are conical scanners that are well suited for vertically developed objects such as buildings or forests. Their application to crops with limited vertical extent and a canopy volume densely populated by leaves and stems or other organs appears limited (Zhang and Grift, 2012; Bareth et al., 2016): the system needs to be moved over a high number of places for large phenotyping platforms. Further, the several microplots may be seen from different distances and angles with impact on the spatial resolution and associated bias introduced between microplots. It seems therefore preferable to observe crops from near nadir directions.

Several manned or semi-autonomous GPS (Geo-Positioning System) navigated vehicles, have been developed in the recent years where vertically scanning LiDARs have been setup. LiDARs provide a full description of the profile of interception, either with single echo (Lisein et al., 2013; James and Robson, 2014) when the resolution is fine enough, or with full wave form systems (Mallet and Bretar, 2009) or an approximation of it with multi-echo systems (Moras et al., 2010). Because of the penetration of the laser beam into most canopies, nadir looking LiDAR techniques provide at the same time the digital surface model corresponding to the top envelope of the crop (called also crop surface model) and the elevation of the background surface called the digital terrain model. Plant height is then simply computed as the difference between the digital surface model and the digital terrain model. Accuracy on plant height measurement using such LiDAR techniques were reported to be better than a few centimeters (Deery et al., 2014; Virlet et al., 2016). Because of their high accuracy, their independency from the illumination conditions and therefore their high repeatability, these LiDAR based techniques are expected to be more accurate than traditional manual height ruler measurements in the field.

RGB image-based retrieval of crop height remains, however, the most widely used approach (Bendig et al., 2013) because of its low cost and high versatility (Remondino and El-Hakim, 2006). Further, the advances in sensors (smaller, lighter and cheaper, increased resolution and sensitivity) and improvements in computer performances along with advances in algorithms have contributed to the recent success of such techniques (Remondino et al., 2014). The 3D dense point clouds are generated by using a large set of high resolution overlapping images. They are processed using structure from motion algorithms implemented in either commercial software (Smith et al., 2015) such as pix4d[1], Agisoft photoscan[2] or in open-source software including micmac (MicMac, IGN, France) or Bundler (Snavely et al., 2006). Nevertheless, accurate retrieval of 3D characteristics of the canopy from structure from motion algorithms requires careful completion of the image acquisition that should provide enough view directions for each point of the scene and with crisp high resolution images to identify the tie points used for the 3D reconstruction of the surface (Turner et al., 2014; Smith et al., 2015). Several factors will thus influence the quality and accuracy of the dense point cloud, including flight configuration (altitude, speed, frequency of acquisitions, trajectory design and sensor orientation) camera setting (resolution, field of view, image quality), illumination and wind conditions, the distribution of ground control points as well as the parameters used to run the structure from motion algorithm (Dandois and Ellis, 2013; Remondino et al., 2014).

Because of the spatial resolution of the images used for the structure from motion algorithms and more importantly because of the occlusions observed when a single point is to be seen from two distinct directions, structure from motion algorithms do not penetrate deeply into dense canopies (Lisein et al., 2013; Grenzdörffer, 2014; Ota et al., 2015). Structure from motion technique provides generally a good description of the digital surface model but accessing the digital terrain model is only possible when the ground is clearly visible (Khanna et al., 2015). This is the case for low canopy coverage or for phenotyping platforms where the ground is visible in the alleys and between the plots (Holman et al., 2016). The identification of ground points can be done directly by the photogrammetric software such as Agisoft Photoscan (Geipel et al., 2014). However, this method will depend on the choice of the classification parameters and the type and stage of vegetation. (Khanna et al., 2015) used the green index (Gitelson, 2004) and applied the Otsu automatic thresholding method (Otsu, 1979) over green crops. For senescent vegetation this approach will not provide good results because of confusions between senescent crop and bare soil. Therefore, the generation of the digital terrain model from the dense point cloud appears to be still a challenge in many situations. The problem could be solved by using a digital terrain model derived from an independent source of information (Bendig et al., 2014; Geipel et al., 2014; Grenzdörffer, 2014), assuming that the digital

[1]www.pix4d.com

[2]www.agisoft.com

terrain model does not vary significantly during the growing season.

The objective of this study is to develop a methodology for estimating plant height of wheat crops from RGB camera aboard UAV or LiDAR aboard a phénomobile (fully automatic rover) in the context of high-throughput field phenotyping. For this purpose, a comprehensive experiment was setup where the field phenotyping platform was sampled several times during the growing season with the UAV and the phénomobile. A definition of the plant height is first provided from the dense point cloud derived from the LiDAR that will constitute the reference. The UAV derived plant height based on the structure from motion algorithm will then be compared with the LiDAR reference plant height, with emphasis on the way the digital terrain model is computed. The flowering date of wheat was estimated from the dynamics of plant height. Finally, the broad-sense heritability of plant height and its correlation with yield and biomass were evaluated.

MATERIALS AND METHODS

Study Area

The field phenotyping platform (**Figure 1B**) is located in Gréoux les Bains (France, 43.7° latitude North, 5.8° longitude East, **Figure 1A**). The platform is approximately 200 m by 250 m size and is mainly flat with a 1 m maximum elevation difference. Wheat was sown on October the 29th 2015 with a row spacing of 17.5 cm and a seed density of 300 seeds·m^{-2}. It was harvested on the 6th July 2016. A total of 1173 microplots of 1.9 m width (11 rows) by 10 m long was considered, each of them corresponding to a given genotype among a total of 550 genotypes grown under contrasted irrigation modalities: about half of the platform was irrigated (WW) while the other part was subjected to water stress (WS modality). A moderate water stress took place in the 2015–2016 season. The cumulated water deficit was 126 mm for the WS modality and 18 mm for the WW modality. A subset of 19 contrasting genotypes was considered here to evaluate the plant height heritability. Each of those genotypes were replicated three times over the WW and WS modalities organized in an alpha plan experimental design.

Plant Height, Biomass and Flowering Stage Ground Measurements

Plant height was manually measured on 12 microplots: on each microplot, the average of 20 height measurements was calculated; each individual sample measurement corresponds to the highest point of the representative plant within an area of 30 cm radius, corresponding either to leaf or to an ear.

The above ground biomass was measured over three segments of 2 m length by two adjacent rows. The first two rows located at the border of the microplots were not considered in the sampling to minimize border effects. The samples were weighed fresh, and a subsample of around 30 plants taken to measure the water content by weighing it fresh and drying it in an oven for 24 h at 80°C. Around stage Zadoks 26, 6 microplots were sampled. At the stage Zadoks 32, 54 microplots were sampled, corresponding to one replicate of 27 genotypes both in WW and WS modalities. Finally at the flowering stage (Zadoks 50), 80 microplots were sampled corresponding to one replicate of 40 genotypes grown under the two irrigation modalities. However, due to measurement errors, the biomass measurement one microplot was missing. The invasive measurements were taken within less than 4 days from the closest LiDAR survey.

The yield of all the microplots corresponding to 19 genotypes times the three replicates in the two irrigation modalities was measured during the harvest: the weight of harvested grain was divided by the microplot area and the grain fresh weight was normalized to 12% relative moisture.

The flowering date was eventually scored visually every 3 days on one replicate for 19 genotypes grown under both irrigation modalities. The usual scoring system was used: flowering stage corresponds to the date when 50% of the ears have their stamina visible.

LIDAR Reference Measurements

The LiDAR on the Phénomobile

The phénomobile, a ground-based high-throughput phenotyping robot rover is equipped with a measurement head (**Figure 1C**) that is maintained automatically at a constant distance from the top of the canopy. The system steps over the microplots with a maximum 1.35 m clearance and an adjustable width of 2 m ± 0.5 m. The phénomobile automatically follows a predefined trajectory in the experimental field using a

FIGURE 1 | (A) Localization of the platform in France; **(B)** Aerial view of the experimental field; **(C)** and the Phénomobile rover robot on which the LiDARs are fixed.

centimetric accuracy real time kinetics GPS and accelerometers. The measurement head is equipped with several instruments including two LMS400 LiDARs (SICK, Germany) operating at 650 nm and scanning downward with ±35° zenith angle in a direction perpendicular to the rows at a frequency of 290 scans per second (Lefsky et al., 2002). The two LiDARs allow getting denser sampling of the scene. As the platform moves forward (**Figure 1C**) at a speed of 0.3 $m{\cdot}s^{-1}$ as recorded with the GPS information, the distance between two consecutive scans of a LiDAR along the row direction is around 1 mm. Measurements are taken every 0.2° along the scanning direction. The size of the footprint will depend on the distance to the sensor that varies from 2.4 mm × 5 mm at 0.7 m minimum measuring distance up to 10.5 mm × 5 mm at 3 m maximum measuring distance. The distance between the sensor and the target is measured from the phase shift principle (Neckar and Adamek, 2011). The intensity of the reflected signal and the distance are recorded at the same time. When the target in the LiDAR footprint is not horizontal or made of elements placed at several heights, the distance and the intensity computed by the LiDAR is approximately the average value over the LiDAR footprint. The nominal error on the distance is 4 mm under our experimental conditions. The scan of one microplot takes about 30 s during which about 3 million points are recorded with associated intensity and x-y-z coordinates. Each plot was sampled 14 times during the entire growth cycle to describe the whole season.

Data Processing and Height Definition

A strip of 0.6 m width located in the center of the microplot was extracted from the 3D point cloud (**Figure 2A**). This corresponds roughly to three rows and allows to limit possible border effects while increasing the probability to get points reflected by the soil by limiting the scan angle. Noise from the resulting points where then filtered using the Matlab implementation of the method proposed by Rusu et al. (2008). This process removed about 1% of the points. They were mainly located in the upper and lower part of the regions of interest.

The 0.6 m width strip was further divided into 20 consecutive non-overlapping elementary cells of 0.5 m length where the canopy height was assessed (**Figure 2A**). This allows accounting for possible variation of the digital terrain model if the microplot is not perfectly flat. This cell size was large enough to get a good description of the z profile (**Figure 2C**) including enough points corresponding to the ground level used to define the digital terrain model. The k-means clustering method (Seber, 1984) with two classes was applied to separate the ground from the vegetation from both the distance and the intensity values (**Figure 2B**). The maximum peak in the z-distribution of the resulted non-vegetation points was assigned as the ground level. The distance of the ground was subtracted from the distance of the 3D point cloud for each elementary cell in the microplot resulting into a distribution of the height values. The height of the canopy is then defined as the height value corresponding to a given R_p of the cumulated height distribution of the vegetation points. The $R_p = 99.5\%$ was selected here to define the vegetation height at the elementary cell level. When considering the later stages where a large heterogeneity of the height is observed at the top layer because of the presence of ears, this corresponds roughly to the area covered by 50 ears for each unit ground area, considering an ear diameter of 1 cm and a typical ear density. The sensitivity of the height to this percentile value will be later discussed in the results section. Finally, the median value of the elementary cells of the microplot was considered as the plant height.

Plant Height Estimates from the UAV

RGB Camera and UAV Flight

A Sony ILCE-6000 digital camera with a 6000 × 4000 pixels sensor was carried by a hexacopter with approximately 20 min autonomy. The camera was fixed on a 2 axes gimbal that maintains the nadir view direction during the flight. The larger dimension of the image was oriented across track to get larger swath. The camera was set to speed priority of 1/1250 s to avoid movement blur. The aperture and ISO were thus automatically adjusted by the camera. The camera was triggered by an intervalometer set at 1Hz frequency that corresponds to the maximum frequency with which RGB images can be recorded on the flash memory card of the camera. The images were recorded in the jpg format. Two different focal lengths were used: 19 and 30 mm with respectively ± 31.0° and ± 21.5° field of view across track. The flight altitude was designed to get around 1 cm GSD for both focal lengths (**Table 1**). Five measurements were completed from tillering to flowering (**Table 1**).

The speed of the UAV was set to 2.5 m/s to provide 90 and 94% overlap between images along the track respectively for the 30 mm and 19 mm focal lengths. The distance between tracks was set to 9 and 11.8 m respectively for the 19 and 30 mm focal lengths to provide 70% overlap across track. Two elevations of 10–15 min were necessary to cover the full area of interest. No images were acquired during the UAV stabilization over the waypoints. In addition, images corresponding to the takeoffs and landings were not used. This resulted in about 600 images for each date. The typical flight plan is shown in **Figure 3**.

Ground Targets and Georeferencing Accuracy

A total of 19 ground targets were evenly distributed over the platform with fixed position for all the flights. They were made of painted PVC disks of 60 cm diameter where the central 40 cm diameter disk was 20% gray level and was surrounded by a 60% gray level color external crown. These gray levels were selected to avoid saturation and allow automatic target detection on the images. Their location was measured with a real time kinetics GPS device ensuring a 1 cm horizontal and vertical accuracy for every flight. Among the 19 targets, 14 were used in the generation of the dense point cloud (ground control points) while the five additional ones were used to evaluate the accuracy of the geo-referencing (Check Points). The spatial distribution of the targets was designed to get some even coverage of the field considered (**Figure 3**).

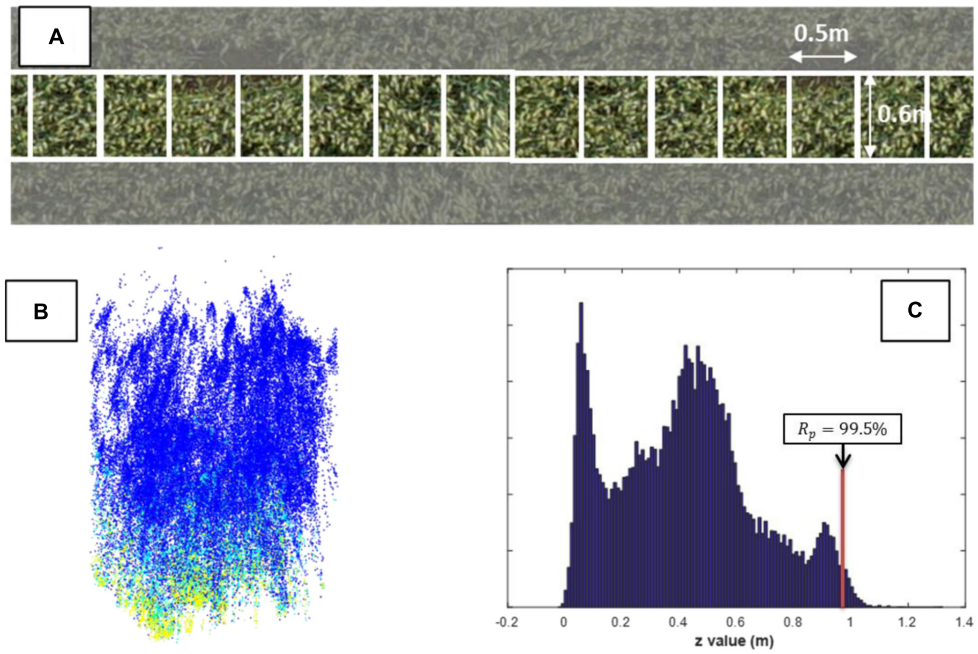


FIGURE 2 | (A) UAV-RGB images of one microplot where a 0.6 m × 0.5 m elementary cell is identified; **(B)** The 3D LiDAR points for an elementary cell. Each point colored with the intensity value of the returned signal [from blue (low intensity) to yellow (high intensity)]; **(C)** The corresponding z-distribution of the 3-D points.

Generation of the 3D Dense Point Cloud from the RGB Images

The ensemble of RGB images was processed with Agisoft Photoscan Professional (V 1.2.6) software. The first step consists in the image alignment performed using the scale invariant feature transform algorithm (Lowe, 2004). An "on-the-job-calibration" was applied to adjust the camera parameters within the structure from motion process. The application of this method was possible because of the high overlap between images (Turner et al., 2014) and the suitable distribution of the ground control points (James and Robson, 2014; Harwin et al., 2015). The Agisoft software generates in a first step a set of tie points, each point being associated with a projection error. As advised by Agisoft, tie points with a projection error higher than 0.3 ground sample distance were removed. A bundle adjustment is then applied (Granshaw, 1980; Triggs et al., 1999). Further, points with a low reconstruction uncertainty (points, reconstructed from nearby photos with small baseline) were then removed. These points are generally observed for small overlapping fraction between images along with a large view zenith angle resulting in larger ground sample distance. The ground control points used in this process were automatically identified using a custom developed pipeline. The check points were not used in the bundle adjustment, the average accuracy on the check points reported in **Table 1** (σx, σy, and σz) were in agreement with the recommendations from (Vautherin et al., 2016): 1–2 times the ground sample distance in x and y directions, and 2–3 times the ground sample distance in the z direction. The dense point cloud is generated from dense-matching photogrammetry using a moderate depth filtering option and the full image resolution as implemented in Photoscan 1.2.6. This filtering process results in more variable density of points of the dense cloud, the mean density of points in the vegetation part of the study area was 2300 points/m^2.

TABLE 1 | Characteristics of the five flights completed over the Gréoux experiment in 2016.

Date (DaS)	Illumination conditions	Wind speed (km/h)	Focal length (mm)	Altitude (m)	GSD (cm)	Overlap (%)		σx (cm)	σy (cm)	σz (cm)
						along	across			
139	Covered	8	30	75	0.98	90	70	2.4	3.1	5.5
152	Sunny	6	30	75	0.98	90	70	4.5	1.3	3.3
194	Sunny	10	19	50	1.04	94	70	5.1	1.3	3.9
216	Cloudy	7	19	50	1.04	94	70	2.1	2.9	2.8
225	Sunny	5	30	75	0.98	90	70	5.0	2.6	3.9

σx, σy, σz, correspond to the standard deviation of the localization of the control points used to quantify the geometric accuracy.

