Recognition of Tomato Late Blight by using DWT and Component Analysis

Hiteshwari Sabrol[1], Satish Kumar[2]
[1]Department of Computer Science & Applications, Panjab University, Chandigarh, Chandigarh 160014, India
[2]Department of Computer Applications, P.U. SSG Regional Centre, Hoshiarpur, 146023, India
(Panjab University, Chandigarh)

Article Info	ABSTRACT
	Plant disease recognition concept is one of the successful and important applications of image processing and able to provide accurate and useful information to timely prediction and control of plant diseases. In the study, the wavelet based features computed from RGB images of late blight infected images and healthy images. The extracted features submitted to Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA) and Independent Component Analysis performed (ICA) for reducing dimensions in feature data processing and classification. To recognize and classify late blight from healthy plant images are classified into two classes i.e. late blight infected or healthy. The Euclidean Distance measure is used to compute the distance by these two classes of training and testing dataset for tomato late blight recognition and classification. Finally, the three-component analysis is compared for late blight recognition accuracy. The Kernel Principal Component Analysis (KPCA) yielded overall recognition accuracy with 96.4%.

Corresponding Author:

Hiteshwari Sabrol,
Department of Computer Science & Applications,
Panjab University,
Chandigarh, Chandigarh 160014, India.
Email: hiteshwarisabrol@gmail.com

1. INTRODUCTION

"Plant disease is an impairment of the normal state of the plant that interrupts or modifies its vital functions" [1]. There are several types of plant disease could cause several losses to the production of crops. The presence of the pathogen depends on the favorable environment conditions and varieties of crops grown, which is the reason for occurrence and prevalence of plant diseases. There are various plant disease management programs that will help to reduce losses in yields and grain quality. An automatic plant disease recognition and diagnosis system can design by using image processing and pattern recognition techniques [2-7].

There is various kind of abnormality states present on the plants leaf which can be identified by mean of manual inspection. The image processing and pattern recognition techniques play the worth full role to convert manual process to automate the process. The automatic diagnosis system based on plant disease features reduces the dependency on experts in the area concerned. In the study, the late blight-infected tomato leaf images and healthy leaf images were taken for the experiment. The two kind of leaf divided into two categories or classes, i.e., late blight infected and healthy. The plant disease features extracted after applying discrete wavelet transformation. Excessive extracted features reduced by using "principal component analysis" (PCA) [8-10], "independent component analysis" (ICA) [11], [12] and "kernel principal component analysis" (KPCA) [13], [14] and some other methods could be used to reduce dimensions of the feature data.

