

Asdrúbal López-Chau¹(✉), Rafael Rojas-Hernández¹,
Farid García Lamont², Valentín Trujillo-Mora¹,
Lisbeth Rodríguez-Mazahua³, and Jair Cervantes²

¹ Universidad Autónoma del Estado de México,
Centro Universitario UAEM Zumpango, Zumpango, Mexico
{a.lchau, r.rojashe}@uaemex.mx

² Universidad Autónoma del Estado de México,
Centro Universitario UAEM Texcoco, Texcoco, Mexico

³ Division of Research and Postgraduate Studies,
Instituto Tecnológico de Orizaba, Orizaba, Mexico

Abstract. In most of classic plant identification methods a dichotomous or multi-access key is used to compare characteristics of leaves. Some questions about if the analyzed leaves are lobed, unlobed, simple or compound need to be answered to identify plants successfully. However, very little attention has been paid to make an automatic distinction of leaves using such features. In this paper we first explore if incorporating prior knowledge about leaves (categorizing between lobed simple leaves, and the unlobed simple ones) has an effect on the performance of six classification methods. According to the results of experiments with more than 1,900 images of leaves from Flavia data set, we found that it is statically significant the relationship between such categorization and the improvement of the performances of the classifiers tested. Therefore, we propose two novel methods to automatically differentiate between lobed simple leaves, and the unlobed simple ones. The proposals are invariant to rotation, and achieve correct prediction rates greater than 98%.

Keywords: Leaf features · McNemar test · Plant identification

1 Introduction

Plant identification is a challenging issue which has aroused researchers' attention in recent years. Classic plant identification methods are based on observing specific features of leaves to categorize leaves. However, in the literature very little attention has been paid to make an automatic distinction between different types of leaves to improve plant identification. This difference is important because most of the classic methods for plant identification use dichotomous or multi-access keys that compare characteristics of the leaves, asking if they are lobed, unlobed, simple or compound, among others features.

In this paper, we first analyze if the relationship between the knowledge about the type of leaf (unlobed simple or lobed simple) and the classification accuracy of

classification methods is significant statistically. As a first approach to explore the relationship, we carried out the following experiment: we extract basic leaf features, and use them with standard classification methods. Then, we add the type of leaf as a feature and test again the same methods. In both cases, the classification accuracies were measured, and then compared applying the McNemar test. According to the results of the experiments, using the type of leaf as a binary feature has a positive impact on the performance of the methods tested, such impact is statistically significant. To be fair, only basic features were extracted from leaves because we are interested in observing the effect of another basic binary leaf feature. Therefore, we propose two novel methods to categorize leaves. The first method presented uses concentric circles to detect the changes of color. The second method uses convex hulls. These methods do not vary neither to scale nor to rotation.

The rest of the paper is organized as follows. Section 2 describes basic leaf features for plant identification and gives a brief review about main works related to extraction of leaf features. Section 3 shows the results of the exploration on the effect of incorporate previous knowledge about the type of leaf to classification methods for plant identification. We present two methods to identify lobed simple leaves in Sect. 4, then Sect. 5 shows experiments and results. Finally, last section of this paper presents conclusions and future works.

2 Related Works

Leaf features are extracted from images previously processed. Then, leaf features are encoded as a set of numbers (vectors) or nominal values, also known as feature descriptors.

Most of leaf features can be categorized into the following six main types [1]. *Geometric*: defined as sets of points that form points, lines, etc.; *Morphological*: related to form and structure of a leaf; *Texture*: these descriptors characterize image textures or regions; *Color*: based on RGB image and its 3 channels; *Shape*: contour of leaves has to be taken into account to describe the structure. Current work is mostly focused on this type of descriptors; *Vein network*: leaf veins are analyzed to extract specific characteristics; *Others*: image descriptors borrowed from computer vision to describe leaves, such as Fourier descriptors, SIFT, and border detectors or filters, for example Gabor. Shape is the most popular feature in literature on plant identification [2], among the six types of descriptors explained.

In [3], authors propose “shape-defining feature” (SDF), by using slopes and distances between two consecutive points. The shape of a leaf along with its fine serrations is retrieved using this method. In order to compute SDF, they draw a total of 400 lines (vertical and horizontal) over the image of a leaf, and then detect the endpoints of these lines. The larger the number of lines, finer is the detail of serrations. For the classification authors use a Neural Network along with AR, CH, Ec and Roundness. A drawback of the method presented in [3] is that the number of features is large (800 features per leaf), compared with the number of images per leaf in data sets.

