CHAPTER 1

INTRODUCTION

Plants exist everywhere we live, as well as places without us. Plant disease is one of the crucial causes that reduces quantity and degrades quality of the agricultural products. Plant Pathology is the scientific study of plant diseases caused by pathogens (infectious diseases) and environmental conditions (physiological factors). It involves the study of pathogen identification, disease etiology, disease cycles, economic impact, plant disease epidemiology, plant disease resistance, pathosystem genetics and management of plant diseases. The primary occupation in India is agriculture. India ranks second in the agricultural output worldwide. Here in India, farmers cultivate a great diversity of crops. Various factors such as climatic conditions, soil conditions, various disease, etc affect the production of the crops. The existing method for plants disease detection is simply naked eye observation which requires more man labor, properly equipped laboratories, expensive devices ,etc. And improper disease detection may led to inexperienced pesticide usage that can cause development of long term resistance of the pathogens, reducing the ability of the crop to fight back.

Tobacco is one of the most successful commercial crops cultivated on this planet. China, India, Brazil and USA are the major producers of tobacco worldwide and these four nations alone contribute around 86 percent of the global production. India is the top two contributors to the global tobacco production and its estimated that around 750 Mkgs of tobacco is being produced in the area around 0.45 M hectares. Tobacco is cultivated in many parts of the world because of its high economic value. Quality inspection of tobacco leaves plays a crucial role in quality assurance of tobacco productions. After curing, the tobacco leaves are inspected and graded according to their color intensity, maturity, leaf structure, body, oil, length, appearance, waste and other characteristics. At present, the grading process is performed manually throughout the world. The grading process is extremely labor-intensive and millions of man-days are required to grade each years crop. A high level of skill is required to the graders, but still many mistakes are made by them because the process is highly subjective. So graders are eager for equipment that can help them grade tobacco leaves. If we can use machine vision technology and design algorithms to grade tobacco leaves automatically, it will be very useful for improving the level and efficiency of tobacco grading, arbitrating the dispute of the quality of tobacco leaves between buyer and seller. Because of the diversity and complexity of tobacco leaves New technology and equipment are needed to automate the quality inspection process of tobacco leaves.

Advantages of Tobacco leaves

- **As a first aid -** One of the best first aid for wounds, ear and tooth ache, even venom stings and snake bites.
- **Clear nasal passages -** They also believed that it helps people with asthma and tuberculosis breathe better.
- **Future treatment for cancer:** With its neuroprotective and tough tumor restraining properties cembranoids is trusted to be the future treatment for chronic cancers.
- **Treats depression:** Nicotine is a rich chemical compound present in tobacco and this research proved the direct link between nicotine and a subjective increase in the discharge of two essential neurotransmitters namely dopamine and serotonin, the lack of which is regarded as a major cause for depression.
- **Bio-Engineering :** Tobacco is used in bio-engineering pharmaceutical labs as a manufacturing platform for the production of wide range of drugs and therepeutic agents
- **Kill pests -** Tobacco is a great pest poison for the garden especially for centipedes, caterpillars and snails that tend to feed on leaves.
- **Relieve allergies -** A poultice of tobacco leaves to put on skin inflammations can help soothe itching and mild pain.
- **Make a natural insect repellent -** Keep the mosquitoes and bed bugs away by boiling a package of chewing tobacco in a gallon of water.

Quality classification of Tobacco Leaves

Fig: Good quality

Fig: Mid quality

Fig: Low Quality

Fig 1.1 Quality classification images

1.1 Introduction to Image Processing

Image processing is a method to convert an image into digital form and perform some operations on it, in order to get an enhanced image or to extract some useful information from it. It is a type of signal dispensation in which input is image, like video frame or photograph and output may be image or characteristics associated with that image.Image processing is one of the rapidly growing technologies. It forms core research area within engineering and computer science disciplines too.

There are two types of methods used for image processing namely,

1.1.1 Analogue image processing: Analog or visual techniques of image processing can be used for the hard copies like printouts and photographs.

• Image analysts use various fundamentals of interpretation while using these visual techniques. The image processing is not just confined to area that has to be studied but on knowledge of analyst.

• It is another important tool in image processing through visual techniques.

• Analysts use various fundamentals of interpretation while using these visual techniques.

1.1.2 Digital image processing:Digital Processing techniques help in manipulation of the digital images by using computers.

• Raw data from imaging sensors from satellite platform contains deficiencies.

• Get over such flaws and to get originality of information, it has to undergo various phases of processing.