This three-component analysis also used for classification purpose by learning technique. There is The tomato late blight disease recognition from healthy leaf images includes eight steps as mentioned in Figure 1. Step 1 Late blight infected: 48 and healthy (Normal): 58. Step 2 Created training and testing dataset. Step 3 Image preprocessing: Perform Adaptive histogram equalization and reshaping the images, Extracting each R, G, B components and Apply whitening for ICA Step 4. Feature Extraction using single level two-dimensional Haar Wavelet. Step 5 Extracting low and high frequency of each R, G, and B component. Step 6 Data summarization and classification by PCA, KPCA & ICA. Step 7 Apply Euclidean Distance to measure the similarity distance between training and testing data by two classes (Late blight infected and healthy. Step 8 Tomato Late Blight Recognition and Classification comparison and cross-validation of three Component Analysis (PCA, KPCA, and ICA).
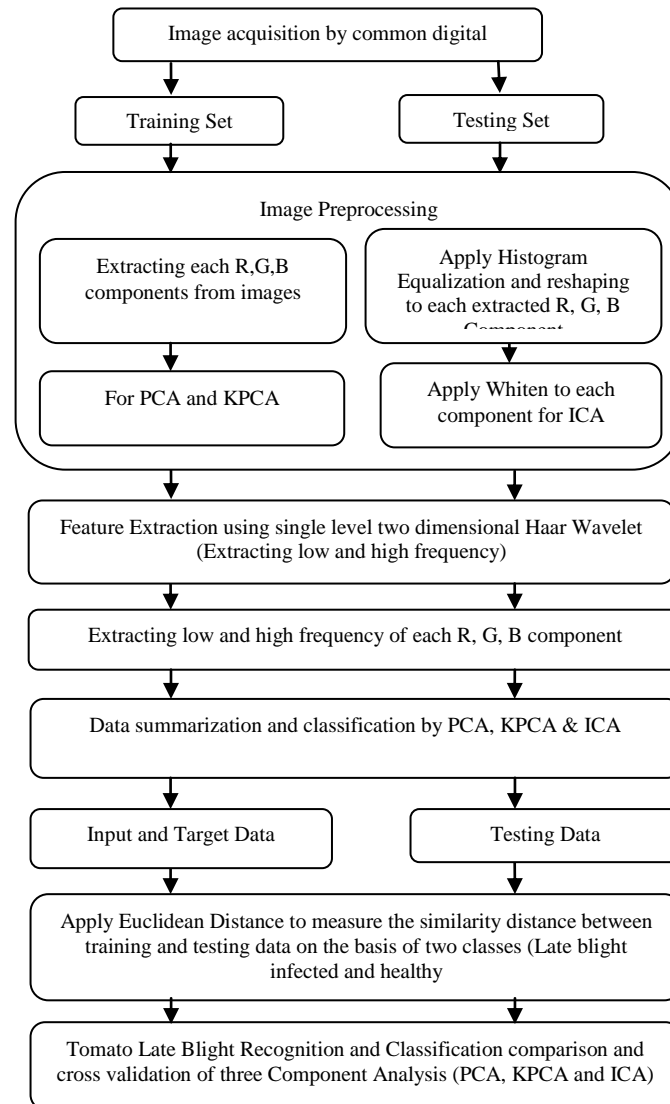
```
                    ┌─────────────────────────────────────┐
                    │  Image acquisition by common digital │
                    └─────────────────────────────────────┘
                         │                          │
                ┌────────────────┐        ┌────────────────┐
                │  Training Set   │        │  Testing Set    │
                └────────────────┘        └────────────────┘
                         │                          │
        ┌──────────────────────────────────────────────────────┐
        │                  Image Preprocessing                   │
        │  ┌──────────────────┐      ┌──────────────────────┐   │
        │  │ Extracting each  │      │ Apply Histogram       │   │
        │  │ R,G,B components │      │ Equalization and      │   │
        │  │ from images      │      │ reshaping to each     │   │
        │  │                  │      │ extracted R, G, B     │   │
        │  │                  │      │ Component             │   │
        │  └──────────────────┘      └──────────────────────┘   │
        │  ┌──────────────────┐      ┌──────────────────────┐   │
        │  │ For PCA and KPCA │      │ Apply Whiten to each  │   │
        │  │                  │      │ component for ICA     │   │
        │  └──────────────────┘      └──────────────────────┘   │
        └──────────────────────────────────────────────────────┘
                                 │
        ┌──────────────────────────────────────────────────────┐
        │ Feature Extraction using single level two dimensional  │
        │ Haar Wavelet (Extracting low and high frequency)       │
        └──────────────────────────────────────────────────────┘
                                 │
        ┌──────────────────────────────────────────────────────┐
        │ Extracting low and high frequency of each R, G, B      │
        │ component                                              │
        └──────────────────────────────────────────────────────┘
                                 │
        ┌──────────────────────────────────────────────────────┐
        │ Data summarization and classification by PCA, KPCA & ICA│
        └──────────────────────────────────────────────────────┘
                    │                          │
        ┌────────────────────┐      ┌────────────────────┐
        │ Input and Target Data│      │  Testing Data      │
        └────────────────────┘      └────────────────────┘
                    │                          │
        ┌──────────────────────────────────────────────────────┐
        │ Apply Euclidean Distance to measure the similarity     │
        │ distance between training and testing data on the      │
        │ basis of two classes (Late blight infected and healthy)│
        └──────────────────────────────────────────────────────┘
                                 │
        ┌──────────────────────────────────────────────────────┐
        │ Tomato Late Blight Recognition and Classification      │
        │ comparison and cross validation of three Component     │
        │ Analysis (PCA, KPCA and ICA)                           │
        └──────────────────────────────────────────────────────┘
```

Figure 1. Steps for Recognition and Classification of Tomato Late Blight by using DWT and Component Analysis

