Leaf Disease Detection using Clustering Optimization and Multi-Class Classifier

DR.S. BHUVANA [1], KAVIYA BHARATI B [2], KOUSIGA P [3], RAKSHANA SELVI S[4]
Professor[1], Student[2]
Department of Computer Science and Engineering[1,2,3,4]
Sri Krishna College of Technology[1,2,3,4], India
bhuvana.harshan01@gmail.com[1],kaviyavaishtpr2510@gmail.com[2],
kousiga.palanivel@gmail.com[3],rakshanaerode@gmail.com[4]

ABSTRACT: *A*griculture is the only passion to cultivate foods, raising a human's life and animals by producing desired plant products. India ranked in the world's five largest producers of over 80% of agricultural produce items, including many cash crops such s rice, guava, tobacco, etc.Identification of the plant diseases is the key to preventing the losses in the yield and quantity of the agricultural product. Health monitoring and disease detection on plant is very critical for sustainable agriculture. It is very difficult to monitor the plant diseases manually. It requires tremendous amount of work, expertise in the plant diseases, and also require the excessive processing time. Consequently, image processing is used for the detection of plant diseases. The proposed system consist of following phases like: image preprocessing, image segmentation using otsu segmentation, clustering of an image using k-means, extract the feature using GLCM feature extraction, classify the image by Multi class SVM classifier. In compared to existing system, the proposed system significantly identify the plant leaf disease at an early disease and improve the accuracy to 98%.

Keywords: image pre-processing, OTSU segmentation, K-means, GLCM, Multi Class SVM

INTRODUCTION

India is a cultivated country and about 70% of the population depends on agriculture. Farmers have large range of diversity for selecting various suitable crops and finding the suitable pesticides for plant. Disease on plant leads to the significant reduction in both the quality and quantity of agricultural products. The studies of plant disease refer to the studies of visually observable patterns on the plants. Monitoring of health and disease on plant plays an important role in successful cultivation of crops in the farm. In early days, the monitoring and analysis of plant diseases were done manually by the expertise person in that field. This needs tremendous amount of work and conjointly requires excessive processing time. The image processing techniques can be used in the plant disease detection. In most of the cases disease symptoms are seen on the leaves, stem and fruit. The plant leaf for the detection of disease is considered which shows the disease symptoms.

In exsisting system, various image processing techniques such as Probabilistic Neural Network, Genetic Algorithm, Support Vector Machine are used.But they have some negative features like the quality of result can vary for different input data, requires tremendous amount of work, expertise in the plant diseases, and also require the excessive processing time.

To overcome these strikes, this paper mainly focus on some image processing techniques like

- prewitt algorithm(Pre-processing) which can resize and convert to black and white image,
- OTSU method which is used to contrast and enhance the effected leaf,
- The image can be clustered according to the

leaf colour,shape and size using K-means clustering,

- The GLCM can extract the texture and colour of the image,
- Finally ,Multi Class SVM classify the effected leaf according to the dataset.

This paper comprised of the following chapters, Literature survey of image processing is decribed in chapter 2. Chapter 3 gives the detailed description of the proposed system. Chapter 4 is to show the experimental result of the proposed system. Comparison and accuracy of the proposed system is given in chapter 5. Finally, chapter 6 gives the conclusion and future work.

2.LITERATURE SURVEY

In this section, image pre-processing methods are discussed. The different classification techniques used for plant leaf disease classification was proposed by Savita N. Ghaiwat [6] by using some classification techniques like k-Nearest Neighbour Classifier, Probabilistic Neural Network, Genetic Algorithm, Support Vector Machine, and Principal Component Analysis, Artificial neural network, Fuzzy logic. A classification technique deals with classifying each pattern in one of the distinct classes and it is used to classify the leaf based on its different morphological features. Selecting classification technique is tricky task because the quality of result can vary for different input data.