Shape context (SC) descriptor, proposed in 2000 by Belongie and Malik [4], is used to compute shape correspondences and similarities between two images. Based on SC,

Zhi et al. [5] proposed “Arc Length Shape Contexts” (ARC-SC). This descriptor is composed of two parts: the sum of Euclidean distances between adjacent points, and the angle between two pixels on the silhouette of leaf. Minimum cost of matches between all ARC-SC of a leaf and the extracted from training set is computed to identify a plant.

Other descriptor that uses leaf shape (specifically, points on the border of the leaf) is Multi scale Distance Matrix (MDM), introduced by Hu et al. [6]. The first step to build MDM is to create a symmetric matrix D in $R^{n \times n}$, whose entry $d_{i,j}$ is the distance between points x_i and x_j , both on the border of the leaf. Then, dimensionality reduction is applied, retaining only unrepeated elements in D .

MDM descriptor is invariant to rotation, scaling and translation; however, to apply MDM, the shape of leaf must be stable, i.e., without noise. Different from the previous methods, Gwo et al. [7] do not use all points on the border, but retain only few ones, compared to other methods. The selection of feature points is realized by comparing distances.

Several methods use distances from a reference points to the border of leaf. Hajjdiab and Al Maskari [8] use the centroid of image as reference. They take 32 points chosen circularly, at equally spaced angles. Shen et al. [9] compute a centroid considering only the points located on the border of the leaf. Then, they subsample the border, obtaining 36 points. This type of methods require the detection of the silhouette of the leaf from clean images.

Kala et al. [10] use the border of leaf in a different way to other works. They compute a sinuosity measure, which expresses the meandering of a curve. An issue of the sinuosity measure is that it requires the silhouette of the leaf to be differentiable and this measure is not rotation invariant.

Texture of leaves has also been used to identify plants. In [11, 12], authors combined shape and texture of a leaf to identify plants. For shape analysis, Beghin et al. [11] extract the contour signature from leaf, and then compute the dissimilarities between all leaves in data set using the Jeffrey-divergence measure. Meanwhile, Chaki et al. [12] apply curvelet transform coefficients together with invariant moments. The method for texture feature extraction presented in [11] uses Sobel directions histogram. Chaki et al. [12] use Gabor filter (GF) and gray level co-occurrence matrix (GLCM).

In [13], authors propose a combination of morphological and geometric features of a leaf. They remove irrelevant features using a fuzzy surface selection method. Few remaining features are used with a Neural Network. Classification is performed quickly by using this simple scheme. However, extraction of features is computationally costly. That method was tested with only four species of plants, all of them have simple leaves.

One of the least used features for plant identification is color. Most of methods for the same purpose work with binary images. de M. Sá Junior et al. [14] use a gravitational approach, which produces success rate above 90%; however, their method requires a manual selection of texture windows and orientation of leaf. A general problem with color features, is that many factors have to be taken into account, for example, illumination conditions, maturity of plant, diseases and environment [14, 15].

3 Studying the Effect of Adding Prior Knowledge About the Type of Leaf on Classification

To explore if the prior knowledge (type of leaf) has an effect of the performance of classification methods, we executed the statistical test of McNemar. It tests consistency in responses across two variables. McNemar test recognizes that some instances will move from incorrectly predicted to correctly predicted and others from correctly predicted to incorrectly predicted just randomly. If the prior knowledge is having no effect on performance of a classification method, the number of instances which move from incorrectly predicted to correctly predicted should be about equal to those who move in the other direction.

Six basic leaf features were extracted from images leaves. We first built and tested six different classifiers with these characteristics of leaves. Then, we manually added a binary leaf feature, assigning a value of true for lobed simple leaves with smooth margins, and a value of false for the rest of the leaves. The six classifiers were trained and tested again. The classification accuracies were measured in both cases. The number of instances correctly/incorrectly predicted before and after adding the binary leaf features were counted to create the contingency tables.

3.1 Materials

One of the most widely used data set for testing plant identification systems is Flavia. It is publicly available at <http://flavia.sourceforge.net>, this set contains 1,907 color images of 32 different species of plants. These images have a dimension of $1,600 \times 1,200$ pixels.