2.   RESEARCH METHOD
2.1. Tomato Leaf Blight

The late blight found on tomato plants is caused by the fungus *Phytophthora infestans. The occurrence of late blight found in later season* or features often not appearing until after blossom. The symptoms or features of late blight first appear on the lower, older leaves as water-soaked, gray-green spots. The image affected from late blight shown in Figure 2(a) and (b).
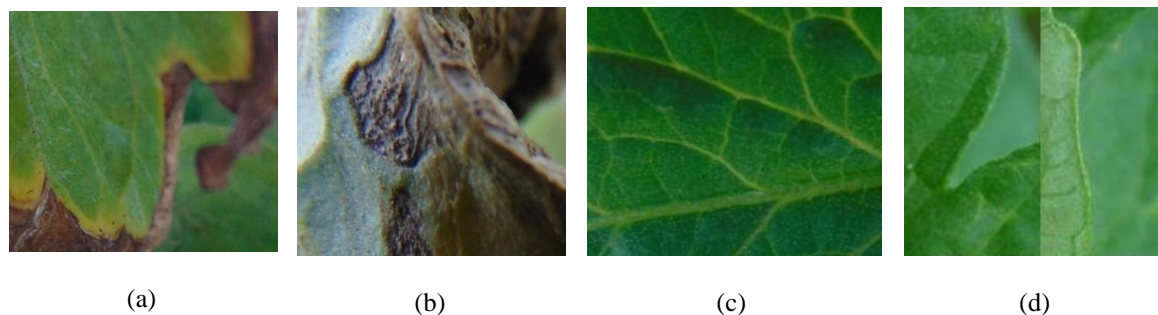
<div align="center">(a)   (b)   (c)   (d)</div>

Figure 2. (a) & (b) Late Blight Affected Tomato Leaves (c) and (d) Healthy Tomato Leaves

2.2. Image DataSet

In this study, a dataset of total 106 digital images of late blight tomato plant disease and healthy were obtained by using the common digital camera. From which the 48 images are late blight infected and 58 images are healthy. Total five times the images are randomly used for training and testing purpose. For training, randomly the set of 90%, 80%, 70%, 60% and 50% images from the total dataset used. For the testing, randomly the set of 10%, 20%, 30%, 40% and 50% images from the total dataset used. The proposed algorithm implemented with MATLAB 2014b.

2.3. Image pre-processing

Apply contrast limited adaptive histogram equalization to enhance the contrast of the intensity and perform reshaping on RGB images of late blight and healthy.

2.4. Feature Extraction using Discrete Wavelet Transformation

Wavelet transformation is used to extract wavelet features from tomato late blight affected and healthy images. As wavelet transformation convert digital signal to wavelet signal or decomposes the digital signal into wavelet function. The wavelet transformation can analysis the detail of image like vertical, diagonal and horizontal subbands of the image[15-17]. The detailed information of the image can be retrieved by low pass and high pass filtering [18-20]. In the experiment, the discrete wavelet transform used for extracting features from the disease affected the area of the plant. First R, G, B components extracted from the images individually. Second compute the low and high frequency of each R, G, B component by using single level two-dimensional discrete Haar wavelet. The Haar Wavelet function defined by the following formula [17].

$$\Psi(t) = \begin{cases} 1, for\ t \in \left(0, \frac{1}{2}\right) \\ -1, for\ t \in \left(\frac{1}{2}, 1\right) \\ 0, otherwise \end{cases} \tag{1}$$

Third compute the standard deviation of each extracted component to be in the scalar form. Finally, combined the three scalar forms of each R, G, and B component and submitted to PCA, KPCA and ICA for data summarization, training, and testing purpose.

2.5. Component Analysis

The two or more independent variables composed in such way to analyze their independence is called Component analysis. To identify the singular component and determine the influenced component due to the changing behavior is the main aim of component analysis. Eliminate ineffective or less effective components for the purpose of reduction of indifferent elements, maintenance and improving generalization. The PCA, KPCA, and ICA are three important feature extraction techniques used for dimensionality reduction and work on self-organizing unsupervised learning approach. We used all these three techniques (PCA, KPCA, and ICA) for data reduction and classification.

The most effective and simple multivariate analysis based to eigenvector is "Principal Component Analysis". PCA an orthogonal linear transformation used to convert the data to a new coordinate system. It also used for signal processing and data reduction in statistical pattern recognition. Self-organized learning provides an adaptive algorithm for PCA. It maximizes the rate of decrease of variance using Eigen

decomposition method or value decomposition theorem. We have used Singular value decomposition theorem better then Eigen decomposition. In mathematical terms, calculate the eigenvalues of the covariance matrix of the original outputs. PCA linearly transform a high-dimensional space which is uncorrelated and orthogonal. First, compute the eigenvalues and eigenvectors, then sort the eigenvalues in descending order and ignore the really small values. Then we transform data into the eigenspace formed by the selected eigenvector. The mean vector of the population is defined as:

$$m_x = \frac{1}{K}\sum_{k=1}^{K} x_k \tag{2}$$

To find the covariance matrix

$$C_x = \frac{1}{K}\sum_{k=1}^{K} x_k x_k^T - m_x m_x^T \tag{3}$$

where C_x is the covariance, K is the samples from random values, and $(x-m_x)(x-m_x)^T$ are the n*n order of matrices.

