Effects of Different Mixtures of Features, Colours and SVM Kernels on Wheat Disease Classification

Punnarai Siricharoen, Bryan Scotney, Philip Morrow and Gerard Parr
School of Computing and Information Engineering
University of Ulster, Coleraine, BT52 1SA
siricharoen-p@email.ulster.ac.uk

David Gibson, Nishan Canagarajah
Department of Computer Science
University of Bristol, Bristol, BS8 5UB
gibson@cs.bris.ac.uk

Abstract

This paper assesses how the combination of different features, colour models and SVM kernels affect the classification performance of wheat disease identification. The basic approach consists of pre-processing, feature extraction, and classification. Five colour models (greyscale, RGB, HSV, YCbCr and L*a*b*), four different feature sets (Haralick, Tamura, First-order statistics and HOG features) and three kernels for a support vector machine (linear, RBF and polynomial) are assessed in terms of overall performance accuracy. Image datasets including non-diseased, Yellow Rust diseased and Septoria diseased leaves have been acquired under controlled conditions. The results show that homogeneity and contrast or energy features combined with basic statistical information such as mean, skewness and kurtosis, and visually perceptual features consisting of directionality, contrast and coarseness, which are extracted from YCbCr images using a classification model based on a linear kernel, produce the highest classification accuracy with low computational complexity

Keywords: Plant Disease Classification, Feature Extraction, Combination of Features, SVM Kernel

1 Introduction

Crop diseases can lead to a substantial decrease in both quantity and quality of agricultural products worldwide. To reduce such losses, early notification or continuous monitoring of crops is required. However, it is expensive, time-consuming and labour-intensive for experts to accurately diagnose the symptoms appearing on a plant, especially in remote areas.

When plants are infected, they can exhibit a range of symptoms, for example, colour spots or colour bands on the leaves, fruit, stems, or seeds. Recently, various image processing techniques have been developed for automated disease detection. Automated systems for plant disease identification have played a role in agriculture not only for rapid detection but also reducing human error. However, to apply an automated classification system to real plant diseases, robust imaging methods are required. The aim of this work is to develop an automated plant disease identification system using image processing. We briefly discuss previous literature on plant disease classification in Section 2, give the details of our methodology and proposed classification system in Section 3, and describe experimentation and evaluation in Section 4. Finally, conclusions are given in Section 5.

2 Literature Review

General automated systems have four main components: pre-processing, segmentation, feature extraction, and classification. Firstly, pre-processing techniques are applied to handle data differences arising from different lighting conditions, or capture devices. The methods used include, but are not limited to, Colour Transformation, Colour Correction using a colour chart, and image enhancement. The choice of colour model is crucial in representing an image for statistical processing. The standard colour model in computer displays is the RGB model. The channels in RGB are highly correlated, so RGB is unlikely to be the best model for describing information [1]. The CIELAB (L*a*b*) model provides a normalised and more visually uniform

chromaticity and more perceptually uniform luminance [1]. The HSV colour space is an intuitive colour model for describing data as well as the HSI colour model, which is designed for image processing [2]. Moreover, YCbCr is a model applied in digital video. The Cb and Cr components are employed as independent two-dimensional distributions which are unaffected by brightness [3]. Secondly, the pre-processed data are clustered into several groups such as background and particularly foreground, as regions of interest will be segmented out and used in the next process. Widely used methods include K-means clustering as a fast and simple technique [4]–[8], Fuzzy c-means [9] as it is more flexible than K-means, and Otsu's thresholding [4], [10], [11] as a robust binary classification. Thirdly, feature extraction is applied to calculate disease pattern representation in segmented areas. Various features have been used including features derived from the greyscale spatial independence matrix [4], [6], [12]–[15], shape and properties of regions of interest [9], [10], [16] and first-order statistical features [9]. Other potential features are widely applied in different applications such as HOG features [17] and visual perception features [18]. However, there is little evidence to demonstrate the effectiveness of features especially for plant disease patterns. In addition, most research studies apply individual sets of features [4], [6], [12], [13], [15], [19] or combine the whole set of different features in the system [10], [16]. The more features that are calculated, the more computational time is required, and so some studies have applied principal component analysis to remove correlated features [16], [20]. Fourthly, the classification models are constructed mainly using Neural Networks (NNs) and Support Vector Machine (SVM). Neural Networks are widely used for many applications in intelligent systems because of their ability to perform non-linear modelling. However, NNs have drawbacks in high computation complexity, a tendency for over-fitting, and lack of explainable relationships amongst inputs, outputs and variables [21]. The Support Vector Machine is a well-known classification method that is generalisable and able to cope with non-separable data [22].