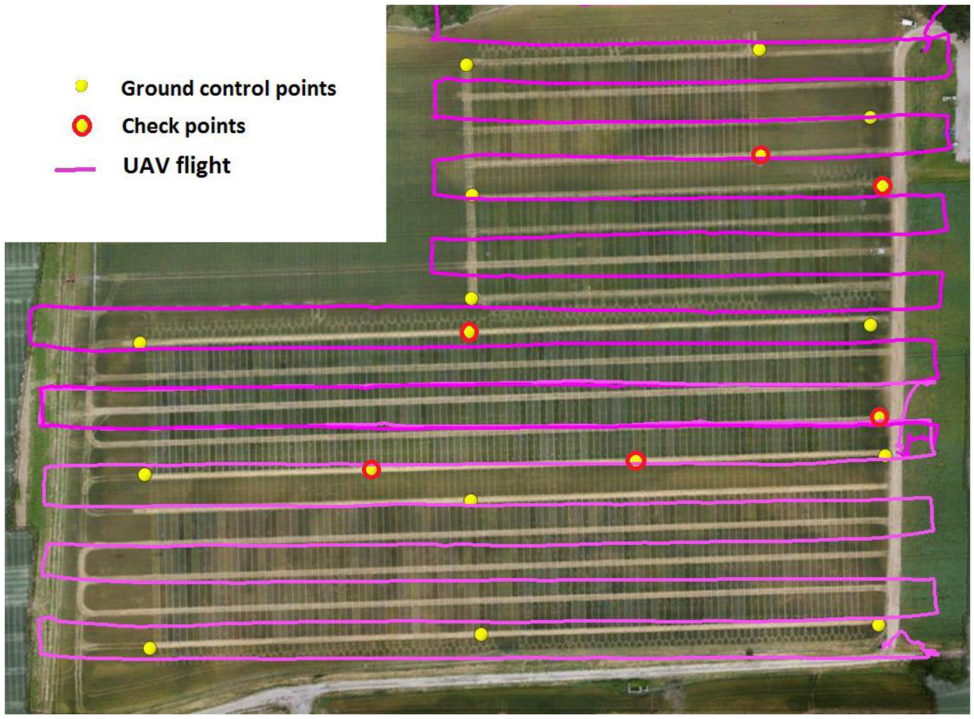


FIGURE 3 | The flight plan with ground control points (yellow circles with red outline) and check points (yellow circles).

Derivation of the Digital Terrain Model

Two methods were used to derive the digital terrain model. The first one is simply based on the collection of the coordinates of the points recorded during sowing by the sowing machine equipped with a centimetric accuracy Real Time Kinematic GPS. The second approach is based on the extraction of ground points from the dense point cloud and interpolation between them to generate the digital terrain model. The phenotyping platform (**Figure 3**) was split into 13 m × 13 m cells with a 75% overlapping (50% in both x and y directions). The size of the cell is a compromise between a small one that allows to get of finer description of digital terrain model variations, and a large one that will ensure to get at least few background points from the dense cloud points. Similarly to the LiDAR processing, a k-means clustering (Seber, 1984) with 2 classes is applied using the z-value and the red and green color associated to each point of the dense cloud. This k-means clustering is iterated over the previous background class if the standard deviation in the background class, σ_b, is lower than 0.14 m. However, if $\sigma_b > 0.14$ m after the 4th iteration the iteration process is stopped and no background *z*-value is assigned to the considered cell. The $\sigma_b > 0.14$ m value corresponds approximatively to the background roughness expected over the 13 m × 13 m cell and was defined after several trial and error tests. Then, ground point cloud was filtered using (Rusu et al., 2008) algorithm to regularize the *z*-values over each cell. Finally, a natural neighbor interpolation (Owen, 1992) was applied to compute the z value for each microplot. Note that here the microplot is assumed to be flat.

Plant Height Estimation

For each plot, the *z*-values of the dense cloud points were subtracted from the *z*-value of the digital terrain model assigned to the microplot. Finally, the microplot is divided into 20 consecutive non-overlapping elementary cells of 50 cm × 60 cm similarly to what was achieved for the LiDAR data. The median value of plant height corresponding to a given R_p of the cumulated z distribution is finally computed and considered as the microplot crop plant height. The selection of the value of the R_p used to define plant height will be discussed later in the Section "Results."

Date When the Maximum Plant Height Is Reached

The flowering date appears roughly when the vegetative growth is completed, i.e., when the stems reached their maximum height. This stage could thus be tentatively estimated using the plant height time course. This requires obviously frequent observations as completed in this study with the LiDAR while the plant height monitoring with the UAV was too sparse. As a consequence, only the LiDAR measurements will be used here for estimating the flowering stage. When expressing the time in GDDs the plant height temporal profile can be approximated by a vegetative growth phase, followed by a plateau during the reproductive phase. The plant height corresponding to the plateau was simply defined by the maximum plant height value over the whole cycle. A second-order polynomial regression was used to describe the plant height during the vegetative growth. The vegetative growth period was assumed to start for GDD = 1000°C.day. It was

then incrementally extended by including additional observation dates for GDD > 1500°C.day if the corresponding plant height elongation rate does not decrease by more than 60% than that of the previous value. The intersection of the elongation curve with the plateau provides the date when plant height reaches its maximum.

RESULTS AND DISCUSSION

LiDAR Measurements of Plant Height

The LiDAR plant height defined using $R_p = 99.5\%$ were compared with the available manual measurements in the field. Results show a strong agreement with a low RMSE of 3.47 cm and small bias (bias = 1.41 cm) (**Figure 4**).

The impact of the R_p value on plant height was further investigated using the difference $\Delta PH = PH_x - PH_{99.5}$ where PH_x and $PH_{99.5}$ represent the plant height values respectively for $R_p = x\%$ and $R_p = 99.5\%$. Results (**Figure 5A**) show that very high values of $R_p = 99.99\%$ increases plant height by more than $\Delta PH = +5$ cm in most situations. Conversely, $R_p = 99.0\%$ decreases plant height by more than $\Delta PH = -5$ cm. The absolute difference ΔPH increases rapidly with plant height for PH < 0.1 (**Figure 5A**). Then, ΔPH increases only slightly with plant height (**Figure 5A**), with, however, significant scatter for the larger plant height values and when R_p is different from the nominal value ($R_p = 99.5\%$). The variability of plant height across the 20 elementary cells within a microplot (**Figure 5B**) shows that it is minimum for $R_p = 99.5\%$ with STD = 3.1 cm. It increases rapidly either for $R_p < 99.5\%$ or for $R_p > 99.5\%$ although the STD value keeps relatively small (STD < 3.7 cm for $R_p = 99.99\%$ or for $R_p = 90\%$). The use of the median values computed over the 20 elementary cells provides in addition a better representativeness of the plant height of a microplot. This appears more important at the tillering stage where the plant height variability within a microplot is the largest. These results confirm that $R_p = 99.5\%$ provides an accurate and precise plant height estimation.

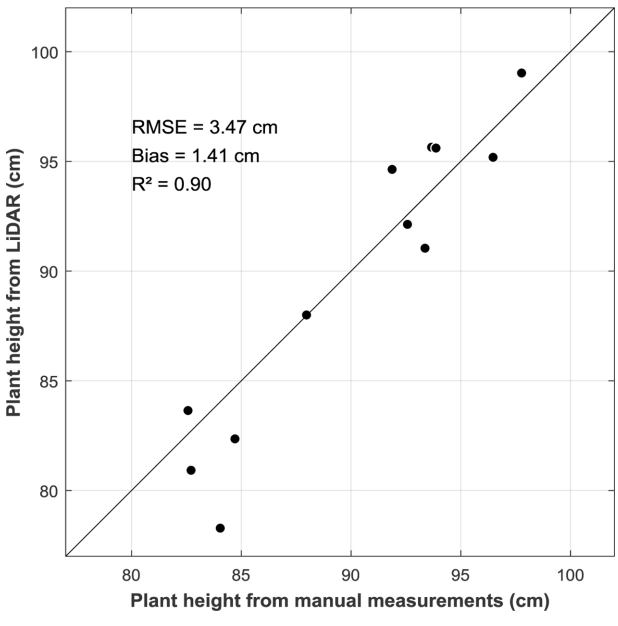


FIGURE 4 | Comparison between plant height derived from LiDAR measurements with plant height measured manually in the field (n = 14). Solid line is the 1:1 line.

Derivation of the Digital Terrain Model with Structure from Motion Algorithm

The digital terrain model extracted from the dense point cloud for each of the 5 flights were compared. In addition, the digital terrain model generated from the real time kinetics GPS placed on the sowing machine during sowing was also used. A mean altitude value of the ground level for the 1173 microplots was then computed for the 7 digital terrain models. Results show that the correlation between the altitudes computed from all the digital terrain model combinations is always very high with $R^2 > 0.97$ (**Table 2**). This indicates that all the digital terrain models were capturing consistently the general topography of the experimental platform.

Results show further that the RMSE values are between 2.6 and 6.8 cm (**Table 2**), except for DaS 152 that shows larger values. No clear explanation was found for the degraded performances of DaS 152. However, better consistency seems to be observed when using a shorter focal length (comparison between DaS 194, DaS 216 and Sowing).

Comparison of Plant Height Derived from Structure from Motion and LiDAR

The LiDAR was more frequently sampling the platform along the growth season as compared to the UAV flights (**Table 1**). Plant height derived from the LiDAR were thus interpolated to the dates of the UAV flights. However, if the LiDAR acquisition of a microplot differs by more than a week from that of the UAV flight, the corresponding microplot was not considered in the comparison. This resulted in a total of 2076 couples of structure from motion and LiDAR plant height. The plant height from structure from motion was first derived using the same R_p as that used for the LiDAR ($R_p = 99.5\%$). Results (**Figure 6**) show that structure from motion plant height are strongly correlated with LiDAR reference plant height across the 5 UAV flights available. This corroborates previous results reported (Bareth et al., 2016; Fraser et al., 2016; Holman et al., 2016). The same level of consistency is observed for plant height derived from a digital terrain model computed from the same dense cloud ($R^2 = 0.97$, RMSE = 7.7 cm) as compared to using the digital terrain model derived from the sowing ($R^2 = 0.98$, RMSE = 8.4 cm). The correlations are generally weaker for the early stages due to the limited range of variation of plant height (DaS 139, DaS 152). Further, using the 30 mm camera focal length (DaS 139 DaS 152 DaS 225) tends to decrease the plant height consistency with the reference LiDAR derived plant height as compared to the 19 mm focal length (**Table 3**). The 19 mm focal length increases the disparity in the view configurations which may help the structure from motion algorithm to get more accurate estimates of the z component in the dense cloud as earlier

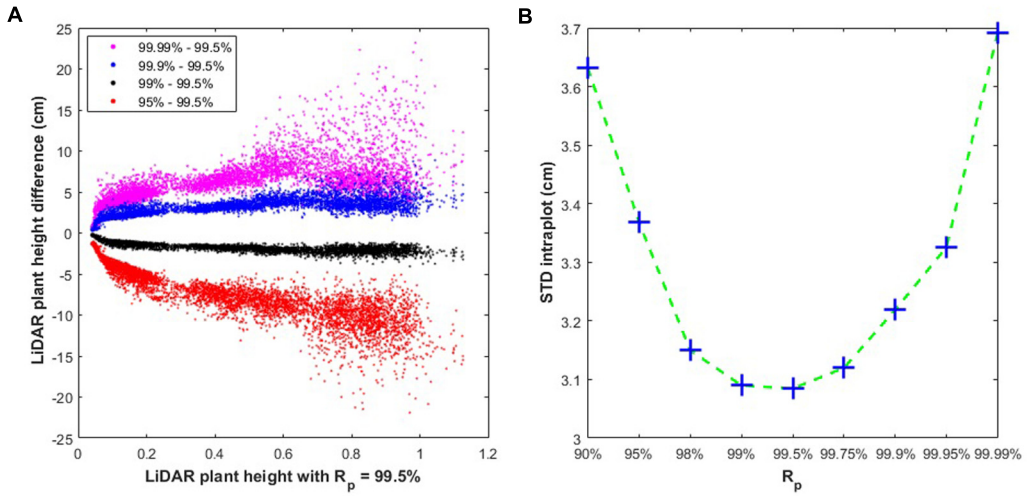


FIGURE 5 | (A) Difference in plant height estimates using several R_p *z*-values (R_p) against estimate using $R_p = 99.5$; **(B)** Standard deviation of plant height (STD intraplot, in cm) computed within a microplot between the 20 elementary cells as a function of the R_p value selected to define plant height.

TABLE 2 | Correlation (R^2, bottom triangle) and RMSE (top triangle) values between the digital terrain models computed over the 1173 microplots for the 5 flights as well as that derived from the real time kinetics GPS on the sowing machine.

R^2 \ RMSE (cm)	Sowing	DaS 139*	DaS 152*	DaS 194	DaS 216	DaS 225*
Sowing	–	2.6	7.2	3.4	2.6	6.0
DaS 139*	1.00	–	7.2	4.5	2.9	5.0
DaS 152*	0.96	0.95	–	9.2	7.4	9.8
DaS 194	0.99	0.99	0.95	–	3.8	6.8
DaS 216	0.99	0.99	0.96	0.99	–	6.0
DaS 225*	0.97	0.97	0.91	0.97	0.96	–

**Indicates that the camera was equipped with the 30 mm focal length instead of the 19 mm one.*

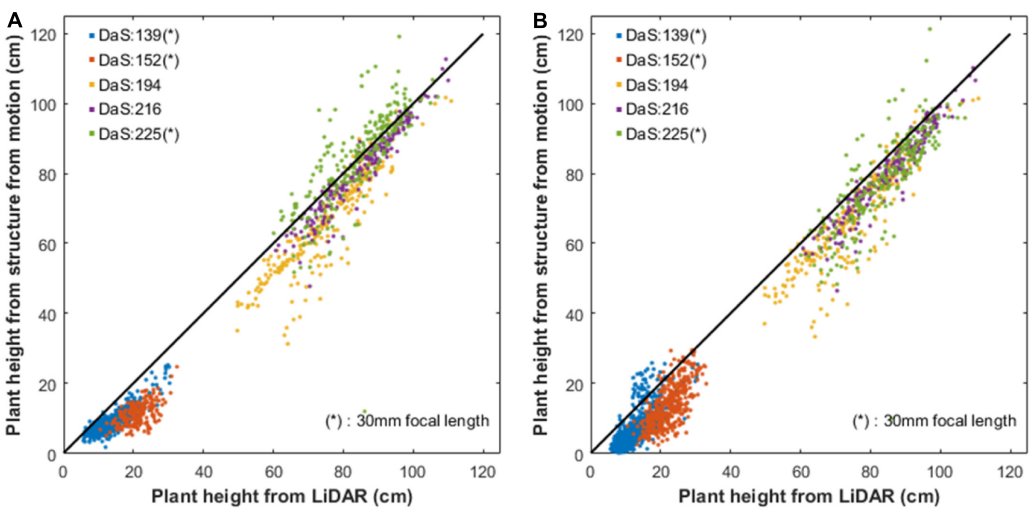


FIGURE 6 | (A) Plant Height computed from the background points identified over each date; **(B)** Plant height computed from the digital terrain model derived from the sowing machine. Each color corresponds to a flight date (DaS). *Indicates that the camera was equipped with the 30 mm focal length instead of the 19 mm one.

TABLE 3 | Agreement between LiDAR and structure from motion derived plant height when the digital terrain model used come either from the same dense cloud or from the Sowing.

	Digital terrain model from the dense cloud			Digital terrain model from Sowing		
DaS	**R^2**	**RMSE (cm)**	**Bias (cm)**	**R^2**	**RMSE (cm)**	**Bias (cm)**
139*	0.76	5.0	−4.4	0.50	6.8	−5.6
152*	0.31	9.2	−8.6	0.45	9.0	−9.0
194	0.84	11.0	−9.4	0.80	9.9	−7.7
216	0.92	5.1	−3.9	0.91	6.2	−5.0
225*	0.59	8.7	−0.38	0.63	9.8	−5.4
All	0.97	7.7	−5.1	0.98	8.4	−6.5

**Indicates that the camera was equipped with the 30 mm focal length instead of the 19 mm one.*

reported (James and Robson, 2014). This result also confirms the ability of Agisoft to model the radial lens distortion of wide field of view lens. However, the calibration of the camera from the bundle adjustment requires an even distribution of a sufficient number ground control points (James and Robson, 2014; Harwin et al., 2015) and a high overlapping between images as done in this study.

A systematic overestimation of the plant height derived from structure from motion is observed as compared to the reference plant height derived from the LiDAR. This agrees with results from other studies (Grenzdörffer, 2014; Bareth et al., 2016; van der Voort, 2016) who found that structure from motion lacked the ability to reconstruct accurately the top of the canopy. This is partly due to the spatial resolution difference between the LiDAR (3–5 mm) and the RGB camera (10 mm) as compared to the size of the objects at the top of the canopy (on the order of the cm). However, increasing the spatial resolution will lead to more noisy dense cloud with more gaps over vegetated areas as reported by Brocks et al. (2016) and as was experienced also in this study (results not shown for the sake of brevity).

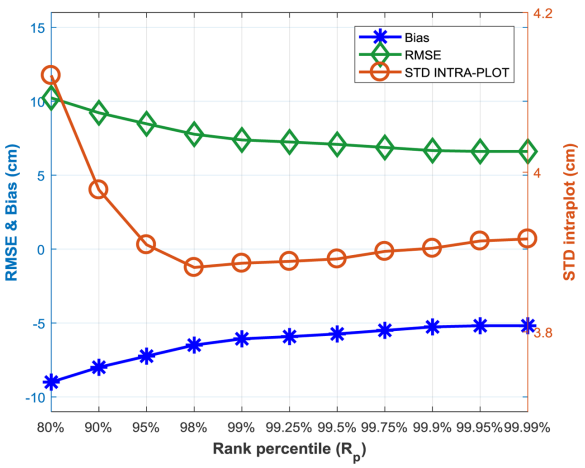


FIGURE 7 | Impact of the rank percentile (R_p) used to defined plant height from the dense cloud derived from structure from motion on RMSE and bias (left *y* axis) and the variability of plant height along the microplot (right *y* axis). The reference plant height used here is that derived from the LiDAR with Rp = 99.5%.

The principles of height measurement are very different between the LiDAR and structure from motion: the structure from motion algorithm uses two different directions to build the dense cloud, limiting the penetration capacity because of possible occultation; conversely LiDAR uses only a single direction with much better penetration in the canopy. As a consequence, the z profiles are expected to be different between LiDAR and structure from motion. The impact of the R_p value used to define plant height from the dense cloud derived from structure from motion was thus further investigated on the 2076 couples of measurements. As expected, increasing the R_p value decreases the bias and thus RMSE with the reference LiDAR plant height (**Figure 7**). However, the decrease seems to be limited after $R_p > 99\%$, reaching 8 cm difference for $R_p = 99.99\%$ Note that the 99.99% percentile corresponds to very few points in the dense point cloud since the cell of 0.5 m × 0.6 m contains around 1000 points. Increasing R_p reduces the variability of plant height between the 20 elementary cells within a microplot up to $R_p = 99.9\%$ (**Figure 8**). This simple sensibility analysis shows that best consistency with the LiDAR reference plant height is obtained for $99.5\% < R_p < 99.99\%$ with actually small improvement for R_p larger than 99.5%. This justifies a posteriori the $R_p = 99.5\%$ value used for plant height estimation from structure from motion.