$$C_y = AC_x A^T \tag{4}$$

where A is the transformation matrix to map x into vector y and the diagonal matrix is C_y and containing elements with the main diagonal are the eigenvalues of C_x. The matrix A_k is designed from k eigenvectors equivalent to the k largest eigenvalues,   producing a transformation matrix of order k*n. The reconstruction of vector by A_k is:

$$\hat{x} = A_k^T y + m_x \tag{5}$$

The *kernel* based method used for nonlinear PCA is a Kernel Principal Component Analysis. The kernel method is used to transforms the original inputs into high dimensional feature space (inner product kernel), where the computation performed on feature space which is non-linearly related to input space. In Kernel PCA, the relation between input space and feature space is nonlinear but implementation relies on linear algebra, so we think that it's an extension of ordinary PCA. In this study, we were using Gaussian-type kernel. Mathematically, we would like to find direction vector in the feature space. All the matters remain same as in PCA, but with kernel matrix

$$K_{ij} = \langle \phi(x_i), \phi(y_i)\rangle = k(x_i, y_i) \tag{6}$$

The projection onto the optimal direction is

$$\langle \phi(x), e\rangle = \langle \phi(x), \sum_i \alpha_i \phi(x_i)\rangle = \sum_i \alpha_i k(x, x_i) \tag{7}$$

Independent Component Analysis: (Cardoaso 2003) stated that "Independent Component Analysis is the decomposition of random vector into linear components that are statistically independent as possible, where the term independence is understood in its strongest statistical sense; ICA goes beyond (second order) decorrelation and therefore requires that the observations representing the data vector be non-Gaussian"[20]. In the study, we used whitening as a preprocessing step for ICA algorithm to find the uncorrelated components but need not be independent, and the variance is equal to the unity. Next, apply discrete Haar wavelet transformation to extract wavelet features.

3.   EVALUATION

We use the Euclidean Distance method to measure the distance between the training set and testing set. Either through supervised learning or unsupervised learning are known, an unknown object, may be called test object can recognize by associating it with one of those class using similarity (or in other words, dissimilarity or distance) measures. To measure the dissimilarity or distance of two classes i.e. Tomato late blight infected or healthy images, we used Euclidean Distance measure as following

$$\sqrt{\sum_{i=1}^{n}(x - y)^2} \tag{8}$$

Using PCA and KPCA, we measure the distance between eigenvalue of training data and testing data. In the case of ICA, the distance measured between eigenvector computed. If the distance is less from a particular class, then the testing image belongs to that class.

4. RESULTS AND DISCUSSION

The recognition and classification of tomato late blight by using discrete wavelet transformation and component analysis yielded accuracy 96.4%. The PCA, KPCA, and ICA are used to the reduction of dimensions of the data using wavelet features and then these techniques used for tomato plant disease recognition and classification purpose. The two classes i.e. late blight infected and healthy used to classify whether the image is affected by late blight disease or not. To calculating the minimum distance between training and testing data, Euclidean distance measure was used to classify that testing image is belongs to which class infected or healthy. Table 1. shows the comparative results of the three-component analysis. The PCA producing second overall recognition accuracy with 89.8% after KPCA. In the first, test dataset of 10%, ICA resulting the worst recognition accuracy and yielding overall 34% recognition only.

5. CONCLUSION AND FUTURE SCOPE

In the study, PCA, KPCA, and ICA used for data reduction and recognition of tomato late blight from a healthy leaf in digital images. The plant disease recognition and classification based on two classes i.e. images affected by late blight and healthy. For feature extraction, we used discrete wavelet transformation based on Haar wavelet. As the results show, extracted feature data from the images of plant diseases by reducing dimensions could reduce the running time for classification, and acceptable recognition results could obtain. The method used in the study proved that PCA, KPCA, and ICA used for extracted feature data reduction from tomato late blight images and also used for recognition purpose. Finally, the distance between training and testing dataset computed by Euclidean distance measure by two classes. If the distance is less from any of two classes, then the particular testing image belongs to one of the two classes either late blight infected or healthy one. The PCA is a kind of linear projection, and it could not correctly handle non-linear data, but it could handle using KPCA. All the properties of ordinary PCA carry over to KPCA. The Kernal PCA is linear in feature space and nonlinear in the input space. As such, it can be applied to all those domains where ordinary PCA used for feature extraction or data reduction, for which nonlinear extension would make sense. As compare to PCA, KPCA and ICA is not only statistical independent up to second order but up to all the individual components of the output vector and involves no constraint of orthogonality.