Our classification system comprises pre-processing using different colour transformations, feature extraction using many types of features, and classification using SVM. The system is proposed to investigate the potential of different colour components, various features and SVM kernels affecting plant disease classification performance. Image segmentation is omitted in the automated system as we assume that the leaves are already segmented out from the background (though we recognise that this is far from a trivial task).

3 Methodology and Proposed Classification System

An automated classification system is proposed and is shown in Figure 1. The system consists of three main steps: pre-processing, feature extraction and classification. We assume that leaves have been segmented out from the background and that only one main leaf in each image is considered.

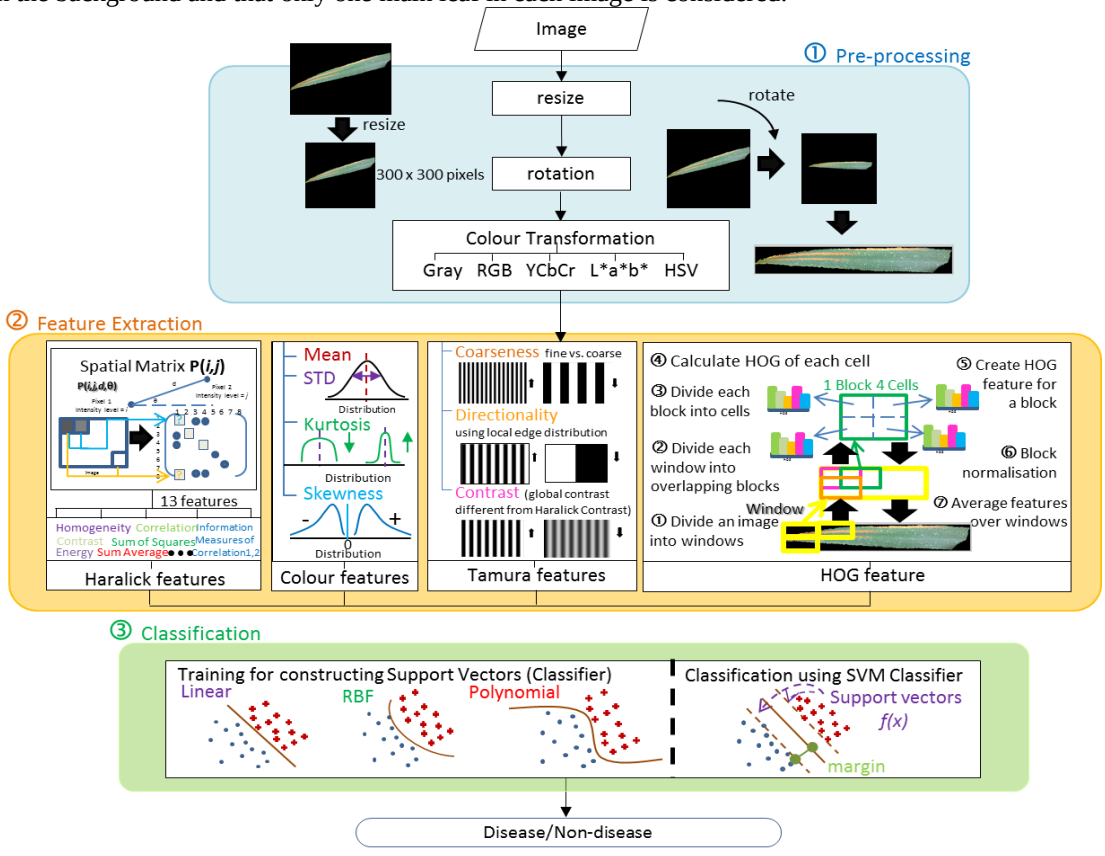


Figure 1 Proposed Classification System

3.1 Pre-processing

Images were obtained from the UK Food and Environment Research Agency [23] by the University of Bristol and the leaves in these images are previously segmented manually from the clutter in order to constrain the background conditions. Because of the diversity of captured images in terms of distance and orientation, pre-processing (see block ① in Figure 1) is then applied to standardise the images.. Firstly, a resize operation is performed to each image using nearest-neighbour interpolation. To mitigate for the differences in direction of angle of the captured leaves, the orientation angles of the leaves are calculated in terms of the directions of the major axes of the leaf ellipses, and then rotation is automatically applied to align the main leaf horizontally. Finally, the rotated leaf is cropped to remove empty space (background in black pixels). Image resizing and cropping is used to reduce processing time of subsequent feature extraction and the rotation process increases the reliability of the extracted features. For this system, five colour spaces, greyscale, RGB, YCbCr, L*a*b* and HSV, are used to evaluate their effects on classification performance.

3.2 Feature Extraction

After the images have been pre-processed, the patterns in the images are extracted in terms of Haralick features, first-order statistical features, Tamura features and HOG features (see block ② in Figure 1).

3.2.1 Haralick Features

