



Volume 2, Issue 1, January 2012

ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Impact of Principal Component Analysis in the Application of Image Processing

Abhishek Banerjee

Information Technology Department, Pailan College Of Management Of Technology ,Pailan, Joka, Kolkata-700104, West Bengal, India

Abstract—Principal component analysis (PCA) is a classical statistical method. It is based on the statistical representation of a random variable. This linear transform has been widely used in data analysis and compression. Impact of PCA is affecting the research work in now a days in the various field like application of Image Processing, pattern recognition ,Neural network and etc . We know that rice is one of the most widely cultivated food crops throughout the world. Demand for rice as a major food item continues to increase and it is estimated that we will have to produce 50% more rice by the year 2025[2]. The several diseases most of them are caused by the bacteria, fungus, virus, and parasite etc affect Rice plants. Diseases affect all the parts of the rice plants including the grain and the root, but mainly in aerial part of the plant i.e. stem and leave. Due to the damages by diseases and pest a large percentages of the production gets lost. Only way to prevent this loss is to timely diagnosis of the field problem and to take the appropriate measure. The diagnosis of the field problems is done manually which may causes improper diagnosis and may not be timely. Thus now a day's using the image processing and soft computing techniques some work has been started to automatically diagnosis the field problem [7]. One of the most important part of this automatic diagnosis processes is to identify the location of the damage caused by the pest or diseases. Thus in my work we have tried to classify the rice leaf and the stem images with the help of image processing techniques. In my work for this reason I have used the gray, green, hue and intensity distribution of the images vertically to the strip of the leaf and the stem at hundred-pixel interval as a feature vector. Among them I have accepted the result of intensity distribution. Because we know that the leaf surfaces are flat the distribution of intensity will be same along the direction but in case of the stem, the intensity will be increase towards the center and again decrease to the center to the boundary, as they are cylindrical in shape. Then I have applied the principal component analysis (PCA) to reduce the dimension of the feature vector to 7×1 . We have used the Baye's Classifier for the classification process with the accuracy of the 70% for leaf and 65% for stem, which is acceptable for the first try.

Keywords— Rice Leaf, Rice Stem, Segmentation, intensity Distribution, Principal Component Analysis, Bayes Classifier

I. INTRODUCTION

We know that stem and leaves are important parts of a rice plant. Though it is very simple to identify the stem or the leaf by a human as they see in the 3-D space but it becomes difficult to classify the stem and the leaf of a rice plant by an automated system that uses the 2-D images of the stem and the leave. In order to detect these diseases automatically by means of image processing and soft computing techniques first step is to detect the location of the disease from the image. For doing this project we have considered various types of material about various types of diseases mainly Brown Spot, Blast, Sheath Blight & Sheath Rot etc.

Brown Spot:- Main symptoms of the disease occur on leaves, glumes of maturing plants and young seedlings and the panicle branches in older plants. The smaller spots are dark brown to reddish brown, and the larger spots have a dark brown margin and reddish brown to gray centers and the spots are circular to oval in shape.

Blast:- Main symptoms of the rice blast include lesions that can be found on all parts of the plant, including leaves, leaf collars, necks, panicles, pedicels, and seeds. Generally the spots have brown/reddish- brown margin and gray-white centers and diamond in shape. Sometimes the shapes are circular, elliptical or spindle shaped. Diamond shaped spots are in 1-1.5cm long with 0.3-0.5 cm width.

Sheath Blight:- Mainly Initial sheath blight symptoms usually occur as water-soaked lesions on the first leaf sheath at or near the water line. The lesion darkens and begins girdling the sheath. In young stage the spots are in green gray and oval to elliptical in shape and in developed stage it is irregular tan to brown in color and the size of the spot in ¼ inch wide and ½ to ¾ inch long.

Sheath Rot:- Infection occurs on the uppermost leaf sheath enclosing the young panicles at late booting stage. Initial symptoms are oblong or somewhat irregular spots or lesions, 0.5-1.5 cm long, with dark reddish brown margins and gray center. Lesions may also consist of diffuse reddish brown discoloration in the sheath.