CHAPTER 2

PROBLEM STATEMENT

At present, the grading process is performed manually throughout the world. The grading process is extremely labor-intensive and millions of man-days are required to grade each years crop. A high level of skill is required to the graders, but still many mistakes are made by them because the process is highly subjective. So graders are eager for equipment that can help them grade tobacco leaves. If we can use machine vision technology and design algorithms to grade tobacco leaves automatically, it will be very useful for improving the level and efficiency of tobacco grading, arbitrating the dispute of the quality of tobacco leaves between buyer and seller. The 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 using specific threshold value followed by segmentation process, the texture statistics are computed for the useful segments, finally the extracted features are passed through the classifier. In classifier the quality of the leaves can be identified.

CHAPTER 3

LITERATURE SURVEY

[1] Zhang, J.; Sokhansanj, S.; Wu, S.; Fang, R.; Yang, W.; Winter, P. A trainable grading system for tobacco leaves. Comput. Electron. Agric. 1997, 16, 231244.

A grading system based on image processing techniques was developed for automatic inspection and grading of flue-cured tobacco leaves. The system used machine vision in extraction and analysis of color, size, shape, surface texture and vein. A two-dimensional feature space was proposed to express feature distribution of tobacco leaves. The space was found to be well confined in an elliptic region. A database was constructed to record the feature distribution of standard contrast tobacco leaves prepared by experts through visual evaluation. The decision on grades was made based on the so-called 'nearest-neighbor' method for which the overall difference among features between the measured tobacco leaves and the standard contrast samples were used as a target parameter for judgment. This system can be easily trained by users with the knowledge of the feature distribution information of different tobacco leaves. It can also be adapted for the inspection of other agricultural products.

[2] MacCormac J. On-line image processing for tobacco grading in Zimbabwe. Proceedings of IEEE International Symposium on Industrial Electronics; Budapest, Hungary. 1–3 June 1993; pp. 327–331

The authors describe how research is currently being undertaken in the Electrical Engineering Department at the University of Zimbabwe in conjunction with the University of Bath with a view to automating the tobacco grading process. The objective of the research is to provide a reliable grading tool to the industry which will give a uniform grading standard across the country. In addition it is intended to increase throughput and provide quantification of the variation in tobacco leaf quality. The heart of the automatic grading system is a high-speed online image processing system developed to process images at a field rate of 50 Hz. The image processing is performed by Texas Instruments TMS 320-E-25 microprocessors operating in a parallel configuration.

[3] Zhang, F.; Fang, R.; Cai, J. Image retrieveal of standary tobacco leaf database. Trans. Chin. Soc.Agric. Mach. 2001, 32, 6667

The authors have recently produced software that uses image segmentation and classification techniques to identify blight and rust lesions on oats and sunflower leaves for the purpose of rapid disease assessment in the field. For tobacco, there has been some research into the estimation of leaf area in Japanese burley and also into a Cuban dark tobacco variety called corojo

[4] Ma W, He L, Xu S, Chen J, Wu Z. Image segmentation based on transmission characteristics of flue-cured tobacco leaves. Trans. Chin. Soc. Agricult. Eng. 2006;22:134–137.

The authors have proposed that Image preprocessing include the detection and restoration of bad lines, geometric rectification or image registration, radiometric calibration and atmospheric correction, and topographic correction. If different ancillary data are used, data conversion among different sources or formats and quality evaluation of these data are also necessary before they can be incorporated into a classification procedure. Accurate geometric rectification or image registration of remotely sensed data is a prerequisite for a combination of different source data in a classification process.

[5] LeCun, Yann, et al. 1998. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE 86 (11): 22782324.

The authors have proposed that Image segmentation performance of tobacco leaves in natural environment, an automatic segmentation model of leaf with active gradient and local information is proposed. Firstly, a segmented monotone decreasing edge composite function is proposed to accelerate the evolution of the level set curve in the gradient smooth region. Secondly, Canny edge detection operator gradient is introduced into the model as the global information. In the process of the evolution of the level set function, the guidance information of the energy function is used to guide the curve evolution according to the local information of the image, and the smooth contour curve is obtained. And the main direction of the evolution of the level set curve is controlled according to the global gradient information, which effectively overcomes the local minima in the process of the evolution of the level set function. Finally, the Heaviside function is introduced into the energy function to smooth the contours of the motion and to increase the penalty function to calibrate the deviation of the level set function so that the level set is smooth and closed. The results showed that the model of tobacco leaf edge profile curve could be obtained in the model of tobacco leaf covered by bare soil. In the complex background, the model can segment the leaves of the tobacco with uneven illumination, shadow and weed background, and it is better to realize the ideal extraction of the edge of the blade.

