Sugar beet and volunteer potato classification using Bag-of-Visual-Words model, Scale-Invariant Feature Transformation, or Speeded Up Robust Feature descriptors and crop row information

Hyun K. Suh*, Jan Willem Hofstee, Joris IJsselmuiden, Eldert J. van Henten

Farm Technology Group, Wageningen University, P.O.box 16, 6700 AA, Wageningen, The Netherlands

Abstract

One of the most important steps in vision-based weed detection systems is the classification of weeds growing amongst crops. In EU SmartBot project it was required to effectively control more than 95% of volunteer potatoes and ensure less than 5% of damage of sugar beet. Classification features such as colour, shape and texture have been used individually or in combination for classification studies but they have proved unable to reach the required classification accuracy under natural and varying daylight conditions. A classification algorithm was developed using a Bag-of-Visual-Words (BoVW) model based on Scale-Invariant Feature Transformation (SIFT) or Speeded Up Robust Feature (SURF) features with crop row information in the form of the Out-of-Row Regional Index (ORRI). The highest classification accuracy (96.5% with zero false-negatives) was obtained using SIFT and ORRI with Support Vector Machine (SVM) which is considerably better than previously reported research although its 7% false-positives deviated from the requirements. The average classification time of 0.10 - 0.11 s met the real-time requirements. The SIFT descriptor showed better classification accuracy than the SURF, but classification time did not vary significantly.

Adding location information (ORRI) significantly improved overall classification accuracy. SVM showed better classification performance than random forest and neural network. The proposed approach proved its potential under varying natural light conditions, but implementing a practical system, including vegetation segmentation and weed removal may potentially reduce the overall performance and more research is needed.

Keywords

Weed classification, Bag-of-Visual-Words, SIFT, SURF, posterior probability

Nomenclature

Abbreviations

BoVW	Bag-of-Visual-Words
SIFT	Scale-Invariant Feature Transformation
SURF	Speeded Up Robust Feature
ORRI	Out-of-Row Regional Index
SVM	Support Vector Machine
kNN	k-Nearest Neighbours
RGB	Red-Green-Blue
EG-RB plane	Excessive Green - Red minus Blue plane
TP	True-Positive
FP	False-Positive
TN	True-Negative
FN	False-Negative

Symbols

m	metre
mm	millimetre
m s^{-1}	metre per second
min	minute
s	second

1. Introduction

Within the EU-funded project SmartBot (SmartBot), a small-sized robot was developed for vision based precision control of volunteer potatoes (weed) in a sugar beet field (Fig. 1). Due to the small size of the robot and its battery operation, the platform design had to refrain from using additional infrastructure and should be able to robustly detect weeds in scenes that are fully exposed to ambient lighting conditions (Suh, Hofstee, & Van Henten, in press). Additional infrastructure such as a hoods and lighting, as for example were used by Nieuwenhuizen et al. (2010) and Haug et al. (2014), was not considered viable.

One of the most important steps in vision-based weed detection is the classification of weeds among crops. The output of this classification is a fundamental element in the subsequent process of weed control either by chemical spraying or mechanical actuation (Behmann, Mahlein, Rumpf, Römer, & Plümer, 2015). In a system for weed detection, vegetation segmentation is followed by classification of the segmented vegetation into weeds and crop. This classification step traditionally involves two aspects: 1) selection of the discriminative features and 2) selection of the classification technique (classifier) to differentiate between weeds and crop.

Regarding the features used for discrimination, many studies have used colour, shape (biological morphology) and texture on an individual basis or in combination (Ahmed, Al-Mamun, Bari, Hossain, & Kwan, 2012; Åstrand & Baerveldt, 2002; Gebhardt & Kühbauch, 2007; Nieuwenhuizen, Tang, Hofstee, Müller, & Van Henten, 2007; Pérez, López, Benlloch, & Christensen, 2000; Persson & Åstrand, 2008; Slaughter, Giles, & Downey, 2008; Swain, Nørremark, Jørgensen, Midtiby, & Green, 2011; Zhang, Kodagoda, Ruiz, Katupitiya, & Dissanayake, 2010). These features are intuitive and easy-to-implement but may have limited discrimination ability under ambient lighting conditions.

In a system that is required work under ambient light conditions, the use of colour features may not yield robust classification (Lee et al., 2010). In the field, illumination constantly changes because of the varying sunlight and weather conditions. These variations in illumination greatly affect the Red-Green-Blue (RGB) pixel values of the acquired field images and lead to an inconsistent colour representation of plants (Sojodishijani, Ramli, Rostami, Samsudin, & Saripan, 2010; Teixidó et al., 2012). Additionally, irrespective of the illumination, it is sometimes hard, if not impossible, to differentiate between volunteer potato and sugar beet using colour features. Usually, volunteer potato has a darker green colour than sugar beet (Fig. 2a) which results in a separable pixel distribution in the EG-RB colour plane (Fig. 2c). However, as is shown in Fig. 2b, volunteer potato occasionally has the same colour as sugar beet which makes them inseparable in the EG-RB colour plane (Fig. 2d). Also, the colour of plants may change depending on their growth stage and nutritional status (Muñoz-Huerta et al., 2013) with plant leaves sometimes even turning yellow in the summer (Fig. 3).

Shape and texture may also not be sufficiently discriminating features for successful classification of sugar beet and volunteer potato in the field. Camargo Neto et al. (2006), Swain et al. (2011), and Rumpf et al. (2012) showed that leaf edge information, plant orientation, and shape could serve as discriminative features. However, results obtained under laboratory conditions in a highly structured environment do not easily translate to real field conditions. Wind, shadow, and specular reflection of sunlight make it difficult for clear recognition of the shape of the plants in the field (Kazmi, Garcia-Ruiz, Nielsen, Rasmussen, & Andersen, 2015). Some studies have shown that texture has the potential to discriminate between broad- and narrow-leaf plants as both have clearly different textural properties (Gebhardt & Kühbauch, 2007; Ishak, Hussain, & Mustafa, 2009; Van Evert, Polder, Van Der Heijden, Kempenaar, & Lotz, 2009). However, sugar beet and volunteer potato have similar textural properties that cannot easily be discriminated (Vollebregt, 2013). Therefore, a solution was needed to classify sugar beet and volunteer potato that would not depend on colour, shape, and textural features.

A potential method to resolve the afore-mentioned issues and meet the performance requirements is to use counter-intuitive features (i.e. local descriptors) extracted by Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) or Speeded Up Robust Features (SURF) (Bay, Ess, Tuytelaars, & Van Gool, 2008). Both SIFT and SURF are invariant to illumination and colour while providing strong performance against noise. The SIFT descriptor has been used for weed classification and recognition in several recent studies (Kazmi et al., 2015; Kounalakis, Triantafyllidis, & Nalpantidis, 2016; Wilf et al., 2016). Using the SIFT descriptor, Wilf et al. (2016) proposed a leaf identification procedure based on a machine learning approach. Although they acquired images under controlled environmental conditions with the manual arrangement of the leaves, their study showed the potential of the SIFT descriptor for leaf classification. Kazmi et al. (2015) used both SIFT or SURF descriptors to classify sugar beet and creeping thistle under field conditions. Their study showed the potential of using local descriptor features for thistle detection. They combined these local descriptors with the features of surface colour and edge shapes. Using k-Nearest Neighbours (kNN) and SVM classifiers a very promising classification performance was achieved. However, their study was limited to detecting creeping thistle in a sugar beet crop, two species having clearly different textural features. Also, the field images were mostly acquired using a cover preventing direct access of sunlight to the scene, quite a distinct difference with the daylight conditions the SmartBot robot is confronted with.

A common way for classifying images using SIFT or SUFT descriptors is to use a Bag-of-Visual-Words (BoVW) approach. The BoVW approach has demonstrated good performance in many computer vision applications such as object and scene classification (Law, Thome, & Cord, 2014; Tsai, 2012; Zhou, Zhou, & Hu, 2013). The BoVW evolved from the original Bag-of-Words methodology which was first proposed in the field of text analysis and information retrieval (Bosch, Muñoz, & Martí, 2007). In text analysis and information retrieval, each appearance of a word is recognised as a feature and is represented in the form of a bag of words, an orderless document representation of vocabulary (Salton & McGill, 1983). Once the Bag-of-Words model learns a vocabulary from all the documents, then each document can be classified by the number of times each word appears (occurrence). The same methodology and concept are applied in image classification in BoVW. The extracted features from an image are treated as a visual word, and the BoVW model is formed based on the occurrence of each visual word. Once the BoVW approach has learned each visual word from all the images, then each image can be classified by the number of times each visual word appears (occurrence).

This paper presents a classification algorithm using a Bag-of-Visual-Words model, SIFT or SURF descriptors. SIFT is known to provide better classification performance than SURF, but it is said to be several times slower than SURF (Csurka, Dance, Fan, Willamowski, & Bray, 2004; Khan, McCane, & Wyvill, 2011; Wu et al., 2013). This research aimed to verify the difference in performance between SIFT and SURF by assessing classification accuracy and computation time on similar datasets (images) obtained in the field in 2015. Since neither SIFT nor SURF uses location related features, crop row information was used as an additional feature and added to the feature set to assess whether that would improve the classification accuracy.

SURF, SIFT and crop row information provide the features but require further processing for classification. Due to the challenging nature of the agricultural environment, and complexity of plant materials, (Suh et al., in press), it is hard to select a-priori one particular classifier which performs best in the classification task at hand. To provide more insight into the performance differences found amongst different classifiers, the Support Vector Machine (SVM), random forest, and neural network classifiers were compared. These classifiers have been used in many agricultural applications (Ahmed et al., 2012; Cho, Lee, & Jeong, 2002; Jeon, Tian, & Zhu, 2011; Lottes, Hörferlin, Sander, & Stachniss, 2016).

To estimate the amount of certainty of the classification output, a posterior probability of the output of the SVM was calculated using a method proposed by Platt (1999). The posterior probability might provide useful information for weed control in practice since the action of removing volunteer potato should only be applied to those potato plants that are classified with a high confidence, while the control action should be skipped for those potato plants that are classified with a low confidence to prevent undesired destruction of the sugar beet.

