Automated Plant Disease Analysis (APDA): Performance Comparison of Machine Learning Techniques

Asma Akhtar, Aasia Khanum, Shoab A. Khan, Arslan Shaukat
College of Electrical and Mechanical Engineering
National University of Sciences & Technology (NUST)
Islamabad, Pakistan
{asmaakhtar, aasia, shoabak, arslanshaukat}@ceme.nust.edu.pk

***Abstract*— Plant disease analysis is one of the critical tasks in the field of agriculture. Automatic identification and classification of plant diseases can be supportive to agriculture yield maximization. In this paper we compare performance of several Machine Learning techniques for identifying and classifying plant disease patterns from leaf images. A three-phase framework has been implemented for this purpose. First, image segmentation is performed to identify the diseased regions. Then, features are extracted from segmented regions using standard feature extraction techniques. These features are then used for classification into disease type. Experimental results indicate that our proposed technique is significantly better than other techniques used for Plant Disease Identification and Support Vector Machines outperforms other techniques for classification of diseases.**

Keywords- Machine Learning; Artificial Intelligence; Classification; Plant Disease Analysis

I. INTRODUCTION

Growing world population has brought a lot of pressure on agricultural resources. It is imperative to obtain maximum yield from crop in order to sustain the population and the economy. Plant diseases are the main source of plant damage which cause economic and production losses in agricultural areas. Owing to distressed climatic and environmental conditions, occurrence of plant diseases is on the rise.

There are various types of diseases in plants, variety of symptoms such as spots or smudge arising on the plant leaves, seeds and stanches of the plant. In order to manage these diseases effectively, there is a need to introduce automatic method of plant surveillance that can scrutinize plant conditions and apply knowledge-based solutions to detect and classify various diseases. Machine Learning is an intrinsically appropriate framework to support this problem. A variety of techniques have been proposed recently for identification and classification of plant diseases from images using Machine Learning. While these automated techniques have paved way for remote monitoring and expert surveillance of plant diseases, there are challenges of accuracy and robustness that need to be addressed for reaping practical benefits from these techniques. This paper presents experimental results of using various Machine Learning techniques for the task of plant disease identification and classification.

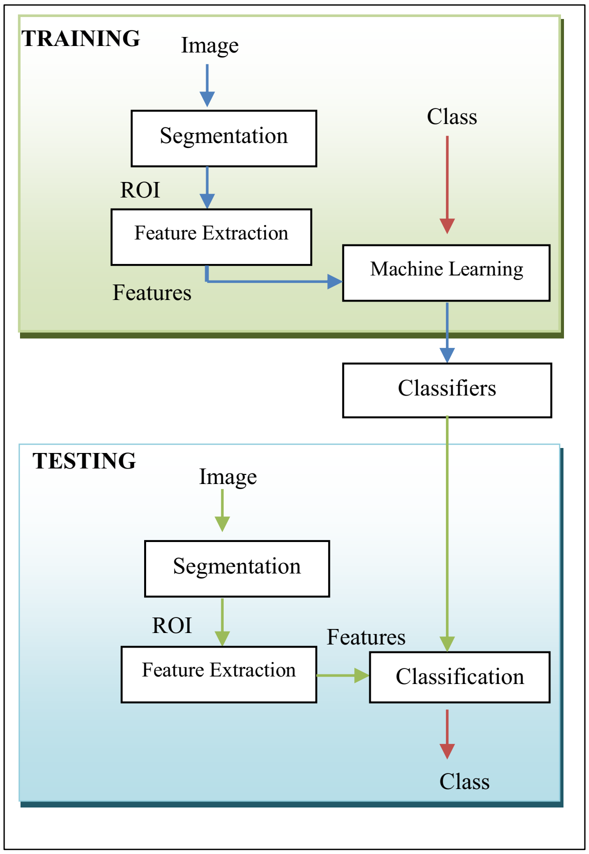

Figure.1 Machine Learning Framework for Automated Plant Disease Analysis

Fig.1 shows the basic setup of automated plant disease analysis using Machine Learning Techniques. In the training section, segmentation is performed from images and features extracted, these features are then used for classification. The paper is structured as follows: section II is an overview of related work in the field, section III explains the applied Machine Learning techniques. Section

978-1-4799-2293-2/13 $31.00 © 2013 IEEE
DOI 10.1109/FIT.2013.19

IV presents the experimental approach, section V discusses the results, and finally section VI concludes the paper.

