Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.778117

EXTRACTION OF STATISTICAL FEATURES THROUGH PCA FOR IMPROVING THE CLASSIFICATION ACCURACY IN LEAF DISEASE DETECTION

Dr M.Banu Priya[1], Dr Praveenkumar G D[2], Dr B.Arivazhagan[3]
[1,2,3] Assistant Professor
[1]Department of Computer Science, 2School of Computer Science
[1]P.K.R Arts College for Women, Gobichettipalayam , [2,3]VET Institute of Arts and Science (Co-Ed) College, Erode
banupriyam@pkrarts.org, erodegd@gmail.com, arivazhaganphd2019@gmail.com

ABSTRACT:
Leaf disease detection is a critical task in agriculture as it enables early identification and intervention, leading to improved crop yield and reduced economic losses. However, accurate and timely detection of leaf diseases remains a challenging problem due to the complexity of leaf images and the presence of various environmental factors. This research proposes a novel approach to improve the classification accuracy in leaf disease detection by utilizing Principal Component Analysis (PCA) for the extraction of statistical features from leaf images. PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space, retaining the most important information while discarding the less significant variance. To achieve accurate and efficient disease classification, the extraction of meaningful statistical features from leaf images is crucial. Principal Component Analysis (PCA) is a widely-used dimensionality reduction technique that can aid in improving the classification accuracy by reducing the feature space.

1.INTRODUCTION

Although novel plant discoveries and also the digitization for species of plant processing are becoming increasingly widespread, leaf disease detection is becoming exceedingly challenging in biological and agricultural. Recognition and identification have become a mechanism in which each specific plant being assigned to a lateral sequence of similar plants based on its specific features. Since botanists have been doing much of the work, the procedure takes a long time. Owing to a shortage of appropriate templates or portrayal systems, a wide range of species of plant variants, including inaccurate picture data pre-processing strategies such as feature extraction and texture extraction, software- aided plant identification remains a difficult challenge in image analysis. The extraction of reliable features of the leaf has been the priority of automated leaf disease recognition.

The very next step under this research is feature extraction, which comes after noise reduction and segmenting the leaf image. The primary objective of this Feature Extraction is to extract leaf characteristics from image data to transform it into a format that enables similarities amongst leaf images. Feature extraction's goal is to extract the much more significant details from the source data and portray it in a higher-dimensional environment. Therefore, in this research, the PCA is used for extracting the features from the segmented leaf image, which uses respectively statistical and directional features to identify disease category in a leaf image.

2. PRINCIPAL COMPONENT ANALYSIS (PCA)

The PCA is a technique for reducing the dimension of a collection of data while maintaining its heterogeneity. Almost every collection of data includes information expressed as vectors of single parameters with integer, binary, or real values in most cases. A geometric point in three-dimensional space, e.g., may be described by a vector of three parameters each of which is aligned with one of the three coordinates axis x, y, and z.A sample could be described in particular by a vector made up of a set of variables. The length of the vectors found in the collection, and therefore the set's dimension, is determined by the number of parameters. Furthermore, a continuum of variability may be specified for each parameter that defines the range of values that only the individual parameter could accept. For example, if the data set includes three-dimensional points delimited by a cube with side 1 and centered in (0, 0, 0), the three parameters describing Cartesian coordinates were bounds to [-1/2, 1/2]. The spectrum of heterogeneity of the three parameters is described by this interval.

The objective of PCA is always to uncover hidden patterns in data and turn them in just such a manner that its differences and similarities are exposed. When the patterns have been discovered, the data may be interpreted as components that are sorted by importance, allowing low-level components to be discarded without losing valuable details. When the actual parameters chosen to describe the data are associated; PCA will minimize the dimension of a collection of data. Now let reconsider the illustration of the three parameters describing the three positions of points inside a cube to further comprehend this principle. The three parameters are associated if, for example, any of the points in the collection fall on an appropriate plane. As PCA has been used to solve this issue, one of the three parameters is transformed into a void parameter. The points in the fresh transformed space could consequently be defined by just two parameters, resulting in a space with a smaller dimension than the first. Since the points are in a two-dimensional region, the details about the third dimension, which is the rejected dimension, are meaningless. This is an oversimplification of the case. The explanations that follow go into the PCA process in greater depth.

