Detection of Leaf Disease Using Principal Component Analysis and Linear Support Vector Machine

Heltin Genitha C
Department of Information Technology
St.Joseph's College of Engineering
Chennai, India
heltingenitha@stjosephs.ac.in

Dhinesh E
Department of Information Technology
St.Joseph's College of Engineering
Chennai, India
dhineshskate007@gmail.com

Jagan A
Department of Information Technology
St.Joseph's College of Engineering
Chennai, India
jaganaravind1997@gmail.com

2019 11th International Conference on Advanced Computing (ICoAC) 978-1-7281-5286-8/20/$31.00 ©2020 IEEE 10.1109/ICoAC48765.2019.246866

***Abstract*— Agriculturalists find difficulties in classifying the diseases in leaves. In olden days, farmers detected the leaf diseases by observing the leaf by its appearance which does not provide accurate results. The simple, efficient and computationally effective plant disease identification system should be introduced to help farmers to detect the leaf diseases with high accuracy. In this paper, the diseases in the leaves are detected by extracting high quality features of the leaf by Principal Component Analysis. It is an image retrieval technique used to extract the high dimension feature of the image without losing its originality. The features such as perimeter, minor axis, area, major axis, equivalence diameter etc., are extracted from the diseased leaf. The input image is pre-processed by removing the noise using median filter, extracting the edges using Sobel filter, identifying the direction of the leaf using Gabor filter and segmenting the image by threshold segmentation. Finally, Linear Support Vector Machine classifier identifies the support vectors to identify the disease. Experiments are carried out using the datasets of tomato plant leaves where tomato is cultivated large scale in Coimbatore, India. Powdery mildew and early blight diseases are the most common disease in tomato which is detected and analysed in this work. The results obtained shows that there is an increase in accuracy when compared to other methodologies.**

Keywords*— Linear support vector machine, Principal component Analysis, Gabor filter.

I. INTRODUCTION

Farmers largely depends only on the agriculture since, they are the essentials in our country. Several issues in agriculture and farming have occurred due to diseases in plant leaves. The diseases in the leaves can easily be detected using multimedia files and image processing techniques. The rapid increase in the size of the multimedia files in recent generation, requires efficient technique to retrieve images from large data set. Conventional search techniques are based on text because the images must be annotated and recorded accordingly. The work of user-based explanation becomes overwhelming due to large sized images, and the text fails to express the actual information of the images. Thus, Content Based Image Retrieval (CBIR) techniques are used to detect the diseases in the leaf. CBIR is the efficient method to retrieve the image with low level principal features resembling colours, shapes, textures [1, 2].

Many researchers have carried out works to detect the leaf disease using image retrieval systems, which is time and cost effective. Such work has been reported by Patil and Kumar [3] to classify diseases in soybean using Local Gray Gabor Pattern (LGGP) filter technique. The performance of LGGP is low because it cannot combine the shape and colour feature. Sivasangari et al., detected the diseases in cotton leaf using Genetic Algorithm, which has the issue on time and result in training and testing [4]. Prithviraj detected unhealthy region of plant leaves using Fuzzy C Means (FCM) algorithm and it is only done for four diseases and further it can be practised in various diseases [5]. Kaliley et al., used a CBIR technique for the recognition of diseases which leads to crop spoliation based on CBIR [6], Madhavi et al., used a creative technique for CBIR with colour and texture feature which has issues in the retrieval of exact characteristics of the sample image [7], Sanjana et al., compared a simple colour difference based approach to fragment the disease affected regions [8]. Oo et al., used K Means clustering algorithm and SVM classification algorithm to detect and classify the leaf disease [9]. Carmel et al., used deep neural network to detect the leaf disease [10].The experts make immense effort in notifying the feedbacks to the farmers personally through mobile phones.

Hence, these works show that there requires a recognition system to classify the diseases in the leaf. In this paper, Principal Component Analysis (PCA) is used to extract the principal features from the diseased leaf. It is an image retrieval method used to retrieve the principal features without loss of the original data. The principal features extracted from PCA is given as input to Linear Support Vector Machine (LSVM) to classify the diseases in the infected leaf.

