Chemometrics and Intelligent Laboratory Systems 216 (2021) 104373

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A generic workflow combining deep learning and chemometrics for processing close-range spectral images to detect drought stress in *Arabidopsis thaliana* to support digital phenotyping

Puneet Mishra [a,*], Roy Sadeh [b], Maxime Ryckewaert [c,d], Ehud Bino [b], Gerrit Polder [e], Martin P. Boer [f], Douglas N. Rutledge [g,h], Ittai Herrmann [b]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[b] The Robert H. Smith Institute for Plant Sciences and Genetics in Agriculture, The Hebrew University of Jerusalem, P.O. Box 12, Rehovot, 7610001, Israel
[c] ITAP, INRAE Montpellier Institut Agro, University Montpellier, Montpellier, France
[d] ChemHouse Research Group, Montpellier, France
[e] Greenhouse Horticulture Group, Wageningen University & Research, P.O. Box 644, 6700AP, Wageningen, the Netherlands
[f] Biometris, Wageningen University and Research Centre, Wageningen, the Netherlands
[g] Université Paris-Saclay, INRAe, AgroParisTech, UMR SayFood, 75005, Paris, France
[h] National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, Australia

ARTICLE INFO

Keywords:
Plant breeding
Non-destructive
Illumination effects
Spectroscopy

ABSTRACT

Close-range spectral imaging (SI) of agricultural plants is widely performed for digital plant phenotyping. A key task in digital plant phenotyping is the non-destructive and rapid identification of drought stress in plants so as to allow plant breeders to select potential genotypes for breeding drought-resistant plant varieties. Visible and near-infrared SI is a key sensing technique that allows the capture of physicochemical changes occurring in the plant under drought stress. The main challenges are in processing the massive spectral images to extract information relevant for plant breeders to support genotype selection. Hence, this study presents a generic data processing workflow for analysing SI data generated in real-world digital phenotyping experiments to extract meaningful information for decision making by plant breeders. The workflow is a combination of chemometric approaches and deep learning. The usefulness of the proposed workflow is demonstrated on a real-life experiment related to drought stress detection and quantification in *Arabidopsis thaliana* plants grown in a semi-controlled environment. The results show that the proposed approach is able to detect the presence of drought just 3 days after its induction compared to the well-watered plants. Furthermore, the unsupervised clustering approach provides detailed time-series images where the drought-related changes in plants can be followed visually along the time course. The developed approach facilitates digital phenotyping and can thus accelerate breeding of drought-tolerant plant varieties.

1. Introduction

Agricultural plant phenotyping is the task of monitoring plant traits during their growth in interaction with their surrounding environmental conditions [1,2]. The main aim of plant phenotyping in plant breeding is to select the best performing genotypes by pre-screening them based on their performance in contrasted environmental conditions [2–5]. Such pre-screening of genotypes allows the plant breeders to select the best performing genotypes for breeding purposes so as to enhance the desired trait in a plant variety [1–4]. For example, a key interest of plant breeders nowadays is to pre-screen genotypes that can tolerate drought stress [6–9] due to climate change depleting water resources and increasing unpredictable drought events [7,8]. To do this, the breeder needs to find the best genotypes by exploring several candidate genotypes [1,10]. A typical exploration for drought resistant genotypes includes growing several genotypes under normal watering conditions and then inducing drought stress [6,7]. During the drought stress period, the genotypes are monitored and those that best resist the drought are identified and taken for further analysis [8]. The decision regarding the best genotypes is based on the observation of physical and functional traits [1,2,10]. The

* Corresponding author.
E-mail address: puneet.mishra@wur.nl (P. Mishra).

https://doi.org/10.1016/j.chemolab.2021.104373
Received 14 May 2021; Received in revised form 25 June 2021; Accepted 28 June 2021
Available online 3 July 2021

physical traits are the number of leaves, height, width, 3D (3 Dimensional) structure and other physically observable traits [10]. The functional traits are more related to the photosynthetic activity of plants and ranges from the concentration of plant pigments, macro and micronutrients, and photosynthesis-related parameters such as transpiration rate, stomatal conductance, as well as herbicide resistance [1–3,10,11].

Nowadays, plant phenotyping is widely performed under both controlled conditions, such as greenhouses, and under field conditions [2,3,11,12]. Preliminary knowledge is usually acquired in the controlled conditions of a greenhouse and is then validated under field conditions. However, the manual monitoring of plants traits throughout the growing season is a cumbersome task as the number of genotypes explored in breeding programs is often large (i.e., hundreds and even thousands) and, in many cases, the climatic conditions are harsh for humans to work for long hours monitoring each plant [2,13–16]. Therefore, the concept of digital plant phenotyping has recently emerged [17] to promote the use of novel non-destructive sensing technologies such as imaging, spectroscopy, fluorometers and advanced data analytics to monitor plant traits in a minimally invasive and automatic way [13–15,17]. Another main aim of digital plant phenotyping is to accelerate plant phenotyping and make it a high-throughput task, where several hundreds or even thousands of genotypes can be explored automatically and rapidly to support the fast development of new plant varieties [3,18–20]. Within the framework of high-throughput digital plant phenotyping, several new automated phenotyping facilities have emerged around the world [4,7,17,21].

Visible and near-infrared (400–1000 nm) spectral imaging (SI) is emerging as a key non-destructive sensor for the rapid assessment of plant traits within the framework of digital phenotyping under semi-controlled conditions as well as in the field [17,22,23]. The advantage of SI is that it allows the simultaneous extraction of morpho-physiological traits by capturing spatially resolved spectral properties of plants, unlike classical RGB colour imaging which allows extraction of morphological traits [17,22]. The visible region (400–700 nm) of the spectrum captures information related to key plant pigments such as chlorophyll, anthocyanins, and carotenoids [17,24]. These pigments are related to the photosynthetic activity of plants and are affected by external influences [24]. The near-infrared (NIR; 700–1000 nm) range of the spectrum captures the molecular properties of samples such as the compounds containing OH, CH, and NH bonds, as well as the internal structural properties of leaves such as the cellular structure and leaf thickness [17,24,25]. The NIR part captures the chemical properties through light reflection affected by absorption and transmission and the internal structure properties through light scattering [26]. However, light absorption and scattering effects are mixed in the NIR signal and data modelling techniques are required to extract the information efficiently [27].

Although much innovation has taken place in the development of low cost, high-quality and easy-to-use spectral camera systems, the main challenge is still related to the extraction of key information relevant to plant breeders for decision-making purposes [17]. Commonly, plant breeders are non-experts in computer vision and in the advanced data analysis approaches required for SI data processing. The huge information present in spectral images is of no interest to plant breeders, as their main interest is the quantifiable patterns in the spectral images which can help them in decision making for genotype selection. However, to extract meaningful patterns and to quantify them, the spectral images must go through several steps of data processing [17]. Several data analysis approaches have been tested [17,22], but there is currently no standard workflow to process spectral images generated in digital phenotyping experiments. In fact, the SI data has often been treated simply as RGB images with extra bands, and have been processed with traditional computer vision approaches (working on the spatial context level) [28–30]. Sometimes the SI data has been considered simply as multivariate spectral data (working at the pixel level on unfolded images) and so classical chemometric approaches were then used [6,7,9,31,32].

Table 1
Summary of activities during the monitoring of Arabidopsis with spectral imaging.