[6] Guru, D.S., et al. 2011. Min-max Representation of Features for Grading Cured Tobacco Leaves. Statistics and Applications 9 (12): 1529.

The authors have proposed that A grading system based on image processing techniques is developed for automatic inspection and grading of tobacco leaves. The system used machine vision in extraction and analysis of color, size, shape, surface texture and vein. A two-dimensional feature space is proposed to express feature distribution of tobacco leaves. The space is found to be well confined in an elliptic region. A database is constructed to record the feature distribution of standard contrast tobacco leaves prepared by experts through visual evaluation. The decision on grades is made based on the so-called nearest-neighbor method for which the over all difference among features between the measured tobacco leaves and the standard contrast samples were used as a target parameter for judgment. This system can be easily trained by users with the knowledge of the feature distribution information of different tobacco leaves.

[7] Prasad, V., T.S. Rao, and M. Babu. 2016. Thyroid Disease Diagnosis Via Hy brid Architecture Composing Rough Data Sets Theory and Machine Learning Algorithms. Soft Computing 20 (3): 11791189.

The authors have proposed that Capturing of a leave on the tree.Converted to gray scale image and then scaled the size to minimize the training time of neural network. After that, noise is added in order to create efficient and different input data for the neural network. Thus, the network has input layer with neurons. After several experiments, it has been found the optimum range for the number of neurons for the hidden layers. Number of output neurons were used in order to recognize number of leaves types.

[8] Lawrence, Steve, et al. 1997. Face Recognition: A Convolutional Neural-Network Approach. Neural Networks, IEEE Transactions on 8 (1): 98113

The authors have proposed that An image processing system of tobacco leaves grading is Actually, the lighted cabinet is the same one as Zhang has used , but all other equipment's, such as computer and camera,has been updated. The image processing system is consisted of a color camera.

All the authors mentioned above have done grading of different fruits and leaves.

In the similar way we are grading the flue-cured tobacco leaves based on the digital image processing and the fuzzy sets theory in this project.

CHAPTER 4

METHODOLOGY

4.1 Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. The aim of machine learning is to allow the computers learn automatically without human intervention and adjust actions accordingly.

Machine learning tasks are typically classified into several broad categories:

4.1.1 Supervised learning : The computer is presented with example inputs and their desired outputs, given by a teacher, and the goal is to learn a general rule that maps inputs to outputs. As special cases, the input signal can be only partially available, or restricted to special feedback.

4.1.2 Semi-supervised learning : The computer is given only an incomplete training signal: a training set with some (often many) of the target outputs missing.

4.1.3 Active learning : The computer can only obtain training labels for a limited set of instances (based on a budget), and also has to optimize its choice of objects to acquire labels for. When used interactively, these can be presented to the user for labeling.

4.1.4 Unsupervised learning : No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself

4.1.5 Reinforcement learning : Data (in form of rewards and punishments) are given only as feedback to the programs actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.

4.2 KNN

K Nearest Neighbor (KNN from now on) is one of those algorithms that are very simple to understand but works incredibly well in practice.This is why it is called the K Nearest Neighbours algorithm.

Most people learn the algorithm and do not use it much which is a pity as a clever use of KNN can make things very simple. It also might surprise many to know that KNN is one of the top 10 data mining algorithms.

The purpose of the k Nearest Neighbors (KNN) algorithm is to use a database in which the data points are separated into several separate classes to predict the classification of a new sample point.

Also it is surprisingly versatile and its applications range from vision to proteins to computational geometry to graphs and so on. We consider each of the characteristics in our training set as a different dimension in some space, and take the value an observation has for this characteristic to be its coordinate in that dimension, so getting a set of points in space.

Can then consider the similarity of two points to be the distance between them I n this space under some appropriate metric. Way in which the algorithm decides which of the points from the training set are similar enough to be considered when choosing the class to predict for a new observation is to pick the k closest data points to the new observation, and to take the most common class among these.

4.3 Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. In July 2018, Van Rossum stepped down as the leader in the language community after 30 years.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a communitybased development model, as do nearly all of Pythons other implementations. Python and CPython are managed by the non-profit Python Software Foundation.

Python uses dynamic typing, and a combination of reference counting and a cycledetecting garbage collector for memory management. It also features dynamic name resolution (late binding), which binds method and variable names during program execution.

4.4 Image Processing

Image processing basically includes the following three steps:

• Importing the image via image acquisition tools.