Haralick features are developed through the grey-level co-occurrence matrix which measures the spatial relationships in image intensity [24]. This relationship is in terms of a matrix of relative frequencies $P(i,j,d,\theta)$ between two neighbouring pixels, one with intensity level i and another with intensity level j, and separated by distance d at directional angle θ. Thirteen features based on a normalised matrix $p(i,j)$ are calculated. Firstly, Angular Second Moment or Energy (EN) is a measure of uniformity of an image:

$$\mathrm{EN} = \sum_i \sum_j \{p(i,j)\}^2 \tag{1}$$

Contrast (CON) measures the variations or spatial frequency of intensity levels for a greyscale image at the reference positions and their neighbours:

$$\mathrm{CON} = \sum_{\substack{n=0 \\ |i-j|=n}}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \right\} \tag{2}$$

where N_g is the number of quantised intensity levels.

Homogeneity (HOM) is a measure of local homogeneity of an image. HOM will be high when an image is homogeneous because it takes the inhomogeneous area into account less than the homogeneous area by use of a weighting factor:

$$\mathrm{HOM} = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j) \tag{3}$$

Other features include Correlation, Sum of Squares, Sum Average, Sum Variance, Entropy, Difference Variance, Difference Entropy, and First and Second Information Measures of Correlation.

3.2.2 Tamura Features

As some features do not correspond to well-explained image patterns, Tamura features, which are based on visual perception, are also introduced [18]. These features include coarseness, directionality, contrast, line-likeness, regularity, and roughness. The first three features are selected in our classification system and the latter three features are omitted as they are combinations of the first three features. Contrast measures the polarization of distribution of black and white in an intensity image. Contrast by Tamura is a global property based on the skewness value of the image which differs from local contrast by Haralick:

$$\mathrm{CON\_T} = \sigma/(\alpha_4)^{1/4} : \textit{where } \alpha_4 = \frac{\mu_4}{\sigma^4} \tag{4}$$

where $\alpha_4 = \frac{\mu_4}{\sigma^4}$, μ_4 is Kurtosis value.

Coarseness (COAR) measures the block size in an image that is frequently repeated. The element size is large when that image is coarse, whereas the image has fine textures when the element size is small. Directionality (DIR) measures the frequency of transition from one colour value to another. An image with high pattern frequency tends to have a higher degree of directionality. Images with the same pattern but different direction have the same degree of directionality.