II. Related Work

Different segmentation and classification techniques have been proposed by different researchers. Some of these are listed below:

Al-Hiary et al [5] has proposed a research work in which he present algorithm for fast and accurate detection and classification of plant diseases. The algorithm starts by acquiring RGB images. In the next step color transformation is applied on RGB images. Images are then segmented using K-means clustering techniques. Texture features are extracted from segmented area using Gray Level Co-occurrence Matrix (GLCM). Neural Network is used as a classification tool. The overall accuracy of this technique is 94%.

Camargo et al [1] proposed an algorithm to identify plant diseases using image processing. First the color transformation of the acquired image is done. These transformed images are then enhanced using Gaussian filter. On this transformed image segmentation is performed in order to segment the disease regions separately by locating optimum threshold. Then the segmented regions are labeled as diseased.

Camargo et al [2] has further used this segmentation technique for classification purpose. Different features such as shape, texture, gray level, fractal dimension, histogram of frequencies are extracted from segmented region. Seven fold cross validation is used for the evaluation. Overall accuracy by combining useful features is 93.1% using Support Vector Machine (SVM) used for the classification purpose.

Zhao et al [4] proposed a method for recognition of maize leaf diseases using machine vision. Image segmentation is based on threshold. Freeman link code is used for feature calculation and diseases are deduced according to binary tree search method. The accuracy of five maize leaf diseases is above 80 %.

III. Machine Learning Techniques

Five (5) different Machine Learning techniques for learning classifier have been investigated in this paper. These techniques are selected due to the reason that these classifiers have performed well in many real applications.

A. K- Nearest Neighbor (KNN)

The K Nearest Neighbor is a kind of lazy learner which means that this classifier train and test at the same time. KNN classifier is instance based classifier that performs classification of unknown instances by relating unknown to known by using distance or similarity function. It takes K nearest points and then assigns class of majority to the unknown instance [11].

B. Naïve Bayes Classifier

Naïve Bayesian Classification is commonly known as a statistical classifier. Its foundation is on Bayes' Theorem, and uses probabilistic analysis for classification. Naïve Bayesian Classifier give more accurate results in less computation time when applied to the large data sets. [15]

C. Support Vector Machine (SVM)

Support Vector Machine is machine learning technique which is basically used for classification. It is a kernel based classifier; initially it was developed for linear separation which was able to classify data into two classes only. SVM has been used for different realistic problems such as face recognition [10], cancer diagnosis [8] voice identification and glaucoma diagnosis.

D. Decision Tree

Decision Tree Classifiers (DTC's) are being successfully used in many areas including medical diagnosis, speech recognition, character recognition etc. Decision tree classifiers have ability to convert the complex decision into easy and understandable decisions. [7]

E. Recurrent Neural Networks

Recurrent Neural Networks (RNN) includes feedback connections. In contrast to feed-forward networks, the dynamical properties are more significant. Neural Network has evolvement within a constant state and the activation values of any units do not change anymore. But in some cases , according to required scenario it is important to change the activation value of the output neurons. [6]

IV. Experimental Setup and Procedure

The proposed system is multistep process as previously illustrated in Fig.1; various aspects are described follows

A. Image Dataset

Data set is prepared by ours manually and used for first time in this research. Rose leaf samples were acquired from Tea Research Institute, Mansehra. These leaves are divided into normal and diseased subsets. The diseased subset contains samples of two types of diseases: Anthracnose and Black Spots. The Nikon camera D90, which is 10 megapixels camera, is used for image acquisition purpose. Program mode is used for more detail green and blues. Aperture is normal about 35 and the lens used is 18-135 mm. Distance of object from lens is about 9 to 12 inches. Images acquired in indoor lighting. There are 40 images in the dataset.

Twenty (20) images of each disease are captured and stored in JPEG format.

