Diagnosis of Diseases on Cotton Leaves Using Principal Component Analysis Classifier

Viraj A. Gulhane
Dept. of Electronics and Telecommunication
Sipna College of Engineering and Tech., Amravati, India
virajgulhane@live.com

Maheshkumar H. Kolekar
Dept. of Electrical Engineering
Indian Institute of Technology, Patna, India
mahesh@iitp.ac.in

***Abstract*—This paper addresses the problem of diagnosis of diseases on cotton leaf using Principle Component Analysis (PCA), Nearest Neighbourhood Classifier (KNN). Cotton leaf data analysis aims to study the diseases pattern which are defined as any deterioration of normal physiological functions of plants, producing characteristic symptoms in terms of undesirable color changes mainly occurs upon leaves; caused by a pathogen, which may be any agent or deficiencies. The predictions of diseases on cotton leaves by human assistance may be wrong in some cases. Using machine vision techniques, it is possible to increase scope for detection of various diseases within visible as well invisible wavelength regions. After implementing PCA/KNN multi-variable techniques, it is possible to analyse the statistical data related to the Green (G) channel of RGB image. Green channel is taken into consideration for faithful feature collection since disease or deficiencies of elements are reflected well by green channel. In most of the cases diseases are seen on the leaves of the cotton plant such as Blight, Leaf Nacrosis, Gray Mildew, Alternaria, and Magnesium Deficiency. The classification accuracy of PCA/KNN based classifier observed is 95%.**

***Index Terms*—Principal Component Analysis, Nearest Neighbourhood Classifier, Cotton Leaves, Diagnosis.**

I. Introduction

The agriculture science plays vital role in day to day life, it is one of the most emerging field so it attracts the thousands of researchers who are actively engaged in studying plant pathology. Now a days there are many challenges like disease control, pest control, lack of production [1]. The lack of production is ultimate cause due to improper disease management and pest control. Using implementable machine visionary techniques, it is possible to detect the diseases in early stages, so ultimate goal may reach. As there are many challenges in finding out diseases manually because each disease show various similarities in patterns due to which it is far difficult to identify similar kind of diseases. The main cause is human visual system, which can't cover the invisible light spectrum, i.e. human visual spectrum is insufficient to recognize the diseases accurately, the range of human visual spectrum is about 390 to 700 nm (Λ). By applying appropriate computer aided machine vision based recognition techniques, it is easy to classify similar kind of diseases which creates a condition of dilemma for human eyes, so spectrum improves about 350 nm to 1000 nm (Λ) or as per requirements. The given spectrum target will be achieved by high resolution camera and its interpretation possible by machine vision algorithms such as Bayesian belief network [2], Hidden markov model [3]. In this paper the main emphasis is given to find out mostly occurring diseases on the cotton leaves. Occurrence of diseases on the cotton plant is reflected mainly by its leaves. In [4] [5] authors have used green color information for detecting field view. We are also using green channel information of RGB image for extracting features because cotton leaves show early symptom of diseases. By choosing appropriate classifier technique like PCA will provide best results to detect the various diseases on leaves of cotton in early stages. Hence, a disease detection using computer aid helps farmers for controlling the diseases. It will therefore ultimately increases the human-computer interaction, by improving the spectrum of observations. Principal Component Analysis is one of the most popular techniques in area of machine vision, having best statistical feature analysis capabilities. Thus PCA is useful for analysis of the given image database described by several sets of parameters. The main goal of PCA is to extract the important features, which may vary accordingly with respect to diseases. These features are considered as a set of orthogonal variables i.e. principal components. The PCA therefore includes (i) extraction of the most significant features from database; (ii) compress the size of the data set by keeping only this significant information; (iii) simplify the description of the data set. Principal components are expressed as linear combinations of the original variables.

