Computers and Electronics in Agriculture 159 (2019) 42–49

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Automated spectral feature extraction from hyperspectral images to differentiate weedy rice and barnyard grass from a rice crop

Yanchao Zhang[a,1], Junfeng Gao[b,1], Haiyan Cen[a], Yongliang Lu[c,*], Xiaoyue Yu[c], Yong He[a,d,*], Jan G. Pieters[b]

[a] College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China
[b] Department of Biosystems Engineering, Ghent University, Coupure Links 653, Ghent 9000, Belgium
[c] State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou 310006, China
[d] Key Laboratory of Spectroscopy Sensing, Ministry of Agriculture, Zhejiang University, Hangzhou 310058, China

ARTICLE INFO

Keywords:
Weed discrimination
Hyperspectral imaging
Support vector machine
Hyperparameter tuning
Feature engineering

ABSTRACT

Barnyard grass (*Echinochloa crusgalli*) and weedy rice (*Oryza sativa f. spontanea*) are two common and troublesome weed species in rice (*Oryza sativa* L.) crop. They cause significant yield loss in rice production while it is difficult to differentiate them for site-specific weed management. In this paper, we aimed to develop a classification model with important spectral features to recognize these two weeds and rice based on hyperspectral imaging techniques. There were 287 plant leaf samples in total which were scanned by the hyperspectral imaging systems within the spectral range from 380 nm to 1080 nm. After obtaining hyperspectral images, we first developed an algorithmic pipeline to automatically extract spectral features from line scan hyperspectral images. Then the raw spectral features were subjected to wavelet transformation for noise reduction. Random forests and support vector machine models were developed with the optimal hyperparameters to compare their performances in the test set. Moreover, feature selection was explored through successive projection algorithm (SPA). It is shown that the weighted support vector machine with 6 spectral features selected by SPA can achieve 100%, 100%, and 92% recognition rates for barnyard grass, weedy rice and rice, respectively. Furthermore, the selected 6 wavelengths (415 nm, 561 nm, 687 nm, 705 nm, 735 nm, 1007 nm) have the potential to design a customized optical sensor for these two weeds and rice discrimination in practice.

1. Introduction

Rice (*Oryza sativa* L.) is one of the main food sources for more than half of the world population, especially in Asia and Latin America (Muthayya et al., 2014). To meet the demand of increasing populations, the global production of rice needs to increase significantly in the next few years with limited area expansion (Chauhan and Johnson, 2011). However, weeds generally infest rice fields and they reduce rice yield and quality by competing for light, space, water and soil nutrients under natural growing conditions. This is especially true for cultivating rice with direct-seeding, which is highly mechanized but more risky of crop yield loss caused by weeds, because of the lack of suppressive effect of flooding on the weeds that emerge either before or along with the rice crop (Rao et al., 2007; Chauhan, 2013) (see Fig. 1).

Barnyard grass (*Echinochloa crusgalli*) and weedy rice (*Oryza sativa f. spontanea*) are two of the most common and troublesome weeds in rice paddy fields (Karim et al., 2004). They both cause severe yield loss of rice (Muthayya et al., 2014). Particularly, weedy rice, also called red rice, is a conspecific weed of cultivated rice (Qiu et al., 2017). The weedy rice plants are very competitive with rice as they are generally taller, have higher growth rate and produce more tillers than cultivated rice (Smith, 1988). The relatively rapid emergence of weedy rice had been observed in several Asian countries and this could cause a severe threat to the rice production (Rao et al., 2007). However, the existing effective weedy rice management options are quite limited given the fact that weedy rice and rice are the same species with genetic and phenotypic similarities. For example, a chemical method with selective herbicide often fails because of the close relationship between weedy rice and rice (Eleftherohorinos and Red, 2002). Mechanical removal of weedy rice is also difficult, as one of the barriers is how to successfully distinguish weedy rice and barnyard grass from rice crop. Given the numerous challenges of weeds pose to rice fields, the exploration of an

* Corresponding authors at: College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China (Y. He).
E-mail addresses: luyongliang@caas.cn (Y. Lu), yhe@zju.edu.cn (Y. He).
[1] These two authors contributed equally to this work and should be considered co-first authors.

https://doi.org/10.1016/j.compag.2019.02.018
Received 6 December 2018; Received in revised form 16 January 2019; Accepted 16 February 2019
Available online 01 March 2019
0168-1699/

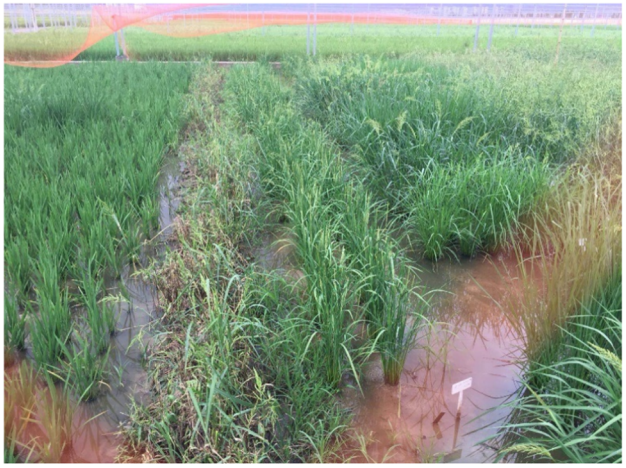

Fig. 1. The experimental field.

effective method to differentiate weeds from rice crop is highly demanded.

Conventional visual cameras are widely used to detect weeds because of their general availability and low cost (Tillett et al., 2008; Gao et al., 2018a). They work well particularly in some circumstances like in the case of relatively large color or shape differences existed between weed and crop. But they provide only limited spectral information as they only record information using three broad bands. Hyperspectral imaging, with hundreds of spectral bands, enables both spectral and spatial information to be captured simultaneously. Every pixel from hyperspectral images has complete spectrum information which has been used for a variety of applications in agriculture (Thenkabail et al., 2013). For example, the study of using hyperspectral imaging for variety discrimination of seeds (Zhang et al., 2012; Gao et al., 2013), and to determine water distribution in meat (Wu et al., 2013). There are also studies about weed and crop recognition using ground-based and drone-based hyperspectral imaging (Wendel and Underwood, 2016; Pantazi et al., 2016; Mirik et al., 2013). However, there is a lack of information and literature to explore the possibility of differentiation of weedy rice and barnyard grass in rice crop.