Plant Height as a Reliable Trait for Wheat Phenotyping

Broad Sense Heritability

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

$$Y = \mu + S + \underline{G} + \underline{L} + \underline{C} + \underline{L{:}C} + \varepsilon$$

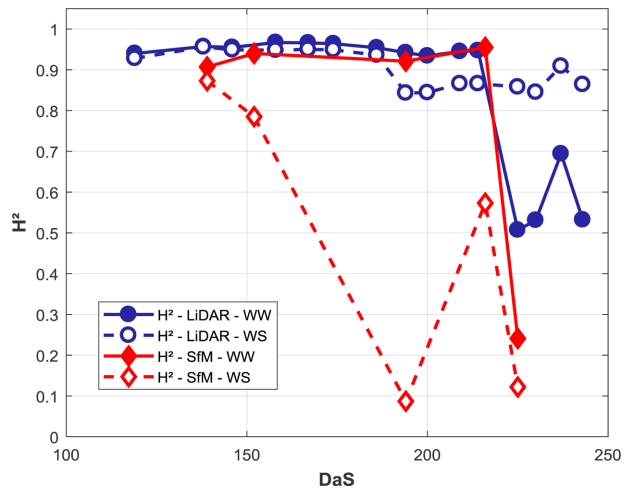


FIGURE 8 | Heritability (H^2) of plant height for different environments conditions and methods along the growth cycle. The two modalities (WW and WS) and plant height derived from LiDAR and structure from motion are individually presented.

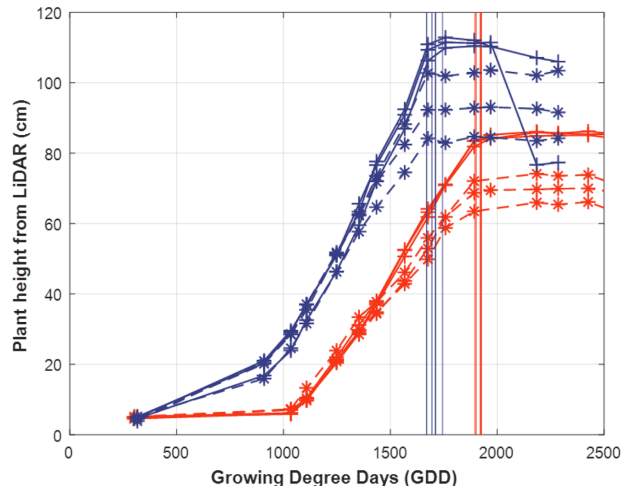


FIGURE 9 | Dynamics of the LiDAR plant height for two genotypes (red and blue lines and symbols) with three replicates in the WW environment ('+' solid lines) and in the WS environment ('*' dashed lines) the date when the maximum plant height is reached is indicated by the vertical line. Time is expressed in Growing Degree Days (GDD, °C.day).

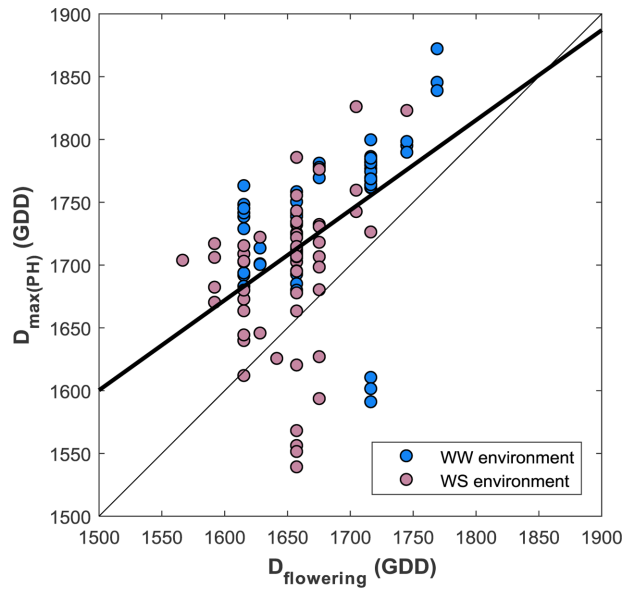


FIGURE 10 | Comparison between the date expressed in Growing Degree Day (GDD) of the maximum plant height growth with the flowering date visually scored (expressed in GDD) (n = 114).

Where S is the soil water holding capacity. G is the random effect of the genotypes. L and C are, respectively, the random lines and column effects in our alpha design plan and L:C is the random sub-block effect. μ is the intercept term (fixed) and ε the random residual error.

The plant height trait derived from the LiDAR shows a high H^2 up to DaS 210 (**Figure 8**) for the WW modality. It drops dramatically at the end of the growth cycle in relation to lodging that was affecting differently the replicates. Conversely, the WS modality keeps relatively stable during the whole growth cycle because no lodging was observed. However, when the water stress starts to impact crop growth around DaS 180, a small decrease of the H^2 is observed: residual environmental effects not accounted for by the alpha experimental plan and the soil water holding capacity were slightly degrading the H^2 value.

The H^2 values computed over the WW modality from structure from motion are close to those observed for the LiDAR, with, however, a slight degradation of the performances. Conversely, the H^2 values computed on the WS modality from structure from motion show the smallest H^2 values. On DaS 194, the H^2 is low for the WS modality. A detailed inspection shows a noisy dense point cloud in the WS part of the field that impacted the height computation and thus H^2. At this specific date and location, the phénomobile was operating during the UAV flights which induces artifacts and problems in the dense point cloud generation from structure from motion.

Plant Height as a Proxy of the Flowering Stage

Due to the reduced observation frequency of the UAV, flowering time was only assessed using the LiDAR plant height. The date when the maximum plant height is reached, $D_{\max(PH)}$, is considered as a proxy of the flowering stage. **Figure 9** shows that $D_{\max(PH)}$ is well identified based on the simple algorithm presented in the methods section. Further, it appears that $D_{\max(PH)}$ is little dependent on the environmental conditions: WW and WS modalities are very close and for the WS modality, there is no difference due to the soil water holding capacity although differences in *max(PH)* are observed.

The flowering dates are well correlated with $D_{\max(PH)}$ (**Figure 10**) (R^2 = 0.24, RMSE = 76, $D_{\max(PH)}$ = 0.7 $D_{flowering}$ + 541). However, the best linear fit shows that the earlier genotypes reach the maximum plant height about 100 GDD after the flowering stage, which corresponds approximately

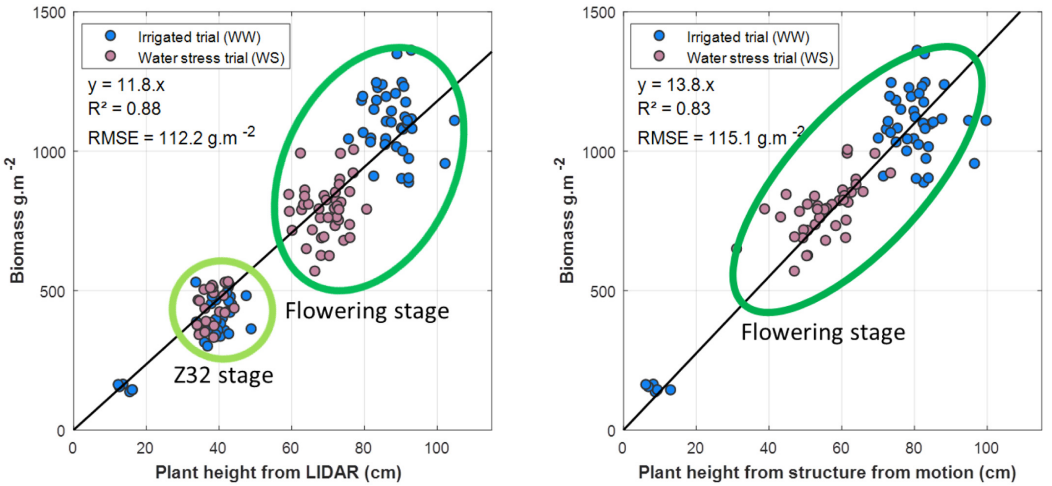


FIGURE 11 | (left) Regression between plant height derived from LiDAR and Biomass (n = 139): (right) regression between plant height derived from structure from motion and Biomass (n = 86). Blue and red symbols correspond respectively to the WW and WS modalities. Unfortunately, no UAV acquisition was conducted for the Zadocks 32 stage. The Z32 and flowering stages are indicated by the corresponding green envelope of the points in the figures.

to 7 days. The late genotypes show less differences, around 20 GDD corresponding to 1 or 2 days after flowering. $D_{\max(PH)}$ appears thus to be a reasonable proxy of the flowering stage considering that the accuracy of its visual scoring date is around 2–3 days. Nevertheless, some genotypes show significant differences from the main relationship as illustrated in **Figure 10**.

The heritability of $D_{\max(PH)}$ was very high, $H^2 = 0.96$ and $H^2 = 0.88$, respectively for the WW and WS modalities. This confirms the small influence of the environment for the genetic expression of this trait.

Relationship with Above Ground Biomass and Yield

Correlations between plant height and biomass along the growing season are very strong (**Figure 11**) both for the LiDAR ($R^2 = 0.88$, RMSE = 112.2 g/m^2) and the structure from motion ($R^2 = 0.91$, RMSE = 98.0 g/m^2). These good relationships confirm observations by several authors (Yin et al., 2011; Bendig et al., 2014; Ota et al., 2015; Tilly et al., 2015). However, these correlations are mainly driven by the variability across stages along the growth cycle. For a given stage, little prediction power of the biomass is observed from plant height (**Figure 11**). The correlation at the flowering stage is relatively low ($R^2 = 0.5$) for both methods. Other variables such as the basal area should be used to improve the correlations. Yield is poorly correlated with maximum plant height both when derived from LiDAR ($R^2 = 0.22$, RMSE = 149.6 g.m^{-2}) and structure from motion ($R^2 = 0.13$, RMSE = 152.3 g.m^{-2}). This is consistent with the poor correlation with biomass observed for a given growing stage, assuming that the harvest index varies within a small range.

DISCUSSION AND CONCLUSION

Since crop surface is very rough, an important point addressed in this study was to propose a definition of plant height from the 3D point cloud retrieved from LiDAR or structure from motion techniques. The 99.5% percentile of the cumulated z-value was found to be optimal for comparison with ground ruler measurements while minimizing the spatial variability over each microplot. However, this definition will probably slightly depend on the canopy surface roughness. As a consequence, the 99.5% percentile used as a reference for wheat should be checked and possibly adapted for other crops as well as a function of the spatial resolution used. LiDAR measurements are based on a single source/view configuration allowing to penetrate into the canopy and reach the ground reference surface. Plant height could then be directly measured because of the availability of ground reference points within a microplot. Conversely, the penetration capacity of structure from motion methods based on the combination of distinct view directions from the UAV is limited because of possible occultation that will increase when the canopy closes. In these conditions, two strategies were compared: (1) either find ground reference points over the whole 3D dense point cloud and interpolate these points to get the digital terrain model; or (2) use and ancillary digital terrain model, that was in this study derived from real time kinetics GPS acquired during the sowing of the crop. The first approach might be limited in the case of a terrain presenting a complex topography when only few ground points are identified. Note that the ground control points could be used as ground level points if the distance to the ground is precisely known. Results show that both methods reach the same level of accuracy. For the two approaches investigated here to define the digital terrain model and extract the plant height of each microplot, the methods presented here were designed to process automatically the original imagery. This includes automatic and direct extraction of the microplots as well as of the digital terrain model from the dense cloud as opposed to earlier studies where plant height was derived from a crop surface model generated from the dense cloud (Bareth et al., 2016).

The comparison between plant height derived from LiDAR and structure from motion shows a very high consistency with strong correlation ($R^2 \approx 0.98$) and small RMSE values (RMSE = 8.4 cm). Most of the RMSE was explained by a significant bias, the plant height being underestimated. This may be partly due to the differences in the spatial resolution of the two systems (about 4 mm for LiDAR and 10 mm for UAV imagery) as well as in differences in canopy penetration capacity. However, plant height derived from structure from motion is systematically lower than that of the LiDAR. Our results further indicate that larger field of view with shorter focal lengths would generate more accurate 3D dense point clouds from structure from motion and thus plant height because of the increased disparity between the several view points. However, complementary study should investigate more deeply this effect as well as the impact of a degraded spatial resolution.

High H^2 (repeatability) of plant height was observed both for LiDAR and structure from motion. The water stress experiment over which the LiDAR and structure from motion techniques were evaluated shows that plant height is a very pertinent trait to characterize the impact of drought before flowering stage: plant height not only quantifies the magnitude of the stress, it allows also to date precisely when the stress started to impact plant growth if sufficiently frequent observations are available. In addition, the date when plant height reaches its maximum was demonstrated to be a reasonable proxy of the flowering date with, however, some slight variability between genotypes. The heritability of the $D_{\max(PH)}$ reached was very heritable since it was demonstrated to be very little dependent on the water stress experienced by the plants in this experiment. The phasing difference between the end of the vegetative growth period and the flowering date might be investigated by breeders as a new trait of interest. Finally, plant height provides obviously a very easy and convenient way to identify plant lodging either based on the temporal evolution of the microplot, or from the variance between the 20 elementary cells considered in each microplot. All these results make the plant height trait very interesting for plant breeders. However, very low correlation with total above ground biomass and yield were observed for a given date of observation while high correlations are found across stages. Additional variables should be used such as the basal area to get the biovolume to get a better proxy of the above ground biomass at harvest.

Plant height derived from UAV using structure from motion algorithms were demonstrated here to lead to similar degree of accuracy as compared to the LiDAR observations from the phénomobile. The affordability and flexibility of UAV platforms and the constant improvement of cameras (better, smaller, lighter, cheaper) will probably make UAVs the basic vehicle to be used for high-throughput field phenotyping of plant height. Further, the recent availability of centimetric knowledge of the camera position for each image based on real time kinetics techniques will ease the structure from motion processing while possibly limiting the number of ground control points to be set up in the field.

AUTHOR CONTRIBUTIONS

SJ manage the field platform. SM and MH design the flight plan. ST, BdS, FB, and AC manage the phénomobile acsuisition. MH pilot the UAV. GC develop some routines for the processing of the images from the UAV. ST develop some routines for the preprocessing of the LiDAR. The algorithm development were mainly accomplished by SM, with the advices and comment from FB, BdS, and DD. SM wrote the manuscript and FB made very significant revisions. All authors participated in the discussion.

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.

ACKNOWLEDGMENTS

We thank very much Olivier Moulin, Guillaume Meloux, and Magalie Camous from the Arvalis experimental station in Gréoux for their kind support during the measurements. We also thank Florent Duyme and Emmanuelle Heritier from Arvalis for important discussions and help about the statistical analysis.

REFERENCES

Bareth, G., Bendig, J., Tilly, N., Hoffmeister, D., Aasen, H., and Bolten, A. (2016). A comparison of UAV- and TLS-derived plant height for crop monitoring: using polygon grids for the analysis of Crop Surface Models (CSMs). *Photogramm. Fernerkund. Geoinf.* 2016, 85–94. doi: 10.1127/pfg/2016/0289

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv:1406.5823

Bendig, J., Bolten, A., and Bareth, G. (2013). UAV-based imaging for multi-temporal, very high Resolution Crop Surface Models to monitor Crop Growth VariabilityMonitoring des Pflanzenwachstums mit Hilfe multitemporaler und hoch auflösender Oberflächenmodelle von Getreidebeständen auf Basis von Bildern aus UAV-Befliegungen. *Photogramm. Fernerkund. Geoinf.* 2013, 551–562. doi: 10.1127/1432-8364/2013/0200

Bendig, J., Bolten, A., Bennertz, S., Broscheit, J., Eichfuss, S., and Bareth, G. (2014). Estimating biomass of barley using Crop Surface Models (CSMs) derived from UAV-based RGB imaging. *Remote Sens.* 6, 10395–10412. doi: 10.3390/rs61110395

Berry, P. M., Sterling, M., Baker, C. J., Spink, J., and Sparkes, D. L. (2003). A calibrated model of wheat lodging compared with field measurements. *Agric. For. Meteorol.* 119, 167–180. doi: 10.1016/S0168-1923(03)00139-4

Blonquist, J. M., Norman, J. M., and Bugbee, B. (2009). Automated measurement of canopy stomatal conductance based on infrared temperature. *Agric. For. Meteorol.* 149, 2183–2197. doi: 10.1016/j.agrformet.2009.10.003

Brocks, S., Bendig, J., and Bareth, G. (2016). Toward an automated low-cost three-dimensional crop surface monitoring system using oblique stereo imagery from consumer-grade smart cameras. *J. Appl. Remote Sens.* 10, 046021–046021. doi: 10.1117/1.JRS.10.046021

Chéné, Y., Rousseau, D., Lucidarme, P., Bertheloot, J., Caffier, V., Morel, P., et al. (2012). On the use of depth camera for 3D phenotyping of entire plants. *Comput. Electron. Agric.* 82, 122–127. doi: 10.1016/j.compag.2011.12.007

Dandois, J. P., and Ellis, E. C. (2013). High spatial resolution three-dimensional mapping of vegetation spectral dynamics using computer vision. *Remote Sens. Environ.* 136, 259–276. doi: 10.1016/j.rse.2013.04.005

Deery, D., Jimenez-Berni, J., Jones, H., Sirault, X., and Furbank, R. (2014). Proximal remote sensing buggies and potential applications for field-based phenotyping. *Agronomy* 4, 349–379. doi: 10.3390/agronomy4030349

Escolà, A., Planas, S., Rosell, J. R., Pomar, J., Camp, F., Solanelles, F., et al. (2011). Performance of an ultrasonic ranging sensor in apple tree canopies. *Sensors* 11, 2459–2477. doi: 10.3390/s110302459

Fraser, R. H., Olthof, I., Lantz, T. C., and Schmitt, C. (2016). UAV photogrammetry for mapping vegetation in the low-Arctic. *Arctic Sci.* 2, 79–102. doi: 10.1139/as-2016-0008

Geipel, J., Link, J., and Claupein, W. (2014). Combined spectral and spatial modeling of corn yield based on aerial images and crop surface models acquired with an unmanned aircraft system. *Remote Sens.* 6, 10335–10355. doi: 10.3390/rs61110335

