Paddy Leaf Disease Classification Using Image Processing

Mr. D. K. Kirange
Dept. of Computer and IT, J T Mahajan College of Engineering,
Faizpur, Tal, Yawal, Dist. Jalgaon, India

Smt Shubhangi D Patil
Department of Information Technology
Government Polytechnic Jalgaon

***Abstract*: The early detection of diseases is important in agriculture for an efficient crop yield. The bacterial spot, late blight, septoria leaf spot and yellow curved leaf diseases affect the crop quality of tomatoes. Automatic methods for classification of plant diseases also help taking action after detecting the symptoms of leaf diseases. This paper presents a performance measure for different feature extraction techniques for paddy leaf disease detection including GLCM, Gabor and SURF and classification techniques including decision trees, SVM, KNN and Naïve Bayes. The dataset contains 500 images of paddy leaves with four symptoms of diseases. We have modeled a system for automatic feature extraction and classification. We have evaluated the performance of the system using different performance measures to conclude with appropriate features set and classification technique for paddy leaf disease classification. The experimental results validate that Gabor features effectively recognizes different types of paddy leaf diseases. Accuracy of SVM is better as compared to other classification techniques.**

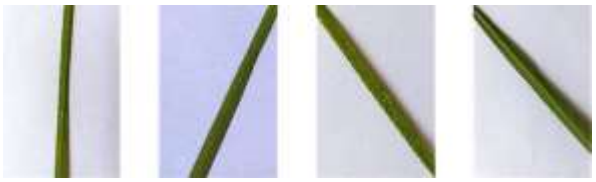
***Keywords*: GLCM, Gabor, SURF, SVM, KNN, Naïve Bayes, Decision Trees**

1. Introduction

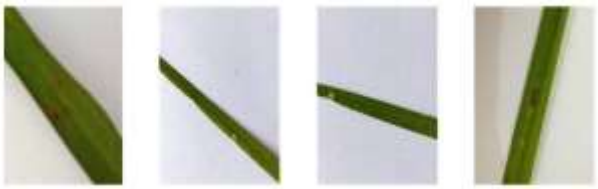
India is an agricultural country wherein most of the population depends on agriculture. Plant diseases have turned into a dilemma as it can cause significant reduction in both quality and quantity of agricultural products. Therefore, plant disease identification is a very important and challenging task. Mostly diseases are seen on the leaves or stems of the plant on the fruits also. Precise quantification of these visually observed diseases, pests, traits has not studied yet because of the complexity of visual patterns. Hence there has been increasing demand for more specific and sophisticated image pattern understanding. In biological science, sometimes thousands of images are generated in a single experiment. These images can be required for further studies like classifying lesion, scoring quantitative traits, calculating area eaten by insects, etc. Almost all of these tasks are processed manually or with distinct software packages. It is not only tremendous amount of work but also suffers from two major issues: excessive processing time and subjectiveness rising from different individuals. Hence to conduct high throughput experiments, plant biologist need efficient computer software to automatically extract and analyse significant content. Here image processing plays important role.Diseases are impairment to the normal state of the plant that modifies or interrupts its vital functions such as photosynthesis, transpiration, pollination, fertilization, germination etc. These diseases are caused by pathogens viz., fungi, bacteria and viruses, and due to adverse environmental conditions. Therefore, the early stage diagnosis of plant disease is an important task. Farmers require continuous monitoring of experts which might be prohibitively expensive and time consuming. Therefore, looking for fast, less expensive and accurate method to automatically detect the diseases from the symptoms that appear on the plant leaf is of great realistic significance. Most leaf diseases are caused by fungi, bacteria and viruses. Fungi are identified primarily from their morphology, with emphasis placed on their reproductive structures. Bacteria are considered more primitive than fungi and generally have simpler life cycles. With few exceptions, bacteria exist as single cells and increase in numbers by dividing into two cells during a process called binary fission viruses are extremely tiny particles consisting of protein and genetic material with no associated protein.