Imply that perhaps the data collection under consideration includes points in a two-dimensional region with coordinates of (-2, -1), (-1, 0), (0, 1), (1, 2), (2, 3). The values of x differ in the range [-2, 2], whereas the values of y vary in the range [-1, 3]. The variation of the parameters x and y is described by these two intervals. These two factors are associated, as can be shown. As the y coordinates rise, the x coordinates rise as well, and a straight line runs between them all. As a result, if one of the two coordinates is identified, the other may be obtained.PCA aims to convert certain parameters so that they are no longer associated. By achieving this, the dimension of the collection of data could be minimized by only considering the parameters with the highest uncertainty and discarding the others. Principal-Components (PC) are still the factors with the most variability. Maybe the first PC could have been considered to reflect the results since they are normally ordered by their heterogeneity. The fact that perhaps a lower order PC exhibits lower variance within the ensemble however does not mean this is irrelevant in regression models.

3. Extraction of Statistical Features through PCA

PCA is most commonly used for locating a lower-dimensional representation of data. That has 2 distinguishing characteristics. During computing, it initially keeps shrinking the measurements of the provided data to a rational and accurate scale. Foremost, it separates the number of distinguishing features from the data input in a quiet manner that the overall dimension is reduced. The important feature characteristics were always present and could be used to identify the actual input details. The covariance-matrix could be found again from the matrix through leveraging the collection of features. The Eigen-values are then calculated using this covariance- matrix. Eigen-vectors were effective in representing complete databases in their nature. Merely some small Eigen-values were considered toward being substantially preferable and greater in importance, whereas the others are substantially quite minimal, and its exposure to data variations is indeed quite limited. As a result, after computing the inner product of the data well with respective Eigen-vectors for its respective Eigen-values, the preferred and greater variance paths were simply maintained.

The following are the general measures of PCA methodology:

Step-1: Compute the covariance-matrix of the specified data input using the following formula:

$$\sum V = \frac{1}{Num}\{(diag(m) - \overline{diag(m)})(diag(n) - \overline{diag(n)})^{T}\}$$

Where, 1≤ m, n ≤ Num

Step-2: As a result of the Eigen-vector matrix (V) and diagonal-matrix (D) of computed Eigen-values,

$$V^{-1}\sum V = D$$

Step-3: To achieve the PC parameter, organize the Eigen-vectors in decreasing order with the accompanying magnitude of Eigen-values.

Step-4: At last, data is transformed in the form of PCs by measuring the inner product of data with relevant Eigen-vectors.

Through specific, the PCA of a given vector ‘v’ associated with the group ‘Y’ is achieved by mapping vector v onto the sub spaces with the length or gaps of corresponding e' Eigen-vectors that relate to the top e' Eigen-values of the auto-correlation matrix ‘R’ in downwards sequence, where e' is lower than e. The above transformation generates a vector of e' coefficients c1,..., ce'. Even so, a linear structure of the Eigen-vectors with its corresponding weights c1,..., ce' is used to describe the vector v. These have segmented images with a resolution of 256x256 as forwarding from segmentation at this point of research. It even gets 32x32 sub-bands of resolution here. Such sub-band images are most often used in PCA to extract PCA parameter values. This PCA function is also used to determine the 13 statistical features mentioned below,

The statistical features calculated as:

Mean

The mean of a signal is calculated. The ‘x0’ through ‘xN-1’ comprise the signal, the 'i' is an iterator that passes through all these measurements, and μ which is the mean.

$$Mean\ \mu = \frac{1}{N} \sum_{i=0}^{N-1} x$$

Standard-Deviation (SD)

The SD of a signal is calculated. The signal has been contained in the variable ‘xi’, where ‘μ’ is the mean, ‘N’ represents the total of trials, and ‘σ’ is the SD.

$$SD\ \sigma^2 = \frac{1}{N-1} \sum_{i=0}^{N-1} (x_i - \mu)^2$$

Entropy

The entropy of a textured image is an indicator of its spatial disorder. It shows which kind of texture is statistically more unstable, with higher entropy indicating strong textures and lower entropy indicating smoother textures.

$$Ent = -\sum_i \sum_j (i,j) \log(P(i,j))$$

Root Mean Square:

The square root of the arithmetic mean of the squares of the numbers, or the square of the feature that determines the continuous waveform, is the RMS value of a series of values or a continuous-time waveform.

$$RMS = \sqrt{\frac{1}{n}} (x_1^2 + x_2^2 + \cdots + x_n^2)$$

Variance:

The amplitude difference across the mean is described as a variance.

$$Var = \sum_i \sum (i - \mu)^2\, p(i,j)$$

Smoothness:

A function's smoothness is defined by the number of constant derivatives would have over a given domain. A function is called "smooth" if it is distinguishable anywhere at the very least.

$$Smooth\ (x) = \begin{cases} x \geq 0, \\ 0\ if x < 0 \end{cases}$$

Kurtosis:

A histogram's flatness is measured as kurtosis.

$$Kurt = \frac{\mu_4}{\sigma^4}$$

Skewness:

A histogram of the image may be distorted up or down norm, depending on whether the Skewness is negative or positive.

$$Skew = (\mu - v/\sigma)$$

Inverse-Difference-Moment (IDM):

The IDM is a tool for determining image homogeneity. Whenever the majority of the instances in GLCM are distributed along the main diagonal, this parameter reaches its maximum value. GLCM contrast is inversely related to IDM.

$$Idm = \sum_i \sum_j$$

Contrast:

The local grey level difference of an image is known as a contrast. This is also known as the linear relationship between adjacent pixels' grey levels.

$$Cont = \sum_{n=0}^{Ng-1} \{ \sum_{i=1}^{Ng} \sum_{\substack{j=1 \\ |i-1|=n}}^{N_g} p(i,j) \}$$

Correlation:

The linear grey level dependency amongst pixels inside an image is represented by correlation. It also denotes interdependence at particular pixel locations.

$$Cor = \frac{\sum_i \sum_j (i,j) P(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

Energy:

The amount of energy in a texture indicates how uniform it is. It tests the localized homogeneity of textures and is, therefore, the polar opposite of Entropy.

$$E = \int_{-\infty}^{\infty} |(t)|^2 dt$$

Homogeneity:

The uniformity metrics of the non-zero entrants in the image are referred to as homogeneity. The opposite of Contrast weight is Homogeneity weight, in which the smaller the homogeneity, the larger the Contrast.

$$Hom = \sum_{i=0}^{2^n} \overline{\sum_{j=0}^{2^n}} \frac{1}{1+(ij)} \times p_{t(i,j)}$$

Figure 6.1 shows the initialization of the feature extraction phase. Figure 6.2 shows the 13 statistical features from the PCA. The 13 statistical features from the segmented image are convoluted in this research by PCA to 3 principal component template models, as shown in Figure 3.1.