Dates (DD-M-YYYY)	Days after sowing	Activities
26/5/2020	0	Seed sowing
06/7/2020	42	SI (1st imaging)
09/7/2020	45	SI
13/7/2020	49	SI
15/7/2020	51	SI
20/7/2020	56	SI
23/7/2020	59	SI + Drought induction
26/7/2020	62	SI
28/7/2020	64	SI
30/7/2020	66	SI
02/8/2020	69	SI
04/8/2020	71	SI (last imaging)

However, SI data is a combination of spatial and spectral information [33,34], and thus requires combinations of computer vision and chemometric approaches at several stages of the processing. For example, the removal of illumination effects from the spectral images of plants requires the use of chemometric normalisation techniques [6,7,35]; the easy handling of huge spectral images requires the use of latent space modelling techniques [34]; the segmentation of plants from the background soil and pots requires advanced computer vision techniques; and finally to quantify the traits into interpretable plots for breeders requires the use of multivariate data processing approaches such as spectral clustering [6,7,35]. To the best of our knowledge, based on recent review articles in the domain [6,17,22,23], the present work is the first of its kind to provide a unified workflow scheme to process SI data generated for the rapid detection of drought stress in Arabidopsis plants which are commonly studied in digital phenotyping experiments.

The data processing workflow presented here is a combination of chemometric and deep learning (DL) approaches. There are four main steps in the workflow: first, the reduction of illumination effects in spectral images using chemometric spectral normalisation techniques [27]; second, reduction of the size of the spectral images by principal components analysis (PCA) [36]; third, separation of plants from background soil and pots using DL-based semantic segmentation; fourth, either unsupervised clustering to capture the changes in the plants in response to the drought onset and progression, or repetitive measure analysis of variance simultaneous component analysis (REP-ASCA) [37] to extract patterns from a designed experiment. The proposed workflow was demonstrated on an experiment related to drought-stress detection and quantification in *Arabidopsis thaliana* plants grown under semi-controlled conditions. *Arabidopsis thaliana* was used as it is a model widely used in plant biology and for screening genes in phenotyping experiments [9,18].

2. Materials and methods

2.1. Plant material

Arabidopsis thaliana (ecotype: Columbia) was grown in the greenhouse facilities of the Hebrew University, Jerusalem, Israel. The soil material was universal potting soil with base fertilizers, the soil was lightweight substrate with porous texture and high water retention. Sowing was performed on a black plastic pot tray with 18–300 cm^3 volume pots with one plant in each pot. 18 plants were well-watered while another 18 plants were subjected to the drought-stress treatment. The semi-controlled conditions were Temperature = 22–24 °C and RH = 40–50%. The plants were sown on 26/May/2020 and were watered regularly until 23/July/2020 at which moment the drought stress was induced in half of the plants (i.e., plants were no longer irrigated). The remaining well-watered class was the control group for comparison with the drought treated plants.

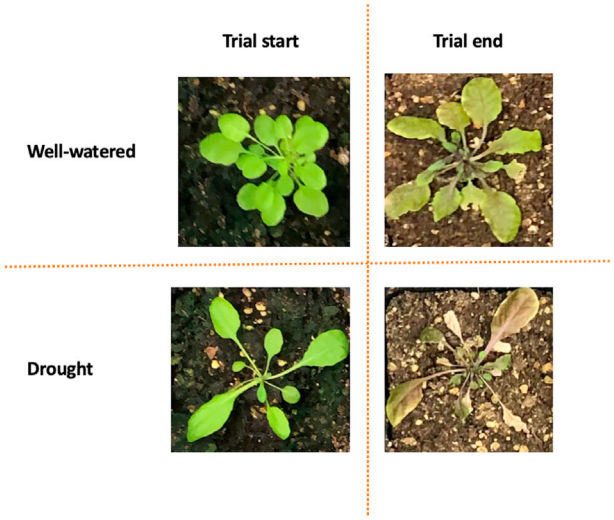

Fig. 1. A typical RGB image of plants before and after the drought stress trial.

2.2. Spectral image acquisition

The plants were monitored by visible and near-infrared spectral imaging at discrete intervals (Table 1) with a HySpex VNIR-1800 (Norsk Elektro Optikk, Oslo, Norway) line scanner that captures a line of 1800 pixels in each shutter opening. The camera was positioned at nadir, 1 m above the top of the canopy. To assemble the whole scene, the pots were placed on a translation stage synchronized with the camera frame rate. The raw data were converted into radiance using a radiometric calibration file specific to the camera and then converted into relative reflectance based on a 50% Zenith Alucore reference panel (Norsk Elektro Optikk, Oslo, Norway). All pre-processing was performed in the HyspexRef software environment (Norsk Elektro Optikk, Oslo, Norway). The spectral range was 407–997 nm with a sampling interval of ~3.26 nm giving a total of 186 bands. The spatial resolution was 6 pixels per mm. During each imaging day, two sets of images were acquired, the first set consisting of images of the 18 well-watered plants in a single scene, the second set being the 18 plants under drought-stress conditions. Hence, two spectral images of size 3958 × 1800 × 186 were generated each day for a total 11 days, giving a total of 22 images. The size of a single image was ~2.5 GB. An example of images from well-watered and drought-stressed plants at start and end of the trial is shown in Fig. 1. It can be noted that on the last day the well-watered plants were greener and healthier than the plants under drought stress.

2.3. Data analysis workflow

A summary of the four-step approach is also illustrated in Fig. 2.

2.3.1. Image preparation and pre-processing

The first step of the workflow is the preparation of the spectral images. This step starts when the reflectance spectral images are ready to be analysed. SI during digital phenotyping experiments involves monitoring several plants over many days during which some images may be badly acquired or human errors may occur. Such spectral images can be considered as outliers and should be identified and removed from the data set. As is often the case in such experiments, only a small number of plants in each group survived for the whole period of the experiment. At first, due to a low gemination rate of the seeds, several of the pots were eliminated, and later, as many pots partially germinated or fell out prior to the drought treatment, other pots were eliminated. In the end this led to a choice of only three plants each for well-watered and drought class. In the following parts of the manuscript, the 3 well-watered plants will be referred to as W1, W2 and W3, and the 3 drought-stressed plants will be referred to as D1, D2 and D3. The pre-processing of spectral reflectance images involves several steps. The first step is to check the spectral signal noise. To improve the smoothness of the spectra for further analysis, a smoothing operation using a 2nd order polynomial was performed using the Savitzky-Golay (SAVGOL) [38] approach with a window size of 13. The next step of spectral image pre-processing is spectral normalisation [27] to eliminate the illumination effects caused by the interaction of the light with the complex plant geometry [17,22]. Various effects such as shadowing, multiple reflections, scattering and a combination of several effects are present in the spectra of plants [17]. These effects are like the additive and multiplicative effects commonly encountered in the chemometrics domain [9,17,31] and the use of spectral normalisation techniques was recently proposed to remove them [6,7,17]. Spectral normalisations are fast and easy to implement and do not require any additional measurements [17]. Of the several spectral normalisation techniques available, two have become popular i.e., standard normal variate (SNV) [39] and variable sorting for normalisation (VSN) [40]. VSN has been shown to outperform SNV [17,35], and for this reason was used in this study to correct the illumination effects in the spectra of the plants [40]. A key point to note is that both the spectral smoothing and normalisation were performed on the unfolded spectral data cubes. After