Within the context of the SmartBot weeding application, following requirements were set by the previous study of Nieuwenhuizen (2009): the resulting automatic weeding system should be able to effectively control more than 95% of the volunteer potatoes as well as ensuring less than 5% of damage of the sugar beet plants. Therefore, classification accuracy should be considerably higher than 95% with a misclassification level of both sugar beet (false-negative) and volunteer potato (false-positive) of less than 5%. In addition, a classification time of less than 1 s per field image is required for feasible real-time field application. In this paper the classification process is evaluated in view of these requirements.

The first section of this paper describes the processing method of the BoVW model construction using the SIFT or SURF features. The following section describes the acquisition and selection of the image dataset, quantitative performance measure, and estimation of the posterior probability of SVM outputs. The experimental results are shown with the corresponding discussions. Lastly, conclusions are drawn.

2. The classification process

The classification process consists of the following procedures: 1) feature extraction using SIFT or SURF descriptors as well as crop row information, 2) feature clustering for visual vocabulary generation, 3) feature quantisation, 4) classification with SVM, random forest or neural network classifiers. The image classification process is shown in Fig. 4, and each component will be described in more detail in the following section.

2.1. Feature extraction with SIFT or SURF descriptors and Out of Row Regional Index (ORRI)

The first step involved the extraction of local features from the training images (Fig. 4a, Fig. 5a). For the selection of the feature extraction point (keypoint) within an image, a regular grid-point based sampling was used as several studies reported that it provided robust performance (Fei-Fei & Perona, 2005; Law et al., 2014; Tsai, 2012). Grid size refers to the density of the feature extraction within a given image. In a preliminary study (Table 4) it was found that a grid size of 3 × 3 pixels proved to perform best for SIFT and 6 × 6 pixels for SURF.

During the generation of the visual vocabulary, the spatial location of the feature within an image was ignored. However, the spatial location may contain some valuable information for weed and crop discrimination. Uijlings et al. (2009) reported that the classification performance of BoVW using SVM was considerably improved when they included spatial information (contextual information) into the algorithm.

In a classification problem with one single object in an image scene, the location of the object within an image may not carry any additional and useful information. However, with weed detection in the field, the location of each plant can play a significant role in the plant recognition. For example, sugar beet plants are cultivated in rows (Åstrand & Baerveldt, 2002). Due to precision seeding, the crop row width and plant spacing within a row are fixed. For this reason, most of the sugar beet are found inside crop rows whilst weeds can be found randomly distributed across the field. Any green plant that is located far away from the crop rows is unlikely to be a crop but very likely to be a weed.

Inspired by the details mentioned above, an out-of-row regional index (ORRI) was generated for each plant on the basis of the out-of-row distance (Fig. 6), a distance between the centre of the plant to the nearest crop row. The ORRI was added to the BoVW feature set. Identifying weeds as weeds when located outside the crop row may sound trivial, which it is. However, it was hypothesised that adding ORRI information during the learning process might add an extra discriminatory dimension, and thus might enhance the discriminative power in the classification. The details of the ORRI generation are described below.

Firstly, the location of three crop rows was manually estimated. Secondly, the distance between the centres of each plant to the nearest crop row, the out-of-row distance, was estimated. Thirdly, each plant received a value for the ORRI from the set [0.3, 0.6, 0.9] based on the following rules:

$$\text{ORRI} = \begin{cases} 0.3 & if \quad \text{out-of-row distance} < 80\, pixels \\ 0.6 & if \quad 80 \leq \text{out-of-row distance} < 160\ pixels \\ 0.9 & otherwise \end{cases} \tag{1}$$

where the out-of-row distance is represented as a pixel value (one pixel corresponds to approximately 1 mm in the field).

For the regional index discrete values of 0.3, 0.6 and 0.9 were used instead of continuous values because it was expected that the estimation of the crop rows and centre point of the plant would be likely to introduce noise.

2.2. Feature clustering for visual vocabulary generation

In this step, extracted features were clustered using k-means clustering, a common method for visual vocabulary generation (Fig. 4b, Fig. 5b). Each cluster centroid determined by k-means clustering was considered as a visual word. Based on a preliminary study, the number of clusters and thus the vocabulary size was set to 500 (Table 4).

If the vocabulary size (number of clusters) is too small, the set of visual words may be too limited to represent all the important features of images, and thus may lead to poor classification performance (Yang, Jiang, Hauptmann, & Ngo, 2007). On the other hand, if the vocabulary size is too large, there is a higher chance of overfitting the training dataset. In addition, a large size of the vocabulary also requires more processing power.

2.3. Feature quantisation

Once the visual vocabulary was generated, the features (descriptors) extracted from each image were assigned to each visual word to construct a histogram of visual word occurrences (Fig. 4c, Fig. 5c). Using the Euclidean distance, each extracted feature was allocated to its nearest visual word (nearest neighbour). A histogram of visual words was then generated by counting the number of features that were assigned to each visual word. The length of the histogram was equal to the number of cluster centres generated by k-means clustering, where the n^{th} value in the histogram was the occurrence of the n^{th} visual word. This process is commonly called feature quantisation (Kato & Harada, 2014). A histogram of visual word occurrence generated from images of sugar beet and volunteer potato is shown in Fig. 7.

2.4. Classification based on supervised learning

Supervised learning was used to train the classifiers for differentiation between sugar beet and potatoes (Fig. 4d, Fig. 5d). Three classifiers were used in this study: SVM, random forest and a neural network. In the SVM, three different polynomial kernels (linear, quadratic and cubic) were assessed. For the evaluation of the classifiers, 10-fold cross-validation was used. Some details of random forest and neural network are described below.

2.4.1 Support Vector Machine (SVM)

The SVM is a supervised learning model based on the theory of statistical learning (Vapnik, 1995). SVM is one of the most widely used classification models in machine learning applications and often reaches high performance in high-dimensional problems with small sample problems (Csurka et al., 2004; Li, 2011). The basic principle of SVM is to find the optimal hyperplane which separates classes with minimum error.

2.4.2 Random forest (Ensemble Classifier)

A random forest classifier, an ensemble method that consists of multiple decision trees, was used for this study. Random forest, as the name says, is constructed from decision trees, more precisely it is a collection of tree-structured classifiers. Each decision tree provides a classification "vote," and the majority vote is selected for the final classification (Chan & Paelinckx, 2008; Liaw & Wiener, 2002; Polikar, 2006). Breiman (2001) reported that the performance of a random forest was superior to other learning algorithms. Rodriguez-Galiano et al. (2012) indicated that the random forest is relatively robust to outliers and noise as well as computationally less expensive than other tree ensemble methods.

2.2.3 Neural Network

An artificial neural network consists of multiple nodes and neurons that are connected in the layers. Compared to other classifiers, according to Behmann et al. (2015), a neural network requires less prior information and is robust to noise thus particularly suitable for the modelling of optical sensor data. In this study, a feed-forward back propagation neural network was used. The neural network used in this research consists of one hidden layer with 150 neurons besides an input and an output layer. In the input layer, histograms of visual words were utilized, and in the output layer, sugar beet was represented by [1, 0] while volunteer potato was represented by [0, 1].

3. Experiment setup

3.1. Field image collection and image dataset

To acquire crop images, a camera was mounted at the height of 1 m perpendicular to the ground on a custom-made frame carried by a mobile platform (Husky A200, Clearpath, Canada) (Fig. 8). A stereo camera (NSC1005c, NIT, France) was equipped with two Kowa 5 mm lenses (LM5JC10M, Kowa, Japan) with a fixed aperture. The camera was set to operate in an automatic acquisition mode with default settings. The camera images from left and right sensors were acquired each having an image resolution of 1280 × 580 pixels. The ground-covered area was 1.3 × 0.7 m per image (pair), corresponding to three crop rows of sugar beet. The acquisition program was implemented in LabVIEW (National Instruments, Austin, TX, USA) and acquired five images per second. Raw format images (TIFF) were initially acquired in the field, and debayer was processed offline to convert the raw format image into RGB colour. Field images were taken while the mobile platform was manually controlled with a joystick and driven along crop rows using a controlled travelling speed of 0.5 m s^{-1}. Sugar beet were sown in April 2015 into sandy and clay soil at Unifarm experimental sites in Wageningen, The Netherlands. One week after sowing the sugar beet, potatoes were planted in random locations throughout the fields. Crop images were acquired for two days in the morning and afternoon on 1 June and 5 June, 2015.

For the labelled image dataset used in this study, a total of 400 individual plant images was manually extracted from selected field images: 200 sugar beet plants and 200 volunteer potato plants. During the selection of this image dataset, images with different illuminations levels were considered as well as images containing shadows. The size of each plant image in the dataset varied from the smallest size of 65 × 65 pixels to the largest of 305 × 315 pixels. Example images in the dataset are shown in Fig. 9.

In the image dataset, all sugar beet were found within crop rows (out-of-row distance < 80 pixels), having an ORRI of 0.3. On the other hand, volunteer potatoes were found inside and outside crop rows. The number of volunteer potatoes found inside the crop row (out-of-row distance < 80 pixels), i.e. ORRI = 0.3, was 55; while the number of volunteer potatoes found outside the crop row (out-of-row distance ≥ 80 pixels), i.e. ORRI > 0.3, was 145.

3.2. Performance measure and system platform

In this study a binary classification was carried out; i.e. sugar beet or volunteer potato. The classification performance measures used in this study are described below.

A confusion matrix (Table 1) was used to assess and compare the classification performances. The classification accuracy was calculated along with training and classification time since this approach should, in the end, yield a real-time field application. Each classifier was validated using 10-fold cross-validation. The classification accuracy and training time were averaged over ten trials with a random split of the dataset. The training time included times for classifier training as well as extracting features and building a visual vocabulary. The classification time was measured for the prediction of one plant image. All images were processed in Matlab 2015b (The MathWorks Inc, Natick, MA, USA) using the Computer Vision System Toolbox™, Neural Network Toolbox™, and VLFeat library for Matlab (Vedaldi & Fulkerson, 2008). Processing time was measured on an Intel® Core™ i7-377T 2.5 GHz processor with 8 GB memory running 64-bit Windows 7.

$$\text{Classification Accuracy} = \frac{TP + TN}{TP + FN + FP + FN} \quad (2)$$

where: TP is true-positive; FP is false-positive; TN is true-negative; FN is false-negative

3.3. *Estimated posterior probability of SVM outputs*

Platt (1999) proposed a method using a sigmoid function to calculate and estimate the posterior probability for SVM classifier. Since then, this method has been used in many applications (Lin, Lin, & Weng, 2007) as it is an useful measure to provide the degree of certainty (belief) of the classification output. In this study, a posterior probability was estimated for the SVM using a linear kernel and employing the ORRI in the feature set.