To the best of our knowledge, this paper is the first to explore the recognition of weedy rice, barnyard grass and rice using machine learning and hyperspectral imaging. The objectives of this study are (1) to develop a pipeline to automatically extract spectral features from line-scanning hyperspectral images, (2) to demonstrate the feasibility of recognizing barnyard grass and weedy rice in rice crop using machine learning algorithms and (3) to determine the most important spectral features for discrimination of barnyard grass, weedy rice and rice.

2. Materials and methods

2.1. Sample preparation

Weedy rice (*Oryza sativa f. spontanea*), barnyard grass (*Echinochloa crusgalli*), and rice (*Oryza sativa*) were planted in the experimental fields of China National Rice Research Institute (CNRRI) Fig. 1. They were planted in respective fields which were close to each other and the infield fertilization and watering were the same. At the growth stage of tiller, all leaf samples were randomly collected from different plants in the experimental fields of CNRRI. Each leaf, near the position of plant canopy, was selected for experimental samples. Immediately after collection, leaf samples were stored inside an icebox at 0 °C in order to slow down their respiration rate and transpiration rate so that the leaves could be entered straight and upright in the hyperspectral imaging cabinet other than curled up due to dehydration. The total numbers of rice, weedy rice, and barnyard grass samples were 100, 81, and 106, respectively.

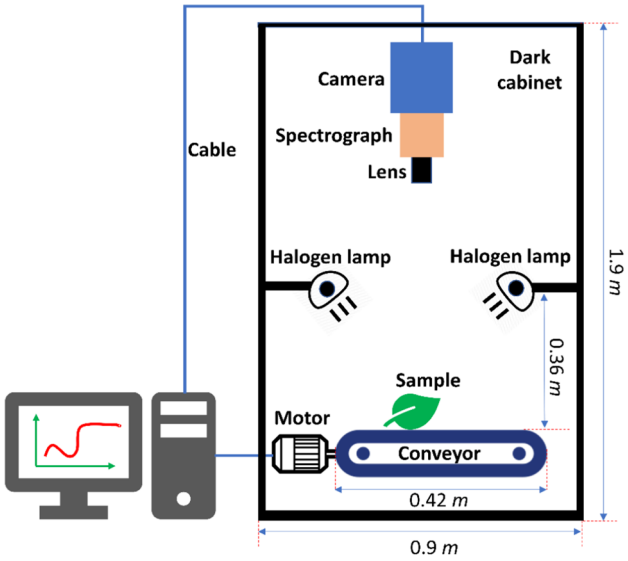

Fig. 2. Scheme diagram of the hyperspectral imaging system.

2.2. Line-scanning hyperspectral imaging system

A line-scanning visual and near-infrared hyperspectral imaging system (Fig. 2), with 512 spectral bands in the range of 380–1024 nm, was used for the acquisition of the hyperspectral image of each sample. The main components of this system include a spectrograph (ImSpector V10E, Specim, Finland), a 672 × 512 (spatial × spectral) charge-coupled device (CCD) camera (C8484–05, Hamamatsu, Japan) with a camera lens (OLE23, Specim, Finland). As the illumination unit, two 150 W halogen lamps (Fiber-Lite DC950 Illuminator, Dolan-Jenner, USA) were deployed symmetrically based on the center of the camera to reduce shadow effect. The leaf samples were placed on the electronically controlled conveyor belt (IRCP0076, Isuzu Optics, China Taiwan) for the image recording. Besides, the system control software (V10E, Isuzu Optics, China Taiwan) was used for setting the optimal parameters (e.g. exposure time, conveyor speed) for imaging. The entire system was installed in a black cabinet (0.9 m × 0.9 m × 1.9 m) for eliminating external light disturbances.

2.3. Image collection and calibration

In our experiment, the conveyor speed was set to be 4.2 mm/*s* in order to synchronize with the scanning of the camera, whose exposure time was adjusted to 0.08 s. The vertical distance between samples and lens was 45 cm. About every 5 leaves were placed side by side with 6 cm central distance on the conveyor of hyperspectral imaging system and each scanning took 4–5 min to finish. The dimension of the obtained hyperspectral images was 672 pixels in the *x*-direction, *n* pixels, depending on how long this image was scanned, in the *y*-direction, and 512 spectral bands in the z-direction, respectively. All the raw images were calibrated using the following equation.

$$I_{calibrated} = \frac{I_{raw} - I_d}{I_w - I_d} \quad (1)$$

where $I_{calibrated}$is the calibrated image, I_{raw} is the raw hyperspectral image, I_d is the dark reference image obtained by covering the lens with a black lens cap, and I_w is the white reference (Teflon white cuboid panel with 99% reflectance, 200 mm × 25 mm × 10 mm) image. Due to the high noise to signal ratio in the spectral range of 380–414 nm and 1009–1080 nm, these bands were removed and the spectral range of 415–1008 nm with a total of 470 bands (features) was considered for

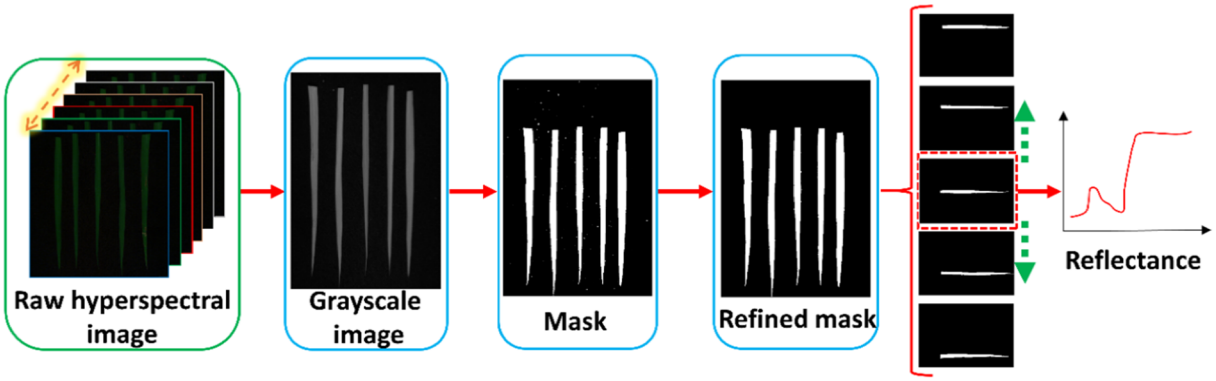

Fig. 3. Automated processing steps of raw hyperspectral imagery.

further analysis.