In general, leaves can be classified according to their blade (simple or compound), edge (smooth, dentate, etc.), petiole (petiolated or sessile), shape of blade, etc. Among these categories, simple, compound, unlobed and lobed are very common in dichotomous keys. For simple leaves, the leaf blade is a single, continuous unit. For compound leaves the blade is divided into two or more leaflets arising from the petiole. Figure 1 shows an example of simple leaf and compound leaf. In this case, it is really easy to categorize these leaves. However, in many other cases this categorization it is really complicated. This is because there are many subtypes of leaves. For example, simple leaves can be unlobed or lobed. For unlobed leaves, the blade is completely undivided. Lobed leaves have projections off the midrib with individual inside veins. Figure 2 shows two examples of simple leaves which are very different from the simple leaf presented in Fig. 1.

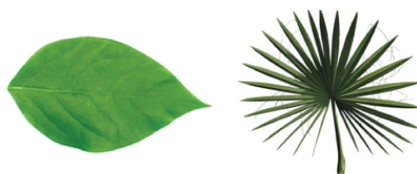


Fig. 1. Example of a simple leaf (left), and a compound leaf (right).

Table 2. Effect prior knowledge (type of leaf as a binary attribute) on six classification methods.

Classification method	Classification accuracy (%)	Classification accuracy (%)	p-value
	<i>Data set 1</i>	<i>Data set 2</i>	
Decision tree C4.5	58.78	60.15	0.0305
KNN (k = 1)	64.13	64.24	0.0478
Random forest	65.44	66.60	0.0216
Multiclass classifier	68.38	71.95	0.0025
Naïve Bayes	55.06	56.58	0.0296
Random tree	56.63	57.63	0.0380

features. This number is lesser and simpler than the used in many other works [12]. Our goal is to compare basic leaf features with the type of leaf, as we consider it a basic leaf feature too.

In order to validate if the improvement in the performance of classifiers is statistically significant, we apply the McNemar test. The p-values achieved are shown in the last column of Table 2. Although the improvement of performances is slight, the p-values values suggest that the difference of frequencies observed in instances correctly classified before and after adding the attribute is not due to randomness. Based on these results, in next Section we propose two methods to detect lobed simple leaves.

4 Proposed Methods to Categorize Leaves

4.1 Method Based on Concentric Circles

The first method that we propose in this paper utilizes concentric circles. By using a preprocessed binary image L of a leaf, we detect changes of color (black to white) along a curve that crosses the image of a leaf. These changes are produced by the leaf or by its leaflets.

First, we create a number of concentric circles over the leaf. Figure 3 shows how to compute the center and the radio of those circles. Then, we count color changes on the trajectory of each circle. To avoid counting noisy pixels, we only consider it a color change when there are at least K pixels of the same color once color variation has been rendered. We empirically determined that a value of $K = 10$ works for most cases.

As a result, we obtained a vector V with A components (A is the number of concentric circles). Each component of V comprises the number of changes of color (from black to white) minus one.

Figure 4 shows an example of two leaves with eight concentric circles over them. One of the leaves is unlobed simple and the other one is lobed simple. For the first leaf $V = [0\ 0\ 0\ 1\ 1\ 1\ 1\ 1]$, for the second leaf $V = [0\ 0\ 0\ 0\ 3\ 3\ 3\ 5]$. Value 0 means that corresponding circle only touches the foreground (black), and never crosses through the background (white), i.e. there are not changes of color along the contour of the circle. The greater the value of a component V , the greater the number of times the contour of the corresponding circle detects changes from black to white.