3.2.3 First-order Statistical Features (Colour Feature)

These features measure basic statistical information in an image. Mean (MEAN, μ) and standard deviation (STD, σ) determine a global average and variation of an intensity image, respectively. Skewness (SKEW)

measures asymmetry of the intensity probability distribution ($p(i)$). Skewness can have negative or positive values:

$$\text{SKEW} = \sigma^{-3} \sum_{i=1}^{N} (i-\mu)^3 p(i) \tag{5}$$

Kurtosis (KUR) measures the shape of the probability distribution relative to the standard normal distribution, and the value is based on the fourth moment of the data. A higher kurtosis value implies a broader and taller distribution shape, whereas a lower kurtosis indicates a narrower and shorter distribution shape:

$$\text{KUR} = \sigma^{-4} \sum_{i=1}^{N} (i-\mu)^4 p(i) \tag{6}$$

3.2.4 Histogram of Oriented Gradients (HOG)

Histogram of Oriented Gradients measures the local edge distribution of an image [17]. HOG features are calculated by the following processes. Firstly, an image is divided into a number of windows, and each window contains overlapping blocks. Each block comprises non-overlapping cells for which histograms of gradient orientation are computed. To cope with a variety of local contrast, normalization is applied to the HOG vector for each block separately. All vector components from each block are combined to create the HOG descriptors. The final HOG descriptor is an average of all vector components from sliding windows in the image.

3.4 Classification

Support Vector Machine (SVM) has been shown to obtain high classification performance and good generalization from various studies. It is a supervised learning method which is based on the basic concept of searching for an optimum hyperplane to separate training data into two classes with maximum margin (see block ③ in Figure 1) [22].

In some classification problems, a simple linear hyperplane cannot be applied to divide groups of data efficiently, especially when handling data that have a number of dimensions. Thus, a non-linear function is represented in the discrimination function instead of a linear function. Assuming the weight vector is a linear combination of training data, the function is expressed in terms of a dot product of kernels. This kernel function, which avoids the explicit mapping of data into the high-dimensional feature space, plays a vital role in improving performance. The widely used types of kernel functions include Radial Basis Function (RBF), Polynomial, and Multilayer perceptron [22].,To determine the hyperplane function which divides particular datasets properly is crucial. The last step of the process aims to discover the best fitting model from three kernel characteristics, namely linear, polynomial, or RBF.

4 Experimental Results

The system proposed in Figure 1 is implemented in MATLAB 2012b. Image data were collected by the UK Food and Environment Research Agency [23] (Figure 2 (a)-(b)). The leaves in each image were initially segmented out from the background manually (Figure 2(c)-(d)). Firstly, the system was tested for binary classification (i.e., yellow rust disease or non-disease) using 5-fold cross-validation on 50 non-diseased and 50 Yellow Rust diseased leaves. For pre-processing, the images are resized into 300x300 pixels, then rotated to arrange the leaves horizontally, and finally cropped to remove major areas of background (now represented as black pixels). RGB images are converted to greyscale, HSV, YCbCr, or L*a*b* models, from which each model is used individually or a combination of different colour models is used.

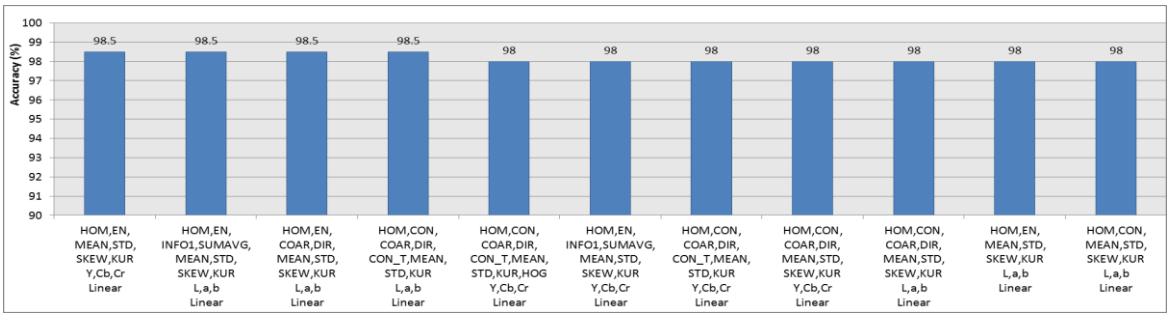
(a) Non-disease (b) Yellow Rust disease (c) Segmented non-disease (d) Segmented Yellow Rust