2.4. Automated feature extraction from raw images

The obtained raw hyperspectral images were first cropped to reduce the redundant surrounding area and the digital values of the pixels were scaled from 0 to 1. Then the single grayscale image from 800 nm wavelength, due to its relatively large digital value difference between green leaves and background in the near-infrared range, was selected to convert into a preliminary mask. The threshold for this procedure was set to be 0.08. After this, small noisy holes less than 100 pixels in the preliminary mask were removed to build a refined mask. As every refined mask contained 4–5 leaves in our experiment, a separation operation was conducted to obtain a sub-mask which only contained one single leaf sample. For each sub-mask, the averaged reflectance from the entire leave was calculated as spectral feature. Fig. 3 shows the main steps for processing the hyperspectral images. We implemented all these procedures with Matlab R2017 software (The Math Works Inc., Natick, MA, USA).

2.5. Wavelet transform for denoising

Hyperspectral data provides both detailed spectral and spatial information of the observed objects, but it generally contains a lot of noise due to the narrow bandwidth of sensors and sampling conditions (Shafri and Yusof, 2009). Wavelet transform is particularly suited to denoise non-linear or non-stationary signals. It applies the transform and zeroing the coefficients below a certain level of threshold. Based on this, noise coefficients would have lower gains than the coefficients corresponding to the studied feature variables. The wavelet transform algorithm to denoise goes as follows (Roy et al., 1999):

(i) differentiate the original signal X(t) to obtain the data x_d (t)

$$x_d(t) = dX(t)/dt \tag{2}$$

(ii) take the discrete wavelet transform of the data x_d (t) and obtain wavelet coefficients $W_{j,k}$ at different dyadic scale j and displacement k with the following equations:

$$W_{j,k} = \int_{-\infty}^{+\infty} x_d(t)\varphi_{j,k}(t)dt \tag{3}$$

$$\varphi_{j,k}(t) = 2^{j/2}\varphi(2^j t - k) \tag{4}$$

where j, k are integers, and where Daubechies's compactly supported orthogonal function (Donoho and Johnstone, 1995) is chosen for the wavelet function $\varphi(t)$.

(iii) Reconstruct the denoised data $\widehat{x_d}(t)$ by taking the inverse transform of the obtained wavelet coefficients $W_{j,k}$.

$$\widehat{x_d}(t) = c_\varphi \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} W_{j,k}\varphi_{j,k}(t) \tag{5}$$

where c_φ is normalization constant given by

$$c_\varphi = 1/\int_{-\infty}^{\infty} \frac{|\hat{\varphi}(\omega)|^2}{\omega} d\omega < \infty \tag{6}$$

with $\hat{\varphi}(\omega)$ as the Fourier transform of the wavelet function $\varphi(t)$.

(iv) obtain denoised signal $\hat{x}(t)$ by Eq. (7).

$$\hat{x}(t) = \int \widehat{x_d}(t)dt \tag{7}$$

We implemented this algorithm with wavelet toolbox in Matlab R2017 software. The wavelet transform for denoising was automatically conducted after obtaining the averaged reflectance from each leaf sample.

2.6. Successive projections algorithm for feature selection

The successive projections algorithm (SPA) is a feature selection technique using minimizing collinearity effects in the calibration data set. Generally, SPA contains three main steps. The first step is to select the variables with minimum collinearity and redundancy as well as maximum projection vector by a simple projection in a vector space. Secondly, the effective variables are determined based on the minimum root mean square error of validation in the validation set of multiple linear regression calibration. Finally, uninformative variables are eliminated without significant loss of prediction ability. The detailed theoretical explanations of SPA are presented in Galvão et al. (2008). In our study, we performed leave one out cross-validation in the training set to run the SPA algorithm.

2.7. Classification models

2.7.1. Random forests

Random Forests (RF), an ensemble learning algorithm, is one of the popular and powerful machine-learning techniques (Breiman, 2001). RF contains a group of classification or regression trees (e.g. 500 decision trees in a single random forest) that are aggregated to compute a classification or regression by means of a majority vote over all trees. Each decision tree is constructed by randomly selecting the subset of features and using a different bootstrap sample from original data, which can reduce the effects of overfitting and improve generalization (Peters et al., 2007). When building a RF, about one-third of the data is left out of the bootstrap samples and not used in the construction of the decision tree. This remaining data, also called out-of-bag samples (OOB), can be used to evaluate the OOB errors as well as to determine the importance of a feature by looking at how much OOB error increase when OOB data for that feature is permuted while all other features are

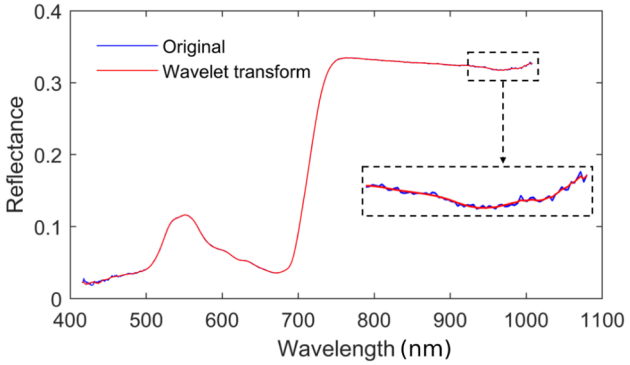

Fig. 4. Noise reduction with wavelet transform.

staying unchanged. The number of trees (m) and the number of randomly selected features (n) to split the tree nodes are two hyperparameters which require being optimized to obtain a minimal RF error.