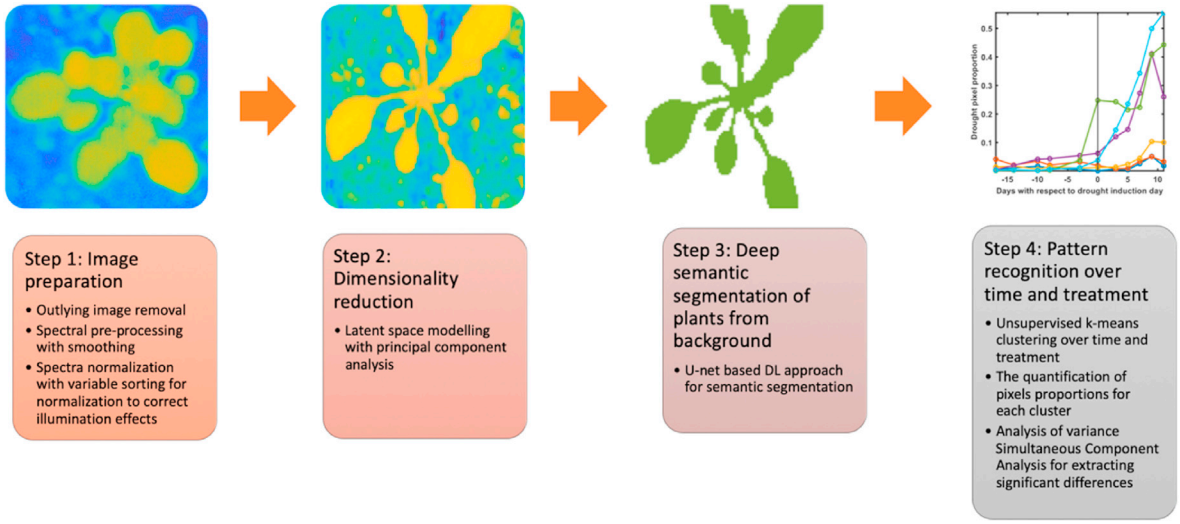

Fig. 2. A summary of the different steps involved in the proposed generic workflow.

the spectral normalisation, the matrices were folded back into 3D (3 Dimensional) cubes.

2.3.2. Dimensionality reduction

Several approaches are available in the machine learning and chemometric domain for dimensionality reduction, but methods based on the use of latent variable are the most practical. The main reason for this is that the latent variable modelling methods not only perform the dimensionality reduction by concentrating all major variations into a small number of latent variables, but also provide loadings and weights vectors which can be used to understand the physicochemical background of these new variables [36]. In this study, principal components analysis (PCA) based latent space modelling was performed on the unfolded data cubes [36], transforming the highly correlated Vis-NIR data into a set of decorrelated variables carrying the important variance in the data [36]. After the PCA analysis, the data matrix was refolded into cubes. The optimal number of PCs was chosen based on the shoulder point of the explained variance plot. Following the PCA, the spectral dimension of the images was reduced from 186 to only 4.

2.3.3. Segmentation of plants from background soil and pots

Once the spectral images pre-processed and their dimensionality reduced, the next step was the segmentation of the plant pixels from the non-plant background which is necessary since the data analysis must be based solely on the pixels of the plants, and the presence of background material could influence the analysis. The segmentation can be performed in several ways. The most common approach is the use of a threshold in the normalized difference vegetation index (NDVI) images [6,7,22]. The values of NDVI ranges from −1 to 1, where a healthy green plant has values near 1, while, the non-plant background material has values close to 0. A simple threshold can be set and a binary segmentation of plant and non-plant background can be attained. However, the main challenge with the NDVI approach is that in many cases the background can contain material that can also have high NDVI values, for example, partially germinated plants and mosses, in which case a suitable threshold cannot be found. Another approach is to train a pixel-based classifier and predict the class for each pixel [22], but this approach is not ideal for two reasons; firstly, it does not use the spatial context of the image and secondly, material with similar spectral properties as the leaves will be misclassified. Recently, a DL-based semantic segmentation approach for plant segmentation was proposed [41]. The approach jointly uses the spatial and spectral information to achieve semantic segmentation. The DL model was based on a U-net, which is widely used for the segmentation of multivariate images. A key point to note is that although a U-net based DL approach is used here, other DL networks can be explored and the best one for each particular application selected. In this study, the ground truth labels were generated manually in MATLAB 2018b (The MathWorks, Natick, MA, USA) using the 'roipoly' function. The U-net was implemented using the Python (3.6) language and Keras/TensorFlow (2.1.0). The model weights were initialized with the 'he normal' initializer in Keras, the adaptive moment (ADAM) optimizer was used to minimize the categorical cross-entropy loss function, and a batch size of 32 was used for model training. Automatic learning rate adaptation based on monitoring the intersection over union (IoU) score was implemented with the 'ReduceLROnPlateau' function from Keras. An early stop was implemented using the 'EarlyStopping' function from Keras, where the training process was automatically stopped if no further significant improvement was noted in the validation loss during the training process. Each model was trained for 1000 epochs but due to the 'EarlyStopping' function, model training was always terminated in ~100 epochs. Checkpointer was used to save the best model weights during the training process.

2.3.4. Pattern recognition during plant growth

2.3.4.1. K-means clustering. Once the plants were segmented from the background, the next step was to reveal the pattern from time-series spectral images related to the plant growth. The patterns also need to be quantified into simple plots or graphics for decision making by plant breeders. An easy approach to process this data is to perform clustering over time and treatment [6,7,9]. To perform such a clustering, a random subset of plants was selected from different time points and treatments i.e., well-watered and drought-stressed. Not all the time points and treatments can be processed simultaneously due to memory limitations. In the domain of spectral image processing spatial or spectral binning can be implemented to reduce data size [34]. However, in this study, in order to keep all the spatial and spectral information, no binning or image compression was performed. The k-means clustering can be performed pixels-wise on the random subset of HS images. In this study, the optimal numbers of k-means clusters were estimated using the Calinski-Harabasz criterion [42] and were implemented in MATLAB with functions available in the 'Statistics and machine learning toolbox'. Once the k-means clustering performed, the cluster centroids were saved and later cluster maps for all the images were obtained using the Euclidean distances. Such clustering allows exploring hidden data patterns that can be visualized as time series trajectories of the cluster pixels. The evolution of clusters can further be quantified as pixel proportions for easy interpretation by plant breeders. In this study, the cluster proportion trajectories over time were quantified using the 'histcount' function in MATLAB individually for each image.

All data analyses were performed on a workstation equipped with a NVidia GPU (GeForce RTX 2080 Ti), an Intel® Core™ i7-4770k @3.5 GHz and 64 GB RAM, running Microsoft Windows 10 OS.

2.3.4.2. REP-ASCA. In this study, REP-ASCA [37] was performed to extract patterns and to study effects of the factors day, drought treatment and their interaction. REP-ASCA as an analysis of variance method for spectral data has the advantage of reducing errors related to the lack of measurement repeatability which may be due to factors that are not identified or nested in the experimental design.

The REP-ASCA approach required, on the one hand, a dataset **X** dedicated to the analysis of variance and, on the other hand, another dataset **W** to describe the repeatability error. In our case, **X**was randomly formed from the large number of spectra available from spectral images. The same number of spectra was used for each modality to strictly respect a balanced design. The remaining spectra were used for **W** to describe the repeatability error. Then, components were obtained from **W** to remove them from **X**.