Table 1. Recognition and Classification of Tomato Late Blight accuracy

Testing Dataset	Principal Component Analysis		Kernel Principal Component Analysis		Independent Component Analysis	
	Late Blight Infected	Healthy	Late Blight Infected	Healthy	Late Blight Infected	Healthy
10%	100%	100%	100%	100%	0%	100%
20%	90%	92%	90%	92%	75%	25%
30%	79%	88%	93%	47%	7%	100%
40%	84%	78%	89%	48%	47%	65%
50%	96%	21%	100%	10%	42%	55%
Overall Accuracy	89.8%	75.8%	96.4%	51.12%	34%	69%

REFERENCES

[1] A. Kelmam, "Plant Diseases Encyclopedia*," Britannica,* 2012.
[2] C. C. Tucker and S. Chakraborty, "Quantitative assessment of lesion characteristics and disease severity using digital image processing," *Journal Phytopathology*, vol. 145, pp. 273–278, 1997.

[3]   I. S. Ahmand, *et al.,* "Color classifier for symptomatic soybean seeds using image processing," *Plant Disease*, vol. 83, pp. 320–327, 1999.

[4]   J. C. Lai, *et al.*, "Advances in research on computer-vision diagnosis of crop diseases," *Scientia Agricultura Sinica*, vol. 42, pp. 1215–122, 2009.

[5]   C. H. Bock, *et al.*, "Automated image analysis of the severity of foliar citrus canker symptoms," *Plant Disease*, vol. 93, pp. 660–665, 2009.

[6]   M. Ramakrishnan and A. S. A. Nisha, "Groundnut Leaf Disease Detection and Classification by using Back Propagation Network," *Proc. IEEE Int. Conf. on Communications and Signal Processing, Melmaruvathur*, pp. 964-968, 2015.

[7]   N. M. Tahir, *et al.*, "Classification of Elaeis Guineensis Disease-Leaf under uncontrolled illumination using RBF Network," *Proc. IEEE Int. Conf. on Control System, Computing, and Engineering, Malaysia*, pp. 617 – 621, 2014.

[8]   B. Li, *et al.*, "Hyperspectral identification of rice diseases and pests based on principal component analysis and probabilistic neural network," *Transactions of the CSAE,* vol. 25, pp. 143–147, 2009.

[9]   H. Jie, "Application of PCA Method on Pest Information Detection of Electronic Nose," in *the proceeding of IEEE international Conference on Information Aqusition.Weihai, Shandong, China*, pp. 1456-1468, 2006.

[10] M. M. Seinl, *et al.*, "Authentications of Myanmar National Registration Card," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI),* vol. 2, pp. 53-58, 2013.

[11] M. Lennon, *et al.*, "Independent Component Analysis as a tool for the dimensionality reduction and the representation of hyperspectral images. Geoscience and Remote Sensing Symposium," *IEEE 2001 International*, vol. 6, pp. 2893-289, 2001.

[12] Y. Fangzhou, *et al.*, "Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets," *BMC Bioinformatics*, pp. 13-24, 2012.

[13] B. Schölkopf, *et al.*, "Kernel principal component analysis," *Artificial Neural Networks ICANN'97, W. Gerstner and A. Germond and M. Hasler and J.-D. Nicoud (Eds.), Springer Lecture Notes in Computer Science*, vol. 1327, pp. 583-588, 1997.

[14]   X. Li, *et al.*, "Determination of dry matter content of tea by near and middle infrared spectroscopy coupled with wavelet-based datamining algorithm," *Computer and Electronics in Agriculture*, vol. 98, pp. 26-53, 2013.

[15] R. Polikar, "The Wavelet Tutorial Part I," http://users.rowan.edu/~polikar/WAVELETS/WTpart1.html.

[16] O. S. Jahromi, *et al.*, "Algebraic theory of optimal filter banks," *IEEE Transactions on Signal Processing*, vol. 51, pp. 442–457, 2003.

[17] C. R. González and R. E. Woods, "Digital Image Processing," *PHI*, 3rdEd, 2008.

[18] A. Akhtar, *et al.*, "Automated Plant Disease Analysis (APDA): Performance Comparison of Machine Learning Techniques", *FIT, Frontiers of Information Technology, Frontiers of Information Technology,* pp. 60-65, 2013.

[19] N. Ghaffarzadeh, "A New Method for Recognition of Arcing Faults in Transmission Lines using Wavelet Transform and Correlation Coefficient," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI),* vol/issue: 1(1), pp. 1-7, 2013.

[20] J. F. Cardaso, "Dependence, correlation and Gaussianity in independent component analysis," *Journal of Machine Learning Research*, vol. 4, pp. 1177-1203, 2003.