Algorithm. (*Generic pseudo-code of a random forest*).

1: **Input**: input data set (contain features and class labels) X
2: **Hyperparameter initialize**: number of trees (m), number of features (n)
3: **For** i = **1** to **m do**:
4: randomly choose a bootstrap subset Xi (two-thirds of instances in X)
5: build a decision tree with randomly selected n features to split nodes
6: **end**
7: **Output**: class label is obtained by the majority vote of the ensemble of m trees

2.7.2. Support vector machines

Support vector machines (SVM) is a supervised learning method which has been employed in the tasks of nonlinear classification, regression and outlier detection. It is based on the concept of hyperplanes that defines decision boundaries. The general idea of SVM is to separate training instances which belong to different classes by tracing maximum margin hyperplanes in the space where the instances are mapped. Thus, SVM only demands training instances near the class boundary, and it is capable of handling high dimensional data even if a small number of training instances are provided (Cavallaro et al., 2015). With kernel trick which maps input data into high-dimensional space, SVM is generalized to non-linear classification problems. SVM solves the following constraint optimization problems:

$$min_{w,\varepsilon_i,b} = \{\frac{1}{2}\|w\|^2 + C\sum_i \varepsilon_i\} \tag{8}$$

subject to

$$y_i(\langle\varnothing(x_i), w\rangle + b) \geq 1 - \varepsilon_i \;\forall\; i = 1, \cdots, n \tag{9}$$

$$\varepsilon_i \geq 0 \;\forall\; i = 1, \cdots, n \tag{10}$$

where $y_i$is a label from data instances and ε_i are the positive slack variables allowing to deal with permitted errors. The regularization parameter (C) affects the generalization capability of SVM. The hyperparameters C, $\varepsilon_i$and kernel function need to be tuned before developing an optimal SVM. Further technical details are presented in Cortes and Vapnik (1995).

In our case, we randomly split the dataset into a training set (221) and a testing set (66) based on a stratified manner. The total number of weedy rice, rice and barnyard grass in the training dataset are 63, 75 and 83, respectively. The total number of weedy rice, rice and barnyard grass in the test dataset are 18, 25 and 23, respectively. Based on 5-fold cross-validation in the training set, the grid-search algorithm (Hsu et al., 2008) was employed to find the optimal parameters to develop the classifiers. For RF, the range of the number of decision trees (m) was set between 400 and 600. The range of the number of selected features (n) to build trees was between 10 and 30. In respect of SVM, linear and radial basis function (RBF) were selected as candidates to decide which one was appropriate. For each kernel function, C was set as a list of [1,10,100,1000], and gamma was set as a list of [0.001, 0.0001] only in case of RBF as kernel type. Python was used to implement hyperparameter tuning and to develop the two classifiers (RF, SVM).

2.8. Model evaluation

The confusion matrix gives a full description of errors made by classifiers (Stehman, 1997). In this matrix, the true labels and the predicted labels are displayed. Overall accuracy (OA) is generally used to evaluate the model overall performance which is calculated as the sum of correctly classified samples divided by the total number of samples. Besides, the recognition rate is used to evaluate the prediction capability for each class. It is a measure of the capability of a classifier to select instances of a certain class from a dataset and corresponds to the true positive rate. Eq. (11) gives its calculation.

$$\text{Recognitionrate}_i = \frac{E_{ii}}{\Sigma_j E_{ij}} \tag{11}$$

In confusion matrix, E_{ii} represents diagonal elements of the i-th class, while $\Sigma_j E_{ij}$ represents the total of true values of the i-th class.

3. Results and discussion

3.1. Wavelet denoising and the averaged reflectance of three species

The result of the wavelet transform to denoise is shown in Fig. 4. From the thumbnail, the denoised reflectance was much smoother than the original raw reflectance which had small abrupt changes, especially at the low (415–500 nm) and high (900–1008 nm) edge of the spectrum region. The result is consistent with the findings of Shafri and Yusof (2009) who reported that wavelet works better than other denoise algorithms. It is necessary to perform noise reduction first before derivative analysis due to its highly sensitivity to local sharp changes in reflectance.

The reflectances of all samples from each class were averaged and the corresponding standard deviation (SD) of each class was calculated. Fig. 5 represents the averaged spectral reflectance curves and their SD of three classes (barnyard grass, weedy rice, rice). It shows that the curves from the three kinds of plants are all the same as typical vegetation spectral responses in the range 415–1008 nm (Knipling, 1970). The blue light (near 440 nm) and red light (near 650 nm) were absorbed by chlorophyll for photosynthesis, resulting in two distinctive absorption valleys. The green light (near 550 nm) was partly reflected by

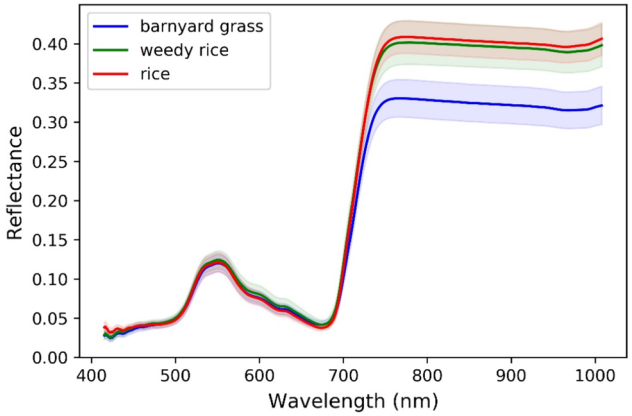

Fig. 5. Averaged reflectance and standard deviations for the three types of plant leaves.

Table 1
Tuned hyperparameters using 5-fold cross-validation.