X was decomposed according to the model established as follows:

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{X}_{\mathrm{T}} + \mathbf{X}_{\mathrm{D}} + \mathbf{X}_{\mathrm{TxD}} + \mathbf{R}$$

Where $\boldsymbol{\mu}$is the average of **X**. The terms $\mathbf{X}_{\mathrm{T}}$and $\mathbf{X}_{\mathrm{D}}$ and $\mathbf{X}_{\mathrm{TxD}}$ are matrices corresponding respectively to treatment, date and interaction between treatment and date. Each of these matrices associates each observation with a mean spectrum of a particular modality. For example, $\mathbf{X}_{\mathbf{T}}$ assigns the average spectrum for all observations of the irrigated condition and the average spectrum for those of the non-irrigated condition. **R** matrix represents the residuals.

REP-ASCA uses a permutation test to estimate the level of significance of each factor. The loadings of each significant factor highlight the spectral regions influenced by that specific factor while the scores can be used to classify observations.

3. Results and discussion

3.1. Step 1: Image pre-processing to remove illumination effects

The spectral images were pre-processed with the VSN algorithm to minimize the illumination effects on *Arabidopsis* leaf reflectance (Fig. 3) giving rise to large variations in the signal intensity within and between

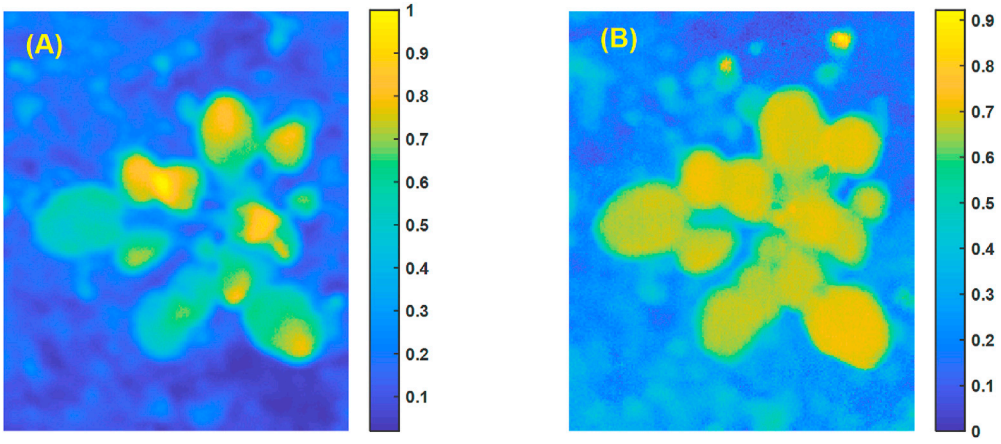

Fig. 3. An example of illumination correction by VSN pre-processing. The image represents the spectral band of 883 nm. (A) Before pre-processing, and (B) after pre-processing with VSN. It can be noted that before pre-processing the image was affected with the illumination effect and the intensity differences between pixels within leaves was high, while after VSN pre-processing the intensity of plant pixels was homogenous.

leaves. In particular, the tip gave higher or lower signal intensities, due to the non-flat geometry of the leaves. Spectral normalisation by VSN reduced the variability in the signal intensities of the leaves (Fig. 3B), in agreement with previous work using VSN in this way [35]. A low variation in the signal intensities of leaves was expected in this case, as that plant was under no stress or external influence which could cause high within-leaf variability in signal intensity.

3.2. Step 2: Dimensionality reduction

After pre-processing, all the spectral images were transformed into principal component (PC) scores images (Fig. 4). The first four PCs (Fig. 4A) were retained, hence the image cubes were transformed from 186 variables to 4. The first PC (Fig. 4B) had near-zero weights in the visible region but increasing weight in the NIR region, which is not related to the plant as chlorophyll gives a peak in the visible region around 550 nm. In the scores image for the 1st PC (Fig. 4C), it can be noted that the plant had scores close to 0, while the other background soil material had higher scores. This confirms that PC1 is not related to the plant in the image. The loadings for PC2 (Fig. 4B) and the corresponding scores image (Fig. 4D) suggest that it is also not related to the plant but to some background material such as green moss present in the soil. PC3 was related to the Arabidopsis as the loading weight vector (Fig. 4B) is a typical spectrum of green vegetation with peaks for photosynthetic pigments, the red-edge region and high signal intensity in the NIR region [43]. The scores image for PC3 (Fig. 4E) also shows that the plant pixels had higher scores than the background. The loading weights for the PC4 (Fig. 4B) were close to zero in the NIR region, but have several peaks in the colour region as well as non-zero signal in the red-edge region. The scores image suggests that PC4 (Fig. 4F) was not related to green vegetation in the scene as the Arabidopsis pixels had scores close to zero.

3.3. Step 3: Deep semantic segmentation of plants

After the PCA transform, the next task of semantic segmentation was performed to distinguish plant pixels from the background. For semantic

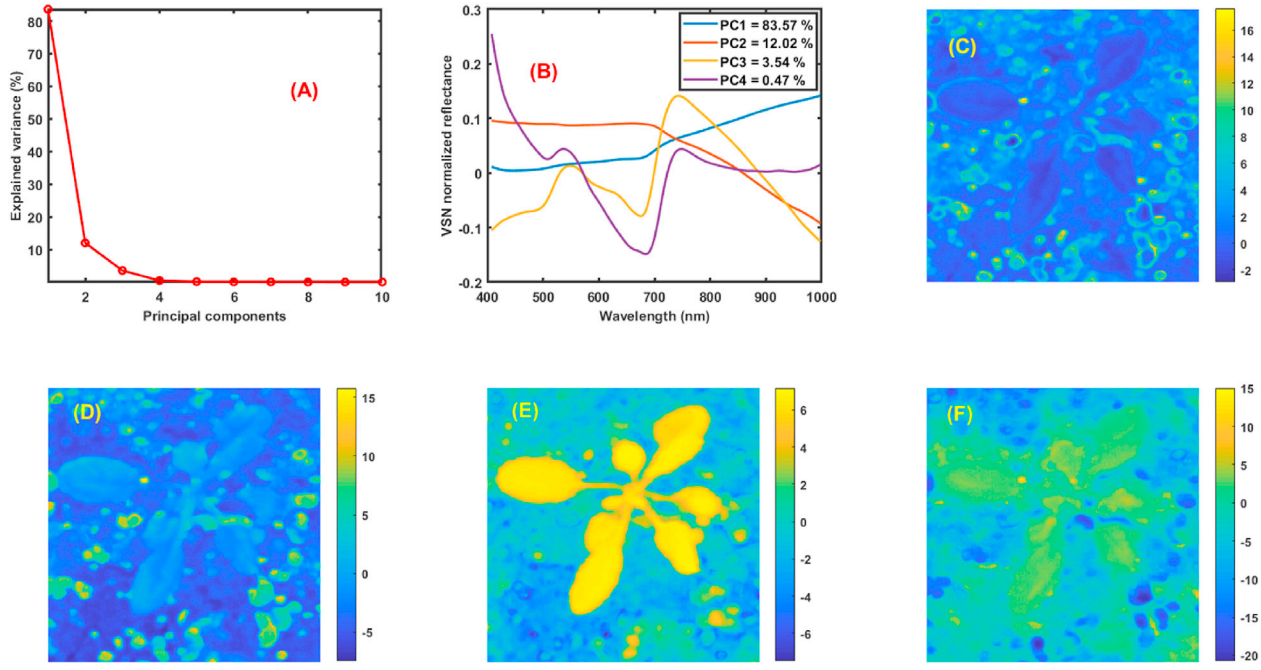

Fig. 4. A summary of the results of the PCA performed on the spectral images. (A) explained variance plot - 4 PCs were retained, (B) loadings weights for the 4 retained PCs, (C) PC1 scores image, (D) PC2 scores image, (E) PC3 scores image, and (D) PC4 scores image.