Fig.1 shows four common diseases on paddy plant. The most common method used by farmers for detection of is by naked eyes, which required experience which totally depends on knowledge of their ancestors. Other method is to refer experts which is followed by only 1% of farmers due to its higher cost. Another economical options is discussed in this paper that is using image processing. All the diseases have common characteristics that they changes some morphological characteristics such as color or shape. These changes can be extracted through classification of different disease through Gabor features.

(a) leaves with BrownSpot disease



(b)paddy leaves with Hispa disease



© paddy leaves with LeafBlast disease

(d) healthy paddy leave

Fig.1: Samples of paddy plant diseases; (a leaves with BrownSpot disease ; (b) paddy leaves with Hispa disease; (c) paddy leaves with LeafBlast disease (d) healthy paddy leave

This research paper aims at following objectives

- To identify deficiency of paddy plant by analyzing its leaf efficiently.
- By identifying deficiency of paddy leaf, we can obtain healthy products (paddys).
- To predict occurrence of disease accurately based on analyzing deficiency symptoms.

The rest of the paper is organized as follows: The dataset used for paddy leaf disease classification is discussed in section 2,related work is presented in the section 3, followed by the proposed method in the section 4, while experimental set-up and results are discussed in the section 5. Finally, we draw our conclusion in the section 6.

2. PlantVillage Image Dataset for Paddy Leaf Disease Classification

We extracted our dataset from the well-known PlantVillage dataset, which contains nearly 50,000 images of 14 crop species and 26 diseases. We choose to work with 9,000 images on Paddy leaves; our dataset contains samples for 4 types of Paddy diseases in addition to healthy leaves, 67classes in total as follow[1]:

- Class (0): leaves with BrownSpot disease
- Class (1): paddy leaves with Hispa
- Class (2): paddy leaves with LeafBlast disease
- Class (3): healthy paddy leave

3. Related Work

In paper [2] there are four steps. Out of them the first one is gathering image from several part of the country for training and testing. Second part is applying Gaussian filter is used to remove all the noise and thresholding is done to get the all green color component. K-means clustering is used for segmentation. All RGB images are converted into HSV for extracting feature The paper [3] presents the technique of detecting jute plant disease using image processing. Image is captured and then it is realized to match the size of the image to be stored in the database. Then the image is enhanced in quality and noises are removed. Hue based segmentation is applied on the image with customized thresholding formula. Then the image is converted into HSV from RGB as it helps extracting region of interest. This approach proposed can significantly support detecting stem oriented diseases for jute plant. According to paper [4] they have proposed for a technique that can be used for detecting paddy plant disease by comparing it with 100 healthy images and 100 sample of disease1 and another 100 sample of disease2. It's not sufficient enough to detect disease or classify it training data is not linearly separable. In paper [5] detection of unhealthy plant leaves include some steps are RGB image acquisition. Converting the input image from RGB to HSI format. Masking and removing the green pixels. Segment the components using Ostu's method. Computing the texture features using color-co-occurrence methodology and finally classifying the disease using Genetic Algorithm. Paper [6] includes paddy disease detection using computer vision. A gray scale image is turned into binary image depending on threshold value. The threshold algorithm is used for image segmentation. The threshold values are given color indices like red, green, blue. But the thresholding is not a reliable method as this technique only distinguishes red paddyes from other colors. It becomes difficult to distinguish ripe and unripe paddyes. For this K-means clustering algorithm is used to overcome the drawbacks. K-means create a particular number of nonhierarchical clusters. This method is numerical, unsupervised, non-deterministic and iterative. Then separating the infected parts from the leaf the RGB image was converted into YcbCr to enhance the feature of the image. The final step is the calculation of the percentage of infection and distinguishing the ripe and unripe paddyes. The methodology for cucumber

disease detection is presented in paper [7]. The methodology includes image acquisition, image preprocessing, feature extraction with Gray level co-occurrence matrix (GLCM) and finally classified with two types: Unsupervised classification and supervised classification. Paddy plant is an important plant in continental region. In paper [8] RGB images are converted into gray scale image using color conversion. Various enhancement techniques like histogram equalization and contrast adjustment are used for image quality enhancement. Different types of classification features like SVM, ANN, FUZZY classification are used here. Feature extract6ion uses different types of feature values like texture feature, structure feature and geometric feature. By using ANN and FUZZY classification, it can identify the disease of the paddy plant. In paper [9] popular methods have been utilizes machine learning, image processing and classification based approaches to identify and detect the disease of agricultural product. In paper [10] image processing technique are used to detect the citrus leaf disease. This system includes: Image preprocessing, segmentation of the leaf using K-means clustering to determine the diseased areas, feature extraction and classification of disease. Uses Gray-Level Co-Occurrence matrix (GLCM) for feature extraction and classification is done using support vector machine (SVM).