II. Literature Review

Though diseases are mostly seen on the leaves of plant, precise quantification of these visually observed disease traits has not well studied yet because of the complexity of visual patterns. Hence there has been increasing demand for more specific and sophisticated image pattern understanding techniques. Computer aided detection of plant diseases is an important research topic as it may prove benefits in detecting diseases automatically from the symptoms that appear on the plant leaves in early stages. In the following context mentioning the various techniques which were used to recognize the disease patterns of the plant. Kim et. al. [6], who had classified the grape fruit peel diseases using color, texture features analysis.

978-1-4799-5364-6/14/$31.00 ©2014 IEEE

The texture features [7] are calculated from the Spatial Gray-level Dependence Matrices (SGDM) and the classification was done using the squared distance technique. Grape fruit peel was infected by several diseases like greasy spot, copper burn. A new approach was proposed for integrating image analysis technique into a diagnostic expert system [8] for three different disorders such as Leaf miner, Powdery and Downey. The proposed approach has greatly reduced error prone dialogue between system and user. The morphological features of leaves are used for plant classification and in the early diagnosis of certain plant diseases. Design and implementation of an artificial vision system which extracts specific geometric and morphological features from plant leaves are possible [9]. The proposed system consists of an artificial vision system (camera), a combination of image processing algorithms and feed forward neural network based classifier.

The weather based prediction models of plant diseases are proposed in [10] and the performance of conventional multiple regression, artificial neural network and Support Vector Machine (SVM) were compared. It was concluded that SVM based regression approach has led to a better description of their relationship between the environmental conditions and disease level which could be useful for disease management. It was proved that just a back-propagation network and shape of the leaf is used to identify the diseases. The Neural network approach [11] for segmentation of agricultural fields in remote sensing data is proposed. A neural network algorithm based on back propagation is used for segmentation of the color images in crop field infected with diseases that changes the usual color of plants.

A Probabilistic Neural Network (PNN) approach for plant leaf recognition was used by Wu et. al. [12], where features are processed by PCA to form input to PNN. It was found that this algorithm works with an accuracy of 90% on 32 kinds of plants. A system which introduces computer management into the cultivation process in the low-tech greenhouse was listed in [13]. A software prototype system for disease detection based on the infected image of various rice plants [14] is proposed where various image segmentation techniques are used to detect infected parts of the plants. In fast and accurate detection of diseases, the novel method was developed which was based on image processing for grading of plant disease [15].

The grape leaf disease was detected in form of the color imagery using hybrid intelligent system [16], where segmented image was filtered using Gabor wavelet for studying disease color of infected leaf. The support vector machines were used to classify types of grape leaf disease. Ying et. al. [17] studied methods of image preprocessing for recognition of crop diseases. They used cucumber powdery mildew, speckle & downy mildews as study samples and reported comparative study.

III. PCA Classifier

PCA is the linear regression method that permits one to recognize the uncorrelated components. In PCA method each feature in the ensemble features being considered as a random feature space. The features plotted in this space form a cloud of features. From the principal axes of this cloud of features, we defined a new coordinate system, where the segment of the original feature vectors along these axes are uncorrelated. The main axis is identified by the eigenvectors of the co-variance matrix of feature space. The scatter of corresponding feature components along each of the axes is given by the eigenvalue of the corresponding eigenvector. Eigenvector with largest eigenvalue are chosen.

Let us consider a data matrix $\mathbf{X}$ with zero empirical mean, where each of the n rows represents a different repetition of the experiment, and each of the p columns gives a particular kind of datum. Mathematically, the transformation is defined by a set of p-dimensional vectors of weights or loading $\mathbf{w}_{(n)} = (\omega_1, ..., \omega_p)_{(n)}$ that map each row vector $\mathbf{x}_{(i)}$ of $\mathbf{X}$ to a new new vector of principal component scores $\mathbf{t}_{(i)} = (t_1, ..., t_p)_{(i)}$, given by $t_{n(i)} = \mathbf{x}_{(i)}\mathbf{w}_{(n)}$ in such a way that the individual variables of t considered over the data set successively inherit the maximum possible variance from $\mathbf{x}$, with each loading vector $\mathbf{w}$ constrained to be a unit vector, thus first vector component $\mathbf{w}_{(1)}$ has to satisfy