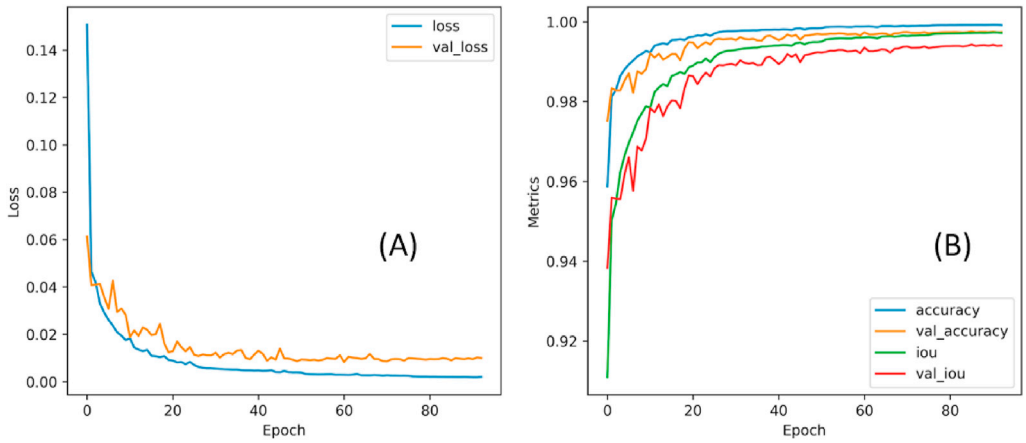

Fig. 5. A summary of the U-net model training and validation performance as a function of the number of epochs. (A) Loss vs. epoch, and (B) Validation accuracy, Intersection over Union score vs. epoch.

segmentation, a U-net was trained. The training and validation performances of the U-net are shown in Fig. 5. Less than 100 epochs were required for the training. The validation and IoU scores were higher than 0.98 (Fig. 5B). The segmentation model managed to cover the entire time series (Fig. 6A). As can be seen in Fig. 6B, in the case of one of the drought-stressed plants, the segmentation model supplied a clear binary segmentation of the plant pixels from the background (Fig. 6A). The segmentation model was trained on the data from the complete time series, covering all the variability in the plant pixels. Thus, the model performed as expected, not only segmenting the green plants from the background, but also segmenting senescing as well as drought-stressed plant parts (Fig. 6A and B).

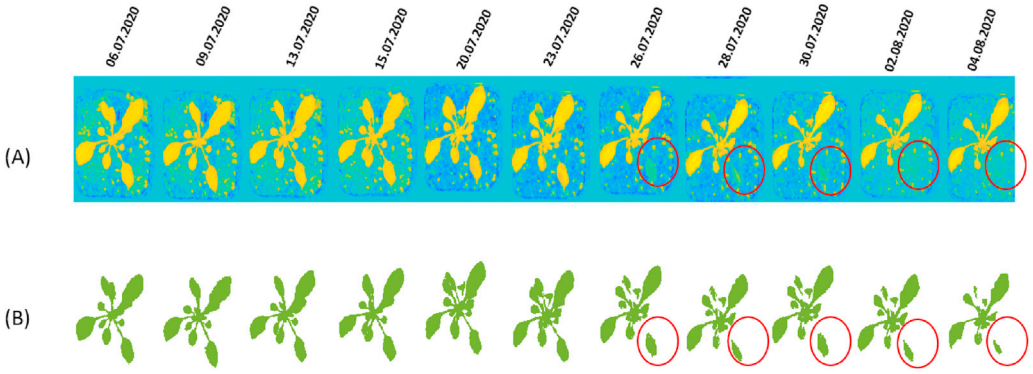

Fig. 6. An example of semantic segmentation performed by the U-net based DL models for the complete time series monitoring of a drought stressed plant. (A) 1st PC image, and (B) segmentation mask. Red circles mark drought-stressed Arabidopsis parts. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

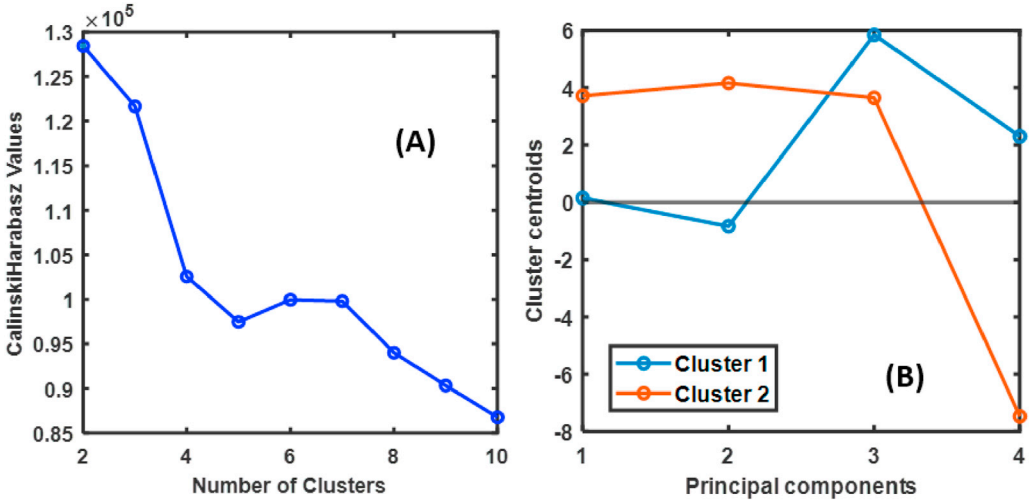

Fig. 7. A summary of the k-means clustering results performed combining time and treatment data. (A) Plot for the selection of 2 as the optimal number of clusters, and (B) cluster centroids. A key point to note is that the k-means clustering was performed on the PCA transformed data to reduce the computational cost.