B. Segmentation

Segmentation is used to separating the diseased region (ROI) from non-disease region in a leaf image. This is done by thresholding the gray scale leaf image such that all gray values below the threshold τ are represented as white and those above τ are represented as black. For the selection of threshold τ Otsu's algorithm [17] is applied. The result of this segmentation procedure is binary image where in the diseased region is represented with a white and non-diseased region is black. Morphological operators 'open' and 'fill' are used for further removing any misleading tiny dots.

Once the binary image has been obtained it is mapped with the original colored image to obtain a masked colored image where the diseased region is represented in color and rest of the image is black.

An illustrative example of all the above Image Processing steps is shown in Fig. 2:

i. Part (a) shows the original image of rose leaf infected by anthracnose and (b) shows original image of rose leaf infected black spot disease
ii. The segmentation process starts by taking the green component of an image. If the green component of an image is greater than 150 and less than 200 then its value is set as 0 (Black) and the pixel of green component which are already having value zero are set to 255(White) and the result is shown in Fig. 2 (c) (d).
iii. The next step is to apply threshold value to segment the disease region from non-disease area.
iv. The threshold value selected through experiments is 80 which give good segmentation results as shown in Fig. 2 (e) (f).

v. Simple thresh holding is not enough as there are many tiny dots which have to be removed so morphological operator open is used for removing small unconnected areas. There are different holes which has to be filled and for that purpose morphological operator fill is used and Fig. 2(g)(h) shows the result of applying open and fill morphological operators.
vi. After applying all the above steps the last step is to map the segmented image with original image and the result of this procedure can be seen in Fig. 2(i)(j).

C. Feature Extraction

In the proposed approach three different feature extraction techniques are used for extracting features from segmented leaf areas. These are Stataistical Features, Discrete Cosine Transform (DCT) and Discrete Wavelet Transform.

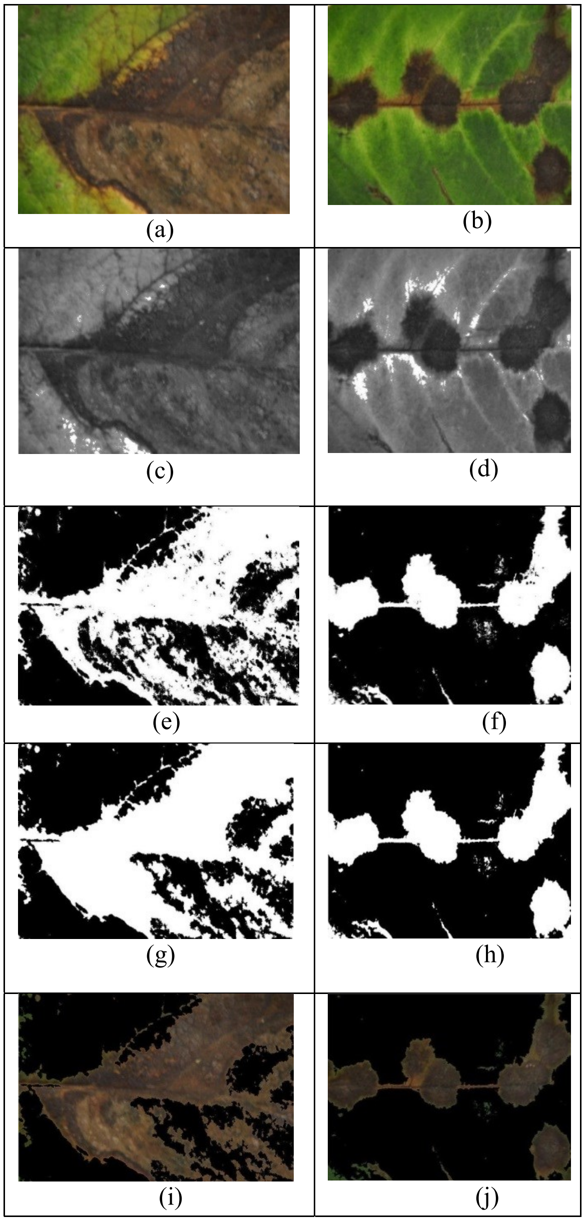

Figure. 2: (a) Anthracnose (b) black spot Original Images of diseases (c) (d) Extracting and thresholding green component (e) (f) Applying threshold=80 (g) (h) Applying morphological operators open and fill (i)(j)Mapping segmented image with original image

i. Statistical Feature

Statistical equations are applied on the dataset for analyzing and interpret the data set. Haralick texture features are used in the experimentation. These are spatial features that indicate pixel relationship based on gray scale intensity and orientation. [19] A total of 11 haralick features are used which are calculated using Gray Level Co-occurrence Matrix (GLCM). Table 1 shows the description of how each texture feature is calculated. In the equations, n represent the number of observed values. X is the sample space and P is the population.