II. STUDY AREA AND DATA USED

Coimbatore and Mettupalyam are the best suitable places for the cultivation of tomato in Tamil Nadu. The tomato plants are with five to nine leaflets which is 10-25 cm long each and pigmentation, colour and cultivation of the fruits are affected by high light intensity. This leads to various bacterial diseases such as southern blight, powdery mildew, verticillium wilt, anthracnose, bacterial speck, blossom end rot, etc. The changing climatic conditions affect the leaves which lead to early blight disease. In this work, the leaf diseases named powdery mildew and early blight is detected from the tomato leaf.

III. METHODOLOGY

This proposed system implements a Principal Component Analysis and high efficient SVM Linear classifier recognition algorithm (Fig. 1). All features required to classify the diseases are extracted. The test image is collected from the sources like digital scanners or cameras and maintained in a database. The sample image is pre-processed prior to the feature extraction process. It converts the greyscale image from the input image which is then converted into binary image later.

The unnecessary noises in the sample leaf are removed using Median filter. Sobel Filter extracts the edges and the

Gabor filter further provides strong response to the direction of the leaf. The intensity values are effectively equalized using histogram equalization. The histogram value determines the texture values and the leaf image is segmented by comparing the values obtained from histogram equalization. Then, some morphological features, such as area, orientation, perimeter, convex area, major axis, eccentricity, minor axis, texture, colour and equivalent diameter are extracted. The PCA extracts the principal features and provides these features to the Linear Support Vector Machine and the diseases are detected by matching the support vectors.

A. Pre-processing

1) Image Collection

Convert the given input leaf image into grayscale image using (eq. 1),

$$f(i,j) = 0.114 * I(i,j,1) + 0.587 * I(i,j,2)0.299 * I(i,j,3) \quad (1)$$

2) Removal of noise - Median Filter

Unnecessary noises in the leaf images are removed using Median filter. Such type of noise removal step increases the accuracy of the results. It preserves the edges in the image. It takes the median value in the pixel which is sorted in increasing order.

3) Edge extraction of the leaf - Sobel Operator

The vertical and horizontal edges in an image are detected by Sobel operator. It is used to estimate the gradient of an image to detect the edges. The corresponding gradient vector for each pixel is operated by Sobel filter.

The following are the steps of the Sobel Operator

Step 1: Input image is accepted.

Step 2: For the input image, mask Gx, Gy is applied, where Gx, Gy are gradients.

Step 3: Sobel edge identification algorithm is applied.

Step 4: Masks Gx, Gy is manipulated independently.

Step 5: The consolidated result is to locate the total extent of the inclination $|G| = \sqrt{Gx^2 + Gy^2}$

Step 6: The output edges are the absolute magnitude.

4) Identify the direction of leaf- Gabor Filter

A Gabor filter is a sinusoidal plane of frequency and orientation, altered by a Gaussian envelope. Shapes, sizes and smoothness of the image are determined by the Gabor filter. It is an efficient tool for extracting texture features [9]. A strong response is provided to a particular direction for locating the target images. The Gabor filter (eq. 2 & eq.3) can be calculated using the formula,

$$G_x[i,j] = B - \frac{(i^2+j^2)}{2\sigma^2}\cos(2\pi f(i\cos\phi + j\sin\phi)) \quad (2)$$

$$G_y[i,j] = C - \frac{(i^2+j^2)}{2\sigma^2}\sin(2\pi f(i\cos\phi + j\sin\phi)) \quad (3)$$

where f is a frequency of texture, ϭ is the sigma or standard deviation of the Gaussian envelope, ϕ is texture orientation, 0≤i≤n; 0≤j≤n, n is the number of pixels, B ,C are constants.