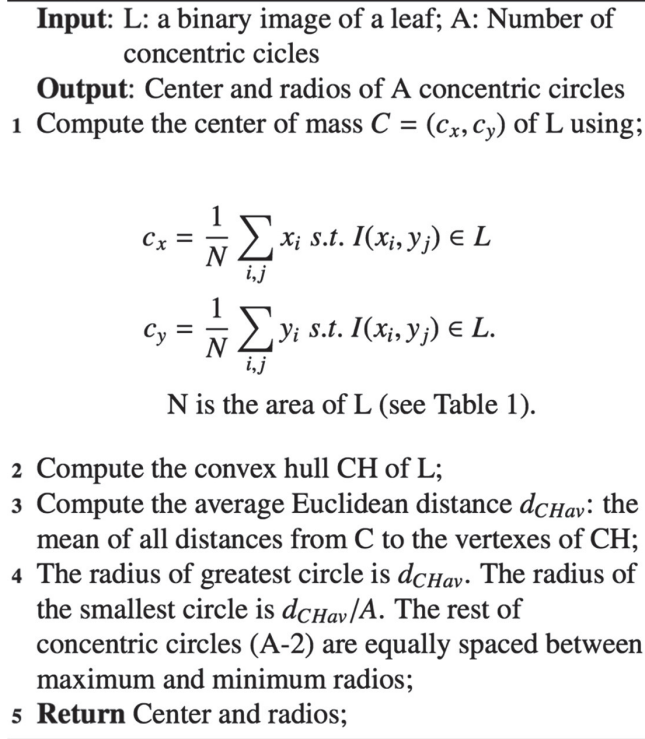


Fig. 3. Pseudocode of algorithm to generate concentric circles on a leaf.

In order to discriminate lobed simple leaves from unlobed simple leaves, we add the number of components of V with a value greater than two, this sum is named S. Value two is based on the observation that lobed leaves have, in general, at least three leaflets, which produce three color changes. Then the criterion shown in Eq. (1) is applied:

$$Lobed = \begin{cases} \text{true} & \text{if } S \geq 2 \\ \text{false} & \text{otherwise} \end{cases} \quad (1)$$

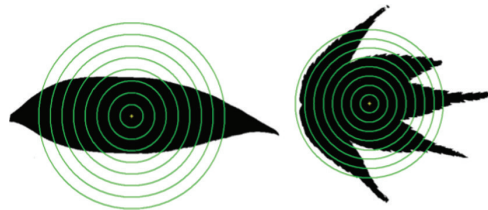


Fig. 4. Concentric circles generated on unlobed simple leaf (left), and lobed simple leaf (right).

The center of mass $C(c_x, c_y)$, convex hull CH and the average distance d_{CHav} are all of them invariant to the orientation of the image. Any circle with radius R centered in C will pass exactly through the same points even if the leaf is rotated, therefore the algorithm proposed in this subsection is invariant to rotation.

4.2 Method Based on Convex Hulls

According to the results presented in Sect. 3.2, incorporation of prior knowledge about the type of leaf (as a feature) improves the performance of classification methods for plant identification. Therefore, a second method to identify lobed simple and unlobed simple leaves was developed based on the concept of convex hull.

Given a binary image L of a leaf, our second method computes the difference between the convex hull of L and the binary image L. Thereafter, the connected components in the resulting image are identified, and the area (number of pixels) of each component is computed. The underlying idea is that the area of the connected components in lobed simple leaves is greater than the areas of the connected components in unlobed simple leaves; an example of this is shown in Fig. 5. The procedure that implements this part of our method is presented in Fig. 6.



Fig. 5. Examples of the difference (white areas) of convex hull and leaf.

Input: Image of a leaf L
Output: Areas of connected components

- 1 $A_L \leftarrow$ Obtain area of L
- 2 $C \leftarrow$ extract the contour of L
- 3 $CH \leftarrow$ Compute the convex hull of C
- 4 $A_{CH} \leftarrow$ Obtain area of CH
- 5 $R_L \leftarrow$ Compute $A_{CH} - A_L$
- 6 $C_c \leftarrow$ Detect connected components of R
- 7 $H \leftarrow \emptyset$
- 8 **for** $c \in C_c$ **do**
- 9 $H \leftarrow H \cup$ Area of c
- 10 **return** H

Fig. 6. Pseudocode of algorithm to compute residual areas.

In general, the set of areas computed by applying the algorithm shown in Fig. 6 has different cardinality for each image. Therefore, we only retain the greatest ten areas. This produces a numeric vector in \mathbb{R}^{10} , which is used to train a Random Forest Classifier. The class of each instance is the label that we manually set for lobed leaves, and that is explained in Subsect. 3.2.

In next Section, we present the results of experiments to measure the performance of our two methods.

5 Evaluation of the Proposed Methods

We measured the performance of our proposals to identify lobed simple leaves. Both methods were tested with images of Flavia data set. In all our experiments, we did not rotate or scale any image.

Because in the literature there are not features specifically designed to identify lobed leaves, we do not compare the obtained results with others methods. Instead, we measure accuracy, specificity and sensitivity of the two introduced methods.