ii. Discrete Cosine Transform (DCT)

DCT is a frequency domain method that is helpful in finding energy of various spectral sub-bands (blocks) of the image. The local DCT method uses a range of 2-Dimensional DCT to construct a feature vector of an image. Where f(x,y) is the intensity of the pixel in row x and column y, u=0,1,2,....N-1 and v=0,1,2,...M-1 .

Table 1: Description of Texture Features

Angularsecond moment(energy)	$\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}[p(i,j)]^2$ (1)
Corelation	$\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\frac{ijp(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$ (2)
Variance	$\sigma^2=\sum_{I=0}^{G-1}(i-\mu)^2p(i)$ (3)
Inverse difference	$\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\frac{p(i,j)}{1+(i-j)^2}$ (4)
Sum Avg	$\sum_{n=0}^{G-1}n^2p_{x-y}(n)$ (5)
Sum variance	$\sum_{i=2}^{2N_g}(i-f6)^2p_{x+y}(i)$ (6)
Sum entropy	$\sum_{i=2}^{2N_g}p_{x+y}(i)log(p_{x+y}(i))$ (7)
Entropy	$\sum_{I=0}^{G-1}\sum_{j=0}^{G-1}p(i,j)\log_2[p(i)]$ (8)
Diff variance	$\sum_{i=0}^{N_{g-1}}(i-\mu_{x-y})^2p_{x-y}(i)$ (9)
Diff entropy	$\sum_{i=2}^{2N_g}p_{x-y}(i)log(p_{x-y}(i))$ (10)
Information measure of correlation 1	$\left[\frac{f9-HXY1}{\max\{HX,HY\}}\right]$ (11)

$$C(u,v)=\alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}f(x,y)cos\left[\frac{\pi(2x+1)u}{2N}\right]cos\left[\frac{\pi(2y+1)v}{2N}\right] \quad (12)$$

here 0≤ u ≤ N, & 0≤ v ≤ N, and

$$a(u)=\begin{cases}\sqrt{\frac{1}{N}}\ for\ \ u=0\\ \sqrt{\frac{2}{N}}\ for\ \ u\neq 0\end{cases} \quad (13)$$

DCT features are extracted from each segmented image. For the selection process; rather than selecting randomly we have applied the zigzag scanning. In our research, a zigzag scanning is performed in order to select the coefficients rather than selecting randomly as shown in the Fig 3. In Zigzag Scanning DCT organize its coefficients in priority order. The first coefficient signifies the highest priority because of its highest variance.

The first main advantage of the DCT is its efficiency. In transforming the spatial domain into frequency domain, blocked DCT is used in which transformation is performed in efficient manner. Secondly, DCT works with entirely real-valued components, in terms of image compression. DCT has good de-correlation and energy compaction characteristics.

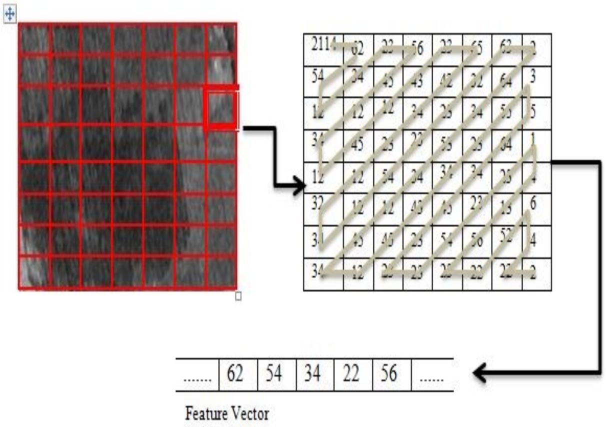

Figure 3: Zig Zag Scanning of Features [18]

iii. Discrete Wavelet Transform (DWT)

Wavelet transform decompose the signal into some basic functions known as wavelets. The capability for multi-resolution analysis wavelet transform can be used for analyzing details of an image at various scales. Detail components contain information of diagonal, vertical and horizontal sub-bands of the image. [13] This information can be extracted by using high pass and low pass filters. [14] **.** Our experimental setup used discrete wavelet transform for extracting features from diseases area of the leaf.