5) Histogram Equalization

Global contrast values of the images are increased using histogram equalization since the data is represented by the nearest contrasting values. The intensities are equalized on the histogram. This method replaces the pixels from lower contrast to higher contrast values. The equalization $h(v)$ can be derived as (eq.4),

$$h(v) = round\left(\frac{cdf(v)-1}{63} \times 255\right) \quad (4)$$

where *cdf(v)* is the non-zero value of the cumulative distribution function, $h(v)$ is the histogram equalization.

6) Threshold Segmentation

In threshold segmentation, each pixel in an image is replaced with a white pixel if the threshold is lesser than the intensity I, or a black pixel if the threshold is greater than the image intensity I. The typical threshold range ought to be between 160-180, which is considered to be normal.

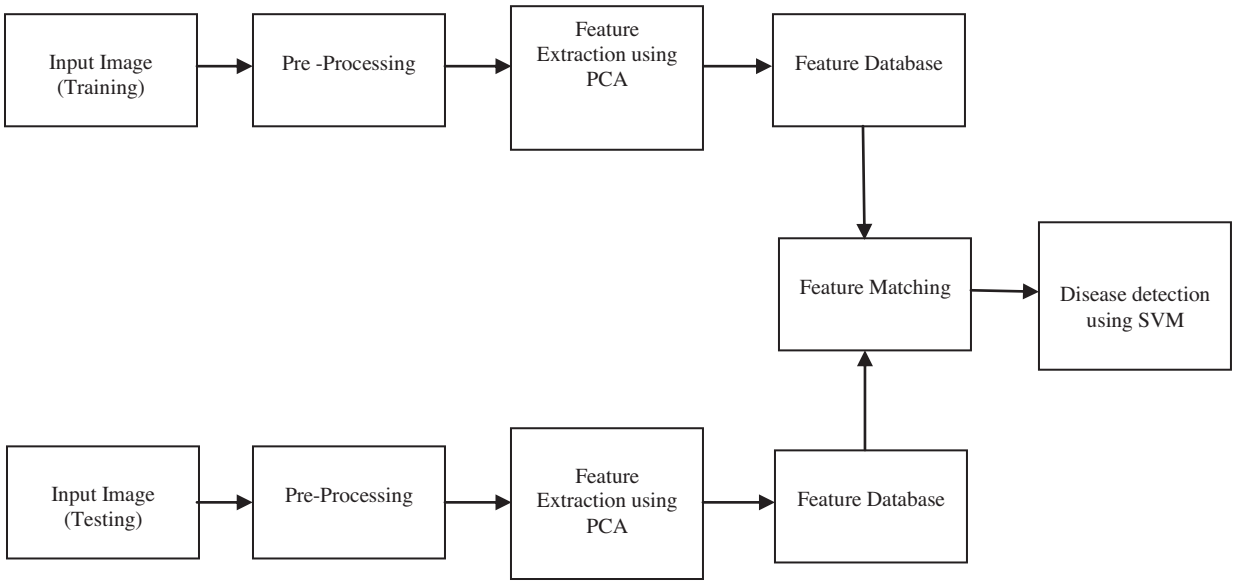

Fig. 1. Methodology diagram for detection of leaf diseases

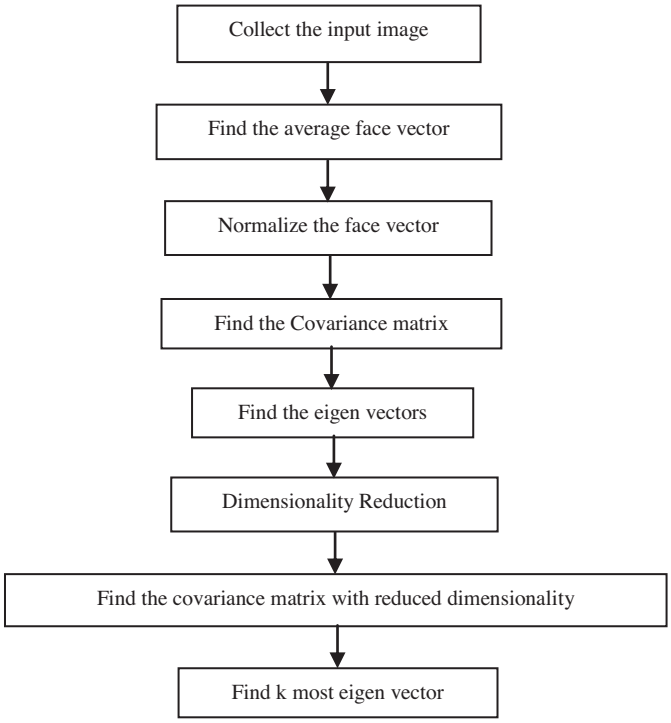

Fig 2. Flowchart of Principal Component Analysis

B. Feature Extraction Using PCA

The principal features extracted from the pre-processed leaf are major axis length, minor axis length, orientation, convex area, filled area, equid-diameter, solidity, colour, texture and size. Fig. 2 shows the flowchart of the working of PCA.