Classifier	Optimal Hyperparameters	Averaged OA	Standard Deviation
RF	m = 520, n = 16	0.851	0.064
SVM	Kernel = 'linear', C = 1000	0.860	0.076

chlorophylls, leading to a reflection peak. A steep slope from 700 nm to 750 nm, also called red edge, shows a rapid change of reflectance of the plants. In the NIR, all the plants kept a relatively high reflectance. Specifically, the reflectances of weedy rice and rice at the NIR region were much higher than barnyard grass. The averaged reflectance of rice was slightly higher than weedy rice at NIR while they presented almost the same spectral responses in other wavelengths. The very similar reflectance characteristics of weedy rice and rice are largely due to the fact that they have high genetic and phenotypic similarity (Dauer et al., 2018), which results in the difficulty to discriminate them. Based on the reflectance patterns of the three plants, it is clear that barnyard grass is the easiest to discriminate among the three plants. It is also observed that the standard deviation values of spectral features of each plant in the NIR region (750–1008 nm) were larger than those in the visual range (415–670 nm).

3.2. Classification results with full wavelengths

The results of optimal hyperparameters are listed in Table 1. With these optimal hyperparameters, we developed the RF and SVM models with all spectral features. Then we used the models to predict samples in the test set. The result is provided in Table 2. It can be seen that the SVM (0.969) performed better than the RF (0.879) in the test set with the entire features.

3.3. Feature selection

The SPA selected the 6 most important spectral features (705 nm, 1007 nm, 735 nm, 687 nm, 561 nm, 415 nm) shown with red dot in Fig. 6. These selected features covered bands from blue (415 nm), green (561 nm), red edge (687 nm, 705 nm, and 735 nm) and near infrared region (1007 nm), especially in red edge region with half of the selected features indicating that the bands from this region play a significant role in quantifying plant characteristics. This finding is also consistent with the results of Pantazi et al. (2016) and Gao et al. (2018b). RF is also popular for feature ranking. Based on the OOB error, every feature importance score, shown in Fig. 7, was computed by the RF classifier. The most important 6 features were 969 nm, 415 nm, 978 nm, 982 nm, 951 nm, 970 nm, respectively. Compared to the selected features from SPA, these spectral features were mainly from the NIR region, highly correlated with each other.

3.4. Model performance with selected features

The SVM and RF classifiers were built again with the selected features from SPA. The hyperparameters were tuned again and the procedures were the same as when modeling with all features. The result of overall accuracy (OA) in the test set is shown in Table 3. It can be seen that the SVM model performed better than the RF model both for the training and the test set. Further prediction details of every sample are depicted by the confusion matrix (Fig. 8(A)). All samples of rice and barnyard grass were correctly classified by the developed SVM model with the features selected by SPA. However, 5 out of 16 weedy rice samples in the test set were wrongly classified as being rice samples, indicating that these samples may have similar spectral features as rice samples in certain wavelengths. This meets the assumptions and real observations which were discussed in Section 3.1. Numerous studies (Pantone and Baker, 1991; Vaughan et al., 2001; Nadir et al., 2017) have highlighted the difficulties of weedy rice and cultivated rice discrimination as they are congeneric and conspecific species. We have not found literature results to directly compare our results.

Table 2
Overall accuracy for two classifiers.

Classifiers	Training set	Test set
RF	1.0	0.879
SVM	0.923	0.969

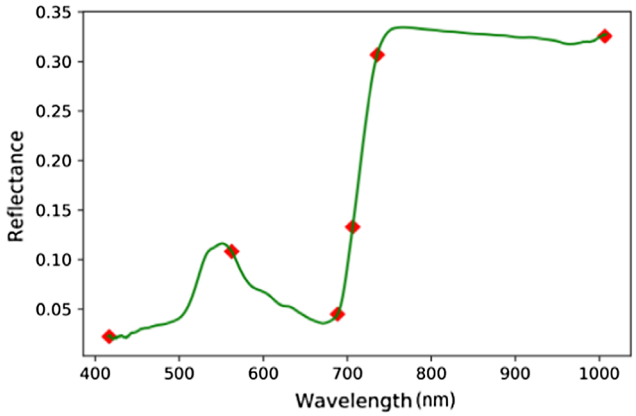

Fig. 6. Selected spectral features shown in red dot in the reflectance curve. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

As the prediction errors in the SVM model come from the weedy rice samples being predicted as rice, we assigned higher sample weights to weed rice samples than barnyard grass and rice samples, which means that the classifier put more emphasis to predict weedy rice samples correctly. The sample weights of barnyard grass and rice were assigned to be 1, while the sample weights of weedy rice were assigned values from 1 to 8 with step of 1. The other parameters and input were kept the same for the previously developed SVM + SPA model. The effects of different sample weights of weedy rice on the prediction accuracy are shown in Fig. 9. It can be seen that both the values of OA in the training set and test set increased first with the increase of sample weights. However, when the sample weights of weedy rice were larger than 4, the values of OA decreased rapidly with the increase of sample weights. The weighted SVM model achieved the highest OA both in the test set (0.970) and in the training set (0.930) when the sample weights of weedy rice were assigned to be 3-fold higher than that of the other two plants. The detailed prediction results in the test set are shown in Fig. 8(B). In the weighted SVM model, all barnyard grass and weedy rice samples were correctly predicted. For rice, 2 out of 25 samples were predicted as being weedy rice. This misclassification error is more acceptable than the error of weeds being predicted as a crop for farmers. The improvement of the weighted SVM model may be attributed to the imbalance of sample weights that rescales the hyperparameter and results in the change of decision boundary of classifiers. Generally, the choice of selecting classifiers strongly depends on specific tasks and objects. Some considered criteria like the number of features and samples, data type, and research purpose can be seen in Behmann et al. (2014).