$$\begin{aligned} \mathbf{w}_{(1)} &= arg_{||\mathbf{w}||=1} max \left\{ \sum_i (t_1)^2{}_{(i)} \right\} \\ &= arg_{||\mathbf{w}||=1} max \sum_i (\mathbf{x}_{(i)}\mathbf{w})^2 \end{aligned} \quad (1)$$

After writing above Eq 1 in matrix form will give,

$$\begin{aligned} \mathbf{w}_{(1)} &= arg_{||\mathbf{w}||=1} max \left\{ ||\mathbf{X}\ \mathbf{w}||^2 \right\} \\ &= arg_{||\mathbf{w}||=1} max \left\{ \mathbf{w}^T \mathbf{X}^T \mathbf{X}\ \mathbf{w} \right\} \end{aligned} \quad (2)$$

Since $\mathbf{w}_{(1)}$ has been defined to be a unit vector, it equally must also satisfy,

$$\mathbf{w}_{(1)} = arg\ max \left\{ \frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X}\ \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\} \quad (3)$$

Here, $\mathbf{X}^T\mathbf{X}$ is called as maximum possible eigenvalue of matrix.The score of first component is given as $t_{1(i)} = \mathbf{x}_{(i)}\mathbf{w}_{(1)}$ Estimating rest of the components, by subtracting $n-1$ principle components from $\mathbf{X}$, n^{th} component can be obtained (Eq 4) and maximum variance data matrix can be obtained (Eq 5).

$$\hat{\mathbf{X}}_{n-1} = \mathbf{X} - \sum_{s=1}^{n-1} \mathbf{X}\ \mathbf{w}_{(s)} \mathbf{w}_{(s)}^T \quad (4)$$

$$\begin{aligned} \mathbf{w}_{(k)} &= arg_{||\mathbf{w}||=1} max \left\{ ||\hat{\mathbf{X}}_{n-1}\mathbf{w}||^2 \right\} \\ &= arg\ max \left\{ \frac{\mathbf{w}^T \hat{\mathbf{X}}_{n-1}^T \hat{\mathbf{X}}_{n-1} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\} \end{aligned} \quad (5)$$

From the principal component, n^{th} component will provide the score as $t_{n(i)} = \mathbf{x}_{(i)}\ \mathbf{w}_{(n)}$, and here $\mathbf{w}_{(n)}$ is the n^{th} eigenvector of matrix $\mathbf{X}^T\mathbf{X}$. The decomposition matrix for the n principal components is given by

$$\mathbf{T} = \mathbf{XW} \tag{6}$$

where $\mathbf{W}$ denotes p-by-p matrix and its columns represents the eigenvectors of $\mathbf{X}^T\mathbf{X}$.

In PCA, larger number of features can make each feature less meaningful. Therefore, we have considered features like variance, standard deviation and mean. Basically the standard deviation is widely used to measure of variability or diversity used for selected block. The term variance is a measure of how far a set of pixel value is spread out. It is one of several descriptors of probability distribution, describing how far the pixel is lies from mean value or expected value. Extracting the features of the data is one of the important task, to attain these tasks, input image pre-processed and partitioned into 20x20 blocks as shown in Fig. 1. The mean and variance of the corresponding to each block will be appended in a column vector. This column vector is nothing but texture feature. Co-variance matrix is formulated by calculating variance for 2-D set using Eq 7.

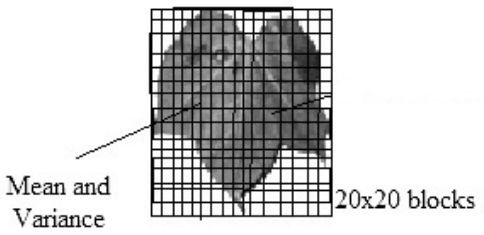

Fig. 1: Partitioning of cotton leaf image into 20×20 blocks.