The detection of plant diseases using their leaves images is explained by Sachin D. Khirade [16]. The studies of the plant diseases mean the studies of visually observable patterns seen on the plant. Health monitoring and disease detection on plant is very critical for sustainable agriculture. It is very tricky to monitor the plant diseases manually. It needs tremendous amount of work and conjointly requires excessive processing time. Hence, image processing is used for the detection of plant diseases. Disease detection involves the steps like image acquisition, image pre-processing, image segmentation, feature extraction and classification. But, the accuracy of the result is 86% only.

The technique to classify and identify the different disease affected plant put forth by Mrunalini R. Badnakhe [14].By using the automated agricultural inspection, Farmer can given potentially better and accurate pro ductivity .The different products can be yield with better quality. The main needs for the agriculture is to predict the infected crop. With the help of this work we are indirectly contributing for the Improvement of the Crop Quality. It is a Machine learning based recognition system which will going to help in the Indian Economy. Digital Analysis of crop color is significant and now it's becoming popular day by day. It is the cost effective method. Because changed in the color are a valuable indicator of crop health and efficiency and survaibility. Then it can be measured with visual scales and inexpensive crop color.

Software solution for automatic detection and classification of plant leaf diseases was proposed by S. Arivazhagan[2]. The developed processing scheme consists of four main steps, first a color transformation structure for the input RGB image is created, then the green pixels are masked and removed by specific threshold value using segmentation techniques, the texture information are computed for the useful segments, finally the extracted features are passed through the classifier. The proposed algorithm's efficiency can successfully detect and classify the examined diseases with an accuracy of 94%. Preparatory outcomes on an informational collection of around 500 plant leaves affirm the fitness of the proposed approach. In order to improve disease identification rate at various stages, the training samples can be increased and shape feature and color feature along with the optimal features can be given as input condition of disease identification.

Anand.H.Kulkarni[1] propose a strategy for recognizing plant sicknesses at beginning period and with precision, by utilizing various picture handling procedures and artificial neural network (ANN).The work begins with capturing the images. Filtered and segmented using Gabor filter. Then, texture and color features are extracted from the result of segmentation and Artificial neural network (ANN) is then trained by choosing the feature values that could distinguish the healthy and diseased samples appropriately. Experimental results showed the classification action by ANN taking feature set is better with an precision of 91%. The results are encouraging and promise the development of a good machine vision system in the area of recognition and classification of plant diseases.

3. PROPOSED SYSTEM

The objective of the proposed system is to detect the plant leaf disease at early stage with 96% accuracy. Fig 1 shows the overall procedure of the proposed system. The following phases are worn in the proposed system.

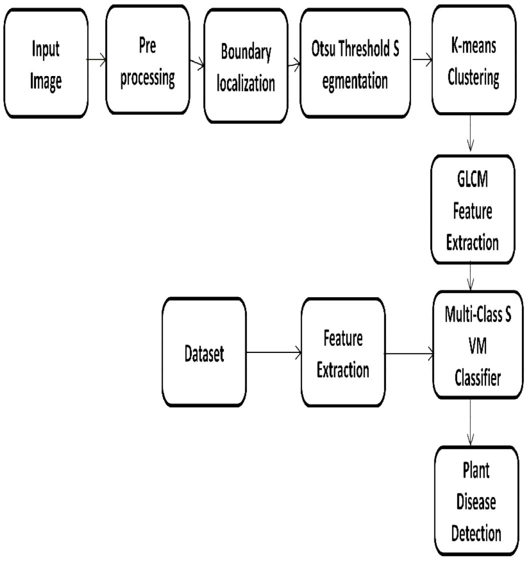


FIG 1 THE OVERALL PHASES OF THE PROPOSED SYSTEM

The input image was pre-processed by converting the image to black and white image[1][6].The second stage is the segmentation, which uses the OTSU Segmentation to contrast and enhance the pre-processed image. The third step is the K-Means Clustering to cluster the image according to the contrast colour. Fourth step is the feature extraction, which uses GLCM feature extraction for extracting the clustered images. Fifth phase is the classification of images using Multi-Class SVM from the extracted images[4]. Finally ,the five phases also done for datasets. The comparison was done for result set and data set, after all the disease name, accuracy and affected percentages was calculated.