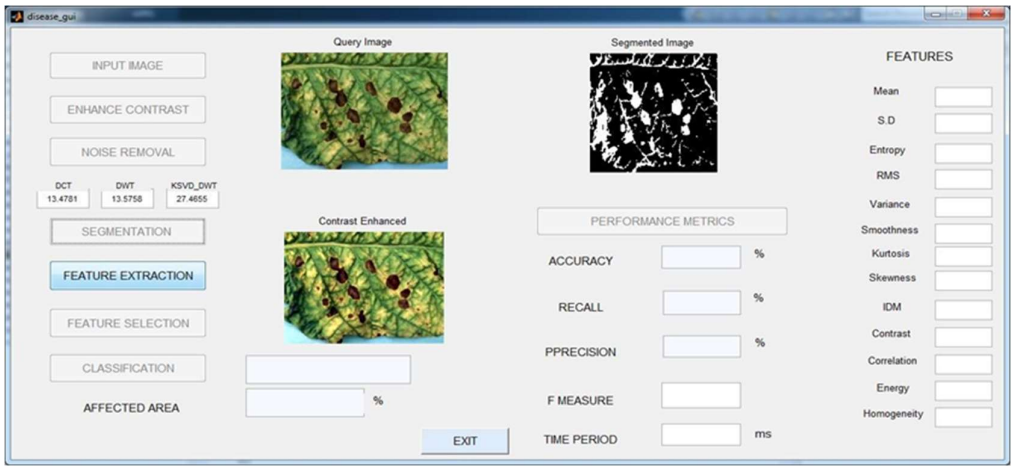

Figure 3.1 Feature Extraction Initialization

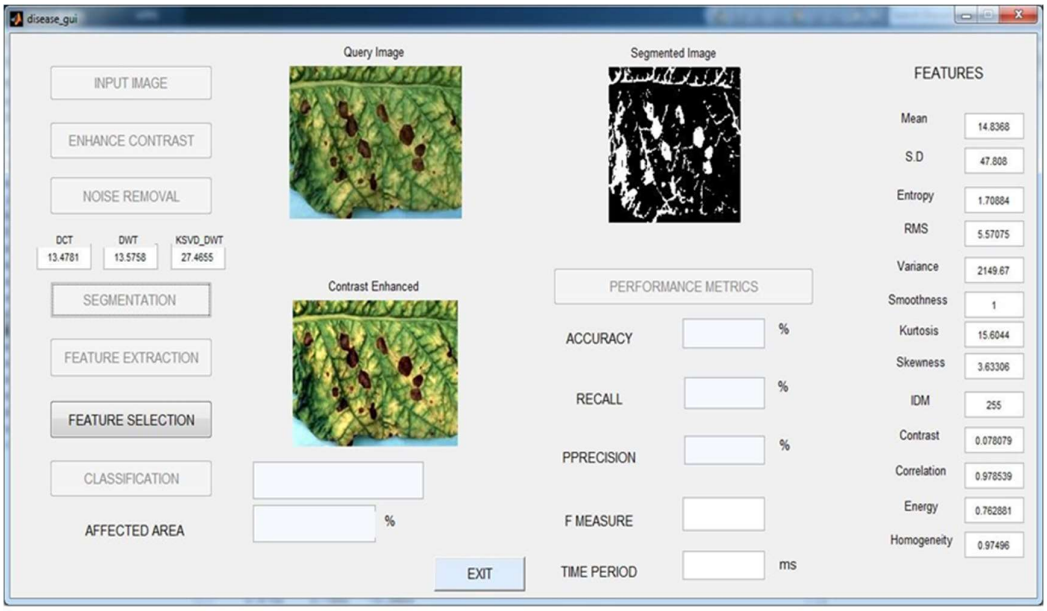

Figure 3.2 Statistical Features from PCA

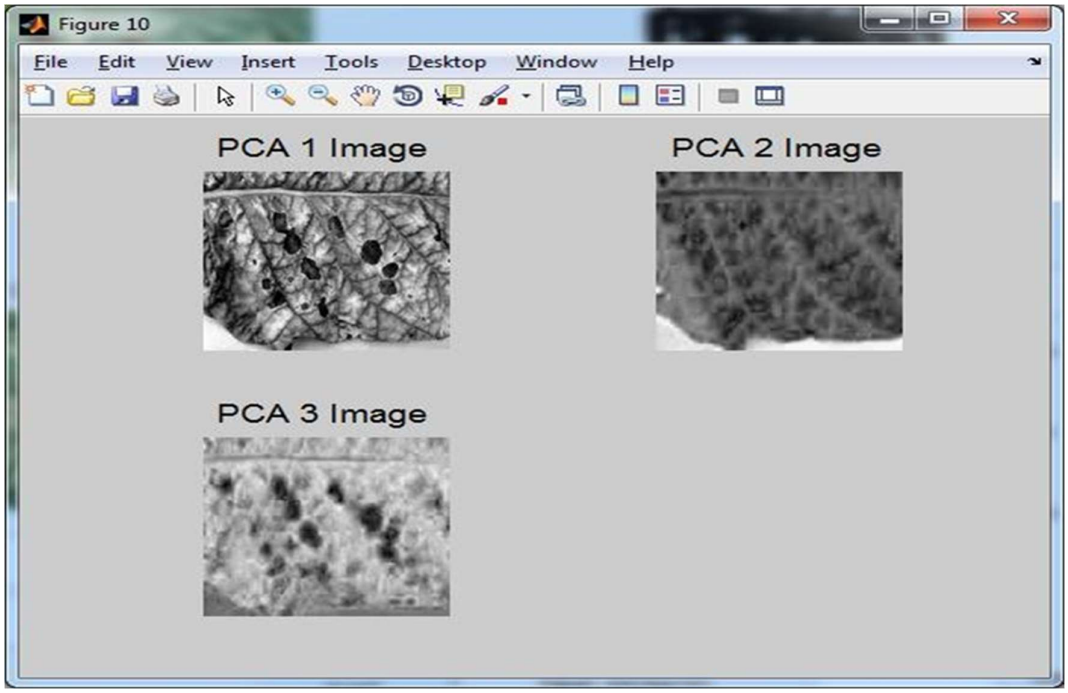

Figure 3.3 Principal Component Template Models

PCA has been used to extract 13 statistical properties from the segmented leaf image and create 3 texture feature vector templates for classifying Leaf's Disease shown in figure 3.2,3.3. Each part of this chapter discusses the strategies and their outcomes to improving the classification accuracy for Leaf Disease detection. Experimental results demonstrate that our PCA-based approach significantly improves the classification accuracy of leaf disease detection compared to traditional methods. The reduced feature space obtained through PCA retains essential information while eliminating redundant features, leading to enhanced discrimination between healthy and diseased leaves. Furthermore, the reduced feature set enables faster processing and reduces computational complexity, making it suitable for real-time applications.

The extraction of statistical features through Principal Component Analysis (PCA) has shown promising results in improving the classification accuracy in Leaf Disease detection. PCA is a dimensionality reduction technique that transforms the original features into a new set of orthogonal components, or principal components, which capture the maximum variance in the data. By applying PCA to the dataset of leaf disease images, we were able to reduce the dimensionality of the feature space while retaining the most important information. This reduction in dimensionality not only helps in overcoming the curse of dimensionality but also reduces the computational complexity of the classification process. The extracted principal components have demonstrated their ability to highlight the essential patterns and structures present in the leaf disease images, leading to better discrimination between healthy and diseased leaves. This has resulted in improved classification accuracy compared to using the original set of features. Furthermore, the use of PCA can also help in handling issues related to co linearity among features and remove redundant or noisy information that might negatively impact the classification performance. However, it is essential to note that the effectiveness of