Figure 2 Wheat diseased and non-diseased images in the experiment

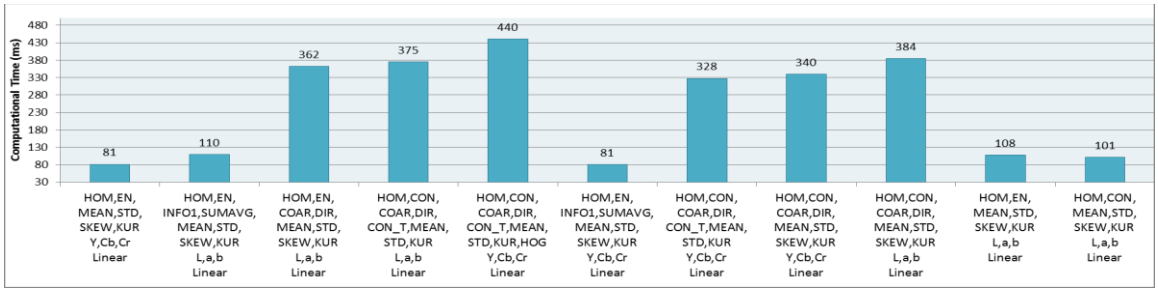
Thirteen Haralick features are created using relative information for pixels in an image with their consecutive neighbouring pixels at a zero directional angle. Also, the intensity levels of an image are quantised into 8 levels, creating an 8x8 spatial matrix for each colour component. Histogram of oriented gradients is calculated using average HOGs from all sliding windows in an image. Optimal window size is empirically set to 32x64 pixels with 50% overlap between blocks. Each block contains 2x2 cells and each cell size is 8x8 pixels. For SVM classifier, the scaling factor of the RBF kernel and the polynomial degree is set empirically to 1 and 2, respectively.

Initially, the combinations of four popular main Haralick features (homogeneity, energy, contrast and correlation) were investigated. It was found that the combinations of homogeneity and energy/contrast produced the highest binary classification accuracy, so these combinations were carried into the next phase of experiments. For the other nine Haralick features, different entropy, sum variance and first information measures of correlation have potential to improve classification accuracy; whereas combination of all four colour features and all three Tamura features have dominant on the classification accuracy. Hence, all combinations of four colour features and three Tamura features were carried into the next phase of combinations among feature sets for bi-class classification and multi-class classification.

Performance of the system for binary classification using a mixture of different features is shown in Figure 3. From the results, YCbCr and L*a*b* colour spaces have a significant influence on classification performance. Similarly, the SVM linear kernel is the best classifier to describe feature distribution for this dataset compared to RBF or polynomial kernels. Homogeneity of Haralick features and colour features (mean, skewness and kurtosis) are effective features to classify disease or non-disease leaf images as it is seen that these features always present in the best eleven accuracy results up to 98.5% in Figure 3(a).