3.1 Pre-Processing

The aim of the pre-processing is to improve the quality of the image like resize the image, convert the input images to black and white etc[1].The Pre-Processing techniques involved the following steps:

- The input image was resized to 256 X 256
- A coloured image was enhance the contrast colour.
- The input image was marked with red colour in the boundary, veins and affected part of the leaf.

3.2 OTSU Segmentation

Otsu's thresholding segmentation method holding a iteration through all the achievable threshold values and measure the pixel levels on each side of the threshold, i.e. the pixels that either fall in foreground or background.[18] The aim is to find the threshold value where the sum of foreground and [20]background spreads at its minimum.

Syntax

[level EM] = graythresh(I)

Description

- Level = graythresh(I) computes a global threshold (level) that can be used to convert an color image to a binary image with im2bw. level is a normalized intensity value that lies in the range [0, 1].
- The graythresh() uses the Otsu's segmentation method, which selects the threshold value to minimize the intra class variance of the black and white pixels.
- [level EM] = graythresh(I) gives the effectiveness metric, EM, as the second output argument. The EM is a value in the range [0 1] that indicates the effectiveness of the threshold value of the input image. The lower boundary is possible only by single gray level images ,were as upper boundary is possible only by two valued images.

3.3 K-Means Clustering

K-means is one of the simplest unsubstantiated learning algorithms that can solve the well known clustering problem. The procedure follows an easy way to classify a given information,which set through an exact range of clusters (assume k clusters) fixed apriori. The main aim is to describe k centers, one for every cluster. These centers should be placed in a cunning way because of different location causes different result[8]. So, the better choice is to place them as much as possible far away from each other[11]. The next step is to require each point belonging to a given data set and associate it to the closest center. When no point is unfinished, the primary step is completed and with early cluster age is also finished. At that time we want to re-calculate k new centroids as bary center of the clusters ensuing from the previous step. After we've these k new centroids, a new binding has to be done between the similar data set points and the nearest new center. A loop has been generated. As a results of this loop we have a addiction to notice that the k centers may change their location step by step until no modification can be made or in other words centers won't move from any more. Finally, this algorithm aims to minimize the objective function know as squared error function given by

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2 \qquad (1)$$

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1. Randomly select 'c' cluster centers.

2. Calculate the distance between each data point and cluster centers.

3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4. Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i \qquad (2)$$

c_i signifies the no. of data points in i^{th} cluster.

5. The separation between all the data points and recently fetched cluster centers was recalculated.

6. If no data point was reassigned then stop, otherwise repeat from step 3.

3.4 GLCM Feature Extraction

GLCM was abbreviated as **Gray-Level Co-Occurrence Matrix.** It is the most classical second-order statistical method for texture analysis. An image is composed of pixels each with an intensity (a specific gray level), [12] the GLCM is a tabulation of how often different combinations of gray levels co-occur in an image or image section. Texture feature make use of the GLCM to provide a measure of the variation in intensity at the pixel of interest.[7] GLCM texture feature operator produces a *virtual variable* which represents a specified texture calculation on a single beam echogram.

Steps for virtual variable creation:

1) Quantize the image data: Each sample on the echogram is treated as a single image pixel and its value is the intensity of that pixel. These intensities are then further quantized into a specified number of discrete gray levels, known as Quantization.

2) Create the GLCM: It will be a square matrix N x N in size where N is the Number of levels specified under Quantization.

Steps for matrix creation are:

a) Let s be the sample under consideration for the calculation.

b) Let W be the set of samples surrounding sample s which fall within a window centred upon sample s of the size specified under Window Size.

c) Define each element i, j of the GLCM of sample present in set W, as the number of times two samples of intensities i and j occur in specified Spatial relationship.

d) The sum of all the elements i, j of the GLCM will be the total number of times the specified spatial relationship occurs in W.

e) Make the GLCM symmetric:

i) Make a transposed copy of the GLCM.
ii) Add this copy to the GLCM itself.

This produces a symmetric matrix in which the relationship i to j is indistinguishable for the relationship j to i.