4. Results

The classification performances of BoVW using SIFT or SURF descriptors are summarised with true-positive (TP), false-negative (FN), false-positive (FP), true-negative (TN), classification accuracy, training time and classification time in Table 2 and Table 3. In these tables, it is also indicated whether the ORRI was used.

4.1 Classification accuracy

In Table 2, using SIFT features and ORRI, the highest classification accuracy obtained was 96.5%; while the lowest classification accuracy obtained was 83.5%. Three classifier models (SVM linear, SVM quadratic, and neural network) showed classification accuracies ≥ 95%, thus meeting the requirements. Likewise, in Table 3, using SURF features and ORRI, the highest classification accuracy obtained was 94.5%; while the lowest classification accuracy obtained was 84.5%. None of the classifier models showed a classification accuracy of ≥ 95%, and thus using SURF features and ORRI did not meet the requirements set at the beginning of this research.

4.2 Misclassification rate (false-positive and false-negative)

The false-negative values obtained for the cases with the highest classification accuracies using SIFT features with ORRI and using SURF features with ORRI were both zero (Table 2 and Table 3). Meeting the requirements, in these cases all the sugar beet plants were correctly classified as a sugar beet, and thus no crop would be eliminated by a weed control operation (0% of undesired control of sugar beet plants). However, in these cases the false-positive values obtained with the highest classification accuracies using SIFT with ORRI, and using SURF with ORRI were 14 (7%) and 22 (11%), respectively. So, 7% and 11% of volunteer potato were classified as sugar beet, respectively, and thus would not be destroyed. These false-positive values do not meet the requirements (misclassification: less than 5%).

4.3 Training and classification time

Training time in this work includes the time needed for training of the classifiers as well as for extracting SIFT or SURF features and building the visual vocabulary. SVMs required 218-222 s and 175-183 s of training time using SIFT with ORRI and SURF with ORRI, respectively; while the neural network required 260 s and 190 s of training time using SIFT with ORRI and SURF with ORRI, respectively. The training times needed by all classifiers were reasonable, considering that training can be done offline and may not have to be repeated very often.

The classification time indicates the time required to classify the class of a single plant image using a trained classifier. For all classifiers, an average time of 0.10 - 0.11 s was needed for classification, which is a reasonable value when the real-time application in the field is considered.

4.4 SIFT compared to SURF

SIFT is known to provide better classification performance than SURF, however, at the expense of more computation time. In view of classification accuracy, this observation was confirmed in this research. Overall, in line with findings reported in the literature, using SIFT features resulted in better classification accuracy than using SURF features. Without ORRI, the accuracy improved on average 6.2% when using SIFT features instead of using SURF features. With ORRI this difference reduced, and on average, the accuracy improved by 2.6% when using SIFT features instead of SURF features. SIFT features required more training time than SURF features. On average 46 s more training time was required when using SIFT instead of SURF. Classification time did not differ much for SIFT and SURF, however, and this result does not match with observations reported in the literature. On average 0.11 s and 0.10 s was needed when using SIFT and SURF, respectively.

4.5 Out-of-Row Regional Index (ORRI)

For all classifiers classification accuracy improved with ORRI. It was earlier hypothesised that adding spatial information (ORRI) during the learning process adds an extra discriminatory dimension which enhances the discriminative power of the classification of sugar beet and volunteer potato. This hypothesis was confirmed by the results, showing that the classification accuracy considerably improved when implementing ORRI in the classification algorithm. Averaged over all classifiers, the improvement in classification accuracy using the ORRI was 4.5% and 8% when using the SIFT and SURF features, respectively.

For comparison, it is worth noting that when using the ORRI as the only feature, a classification accuracy of 86.3% was obtained in all classifiers with TP, FN, FP and TN of 200, 0, 55, 145, respectively. This is a relevant result because, as mentioned earlier, in the dataset a total of 255 plants (200 sugar beet and 55 volunteer potatoes) were found inside crop rows (out-of-row distance < 80 pixels, having an ORRI of 0.3). In Table 3, it can be seen that adding ORRI to SURF and classifying with a SVM and a linear kernel results in a change of classification for 45 plants (FN:25→0, FP:42→22). Further analysis of the individual images revealed that 29 of these 45 images had an ORRI 0.3, so were inside crop rows: 25 of them were sugar beet plants, and four of them were volunteer potato plants. Interestingly enough, these 25 sugar beet, though being inside the crop rows, were not properly classified by SURF only (without ORRI). This is unsurprising because SURF does not employ any locational feature. More interesting is to note that four of the images were volunteer potato plants. So, by adding a location feature in training improved the classification for volunteer potato inside crop rows, which is a real challenge in weed classification.

When training time is considered with ORRI, SIFT required on average 6.7 s more time when training without ORRI. Likewise, training with ORRI using SURF required on average 7.2 s more time than training without ORRI. When it comes to classification, however, the use of ORRI did not lead to a considerable increase in calculation time.

4.6 Comparison of SVM, Random Forest and Neural Network classifiers

SVM classifiers with a linear and quadratic kernel showed better classification accuracies than random forest and neural network, though the SVM and neural network did not differ much. In Table 2, using SIFT features and ORRI, the highest classification accuracy of 96.5% was obtained with a SVM and a quadratic kernel; while the lowest classification accuracy of 90.5%

was obtained with the random forest. In Table 3, using SURF features and ORRI, the highest classification accuracy of 94.5% was obtained with a SVM and both a linear and a quadratic kernel; while the lowest classification accuracy of 84.5% was obtained with the random forest.

4.7 Grid size and vocabulary size

Classification accuracy with different sizes of grid and vocabulary are compared in Table 4. Using small grid sizes tended to produce better result than large grid sizes. However, vocabulary size did not seem to produce any regular pattern of performance. In fact, grid and vocabulary size are not formally related, but a certain combination (in this case, a grid size of 6 × 6 and vocabulary size of 500) showed a better performance than others in this study. Therefore, a grid size of 6 × 6 pixels and vocabulary size of 500 were used as an optimal combination when employing the SURF descriptor because the highest classification accuracy (94.5%) was achieved with these settings. For the SIFT descriptor, a grid size of 3 × 3 pixels with a vocabulary size of 500 was used as the highest classification accuracy was achieved with these settings.

4.8 Estimated posterior probability

The posterior probabilities of the SVM with linear kernel using SIFT features and ORRI were calculated and visualized in the form of a box-and-whiskers plot in Fig. 10. All sugar beet images were correctly classified as sugar beet (true-positive), and on average the posterior probability was 0.96 with a standard deviation of 0.09. A total of 180 volunteer potato images (out of 200) was correctly classified as volunteer potatoes (true-negative), and for these images the average posterior probability was 0.98 with a standard deviation of 0.02. However, 20 volunteer potato images were incorrectly classified as sugar beet (false-positive). With an average value of 0.49 and standard deviation of 0.27, in these cases, the average posterior probability was lower than in the true-positive and true-negative cases. These results indicate that the classifier was more confident in case of correct classification than when making a false classification.

The above results show that the posterior probability might provide useful information for weed control in practice. Using the posterior probability, the action to remove volunteer potato should only be applied to those plants that are classified with a high confidence. Figure 11, for example, shows the classification results with the posterior probability with a field image. Plants 1 to 6 are sugar beet whereas plants 7 to 9 are volunteer potatoes (Fig. 11a and 11b). In Figure 11c, plants 2 to 6 are correctly classified as sugar beet with a posterior probability of 0.86 and higher; and plants 7 to 9 are correctly classified as volunteer potatoes with a posterior probability of 1.0. However, plant 1 (sugar beet) is incorrectly classified as a volunteer potato (false-negative). In this case, the posterior probability is 0.54 and considerably lower than the others. In such a case, based on the lower posterior probability, it might be beneficial to skip the weed control action because since it would lead to the destruction of the crop.

5. Discussion

5.1 Classification accuracy

The classification accuracy obtained using BoVW approach with ORRI exceeded previously reported accuracies; e.g. Nieuwenhuizen et al. (2010) and Persson & Åstrand (2008). Considering the different illuminations levels and shadows in the image dataset, the highest classification accuracy (96.5%) obtained in this study is considerably better than any other approaches with colour, shape, and texture features in the literature for weed classification. However, the overall performance of weed control also depends on the performance of vegetation segmentation as well as the actuation performance of the weeding device. If the individual performance of either one of these two operations would be < 100%; thus the classification accuracy should be considerably higher than 95% in order for the automatic weeding system to effectively control more than 95% of the volunteer potatoes in the field. In this regard, the highest classification accuracy achieved in this study (96.5%) may not be enough to satisfy the overall performance of volunteer potato control since it is not significantly higher than 95%.

The obtained results were based on manually extracted plant images. Thus, the proposed approach itself does not lead to the precise detection of volunteer potato in field images. To make a complete system for the use of weed control in the field, vegetation segmentation and weed removal operation needs to be integrated. During integration, overlapping plant cases need to be considered as well.

5.2 Misclassification rate (false-positive and false-negative)

For weed control in practice, it is critical to have a large as possible number of true-positives as well as a large as possible number of true-negatives. Not only that, but it is also important to consider both the number of false-negatives (the number of sugar beet plants that are classified as volunteer potatoes) and the number of false-positives (the number of volunteer potato plants that are classified as sugar beet). The false-negatives lead to the removal of the cash crop caused by the misclassification, thus keeping the number of false-negatives as small as possible is critical (Lottes et al., 2016). However, it is desirable to keep the number of false-positives as small as possible. If there are many left over volunteer potato plants caused by misclassification, then a weed control robot may need to drive repetitively across the field to meet the statutory regulation in the Netherlands (Nieuwenhuizen, 2009). The economic consequences of false-negatives and false-positive detections require further research.