Basic Observation: -

- ❖ The surface of leaf belt is flat where stem is cylindrical.
- ❖ There exist strips in the leaf and stem, which are parallel to the length of the object.
- ❖ The strips are blackish in color.

By depending on these observations we found that the intensity distribution of the images perpendicular to the strip may be used as a feature for classification. Because the leaf surfaces are flat the distribution of intensity will be same along the direction but in case of the stem, the intensity will be increase towards the center and again decrease to the center to the boundary, as they are cylindrical in shape.

This paper has been divided into five sections. Section II reviews the previous work while Section III describes the Design Procedure of the work. Section IV discusses the results of the work and section E arrives at conclusions.

II. PAPER REVIEW SECTION

The following discussion is related with some journal papers, which is connected with this work. We know that the rice plants are infected by blast and the spectral characteristics curves draw the percentage of infection. For this reason many techniques have been introduced in now a days like [2] an image analysis based techniques for detection of possible changes in rice fields for mineral deficiency. Minerals are: - Boron, Iron, Magnesium, Manganese, Nitrogen and Potassium. The observation and examination of the applicability of broadband high- spatial-resolution ADAR (airborne data acquisition and registration) remote sensing data in visible and near infrared regions for rice disease detection is described in Paper [3]. We also know that 3-D curvature information of any free form surface encoded into a 2-D image corresponding to a certain point on the surface [4] but now a day's many ideas like geometric recognition algorithm was developed to identify molecular surface [5] and another approach [6] investigates surface recognition using "one-dimensional" data, specifically

points sampled along three concurrent curves on the surface of an object. But now a day's using the image processing and soft computing techniques some work has been started to automatically diagnosis the field problem [7]. Principal Component Analysis (PCA) method is a mathematical procedure that extracts relevant information from a large data set [12][13]. Karl Pearson invented this method in 1901. I have previously said Principal component analysis (PCA) is a classical statistical method. This linear transform has been widely used in data analysis and compression. Principal component analysis is based on the statistical representation of a random variable. PCA involves the calculation of the eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute. It is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way which best explains the variance in the data. There are some steps for implementing Principal Component Analysis. They are:-

Step-1-Take an original data set and calculate mean of the data set

Taking $\mathbf{X}_1 \cdot \cdot \cdot \mathbf{X}_N$ as column vectors, each of which has M rows. Place the column vectors into a single matrix \mathbf{X} of dimensions $M \times N$.

Step-2-Subtract off the mean for each dimension.

Find the empirical mean along each dimension $m = 1, \dots, M$ of each column. Place the calculated mean values into an empirical mean vector \mathbf{u} of dimensions $M \times 1$.

$$\mathbf{u} [m] = (1/N) \sum_{n=1}^N X [m, n]$$

Mean subtraction is an integral part of the solution towards finding a principal component basis that minimizes the mean square error of approximating the data. There are two steps: -

1. Subtract the empirical mean vector \mathbf{u} from each column of the data matrix \mathbf{X} .
2. Store mean-subtracted data in the $M \times N$ matrix \mathbf{B} .

$$\mathbf{B} = \mathbf{X} - \mathbf{u}\mathbf{h}$$

[Where \mathbf{h} is a $1 \times N$ row vector of all 1's : $h[n]=1$ for $n=1 \dots N$]

Step-3-Calculate the covariance matrix

Find the $M \times M$ empirical covariance matrix \mathbf{C} from the outer product of matrix \mathbf{B} with itself: -

$$\mathbf{C} = E [\mathbf{B} \times \mathbf{B}] = E [\mathbf{B} \cdot \mathbf{B}^*] = (1/N) \sum \mathbf{B} \cdot \mathbf{B}^*$$

Where E is the expected value operator, \times is the outer product operator, and $*$ is the conjugate transpose operator. Note that if \mathbf{B} consists entirely of real numbers, which is the case in

many applications, the "conjugate transpose" is the same as the regular transpose.