By obtaining the co-variance matrix, the next step includes the calculation of corresponding scatter matrix (Eq 7),

$$\begin{aligned} S &= \sum_{j=1}^{n}(X_j - \hat{X})(X_j - \hat{X})^T \\ &= \sum_{j=1}^{n}(X_j - \hat{X}) \otimes (X_j - \hat{X})^T \\ &= \sum_{j=1}^{n} X_j {X_j}^T - n\ \hat{X}(\hat{X})^T \\ \therefore S_{m\times m} &= X\ C_n X^T \end{aligned} \tag{7}$$

Where X_j signifies j^{th} column of X, and C_n is called as centering matrix. We are getting the eigenvalues from the diagonal of scatter matrix and corresponding eigenvectors W. Here matrix S is of dimension $m \times m$, which is positively valued and partial definite in nature. In order to obtain a desired estimate for the co-variance matrix for the given feature, we must use eigenvalues matrix W obtained from Eq 7. Hence formulating the principal components for the given scatter matrix as

$$\begin{aligned} Q(P\ C_{(j)}, P\ C_{(k)}) &\propto (XW_{(j)}) \times (XW_{(k)}) \\ &= W_{(j)}^T X^T X\ W_{(k)} \\ &= W_{(j)}^T \lambda_{(k)} W_{(k)} \end{aligned} \tag{8}$$

Where $W_{(j)}$ and $W_{(k)}$ represents eigenvalues. Fig. 2 shows plot of eigenvalue Vs. corresponding Eigenvectors in descending order.

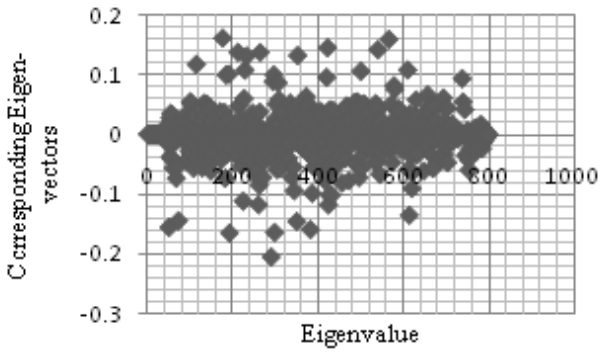

Fig. 2: Plot of Eigenvalue Vs. Eigenvectors.

IV. KNN Classifier

In KNN method, partitioning n observation into k clusters and we get $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ set of observations of dimension d, now partitioning those n observations into k sets, given as, $(k \leq n)$ $\mathbf{S} = \{S_1, S_2, ..., S_k\}$, also, minimizing cluster size (Eq 9), where mean of points in S_i is denoted by μ_i.

$$arg\ \mathbf{S}\ min \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} ||\mathbf{x}_j - \mu_i||^2 \tag{9}$$

Leaves data was classified into the required clusters, where maximum cosine distance represents the possible match between test feature matrix and train feature matrix. The cosine distance is given by:

$$\cos(\theta) = \frac{n(\mathbf{x} \cap S)}{\sqrt{n(\mathbf{x}) \times n(S)}} \tag{10}$$

where $\mathbf{x}$ is set of observations and S is feature space.

V. Result and Discussion

In order to acquire maximum possible features from 110 samples, we considered green channel approach because entire color of the leaf is decided by itself. The normal cotton leaves are basically have the green color pigments which is the dominant color pattern (Fig 3). In disease condition a non uniform patterns appear on the green channel, which would be indication of disease present on the leaf. We can say that the combination of different abnormal color patterns lead to

different diseases (Fig 4, 5, 6, 7).

Fig. 3: Normal (non-diseased) leaves.