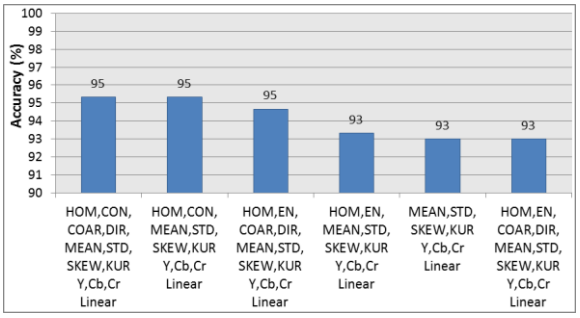
(a) Classification Accuracy



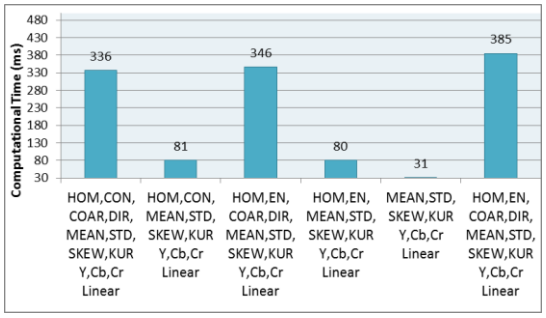
(b) Computational time

Figure 3 Top eleven accuracies for wheat binary classification (disease/non-disease)

Computational complexity for each classification is shown in Figure 3(b). It is simple to calculate Haralick and colour features, so the combinations containing only those features show low computational times as shown in the first bar in the chart in Figure 3(b); the classification time is less than 100 milliseconds. The computational times of Tamura and HOG features are relatively higher than the other two feature sets as they require many steps to calculate each feature; the fifth bar in the chart shows that with these additional two feature sets the computational time increases by about 300 milliseconds (from 81 msec to 440 msec).



(a) Classification Accuracy



(b) Computational time

Figure 4 Top six accuracies for wheat non-disease and disease

These feature sets were also investigated for multi-class classification (non-disease, Yellow Rust and Septoria diseases) as shown in Figure 4. Similar to binary classification, the best six mixtures of features, including two Haralick features (homogeneity and contrast or energy), two Tamura features (coarseness and directionality) and colour features, give the highest accuracy of up to 95%. When we combine HOG with other features we found that there is little improvement in accuracies compared to the use of HOG alone, and classification accuracies are less than for the feature combinations without HOG. The processing time for

computing only the colour feature set is the least at approximately 30 milliseconds, and the classification accuracy shows a high value at 93% (fifth bar in the chart in Figure 4(a)-(b)).

5 Conclusion

Since using too many features may cause over-fitting and also require more computational time, the classification system discussed in this paper is proposed to assess effectiveness of features from different feature sets, such as Haralick features, first-order statistical features, Tamura features and HOG features. The results show that the most efficient features include, but are not limited to, two Haralick features (mixture of homogeneity and contrast or energy), three Tamura features and colour features. Use of only the HOG features is also a potential approach for classification, but the combination of features mentioned above performs better in both binary and multi-class classification. Also, processing times in calculating Haralick and colour features are less than for Tamura and HOG features. Following the experiments, it is planned that the effective sets of features be implemented in lightweight systems, such as a mobile application for real plant disease classification.

Acknowledgement:

This work was supported by a University of Ulster Vice Chancellor's Research Scholarship. We are grateful to EPSRC-DST funded India-UK Advanced Technology Centre (IU-ATC) project for providing some crop disease databases.

References:

[1] A. McAndrew, *Introduction to Digital Image Processing with MATLAB*. Boston: Thomson Course Technology, 2004.

[2] R. C. Gonzalez, R. E. Wood, and Steven L. Eddins, *Digital Image Processing using MATLAB*. New Jersey: Pearson Education, Inc., 2004.

[3] S. Kai, L. Zhikun, S. Hang, and G. Chunhong, "A Research of maize disease image recognition of Corn Based on BP Networks," in *2011 Third Int. Conf. Measuing Technol. Mechatronics Autom.*, 2011, pp. 246–249.

[4] H. Al Hiary, S. Bani Ahmad, M. Reyalat, M. Braik, and Z. ALRahamneh, "Fast and Accurate Detection and Classification of Plant Diseases," *Int. J. Comput. Appl.*, vol. 17, no. 1, pp. 31–38, 2011.