Step-4-Calculate eigen vector and eigen value of the covariance matrix

Compute the matrix **V** of eigenvectors which diagonalizes the covariance matrix **C**:

$$V^{-1}CV=D$$

where **D** is the diagonal matrix of eigenvalues of **C**. This step will typically involve the use of a computer-based algorithm for computing eigenvectors and eigenvalues.

Step-5-Extract diagonal of matrix as vector: -

Matrix **D** will take the form of an $M \times M$ diagonal matrix, where

$$D[p,q]=\lambda_m \text{ for } p=q=m$$

is the m th eigenvalue of the covariance matrix **C**, and

$$D[p,q]=0 \text{ for } p \neq q$$

Matrix **V**, also of dimension $M \times M$, contains M column vectors, each of length M , which represent the M eigenvectors of the covariance matrix **C**.

Step-6-Sorting in variance in decreasing order

Sort the column of the eigen vector matrix **V** and eigen value matrix **D** in order of decreasing eigen value

Step-7-Choosing components and forming a feature vector

Here is where the notion of data compression and reduced dimensionality comes into it. If we look at the eigenvectors and eigenvalues from the previous section, we will observe that the eigenvalues are quite different values. In fact, it turns out that the eigenvector with the highest eigenvalue is the principle component of the data set. What needs to be done now is you need to form a feature vector. Taking the eigenvectors that we want to keep from the list of eigenvectors, and forming a matrix with these eigenvectors in the columns construct this: -

Feature Vector = (eig₁ eig₂ eig₃.....eig_n)

Step-8-Deriving the new data set

This is the final step in PCA and it is also the easiest. Once we have chosen the components (eigenvectors) that we wish to keep in our data and formed a feature vector, we simply take the transpose of the vector and multiply it on the left of the original data set, transposed.

Final Data=Row Feature Vector*Row Data Adjust

where Row Feature Vector is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvectors at the top, and Row Data Adjust is the mean-adjusted data transposed, i.e. the data items are in each column, with each row holding a separate dimension.

III. DESIGN PROCEDURE OF THE WORK

To make this project easier the total work should be divided into five steps.

III.a. Image acquisition and preprocessing:

Acquisition could be as simple as being given an image that is already in digital form. For our work we have collected some stem or leaf images. For that particular reason we are taking help of Adobe Photoshop. By the help of crop tool of Adobe Photoshop we are cropping the required image from the leaf or stem image.

III.b. Segmentation of object from the image:

We know that segmentation is a process that partitions a digital image into multiple regions. A number of general-purpose algorithm and techniques have been developed for image segmentation. For our work Edge detection techniques have been used for segmentation. We have used edge detection techniques for getting edges because we will identify the direction of the strip In edge detection technique we are using prewitt and sobel method.

- ❖ Here we are using the mask along x-axis that is gradient of x. We know that if the first mask were moved around an image it would respond horizontally.

-1	-1	-1
0	0	0
1	1	1

- ❖ Again using masking for converting the thick edge into thin edges along x-axis. With a constant background, the maximum response would result when the line passed through the middle row of the mask.

-2	-1	-2
3	3	3
-2	-1	-2

- ❖ Again using masking for getting edges along y-axis that is gradient of y.

$$\theta = \tan^{-1} (Gy/Gx)$$

Where Gy=Gradient of y and Gx=Gradient of x.

-1	0	1
-1	0	1
-1	0	1

III.c By identifying the direction of the strip we determine to rotate the object to make the strip parallel to the length of the object

Here we will discuss that how should we find out the direction of the strip? In the previous section we detect the gradient of x and gradient of y. Then we are dividing the gradient of y to gradient of x and store in the variable called **theta**. Now we detect the angle of every pixel of the said variable. After detection of the angle of every pixel I was forming two groups. One for 0 degrees and another is for 90 degrees. Counting the maximum no of pixel in the each group. If the maximum no of pixel in a particular image is 0 degrees than the maximum no of pixel in a particular image is 90 degrees then the edge direction is horizontal else that is vertical.