PCA heavily depends on the choice of the number of principal components to retain. Selecting an appropriate number of components is crucial, as retaining too few may result in loss of important information, while retaining too many may lead to overfitting and increased computational overhead.

4.Conclusion

The proposed approach effectively addresses the challenges of leaf disease detection and provides a robust and efficient solution for accurate and early identification of leaf diseases in agricultural fields. The integration of PCA for statistical feature extraction demonstrates promising results and lays the foundation for further advancements in automated plant disease diagnosis systems. The proposed approach can be integrated into agricultural systems and crop management practices to aid in the timely identification and control of leaf diseases, thereby contributing to sustainable agriculture and food security.

References

[1] Jain, S.; Sahni, R.;Khargonkar, T.; Gupta, H.; Verma,O.P.; Sharma, T.K.; Bhardwaj, T.; Agarwal, S.; Kim, H. Automatic Rice Disease Detection and Assistance Framework Using Deep Learning and a Chatbot, Electronics 2022, volume 11, 2110.

[2] Jayanthi M G, D.R.Shashikumar , " A model for early detection of paddy leaf disease using optimized fuzzy inferencre system, IEEE-ICSST-2019, pages 206-211.

[3] K. Ferentinos. Deep learning models for plant disease detection and diagnosis, Comput. Electron. Agric. volume 145, 2018, pages311–318.

[4] G.D. Praveenkumar, R. Nagaraj. Regularized Anisotropic Filtered Tanimoto Indexive Deep Multilayer Perceptive Neural Network Learning for Effective Image Classification, Neuroscience Informatics, Elsevier, 2022, 100063.

[5] G.D.Praveenkumar, R.Nagaraj, Intelligent Adaptive Anisotropic Diffusion Filtered Deep Neural Network with Gaussian Activation Function for Image Classification, In.Con.Proc., IEEE-ICCMC-2022,Pages1377-1382,2022.

[6] G.D.Praveenkumar , M.Kiruthika , S.Tamilselvi ,Image Recognition of Paddy Leaf Disease Identification and Classification Using Deep Neural Network Model, Industrial Engineering Journal, ISSN: 0970-2555 ,Volume : 52, Issue 6, No. 3, June : 2023