Gitelson, A. A. (2004). Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation. *J. Plant Physiol.* 161, 165–173. doi: 10.1078/0176-1617-01176

Granshaw, S. I. (1980). Bundle adjustment methods in engineering photogrammetry. *Photogramm. Rec.* 10, 181–207. doi: 10.1111/j.1477-9730.1980.tb00020.x

Grenzdörffer, G. J. (2014). Crop height determination with UAS point clouds. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* 1, 135–140. doi: 10.5194/isprsarchives-XL-1-135-2014

Harwin, S., Lucieer, A., and Osborn, J. (2015). The impact of the calibration method on the accuracy of point clouds derived using unmanned aerial vehicle multi-view stereopsis. *Remote Sens.* 7, 11933–11953. doi: 10.3390/rs70911933

Hoffmeister, D., Waldhoff, G., Korres, W., Curdt, C., and Bareth, G. (2015). Crop height variability detection in a single field by multi-temporal terrestrial laser scanning. *Precis. Agric.* 17, 296–312. doi: 10.1007/s11119-015-9420-y

Holland, J. B., Nyquist, W. E., and Cervantes-Martínez, C. T. (2002). "Estimating and Interpreting heritability for plant breeding: an update," in *Plant Breeding Reviews*, ed. J. Janick (Hoboken, NJ: John Wiley & Sons, Inc.), 9–112. doi: 10.1002/9780470650202.ch2

Holman, F. H., Riche, A. B., Michalski, A., Castle, M., Wooster, M. J., and Hawkesford, M. J. (2016a). High throughput field phenotyping of wheat plant height and growth rate in field plot trials using UAV based remote sensing. *Remote Sens.* 8:1031. doi: 10.3390/rs8121031

James, M. R., and Robson, S. (2014). Mitigating systematic error in topographic models derived from UAV and ground-based image networks. *Earth Surf. Process. Landf.* 39, 1413–1420. doi: 10.1002/esp.3609

Khanna, R., Moller, M., Pfeifer, J., Liebisch, F., Walter, A., and Siegwart, R. (2015). "Beyond point clouds - 3D mapping and field parameter measurements using UAVs," in *Proceedings of the IEEE 20th Conference on Emerging Technologies & Factory Automation (ETFA)*, Luxembourg, 1–4. doi: 10.1109/ETFA.2015.7301583

Lefsky, M. A., Cohen, W. B., Parker, G. G., and Harding, D. J. (2002). Lidar remote sensing for ecosystem studies lidar, an emerging remote sensing technology that directly measures the three-dimensional distribution of plant canopies, can accurately estimate vegetation structural attributes and should be of particular interest to forest, landscape, and global ecologists. *BioScience* 52, 19–30. doi: 10.1641/0006-3568(2002)052[0019:LRSFES]2.0.CO;2

Lisein, J., Pierrot-Deseilligny, M., Bonnet, S., and Lejeune, P. (2013). A photogrammetric workflow for the creation of a forest canopy height model from small unmanned aerial system imagery. *Forests* 4, 922–944. doi: 10.3390/f4040922

Llorens, J., Gil, E., Llop, J., and Escolà, A. (2011). Ultrasonic and LIDAR sensors for electronic canopy characterization in vineyards: advances to improve pesticide application methods. *Sensors* 11, 2177–2194. doi: 10.3390/s110202177

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 91–110. doi: 10.1023/B:VISI.0000029664.99615.94

Mallet, C., and Bretar, F. (2009). Full-waveform topographic lidar: state-of-the-art. *ISPRS J. Photogramm. Remote Sens.* 64, 1–16. doi: 10.1016/j.isprsjprs.2008.09.007

Moras, J., Cherfaoui, V., and Bonnifait, P. (2010). "A lidar perception scheme for intelligent vehicle navigation," in *Proceedings of the 11th International Conference on Control, Automation, Robotics and Vision*, Singapore, 1809–1814. doi: 10.1109/ICARCV.2010.5707962

Neckar, P., and Adamek, M. (2011). Software and hardware specification for area segmentation with laser scanner SICK LMS 400. *J. Syst. Appl. Eng. Dev.* 5, 674–681.

Ota, T., Ogawa, M., Shimizu, K., Kajisa, T., Mizoue, N., Yoshida, S., et al. (2015). Aboveground biomass estimation using structure from motion approach with aerial photographs in a seasonal tropical forest. *Forests* 6, 3882–3898. doi: 10.3390/f6113882

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* 9, 62–66. doi: 10.1109/TSMC.1979.4310076

Owen, S. J. (1992). *An Implementation of Natural Neighbor Interpolation in Three Dimensions*. Masters thesis, Brigham Young University, Provo, UT.

Rawson, H., and Evans, L. (1971). The contribution of stem reserves to grain development in a range of wheat cultivars of different height. *Aust. J. Agric. Res.* 22, 851. doi: 10.1071/AR9710851

Remondino, F., and El-Hakim, S. (2006). *Image-based 3D Modeling: A Review*. Available at: http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=8913373 [accessed March 22, 2016].

Remondino, F., Grazia, M., Nocerino, E., Menna, F., and Nex, F. (2014). State of the art in high density image matching. *Photogramm. Rec.* 29, 144–166. doi: 10.1111/phor.12063

Rusu, R. B., Marton, Z. C., Blodow, N., Dolha, M., and Beetz, M. (2008). Towards 3D point cloud based object maps for household environments. *Rob. Auton. Syst.* 56, 927–941. doi: 10.1016/j.robot.2008.08.005

Schima, R., Mollenhauer, H., Grenzdörffer, G., Merbach, I., Lausch, A., Dietrich, P., et al. (2016). Imagine all the plants: evaluation of a light-field camera for on-site crop growth monitoring. *Remote Sens.* 8:823. doi: 10.3390/rs8100823

Seber, G. A. F. (ed.) (1984). "Multivariate distributions," in *Multivariate Observations*, (Hoboken, NJ: John Wiley & Sons, Inc.), 17–58. doi: 10.1002/9780470316641.ch2

Smith, M. W., Carrivick, J. L., and Quincey, D. J. (2015). Structure from motion photogrammetry in physical geography. *Prog. Phys. Geogr.* 40, 247–275. doi: 10.1177/0309133315615805

Snavely, N., Seitz, S. M., and Szeliski, R. (2006). *Photo Tourism: Exploring Photo Collections in 3D. in ACM Transactions on Graphics*, 835–846. Available at: http://dl.acm.org/citation.cfm?id=1141964 [accessed January 3, 2017].

Tilly, N., Aasen, H., and Bareth, G. (2015). Fusion of plant height and vegetation indices for the estimation of barley biomass. *Remote Sens.* 7, 11449–11480. doi: 10.3390/rs70911449

Triggs, B., McLauchlan, P. F., Hartley, R. I., and Fitzgibbon, A. W. (1999). "Bundle adjustment—a modern synthesis," in *Proceedings of the International workshop on Vision Algorithms*, eds B. Triggs, A. Zisserman, and R. Szeliski (Berlin: Springer), 298–372.

Tumbo, S. D., Salyani, M., Whitney, J. D., Wheaton, T. A., and Miller, W. M. (2002). Investigation of laser and ultrasonic ranging sensors for measurements of citrus canopy volume. *Appl. Eng. Agric.* 18, 367–372. doi: 10.13031/2013.8587

Turner, D., Lucieer, A., and Wallace, L. (2014). Direct georeferencing of ultrahigh-resolution UAV imagery. *IEEE Trans. Geosci. Remote Sens.* 52, 2738–2745. doi: 10.1109/TGRS.2013.2265295

Turner, P., Tubana, B., Girma, K., Holtz, S., Kanke, Y., Lawles, K., et al. (2007). *Indirect Measurement of Crop Plant Height*. Stillwater, OK: Oklahoma State University.

van der Voort, D. (2016). *Exploring the Usability of Unmanned Aerial Vehicles for Non-Destructive Phenotyping of Small-Scale Maize Breeding Trials*. Wageningen: Wageningen University and Research Centre.

Vautherin, J., Rutishauser, S., Schneider-Zapp, K., Choi, H. F., Chovancova, V., Glass, A., et al. (2016). Photogrammetric accuracy and modeling of rolling shutter cameras. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* 3, 139–146. doi: 10.5194/isprs-annals-III-3-139-2016

Virlet, N., Sabermanesh, K., Sadeghi-Tehran, P., and Hawkesford, M. J. (2016). Field Scanalyzer: an automated robotic field phenotyping platform for detailed crop monitoring. *Funct. Plant Biol* 44, 143–153. doi: 10.1071/FP16163

Yin, X., McClure, M. A., Jaja, N., Tyler, D. D., and Hayes, R. M. (2011). In-season prediction of corn yield using plant height under major production systems. *Agron. J.* 103:923. doi: 10.2134/agronj2010.0450

Zhang, L., and Grift, T. E. (2012). A LIDAR-based crop height measurement system for *Miscanthus giganteus*. *Comput. Electron. Agric.* 85, 70–76. doi: 10.1016/j.compag.2012.04.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

From this first study we have developed a methodology to derive the height through UAV and ground vehicle platforms. The pipeline presented in this chapter was also applied in many other studies. A low correlation was found between the height and the biomass at a given stage. This can be easily explained by the variation in plant and ear density and the basal area. This is going to be the next topics of research in the following chapters.

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

Some key phenomic information can be derived by image analysis. The emergence of deep learning have motivated us to investigate the potential of convolutional neural network for the identification and counting of wheat ears. This trait is known to be challenging one wich more traditional images processing pipeline was not able to tackle. Traditional in-situ measurements were also prone to error measurements due to sampling problems as highlighted in this study.

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

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.

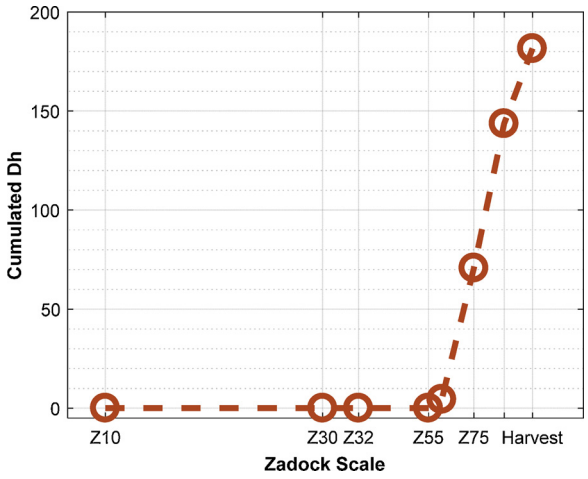


Fig. 1. Cumulated Dh for different zadocks development stages. Dh is the difference between the actual and maximum evapostranspiration values.

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

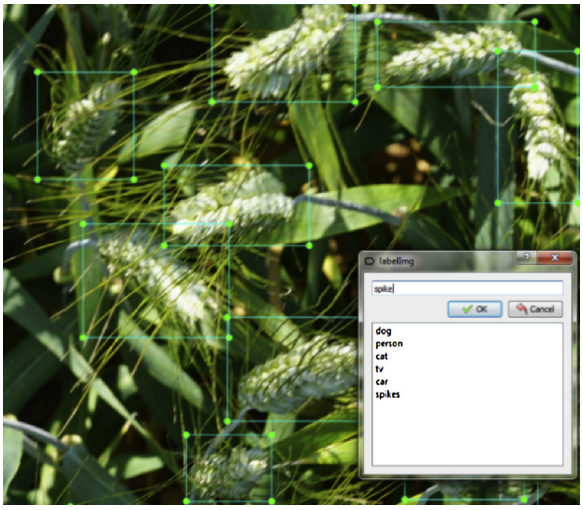


Fig. 2. Example of bounding boxes interactively drawn using the labelimg software.

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

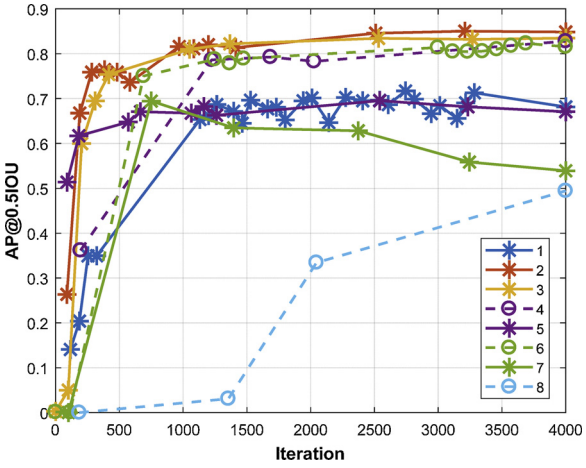


Fig. 3. AP@0.5IOU as a function of the number of iterations during the training process. The several curves correspond to the models presented in Table 2.

(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.

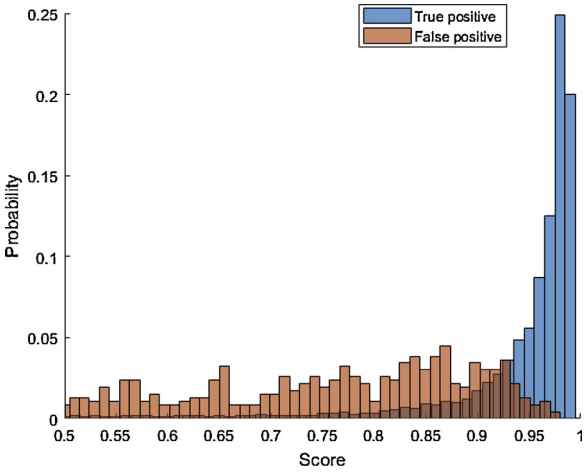


Fig. 5. Distribution of the associated scores for the true positive (blue) and the false positive (red) bounding boxes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 (≈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).

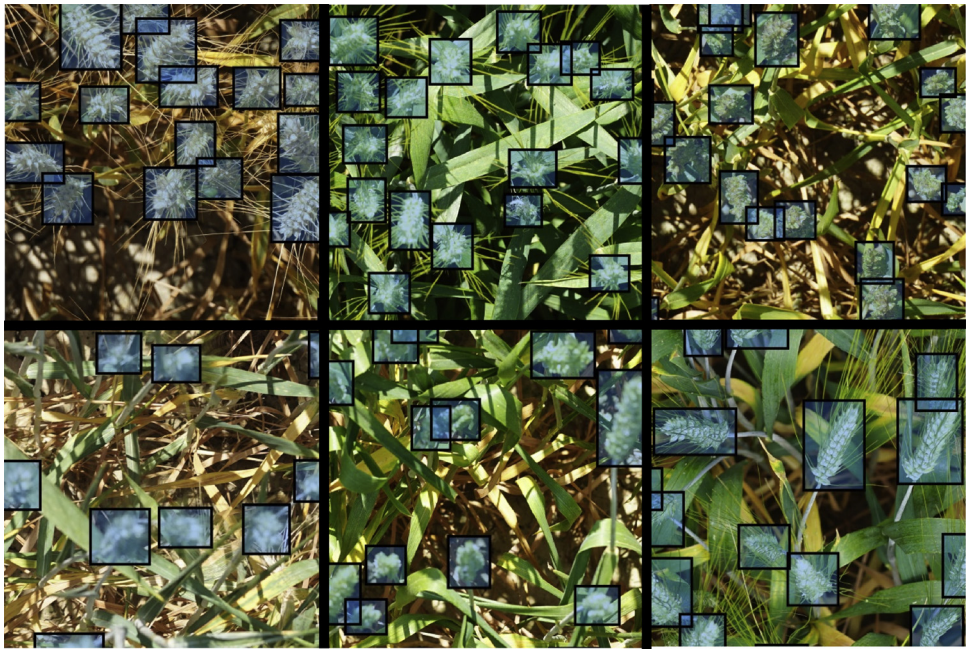


Fig. 6. Examples of sub-images (500 × 500 pixels) with ears detected by model #2.

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.

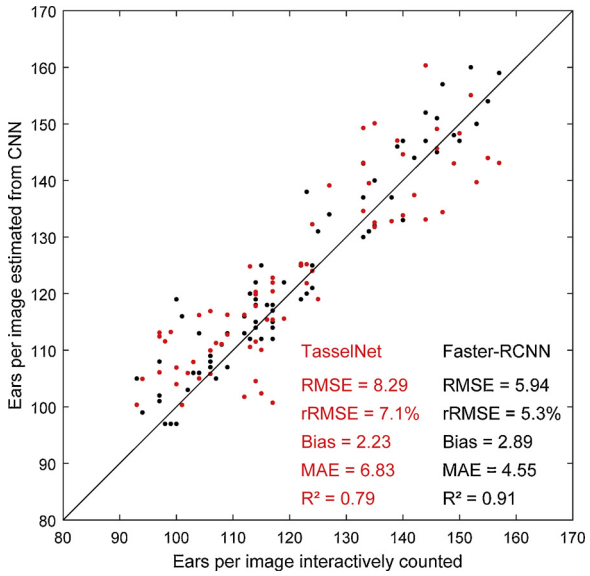


Fig. 8. Comparison between the number of ears in each image visually labelling and that estimated using either the Faster-RCNN (model #2 black dots) or the Tasselnet (model #9 red dots). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of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

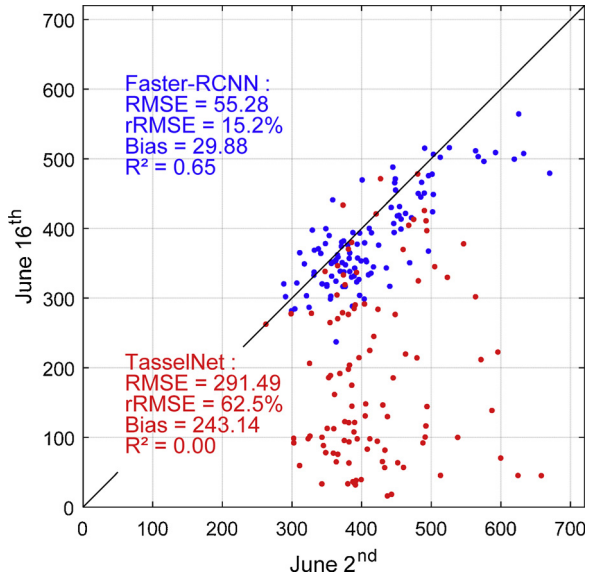


Fig. 9. Comparison between the average ear density estimated for the first (June 2nd) and the second (June 16th) dates using Faster-RCNN (blue dots) and TasselNet (red dots) models trained over the first date only. (For interpretation of the references to colour in this figure legend, the reader is referred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

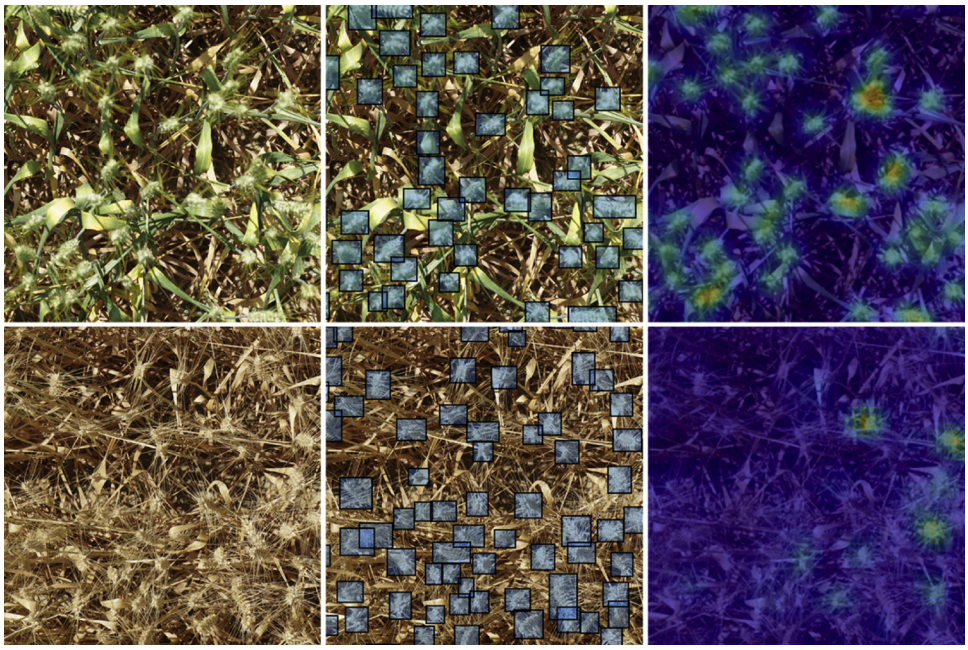


Fig. 10. Example of two images selected from the same microplot on the first and second date over which Faster-RCNN and TasselNet models were applied. The models were trained on the first date and the genotype was in the validation dataset. The top images corresponded to microplot showing little senescence, 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.