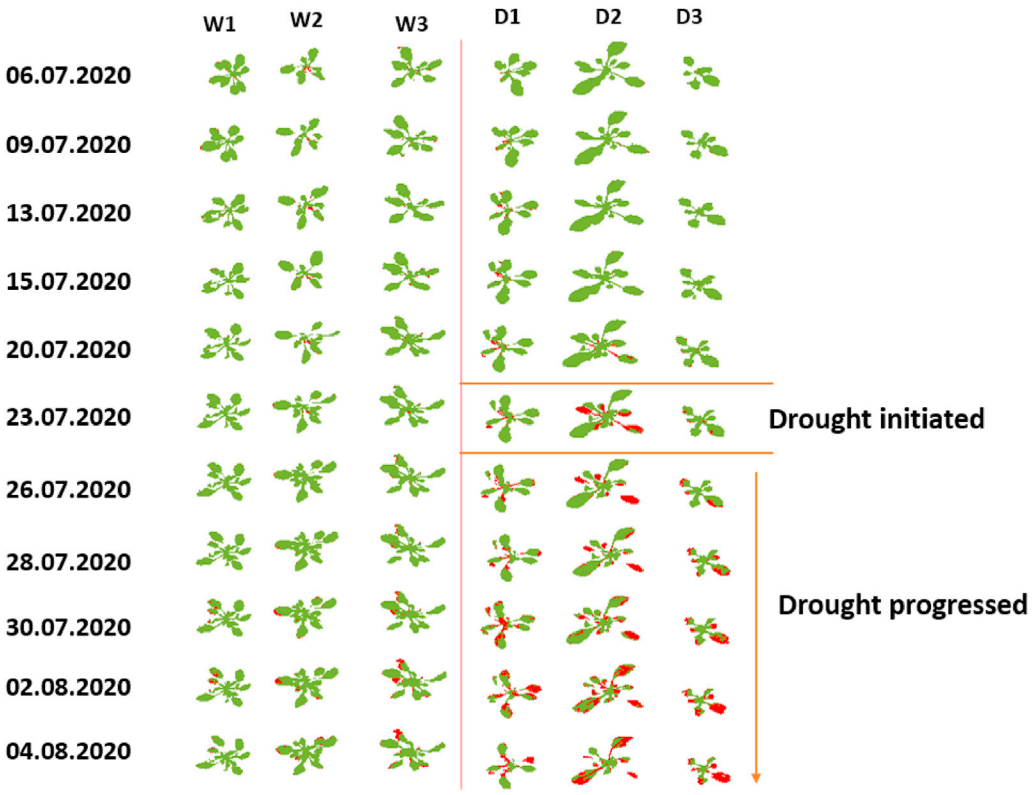

Fig. 8. Cluster maps for the complete time series of 3 well-watered (W1, W2, and W3) and 3 drought-stressed (D1, D2, and D3) plants. The drought stress was initiated at 23/07/2020. There were only two clusters identified in the k-means modelling. The green and red pixels correspond to the two clusters. The red pixels are indicative of the drought effect and can be seen as progressively increasing after the initiation of drought conditions. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.4. Step 4: Pattern recognition over time

3.4.1. k-mean clustering along time

Once the segmentation of plants was done, the pattern explaining differences in time and water status (i.e., well-watered and drought-stressed) was extracted from the data. In this study, k-means clustering of the PCA transformed data was used for unsupervised pattern extraction. The Calinski-Harabasz criterion indicated that just 2 clusters were sufficient to explain the pattern in the data (Fig. 7A). The first cluster has zero scores for PC1 and PC2, and high scores for PC3 (Fig. 7B). Since PC3 is related to the green healthy plant pixels, a higher weight of PC3 for cluster 1 suggest that cluster 1 was related to green healthy plant pixels. On the other hand, cluster 2 had positive weights for PC1 and PC2, and high negative weights for PC4, all were related to unhealthy plant pixels.

The differences between well-watered and drought-stressed Arabidopsis plants are presented in a cluster map for the complete time series (Fig. 8). For the well-watered plants, cluster 1 dominated over the complete time series, indicating that the well-watered plants stayed healthy. For the plants under drought stress, a progressive increase in cluster 2 was noted after the drought induction. Hence, the cluster maps provided a clear understanding of the effect of the drought-stress compared to the well-watered plants. A point to note is that the well-watered plants in the late time points also had a few pixels corresponding to cluster 2 (red), but they were not dominant as in the case of the plants under drought stress. The small number of cluster 2 (red) pixels in the late time point can be related to the natural senescence of the plants.

The cluster maps gave visual insight into the patterns in the data that need to be quantified so that the plant breeders can use them for easy

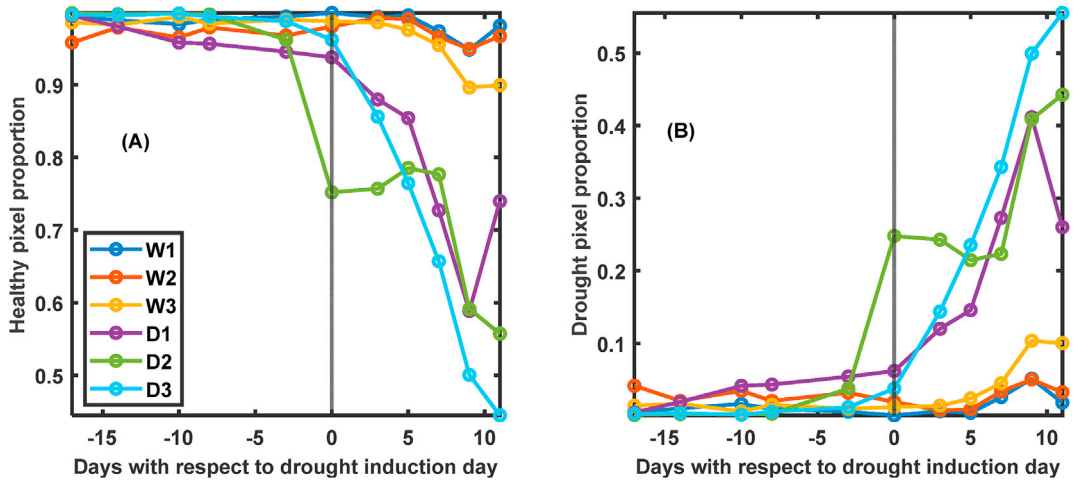

Fig. 9. Plots showing the evolution of cluster proportions. (A) Evolution of healthy plant pixels i.e., cluster 1, and (B) Evolution of drought stress related pixels i.e., cluster 2. Day 0 in the x-axis indicates the induction of drought stress.

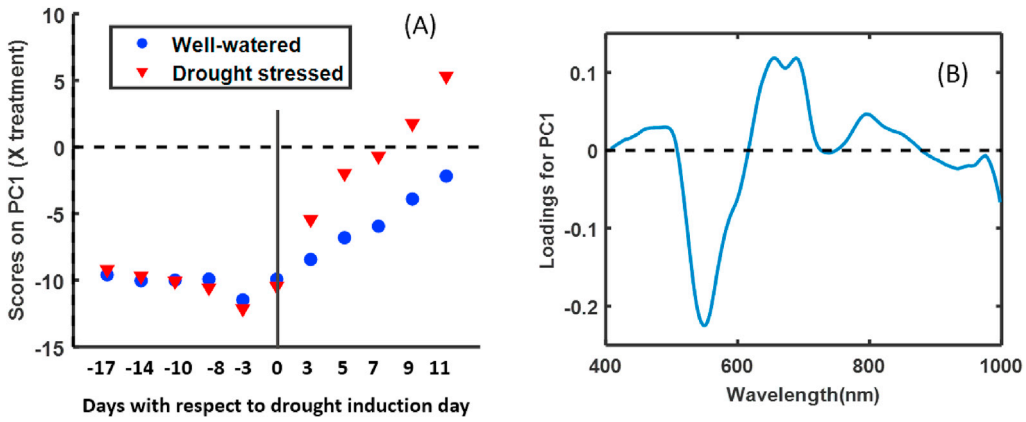

Fig. 10. A summary of the REP-ASCA analysis on the spectral data of plants. (A) Scores corresponding to the factor related to treatment, and (B) loadings vector corresponding to the effect of drought treatment.

comparison of plants under study. To achieve this quantification, it was proposed to use the evolution of cluster pixels over time, where the proportion of pixels for each cluster map is plotted as time series trajectories (Fig. 9). Fig. 9A and B shows the trajectories for cluster 1 and 2, respectively. In Fig. 9A, it can be noted that the cluster proportions for well-watered plants (W1, W2 and W3) remained >0.90 for cluster 1 during the complete time series, while for the plants under drought stress (D1, D2 and D3) the cluster proportion decreased after the day of drought induction i.e., 23/07/2020. Similarly, for cluster 2, the well-watered plants kept cluster proportions <0.1, while for the plants under drought stress the cluster proportions drastically increased after the day of drought induction. Either Fig. 9A or B can be used for decision making purposes as they both provide the same trend: Fig. 9A showing how the healthy pixels decreased; Fig. 9B showing how the pixels related to drought increased along the time of drought progression. The well-watered and drought-stressed plants can be separated from day 3 onwards. A key point to note in Fig. 9 is that, for plant D1, it seemed that on the last day the total number of healthy pixels increased. However, this was not the case. A reason we can give for an apparent slight increase in healthy pixels in drought stress suffering plants on the last day of the experiment could be related to the shrinking of dehydrated leaves. Such a shrinkage can reduce the proportion of plant pixels suffering from drought-stress with respect to the total number of plant pixels.