1) Calculation of Eigen vectors

Step 1: Acquire M training images $i_1, i_2, \ldots . i_m$,it is very important that the leaf images are centered.

Step 2: Represent to each image $i_1$as a vector, Γ_i, as the issues in developing matrix with numerous segments using formulae (eq.5),

$$i_i = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix} \dashrightarrow \begin{bmatrix} a_{11} \\ . \\ a_{NN} \end{bmatrix} \quad (5)$$

Step 3: Locate the normal face vector, ψ using formulae (eq. 6)

$$\psi = \frac{1}{M}\sum_{i=1}^{M} \Gamma_i \quad (6)$$

Step 4: The mean is detected from each face vector Γ_i to get a set of vectors ϕ (eq. 7).

$$\phi = \Gamma_i - \psi \quad (7)$$

Step 5: Covariance matrix C is calculated using (eq.8),

$$C = AA^T \text{ where } A = [\phi_1, \phi_2, \phi_M] \quad (8)$$

Step 6: The computation of covariance matrix with high dimensionality will makes the system to run out of memory. In order to provide solution to the problem, covariance

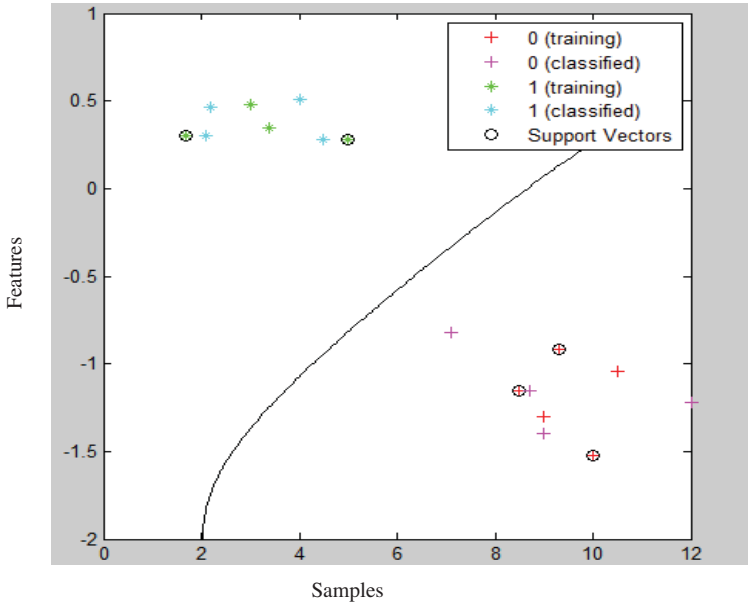

Fig. 3. Representation of the principal features such as texture, colour and orientation

matrix is calculated as **C**=$A^T A$.
Step 7: Calculate the M largest Eigenvectors of $A^T A$.
Step 8: Select the best K Eigenvectors.