III.d Determine the intensity distribution parallel to the y-axis

In this section we take the intensity level distribution along y-axis in column wise. We are taking intensity level distribution around one hundred-pixel interval. Now plotting the required result. After plotting the graph we are getting the zigzag line or curve. Then we are reducing the dimension of intensity level distribution using principal component analysis method. We have applied the principal component analysis (PCA) to reduce the dimension of the feature vector to 7*1.

III.e. Classify the stem and leaves based on the vectors representing intensity distribution parallel to the y-axis: -

We know that the leaf surfaces are flat the distribution of intensity will be same along the direction but in case of the stem, the intensity will be increase towards the center and again decrease to the center to the boundary, as they are cylindrical in shape. For classification purpose we are using Baye’s Classifier.

About Baye’s Classifier: -

Baye’s classifier determines the belongingness of a pattern in a particular class depending on the likelihood of the pattern with that class. The probability that pattern x belongs to class ω_i is denoted by p (ω_i/x). Assume pattern x actually belongs to class ω_i but the classifier detects that x comes from class ω_j then it incurs loss equal to L_{ij}. Since the pattern x may belong to any of the M classes under consideration, the expected loss incurred to assigning observation on x to class ω_j is formulated in equation (i).

$$r_j(x) = \sum_{i=1}^M L_{ij} p(\omega_i/x) \dots \dots \dots (i)$$

This is often referred to as the conditional average risk in decision theory terminology. The classifier has M possible categories to classify each pattern. Equation (i) computes the quantities r₁(x), r₂(x),..... R_M(x), for each x, and assign each pattern to the class within the smallest conditional loss, which produces total expected loss with respect to all decision is minimum. The classifier, which minimizes the total expected loss called the Bayes’ classifier. From the statistical point of view that Bayes’ classifier represents the optimum measure to performance using Bayes’ formula given in equation (ii)

$$p(\omega_i/x) = \frac{p(\omega_i)p(x/\omega_i)}{p(x)} \dots \dots (ii) \text{ is rewritten in equation (iii)}$$

$$r_j(x) = \frac{1}{p(x)} \sum_{i=1}^M L_{ij} p(\omega_i)p(x/\omega_i) \dots \dots \dots (iii)$$

Where p (x/ω_i) is called the likelihood function of class ω_i. Since 1/p(x) is a common factor so dropped from equation (iii) and the expression of average loss then reduce to

$$r_j(x) = \sum_{i=1}^M L_{ij} p(x/\omega_i) p(\omega_i) \dots \dots \dots (iv)$$

Therefore, a pattern x is assigned to a class ω_i if r_i(x) < r_j(x) for j=1,2,3,.....M, j≠i. However, a special type of loss function has been used which computes loss to zero for the correct decision and same for all erroneous decisions. The loss function is expressed as

$$L_{ij} = 1 - \delta_{ij} \dots \dots \dots (v)$$

Where δ_{ij}=1 when i=j and δ_{ij}=0 when j≠i. If we substitute the loss function of the equation (xii) using the equation (xiii), it becomes as follows:

$$r_j(x) = \sum_{i=1}^M (1 - \delta_{ij}) p(x/\omega_i) p(\omega_i) = p(x) - p(x/\omega_j) p(\omega_j) \dots \dots \dots (vi)$$

Therefore the Bayes’ classifier assign a particular pattern x to class ω_i if

$$p(x) - p(x/\omega_i) p(\omega_i) < p(x) - p(x/\omega_j) p(\omega_j) \text{ Or } p(x/\omega_i) p(\omega_i) > p(x/\omega_j) p(\omega_j), j=1,2,3, \dots \dots M; j \neq i \dots \dots \dots (vii)$$