5.3 Calculation time

From general observations of the field images, there is an average of 6-8 plants in one image. Based on these number of plants in an image, the whole plant classification of one field image may take up to 0.8 seconds (including all other steps in the image processing) using SVM classifiers, which is acceptable for our real-time application (<1 s for one field image). The classification time, of course, depends on the size of each plant found in an image, and can be further improved with a parallel-processing approach. In addition, the size of the grid and vocabulary also influences the classification and processing time. If the processing time is highly critical for certain applications, grid and vocabulary size can be changed to reduce the processing time at the expense of classification accuracy.

5.4 SIFT and SURF

Several studies have indicated that SURF is rapid for computation and matching (Khan et al., 2011; Panchal, Panchal, & Shah, 2013; Wu et al., 2013; Zagoris et al., 2014). In this research SIFT required more training time than SURF. However, the classification times required for SIFT and SURF were not considerably different in this study. This result is not accord with the literature. In this study, the different grid sizes used for SIFT and SURF may have caused classification times to be similar.

There is room for improvement in terms of the classification accuracy. During the extraction of SIFT and SURF descriptors, dataset images were converted to greyscale ignoring all the colour information (RGB) because SIFT and SURF operate on intensity information only. However, colour may carry some discriminative information for the classification of sugar beet and volunteer potato. To overcome the abovementioned weakness of SIFT and SURF descriptors, several variations of SIFT and SURF have been proposed in the literature using colour features such as rgSIFT, Transformed colour SIFT, and Color-SURF (Fan, Men, Chen, & Yang, 2009; Van De Sande, Gevers, & Snoek, 2008) to improve classification accuracy. Similarly, Rassem and Khoo (2011) proposed not to convert RGB image to greyscale but to apply the feature extraction on each RGB channel. The extracted features from the individual colour channels may add extra discriminative power for classification, and validating this hypothesis is, therefore, a topic of a future study. As indicated in Fig. 2, the added value of using also colour might be limited in cases where, as here, crop and weed plants have similar colour values.

5.5 Out-of-Row Regional Index (ORRI)

Combining ORRI considerably improved the classification accuracy enhancing the discriminative power of the classification. However, spatial information of each plant (ORRI) including crop rows and out-of-row distance was manually estimated in this study. For an automated field application using a mobile robot, the estimation of crop rows and out-of-row distance should be automated as well. Algorithms for crop row detection have been presented in several studies (Guerrero et al., 2013; Hiremath, Van Evert, Braak, Stein, & Van der Heijden, 2014; Kise, Zhang, Rovira Más, & Mas, 2005; Leemans & Destain, 2006; Romeo et al., 2012; Søgaard & Olsen, 2003), but these algorithms are likely to introduce noise. Thus, in the current approach, regional index (0.3, 0.6 and 0.9) was used instead of a precise number for the out-of-row distance to compensate any potential noise.

5.6 Classifiers

Based on the results obtained in this study SVM classifiers would be an easy and plain choice for field applications, not only because SVM classifiers showed better classification performance in most cases than random forest and neural network, but also because SVMs are easier to implement than other classifiers. However, the neural network also performed quite well, showing similar classification performance as SVMs, although a simple network structure (1 hidden layer) was used in this study. Kanellopoulos & Wilkinson (1997) indicated that multi-layer network architecture might be potentially more powerful than a simple network. This has been confirmed over the past few decades in various applications (LeCun, Bengio, & Hinton, 2015). Thus, adding more layers is likely lead to better classification performance.

5.7 Posterior probability

The posterior probability estimated by Platt's method offers additional information during the weed control action, which can be useful in practice. Using this posterior probability, the action to remove volunteer potato should only be applied to those volunteer potato plants that are classified with a high confidence. Volunteer potato plants that are classified with lower confidence might be better skipped because it might lead to the undesired destruction of the sugar beet. However, the characteristics and applicability of this approach need further study.

Two studies have indicated that probability estimation using Platt's method could be ineffective in some cases especially for large datasets (Niculescu-Mizil & Caruana, 2005; Perez-Cruz, Martinez-Olmos, & Murillo-Fuentes, 2007). To compensate for the weakness of Platt's method, Lin et al. (2007) proposed an improved algorithm which theoretically avoids numerical difficulties. When large datasets are concerned, their proposed method for probability estimation might be a better choice.

In this study, the posterior probability was estimated only for SVM classifier. However, the posterior probability for other classifiers, such as random forest and neural network, can also be estimated using a method proposed by Niculescu-Mizil and Caruana (2005). They reported that random forest and the neural network classifiers provided well-calibrated probabilities having no bias compared to SVM. Investigating the posterior probability for other classifiers would be a future study topic.

5.8 Reflection on contribution to weed control

In this study, binary classification (between sugar beet and volunteer potato) was proposed based on the assumption that in most cases plants found in sugar beet fields are either sugar beet or volunteer potato. However, in an agricultural field, a variety of different weed species is likely to be found. A future study topic might include a multiclass classification of weed species within a crop. Classification of other crop species may also benefit from the proposed approach.

6. Conclusions

In this study, an algorithm using a Bag-of-Visual-Words model and SIFT or SURF descriptors as well as crop row information in the form of the ORRI was proposed for the classification of sugar beet and volunteer potato under natural and varying daylight conditions. In EU SmartBot project it was required to effectively control > 95% of volunteer potatoes (weed) and to ensure < 5% of undesired control of the sugar beet crop. Considering the different illuminations levels and shadows in the image dataset, the highest classification accuracy of 96.5% with false-negative of 0% which was obtained using SIFT features and ORRI with SVM classifier is considerably better than any other approaches found in the literature that used colour, shape and textural features. Therefore, the proposed approach proved its potential under ambient light conditions although the false-positive rate of 7% deviates from the requirements (misclassification: < 5%). An average time of 0.10 - 0.11 s was needed for classification, which is a reasonable value when the real-time application in the field is considered and is well within the required 1 s. However, implementing a full pipeline including vegetation segmentation and weed removal operation may potentially reduce the overall performance. The SIFT descriptor showed better classification accuracy than using the SURF descriptor. Using SIFT required more training time than SURF, but the classification time required for SIFT and SURF was not considerably different.

Adding crop row information as an additional feature (ORRI) significantly improved the overall classification accuracy. However, for an automated field application using a weed control robot, the estimation of crop rows and out-of-row distance should be automated and might potentially introduce noise.

In this application, SVM classifiers showed better classification performance than random forest and neural network. However, a neural network with multi-layer architecture would potentially improve the performance.

The posterior probability estimation can be useful in practice which provides an another decision moment for weed control action, but characteristics and applicability of it need further study.

This study has shown the potential benefit of using counter-intuitive features such as SIFT and SURF instead of colour, shape and texture for weed classification under natural daylight conditions.

7. Acknowledgements

The work presented in this paper was part of the Agrobot part of the SmartBot project and funded by Interreg IVa, European Fund for the Regional Development of the European Union and Product Board for Arable Farming. We thank Gerard Derks at experimental farm Unifarm of Wageningen University for arranging and managing the experimental fields.

References

Ahmed, F., Al-Mamun, H. A., Bari, A. S. M. H., Hossain, E., & Kwan, P. (2012). Classification of crops and weeds from digital images: A support vector machine approach. *Crop Protection*, *40*, 98–104.

Åstrand, B., & Baerveldt, A. J. (2002). An agricultural mobile robot with vision-based perception for mechanical weed control. *Autonomous Robots*, *13*(1), 21–35.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359.

Behmann, J., Mahlein, A. K., Rumpf, T., Römer, C., & Plümer, L. (2015). A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precision Agriculture*, 16(3), 239-260.

Bosch, A., Muñoz, X., & Martí, R. (2007). Which is the best way to organize/classify images by content? *Image and Vision Computing*, *25*(6), 778–791.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Camargo Neto, J., Meyer, G. E., Jones, D. D., & Samal, A. K. (2006). Plant species identification using Elliptic Fourier leaf shape analysis. *Computers and Electronics in Agriculture*, *50*(2), 121–134.

Chan, J. C. W., & Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, *112*(6), 2999–3011.

Cho, S. I., Lee, D. S., & Jeong, J. Y. (2002). Weed-plant Discrimination by Machine Vision and Artificial Neural Network. *Biosystems Engineering*, *83*(3), 275–280.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, (ECCV 2004)*, (pp. 1–22). Prague, Czech Republic: Springer-Verlag.

Fan, P., Men, A., Chen, M., & Yang, B. (2009). Color-SURF: A surf descriptor with local kernel color histograms. In *IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC 2009)* (pp. 726–730). Beijing, China: IEEE.

Fei-Fei, L., & Perona, P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)* (Vol. 2, pp. 524–531). San Diego, USA: IEEE

Gebhardt, S., & Kühbauch, W. (2007). A new algorithm for automatic Rumex obtusifolius detection in digital images using colour and texture features and the influence of image resolution. *Precision Agriculture*, *8*(1), 1–13.

Guerrero, J. M., Guijarro, M., Montalvo, M., Romeo, J., Emmi, L., Ribeiro, A., & Pajares, G. (2013). Automatic expert system based on images for accuracy crop row detection in maize fields. *Expert Systems with Applications*, *40*(2), 656–664.

Haug, S., Michaels, A., Biber, P., & Ostermann, J. (2014). Plant classification system for crop /weed discrimination without segmentation. In *IEEE Winter Conference on Applications of Computer Vision (WACV 2014)* (pp. 1142–1149). Steamboat Springs, USA: IEEE.

Hiremath, S., Van Evert, F. K., Ter Braak, C., Stein, A., & Van der Heijden, G. (2014). Image-based particle filtering for navigation in a semi-structured agricultural environment. *Biosystems Engineering*, *121*, 85–95.

Ishak, A. J., Hussain, A., & Mustafa, M. M. (2009). Weed image classification using Gabor wavelet and gradient field distribution. *Computers and Electronics in Agriculture*, *66*(1), 53–61.

Jeon, H. Y., Tian, L. F., & Zhu, H. (2011). Robust crop and weed segmentation under uncontrolled outdoor illumination. *Sensors*, *11*(6), 6270–6283.

Kato, H., & Harada, T. (2014). Image reconstruction from bag-of-visual-words. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2014)* (pp. 955–962). Columbus, USA: IEEE.

Kazmi, W., Garcia-Ruiz, F., Nielsen, J., Rasmussen, J., & Andersen, H. J. (2015). Exploiting affine invariant regions and leaf edge shapes for weed detection. *Computers and Electronics in Agriculture*, *118*, 290–299.