2) Calculated Weights

Step 1: Linear combination, ϕ of the Eigenvectors u_i is (eq.9)

$$\phi_i = \sum_{j=1}^{K} w_j u_i \quad (9)$$

Step 2: The weight of each vector can be represented as (eq.10),

$$w_j = u_j^T \phi_i \quad (10)$$

Step 3: The normalized training is represented using (eq.11),

$$\Omega_i = \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ w_3 \end{bmatrix} \quad (11)$$

3) Recognization Task

If an unknown leaf image is to be recognized, then the steps are to be followed are (i) Normalize the incoming probe, (ii) Project the normalized probe to the Eigen space and find the weights and (iii) Classification task.

Fig. 4. (a) Input leaf image (b) Segmented image and (c) classified disease.

TABLE 1. CONFUSION MATRIX OF THE CLASSIFIED OUTPUT OF TOMATO LEAVES USING LSVM

Diseases	No of Sample	Early blight	Powdery mildew	No Disease	User's Accuracy %
Early blight	75	66	5	4	88
Powdery mildew	75	4	68	3	90.66
No Disease	150	10	8	132	88
Producers Accuracy %		82.5	83.95	94.96	-

C. *Disease Recognition Using SVM*

The given leaf image is recognized by using SVM classifier technique. Linear support vector machine is used because the classes are linearly separated. In LSVM classification, the graph (Fig. 3) represents the features extracted by the PCA algorithm in the corresponding axes. The graph is drawn on the basis of samples versus features and it clearly shows how the samples are classified.

The data points nearest to the hyper plane are called support vectors and only the support vectors are essential for classification where as other training vectors are ignored. Finally, the separated support vectors are matched with the trained images to identify the diseased leaf.

IV. RESULTS AND DISCUSSION

Experiments are carried out for tomato leaves using image processing techniques such as Principal Component Analysis and Linear Support Vector Machine. A dataset of 1000 samples were used in this work for training. PCA extracts the features after several pre-processing steps. The extracted features are provided to the Linear Support Vector Machine to identify the disease using the required dataset. The diseases classified for tomato leaves are the common bacterial disease early blight and powdery mildew. These diseases are the dangerous bacterial disease which restricts the cultivation of tomato in agriculture.

The collected sample image is processed before extracting the features (Fig. 4(a)). The image is transformed into grey image and the unnecessary noises are removed using Median filter and Gabor filter. It filters the noises and retrieves the base detail of the image. Histogram equalization was performed to increase the luminance of the pixels and calculate the luminance constant through which threshold segmentation is proceeded (Fig. 4(b)). The symptoms of disease is analyzed by using segmentation.

These processing steps are resulted to extract the features of the leaf image. The leaf image has several vectors with respect to the resolution. Several steps are performed to find the principal vectors and the weightage of each vector is calculated. The principal features extracted are textures, colours and orientation. Principal Component Analysis and Linear Support Vector Machine is one of the proven technique for feature extraction and cclassification respectively. In this work, we have hybridized these two techniques and the results are tabulated (Table I).

Wolf et al., used texture analysis methods such as Gabor filters and fractals for retrieving the image which is based on image transformation [12]. Shinde et al., used sobel filter and K-means clustering techniques for the soya bean leaf disease detection [13]. GangHou et al., conveys that in feature representation, the feature fusion mechanism merges the colour and shape to get a high retrieval performance [14]. However, these techniques have its own limitations. But, PCA extracts only the principal vectors which explain the entire data of the original image without any data loss.

Since the principal features are calculated linearly from the PCA technique, Linear SVM classifier is used in the classification of the disease that depends on the data set collected (Fig. 4(c)).The principal vectors such as texture, colours, orientations are plotted in the graph and the hyper plane is used in the separation of the vectors. The vectors near the hyper plane margins are termed to be support vector. The distance is calculated and if the threshold value is higher than the distance, the disease is classified by the dataset else, the disease is not detected and it requires further training. Rani et al., [15] demonstrated a modified SVM technique to classify the image similar to that of the sample image. In this system, the real world images are automatically processed to retrieve the exact information in real time basis by software perspective.