Feature selection is an important procedure to filter out features that are not significant for modelling. Thus, it is useful to gain a better understanding of the relationships of features and responsive variables. Besides, it can reduce overfitting and improve the generalization of models. Hyperspectral data, with a high number of narrow spectral bands, contains a high redundancy and multicollinearity between bands. Performing spectral feature selection can be extremely useful to data interpretation and sensor design (Stellacci et al., 2016). Although

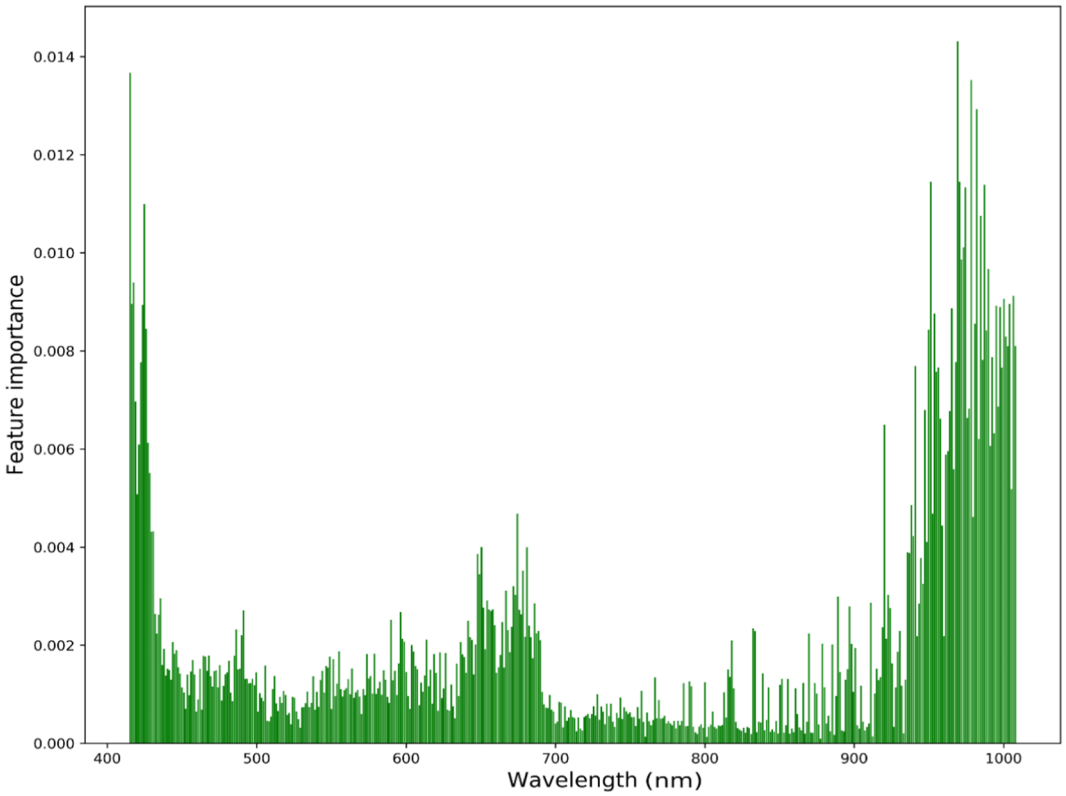

Fig. 7. Feature importance score evaluated by RF.

Table 3
Overall accuracy (OA) in the training and test set with selected features by SPA.

Classifiers	Parameters	Training set	Test set
RF + SPA	m = 100, n = 3	0.814	0.758
SVM + SPA	linear kernel, C = 1000	0.919	0.924
Weighted SVM + SPA	linear kernel, C = 1000	0.930	0.970

commercial multispectral cameras like Micro-MCA 6 produced by Tetracam company, already provide 6 multispectral channels for general applications, these important selected bands can facilitate the design of a customized camera specifically for these two weeds recognition from a rice crop. Comparing the spectral bands of Micro-MCA 6 (490 nm, 550 nm, 680 nm, 720 nm, 800 nm, 900 nm), it can be seen that several bands from the selected wavelengths (415 nm, 561 nm, 687 nm, 705 nm, 735 nm, 1007 nm) are quite close to the generalized bands in Micro-MCA 6. When considering the potentials in real situation, the possible difficulty may be the lighting conditions. In laboratory environment, lighting source is placed at fixed position in a closed cabinet which has been proved effective. However, the light in fields is arbitrary and may come from different directions. In our study, we placed all sample leaves horizontally on the conveyor for imaging. This simplifies the sampling process in the real situation which could lead to inconsistent reflectance due to large variations in field conditions.

The fundamental goal of a machine learning model is to generalize well from training data to any new data from the problem domain. Poor generalization performance of a machine learning model mainly results

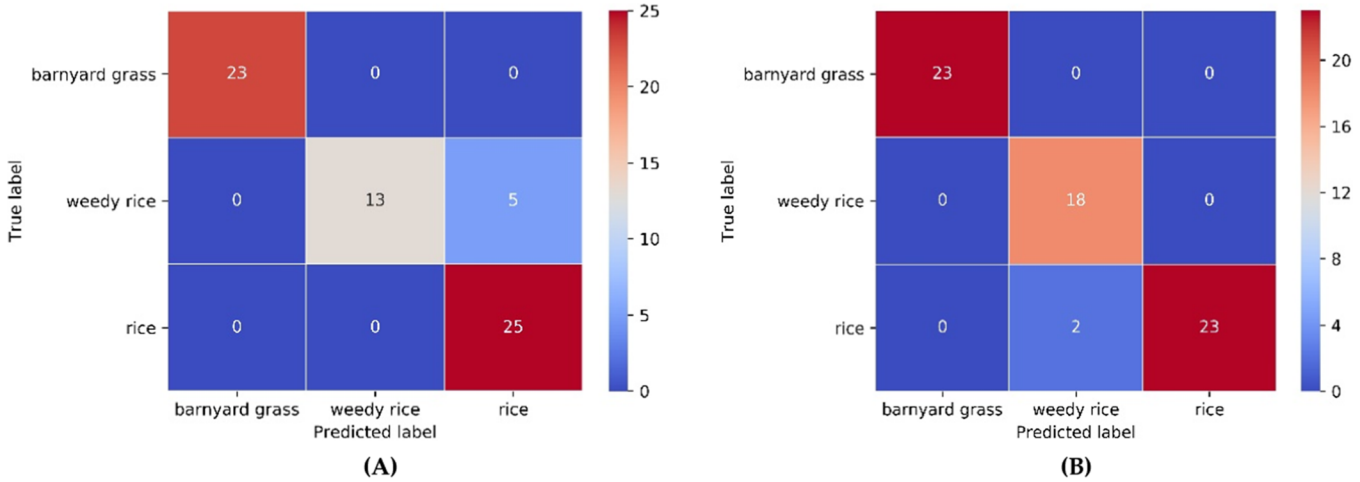

Fig. 8. Confusion matrix of the SVM with selected features in the test dataset, (A) same sample weights; (B) imbalanced sample weights (3-fold sample weights for weedy rice when compared to the other two types of plants).