Khan, N. Y., McCane, B., & Wyvill, G. (2011). SIFT and SURF performance evaluation against various image deformations on benchmark dataset. In *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA 2011)* (pp. 501–506). Noosa Heads, Australia: IEEE.

Kise, M., Zhang, Q., Rovira Más, F., & Mas, F. R. (2005). A stereovision-based crop row detection method for tractor-automated guidance. *Biosystems Engineering*, *90*(4), 357–367.

Kounalakis, T., Triantafyllidis, G. A., & Nalpantidis, L. (2016). Weed recognition framework for robotic precision farming. In *Proceedings of 2016 IEEE International Conference on Imaging Systems and Techniques (IST 2016)* (pp.466–471). Chania, Greece: IEEE.

Law, M. T., Thome, N., & Cord, M. (2014). Fusion in Bag-of-words image representation: Key ideas and further insight (Chapter 2). In *Fusion in Comuter Vision: Understanding complex visual content* (pp. 29–52). Springer International Publishing.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, W. S., Alchanatis, V., Yang, C., Hirafuji, M., Moshou, D., & Li, C. (2010). Sensing technologies for precision specialty crop production. *Computers and Electronics in Agriculture*, *74*(1), 2–33.

Leemans, V., & Destain, M.-F. (2006). Application of the Hough Transform for Seed Row Localisation using Machine Vision. *Biosystems Engineering*, *94*(3), 325–336.

Li, G. Z. (2011). Machine learning for clinical data processing (Chapter 4.9). In *Machine Learning: Concepts, Methodologies, Tools and Applications* (pp. 876–878). IGI Global.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2/3*(December), 18–22.

Lin, H. T., Lin, C. J., & Weng, R. C. (2007). A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, *68*(3), 267–276.

Lottes, P., Hörferlin, M., Sander, S., & Stachniss, C. (2016). Effective Vision-based Classification for Separating Sugar Beets and Weeds for Precision Farming. *Journal of Field Robotics*, 1-19.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Muñoz-Huerta, R. F., Guevara-Gonzalez, R. G., Contreras-Medina, L. M., Torres-Pacheco, I., Prado-Olivarez, J., & Ocampo-Velazquez, R. V. (2013). A review of methods for sensing the nitrogen status in plants: advantages, disadvantages and recent advances. *Sensors*, *13*(8), 10823–10843.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML 2005)*, (pp. 625–632). Bonn, Germany: ACM.

Nieuwenhuizen, A. T. (2009). *Automated detection and control of volunteer potato plants*. PhD thesis, Wageningen University, The Netherlands.

Nieuwenhuizen, A. T., Hofstee, J. W., & Van Henten, E. J. (2010). Performance evaluation of an automated detection and control system for volunteer potatoes in sugar beet fields. *Biosystems Engineering*, *107*(1), 46–53.

Nieuwenhuizen, A. T., Tang, L., Hofstee, J. W., Müller, J., & Van Henten, E. J. (2007). Colour based detection of volunteer potatoes as weeds in sugar beet fields using machine vision. *Precision Agriculture*, *8*(6), 267–278.

Panchal, P. M., Panchal, S. R., & Shah, S. K. (2013). A Comparison of SIFT and SURF. *International Journal of Innovative Research in Computer and Communication Engineering*, *1*(2), 323–327.

Perez-Cruz, F., Martinez-Olmos, P., & Murillo-Fuentes, J. J. (2007). Accurate posterior probability estimates for channel equalization using gaussian processes for classification. In *Proceedings of the 8th Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2007)* (pp. 1–5). Helsinki, Finland: IEEE.

Pérez, A. J., López, F., Benlloch, J. V., & Christensen, S. (2000). Colour and shape analysis techniques for weed detection in cereal fields. *Computers and Electronics in Agriculture*, *25*(3), 197–212.

Persson, M., & Åstrand, B. (2008). Classification of crops and weeds extracted by active shape models. *Biosystems Engineering*, *100*(4), 484–497.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers* (pp. 61–74). MIT Press.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–45.

Rassem, T. H., & Khoo, B. E. (2011). Object class recognition using combination of color SIFT descriptors. In *Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST 2011)* (pp. 290–295). Batu Ferringhi, Malaysia: IEEE.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*(1), 93–104.

Romeo, J., Pajares, G., Montalvo, M., Guerrero, J. M., Guijarro, M., & Ribeiro, A. (2012). Crop row detection in maize fields inspired on the human visual perception. *The Scientific World Journal*, *2012*, 1-10.

Rumpf, T., Römer, C., Weis, M., Sökefeld, M., Gerhards, R., & Plümer, L. (2012). Sequential support vector machine classification for small-grain weed species discrimination with special regard to Cirsium arvense and Galium aparine. *Computers and Electronics in Agriculture, 80,* 89–96.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York, USA: McGraw-Hill, Inc.

Slaughter, D. C. C., Giles, D. K. K., & Downey, D. (2008). Autonomous robotic weed control systems: A review. *Computers and Electronics in Agriculture*, *61*(1), 63–78.

SmartBot. https://www.keep.eu/keep/project-ext/39219. (Accessed 20 November 2017).

Søgaard, H. T., & Olsen, H. J. (2003). Determination of crop rows by image analysis without segmentation. *Computers and Electronics in Agriculture, 38*(2), 141–158.

Sojodishijani, O., Ramli, A. R. R., Rostami, V., Samsudin, K., & Saripan, M. I. I. (2010). Just-in-time outdoor color discrimination using adaptive similarity-based classifier. *IEICE Electronics Express*, *7*(5), 339–345.

Suh, H. K., Hofstee, J. W., & Van Henten, E. J. (in press). Improved vegetation segmentation with ground shadow removal using an HDR camera. *Precision Agriculture*. https://doi.org/10.1007/s11119-017-9511-z.

Swain, K. C., Nørremark, M., Jørgensen, R. N., Midtiby, H. S., & Green, O. (2011). Weed identification using an automated active shape matching (AASM) technique. Biosystems Engineering, 110(4), 450–457.

Teixidó, M., Font, D., Pallejà, T., Tresanchez, M., Nogués, M., & Palacín, J. (2012). Definition of linear color models in the RGB vector color space to detect red peaches in orchard images taken under natural illumination. *Sensors*, *12*(6), 7701–7718.

Tsai, C. F. (2012). Bag-of-Words Representation in Image Annotation: A Review. *ISRN Artificial Intelligence*, *2012*, 1–19.

Uijlings, J. R. R., Smeulders, A. W. M., & Scha, R. J. H. (2009). What is the Spatial Extent of an Object? In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)* (pp. 770–777). Miami, USA: IEEE.

Van De Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2008). A Comparison of Color Features for Visual Concept Classification. In *Proceedings of the International Conference of*

Image and Video Retrieval (CIVR 2008) (pp. 141-150). Niagara Falls, Canada: ACM.

Van Evert, F. K., Polder, G., Van Der Heijden, G. W. A. M., Kempenaar, C., & Lotz, L. A. P. (2009). Real-time vision-based detection of Rumex obtusifolius in grassland. *Weed Research*, *49*(2), 164–174.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc.

Vedaldi, A., & Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM International Conference on Multimedia (MM 2010)* (pp. 1469-1472). Firenze, Italy: ACM.

Vollebregt, M. J. P. (2013). *Texture based discrimination between sugar beet and volunteer potato*. MSc thesis, Wageningen University.

Wilf, P., Zhang, S., Chikkerur, S., Little, S. A., Wing, S. L., & Serre, T. (2016). Computer vision cracks the leaf code. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(12), 3305–3310.

Wu, J., Cui, Z., Sheng, V. S., Zhao, P., Su, D., & Gong, S. (2013). A Comparative Study of SIFT and its Variants. *Measurement Science Review*, *13*(3), 122–131.

Yang, J., Jiang, Y. G., Hauptmann, A. G., & Ngo, C. W. (2007). Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR 2007)* (Vol. 63, pp. 197–206). Augsburg, Germany: ACM.

Zagoris, K., Pratikakis, I., Antonacopoulos, A., Gatos, B., & Papamarkos, N. (2014). Distinction between handwritten and machine-printed text based on the bag of visual words model. *Pattern Recognition*, *47*(3), 1051–1062.

Zhang, Z., Kodagoda, S., Ruiz, D., Katupitiya, J., & Dissanayake, G. (2010). Classification of Bidens in wheat farms. *International Journal of Computer Applications in Technology*, *39*, 123–129.

Zhou, L., Zhou, Z., & Hu, D. (2013). Scene classification using a multi-resolution bag-of-features model. *Pattern Recognition*, *46*(1), 424–433

Fig. 1. The robotic platform for volunteer potato control in a sugar beet field.