Now consider M pattern classes governed by the multivariate normal density functions known as likelihood function. Each density is completely specified its mean vector m_i and covariance matrix C_i, which is defined as:

$$p(x/\omega_i) = \frac{1}{(2\pi)^{n/2} |C_i|^{1/2}} \exp [-1/2(x-m_i)^t C_i^{-1} (x-m_i)]; \dots \dots \dots (viii)$$

$$\text{Where } i=1,2, \dots \dots M; \dots \dots m_i = E_i \{x\} \dots \dots \dots (ix)$$

$$C_i = E_i \{ (x-m_i)(x-m_i)^t \} \dots \dots \dots (x)$$

Where E_i { . } denotes the expectation operator over the pattern of class ω_i. In equation (viii), n is the dimensionality of the pattern vectors and |C_i| indicates the determinant of the covariance matrix C_i. The mean vector is defined as

$$MN = E \{x\} = \int xp(x) dx \dots \dots \dots (xi)$$

Where x = { x₁, x₂,..... x_n } and MN = { m₁, m₂,..... m_n };

The expected value is approximately average and the corresponding mean vector is given in equation (xii)

$$MN=E \{x\} \approx \frac{1}{N} \sum_{j=1}^N x_j \dots\dots\dots(xii)$$

Where N is the no of samples. The covariance is given by

$$C = \begin{pmatrix} C_{11} & C_{12} & \dots & C_{1n} \\ C_{21} & C_{22} & \dots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \dots & C_{nn} \end{pmatrix}$$

$$C_{1k} = E_i \{(x_1 - m_1)(x_k - m_k)\}$$

$$= \int_{-\alpha}^{\alpha} \int_{-\alpha}^{\alpha} (x_1 - m_1)(x_k - m_k) p(x_1, x_k) dx_1 dx_k \dots\dots\dots(ix)$$

$$\begin{aligned} C &= E \{(x-m)(x-m)'\} \\ &= \{xx' - 2xm' + mm'\} \\ &= E \{xx'\} - mm' \dots\dots\dots(xiv) \end{aligned}$$

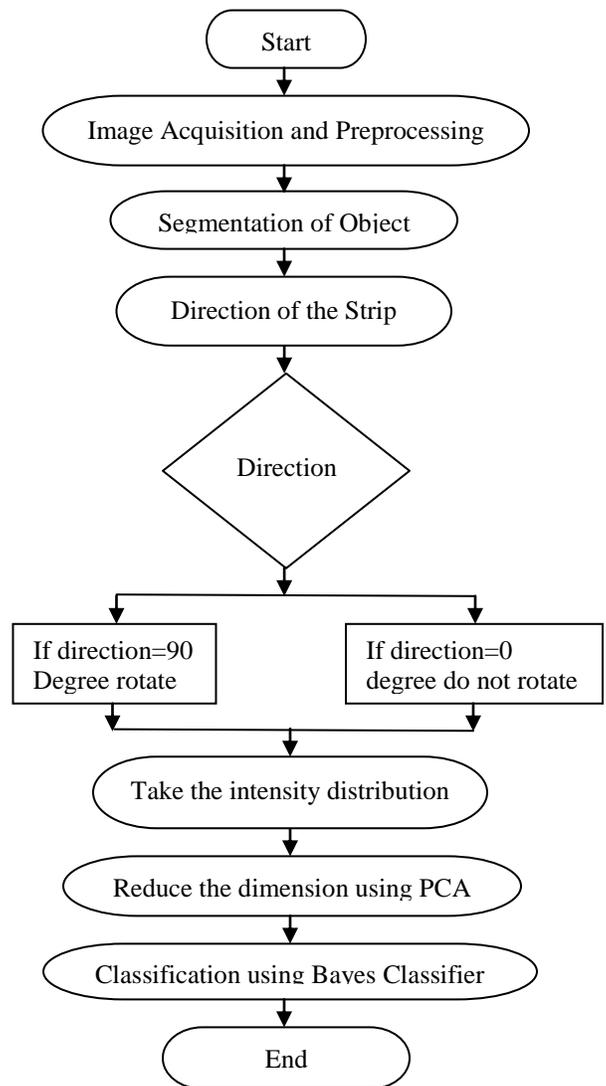
Approximating the expected value by the sample average