Henceforth, the method based on circles will be referred as M_{Circ} , whereas the method based on convex hull will be referred as M_{ConvexH} .

In order to measure the performance of M_{Circ} and M_{ConvexH} , we use the whole Flavia data set. The confusion matrices obtained are presented in Tables 3 and 4. The positive cases correspond to lobed simple leaves, whereas the negative cases are the unlobed simple ones. Based on these matrices, the following measures are obtained:

Table 3. Confusion matrix for the method based on circles M_{Circ} .

Real class	Prediction	
	Lobed = false	Lobed = true
Lobed = false	1,629	64
Lobed = true	6	208

Table 4. Confusion matrix for the method based on convex hulls M_{ConvexH} .

Real class	Prediction	
	Lobed = false	Lobed = true
Lobed = false	1,672	21
Lobed = true	0	214

- **Accuracy:** the proportion of the total number of predictions (positive and negative) that were correct.
- **Sensitivity or Recall:** the proportion of actual lobed simple leaves which are correctly identified.
- **Specificity:** the proportion of actual unlobed simple leaves which are correctly identified.

It can be observed in Tables 3 and 4 that most of the prediction errors are committed in actual unlobed simple leaves, which are incorrectly identified as lobed ones. The method M_{Circ} produces some errors in the identification of lobed simple leaves, whereas M_{ConvexH} does not commit this error in this type of leaves.

Comparing the performances of both methods (Table 5), it is possible to claim proposal with the highest performance is M_{ConvexH} . One more time, we tested the performances of the six classification methods using the outcomes of M_{ConvexH} . Table 6 shows the classification accuracies obtained. The performances are quite similar – although slightly lower - to those show in Table 2. This is could be due to the method M_{ConvexH} does not identify the 100% of leaves correctly.

Table 5. Performances of proposed methods.

Real class	Accuracy (%)	Recall	Specificity
M_{Circ}	96.33	0.9720	0.9963
M_{ConvexH}	98.89	1.000	0.9876

Table 6. Performance of classification method using M_{ConvexH}

Classification method	Classification accuracy (%) <i>Data set 2</i>
Decision tree C4.5	60.01
KNN (k = 1)	64.14
Random forest	66.61
Multiclass classifier	71.87
Naïve Bayes	56.45
Random tree	57.61

It can be seen in Tables 2 and 6 that the method with best performance is Multiclass classifier. It transforms a multiclass problem into several two-class ones, each one of these problems is solved with logistic regression. The second best method is Random forest, which uses a number of decision trees to solve the multiclass problem. These two classification methods are more suitable for plant identification with our methods.

6 Conclusions and Future Work

Many classic plant identification methods use dichotomous keys that take into account specific features of leaves, such as lobed, unlobed, simple or compound. However, state-of-the-art methods are not oriented to detect these leaf features. In this paper, we firstly explore if adding the type of leaf (distinguishing between lobed simple leaves and unlobed simple ones) as a basic binary feature can improve the performance of six classification methods. We found that incorporating this previous knowledge is beneficial for classifier, although the improvement is slight.

Motivated by the results obtained, we designed two new methods to discriminate automatically between unlobed simple and lobed simple leaves. The first method detects changes from black-to-white (and vice versa) in binary images. The second method uses the differences of areas between convex hull and the leaf.

Both methods were tested with color images from Flavia data set. The correct prediction rate is above 96% for the method based on circles, and greater than 98% for the method based on convex hull. These methods are invariant to rotation of images.

Adding the prior knowledge about the type of the leaf for creating a complete plant identification system is out of the scope of the proposals presented in this paper; however, preliminary results of experiments with Flavia data set have shown our methods can help to improve the performance of such type of systems, although at this moment the improvement achieved is not statistically significant yet. This is due to our methods only discriminate between unlobed simple and lobed simple leaves, the proportion between these types of leaves is 8:1 in Flavia data set. The number of lobed leaves is very small compared to the number of unlobed leaves, Currently, we are working in an improved version of our methods, to discriminate between more types of leaves.