3.4.2. REP-ASCA

In addition to the previous analysis, the analysis of variance REP-ASCA method studies the effect of factors on the variance of the observations. On this dataset all factors, i.e., day, drought treatment and their interaction, are significant (p-value<0.05). In order to target drought detection, scores and loadings for the treatment factor are shown in Fig. 10A and Fig. 10B, respectively.

Before the induction of drought stress, the scores obtained for the well-watered condition are negative and close to those obtained for drought stressed condition. In contrast, after induction, the separation between the two conditions gradually increases and scores of the drought-condition become positive while those for well-watered condition remain negative. This temporal evolution shows that the installation of stress is progressive. This dissociation between the two conditions appears as early as the third day after induction.

The loadings vector (Fig. 10B) related to the treatment factor highlights the spectral regions involved in this dissociation. On this component, a peak is visible at 550 nm corresponding to plant pigments [17] and more particularly to anthocyanins. Other information is carried by this component such as the red edge regions 670–720 nm [17] and the 3rd overtones of OH bonds [44] related to moisture differences in well-watered and drought-stressed plants. All these key spectral bands are likely to be indicators for detecting drought stress (Fig. 10B).

4. Summary and conclusions

This study proposed a generic four-step scheme to process SI data generated during digital plant phenotyping experiments: pre-processing, dimensionality reduction, segmentation, and pattern recognition. A demonstration case was presented related to drought stress detection in *Arabidopsis thaliana* grown under semi-controlled conditions and monitored over a period of ~2 months. The results showed that the generic scheme allowed an early detection of drought stress. Furthermore, the drought symptoms were quantified as the proportion of healthy and unhealthy pixels to non-destructively identify the effect of drought 3 days after induction. REP-ASCA was used to confirm the detection of drought from 3 days onwards. Although the demonstration case was related to drought stress detection, the generic scheme can be applied to any of the other types of stress or traits commonly explored in digital phenotyping experiments. Furthermore, the generic scheme proposed here being flexible, users can easily replace various steps within it. If, for example, new chemometric pre-processing approaches are developed, or if new DL segmentation approaches appear then the user can introduce them. A key limitation of the study is that the demonstration was performed on a single genotype of *Arabidopsis thaliana*, hence, no genotype selection was performed. Future work will involve testing this generic scheme for new experiments involving multiple genotypes. Although this data processing pipeline covers several steps required for modelling spectral images related to monitoring plants under drought stress, most plant breeders are non-experts in data processing and usually require easy to use tools to gain abstract understanding of their breeding trials. In that case, the most informative and easiest to understand part of the analysis was REP-ASCA which provides an objective way to compare the effects of treatments on the plants. Another key point to note is that, although it may seem complicated to do the DL modelling in Python and the chemometric analysis processing in MATLAB, this was done because DL modelling was easier in Python than in MATLAB. However, MATLAB has also recently provided options to use DL models developed in Python, hence, it can be expected that in future implementations all the modelling presented in this work will be performed within a single programming environment.

Author statement

Puneet Mishra: Conceptualization; Methodology; Software; Writing - Original Draft; Data Curation.

Roy Sadeh: Conceptualization; Methodology; Software; Writing - Original Draft; Data Curation.

Maxime Ryckewaert: Methodology; Software; Writing - Original Draft.

Ehud Bino: Conceptualization; Methodology; Software; Writing - Original Draft; Data Curation.

Gerrit Polder: Conceptualization; Methodology; Writing - Review & Editing.

Martin P. Boer: Conceptualization; Methodology; Writing - Review & Editing.

Douglas N. Rutledge: Conceptualization; Methodology; Writing - Review & Editing.

Ittai Herrmann: Conceptualization; Methodology; Writing - Review & Editing.

Declaration of competing interest

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

Acknowledgements

This research was supported by the Wageningen University & Research knowledge base program 'Data Driven & High Tech.'

Dr. Idan Efroni, *The Hebrew University of Jerusalem* – supplying the plant material.