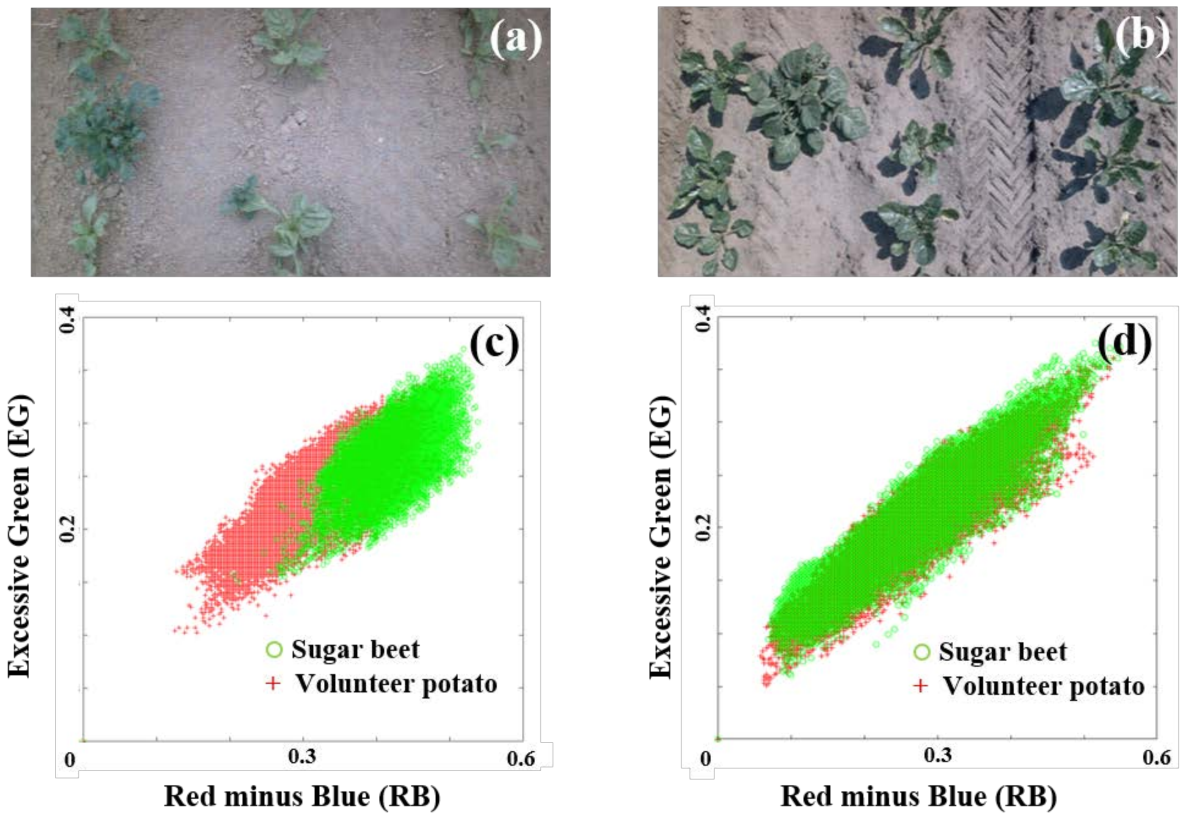

Fig. 2. In general, volunteer potato has a darker green colour than sugar beet (a). In such a case, sugar beet and volunteer potato are separable (based on the colour) in the EG-RB plane (c). An example case of volunteer potato having the same colour distribution as sugar beet (b). Sugar beet and volunteer potato are then visually inseparable in the EG-RB plane (d). To compare the colour difference between sugar beet and volunteer potato, the EGRBI transformation was used (Nieuwenhuizen et al., 2007).

Fig. 3. Example plant images in the field. The plant leaves often turn yellow in the summer as indicated in squares.

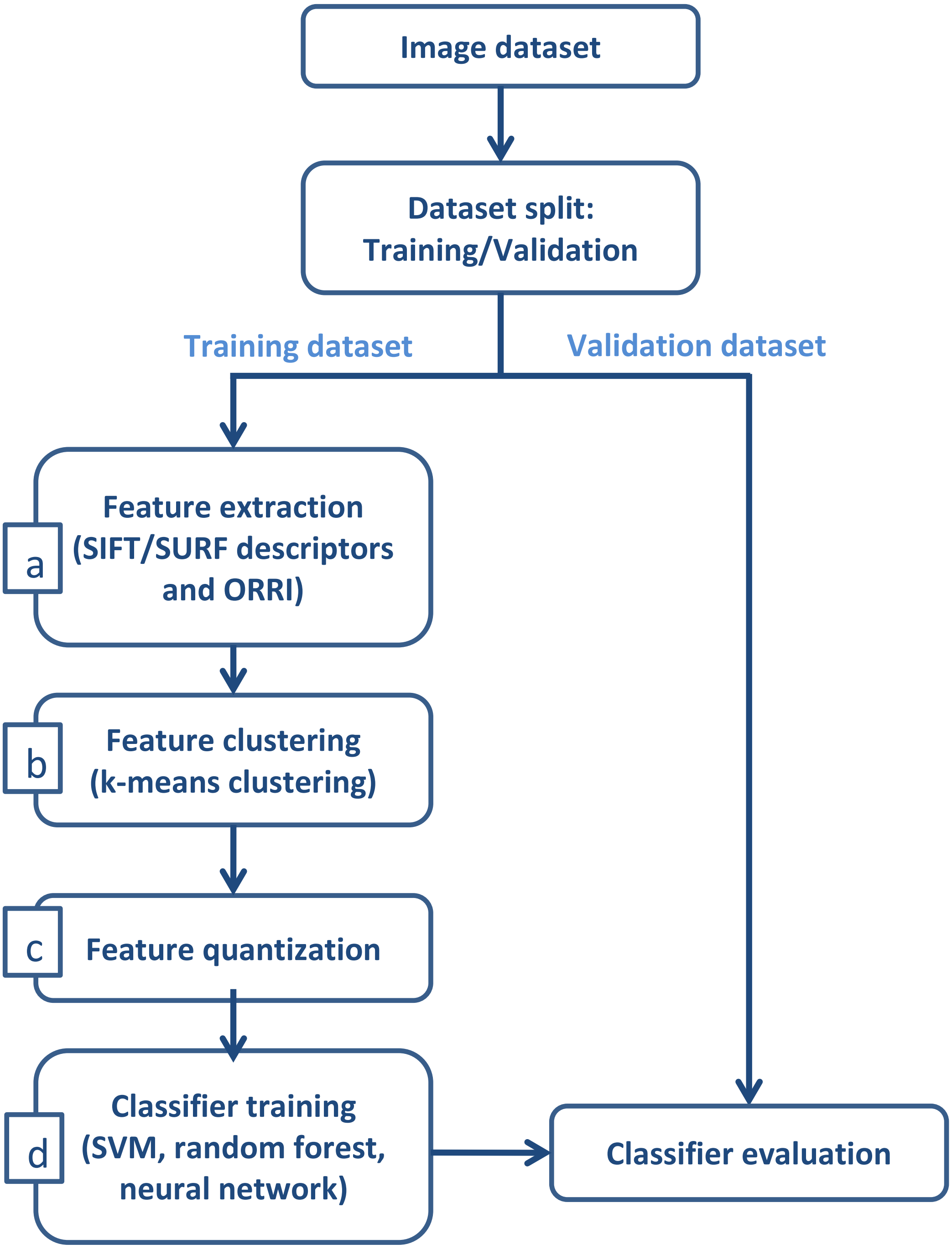

Fig. 4. Flowchart of image classification using Bag-of-Visual-Words.

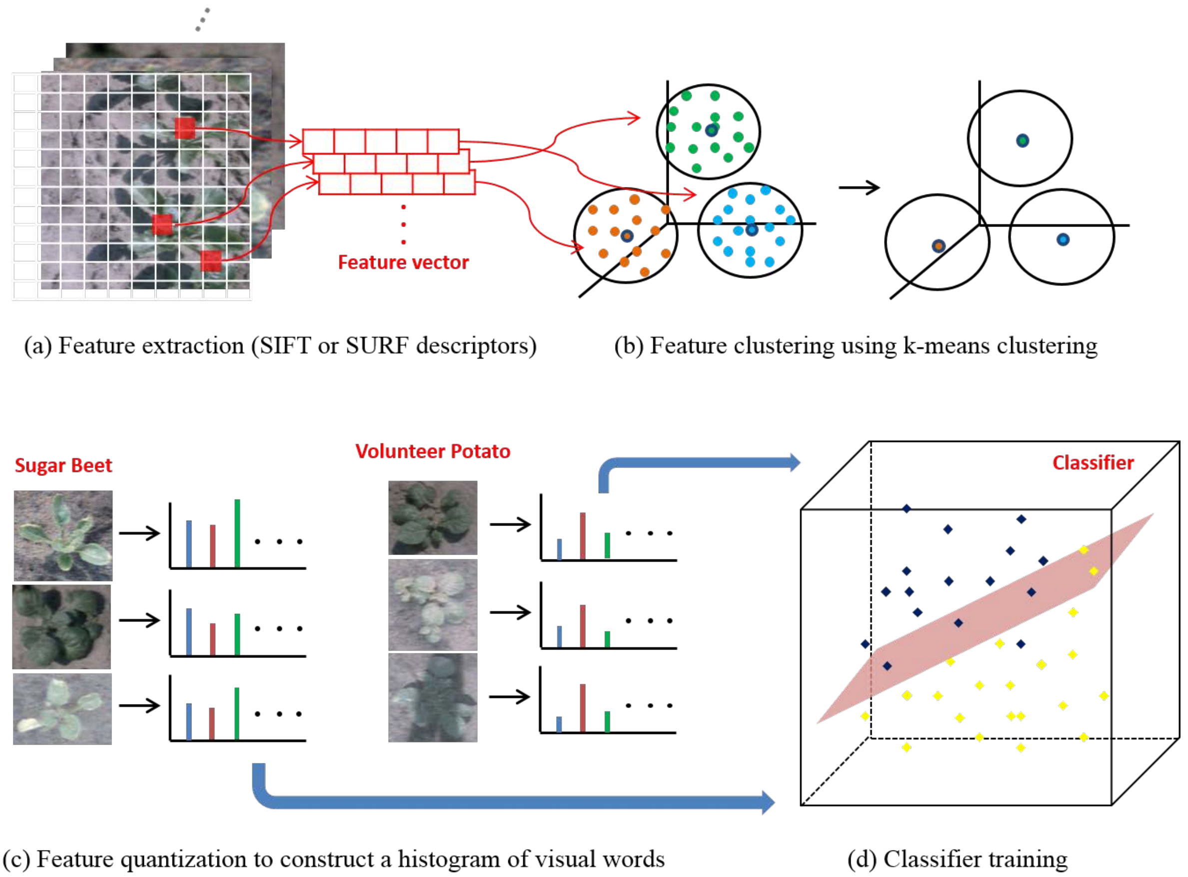

Fig. 5. Overview of BoVW model generation. SIFT or SURF features (local descriptors) were extracted from the training images (a). The extracted features were then clustered for visual vocabulary generation using k-means clustering (b). A histogram of visual words was constructed from each training image (c), which was used for classifier training (d).