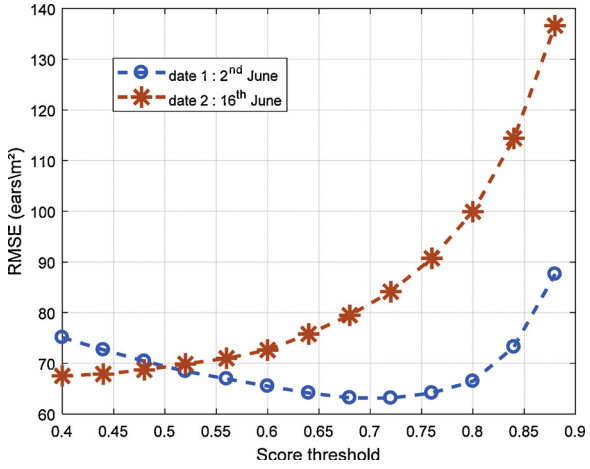


Fig. 11. RMSE of the estimated ear density as a function of the score threshold value.

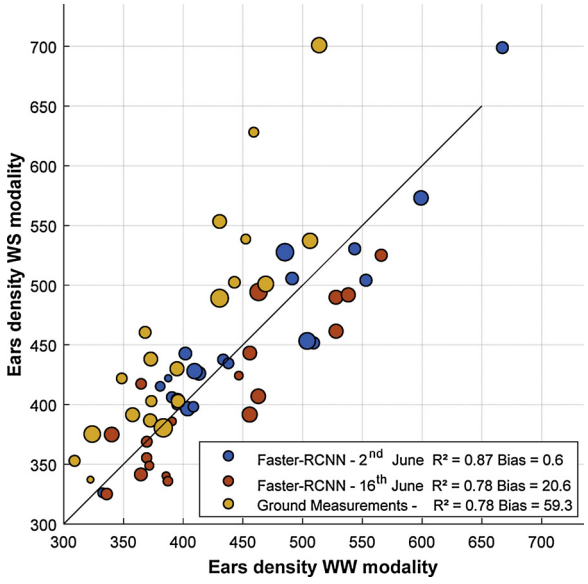


Fig. 12. Comparison between ear density estimated in the WW modality and in the WS modality (The size of the circles represented the variability of the measurements for the different replicates).

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

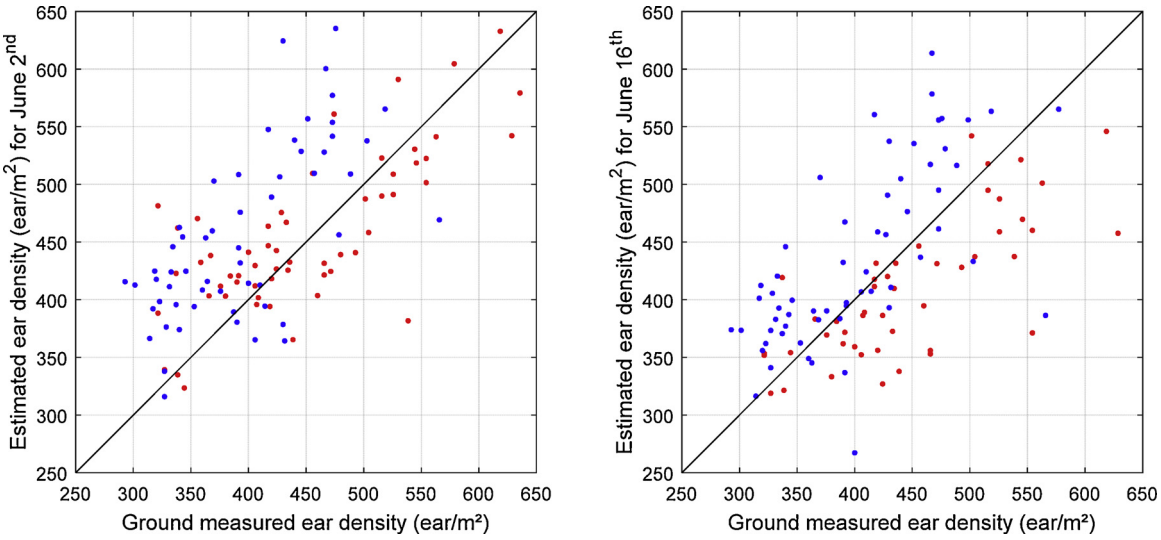


Fig. 13. Comparison between ear density estimated from the RGB images and the ground-measurements for the two modalities (red : WS; blue : WW). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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.

In the third chapter we have shown that it is possible to derive the ear density from 2D RGB images. Further studies are planned to confirm the results and to report the robustness and limitations of the methods. The training dataset is expected to grow and many other improvements are possible. Pre and post processing can be for instance also easily implemented. The regression-based approach can also be improved and studies are currently being conducted on this aspect. The two acquisition campaign had a slight change regarding the configuration of the acquisition and crops were also at a different growth stage. This is a good illustration of the problem of the non-identical distribution between the training and testing dataset. This cross-domain robustness issue also known as domain adaptation has been widely studies in the computer vision community. As discussed in the introduction, in field crop phenotyping each acquisition is subjected to a lot of parameters which make every campaign acquisition domain dependent. This highlights the urgent need to build domain adaptation algorithms for the processing of such networks in the wild.

4 High-throughput measurements of stem characteristics to estimate ear density and above ground biomass

In the previous chapters we have derived several traits from RGB images the height and number of ears through their identification. It is interesting to note that by following the plant height dynamics we were able to have an estimate of flowering time with a reasonable accuracy. However it doesn't look practicable to have a high frequency of observation with UAV. The new affordability of new IOT (Internet of Things) sensors may help to compensate for this lack of temporal information. Indeed, high resolution images can be acquired and so on the identification of the ears can be derived through this technology.

In order to have good estimates of the above ground biomass more traits and information about the plant architecture should be computed. One such trait is the stem diameter. Stem diameter is rarely measured because it is tedious and time consuming. This motivated us to transfer the methodology developed in the third chapter to the study of the stem. Indeed, it is possible to derive the stem diameter by identifying them. For this purpose, 2D images were acquired after harvest. We finally use all the traits computed and presented in this thesis do derive the above ground biomass.

High-throughput measurements of stem characteristics to estimate ear density and above ground biomass

Xiuliang Jin[a,d], Simon Madec[a], Dan Dutartre[b], Benoit de Solan[c], Alexis Comar[b] and Frédéric Baret[a]

[a]INRA EMMAH, UMR 1114 228 route de l'aérodrome, 84914 Avignon, France
[b] HIPHEN, Rue Charrue, 84000 Avignon, France
[c]ARVALIS-Institut du végétal, Station expérimentale, 91720 Boigneville, France
[d]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences/Key Laboratory of Crop Physiology and Ecology, Ministry of Agriculture, Beijing 100081, China

Correspondence information: Institute of Crop Sciences, Chinese Academy of Agricultural Sciences/Key Laboratory of Crop Physiology and Ecology, Ministry of Agriculture, Beijing 100081, China
E-mail address: jinxiuxiuliang@126.com

Abstract

Total above ground biomass at harvest and ear density are two important traits to characterize wheat genotypes. Two experiments were carried out in two different sites where several genotypes were grown under contrasted irrigation and nitrogen treatments. A high spatial resolution RGB camera was used to image the residual stems standing straight after the cutting by the combine machine during harvest. It provided a ground spatial resolution better than 0.2 mm. A Faster Regional Convolutional Neural Network (Faster-RCNN) deep-learning model was first trained to identify the stems cross section. Results showed that the identification provided precision and recall close to 95%. Further, the balance between precision and recall allowed getting accurate estimates of the stem density with a relative RMSE close to 7% and robustness across the two experimental sites. The estimated stem density was also compared with the ear density measured in the field with traditional methods. A very high correlation was found with almost no bias, indicating that the stem density could be a good proxy of the ear density. The heritability/repeatability evaluated over 16 genotypes in one of the two experiments was slightly higher (80 %) than that of the ear density (78 %). The diameter of each stem was computed from the profile of gray values in the extracts of the stem cross section. Results show that the stem diameters follow a gamma distribution over each microplot with an average diameter close to 2.0 mm. Finally, the biovolume computed as the product between the average stem diameter, the stem density and plant height were closely related to the above ground biomass at harvest with a relative RMSE of 6 %. Possible limitations of the findings and future applications are finally discussed.

Keywords: Stem density, Stem diameter, RGB image, Faster-RCNN, deep learning, Ear density, above ground Biomass

1 Introduction

Ear density (the numbers of ears per m^2) is generally well correlated with above ground biomass and grain yield at maturity of wheat [1,2]. However, the correlation may depend on environmental conditions as well as genotypes. Most stems observed at harvest bear an ear: stem density (the numbers of stems per m^2) appears thus as a good proxy of the ear density [3]. Stem density depends thus both on plant density and on the number of stems per plant which is quantified by the tillering coefficient. The environmental conditions experienced by the crop and the genotype control the tillering coefficient [4]. Therefore, several studies report the interest of the ear and stem density as traits to be used in the selection process of wheat genotypes [4,5]. Further, plant height and stem diameter are highly correlated with the above ground biomass in wheat [6-10]. Therefore, stem density, ear density, plant height, and stem diameter are thus highly desired to score the performances of a genotype in wheat crop breeding programs.

The number of stems per plant is difficult to evaluate when plants start to produce tillers since plants are often intricated and hardly identifiable. Further, the number of stems per plant may change with time due to possible tiller regression during tillering and stem elongation stages. After the flowering stage, most stems bear an ear and the stem density provides therefore a good proxy of the ear density. Ear and stem densities are therefore usually measured at maturity by manual counting over a given sample area. The stem diameter is rarely measured since it is very tedious and time consuming. Similarly, above ground biomass is rarely measured extensively for the same reasons. Crop height at harvest is most frequently measured in the field using a ruler. In addition to the limits of these low-throughput invasive measurements that require large human resources to be completed, the small sampling area used and errors associated to the manual measurements may result in significant uncertainties on these variables that would limit the repeatability and heritability as computed from the experimental observations. It appears therefore necessary to develop new methods for accurate measurements of the stem density, crop height and stem diameter for wheat crops within large field phenotyping experiments.

The recent advances in high resolution imaging systems, computing capacity as well as image processing algorithms offer great opportunities to develop nondestructive high-throughput methods. Jin et al. [11] and Liu et al. [12] have demonstrated that the plant density could be estimated at early stages in wheat crops from high-resolution imagery. Direct estimates of the tillering coefficient at the end of the tillering stage were investigated by several authors with application to the management of nitrogen fertilization for stable crops. Vegetation indices computed from the reflectance measurements have been related empirically to the tiller density [13-15]. However, reflectance measurements are mainly sensitive to the amount of green foliage, which is loosely related to the stem density. Alternatively, several authors have developed algorithms for estimating wheat stem density at early stages from high-resolution imagery [16]. Unfortunately, this method applied to plants in pots grown under greenhouse conditions is difficult to transfer to field conditions. Further, the number of stems at relatively early stages may overestimate the actual stem density at harvest because of possible tiller regression as already pointed out. Previous scientists have used algorithms for estimating wheat ear density in-field conditions using RGB or thermal imagery [17-20]. However, these techniques operated from the top of the canopy before harvest may be limited when a significant number of ears are laying in the lower layers of the canopy. Previous studies have also demonstrated that above ground biomass (AGB) can be estimated using different remote sensing platforms [21-27]. However, the correlative nature of these relationships questioned their robustness when applied outside the domain where they have been calibrated.

The aim of this study is to develop and evaluate a method to estimate stem density after the harvest. Images of the remaining stems cut by the combine machine during harvest show a clear circular cross section at their tip that could be identified by machine vision techniques. Further, the diameter of the stem could be also measured to tentatively estimate the AGB by combining the average stem diameter with the stem density and plant height. High throughput estimates of plant height have become now a standard trait easy to compute from 3D point clouds derived from LiDAR or standard cameras aboard drone [28]. The main objectives of this study are therefore: (1) to develop a method for identifying stems from post-harvest sub-millimetric RGB images and compute the stem density; (2) to compare the estimated stem density with the ear density measured with traditional invasive methods; (3) to estimate the stem diameter and describe their distribution; (4) to investigate the capacity of stem density, stem diameter and plant height to provide a proxy of AGB. The field experiments and data

acquisition are first described. The developed methods are then presented and their performances to estimate stem density, stem diameter and AGB are finally evaluated and discussed.

2 Materials and methods

2.1 Experimental sites, and ground measurements

The Gréoux and Clermont sites located in France (Table 1) were hosting wheat phenotyping experiments with about one thousand microplots of 13 rows by 10 m length (Gréoux) or 8 rows by 2.5 m length (Clermont). For both sites, rows were spaced by 17.0 cm. A subsample of microplots (Table 1) was selected in both sites for the development and validation of the method. They included genotypes with contrasted tillering capacity and plant architecture as well as variation in irrigation (Gréoux) and nitrogen (Clermont) crop management.

Table 1 Characteristics of the Gréoux and Clermont experimental sites

Sites	Latitude	Longitude	Number of plots	Sowing date	Sowing density (seeds/m^2)
Gréoux	43° 45′ N	5° 53′ E	66	27/10/2015	300
Clermont	45° 46′ N	3° 70′ E	123	02/11/2016	280

The ear density (ears/m²) was measured in the field at maturity for the 66 (Gréoux) or 123 (Clermont) microplots considered, by counting the ears over three samples of two rows by 1.0 m length corresponding to a 1.02 m² sampled area. The AGB (g/m²) was measured in Gréoux over 37 microplots by collecting all the plants within three samples of two rows by 1.0 m length. The samples were then oven-dried at 70 ℃ for three days and finally weighed. The height (cm) of the plants was measured using two LMS400 LiDARs (SICK, Germany) fixed on a phenomobile, i.e. a robot rover that automatically moved in the field and collected the measurements. More details on height measurements are given in Madec et al. [28].

2.2 Image acquisition and visual labelling of stems

A Canon EOS 550D RGB camera with a resolution of 5184 by 3456 pixels equipped with a 29 mm focal lens was fixed on a pole and maintained at a 1.2 m distance from the ground at the Gréoux experimental site. The camera was set to speed priority. The same operating mode was used in Clermont, except that the camera was a Sony ILCE-6000 with 6000 by 4000 pixels equipped with a 60 mm focal length lens and maintained at 1.8 m from the ground. The images were recorded in JPG format on the SD memory card. Measurements were completed under cloudy illumination conditions with light wind. Three (Gréoux) or four (Clermont) images were taken over each microplot. A subsample corresponding to four rows by 0.6 m length for Gréoux and four rows by 1.6 m length for Clermont (Table 2) was extracted in the center of each image. This offered the advantage to minimize image deformation observed mostly on the borders of the whole image. The Gréoux images were first resampled using a bicubic interpolation algorithm to provide the same resolution as that of Clermont.

Table 2 characteristics of the images taken over the two experimental sites.

Sites	Date of images	Distance to ground (m)	Ground resolution (mm)	Sampled area (m^2)
Gréoux	12/07/2016	1.20	0.18	1.23
Clermont	18/07/2017	1.80	0.12	4.36

The good quality of images provided strong confidence on the visual identification of the stems (Figure 1). A bounding box was interactively drawn around each stem identified in the images. The bounding box used to identify each stem was designed to include enough elements surrounding the stem (Figure 1). A total of 822 images were visually annotated to be used for the calibration and validation of the method.

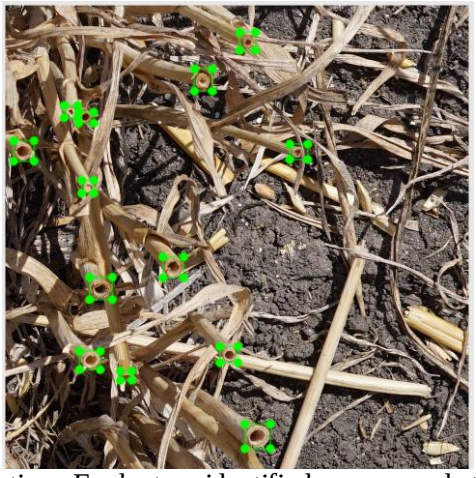


Figure 1. Visual stem identification. Each stem identified corresponds to a green bounding box. Note: The image is actually cropped from original image by 1000x1000 pixel.