Table 1 shows the confusion matrix of the classified output of tomato leaves using LSVM. The users accuracy for powdery mildew and early blight are 90.66% and 88% respectively. The producers accuracy for powdery mildew and early blight are 83.95% and 82.5% respectively.

V. CONCLUSION

The proposed system provides a method for extracting the principal features of the tomato leaf listed as colour, texture, orientation using PCA technique. The high efficient recognition algorithm, Linear Support Vector Machine classifier detects the diseases in the leaf image. The overall accuracy is 88.67% and the kappa coefficient is 082. PCA technique provides high accuracy since the features are linearly obtained from the diseased leaf. Thus, linear SVM easily classifies the disease in plant leaves. Further, this algorithm can be practised for the future enhancement of detecting different diseases in plant leaves. In addition, this work is merged with hardware devices to identify the diseases in real time without training the images.

REFERENCES

[1] J. Yue, Z. Li, L. Liu and Z. Fu, "Content-based image retrieval using color and texture fused features", Mathematical and Computer Modelling, Vol. 54. No. 3-4, pp.1121-1127, 2011.

[2] A. W. Smeulders, M. Worring, S. Santini, S, A. Gupta, A. and R. Jain, "Content-based image retrieval at the end of the early years". IEEE Trans on Pattern Analysis & Machine Intelligence, Vol. 12, pp.1349-1380, 2000.

[3] J. K. Patil and R. Kumar, "Comparative analysis of content based image retrieval using texture features for plant leaf diseases". International Journal of Applied Engineering Research, Vol. 11, No. 9, pp.6244-6249, 2016.

[4] A. Sivasangari, K. Priya, "Cotton Leaf Disease Detection and Recovery Using Genetic Algorithm", International Journal of Engineering Research and General Science, Vol. 117, pp. 119-123, 2014.

[5] D. Prithviraj, “Leaf Infection Detection and Diagnosis using Image Processing”, International Journal of Applied Engineering Research, 2014.

[6] K.S. Kailey and G. S. Sahdra, “Content-based image retrieval (CBIR) for identifying image based plant disease”, International Journal of Computer Technology and Applications,Vol 3, No. 3, pp. 1099-1104 2012.

[7] K. V. Madhavi, R. Tamilkodi, R. B. Dinakar and K. JayaSudha, “An innovative technique for content based image retrieval using color and texture features”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, No. 5, pp.1257-1263, 2013.

[8] Y.Sanjana, AshwathSivasamy, “Plant Disease Detection using Image Processing”, International Journal of Engineering Research and General Science, Vol. 9, No. 4, pp. 107-108, 2015.

[9] Y. M. Oo, and C. Htun. "Plant Leaf Disease Detection and Classification using Image Processing.", International Journal of Research and Engineering, Vol. 5, No. 9, pp. 516-523, Oct 2018.

[10] K. Prema and Carmel Mary Belinda, “Smart Farming: IoT Based Plant Leaf Disease Detection and Prediction using Deep Neural Network with Image Processing”, International Journal of Innovative Technology and Exploring Engineering, Vol. 8, No. 9, pp.3081-3083, July, 2019

[11] D. Zhang, A. Wong, M. Indrawan, and G. Lu, “Content-based image retrieval using Gabor texture features”, IEEE Transactions Pami, 2000.

[12] C. Wolf, J. M. Jolion, W. Kropatsch and H. Bischof, “Content based image retrieval using interest points and texture features”, In Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, Vol. 4, pp. 234-237, 2000.

[13] C. Ravi Shinde, C. Jibu Mathew and Y. Patil, “Segmentation Technique for Soybean Leaves Disease Detection”. International Journal of Advanced Research, Vol. 5, No. 3, pp. 522-528 2015.

[14] GangHou and JunKong, “CBIR using Texture structure Histogram”, Journal of Network Communications and Emerging Technologies, pp. 1356-1363, 2013.

[15] D. Rani. and M. Goyal, “A Research Paper on Content-based Image Retrieval System using Improved SVM Technique”, International Journal of Engineering and Computer Science (IJECS), Vol. 3, No. 12, pp.9755-7760, 2014.