$$C \approx \frac{1}{N} \sum_{j=1}^N x_j x_j' - mm' \dots\dots\dots(xv)$$

After identifying stem or leaves of rice plants colour distribution has been done to examine whether the leaves are spot or not. At first minimum pixel value as well as maximum pixel value has been calculated. After that difference has been calculated. If the difference is minimum then the leaves or stem is cleared but when the difference is maximum then there is spot and colour distribution is maximum.

Here two classes are considered i.e. M=2 (leaf and Stem) and the features are considered by depending on the result of PCA.

III. A. FLOW CHART OF THE DESIGN



IV. RESULTS OF THE PROJECT WORK

IV.A.IMAGE ACQUISITION AND PREPROCESSING.

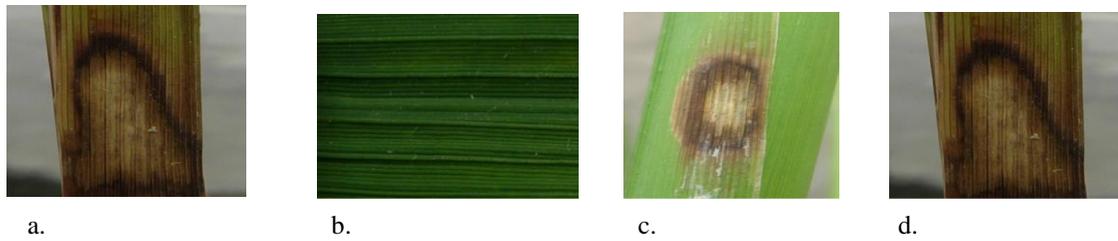


Fig1: Showing the acquired image.

IV.B. SEGMENTATION OF OBJECT FROM THE IMAGE

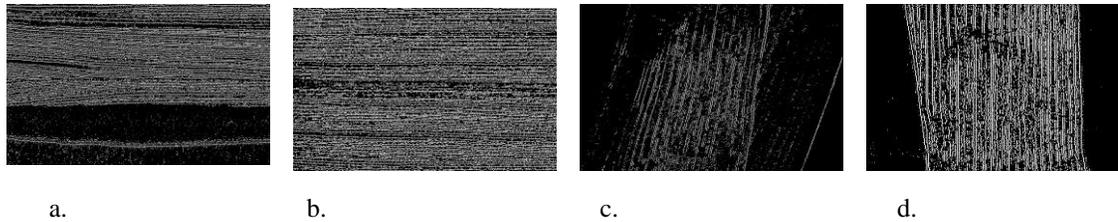


Fig.2. Showing the image after segmentation of Fig1.

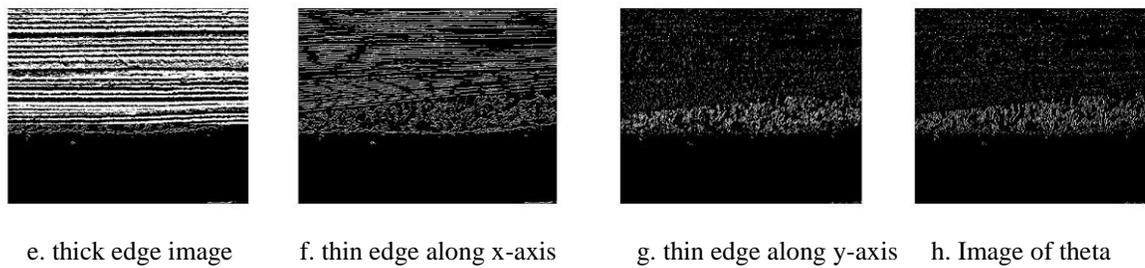


Fig: 3: -Showing the different stage of masking for figure1.a

IV.C. BY IDENTIFYING THE DIRECTION OF THE STRIP WE DETERMINE TO ROTATE THE OBJECT TO MAKE THE STRIP PARALLEL TO THE LENGTH OF THE OBJECT.