• Analysing and manipulating the image.

• Output in which result can be altered image or report that is based on image analysis.

4.4.1 Image acquisition involves capturing the images with the help of digital camera.

4.4.2 Image Pre- processing is carried out to improve the quality of the image and remove the unwanted noise in image followed by clipping and smoothing of the image. The image enhancement is carried out to increase the contrast. The RGB images are converted into grey images using colour conversion techniques

4.5 Feature Extraction

Features are the information that are extracted from an image. These are realvalued numbers (integers, float or binary). There are a wider range of feature extraction algorithms in Computer Vision. When deciding about the features that could quantify leaves, we could possibly think of Color, Texture and Shape as the primary ones. This is an obvious choice to globally quantify and represent the leaves.

4.5.1 Global Feature Descriptors

These are the feature descriptors that quantifies an image globally. These dont have the concept of interest points and thus, takes in the entire image for processing. Some of the commonly used global feature descriptors are

• Color - Color Channel Statistics (Mean, Standard Deviation) and Color Histogram.

• Shape - Hu Moments, Zernike Moments. • Texture - Haralick Texture, Local Binary Patterns (LBP)

4.5.2 Local Feature Descriptors

These are the feature descriptors that quantifies local regions of an image. Interest points are determined in the entire image and image patches/regions surrounding those interest points are considered for analysis. Some of the commonly used local feature descriptors are

• SIFT (Scale Invariant Feature Transform)

• SURF (Speeded Up Robust Features)

• ORB (Oriented Fast and Rotated BRIEF)

4.5.3 Combining Global Features

There are two popular ways to combine these feature vectors. For global feature vectors, we just concatenate each feature vector to form a single global feature vector. This is the approach we will be using in this tutorial. For local feature vectors as well as combination of global and local feature vectors, we need something called as Bag of Visual Words (BOVW). This approach is not discussed in this tutorial, but there are lots of resources to learn this technique. Normally, it uses Vocabulory builder, K-Means clustering, Linear SVM, and Td-Idf vectorization.

4.6 Functions for global feature descriptors

4.6.1 Hu Moments

To extract Hu Moments features from the image, we use cv2.HuMoments() function provided by OpenCV. The argument to this function is the moments of the image cv2.moments() flatenned. It means we compute the moments of the image and convert it to a vector using flatten(). Before doing that, we convert our color image into a grayscale image as moments expect images to be grayscale.

4.6.2 Haralick Textures

To extract Haralick Texture features from the image, we make use of mahotas library. The function we will be using is mahotas.features.haralick(). Before doing that, we convert our color image into a grayscale image as haralick feature descriptor expect images to be grayscale.

4.6.3 Color Histogram

To extract Color Histogram features from the image, we use cv2.calcHist() function provided by OpenCV. The arguments it expects are the image, channels, mask, histSize (bins) and ranges for each channel [typically 0-256). We then normalize the histogram using normalize() function of OpenCV and return a flattened version of this normalized matrix using flatten().

4.7 Training classifiers

After extracting, concatenating and saving global features and labels from our training dataset, its time to train our system. To do that, we need to create our Machine Learning models. For creating our machine learning models, we take the help of scikit-learn.

We will choose Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Decision Trees, Random Forests, Gaussian Naive Bayes and Support Vector Machine as our machine learning models. To understand these algorithms, please go through Professor Andrew NGs amazing Machine Learning course at Coursera or you could look into this awesome playlist of Dr.Noureddin Sadawi.

Furthermore, we will use train test split function provided by scikit learn to split our training dataset into train data and test data. By this way, we train the models with the train data and test the trained model with the unseen test data. The split size is decided by the test size parameter.

We import all the necessary libraries to work with and create a models list. This list will have all our machine learning models that will get trained with our locally stored features. During import of our features from the locally saved .h5 file format, it is always a good practice to check its shape. To do that, we make use of np.array() function to convert the .h5 data into a numpy array and then print its shape. Finally, we train each of our machine learning model and check the cross-validation results.

4.8 Libraries and Tools

4.8.1 Spyder:

Spyder, the Scientific Python Development Environment, is a free integrated development environment (IDE) that is included with Anaconda. It includes editing, interactive testing, debugging and introspection features.

Spyder is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts.

It features a unique combination of the advanced editing, analysis, debugging and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection and beautiful visualization capabilities of a scientific package.