Fig.4: Leaf Spot (Blight) on leaves.

Fig. 5: Magnesium Deficiency condition.

Fig. 6: Leaf Nacrosis Disease on leaves.

Fig. 7: Gray Mildew Disease on leaves.

In Fig. 3, leaves were assumed to be non-diseased $(normal)$ because there was no drastic color changes occurred on green channel. As the leaves were infected by the pathogens, the green channel of the leaves had shown color changes (Fig. 4). Similarly, the drastic pigmentation change is shown in the Fig. 5, the diseases called as $Lalya$ in local language, scientific name is $MagnesiumDeficiency$ disease, where the entire green channel is red and green shade. Similarly, Fig. 6 & 7 shows the $Leaf\ Narcosis$, $Gray\ Mildew$ diseases. Suppose if we consider diseases like $leaf\ spots\ or\ Blight$ & $Magnesium\ Deficiency$, as shown in Fig. 4 & Fig. 5 respectively. In those cases, there is not much variation in the color pattern and which is not easily recognizable to the human eye, now if $Lalya$ disease is in its early phase, so in this case it will be more difficult to predict, also if same disease is in its final stage, again it will be difficult to differentiate it from $LeafNacrosis$ (Fig. 6). After applying the PCA/KNN classifier technique the following diseases were well recognized with recognition accuracy 95%, which was more than human eye (Table 1). Here, Fig. 8 denotes the curve between the frequency of diseases (%) occurring over the 110 test samples. Accuracy of PCA/KNN classifier is expressed (Eq 11) by considering true positive samples in each case.

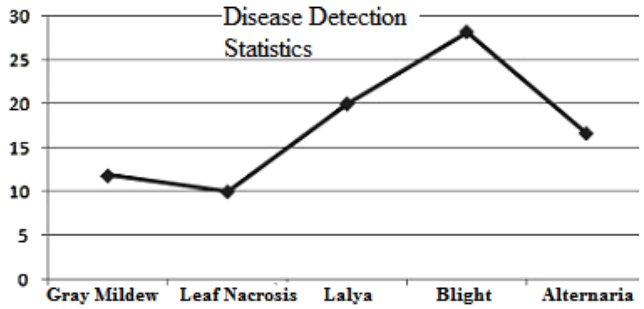

Fig. 8: Frequency of detected diseases over cotton leaves [%]

Table 1: Recognition rate (%) of PCA/KNN classifier:

Sr. No.	Name of the disease.	Classification Accuracy %	
		Manual	PCA/ KNN
1	Leaf Spot / Blight	76	94
2	Magnesium Deficiency	89	92
3	Leaf Nacrosis	70	96
4	Gray Mildew	74	97
5	Alternaria	78	94
6	Non Disease	98	97
Overall (%) Classification Accuracy		81	95

$$Accuracy = \frac{True\ Positive\ Samples}{Total\ No\ of\ Samples} \tag{11}$$

VI. Conclusion and Future Scope

In this paper, we have presented PCA/KNN classifier for faithful detection of diseases on cotton leaves. The frequency of detected diseases on cotton leaves are shown in Fig. 8. Cotton leaf disease Blight has more frequency $i.e.$ 28%. It was found that similar pattern diseases are having more cosines distances during KNN classification due to which there will be chance of mis-classification, $i.e.$ some diseases are having similarities in their color patterns due to which disease patterns are not well recognized. The overall disease recognition accuracy analysis as mentioned in Table 1 is about 95%, which is 14% more than that of manual observations. It was observed that recognizing the disease on cotton leaf is sometimes tedious task because photosynthesis process mostly hamper recognition rate of real time leaf disease recognition system. In future, we will design more robust classifier considering features like texture [18], leaf shape.

References

[1] K. Kranthi, D. Jadhav, S. Kranthi, R. Wanjari, S. Ali, and D. Russell, "Insecticide resistance in five major insect pests of cotton in india," *Crop Protection*, vol. 21, no. 6, pp. 449–460, 2002.

