Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery

Junfeng Gao [a], *David Nuyttens* [b], *Peter Lootens* [c], *Yong He* [d,*], *Jan G. Pieters* [a]

[a] *Department of Biosystems Engineering, Ghent University, Coupure Links 653, 9000, Gent, Belgium*
[b] *Technology and Food Science Unit, ILVO, Burg. Van Gansberghelaan 115, 9820, Merelbeke, Belgium*
[c] *Plant Sciences Unit, ILVO, Caritasstraat 39, 9090, Melle, Belgium*
[d] *College of Biosystems Engineering and Food Science, Zhejiang University, Yuhangtang Road 866, 310058, Hangzhou, Zhejiang, China*

ARTICLE INFO

Article history:
Received 14 August 2017
Received in revised form
2 February 2018
Accepted 14 March 2018

Keywords:
Snapshot hyperspectral imaging
Machine learning
Plant classification
Hyperparameter tuning
Feature selection
Cross validation

This study explores the potential of a novel hyperspectral snapshot mosaic camera for weed and maize classification. The image processing, feature engineering and machine learning techniques were discussed when developing an optimal classification model for the three kinds of weeds and maize. A total set of 185 spectral features including reflectance and vegetation index features was constructed. Subsequently, the principal component analysis was used to reduce the redundancy of the constructed features, and the first 5 principal components, explaining over 95% variance ratio, were kept for further analysis. Furthermore, random forests as one of machine learning techniques were built for developing the classifier with three different combinations of features. Accuracy-oriented feature reduction was performed when choosing the optimal number of features for building the classification model. Moreover, hyperparameter tuning was explored for the optimal selection of random forest model. The results showed that the optimal random forest model with 30 important spectral features can achieve a mean correct classification rate of 1.0, 0.789, 0.691 and 0.752 for *Zea mays*, *Convolvulus arvensis*, *Rumex* and *Cirsium arvense*, respectively. The McNemar test showed an overall better performance of the optimal random forest model at the 0.05 significance level compared to the k-nearest neighbours (KNN) model.

* *Corresponding author*. Fax: +86 571 88982143.
E-mail address: yhe@zju.edu.cn (Y. He).

https://doi.org/10.1016/j.biosystemseng.2018.03.006
1537-5110/

Nomenclature

d	Dimension of features in the dataset
E_{ii}	Diagonal value of confusion matrix
F_1	A weighted average of precision and recall
α	One sample of the dataset
m	The number of trees to build Random Forests
M	Eigenvector Matrix
n	The number of selected features to build Random Forests
q_{01}	Number of samples misclassified by KNN but not by RF
q_{10}	Number of samples misclassified by RF but not by KNN
μ	Eigenvalues
β	Eigenvector
$R_{\text{calibrated}}(\lambda)$	Calibrated reflectance at wavelength λ
$Raw(\lambda)$	Uncalibrated digital number of pixel at wavelength λ
$W(\lambda)$	The digital number of calibration panel at wavelength λ
X_i	A bootstrap subset
Ψ	Statistical result from the McNemar test

Abbreviations

CV	Cross Validation
DC	Dark Current value
KNN	K-Nearest Neighbour
NDVI	Normalised Difference Vegetation Index
NIR	Near Infrared
OOB	out of bag samples
RF	Random Forests
ROI	Region of Interest
RVI	Ratio Vegetation Index
SSWM	Site-Specific Weed Management
VIs	Vegetation Indices
VB	Visual band

1. Introduction

Maize (*Zea mays*), one of the main cereals for food, forage and processed industrial products, is widely grown worldwide and a greater amount of maize is produced every year than any other grain (Ostrý, Malíř, & Pfohl-Leszkowicz, 2015). Although the maize yield increased to 1080 million tonnes in 2016 according to the statistics of the Food and Agriculture Organization of United Nations (FAO),[1] the quality of this crop still faced many problems such as weed infestation, animal pests and pathogens (Oerke, 2006). Weeds are one of the most important factors to limit maize production. They cause significant yield losses worldwide with an average of 29.2% if no weed control is applied (Dogan, Ünay, Boz, & Albay, 2004; Oerke & Steiner, 1996). Generally, most fields are infested with multiple weeds. For maize fields, *Convolvulus arvensis* and *Cirsium arvense* are the common weeds in central and western Europe (Meissle et al., 2010). In some certain circumstances, *Rumex* is also germinated among maize seedlings due to the easy propagation of its seeds. Besides, they are all perennial dicotyledons, which are suitable to control using chemical or mechanical ways (Macías, Castellano, & Molinillo, 2000; Zhang, 2003). The common weed management methods include prevention and cultural, mechanical, biological and chemical approaches (Harker & O'Donovan, 2013). Chemical methods such as spraying effective herbicides are the dominant management techniques for weed control in modern agriculture (Harker & O'Donovan, 2013). In most weed control methods, it is generally accepted to be most effective to control weeds in their early growth stage (López-Granados, 2011). Especially for maize crop, it is difficult to spray in practices in late growth stages due to the height of maize plants.

Under natural growing conditions, weeds are generally distributed in small patches, but farmers often uniformly spray herbicide in their fields, which is not in agreement with sustainable agriculture development and increases the cost of crop production. Site-specific weed management (SSWM), a precision agriculture approach, refers to the spatially variable application of weed control strategies for achieving the minimisation of herbicide usage (Shaw, 2005). It is useful in monitoring and managing weed patches at early growth stages (Shaner & Beckie, 2014). However, one of the main technical challenges of implementation lies in weed detection or classification (Shaner & Beckie, 2014; Slaughter, Giles, & Downey, 2008).

Currently, most weed detection studies can be classified into two groups. The first group utilises geometric differences for identification, such as leaf shape, texture, crop location. The second group differentiates weeds from crops using spectral reflectance characteristics (Slaughter et al., 2008; Thompson, Stafford, & Miller, 1991). Based on the two principles, various sensors, both imaging and non-imaging ones, have been applied in the investigation of weed detection in recent years. RGB cameras are widely applied for weed detection due to their general availability and low cost (Romeo et al., 2013; Tellaeche, Pajares, Burgos-Artizzu, & Ribeiro, 2011; Torres-Sánchez, López-Granados, De Castro, & Peña-Barragán, 2013; Gao et al., 2018). However, RGB cameras provide only limited spectral information as they only record information using three broad bands. To obtain more spectral information, a hyperspectral camera was introduced in classification applications (Gao, Li, Zhu, & He, 2013). Hyperspectral imaging sensors often involve more and narrower bands to gain detailed spectral information. 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 applications of line-scanning hyperspectral imagery for weed species recognition were presented by Okamoto, Murata, Kataoka, and Hata (2007) and Pantazi, Moshou, and Bravo (2016). Wendel and Underwood (2016) also developed a self-supervised training data generation and weed detection system for vegetable fields. However, these systems, based on line-scanning hyperspectral sensors, are negatively affected by the rapid motion of platforms or objects because of the need to scan image. A snapshot hyperspectral system, without scanning,

[1] FAOSTAT data website, http://www.fao.org/faostat/en/#home; accessed 21 January 2018.