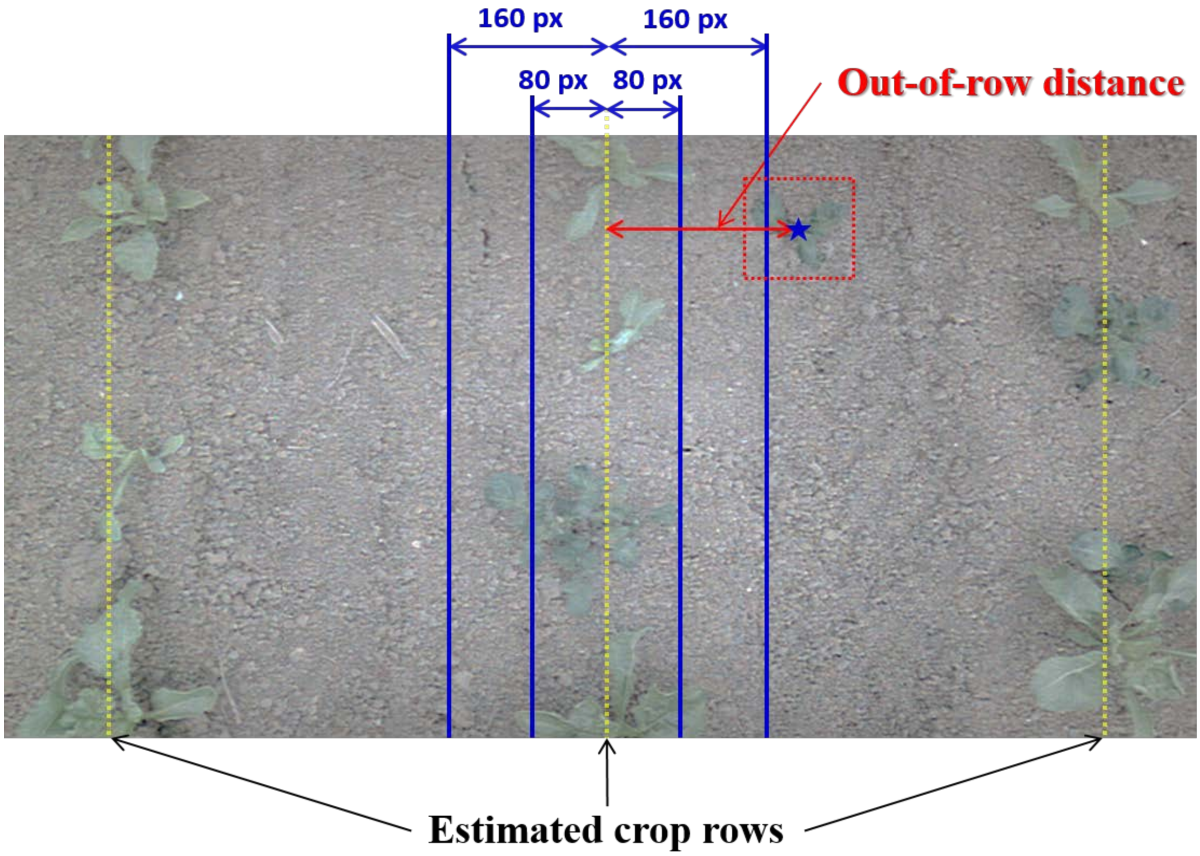

Fig. 6. The location of the three crop rows in the field of view was manually estimated (yellow dotted lines). An individual plant was extracted, then the distance between the centre position of a plant (marked as a star) to the nearest crop row, the out-of-row distance, was estimated. Two distances from the central crop row (80 and 160 pixels) are shown (blue lines).

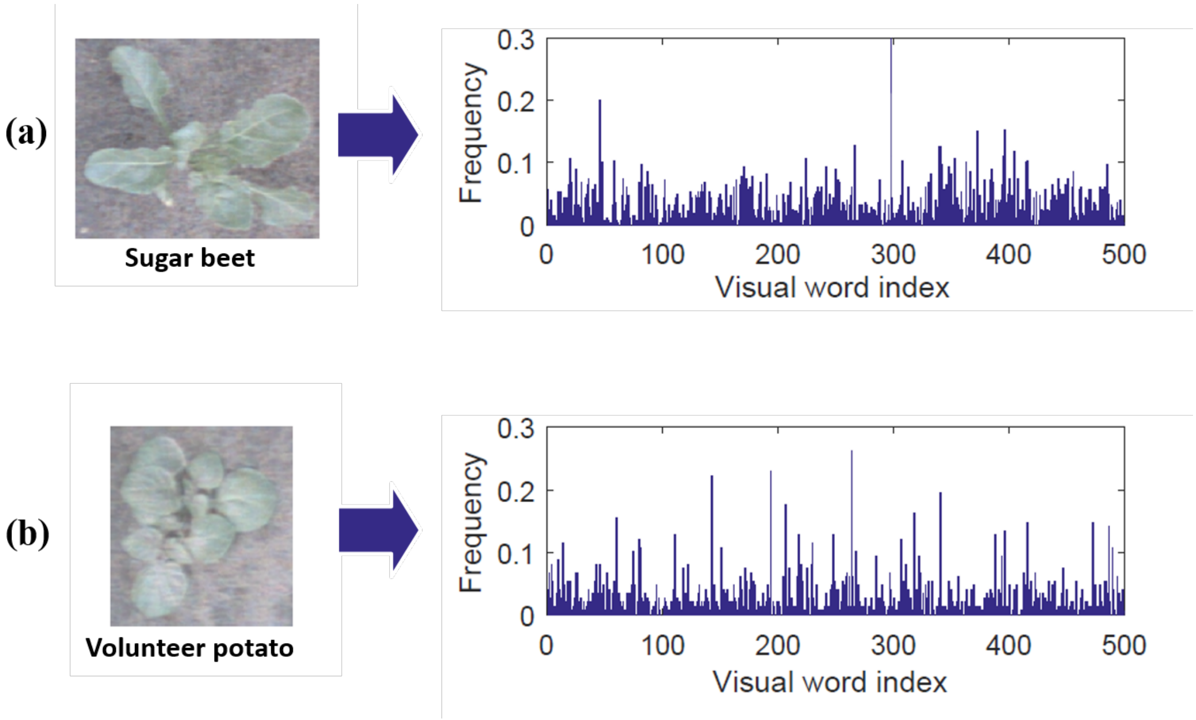

Fig. 7. Images of sugar beet (a) and volunteer potato (b) on the left, with the associated histograms of visual word occurrences on the right.

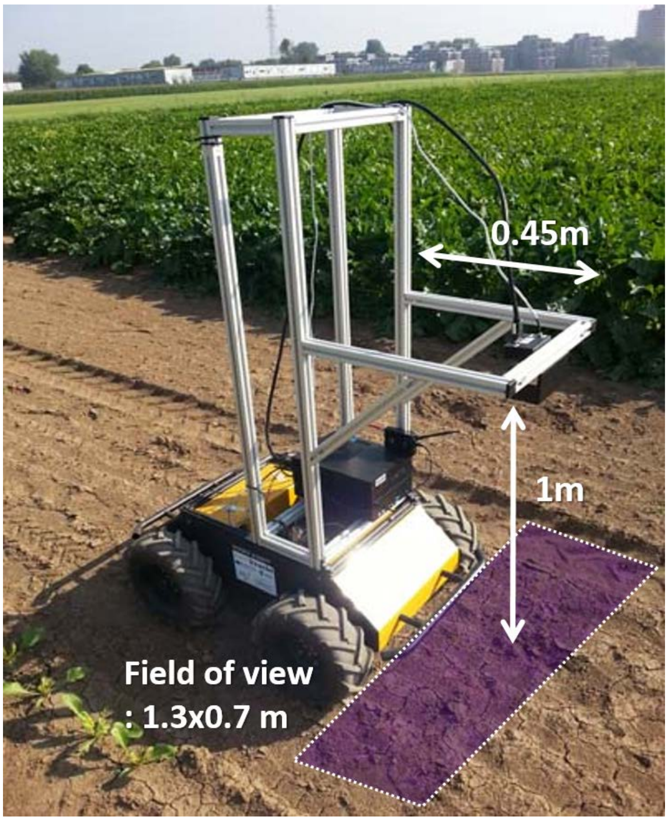

Fig. 8. Field images were acquired with a stereo camera mounted at the height of 1 m viewing perpendicular to the ground surface resulting in a field of view of 1.3 × 0.7 m. A mobile platform, Clearpath Husky, was manually controlled with a joystick and driven along crop rows using a controlled traveling speed of 0.5 m s^{-1}.

Fig. 9. Example images from the field image dataset containing a total of 400 plant images with 200 sugar beet (top) and 200 volunteer potatoes (bottom). During the generation of this dataset, images with different illumination levels were selected as well as images containing shadows.

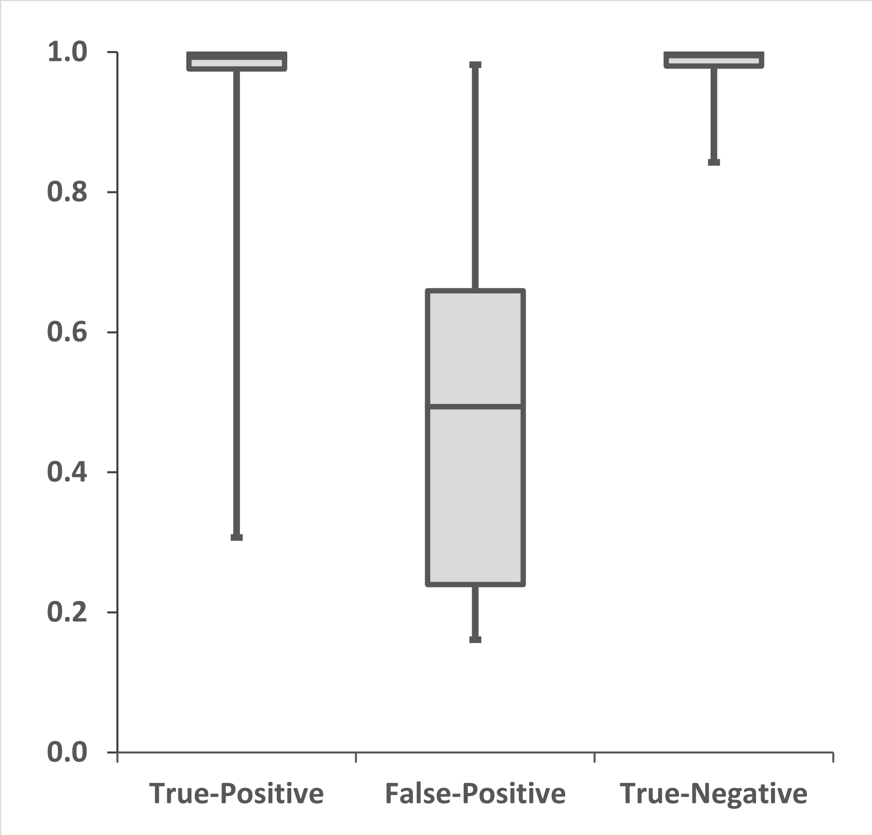

Fig. 10. A box-and-whisker plot of the estimated posterior probabilities of true-positive, false-positive and true-negative classifications using the SVM with linear kernel on SIFT features and ORRI. All sugar beet images were correctly classified as sugar beet (true-positive) with an average posterior probability of 0.96. A total of 180 volunteer potato images (out of 200) was correctly classified as volunteer potatoes (true-negative) with an average posterior probability of 0.98. However, 20 volunteer potato images were incorrectly classified as sugar beet (false-positive) with a Q1 (1st quartile), median and Q3 (3rd quartile) of 0.24, 0.49 and 0.66, respectively.