References

1. Pahikkala, T., Kari, K., Mattila, H.: Classification of plant species from images of overlapping leaves. *Comput. Electron. Agric.* **118**, 186–192 (2015). doi:[10.1016/j.compag.2015.09.003](https://doi.org/10.1016/j.compag.2015.09.003)
2. Jamil, N., Hussin, N.A.C., Nordin, S., Awang, K.: Automatic plant identification: is shape the key feature? *Procedia Comput. Sci.* **76**, 436–442 (2015). doi:[10.1016/j.procs.2015.12.287](https://doi.org/10.1016/j.procs.2015.12.287)
3. Aakif, A., Khan, M.F.: Automatic classification of plants based on their leaves. *Biosyst. Eng.* **139**, 66–75 (2015). doi:[10.1016/j.biosystemseng.2015.08.003](https://doi.org/10.1016/j.biosystemseng.2015.08.003)
4. Belongie, S., Malik, J.: Matching with shape contexts. In: *IEEE Work on Proceedings of the Content-Based Access Image Video Libraries*, pp. 20–26 (2000)
5. Zhi, Z.-D., Hu, R.-X., Wang, X.-F.: A new weighted ARC-SC approach for leaf image recognition. In: Huang, D.-S., Ma, J., Jo, K.-H., Gromiha, M.M. (eds.) *ICIC 2012. LNCS*, vol. 7390, pp. 503–509. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-31576-3_64](https://doi.org/10.1007/978-3-642-31576-3_64)
6. Hu, R., Jia, W., Ling, H., Huang, D.: Multiscale distance matrix for fast plant leaf recognition. *IEEE Trans. Image Process.* **21**, 4667–4672 (2012)
7. Gwo, C.Y., Wei, C.H., Li, Y.: Rotary matching of edge features for leaf recognition. *Comput. Electron. Agric.* **91**, 124–134 (2013). doi:[10.1016/j.compag.2012.12.005](https://doi.org/10.1016/j.compag.2012.12.005)
8. Hajjdiab, H., Al Maskari, I.: Plant species recognition using leaf contours. In: *2011 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 306–309 (2011)
9. Shen, Y., Zhou, C., Lin, K.: Leaf image retrieval using a shape based method. In: Li, D., Wang, B. (eds.) *AIAI 2005. ITIFIP*, vol. 187, pp. 711–719. Springer, Boston (2005). doi:[10.1007/0-387-29295-0_77](https://doi.org/10.1007/0-387-29295-0_77)
10. Kala, J.R., Viriri, S., Moodley, D.: Sinuosity coefficients for leaf shape characterisation. In: Pillay, N., Engelbrecht, A.P., Abraham, A., du Plessis, M.C., Snášel, V., Muda, A.K. (eds.) *Advances in Nature and Biologically Inspired Computing. AISC*, vol. 419, pp. 141–150. Springer, Cham (2016). doi:[10.1007/978-3-319-27400-3_13](https://doi.org/10.1007/978-3-319-27400-3_13)

11. Beghin, T., Cope, J.S., Remagnino, P., Barman, S.: Shape and texture based plant leaf classification. In: Blanc-Talon, J., Bone, D., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2010. LNCS, vol. 6475, pp. 345–353. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-17691-3_32](https://doi.org/10.1007/978-3-642-17691-3_32)
12. Chaki, J., Parekh, R., Bhattacharya, S.: Plant leaf recognition using texture and shape features with neural classifiers. *Pattern Recogn. Lett.* **58**, 61–68 (2015). doi:[10.1016/j.patrec.2015.02.010](https://doi.org/10.1016/j.patrec.2015.02.010)
13. Tzionas, P., Papadakis, S.E., Manolakis, D.: Plant leaves classification based on morphological features and a fuzzy surface selection technique. In: Fifth International Conference on Technology and Automation, Thessaloniki, Greece, pp. 365–370 (2005)
14. de M. Sá Junior, J.J., Backes, A.R., Cortez, P.C.: Plant leaf classification using color on a gravitational approach. In: Wilson, R., Hancock, E., Bors, A., Smith, W. (eds.) CAIP 2013. LNCS, vol. 8048, pp. 258–265. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-40246-3_32](https://doi.org/10.1007/978-3-642-40246-3_32)
15. McCarthy, C.L., Hancock, N.H., Raine, S.R.: Applied machine vision of plants: a review with implications for field deployment in automated farming operations. *Intell. Serv. Robot.* **3**, 209–217 (2010). doi:[10.1007/s11370-010-0075-2](https://doi.org/10.1007/s11370-010-0075-2)