Due to summation of all the elements i, j of the GLCM will now be twice the total number of times the specified spatial relationship occurs in W.

f) Normalize the GLCM:

Divide each element by the sum of all elements.

The elements of the GLCM may now be considered probabilities of finding the relationship *i, j (or j, i)* in W.

3) Calculate the selected Feature. This estimation uses only the values in the GLCM. The sample s in the resulting virtual variable is replaced by the value of this calculated feature.

GLCM directions of Analysis

- *Horizontal (0^0)*
- *Vertical (90^0)*
- *Diagonal*
- *Bottom left to top right (-45^0)*
- *Top left to bottom right (-135^0)*
- Denoted P_{0}, P_{45},P_{90}, & P135 *respectively.*
- Ex. $P_0(i, j)$
- GLCM of an image is computed using a displacement vector d, defined by its radius δ and orientation θ.

Initially the matrix was builted, based on the orientation and distance between image pixels. Then meaningful statistics are extracted from the matrix as the texture representation . Haralick proposed the following texture features :

- Energy
- Contrast
- Correlation
- Homogenity
- Entropy

Energy: It is a gray-scale image texture measure of homogeneity changing, reflecting the distribution of image gray-scale uniformity of weight and texture..

$$E = \sum_{x}\sum_{y} p(x, y \quad (3)$$

p(x,y) is the GLCM

Contrast: Contrast is that diagonal close to the rotational inertia, that can measure the value of the matrix is scattered and local changes in number , reflect the image clarity and texture of shadow depth.

$$\text{Contrast} \qquad I = \sum\sum (x-y)^2 p(x,y) \quad (4)$$

Entropy: It measures image texture randomness, when the space co-occurrence matrix for all values are equal, it achieved the minimum value.

$$S = -\sum_x \sum_y p(x,y) \log p(x,y) \qquad (5)$$

Correlation Coefficient: Measures the joint probability occurrence of the specified pixel pairs.

Correlation:
sum(sum((x-μx)(y-μy)p(xy)/$\sigma_x\sigma_y$)) (6)

Homogeneity: Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

Homogenity =
sum(sum(p(x , y)/(1 + [x-y]))) (7)

Entropy Measurement
The Entropy will be finding for all sub band coefficients entropy measurement was used to extract significant information for texture pattern. Entropy measurement is giving by equation

$$\text{Entropy} = -\sum_{i,j} C(i,j) \log C(i,j) \qquad (8)$$

$C(i, j)$ is the normalized histogram that is applied to coefficients resulting from wavelet decomposing. As a result, a vector of each image sample on the database was produced. The resultant vectors then will be saved in the memory in indexing feature vector which contains the indexes to both the names and the images of training database.

3.5 Multi-Class SVM Classifier

Multi-class SVM provide a dense set of constraints, the number of variables in its dual problem is still $l \times k$ This value may explode even for small datasets. For instance, an English letter recognition problem with 2,600 samples (100 samples per letter) would require solving a QP of size 2,600×26, which will result in a large computational complexity. Here, we follow Crammer and Singer's work and further introduce a simplified method named Sim M-SVM for relaxing its constraints By doing so, solving one single l-variable QP is enough for a multi-class classification task. Before we describe the new direct multi-class SVM method we first compare the loss function[4] of the above two "all-together" approaches.For a training example **x***i*, we let. SVMs as binary classifiers have drawn much attention because of their high classification performance and thorough mathematical foundations rooted in statistical learning theory.The data in such a space lies on a unit hypersphere.[15].A modification in the SVM algorithm taking advantage of this geometrical property is proposed in [6] where the offset of the optimal separating hyperplane (OSH) is shifted. In the standard SVM algorithm, the OSH is placed in the middle of the margins.