4.8.2 OpenCV:

OpenCV supports a wide variety of programming languages such as C++, Python, Java, etc., and is available on different platforms including Windows, Linux, OS X, Android, and iOS. Interfaces for high-speed GPU operations based on CUDA and OpenCL are also under active development.

OpenCV-Python is the Python API for OpenCV, combining the best qualities of the OpenCV C++ API and the Python language.

OpenCV-Python makes use of Numpy, which is a highly optimized library for numerical operations with a MATLAB-style syntax. All the OpenCV array structures are converted to and from Numpy arrays. This also makes it easier to integrate with other libraries that use Numpy such as SciPy and Matplotlib.

CHAPTER 5

OBJECTIVES

Human eyes may fail to recognize the minute differences in color, shape or texture of the leaf. But the system is developed to recognize quality of leaves more accurately. The separation of cured Tobacco leaves manually consumes more time and also it requires human resource. If it is done by the system, then the accuracy will be improved and without any manual intervention the quality of the leaves may be detected.

The system classifies the leaves as Quality 1, Quality 2 based on shape, colour and texture. As the quality of the leaf will be known prior, one can have idea about further processes and curing leaves. As the leaves are classified by the system, Time taken by the system will be very less. System may minimize the errors which could be common in manual method of classification. This system reduces human efforts and manual intervention.It will be very useful for improving the level and efficiency of tobacco grading, arbitrating the dispute of the quality of tobacco leaves between buyer and seller Based on this result, one can decide how much temperature is given to a particular leaf. Increasing marketing efficiency of Tobacco leaves hence increasing the India’s economy coming from tobacco production.

CHAPTER 6

SYSTEM DESIGN

6.1 Existing System

In this System using support vector machine method for classification and classification made based on color. The main disadvantage of the SVM algorithm is that it has several key parameters that need to be set correctly to achieve the best classification results for any given problem. Parameters that may result in an excellent classification accuracy for problem A, may result in a poor classification accuracy for problem B. The user may, therefore, have to experiment with a number of different parameter settings in order to achieve a satisfactory result. The main parameters that the user should experiment with are the SVM kernel type (which can be set by the set Kernel Type (UINT kernel Type) method), the SVM type (which can be set by the set SVM Type (UINT svm Type) method), and the kernel-specific parameters (such as gamma, degree, nu, etc.).

Choosing a good kernel function is not easy. Long training time on large data sets Difficult to understand and interpret the final model, variable weights and individual impact Since the final model is not so easy to see, we can not do small calibrations to the model hence its tough to incorporate our business logic

6.3 Proposed System

The method of procedure in this project is summarized in Figure 6.2. Given the statement of the problem and definition of the aim of the project, as stated in introduction,In this System using k nearest neighbour method for classification and classification made based on color,texture and shape.

Each leaf was then prepared and photographed under suitable and stringently consistent conditions of lighting and scale. Later, each photographic image was digitized for use in the computer classifier algorithms. To ensure the highest standards of consistency, images were pre-processed prior to the isolation (segmentation) of the main objects in each image.

Once each object had been unambiguously recognized, features could be extracted from it. Such features may be as simple to extract- as, say, the length or pixel area of the segmented object, or they may be. The results of quite extensive mathematical derivation. Whichever the case, the process of choosing appropriate features for a given classification purposes is notoriously heuristic and often adhoc and this dissertation will cover in detail the reasons behind the feature choices that were made in this project.

Arising from these results, it has been possible to give some objective criteria for the grading of tobacco leaves by colour,shape and texture and to draw certain conclusions regarding the feasibility of the machine vision grading of flue-cured tobacco leaves.

6.3 SYSTEM ARCHITECTURE

Training phase:

Testing Phase:

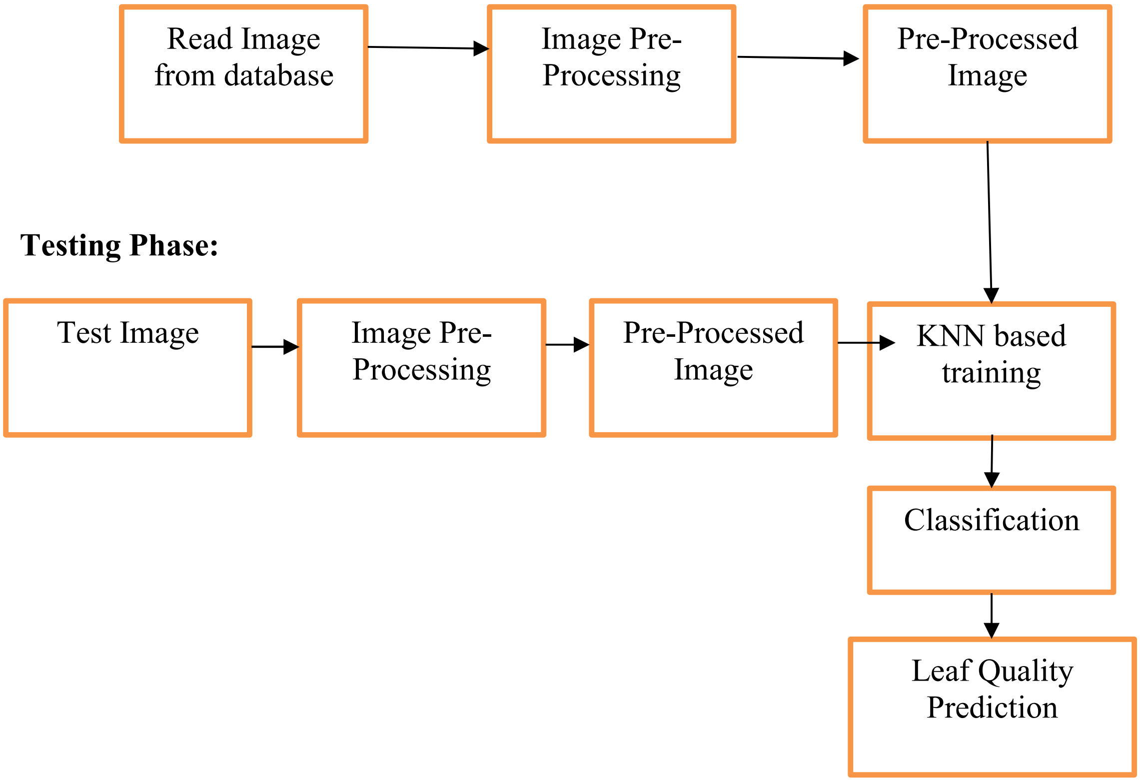

Fig. 6.1 System Architecture

In training phase first we read the images that are stored in the database.The image will be pre-processed and that image will be sent to testing phase.

In testing phase test image will be read from the database.This image will be sent to image pre-processing.Then pre-processed test image and pre processed training image both are sent to KNN classification.Then the KNN algorithm will classify the images and from this classification quality of the leaves are predicted.

CHAPTER 7

REQUIREMENTS SPECIFICATION

A System Requirements Specification (SRS) (also known as a Software Requirements Specification) is a document or set of documentation that describes the features and behaviour of a system or software application. It includes a variety of elements (see below) that attempts to define the intended functionality required by the customer to satisfy their different users.

Depending on the methodology employed the level of formality and detail in the SRS will vary, but in general a SRS should include a description of the SRS will vary, but in general a SRS should include a description of the functional requirements, system requirements, constraints, assumptions and acceptance criteria.

7.1 Functional Requirements

Functional requirements may involve calculations, technical details, data manipulation and processing, and other specific functionality that define what a system is supposed to accomplish. Functional requirements are supported by non-functional requirements (also known as ”quality requirements”), which impose constraints on the design or implementation (such as performance requirements, security, or reliability). Generally, functional requirements are expressed in the form ”system must do requirement”. Functional requirements specify particular results of a system. This should be contrasted with non-functional requirements, which specify overall characteristics such as cost and reliability. Functional requirements drive the application architecture of a system.

7.2 Non-functional Requirements

A non-functional requirement (NFR) is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. They are contrasted with functional requirements that define specific behavior or functions.The plan for implementing non-functional requirements is detailed in the system architecture, because they are usually architecturally significant requirements. Non-functional requirements are in the form of ”system shall be “requirement”, an overall property of the system as a whole or of a particular aspect and not a specific function. The system’s overall properties commonly mark the difference between whether the development project has succeeded or failed.Non-functional requirements are often called ”quality attributes” of a system. Other terms for non-functional requirements are “qualities”, “quality goals”, “quality of service requirements”, “constraints”, “non behavioral requirements”, or “technical requirements”.

7.3 Hardware requirement

- Processor : Pentium or above
- Processor Speed : 1.6Ghz and above
- RAM : 4GB and above
- Storage Space : Approx. 5 GB

7.4 Software requirements

- Operating System : Windows7 & above or Linux
- Language : Python
- Libraries : Python Libraries
- IDE : Anaconda / Spyder

7.5 User Requirements

- Throughout knowledge of the compiling python programs.
- Knowledge of converting train data into dataset.