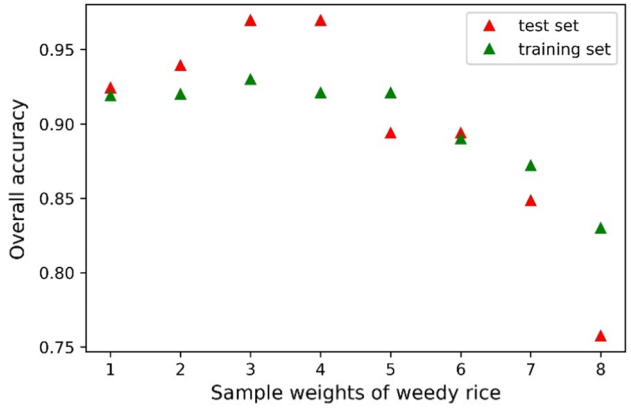

Fig. 9. Overall accuracy as a function of sample weights.

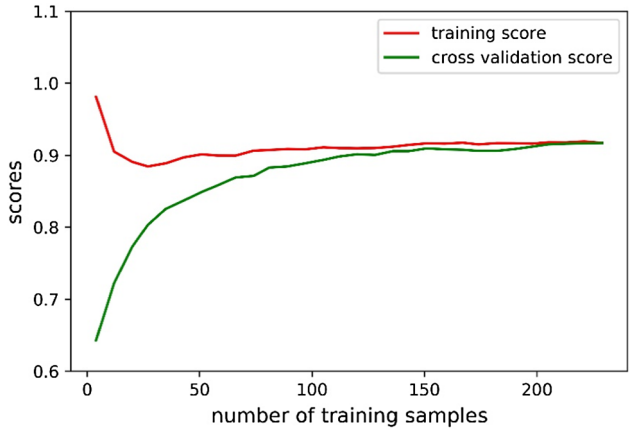

Fig. 10. Learning curves of the weighted SVM with selected features by SPA.

from underfitting or overfitting (Domingos, 2012). Learning curves represent the generalization performance of a machine learning model as a function of the number of training samples. In this study, we used 5-fold cross-validation to plot the learning curves of the weighted SVM model (weighted SVM + SPA). From Fig. 10, the gap between training score and cross-validation score becomes narrower with the increase of training samples. The averaged training and cross-validation accuracy scores both finally plateau around 0.9. Based on the narrow gap and high cross-validation score, it is concluded that the developed model suffers neither underfitting nor overfitting.

Whilst the weighted SVM model provided a good result in recognizing barnyard grass (100%), weedy rice (100%) and rice (92%), it is very important to consider the other aspects to improve model robustness and accuracy. In our study, we only utilized spectral features from plant leaves to build classification models. It is worthy to extract textural and geometry features in spite of their similar shape traits. Kwon et al. (1992) reported that weedy rice tends to have long, hispid, pale and droopy leaves and more culms, forming a more open canopy structure than cultivated rice. It might be useful to explore the combination of spectral and morphological features from plant canopy for weed and crop recognition. Besides, it is meaningful to determine which growth stage is best and reliable for site-specific weed management in practice.

4. Conclusions

In our study, we developed a pipeline to automatically extract spectral features from line scan hyperspectral images for weedy rice, barnyard grass, and rice recognition. Wavelet algorithm was used to denoise the raw spectral features. Subsequently, random forests and support vector machine classifiers were developed with optimized hyperparameters to compare their performances in the test set. Feature selection was explored through successive projection algorithm (SPA) and random forests. The best model in our case was the linear kernel-based support vector machine (SVM) with 6 spectral features (415 nm, 561 nm, 687 nm, 705 nm, 735 nm, 1007 nm) selected by SPA. The imbalance of sample weights, namely 3-fold sample weights of weedy rice with respect to weights for barnyard grass and rice samples, boosted the performance of the SVM model. It was shown that the weighted SVM model achieved 100%, 100%, and 92% recognition rates for barnyard grass, weedy rice and rice, respectively.

Author contributions

Y.Z., J.G., Y.L., H.C., Y.H., and J.P., contributed to the overall study design and supervised all research. Y.Z., J.G. and Y.H finished the experiment of design and data collection. J.G. and Y.Z. contributed to the data analysis and the manuscripts. All authors reviewed and approved the manuscript.

Funding

This research is supported by Science and Technology Development of Zhejiang Province Program (2015C02007), National Key Research and Development Program of China (2016YFD0700304), and the Special Research Fund (BOF) of the Ghent University (No. 01SC3616).

Conflict of interest

The authors declare no conflict of interest.