2.3 Object detection using Faster-RCNN

Convolutional Neural Networks (CNNs) are powerful machine learning methods [29]. They are widely used to extract imagery information features and then classify objects. CNNs were trained using large collections of diverse images to extract more effectively rich feature representations. These CNNs features often outperform handcrafted ones such as histogram of oriented gradients (HOG), local binary patterns, or speeded up robust features [30]. The TensorFlow (https://www.tensorflow.org/) implementation of Faster Regional Convolutional Neural Network (Faster-RCNN) by the object detection application programming interface (API) [31] was implemented here. Faster-RCNN has been widely used to detect objects [32]. The region proposal network (RPN) branch was inserted between the conv4 and conv5 blocks. The Inception-Resnet-V2 model was used as it obtained the best accuracy among several modern object detectors [31]. An anchor was set at each location considered by the convolution maps of the RPN layer. Each anchor was associated with a size and aspect ratio. A set of 12 anchors with different size and aspect ratio were assigned at each location, following the default setting. The number of proposed regions per patches was set to 300, which was consistent with the expected number of stems per patch. Note that the images used here were 1000 x 1000 pixels since the memory requirements were too demanding for larger images. The original images were thus split into 1000 x 1000 patches, keeping 50% overlap between neighboring patches to minimize possible problems associated to the borders. The batch size was fixed to 1 and the threshold value for the non-maxima suppression with an IOU (Intersection Over Union) was set to 0.2. The model was trained with a learning rate of 0.0003 and a momentum of 0.90. The model was first pretrained on the Common Objects in COntext (COCO) dataset to provide the starting point. The COCO dataset [31] contains 0.33 million images with 1.5 million of object instances belonging to 80 object categories. More details on the Faster RCNN used could be found in Madec et al. [20]. The pretrained model was then fine-tuned over the calibration image extracts. It identified and localized stems using a bounding box associated with a confidence score varying between 0.0 and 1.0.

The trained model was finally applied to all the image extracts available. When identified stem bounding boxes were overlapping, a minimum 0.75 overlap fraction was used to eliminate one of the overlapping bounding boxes. Finally, bounding boxes with a confidence score value smaller than 0.80 were not considered as stems. This score threshold value was optimized to get the best stem density estimation performances. An example of the Faster-RCNN stem detection result is presented in Figure

2. The estimated stem density (stem/m²) was eventually computed by dividing the number of stems identified over the image extracts of a microplot by the size of the extracts (
Table 2).

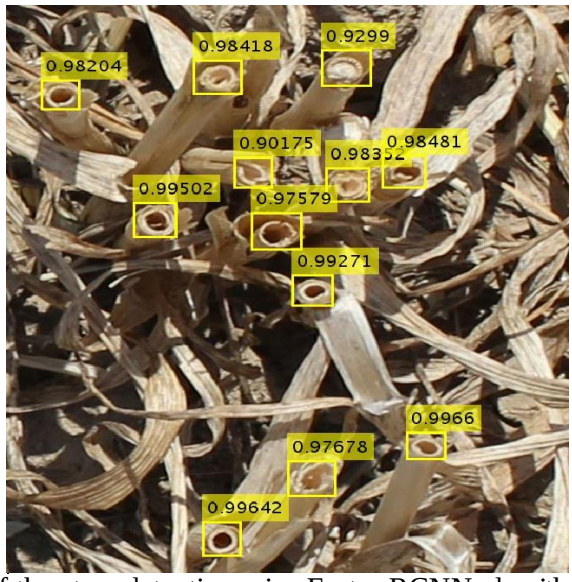


Figure 2. Application of the stem detection using Faster RCNN algorithm to an image extract in Gréoux. Each yellow bounding box corresponds to the identified stem and associated to its score corresponding to the probability to contain a stem.

2.4 Estimating the stem diameter and biovolume

The bounding box of the identified stems was first transformed into gray images using the value (V) component of the HSV transform [33]: V=0.2989R+0.5870G+0.1140B, where R, G, and B were respectively the red, green and blue components of the RGB images coded in 8 bits. The gray value profiles were then extracted along four compass directions: 0°, 45°, 90° and 135° (Figure 3). The gray level profiles show typical patterns with high values corresponding to the border of the stem and lower values outside and inside the stem (Figure 3). The two borders of the stem were thus identified using the two maximum gray value. The distance between the maximums was computed and then averaged over the four compass directions to provide an estimate of the diameter.

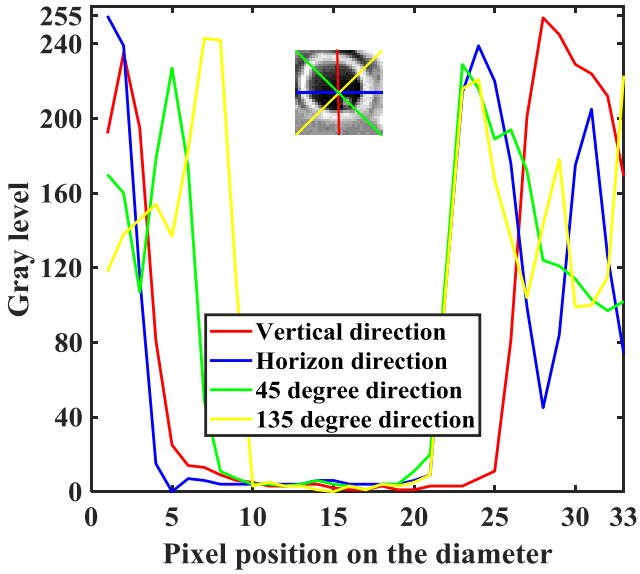


Figure 3. The extraction of stem diameter of each sub-window image using diameter gray level profile.

The stem diameter was used to compute the area of the section of the stem. The basal area of each microplot was then computed as the average area of the stem section multiplied by the stem density.

Finally, the biovolume was computed as the product of the basal area by the plant height as derived from the LiDAR measurements.

2.5 Statistical analysis

Both Gréoux and Clermont datasets were randomly split into 2/3 for model calibration and 1/3 for validation. A first global training (Called here Cgc) was investigated by pooling the calibration datasets of Gréoux and Clermont sites. The same was done for the validation datasets Vgc. The performances of this global calibration (Cgc) were also evaluated on both the Greoux (Vg) and Clermont (Vc) validation datasets. Then, a cross validation was also investigated to better evaluate the robustness of the classification: the calibration was completed on the Gréoux (Cg) or Clermont (Cc) calibration datasets and validated on the Gréoux (Vg) and Clermont (Vc) validation datasets. Table 3 presents the several cases considered.

Table 3. Characteristics of the data sets used for the calibration and validation of the algorithm. Statistics of the stem density are indicated for each data set, including minimum (Min), mean (Mean), maximum (Max), range (Range), standard deviation (SD) and coefficient of variation (CV) of the stem density.

		Number of image extracts		Stem density (stem/m²)					
Dataset	Name	Gréoux	Clermont	Min	Mean	Max	Range	SD	CV (%)
Calibration	Cgc	132	416	112	493	991	879	132	26.77
Validation	Vgc	66	208	161	561	906	745	179	31.91
Calibration	Cg	132		112	357	605	493	77	21.57
Validation	Vg	66		161	352	601	440	78	22.16
Calibration	Cc		416	308	549	991	683	101	18.04
Validation	Vc		208	373	662	906	533	116	17.52

A detected stem bounding box (i.e. with a score >0.8) was considered correct (true positive, TP) if its IOU with a labelled stem bounding box was larger than the IOU threshold value. Otherwise, the detected stem bounding box was considered as false positive (FP). The proposed bounding boxes with a score<0.8 (i.e. not considered as stems) with IOU larger than the IOU threshold value were considered false negative (FN). The IOU threshold value was set to the usual value of 0.5. The precision (TP/ (FP+TP)), recall (TP/ (FN+TP)), and bias (1-(Precision / Recall)) were also calculated.

2.6 Heritability computation

The broad sense heritability (H^2) evaluates the repeatability of the stem or ear density estimates. It was computed as the percentage of the genotypic variance, Vg, to the total variance, Vg+Ve, where Ve is the variance due to the environment [34]. The heritability of the stem density and ear density was computed over sixteen wheat genotypes (122 plots) selected from the Clermont experimental site where each genotype was replicated six to fifteen times.

3 Results and discussion

3.1 Stems are accurately identified using the Faster-RCNN model

To evaluate the robustness of the RCNN model, it was calibrated on the Greoux (Cg), Clermont (Cc) or on both datasets (Cgc). Performances computed over the validation datasets were very good with $0.91 < precision < 0.96$ and $0.93 < recall < 0.97$ (Table 4). Precision and recall were well balanced with a small bias: $-0.03 < bias < 0.01$.

Table 4. Accuracy of stem identification using the Faster-RCNN method. Results are presented for three calibration datasets (Cg, Cc and Cgc). The evaluation is achieved on the validation data set (Vg, Vc, Vgc).

Calibration dataset	Validation dataset	Precision	Recall	Bias
Cg	**Vg**	0.95	0.97	-0.02
	Vc	0.91	0.93	-0.02
	Vgc	0.92	0.94	-0.02
Cc	**Vg**	0.92	0.95	-0.03
	Vc	0.96	0.95	0.01
	Vgc	0.96	0.95	0.01
Cgc	**Vg**	0.94	0.95	-0.01
	Vc	0.96	0.95	0.01
	Vgc	0.95	0.95	0.00

The results showed that the classification accuracy of stem identification was very high based on the precision and recall values over the same experiments (Table 4). The robustness of the classification was further investigated by comparing the precision and recall values computed over the validation datasets coming from the other experiments. Results show that the classification evaluated over the same experiment used to calibrate the model was always performing the best (Table 4). The classification performances decreased significantly when the model calibrated over a single experiment was validated on the other experiment. This may be explained both by the limited sample size of the calibration dataset, but also by the specific features associated to each experiment, including the spatial resolution (Tables 2 and 3). However, when the calibration was completed over the pooled experiments (Cgc), the precision and recall values decreased only slightly when evaluated over each individual experiment (Vg or Vc) (Table 5). The model captured the key information common to the two experiments to provide a consistent stem identification. It confirmed the efficiency and robustness of the Faster-RCNN method.

3.2 Stem density is accurately estimated

The consequences of the identification performances of the Faster-RCNN model discussed previously were evaluated in terms of plant density at the image extract level. For the sake of consistency, the several calibration and validation datasets were considered to further evaluate the robustness of the model. Results showed RRMSE values ranging from 6.08 % to 9.19 %. Calibrating over the pooled datasets (Cgc, Table 5) provided the best performances with RRMSE lower than 7%. A slight degradation of the performances was observed when calibrating over a single data set. Calibrating over the Greoux dataset provided the worst performances when validated over the Clermont dataset (RRMSE=9.2%) because of the smaller sample size and variation in the cutting height and inclination of the stems during harvest between Gréoux and Clermont sites. In the following, we have used the Faster RCNN trained over the pooled Gréoux and Clermont calibration datasets (Cgc) that provided the more robust performances.

Table 5. Performances of the stem density estimation when using Faster-RCNN method for the post-classification step. The evaluation is achieved on the three validation data sets (Vg, Vc and Vgc).

Calibration dataset	Validation dataset	Sample size	slope	intercept	R^2	RMSE (stems/m^2)	RRMSE (%)
Cg	**Vg**	66	0.98	25.63	0.95	24.45	6.95
	Vc	208	0.94	41.25	0.85	60.82	9.19
	Vgc	274	0.95	38.12	0.90	40.23	7.17
Cc	**Vg**	66	0.95	20.85	0.88	30.52	8.67
	Vc	208	0.97	17.14	0.96	40.25	6.08
	Vgc	274	0.96	30.28	0.92	39.66	7.07

Cgc	**Vg**	66	0.95	33.12	0.91	25.46	7.23
	Vc	208	0.97	45.33	0.92	42.81	6.47
	Vgc	274	0.98	38.16	0.94	38.45	6.85

When considering the calibration over the pooled dataset (Cgc) that provided the overall best performances, very small biases were observed with points closely distributed around the 1:1 line (Figure 4). The scatter around the 1:1 line appeared to be relatively independent from the stem density (Figure 4).

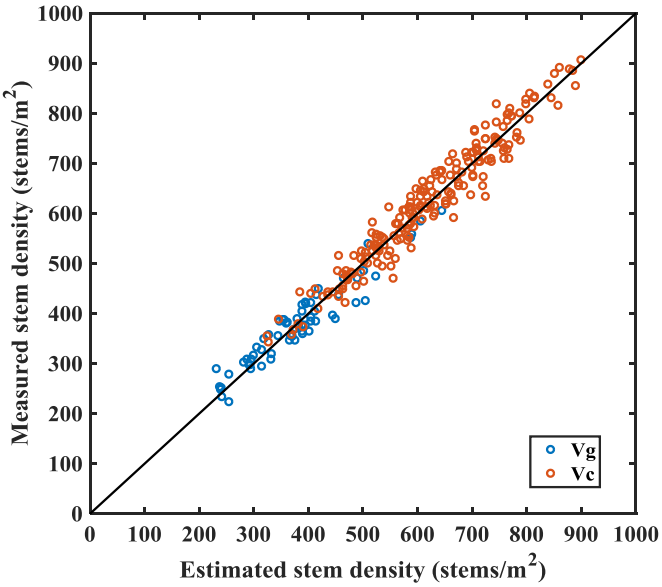


Figure 4. Comparison between the stem density estimated using Faster-RCNN method calibrated over the pooled Cgc dataset and the stem density evaluated visually over the images. The black line corresponds to the 1:1 line, the red and blue circles corresponding respectively to the Gréoux (Vg) and Clermont (Vc) validation datasets.

3.3 The stem density is a good proxy of the ear density

The stem density estimated with the Faster-RCNN model calibrated over the Cgc dataset was compared to the ear counted visually at the ground level. Both quantities were evaluated on different samples expected however to represent the average microplot value. Results showed that the estimated stem density based on the Faster-RCNN model was very consistent (Figure 5) with the measured ear density at the Gréoux (Table 6) and Clermont (Table 6) experimental sites. The scatter between ear and stem densities appeared to increase with the density: this was obvious between the Gréoux (250<density<550) and Clermont (400<density<800) sites. Part of the larger scatter observed over the Clermont site might come from a smaller sample size for the ear density visual counting (1.02 m²). The scatter between ear and stem densities seems to increase with the density within the Clermont site between the low and high densities (Figure 5). Nevertheless, the good agreement found between ear and stem densities was thus confirming the results of Siddique et al. [3].

Previous studies demonstrated that the RGB imagery can be used to estimate ear density using image processing algorithms [17-20]. However, ear density estimation performances were generally limited to a comparison between the ears detected by the machine learning algorithm and those that can be visually identified by and operator on the image. Some discrepancies could appear when comparing with the actual ear density, particularly when some ears are lying in the lower canopy layers and could not be easily seen from the top of the canopy. Counting ears from the stem sections appears therefore preferable under such conditions. Further, stem sections are relatively simpler objects to identify as compared to ears that may show a large aspect variability. Additionally, ears can frequently overlap in the field, making their identification more complex as compared to stem sections that never overlap.

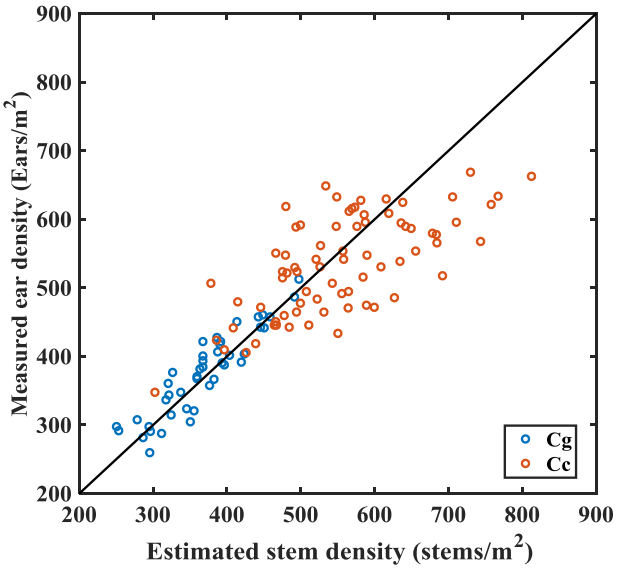


Figure 5. Relationship between the estimated stem density and the measured ear density at the Gréoux (blue dots) and Clermont (red dots) datasets. The black line corresponds to the 1:1 line.

Table 6. Statistics of the relationships between the estimated stem density and the measured ear density. The Faster-RCNN was trained over the Cg+Cc dataset.

Datasets	Slope	Intercept	R^2	RMSE (stems/m²)	RRMSE (%)
Gréoux	0.92	32.04	0.83	24.67	6.54
Clermont	0.56	236.08	0.51	53.32	9.51
Gréoux & Clermont	0.76	111.00	0.80	45.73	9.22

3.4 Stem and ear densities are highly heritable

The Heritability (H^2) values of the stem density and ear density were compared at the Clermont experimental site where several replicates of 16 genotypes were available. Results show that the H^2 values of stem density (80.1%) and ear density (78.3 %) were high and close together. This is consistent with the strong relationship found between both quantities (Figure 5). These heritability values agreed well with the values provided by Madec et al. (2019) [20]. The H^2 value of the stem density was slightly higher than that of the ear density, probably because of the larger sample size used for estimating the stem density from the RGB images, which makes the values more repeatable. The high values of heritability found suggested that the proposed method will be well suited to serve the breeders needs.

3.5 Stem diameter follows a gamma distribution

The distribution of stem diameter was investigated at Gréoux and Clermont experimental sites respectively on 66 and 156 microplots. The distribution of the stem diameter may be a pertinent trait describing the structure of the tiller population that may be impacted by the growth conditions. The distribution of the stem diameter of each microplot was adjusted either to a normal or to a gamma distribution. The corresponding p values associated to the fit of each distribution was computed. Results show that the p value of the gamma distribution was larger than that of the normal distribution for 80% of the microplots for Gréoux and 84% of the microplots for Clermont. The gamma

distribution characterized by a scale and a shape parameter was therefore selected to describe the stem diameter distribution over each microplot.