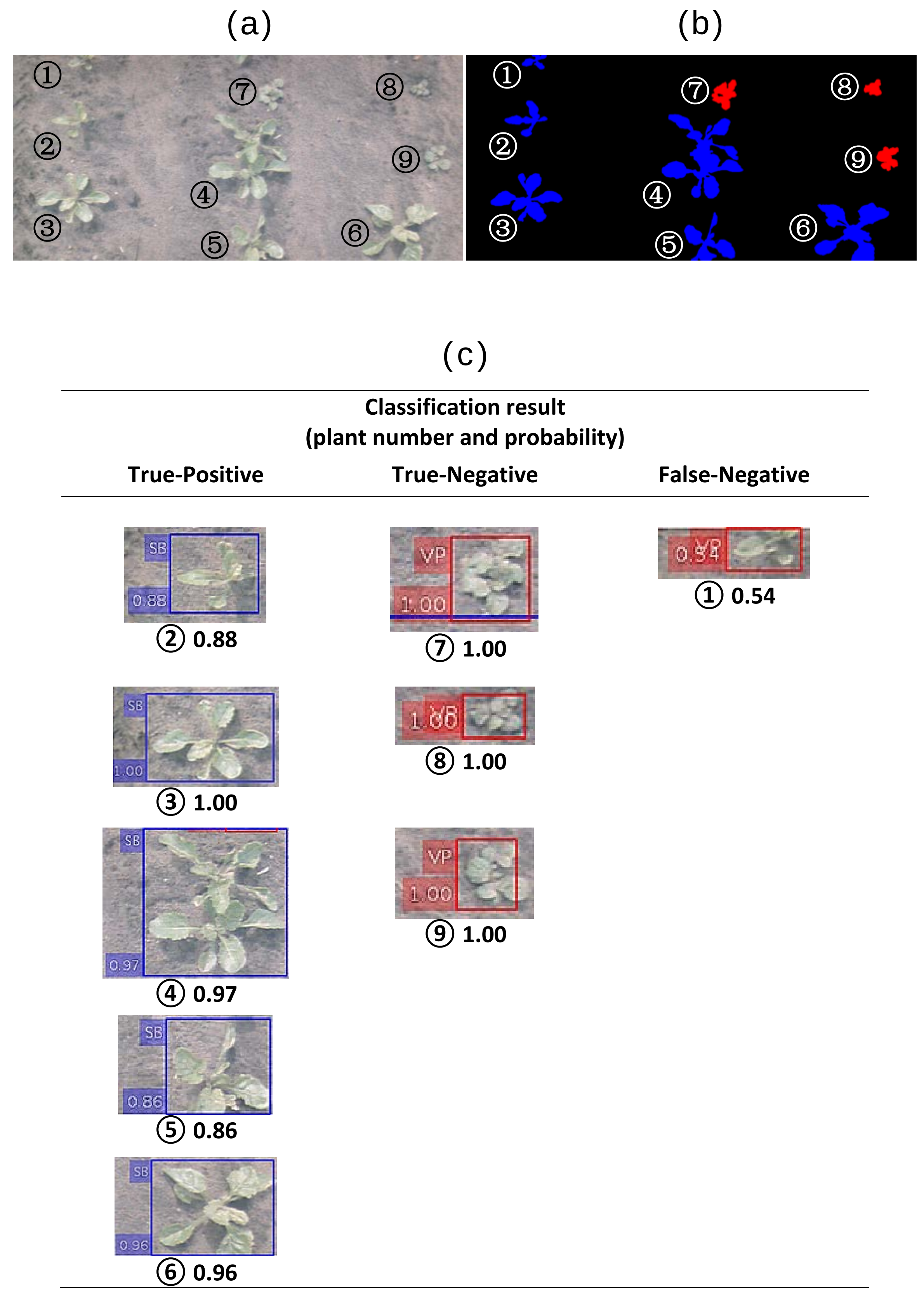

Fig. 11. An example of the classification results with posterior probability with a field image. (a) A field image with plant number. Each plant was manually extracted, and then put into the classification algorithm proposed in this study. (b) The ground truth of the given image. Plants 1 to 6 are sugar beet, and plants 7 to 9 are volunteer potatoes. (c) Classification results

with posterior probability. Plants 2 to 6 are correctly classified as sugar beet (true-positive) with a posterior probability of 0.86 and higher, and plants 7 to 9 are correctly classified as volunteer potatoes (true-negative) with a posterior probability of 1.0. However, plant 1 is incorrectly classified as a volunteer potato (false-negative) and results in a posterior probability of 0.54.

Table 1. Confusion matrix used for sugar beet and volunteer potato classification. (TP: true-positive, TN: true-negative, FP: false-positive, and FN: false-negative)

		Predicted Class	
		Sugar Beet (SB)	Volunteer Potato (VP)
Class	Sugar Beet (SB)	TP	FN
	Volunteer Potato (VP)	FP	TN

Table 2. The classification performance using SIFT features is shown. The classifiers were trained and validated with a total of 400 images (200 of sugar beet and 200 of volunteer potato) using 10-fold cross-validation. The final classification performance was averaged over ten repetitions. The training time includes the time for training of the classifier as well as for extracting SIFT features and building a visual vocabulary. The classification time includes the time required to classify the class of a single plant image using the trained classifier.

(TP: true-positive, TN: true-negative, FP: false-positive, and FN: false-negative)

Classifier models			TP	FN	FP	TN	Classification Accuracy (%)	Training time (s)	Classification time (s/image)
				(% of total)					
SVM	Linear	without ORRI*	183 (91.5)	17 (8.5)	20 (10)	180 (90)	90.8	218.6	0.107
		with ORRI	200 (100)	0 (0)	20 (10)	180 (90)	95.0	221.4	0.108
	Quad-ratic	without ORRI	186 (93)	14 (7)	17 (8.5)	183 (91.5)	92.3	216.6	0.106
		with ORRI	**200 (100)**	**0 (0)**	**14 (7)**	**186 (93)**	**96.5**	**218.8**	**0.107**
	Cubic	without ORRI	188 (94)	12 (6)	18 (9)	182 (91)	92.5	219.3	0.106
		with ORRI	196 (98)	4 (2)	17 (8.5)	183 (91.5)	94.8	222.6	0.106
Random Forest		without ORRI	172 (86)	28 (14)	38 (19)	162 (81)	83.5	228.9	0.109
		with ORRI	183 (91.5)	17 (8.5)	21 (10.5)	179 (89.5)	90.5	238.9	0.108
Neural Network		without ORRI	187 (93.5)	12 (6)	23 (11.5)	177 (88.5)	91.2	245.4	0.125
		with ORRI	195 (97.5)	5 (2.5)	12 (6)	188 (94)	95.8	260.5	0.130

* ORRI: Out-of-Row Regional Index

Table 3. The classification performance using SURF features is shown. The classifiers were trained and validated with a total of 400 images (200 of sugar beet and 200 of volunteer potato) using 10-fold cross-validation. The final classification performance was averaged over ten repetitions. The training time includes the time for training of the classifier as well as for extracting SIFT features and building a visual vocabulary. The classification time includes the time required to classify the class of a single plant image using the trained classifier.

(TP: true-positive, TN: true-negative, FP: false-positive, and FN: false-negative)

Classifier models			TP	FN	FP	TN	Classification Accuracy (%)	Training time (s)	Classification time (s/image)
				(% of total)					
SVM	**Linear**	without ORRI*	175 (87.5)	25 (12.5)	42 (21)	158 (79)	83.3	175.8	0.099
		with ORRI	**200 (100)**	**0 (0)**	**22 (11)**	**178 (89)**	**94.5**	**182.9**	**0.099**
	Quad-ratic	without ORRI	179 (89.5)	21 (10.5)	35 (17.5)	165 (82.5)	86.0	175.7	0.099
		with ORRI	**196 (98)**	**4 (2)**	**18 (9)**	**182 (91)**	**94.5**	**182.9**	**0.105**
	Cubic	without ORRI	176 (88)	24 (12)	29 (14.5)	171 (85.5)	86.8	175.7	0.099
		with ORRI	195 (97.5)	5 (2.5)	20 (10)	180 (90)	93.8	183.1	0.101
Random Forest		without ORRI	170 (85)	30 (15)	55 (27.5)	145 (72.5)	78.8	178.9	0.106
		with ORRI	179 (89.5)	21 (10.5)	42 (21)	159 (79.5)	84.5	186.2	0.104
Neural Network		without ORRI	165 (92.5)	35 (17.5)	27 (13.5)	173 (86.5)	84.5	195.1	0.115
		with ORRI	190 (95)	10 (5)	21 (10.5)	179 (89.5)	92.3	190.1	0.119

* ORRI: Out-of-Row Regional Index

Table 4. Comparison of classification accuracy (%) with different grid and vocabulary sizes. Using SURF descriptor, the classification accuracy of SVM linear with ORRI is shown.

		Vocabulary size					
		100	200	300	400	500	600
Grid size	4×4	92.5	93.2	91.2	93.7	93.8	92.7
(pixels)	6×6	91.7	93.0	93.5	92.5	**94.5**	92.7
	8×8	90.5	91.0	93.7	92.0	90.7	92.2
	10×10	91.2	92.7	93.6	93.2	92.7	92.7
	12×12	91.2	91.5	91.5	91.5	91.0	91.5