[2] M. H. Kolekar, "Bayesian belief network based broadcast sports video indexing," *Multimedia Tools and Applications*, vol. 54, no. 1, pp. 27–54, 2011.

[3] M. H. Kolekar and S. Sengupta, "Hidden markov model based video indexing with discrete cosine transform as a likelihood function," *IEEE INDICON Conference*, pp. 157–159, 2004.

[4] M. H. Kolekar and K. Palaniappan, "A hierarchical framework for semantic scene classification in soccer sports video," *IEEE Region Ten (TENCON) Int. Conference*, 2008.

[5] M. H. Kolekar, K. Palaniappan, S. Sengupta, and G. Seetharaman, "Semantic concept mining based on hierarchical event detection for soccer video indexing," *Journal of multimedia*, vol. 4, no. 5, pp. 298–312, 2009.

[6] D. G. Kim, T. F. Burks, J. Qin, and D. M. Bulanon, "Classification of grapefruit peel diseases using color texture feature analysis," *International Journal of Agricultural and Biological Engineering*, vol. 2, no. 3, pp. 41–50, 2009.

[7] M. H. Kolekar, S. Talbar, and T. Sontakke, "Texture segmentation using fractal signature," *IETE J RES*, vol. 46, no. 5, pp. 319–323, 2000.

[8] M. Ei-Helly, A. Rafea, S. Ei-Gamal, and R. A. E. Whab, "Integrating diagnostic expert system with image processing via loosely coupled technique," *Central Laboratory for Agricultural Expert System*, 2004.

[9] P. Tzionas, S. E. Papadakis, and D. Manolakis, "Plant leaves classification based on morphological features and a fuzzy surface selection technique," *Fifth International Conference on Technology and Automation, Thessaloniki, Greece*, pp. 365–370, 2002.

[10] R. Kaundal, A. S. Kapoor, and G. P. Raghava, "Machine learning techniques in disease forecasting: a case study on rice blast prediction," *BMC bioinformatics*, vol. 7, no. 1, p. 485, 2006.

[11] B. J. Woodford, N. K. Kasabov, and C. H. Wearing, "Fruit image analysis using wavelets," *Proceedings of the ICONIP/ANZIIS/ANNES*, vol. 99, pp. 88–91, 1999.

[12] S. G. Wu, F. S. Bao, E. Y. Xu, Y.-X. Wang, Y.-F. Chang, and Q.-L. Xiang, "A leaf recognition algorithm for plant classification using probabilistic neural network," *Signal Processing and Information Technology IEEE International Symposium*, pp. 11–16, 2007.

[13] M. Maliappis, K. Ferentinos, H. Passam, and A. Sideridis, "Gims: A web based greenhouse intelligent management system," *World Journal of Agricultural Sciences*, vol. 4, no. 5, pp. 640–647, 2008.

[14] S. Phadikar and J. Sil, "Rice disease identification using pattern recognition techniques," *Computer and Information Technology, ICCIT 11th International Conference*, 2008.

[15] S. Weizheng, W. Yachun, C. Zhanliang, and W. Hongda, "Grading method of leaf spot disease based on image processing," *IEEE International Conference on Computer Science and Software Engineering*, vol. 6, pp. 491–494, 2008.

[16] A. K. Meunkaewjinda A., Kumsawat P. and A. Srikaew, "Grape leaf disease detection from color imagery using hybrid intelligent system," *5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, vol. 1, pp. 513–516, 2008.

[17] G. Ying, L. Miao, Y. Yuan, and H. Zelin, "A study on the method of image pre-processing for recognition of crop diseases," *ICACC'09. International Conference*, pp. 202–206, 2009.

[18] M. H. Kolekar, "An algorithm for designing optimal gabor filter for segmenting multi-textured images," *IETE journal of research*, vol. 48, no. 3-4, pp. 181–187, 2002.