The average stem diameter of each microplot ranged between 1.8 to 2.5 mm, with a median value close to 2.0 mm for both sites (Figure 6). The average stem diameter was loosely but positively correlated to the stem density (Figure 6): the stress experienced by the plants was affecting both the density and the diameter of the stems, with no apparent compensations between these two traits. The stem diameter distributions for each microplot as described by a gamma function was further investigated: shape parameters were slightly smaller for the Gréoux site (4<shape<8) as compared to those of the Clermont site (5<shape<10). Conversely, scale parameters were slightly larger for the Gréoux site (0.3<scale<0.6) as compared to the Clermont site (0.2<scale<0.5). The distribution of the diameters was more concentrated around the average for the Clermont site as compared to the Gréoux site where a larger range of diameters was observed. This may be related to the stress conditions that were stronger in Gréoux, particularly during the stem elongation phase. This was also reflected by the stem density that was more impacted in Gréoux. The scale and shape parameters were negatively correlated for both sites, with a stronger correlation for Clermont (Figure 6). Since the average of a gamma distribution is defined by the product between the shape and scale parameters, the negative correlation between the two parameters was explained by the constraint to keep the average close to 2.0 mm. Therefore, both parameters could be equally used to describe the 'flatness' of the stem diameter distribution.

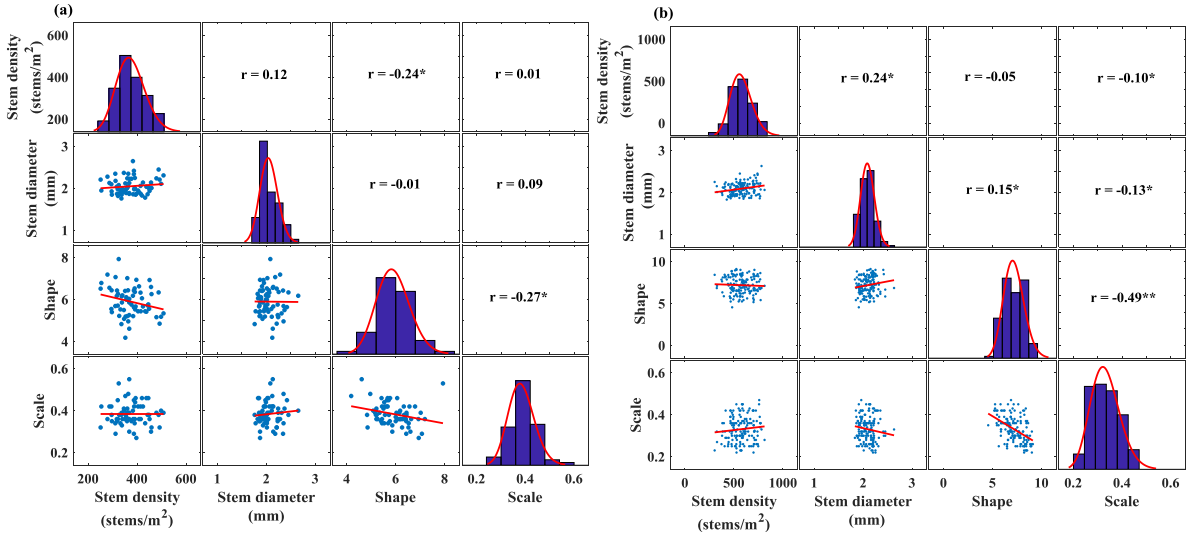


Figure 6. Correlation and distribution between stem density, average stem diameter, shape, and scale parameters over Gréoux (a) and Clermont (b) experimental sites. The correlation coefficient, *r*, is given in the upper trianglar matrix with ** and * corresponding respectively to significant values at 0.01 and 0.05 probability levels.

3.6 The biovolume is a good proxy of the above ground biomass

A total of 37 microplots from the Gréoux dataset was used to relate the measured AGB with the ear density and the four structural traits derived from high-throughput measurements: stem density, stem basal area computed as the product between the average stem diameter with the stem density, plant height, and biovolume computed as the product between the basal area and plant height. Results show that all these traits are strongly correlated with AGB (Figure 7 and Table 8). The best relationship is however obtained using the biovolume that combines the three main original traits: stem density and average stem diameter that combined into the basal area, and plant height. Note that these traits are relatively independent: stem density and plant height are loosely correlated (Figure 6, r^2=0.16); plant height and basal area are also loosely correlated (r^2=0.15).

Because the data set used was limited, the predictive performances of the relationships observed between the AGB and the five traits was evaluated using a leave-one-out cross-validation method (Jin et al., 2018). The best determination coefficients were observed consistently for the biovolume (Table 8) with a relative error of 5.8%, i.e. within the order of magnitude of the accuracy with which AGB

was measured. Our results are very consistent with those presented by Aziz et al. (2004) and Pittman et al. (2015).

To further improve the predictive model, we used all the five traits together within a multiple linear regression model. Marginal improvement of the model performances was observed (Figure 7 and Table 8). This may be explained by the strong relationships between the five traits used, as well as the decrease in the degree of freedom induced by the increase of the number of coefficients to adjust (six coefficients instead of two needed when using only the biovolume). The biovolume appeared therefore as a very sound proxy of the AGB. Previous results suggested that AGB could be estimated using different optical techniques and technologies [21-27]. Our study further confirmed these results. The results demonstrated that the estimation accuracy of AGB could be improved by combining Lidar data and RGB imagery. However, the stability of the relationship found over the limited sample used in this study should be further evaluated with emphasis on the possible dependency on the environmental and management conditions, as well as on differences between genotypes.

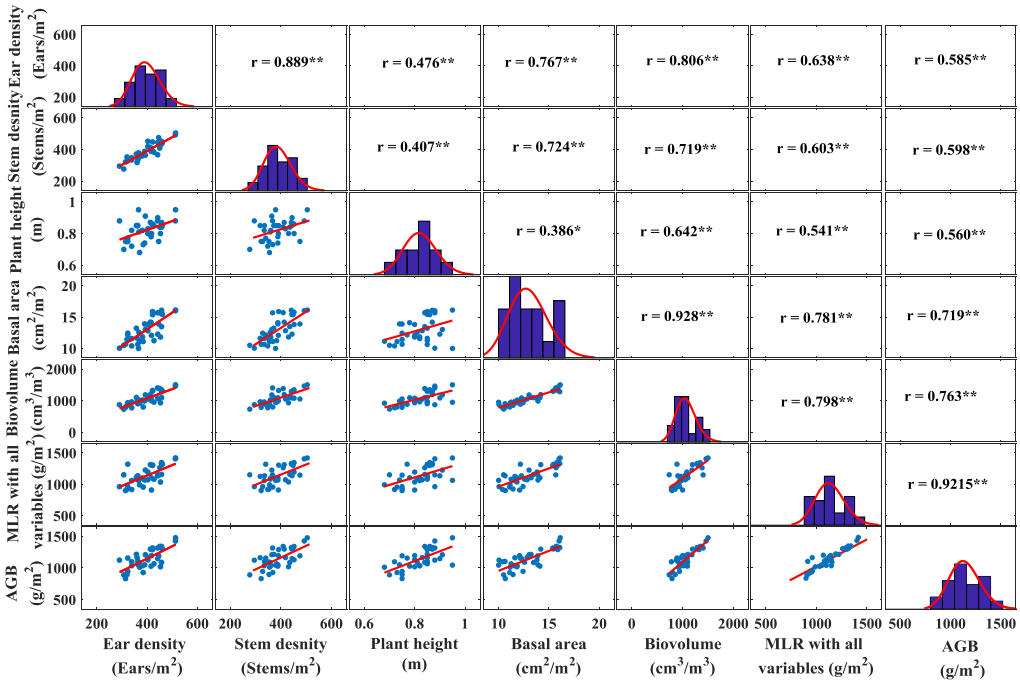


Figure 7. Correlation matrix between the AGB and the six variables investigated. Note: ** and * correlation significant at the 0.01 and 0.05 level of probability, respectively.

Table 7. Biomass regression models derived from stem density, ear density, stem area, height, and biovolume at the Gréoux experimental site. Note: ** Model significant at the 0.01 level of probability. The R^2, RMSE, and RRMSE are averaged R^2, RMSE, and RRMSE values of leave one-out cross-validation methods.

Variables	R^2	RMSE (g/m^2)	RRMSE (%)
Stem density	0.43**	94	8.0
Ear density	0.51**	84	8.1
Plant height	0.44**	110	9.8
Basal area	0.64**	77	7.3
Biovolume	0.81**	62	5.8
All	0.85**	58	5.1

4 Conclusion

This study demonstrated that the identification of the stems after the harvest was possible using deep-learning approaches applied to RGB images. This requires the spatial resolution to be sufficient, i.e. around 0.2 mm since the stem diameters are around 2.0 mm. It ensures that the object to be identified within the image are represented with an optimal number of pixels comprised between 40 and 120 pixels as advised by Madec et al. [20]. Such high resolution could be achieved using either a high-resolution RGB camera fixed on a pole, on a cart, on a phenomobile or even on a UAV flying at low altitude as already demonstrated by Jin et al. [11]. Alternatively, a set of RGB cameras could be mounted on the combine machine and provide in near real time an estimate of the stem density.

The method requires the stems not to be covered by the straw rejected by the combine machine. Further, too inclined stems due to the harvest process or some post-harvest practice may result in degraded performances since the sections of the tip of the stems will not be viewed by the camera, or will be strongly deformed. Further, the proposed method may be not suitable under stem lodging situations where the stem sections will show unexpected patterns. Nevertheless, the objects to be identified are relatively simple, which would indicate that the Faster RCNN model trained over the data used in this study would be robust. Changes in the illumination conditions may have little impact of the stem identification since the objects are mostly identified by the relative brightness of the pixels, the color bringing itself very little information. We demonstrated therefore that the stem density is accessible with high-throughput, relatively low cost and with a very good accuracy. Further, the capacity to sample large area to estimate the stem density will minimize the impact of the spatial variability within a microplot.

Although Madec et al. [20] amongst others demonstrated that similar deep learning techniques could be applied efficiently to estimate the ear density, ear identification is more complex because of strong differences of the ear aspect between cultivars and environment, as well as because of possible occlusion of some ears by the top ears or the top leaves. We demonstrated in this study that the stem density was a very close proxy of the ear density although some discrepancy is expected under specific environmental conditions. In such circumstances, the distribution of the diameter of the stems could potentially provide the necessary information to get a better estimate of the ear density from the stem density and diameter distribution.

Once the stem is identified, we demonstrated that the diameter could be easily measured. The distribution of the stem diameters followed a gamma function with an average diameter close to 2.0 mm. The distribution of the stem diameters may be indicative of the structure of the tiller population that may be governed by the genetics in interaction with the sowing density and pattern as well as the environmental conditions experienced by the plants. Finally, the biovolume computed as the product between the average stem diameter, the stem density and plant height were demonstrated to be a close proxy of the above ground biomass. This opens very attractive potentials to the breeders for getting a high-throughput estimates of the total plant biomass at harvest and possibly quantify the radiation use efficiency and the harvest index assuming that the yield will be measured anyway. Nevertheless, these promising results should be verified under a much larger number of situations to verify that the correlations are not too dependent on the environmental conditions as well as on the genotype.

Aknowledgements

This study was supported by “Programme d’investissement d’Avenir” PHENOME (ANR-11-INBS-012) with participation of France Agrimer and “Fonds de Soutien à l’Obtention Végétale”. We thank also the people from Gréoux, and Clermont who participated to the experiments. The work was completed within the UMT-CAPTE funded by the French ministry of Agriculture.

References

[1] K. Joseph, M. Alley, D. Brann, and W. Gravelle, “Row spacing and seeding rate effects on yield and yield components of soft red winter wheat,” *Agronomy Journal*, vol. 77, no.2, pp.211-214, 1985.

[2] J. Whaley, D. L. Sparkes, M. J. Foulkes, J. H. Spink, T. Semere, and R. K. Scott, "The physiological response of winter wheat to reductions in plant density," *Annals of Applied Biology*, vol.137, no.2, pp. 165-177, 2000.
[3] K. Siddique, E. Kirby, M. Perry, "Ear: stem ratio in old and modern wheat varieties; relationship with improvement in number of grains per ear and yield," *Field Crops Research*, vol.21, pp. 59-78, 1989.
[4] J. Hiltbrunner, B. Streit, and M. Liedgens, "Are seeding densities an opportunity to increase grain yield of winter wheat in a living mulch of white clover?" *Field Crops Research*, vol.102:, pp. 163-171, 2007.
[5] C. M. T. Donald, "The breeding of crop ideotypes," *Euphytica*, vol.17, pp. 385-403, 1986.
[6] C. Law, J. Snape, and A. Worland, "The genetical relationship between height and yield in wheat," *Heredity*, vol.40, pp. 133-151, 1978.
[7] C. Law, J. Snape, and A. Worland, "Aneuploidy in wheat and its uses in genetic analysis." *Wheat Breeding*. Springer. pp. 71-108, 1987.
[8] A. J. King, L. R. Montes, J. G. Clarke, et al., "Identification of QTL markers contributing to plant growth, oil yield and fatty acid composition in the oilseed crop Jatropha curcas L," *Biotechnology for Biofuels*, vol.8. no.1, pp.160, 2015.
[9] S. Arnoult, M. C. Mansard, and M. Brancourt-Hulmel, "Early Prediction of Biomass Production and Composition Based on the First Six Years of Cultivation," *Crop Science*, vol.55, pp. 1104-1116, 2015.
[10] J. Subira, K. Ammar, F. Álvaro, L. F. G. del Moral, S. Dreisigacker, and C. Royo, "Changes in durum wheat root and aerial biomass caused by the introduction of the Rht-B1b," *Plant and Soil*, vol.403, pp.291-304, 2016.
[11] X. L. Jin, S.Y. Liu, F. Baret, M. Hemerlé, and A. Comar, "Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery," *Remote Sensing of Environment*, vol.198, pp.105-114, 2017.
[12] S. Y. Liu, F. Baret, D. Allard, et al., "A method to estimate plant density and plant spacing heterogeneity: application to wheat crops," *Plant Methods*, vol.13, p.38, 2017.
[13] M. Flowers, R. Weisz, and R. Heiniger, "Remote sensing of winter wheat tiller density for early nitrogen application decisions," *Agronomy Journal*, vol.93, no.4, pp.783-789, 2001,
[14] S. Phillips, D. Keahey, J. Warren, and G. Mullins, "Estimating winter wheat tiller density using spectral reflectance sensors for early-spring, variable-rate nitrogen applications," *Agronomy Journal*, vol.96, no.3, pp. 591-600, 2004.
[15] I. Scotford, and P. Miller, "Estimating tiller density and leaf area index of winter wheat using spectral reflectance and ultrasonic sensing techniques," *Biosystems engineering*, vol.89, no.4, pp. 395-408, 2004.
[16] R. D. Boyle, F. M. K. Corke, J. H. Doonan, "Automated estimation of tiller number in wheat by ribbon detection," *Machine Vision and Applications*, vol.27, no.5, pp. 637-646, 2016.
[17] J. A. Fernandez-Gallego, S. C. Kefauver, N. A. Gutiérrez, M. T. Nieto-Taladriz, and J. L. Araus, "Wheat ear counting in-field conditions: high throughput and low-cost approach using RGB images," *Plant methods*, vol.14, no.1, pp.22, 2018.
[18] J. A. Fernandez-Gallego, M. Buchaillot, N. Aparicio Gutiérrez, M. T. Nieto-Taladriz, J. L. Araus, and S. C.Kefauver, " Automatic wheat ear counting using thermal imagery," *Remote Sensing*, vol.11, no.7, pp. 751, 2019.
[19] T. Liu, C. M. Sun, L. Wang, X. C. Zhong, X. K. Zhu, and W. S. Guo, "In-field wheatear counting based on image processing technology," *Nongye Jixie Xuebao/ Transactions of the Chinese Society for Agricultural Machinery*, vol. 45, pp. 282-290, 2014.
[20] S. Madec, X. L. Jin, H. Lu, et al., "Ear density estimation from high resolution RGB imagery using deep learning technique," *Agricultural and Forest Meteorology*, vol.264, pp.225-234, 2019.
[21] R. Ballesteros, J. F. Ortega, D. Hernández, and M. A. Moreno, "Applications of georeferenced high-resolution images obtained with unmanned aerial vehicles. Part I: Description of image acquisition and processing," *Precision Agriculture*, vol.15, no.6, pp. 579-592, 2014.
[22] S. Gao, Z. Niu, N. Huang, and X. H. Hou, "Estimating the leaf area index, height and biomass of maize sing HJ-1 and RADARSAT-2," *International Journal of Applied Earth Observation and Geoinformation*, vol.24, pp.1-8, 2013.

[23] X. L. Jin, G. J. Yang, X. G. Xu, et al., "Combined multi-temporal optical and radar parameters for estimating LAI and biomass in winter wheat using HJ and RADARSAR-2 data," *Remote Sensing*, vol.7, pp.13251-13272, 2015.
[24] X. L. Jin, Z. H. Li, G. J. Yang, et al., "Winter wheat yield estimation based on multi-source high-resolution optical and radar imagery and AquaCrop model using particle swarm optimization algorithm," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol.126, pp.24-37, 2017.
[25] J. Liu, E. Pattey, J. R. Miller, H. McNairn, A. Smith, and B. Hu, "Estimating crop stresses, aboveground dry biomass and yield of corn using multi-temporal optical data combined with a radiation use efficiency model," *Remote Sensing of Environment*, vol.114, 1pp.167-1177, 2010.
[26] N. Tilly, D. Hoffmeister, Q. Cao, et al., "Multitemporal crop surface models: Accurate plant height measurement and biomass estimation with terrestrial laser scanning in paddy rice," *Journal of Applied Remote Sensing*, vol.8, no.1, pp.83671, 2014.
[27] J. B. Yue, G. J. Yang, C. C. Li, et al., "Estimation of winter wheat above-ground biomass using unmanned aerial vehicle-based snapshot hyperspectral sensor and crop height improved models," *Remote Sensing*, vol.9, no.7, pp.708, 2017.
[28] S. Madec, F. Baret, B. de Solan, et al., 2017. "High-Throughput Phenotyping of Plant Height: Comparing Unmanned Aerial Vehicles and Ground LiDAR Estimates," *Frontiers in Plant Science*, vol.8, pp.2017, 2017.
[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097-1105, 2012.
[30] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009.
[31] J. Huang, V. Rathod, C. Sun, M. et al., "Speed/accuracy trade-offs for modern convolutional object detectors," ArXiv161110012 Cs [Internet]. 2016, Available from: http://arxiv.org/abs/1611.10012.
[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *In Advances in Neural Information Processing Systems*, pp. 91-99, 2015.
[33] G. W. Meyer, and D. P. Greenberg, "Perceptual color spaces for computer graphics," *ACM SIGGRAPH Computer Graphic*, vol.14, no.3, pp.254-261, 1980.
[34] M. R. Dohm, "Repeatability estimates do not always set an upper limit to heritability," *Functional Ecology*, vol.16, pp. 273-280, 2002.