Paper [11] presents classification and detection techniques that can be used for plant leaf disease classification. Here preprocess is done before feature extraction. RGB images are converted into white and then converted into grey level image to extract the image of vein from each leaf. Then basic Morphological functions are applied on the image. Then the image is converted into binary image. After that if binary pixel value is 0 its converted to corresponding RGB image value. Finally by using pearson correlation and Dominating feature set and Naïve Bayesian classifier disease is detected.

4. Paddy Leaf Disease Classification

The system contains five major modules: paddy leaf disease input database, preprocessing/ Noise removal, feature extraction, classifier & recognized output as illustrated in fig.2. Overall, the system is based on preprocessing/ noise removing mechanism of leaf disease images, extracting some of features which contain information about textural features of the image & taking appropriate pattern recognition model to identify the type of paddy leaf image disease

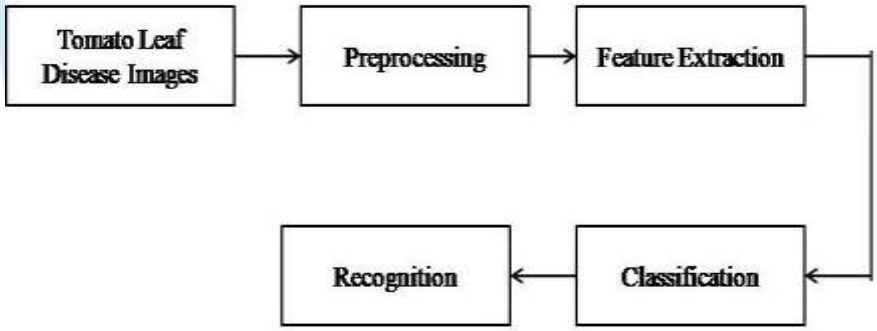


Fig.2. Structure of Paddy Leaf Image Disease Classification

Various steps for classification of paddy disease from disease leaf images as follows:

1. Database Gathering and Preprocessing:Obtaining the paddy leaf disease images and Normal paddy leaf images Data from Plant Village Dataset
2. Preprocessing and noise removal of paddy leaf images using Median Filter.
3. Feature Extraction : GLCM, Gabor and SURF features
4. Design and development of the system for classification of paddy leaf images as normal or diseased containing 7 types of paddy diseases.
5. Evaluate the performance of different classifiers
   - SVM
   - KNN
   - Naïve Bayes
   - Decision Trees

a. Gabor Features Extraction

In image processing, a Gabor filter, named after Dennis Gabor, is a linear filter used for texture analysis, which means that it basically analyzes whether there are any specific frequency content in the image in specific directions in a localized region around the point or region of analysis. Frequency and orientation representations of Gabor filters are claimed by many contemporary vision scientists to be similar to those of the human visual system, though there is no empirical evidence and no functional rational to support the idea. They have been found to be particularly appropriate for texture representation and discrimination. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave.

b. SURF Features

The SURF method (Speeded Up Robust Features) is a fast and robust algorithm for local, similarity invariant representation and comparison of images. The main interest of the SURF approach lies in its fast computation of operators using box filters, thus enabling real-time applications such as tracking and object recognition.

c. Statistical Features

stats = graycoprops(glcm, properties) calculates the statistics specified in properties from the gray-level co-occurence matrix glcm. glcm is an m-by-n-by-p array of valid gray-level co-occurrence matrices. If glcm is an array of GLCMs, stats is an array of statistics for each glcm.