D. Classification

Classification is performed by using 5 different Machine Learning techniques K Nearest Neighbor , Support Vector Machine, Decision Tree , Naïve Bayesian and Recurrent Neural Network. The value of k = 1 used in K-NN and linear kernel used in SVM in our experments. The classifiers are tested with individual features (texture, DCT and DWT) as well as different feature combinations (DCT+DWT, DCT+Texture, DWT+Texture) and performance is evaluated by 10-fold cross validation.

V. RESULT AND DISCUSSION

MATLAB is used for the experimentation of proposed system. [20] Table 2 shows a comparison of the accuracy of five classifiers on individual features. It is clear that on the average, Decision Tree is the best classifier, followed at a close distance by KNN and SVM. Decision Tree gives its best performance on DCT features followed by DWT. KNN gives its best performance on DWT followed by Texture and SVM gives its best performance on Texture features followed by DWT. The best performing features on the average accuracy is Discrete Wavelet Transform (DWT) feature.

Table 2: Accuracy of Classifiers Using Individual Features

Features	DCT	DWT	Texture	Average
KNN	75 %	91.85%	75.45%	80.77%
Decision Tree	91.95%	83%	75%	83.31%
Naïve Bayes	66%	75%	50%	63.67%
RNN	66.25%	75%	83%	74.45%
SVM	75%	83.67%	90.45%	83.04%
Average	74.84%	81.70%	74.87%	

Table 3 : Accuracy of Combinition of Classifers

Features	DCT+ DWT	DWT+ Texture	DCT+ Texture	Average
KNN	91.25%	94%	66%	83.75%
Decision Tree	83.38%	75%	93%	83.78%
Naïve Bayes	83%	50%	50%	61%
RNN	75%	75%	91%	80.33%
SVM	94.45%	83%	75%	84.15%
Average	85.41%	75.4%	75%	

After performing classification separately on three feature extraction techniques we combined these features to check the accuracy and in the result accuracy is improved. Table 3 shows the accuracy of different feature combinations.

Table 4: Accuracy of Feature Combination

Reference	Segmentation	Features	Classifier	Accuracy
Camargo et.al, 2008 [1]	Threshold Segmentation	Shape+ Texture+ Dispersio+Grayleve+Histogram of Frequencies	SVM	93.1%
Al-Hiary et.al 2011 [5]	K-Means Clustering	Texture Features	Neural Networks	94%
Guru et.al,2011 [3]	Morphological Operation Technique	Texture Features	Probabilistic Neural Network	88.59%
Bashish et.al,2010 [16]	K-means Segmentation	Texture Features	Neural Networks	93%
Present Experimental Finding	Otsu Segmentation	DCT+DWT	SVM	94.45%

Four classifiers i.e. KNN, RNN, SVM, and Decision Tree show improvement in results. While KNN, RNN and SVM show clear improvement in average performance, Decision Tree has marginal improvement in average accuracy of DCT+DWT+SVM; whereas in case of individual features best accuracy was 91.95 % with DCT + Decision Tree. Comparing Table 2 and Table 3, the best choice of features is 85.41% with DCT+DWT, followed by 81.70% with DWT alone. Experimental results indicate that our technique is more accurate results as compare to other techniques for Plant Identification and Classification. As noticed the best average accuracy discovered in the experiments reported here is 94.45% with combination of DCT, DWT, and SVM.

The best accuracy, i.e. 94.45% is achieved by a combination We compare this finding with previous research results reported in section 2 above. Table 4 shows the comparison.