$$\zeta_{i,m} = 1 - f_{y_i}(\mathbf{x}_i) + f_m(\mathbf{x}_i), \qquad (9)$$

for $m = \{1, \ldots, k\} \backslash yi$. If $\zeta i,m > 0$, it depicts the pairwise margin violation between the true class yi and some other class m. In Weston and Watkins' work, their loss function adds up all positive margin violation ($\zeta i,m > 0$),

$$\xi_i^{(1)} = \sum_{m \neq y_i} [\zeta_{i,m}]_+, \qquad (10)$$

where [·]+ ≡ max(·,0). In the original work proposed by Weston and Watkins the term ζ adopts the "2" rather than "1" as follows:

$$\zeta_{i,m} = 2 - f_{y_i}(\mathbf{x}_i) + f_m(\mathbf{x}_i), \qquad (11)$$

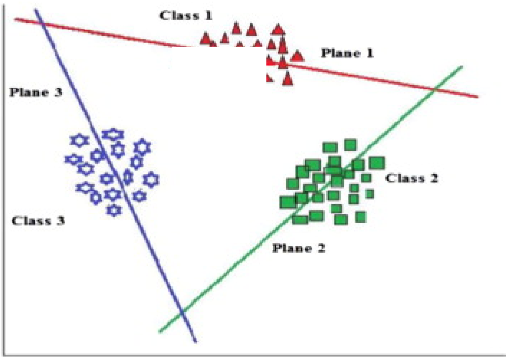


FIG 2.MULTI CLASS SVM

Here in order to compare the work proposed by Weston and Watkins with the other methods consistently, we scale the“2” into “1,” i.e. adopt the .[18] As for Crammer and Singer’s approach, sample loss counts only the maximum positive margin violation.

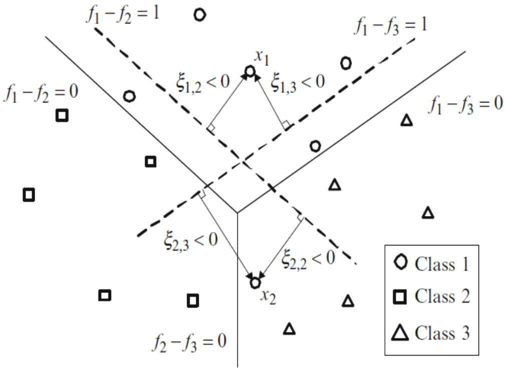


FIG 3. Multi-class classification visualization.

Class 1, 2, and 3 are symbolizing the *circles*, *rectangles*, and *triangles*, respectively. The *bold lines* represent some possible class boundaries. The two *dash lines* are two positive margins for the first class. The couple wise margin abuse for two examples from the first class, ζ1,2 and ζ1,3 for **x**1, and ζ2,2 and ζ2,3 for **x**2, are depicted in the fig 3

$$\xi_i^{(2)} = \left[\max_{m\neq y_i} \zeta_{i,m}\right]_+ . \qquad (12)$$

gives a multi-class classification graphical illustration, where three classes are represented as circles, rectangles, and triangles, respectively. The bold lines represent some possible decision boundaries. We plot the two positive pair wise margins for the first class (shown in dash lines).

In order to reduce the problem size, the number of constraints should be proportional to *l* instead of *l* × *k*. To construct the multi-class predictors of Sim M SVM, we introduce the following relaxed bound.

4. EXPERIMENTAL RESULT

The proposed system is implemented using MATLAB 8.3.0.532 (R2014a). Fig 4 shows the sample dataset images.