Fig: 4: -image showing the angle

IV.D. DETERMINE THE INTENSITY DISTRIBUTION PARALLEL TO THE Y-AXIS.

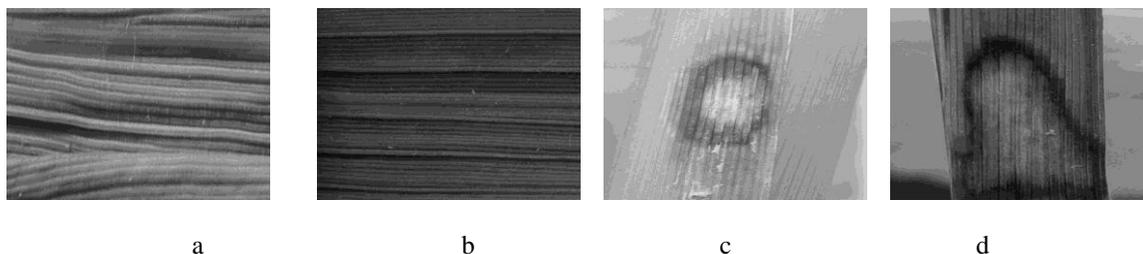


Fig: 5: -image showing after taking intensity distribution

We have used the gray, green, hue and intensity distribution of the images vertically to the strip of the leaf and the stem at **hundred-pixel interval** as a feature vector. Among them we have accepted the result of intensity distribution. Because we know that the leaf surfaces are flat the distribution of intensity will be same along the direction but in case of the stem, the intensity will be increase towards the center and again decrease to the center to the boundary, as they are cylindrical in shape. Then we have applied the principal component analysis (PCA) to reduce the dimension of the feature vector to 7*1.

IV.E.CLASSIFY THE STEM AND LEAVES BASED ON THE VECTORS REPRESENTING INTENSITY DISTRIBUTION PARALLEL TO THE Y-AXIS.

For classification purpose we are using Bayes Classifier. Some steps are required for implementing the Bayes classifier. They are: -

1. Take the mean of “**final data**” [steps of PCA: -step-8] for both of two classes (leaf and stem) &store it in a particular vector.
2. Find the co-variance matrix of both the two mean vector (for both leaf and stem).
3. Calculate the likelihood function for both the two classes depending upon the mean and covariance matrix i.e. likelihood of leaf and the likelihood of stem.
4. Take the pattern and maintain the process from [i-iii].
5. If the likelihood of leaf is greater than likelihood of stem then the pattern is leaf else it is stem.

Here Baye’s Classifier is used for the classification process with the accuracy of the 70% for leaf and 65% for stem, which is acceptable for the first try. See Table1.

No	Name Of the Distribution	Place	Result
1.	Gray Level Distribution	Leaf	49%
2.	Gray Level Distribution	Stem	56%
3.	Green Distribution	Leaf	47%
4.	Green Distribution	Stem	46%
5.	Intensity Level Distribution	Leaf	70%
6.	Intensity Level Distribution	Stem	65%
7.	Hue Level Distribution	Leaf	50%
8.	Hue Level Distribution	Stem	44%

Table 1: Shows the Classification Results

for the different distribution of the Feature Set.