graycoprops normalizes the gray-level co-occurrence matrix (GLCM) so that the sum of its elements is equal to 1. Each element (r,c) in the normalized GLCM is the joint probability occurrence of pixel pairs with a defined spatial relationship having gray level values r and c in the image. graycoprops uses the normalized GLCM to calculate properties

Different statistical features considered here are contrast, correlation, Energy and Homogeneity

d. Support Vector Machines

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. When data are unlabelled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.

e. KNN Classification

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

f. Naïve Bayes Classification

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Naïve Bayes is a classification algorithm that relies on strong assumptions of the independence of covariates in applying Bayes Theorem. The Naïve Bayes classifier assumes independence between predictor variables conditional on the response, and a Gaussian distribution of numeric predictors with mean and standard deviation computed from the training dataset.

Naïve Bayes models are commonly used as an alternative to decision trees for classification problems. When building a Naïve Bayes classifier, every row in the training dataset that contains at least one NA will be skipped completely. If the test dataset has missing values, then those predictors are omitted in the probability calculation during prediction.

g. Decision Trees

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (**S**), each belonging to one of the classes **C**1, **C**2, …, **C**k is as follows:

Step 1: If all the objects in **S** belong to the same class, for example **C**i, the decision tree for **S** consists of a leaf labelled with this class.
Step 2: Otherwise, let T be some test with possible outcomes *O*1, *O*2,…, *O*n.

Each object in **S** has one outcome for T so the test partitions **S** into subsets **S**1, **S**2,… **S**n where each object in **S**i has outcome *O*i for T. **T** becomes the root of the decision tree and for each outcome *O*i we build a subsidiary decision tree by invoking the same procedure recursively on the set **S**i.

5. Result Analysis

Matlab-based GUI-driven tool is developed for effective classification of paddy leaf diseases. Fig.3 shows graphical user interface(GUI) developed for proposed algorithm before execution. GUI for this software is divided into number of subgroups according to their functionality.

a. Database Selection and Preprocessing:

Paddy leaf disease images training database is selected. Then for preprocessing, median filter is used for noise removal.

b. Features Extraction

From the preprocessed training images Gabor, SURF and statistical features are extracted. Features matrix is constructed.

c. Classification

Different classifiers including SVM, KNN, Naïve Bayes and Decision Trees are trained with various features for paddy leaf disease classification. This module deals with paddy leaf disease detection and classification. The performances of different classifiers have been evaluated by considering different number of training images. Four parameters are used for evaluating performance of the algorithm. Those are accuracy, precision, recall and F measure. These parameters are defined using 4 measures True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)

True Positive: DR detection coincides with actual labelled data
True Negative: both classifier and actually labelled absence of DR
False Positive: system labels a healthy case as an DR one
False Negative: system labels DR image as healthy

Accuracy: Accuracy is the ratio of number of correctly classified cases, and is given by

$$Accuracy = (TP+TN) / N$$

Total number of cases are N

Precision is the fraction of retrieved images that are relevant to the query. Precision takes all retrieved images into account, but it can also be evaluated at a given cut-off rank, considering only the results returned by the system

Precision : Precision is defined as

$$Precision = TP/ (TP+FP)$$

Recall: Recall is the fraction of the relevant images that are successfully retrieved. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by also computing the precision.

Recall is defined as

$$Recall = TP / (TP+FN)$$

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, F1 score is 0.701.

$$F1\ Score = 2*(Recall * Precision) / (Recall + Precision)$$

TABLE I. PERFORMANCE MEASURE WITH DECISION TREE

Decision Trees					
	'Accuracy'	'Precision'	'Recall'	'F Measure'	Accuracy
GLCM	0.6352	0.1825	0.722	0.2772	0.5626
Gabor	**0.7459**	**0.2278**	**0.6961**	**0.5325**	**0.7625**
SRF	0.5323	0.1256	0.4526	0.1245	0.6253

TABLE II. PERFORMANCE MEASURE WITH SVM

SVM					
	'Accuracy'	'Precision'	'Recall'	'F Measure'	Accuracy
GLCM	0.4095	0.1306	0.9458	0.2295	0.2958
Gabor	**0.7339**	**0.2525**	**0.9492**	**0.3989**	**0.6589**
SRF	0.326	0.1084	0.8644	0.1926	0.4563