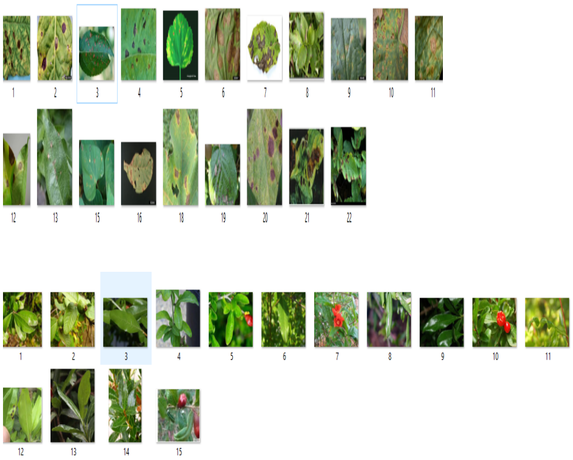
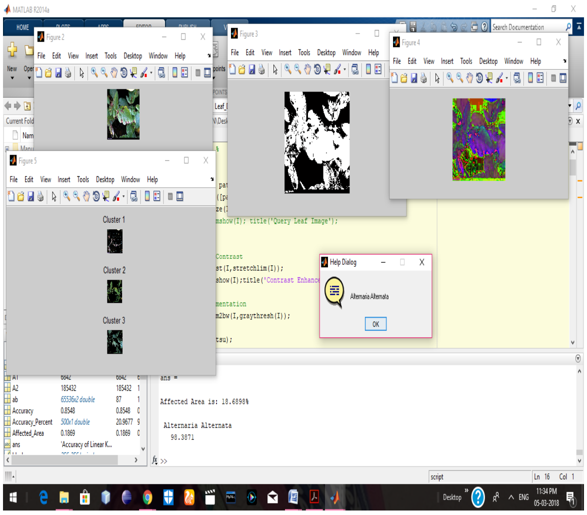


Fig 4 Sample Dataset images

The dataset holds 25 different plant leaves images of JPEG format, divided into five classes as follows: Alternaria Alternaat, Anthracnose, Bacteria Blight, Cercospora Leaf Spot and Healthy Leaves. Each image of encapsulates 22 images of 386*256.In proposed system, the image is rescaled to 256*256.

An image from dataset was pre-processed and convert to black and white image and clustered according to their texture, color, shape. Image was classified by multi-class svm classifier, after this process the disease name was detected with accuracy and affected percentage.Fig 5 shows the retrived result of a input image.



```
mshow(I); title('Query Leaf Image');
Contrast
st(I,stretchlim(I));
show(I);title('Contrast Enhance
mentation
m2bw(I,graythresh(I));
tsu);
```

Alternaria Alternata

A1	6842	6842
A2	185432	185432
ab	65536x2 double	87
Accuracy	0.8548	0.8548
Accuracy_Percent	500x1 double	20.9677
Affected_Area	0.1869	0.1869
ans	'Accuracy of Linear K...	

```
ans =

Affected Area is: 18.6898%

 Alternaria Alternata
   98.3871
```

Fig 5. Retrieval result

5. PERFORMANE EVALUATION

The disease name was detected from the dataset with retrival affected and accuracy percentage. Fig 6 shows the accuracy of the plant leaf disease.

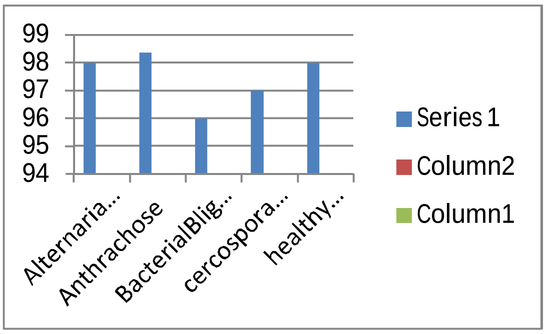


Fig 6. Accuracy of the plant leaf disease.

The comparison of the proposed system and existing system for the plant leaves on the dataset was shown in the fig 7.

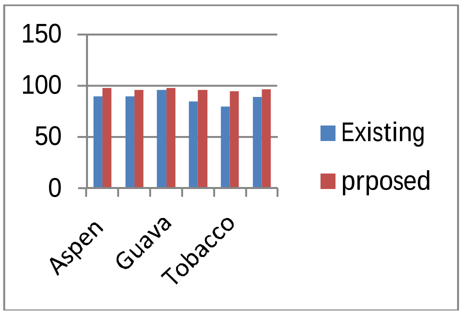


Fig 7. Comparison of proposed system and existing system.

6.CONCLUSION & FUTUTRE WORK

The accurate detection and classification of the plant disease is very important for the flourishing cultivation of crop and this can be done using image processing. There are various techniques to segment the plant disease. There are some Feature extraction and classification techniques to extract the features of infected leaf and the classification of plant diseases. The use of SVM methods for classification of disease in plants such as self-organizing feature map, back propagation algorithm, SVM etc. can be efficiently used. From these methods, we can accurately identify and classify various plant diseases using image processing techniques.