FIG: 6: -GRAPHICAL SUCCESS RATE FOR DIFFERENT DISTRIBUTION FOR LEAF IMAGE AND STEM IMAGE

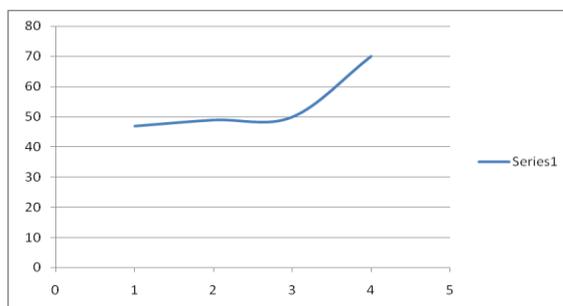
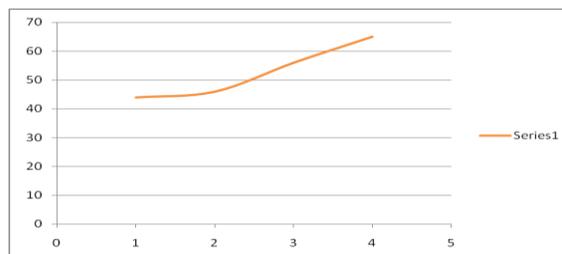


Fig: 6.A:- FOR LEAF IMAGE



[11]. Rice Doctor, International Rice Research Institute, Philipines, www.irri.org,2007.

[12]. Jonathon Shiens, "A Tutorial on Principal Component Analysis" Center for Neural Science, New York University, 22nd April2009, version3.01.

[13]. Lindsay I Smith, "A Tutorial on Principal Component Analysis", February 26,2002.

FIG. 6B-FOR STEM IMAGE

V. DISCUSSIONS AND FUTURE PLANNING

This work aims to develop a system that will automatically classify the rice plant images either as stem or leaf image by the help of application of image processing. After classifying the stem or leaf images our future planning is to examine that if there were any spot or not. After detecting the spot of leaf or stem it is important to know that which type of spot are in there. After that our aim is to detect the type of disease like Brown Spot, Blast, Sheath Blight Sheath Rot or etc, which is our ultimate goal.

REFERENCES

- [1]. Hashem EI-Khatib, Fetoh Hawela, Hasan Hamdi and Nabit EI-Mowelhi, "Spectral characteristics curves of the rice plants infected by blast." Soil and Water Research Institute, Remote Sensing Unit, Giza, Egypt, 18-21 August 1993 pp.526-528.
- [2]. P.Sanyal, U.Bhattachrya, S.K.Parui, S.K.Bandopadhyay & S.Patel, "Color Texture analysis of Rice Leaves to Diagnose Deficiency in the balance of Mineral levels towards Improvement Of Crop Productivity." 2007-IEEE 10th International Conference on Information Technology, pp. 85-90,2007.
- [3]. Zhihao.Qin, Minghua Zhang, Thomas Christensen, Wenjuan Li, & Huajan Tang," Remote Sensing Analysis Of Rice Disease Stresses For Farm Pest Management Using Wide band Airborne Data." 2003-IEEE, pp.2215-2217.
- [4]. Samesh M.Yamany, Aly A.Farag and Ahmed EI-Bialyz, "Free-Form Object Recognition And Registration Using Surface Signature."1999-IEEE, pp.457-461.
- [5]. Ephraim Katchalski-Katzir, Isaac Shavir, Miriam Eisenstein, Asher A.Friesem, Claude Aflalo and Ilya A.Vakser. "Molecular Surface Recognition: Determination of Geometric fit between Proteins and their Ligands by Correlation Techniques." March 1992- Proc. Natl. Acad.Sci.USA- Biophysics, vol.89, and pp.2195-2199.
- [6]. Rinat Ibrayev, Yan -bin Jia, "Surface Recognition by Registering Data Corves from Touch."2006-IEEE, pp.55-60.
- [7]. Santanu Phadikar and Jaya Sil, "Rice Disease Identification using Pattern Recognition Techniques" 11th International Conference, 24-27th December2008- ICCIT, pp 420-423.
- [8]. Gonzalez & Woods, "Digital Image Processing", New Delhi: Pearson Education-2nd edition, 2007.
- [9]. T.Reddy & S.Reddy, "Principles Of Agronomy", New Delhi: Kalyani Publishers, 2007.
- [10]. S.H.OU, "Rice Diseases", England: Commonwealth Agriculture Burcaux, 1972.