TABLE III. PERFORMANCE MEASURE WITH KNN

KNN					
	'Accuracy'	'Precision'	'Recall'	'F Measure'	Accuracy
GLCM	0.633	0.1701	0.7593	0.2779	0.6325
Gabor	**0.732**	**0.2555**	**0.9831**	**0.4056**	**0.782**
SRF	0.5485	0.1462	0.7966	0.2471	0.4256

TABLE IV. PERFORMANCE MEASURE WITH NAÏVE BAYES

NB					
	'Accuracy'	'Precision'	'Recall'	'F Measure'	Accuracy
GLCM	0.6589	0.1919	0.8305	0.3117	0.3759
Gabor	**0.675**	**0.2187**	**0.9695**	**0.3568**	**0.6324**
SRF	0.3493	0.0876	0.6373	0.1541	0.3256

As depicted in tables I to IV, Gabor features with all classifiers give promising result for paddy leaf disease classification. So for paddy leaf diseases classification, KNN classification is better better performance.

6. Conclusion

In this paper, KNN classification framework with Gabor features is used for paddy leaf disease classification. Different features like SURF, Statistical and Gabor are used. The different classifiers including SVM, KNN, Naïve Bayes and decision trees are trained to carry out the final classification. Main focus of this study is to preprocess the paddy leaf images for noise removal. After preprocessing and features extraction, classification of the selected four different paddy leaf diseases is performed. For PlantVillage data KNN with Gabor features gives better accuracy in terms of precision, recall and F measure. The experimental results have demonstrated the effectiveness of our proposed algorithm to be good enough to be employed in real time applications. The classification results showed that KNN is resulting good accuracy. The method proposed in the paper could also use for plant disease image recognition and classification. There are more sophisticated techniques are available for classification like Adaptive neuro fuzzy, Neural Networks, Genetic algorithm. etc. for image classification. These techniques can also use for plant image recognition and classification.

References

[1] https://www.kaggle.com/emmarex/plantdisease, was retrieved 20/12/2018.
[2] Pranjali B. Padol, Prof. AnjilA.Yadav, "SVM Classifier Based Grape Leaf Disease Detection" 2016 Conference on Advances in Signal Processing(CAPS) Cummins college of Engineering for Women, Pune. June 9-11, 2016.
[3] Detecting jute plant disease using image processing and machine learning 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)
[4] Tejoindhi M.R, Nanjesh B.R, Jagadeesh Gujanuru Math, AshwinGeetD'sa "Plant Disease Analysis Using Histogram Matching Based On Bhattacharya's Distance Calculation" International Conference on Electrical, Electroniocs and Optimization Techniques(ICEEOT)-2016
[5] Detection of unhealthy plant leaves using image processing and genetic algorithm with Arduino2018 International Conference on Power, Signals, Control and Computation (EPSCICON)
[6] Tanvimehera, vinaykumar,pragyagupta "Maturity and disease detection in tomato using computer vision" 2016 Fourth international conference on parallel, distributed and grid computing(PDGC)
[7] Ms.Poojapawer ,Dr.varshaTukar, prof.parvinpatil "Cucumber Disease detection using artificial neural network"
[8] Detection and measurement of paddy leaf disease symptoms using image processing
[9] Mukesh Kumar Tripathi, Dr.Dhananjay, D.Maktedar" Recent Machine Learning Based Approaches for Disease Detection and Classification of Agricultural Products" International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT)-2016.
[10] Detection of leaf diseases and classification using digital image processing2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)
[11] Dhiman Mondal, Dipak Kumar Kole, Aruna Chakraborty, D. Dutta Majumder" Detection and Classification Technique of Yellow Vein Mosaic Virus Disease in Okra Leaf Imagesusing Leaf Vein Extraction and Naive Bayesian Classifier., 2015, International Conference on Soft Computing Techniques and Implementations- (ICSCTI) Department of ECE, FET, MRIU, Faridabad, India, Oct 8-10, 2015.