The future work principally worries with the extensive database and propel advance feature of colour extraction that contains a better result of detection.. Another work worries with look into work in a specific field with thrust highlights and innovation.

REFERENCE

[1] Anand.H.Kulkarni, Ashwin Patil R.K.Applying image processing technique to detect plant diseases International Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.2, Issue.5, Sep-Oct. 2012 pp-3661-3664 ISSN: 2249-6645.

[2] Arivazhagan S,Newlin Shebiah R,Ananthi S.Detection of unhealthy region of plant leaves and classification of plant leaf disease usi g Texture features Agric Eng Int CICR 2013;15(1):211-7.

[3] Bashir Sabah Sharma Navdeep. Remote area plant disease detection using Image processing .IOSR J Electron Communication Eng 2012;2(6):31-4.ISSN:2278-2834.

[4] Bhagya Patil, AnupamabPatlanshetty, Suvarna Nandyal. Plant classification using SVM classifier,Computational Intelligence and Info Technol,2013.

[5] Bhanu B,Peng J.Adaptive Integrated image segmentation and object recognition. IEEE Trans Syst Man Cybern Part C 2000;30:427-41.

[6] Ghaiwat Savita N,Arora Parul.Detection and classification of plant leaf disease using Image processing techniques: a review Int J Recent Adv Eng Technol 2014;2(3):2347-812.ISSN.

[7] L.Gurukumar,P.Sathyanarayanan, I.age Texture Feature Extraction using GLCM Approach ,International Journal of Scientific and Res,5 Jan 2013.

[8] Jianpeng Qi,Yanwei Yu,Lithong Wang.k means :An Effective and efficient k means clustering .2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computingand Communication(SustainCom) (BDCloud-SocialCom-SustainCom).

[9] Karnal O Hajari,Miran R Gavahio.Unhealthy region of citrus leaf detection using Image processing techniques Convergence of Technology,(12CT)2014.

[10] Kumbhar Nithin P.Agricultural plant disease detection using Image processing .Int J.Adv Res Electr Electron Instrument Eng 2013.

[11] Liu Xumin,Than Yong.Research on k means clustering Algorithm :An improved k-means clustering Algorithm; a review Int J Recent Adv Eng Technol 2010.

[12] Maroune Ben Haj Ayech,Hamid Amiri.Texture Description using Stasitical Feature Extraction,Int Computer Science Telecommun 2016;3(6).

[13] Ma Z Tavares,Image processing and analysis: application and trends AES-ATEMA'2010,Fifth International conference Canada,ISBN:978-0-9780479-7-9.

[14] Mrunalini R Badnakhe,Deshmukh Prashant R.An application. Of K- means clustering and artificial in pattern recognition for crop disease .Int Find Adv INF Technology 2011;20.2011 IPCSIT.

[15] Patil Sanjay B et AL.Leaf disease severity measurement using Image processing. Int J Eng Technol 2011;3(5):297-301.

[16] Sachin D khirade.Plant Disease detection using Image processing:a review .Int J Recent Adv Eng Technol2015,ICCUBEA,27 Feb 2015.

[17] Sofine Mouine,Other Yahioui,Anne Verroust-Blondet.Plant species recognition using spatial correlation between the leaf Margin and leaf salient points.,Int Agri. ICIP ,18 Sept 2013.

[18] Rathod Arti N,Tanawal Bhavesh,Shah Vatsal.Image processing techniques for detection of leaf disease.Int J Adv Res Comput Sci Soft Eng 2013;3(11).

[19] Woods Keri.Genetic algorithms: colour image segmentation literature review, 2007.

[20] Zhen MATavares JMRS,Natal JorgeRM,Areview on the current segmentation algorithm for medical image.In st inter national conference on imaging theory and applications, Portugal;2009,ISBN:978-989-8111-68-5;