[5] S. Bashir and N. Sharma, "Remote Area Plant Disease Detection Using Image Processing," *IOSR J. Electron. Commun. Eng.*, vol. 2, no. 6, pp. 31–34, 2012.

[6] D. Al Bashish, M. Braik, and S. Bani-Ahmad, "A framework for detection and classification of plant leaf and stem diseases," in *Signal Image Process. ICSIP 2010 Int. Conf. on*, 2010, pp. 113–118.

[7] S. R. Dubey and A. S. Jalal, "Detection and Classification of Apple Fruit Diseases Using Complete Local Binary Patterns," in *2012 Third Int. Conf. Comput. Commun. Technol.*, 2012, pp. 346–351.

[8] H. Wang, G. Li, Z. Ma, and X. Li, "Image recognition of plant diseases based on backpropagation networks," in *Image Signal Process. (CISP), 2012 5th Int. Congr.*, 2012, pp. 894–900.

[9] M. El-Helly, R. Ahmed, and S. El-Gammal, "An Integrated Image Processing System for Leaf Disease Detection and Diagnosis," in *Proc. 1st Indian Int. Conf. Artif. Intell. IICAI 2003*, 2003, pp. 1182–1195.

[10] Q. Yao, Z. Guan, Y. Zhou, J. Tang, Y. Hu, and B. Yang, "Application of Support Vector Machine for Detecting Rice Diseases Using Shape and Color Texture Features," *2009 Int. Conf. Eng. Comput.*, pp. 79–83, 2009.

[11] J. Pang, Z. Bai, J. Lai, and S. Li, "Automatic segmentation of crop leaf spot disease images by integrating local threshold and seeded region growing," *2011 Int. Conf. Image Anal. Signal Process.*, pp. 590–594, Oct. 2011.

[12] S. Ananthi and S. V. Varthini, "Detection and Classification of Plant Leaf Diseases," *Int. J. Res. Eng. Appl. Sci.*, vol. 2, no. 2, pp. 763–773, 2012.

[13] S. B. Dhaygude and N. P. Kumbhar, "Agricultural plant Leaf Disease Detection Using Image Processing," *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.*, vol. 2, no. 1, pp. 599–602, 2013.

[14] D. G. Kim, T. F. Burks, J. Qin, and D. M. Bulanon, "Classification of grapefruit peel diseases using color texture feature analysis," *Int. J. Agric. Biol. Eng.*, vol. 2, no. 3, pp. 41–50, 2009.

[15] R. Pydipati, T. F. Burks, and W. S. Lee, "Identification of citrus disease using color texture features and discriminant analysis," *Comput. Electron. Agric.*, vol. 52, no. 1–2, pp. 49–59, Jun. 2006.

[16] H. Wang, G. Li, Z. Ma, and X. Li, "Image recognition of plant diseases based on principal component analysis and neural networks," *2012 8th Int. Conf. Nat. Comput.*, pp. 246–251, May 2012.

[17] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, pp. 886–893, 2005.

[18] H. Tamura, S. Mori, and T. Yamawaki, "Textural Features Corresponding to Visual Perception," *Syst. Man Cybern. IEEE Trans.*, vol. 8, no. 6, pp. 460–473, 1978.

[19] T. A. Pham, "Optimization of Texture Feature Extraction Algorithm," 2010.

[20] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–45, Sep. 2010.

[21] A. K. Jain, J. Mao, H. Road, and S. Jose, "Artificial Neural Networks : A Tutorial," in *IEEE Comput. Spec. Issue Neural Comput.*, 1996, pp. 1–49.

[22] N. Christianini and J. Shawe-Taylor, *An Introduction to Support Vector Machine*. Cambridge University Press, 2000.

[23] "The Food & Environment Research Agency." [Online]. Available: http://fera.co.uk/.

[24] R. M. Haralick, K. Shanmugam, and DinsteinIts'shak, "Textural Features for Image Classification," *Syst. Man Cybern. IEEE Trans.*, vol. SMC-3, no. 6, pp. 613–621, 1973.