References

[1] S. Dhondt, N. Wuyts, D. Inzé, Cell to whole-plant phenotyping: the best is yet to come, Trends Plant Sci. 18 (2013) 428–439.
[2] R. Pieruschka, U. Schurr, Plant phenotyping: past, present, and future, Plant Phenomics (2019) 6, 2019.
[3] C. Zhao, Y. Zhang, J. Du, X. Guo, W. Wen, S. Gu, J. Wang, J. Fan, Crop phenomics: current status and perspectives, Front. Plant Sci. 10 (2019) 714.
[4] W. Yang, H. Feng, X. Zhang, J. Zhang, J.H. Doonan, W.D. Batchelor, L. Xiong, J. Yan, Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives, Mol. Plant 13 (2020) 187–214.
[5] F.A. van Eeuwijk, D. Bustos-Korts, E.J. Millet, M.P. Boer, W. Kruijer, A. Thompson, M. Malosetti, H. Iwata, R. Quiroz, C. Kuppe, O. Muller, K.N. Blazakis, K. Yu, F. Tardieu, S.C. Chapman, Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding, Plant Sci. 282 (2019) 23–39.
[6] M.S.M. Asaari, P. Mishra, S. Mertens, S. Dhondt, D. Inze, N. Wuyts, P. Scheunders, Close-range hyperspectral image analysis for the early detection of stress responses in individual plants in a high-throughput phenotyping platform, ISPRS J. Photogrammetry Remote Sens. 138 (2018) 121–138.
[7] M.S.M. Asaari, S. Mertens, S. Dhondt, D. Inze, N. Wuyts, P. Scheunders, Analysis of hyperspectral images for detection of drought stress and recovery in maize plants in a high-throughput phenotyping platform, Comput. Electron. Agric. 162 (2019) 749–758.
[8] N. Briglia, G. Montanaro, A. Petrozza, S. Summerer, F. Cellini, V. Nuzzo, Drought phenotyping in Vitis vinifera using RGB and NIR imaging, Sci. Hortic. 256 (2019), 108555.
[9] P. Mishra, T. Feller, M. Schmuck, A. Nicol, A. Nordon, Early Detection of Drought Stress in Arabidopsis Thaliana Utilsing a Portable Hyperspectral Imaging Setup, 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, WHISPERS), 2019, pp. 1–5.
[10] A. Walter, F. Liebisch, A. Hund, Plant phenotyping: from bean weighing to image analysis, Plant Methods 11 (2015) 14.
[11] F. Fiorani, U. Schurr, Future scenarios for plant phenotyping, Annu. Rev. Plant Biol. 64 (2013) 267–291.
[12] I. Herrmann, E. Bdolach, Y. Montekyo, S. Rachmilevitch, P.A. Townsend, A. Karnieli, Assessment of maize yield and phenology by drone-mounted superspectral camera, Precis. Agric. 21 (2020) 51–76.
[13] L. Li, Q. Zhang, D.F. Huang, A review of imaging techniques for plant phenotyping, Sensors 14 (2014) 20078–20111.
[14] N. Fahlgren, M.A. Gehan, I. Baxter, Lights, camera, action: high-throughput plant phenotyping is ready for a close-up, Curr. Opin. Plant Biol. 24 (2015) 93–99.
[15] T. Roitsch, L. Cabrera-Bosquet, A. Fournier, K. Ghamkhar, J. Jiménez-Berni, F. Pinto, E.S. Ober, Review: new sensors and data-driven approaches—a path to next generation phenomics, Plant Sci. 282 (2019) 2–10.
[16] G. Perich, A. Hund, J. Anderegg, L. Roth, M.P. Boer, A. Walter, F. Liebisch, H. Aasen, Assessment of multi-image unmanned aerial vehicle based high-throughput field phenotyping of canopy temperature, Front. Plant Sci. 11 (2020) 150.
[17] P. Mishra, S. Lohumi, H. Ahmad Khan, A. Nordon, Close-range hyperspectral imaging of whole plants for digital phenotyping: recent applications and illumination correction approaches, Comput. Electron. Agric. 178 (2020), 105780.
[18] C. Costa, U. Schurr, F. Loreto, P. Menesatti, S. Carpentier, Plant phenotyping research trends, a science mapping approach, Front. Plant Sci. 9 (2019) 1933.
[19] S. Jarolmasjed, S. Sankaran, A. Marzougui, S. Kostick, Y.S. Si, J.J.Q. Vargas, K. Evans, High-throughput phenotyping of fire blight disease symptoms using sensing techniques in apple, Front. Plant Sci. 10 (2019).
[20] A. Mazis, S.D. Choudhury, P.B. Morgan, V. Stoerger, J. Hiller, Y. Ge, T. Awada, Application of high-throughput plant phenotyping for assessing biophysical traits and drought response in two oak species under controlled environment, For. Ecol. Manag. 465 (2020), 118101.
[21] U. Lee, S. Chang, G.A. Putra, H. Kim, D.H. Kim, An automated, high-throughput plant phenotyping system using machine learning-based plant segmentation and image analysis, PloS One 13 (2018) e0196615.
[22] P. Mishra, M.S.M. Asaari, A. Herrero-Langreo, S. Lohumi, B. Diezma, P. Scheunders, Close range hyperspectral imaging of plants: a review, Biosyst. Eng. 164 (2017) 49–67.
[23] P. Mishra, G. Polder, N. Vilfan, Close Range Spectral Imaging for Disease Detection in Plants Using Autonomous Platforms: a Review on Recent Studies, Current Robotics Reports, 2020.
[24] J.-B. Féret, K. Berger, F. de Boissieu, Z. Malenovský, PROSPECT-PRO for estimating content of nitrogen-containing leaf proteins and other carbon-based constituents, Rem. Sens. Environ. 252 (2021), 112173.
[25] S. Jacquemoud, F. Baret, PROSPECT: a model of leaf optical properties spectra, Rem. Sens. Environ. 34 (1990) 75–91.
[26] R.F. Lu, R. Van Beers, W. Saeys, C.Y. Li, H.Y. Cen, Measurement of optical properties of fruits and vegetables: a review, Postharvest Biol. Technol. 159 (2020).
[27] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, Trac. Trends Anal. Chem. 132 (2020), 116045.
[28] J. Behmann, A.K. Mahlein, S. Paulus, J. Dupuis, H. Kuhlmann, E.C. Oerke, L. Plumer, Generation and application of hyperspectral 3D plant models: methods and challenges, Mach. Vis. Appl. 27 (2016) 611–624.
[29] J. Behmann, A.K. Mahlein, S. Paulus, H. Kuhlmann, E.C. Oerke, L. Plumer, Calibration of hyperspectral close-range pushbroom cameras for plant phenotyping, ISPRS J. Photogrammetry Remote Sens. 106 (2015) 172–182.
[30] C. Romer, M. Wahabzada, A. Ballvora, F. Pinto, M. Rossini, C. Panigada, J. Behmann, J. Leon, C. Thurau, C. Bauckhage, K. Kersting, U. Rascher, L. Plumer, Early drought stress detection in cereals: simplex volume maximisation for hyperspectral image analysis, Funct. Plant Biol. 39 (2012) 878–890.
[31] P. Mishra, M. Schmuck, S. Roth, A. Nicol, A. Nordon, Homogenising and Segmenting Hyperspectral Images of Plants and Testing Chemicals in a High-Throughput Plant Phenotyping Setup, 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, WHISPERS), 2019, pp. 1–5.
[32] I. Herrmann, U. Shapira, S. Kinast, A. Karnieli, D.J. Bonfil, Ground-level hyperspectral imagery for detecting weeds in wheat fields, Precis. Agric. 14 (2013) 637–659.
[33] J.M. Amigo, I. Martí, A. Gowen, F. Marini, Chapter 9 - Hyperspectral Imaging and Chemometrics: A Perfect Combination for the Analysis of Food Structure, Composition and Quality, Data Handling in Science and Technology, Elsevier2013, pp. 343-370.
[34] N. Mobaraki, J.M. Amigo, HYPER-Tools. A graphical user-friendly interface for hyperspectral image analysis, Chemometr. Intell. Lab. Syst. 172 (2018) 174–187.
[35] P. Mishra, G. Polder, A. Gowen, D.N. Rutledge, J.-M. Roger, Utilising variable sorting for normalisation to correct illumination effects in close-range spectral images of potato plants, Biosyst. Eng. 197 (2020) 318–323.
[36] R. Bro, A.K. Smilde, Principal component analysis, Analytical Methods 6 (2014) 2812–2831.
[37] M. Ryckewaert, N. Gorretta, F. Henriot, F. Marini, J.-M. Roger, Reduction of repeatability error for analysis of variance-Simultaneous Component Analysis (REP-ASCA): application to NIR spectroscopy on coffee sample, Anal. Chim. Acta 1101 (2020) 23–31.
[38] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[39] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.
[40] G. Rabatel, F. Marini, B. Walczak, J.-M. Roger, VSN: variable sorting for normalization, J. Chemometr. 34 (2020) e3164.
[41] P. Mishra, R. Sadeh, E. Bino, G. Polder, M.P. Boer, D.N. Rutledge, I. Herrmann, Complementary chemometrics and deep learning for semantic segmentation of tall and wide visible and near-infrared spectral images of plants, Comput. Electron. Agric. 186 (2021), 106226.
[42] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. 3 (1974) 1–27.
[43] M. Matzrafi, I. Herrmann, C. Nansen, T. Kliper, Y. Zait, T. Ignat, D. Siso, B. Rubin, A. Karnieli, H. Eizenberg, Hyperspectral technologies for assessing seed germination and trifloxysulfuron-methyl response in Amaranthus palmeri (palmer amaranth), Front. Plant Sci. 8 (2017) 474.
[44] B.G. Osborne, Near-Infrared Spectroscopy in Food Analysis, Encyclopedia of Analytical Chemistry, 2006.