5 Conclusion and perspectives

The main objective of plant breeders is to create and identify genotypes that are well adapted to the targeted environments including future growing conditions. The identification of the corresponding genotypes or the QTLs associated to the desired structural and functional traits depends on the quality of the phenotype predictions. The private and public investment in phenotyping infrastructures has paved the way for collecting large amount of data. The bottleneck has now shifted to the extraction of the pertinent traits [1]. Focus should therefore be put on the development of methods for trait derivation from the phenotyping data available, to evaluate their performances and benchmark the associated algorithms.

The main objective of this thesis was to develop methods to estimate few main structural traits in the context of plant breeding. Further, this thesis was undertaken in the context of rapid hardware and software technological advancements illustrated by the increasing accessibility to UAV and UGV platforms, the decreasing cost of the processing units (GPUs, cloud computing) and the boom in the development of deep learning algorithms. In the first section, we will sum up the main findings and limitations of the proposed workflows.

5.1 Summary of the pipeline developed and the methodologies used

In the second chapter, we focused on the development and comparison of high throughput methodologies for height estimation from RGB images and structure from motion (SFM) algorithms. The technique of structure from motion is now well developed and mature. Acquired from a UAV platform, the throughput, the accuracy and precision of the retrieved traits and the cost of the method are now suitable to be integrated in most breeding programs. Manual observations are labor intensive and prone to errors. We therefore relied on frequent LiDAR measurements as a reference for height estimation from the UAV observations. Best results with the RGB images were achieved over wheat crops with a spatial resolution of 1 cm. Our workflow differs from the literature because it directly extracts the height and soil level from the 3D dense cloud. This avoids sampling, filtering and other post-processing steps from GIS software to compute the digital surface model (DSM). We have defined the height as the 99.5 % rank percentile and the height distribution within a 0.5x0.6 m² subsample area from which the microplot value is computed as the median of all the corresponding subsamples. This strategy was also further applied in our other research and [2]. Many other studies have concluded positively about the capability of UAV platform for plant height estimation for genomic selection. This confirms that the broad-sense heritability computed with these methods is suitable in the context of plant breeding.

It is interesting to note that slight underestimation of the height was observed which was further analyzed in another study (experiments) [3]. Bias introduced by the used methodology will be discussed in the next section. A critical step in the workflow is the soil level estimation. Our study has shown that it is either possible to extract the soil level from another or from the same

acquisition campaign. Separation between ground and vegetation points was performed with a semi-supervised machine learning algorithm (k-means). This classification step takes as input the RGB value of the points along with their heights. Contextual information may also add significant information which can be done using more complex features or the use of convolutional neural network (CNN).

The next two chapters describe how to employ CNN to derive wheat ear and stem density. The wheat organ density estimation is a challenging trait for plant breeders. CNN algorithms are now surpassing hand-crafted algorithms in computer vision. A methodology was proposed to evaluate their feasibility for this phenotyping activity. We compared two different approaches: counting by detection and counting by regression. In the first case the output of the algorithms are boxes around the ears along with a classification score, whereas in the second case the output of the CNN is a direct count. Both approaches achieved impressive results on the test datasets, suggesting that RGB-based images could replace in-situ measurements. For the Faster-RCNN object detection used here, the best results were obtained for a spatial resolution of around 0.3mm while the regression approach did not require such high spatial resolution. In machine learning and especially in deep learning, the choice of the training data fundamentally influences a model's behavior. For this reason we have tested the model on another acquisition campaign where the acquisition procedure was slightly different and the visual appearance of the ears were different as well (more senescent). The Faster-RCNN detection pipeline was able to generalize in a new scene while the regression approach failed. This result shows the importance of the "domain" concept and suggests that transfer learning where large datasets are used to calibrate the model helps to increase the robustness of the model.

In the third chapter we addressed the challenge of stem counting. The stem detection was performed by imaging after harvest and was based on the same methodology presented in Chapter 2. We showed that it was possible to count the residual stems standing straight after the harvest by the combine machine from high resolution RGB images. Results showed good correlation with the in-situ ear density with an RMSE of 45.73 ears/m². This was lower than the ear density estimation by ear detection of 68.7 ears/m² observed for the experiment presented in the Chapter 2 although the experiments were different. In this third study, we also derive the basal area by computing the gray level profiles of the stems. Although we think that more experiment should be conducted to validate the proposed methodology, standing stem detection after the harvest seems to be a promising phenotyping activity.

5.2 Replacing in situ-measurements – Bias – Heritability

In the three studies presented, results were really close to the ground truth. Indeed, the ground truth and manual/visual measurement were often examined and questioned. Plant height derived from UAV was demonstrated to yield similar degree of accuracy as compared to LiDAR observations from the Phénomobile. The ear and stem detection were almost as good as the human interactive count while sometimes outperforming them. Our study suggests that image-based phenotyping can replace manual measurements for certain traits. This automation of the phenotyping activities with the development and cost-efficiency of the image-based phenotyping platform allows us to estimate the traits using larger sampling area. This has the advantage to increase the repeatability of the derived traits. This was the reason why the high broad sense heritability (H^2) was computed in the three studies. H^2 from the manual in-situ measurements were always worse than the values computed with our algorithms.

However pipelines were always introducing biases in the results. This was the case with the plant height estimated from structure from motion algorithms where an underestimation was present. In the third chapter, uncertainties attached to the ground measurements had complicated the analysis of a potential Bias. Nevertheless, when comparing the density estimates from two different acquisition campaign, a Bias was present which was confirmed in further unpublished studies. Phenotyping increasingly depends on machine learning models, while the model behavior will subsequently depend strongly on the training data. This can lead to Biases which are not easy to control. For example, the regression-based counting models surprisingly failed while counting ears with advanced senescence; further, our counting models were detecting more easily ears with awns with potential biases introduced between genotypes with and without awns. In other disciplines where deep learning is in a more advanced stage such as for gender classification, words and articles embedding, machine learning models have been shown to exhibit racial biases [4], [5] leading to hiring discrimination [6]. It is therefore important to put effort towards creating a detailed annotation, avoiding the lack of diversity in our phenotyping information test dataset. Larger broad-sense heritability may also be computed if a genotype-specific biases and error is computed. The future study should also report results on those datasets. This was one of the motivations behind the freely accessible datasets which were published during my INRA contract [7]. This development of public dataset is a remaining challenge regarding the private economical context of crop phenotyping development.

Under field conditions, measurements are often subject to a lot of variations. In addition, the complex GxExM interactions add large variability to the visual aspect of the crops. The illumination and uncontrolled conditions of acquisition make two different acquisition campaigns domain-specific. This may be one of the main reasons why empirical relationships are often proposed to assess canopy state variables in the literature [8], [9] as self-calibration is a pragmatic and sure approach for phenotyping. It also has a twofold benefit because it could then help the continuous increase in the size and diversity of the annotated datasets. To get rid of this self-calibration our networks and pipeline have to be able to generalize. I believe that our deep learning technology is not mature enough to apply the models in the wild. Possible improvement and developments regarding the convolution neural networks workflow will be discussed in the next section.

5.3 Perspectives

5.3.1 Convolutional neural networks

There is a huge and growing interest in the deep learning technology. A keyword-based search from the web scientific indexing service Web of Science using the following query was performed and illustrated:

[Deep Learning] or [CNN] or [Convolution neural network]

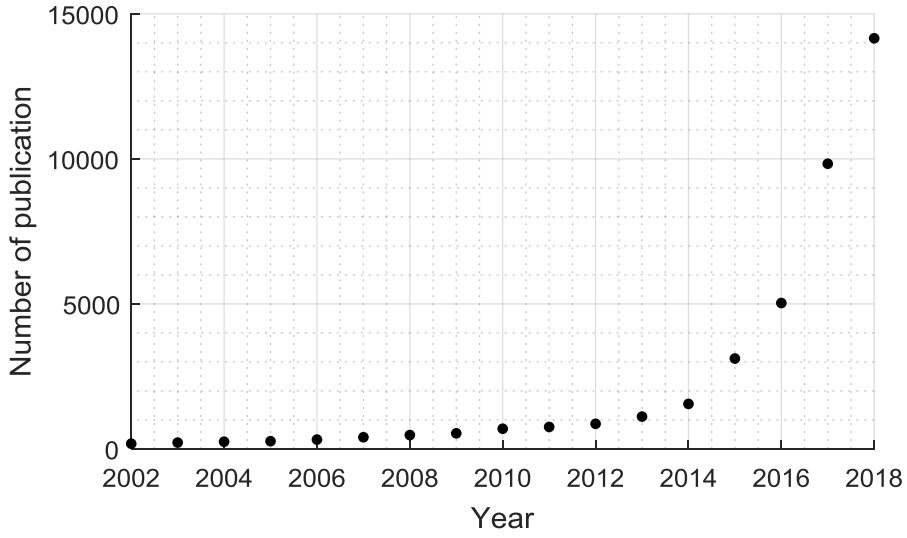


Figure 1 : Illustration of the increased interest of deep learning in research

As reviewed in the introduction and by this illustration, model architecture and their hyper-parameters optimization are evolving rapidly. For instance in the object detection field, improvements were recently made by incorporating multi-scale training [10] and deformable layers [11], [12], improvement of the loss definition and the Non-Maxima-Suppression (NMS) [13].

In our third chapter; counting was performed within local patches. This approach failed when incomplete objects were present within the patches. Accurate number of ears is hard to obtain without contextual information. Future model architecture should take into account this information. Further, robustness can be also added to regression-based counting approaches by adopting transfer-learning (pre-training the models on large scale dataset). These developments are the subject of a paper recently submitted thanks to collaborations with Huazhong University of Science.

Regarding the training strategy, it is certainly interesting to train the models on many categories from numerous datasets. This is supposed to improve the definition of what background is [14]. It can then be relevant to merge datasets from varied crops to build universal models with diverse categories. This is one aspect of one of our ongoing studies, where a methodology for the identification and counting of plants at emergence is being developed.

As discussed in the last section, bias and errors are introduced by the machine learning models when applied to a dataset from new domains. A strategy can be devised to solve this domain shift problem and to improve the performances of deep learning models when applied in the wild. Numerous studies have already tackled this problem, as in the case of data distillation [15] and domain adaptation [16]. More attention should also be paid to adversarial networks. They can for instance be used to learn the mapping between input and output image. They were recently used to learn to translate and image from a source domain X to a target domain Y [17] in the absence of paired examples.

The use of synthetic dataset may also help the training data constitution. The main issue we are currently facing is the gap between the synthetic and real images as illustrated in the figure below. This problem has already been tackled in other domains [18]–[22].

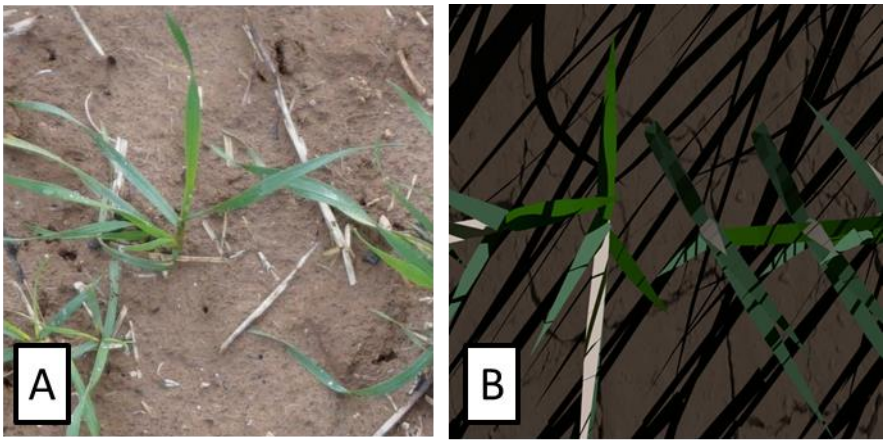


Figure 2 : Illustration of the gap between real (A) and synthetic (B) images

Image segmentation is another example of an activity of interest in phenotyping. The segmentation goal generally refers to a pixel-level classification between vegetation and background. For instance, Leaf area, fraction of vegetation cover and stay-green can be derived through this process. CNNs have the ability to provide additional value because of their abilities to learn about the texture of the image [23]. This is a change of paradigm when compared to the classic approach where colors/spectral indices thresholding are performed at the pixel level. For this purpose we use deconvolutional layers, the model architecture is an encoder – decoder [24]. This is another example of the activities conducted within the CAPTE unit.

In plant phenotyping we cannot rely on public dataset. Despite the increasing efforts to digitize agriculture, the increasing public/private investments, development of image acquisitions platforms and related publications; there is still a lack of annotated public datasets. Since annotating datasets is a time-consuming task and often requires contribution from experts, there is a need to optimize this effort by selecting the best images to annotate. Such an effort would particularly facilitate to build datasets with high diversity towards models that can generalize better without further fine-tuning.

The phenotyping community is also facing the reproducibility crisis [25]. The non-interpretability and black-box nature of neural networks is not expected to help tackling this issue. Open science contributions thus have to be encouraged and promoted.

5.3.2 Use of the methodology for other crops

The developed methodologies in this thesis can be easily transferred to others crops. Wheat was indeed one of the most challenging crop that we were working on. The high density of ears, small diameter of the stem, irregular spacing among the plants and high overlap between them make wheat a challenging crop to phenotype.

5.3.3 Interest of the dynamics

In crop phenotyping, the dynamics of a trait is interesting for breeders. The phenology, i.e. the date of appearance of ears, the flowering status or other important growth stages, are key informations. This is particularly related to the abiotic stress tolerance which is of interest with respect to climate change studies context. In this thesis, we have proposed a methodology to derive the flowering stage with the plant height information. This was roughly estimated when

the plant growth slows down and the height starts to reach a plateau. The heritability of this variable was found to be high. This was a reasonable proxy to the flowering date (RMSE = 76 GDD), considering that the scoring was performed visually with a sampling time interval of 2-3 days. It is also interesting to note that the RMSE were different between the cultivars. Indeed, some cultivars appear to have specific growing patterns which can be interesting for breeders. The visual identification of ears in the RGB images presented in the second study would be the first step before identification or anthers used to get a more precise flowering status. We also demonstrated that precise plant height monitoring reveals when the stress starts to impact the plant growth when it occurs before flowering, with a quantification of its magnitude at least within each genotype.

5.3.4 Robustness of allometric relationships for biomass estimation

The total above ground biomass at harvest is a very important trait for breeders, since it first provides a way to evaluate the photosynthetic efficiency of the crop as well as the harvest index if the yield is independently measured. However, the biovolume used as a proxy of the biomass does not change from the flowering stage to harvest, although the total biomass is expected to increase significantly. This should be further investigated as well as the robustness and accuracy of such allometric relationship. The stem diameter, ear and stem density were demonstrated to be derivable with high-throughput. Combine machines should therefore be equipped with height devices and cameras to systematically derive such information.

References

[1] R. T. Furbank and M. Tester, "Phenomics – technologies to relieve the phenotyping bottleneck," *Trends in Plant Science*, vol. 16, no. 12, pp. 635–644, Dec. 2011.

[2] "(PDF) Accuracy assessment of plant height using an unmanned aerial vehicle for quantitative genomic analysis in bread wheat," *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/332414305_Accuracy_assessment_of_plant_height_using_an_unmanned_aerial_vehicle_for_quantitative_genomic_analysis_in_bread_wheat. [Accessed: 03-May-2019].

[3] X. Han *et al.*, "Measurement and calibration of plant-height from fixed-wing UAV images," *Sensors*, vol. 18, no. 12, p. 4092, 2018.

[4] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[5] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, 2016, pp. 4349–4357.

[6] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[7] S. Madec, *Dataset from the Ear density estimation from high resolution RGB imagery using deep learning technique paper: simonMadec/Wheat-Ears-Detection-Dataset*. 2019.

[8] P. Hu *et al.*, "Estimation of plant height using a high throughput phenotyping platform based on unmanned aerial vehicle and self-calibration: Example for sorghum breeding," *European Journal of Agronomy*, vol. 95, pp. 24–32, Apr. 2018.

[9] F. Baret *et al.*, "Leaf-rolling in maize crops: from leaf scoring to canopy-level measurements for phenotyping," *J Exp Bot*, vol. 69, no. 10, pp. 2705–2716, Apr. 2018.
[10] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: Efficient Multi-Scale Training," *arXiv:1805.09300 [cs]*, May 2018.
[11] J. Dai *et al.*, "Deformable Convolutional Networks," *arXiv:1703.06211 [cs]*, Mar. 2017.
[12] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More Deformable, Better Results," *arXiv:1811.11168 [cs]*, Nov. 2018.
[13] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding Box Regression with Uncertainty for Accurate Object Detection," *arXiv:1809.08545 [cs]*, Sep. 2018.
[14] A. Rame, E. Garreau, H. Ben-Younes, and C. Ollion, "OMNIA Faster R-CNN: Detection in the wild through dataset merging and soft distillation," *arXiv preprint arXiv:1812.02611*, 2018.
[15] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data Distillation: Towards Omni-Supervised Learning," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4119–4128.
[16] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," *arXiv preprint arXiv:1803.03243*, 2018.
[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
[18] J. Tremblay *et al.*, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 969–977.
[19] H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets deep learning for car instance segmentation in urban scenes," in *British Machine Vision Conference*, 2017, vol. 1, p. 2.
[20] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaption for synthetic medical images via adversarial training," *arXiv preprint arXiv:1711.06606*, 2017.
[21] F. C. Nogue, A. Huie, and S. Dasgupta, "Object Detection using Domain Randomization and Generative Adversarial Refinement of Synthetic Images," *arXiv preprint arXiv:1805.11778*, 2018.
[22] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from Simulated and Unsupervised Images through Adversarial Training.," in *CVPR*, 2017, vol. 2, p. 5.
[23] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018.
[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597 [cs]*, May 2015.
[25] M. Baker, "Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the'crisis rocking science and what they think will help," *Nature*, vol. 533, no. 7604, pp. 452–455, 2016.