References

Behmann, J., Mahlein, A.-K., Rumpf, T., Römer, C., Plümer, L., 2014. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. Precis. Agric. 1–22.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J.A., On, Plaza A., 2015. Understanding big data impacts in remotely sensed image classification using support vector machine methods. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 4634–4646.
Chauhan, B.S., 2013. Strategies to manage weedy rice in Asia. Crop Prot. 48, 51–56.
Chauhan, B.S., Johnson, D.E., 2011. Row spacing and weed control timing affect yield of aerobic rice. F. Crop. Res. 121, 226–231.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.
Dauer, J., Hulting, A., Carlson, D., Mankin, L., Harden, J., Mallory-Smith, C., 2018. Gene flow from single and stacked herbicide-resistant rice (*Oryza sativa*): modeling occurrence of multiple herbicide-resistant weedy rice. Pest Manag. Sci. 74, 348–355.
Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55, 78–87.
Donoho, D.L., Johnstone, I.M., 1995. Adapting to unknown smoothness via wavelet shrinkage. J. Am. Stat. Assoc. 90, 1200–1224.
Eleftherohorinos, I.G., Red, Dhima K. V, 2002. Rice (*Oryza sativa*) control in Rice (*O. sativa*) with preemergence and postemergence herbicides. Weed Technol. 16, 537–540.
Galvão, R.K.H., Araújo, M.C.U., Fragoso, W.D., Silva, E.C., José, G.E., Soares, S.F.C., Paiva, H.M., 2008. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. Chemom. Intell. Lab. Syst. 92, 83–91.
Gao, J., Li, X., Zhu, F., He, Y., 2013. Application of hyperspectral imaging technology to discriminate different geographical origins of *Jatropha curcas* L. seeds. Comput. Electron. Agric. 99, 186–193.
Gao, J., Liao, W., Nuyttens, D., Lootens, P., Vangeyte, J., Pižurica, A., He, Y., Pieters, J.G., 2018a. Fusion of pixel and object-based features for weed mapping using unmanned aerial vehicle imagery. Int. J. Appl. Earth Obs. Geoinf. 67, 43–53.
Gao, J., Nuyttens, D., Lootens, P., He, Y., Pieters, J.G., 2018b. Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery. Biosyst. Eng. 170, 39–50.
Hsu, C., Chang, C., Lin, C., 2008. A practical guide to support vector classification. BJU Int. 101, 1396–1400.
Karim, R.S.M., Man, A.B., Sahid, I.B., 2004. Review Paper: Weed problems and their management in rice fields of Malaysia: an overview. Weed Biol. Manag. 4, 177–186.
Knipling, E.B., 1970. Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation. Remote Sens. Environ. 1, 155–159.
Kwon, S.L., Roy, J., Smith, J., Talbert, R.E., 1992. Comparative growth and development

of red rice (*Oryza sativa*) and Rice (*O. sativa*). Weed Sci. 40, 57–62.
Mirik, M., Ansley, R.J., Steddom, K., Jones, D.C., Rush, C.M., Michels, G.J., Elliott, N.C., 2013. Remote distinction of a noxious weed (Musk Thistle: Carduus Nutans) using airborne hyperspectral imagery and the support vector machine classifier. Remote Sens. 5, 612–630.
Muthayya, S., Sugimoto, J.D., Montgomery, S., Maberly, G.F., 2014. An overview of global rice production, supply, trade, and consumption. Ann. N. Y. Acad. Sci. 1324, 7–14.
Nadir, S., Xiong, H.B., Zhu, Q., Zhang, X.L., Xu, H.Y., Li, J., Dongchen, W., Henry, D., Guo, X.Q., Khan, S., Suh, H.S., Lee, D.S., Chen, L.J., 2017. Weedy rice in sustainable rice production. A review. Agron. Sustain. Dev. 37.
Pantazi, X.E., Moshou, D., Bravo, C., 2016. Active learning system for weed species recognition based on hyperspectral sensing. Biosyst. Eng. 1–10.
Pantone, D.J., Baker, J.B., 1991. Weed-crop competition models and response-surface analysis of red rice competition in cultivated rice: a review. Crop Sci. 31, 1105–1110.
Peters, J., Baets, B.De., Verhoest, N.E.C., Samson, R., Degroeve, S., Becker, P.De., Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modelling. Ecol. Modell. 207, 304–318.
Qiu, J., Zhou, Y., Mao, L., Ye, C., Wang, W., Zhang, J., Yu, Y., Fu, F., Wang, Y., Qian, F., Qi, T., Wu, S., Sultana, M.H., Cao, Y.-N., Wang, Y., Timko, M.P., Ge, S., Fan, L., Lu, Y., 2017. Genomic variation associated with local adaptation of weedy rice during de-domestication. Nat. Commun. 8, 15323.
Rao, A.N., Johnson, D.E., Sivaprasad, B., Ladha, J.K., Mortimer, A.M., 2007. Weed management in direct-seeded rice. Adv. Agron. 93, 153–255.
Roy, M., Kumar, V.R., Kulkarni, B.D., Sanderson, J., Rhodes, M., Stappen, M., 1999. Vander A simple denoising algorithm using wavelet transform. Am. Inst. Chem. Eng. J. 45, 2461–2466.
Shafri, H.Z.M., Yusof, M.R.M., 2009. Trends and issues in noise reduction for hyperspectral vegetation reflectance spectra. Eur. J. Sci. Res. 29, 404–410.
Smith, R.J.J., 1988. Weed thresholds in Southern U.S. Rice, *Oryza sativa*. Weed Technol. 2, 232–241.
Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens. Environ. 62, 77–89.
Stellacci, A.M., Castrignanò, A., Troccoli, A., Basso, B., Buttafuoco, G., 2016. Selecting optimal hyperspectral bands to discriminate nitrogen status in durum wheat: a comparison of statistical approaches. Environ. Monit. Assess. 188, 1–15.
Thenkabail, P.S., Mariotto, I., Gumma, M.K., Middleton, E.M., Landis, D.R., Huemmrich, K.F., 2013. Selection of hyperspectral narrowbands (hnbs) and composition of hyperspectral twoband vegetation indices (HVIS) for biophysical characterization and discrimination of crop types using field reflectance and hyperion/EO-1 data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 427–439.
Tillett, N.D., Hague, T., Grundy, A.C., Dedousis, A.P., 2008. Mechanical within-row weed control for transplanted crops using computer vision. Biosyst. Eng. 99, 171–178.
Vaughan, L.K., Ottis, B.V., Prazak-Havey, A.M., Bormans, C.A., Sneller, C., Chandler, J.M., Park, W.D., 2001. Is all red rice found in commercial rice really *Oryza sativa*. Weed Sci. 49, 468–476.
Wendel, A., Underwood, J., 2016. Self-supervised weed detection in vegetable crops using ground based hyperspectral imaging. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 5128–5135.
Wu, D., Wang, S., Wang, N., Nie, P., He, Y., Sun, D.-W., Yao, J., 2013. Application of time series hyperspectral imaging (TS-HSI) for determining water distribution within beef and spectral kinetic analysis during dehydration. Food Bioprocess Technol. 6, 2943–2958.
Zhang, X., Liu, F., He, Y., Li, X., 2012. Application of hyperspectral imaging and chemometric calibrations for variety discrimination of maize seeds. Sensors 12, 17234–17246.