VI. Conclusion and Future Work

We have found that DCT, DWT and Texture feature extraction techniques give good results in classification. The proposed approach of combining DCT+DWT features for classification with Support Vector Machine (SVM) gives maximum accuracy of 94.45%. Our proposed technique indicates more enriched results as compare to other techniques for Plant Disease Identification and Classification. Dataset used in these experiments are very limited. In the future we will work with appropriate quantity of data for experiments.

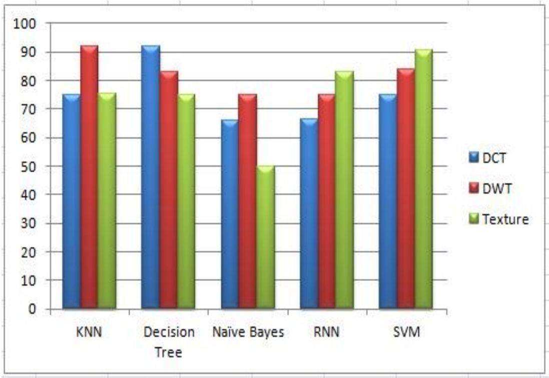

Figure 4: Accuracy of Classifiers using Individual Classifers

Accuracy can also be increase by enhancing the other Image processing and Classification techniques. We will also change the feature extraction techniques to improve the classification on this dataset.

ACKNOWLEDGEMENT

The authors would like to thank the Agronomist Dr. S.M. Mughal [1] for his help in collection of leaf dataset. The authors are also grateful to the reviewers for their valuable and appreciated comments and suggestions.

REFERENCES

[1] Camargo, A. and J. S. Smith. 2008. An image-processing based algorithm to automaticallyidentify plant disease visual symptoms. Bio.Sys. Eng., 102: 9 – 21.

[2] Camargo, A. and J. S. Smith. 2009. Image pattern classification for the identification of disease causing agents in plants. Com. Elect. Agr. 66: 121–125.

[3] Guru, D. S., P. B. Mallikarjuna and S. Manjunath. 2011. Segmentation and Classification of Tobacco Seedling Diseases. Proceedings of the Fourth Annual ACM Bangalore Conference.

[4] Zhao, Y. X., K. R. Wang, Z. Y. Bai, S. K. Li, R. Z. Xie and S. J. Gao. 2009. Research of Maize Leaf Disease Identifying Models Based Image Recognition. Crop Modeling and Decision Support.Tsinghua uni.press. Beiging. pp. 317-324.

[5] Al-Hiary, H., S. Bani-Ahmad, M. Reyalat, M. Braik and Z. ALRahamneh. 2011. Fast and Accurate Detection and Classification of Plant Diseases. Int. J. Com. App., 17(1): 31-38.

[6] PearlMutter, B. A. 1990. *Dynamic Recurrent Neural Network*

[7] Aly, M. 2005. *Survey on Multiclass Classification Methods*

[8] Fury, T. S., N. Cristianini and N. Duffy. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Proc. BioInfo., 16(10): 906-914.

[9] Scholkopf, B. and A. J. Smola. 2001. Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge.

[10] Huang, J., V. Blanz and B. Heisele. 2002. Face Recognition Using Component-Based SVM Classification and Morphable Models, pp. 334– 341.

[11] Mohammed J. Islam ., Q. M. Jonathan Wu, MajidAhmadi, A.Maher and Sid-Ahmed.2007. Investigating the Performance of Naive- Bayes Classifiers and K-Nearest Neighbor Classifiers.ICCI Proceedings of International Conference on Convergence Information Technology .IEEE Computer Society

[12] Bock, C. H., G. H. Poole, P. E. Parker and T. R. Gottwald. 2010. Plant Disease Severity Estimated Visually, by Digital Photography and Image Analysis, and by Hyperspectral Imaging. Cri. Rev. Pla. Sci., 29: 59–107.

[13] http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.com (Accessed: 25th April 2013)

[14] Naveed N., T. S., Choi and A .Jaffa .Malignancy and Abnormality Detection of Mammograms using DWT features and ensembling of classifiers, International Journal of the Physical Sciences ,Vol.6(8)

[15] Duda, R. O., P. E. Hart and D. G. Stork. 2001. *Pattern classification*, 2nd edition, John Wiley and Sons, New York.

[16] Al Bashish, D., M.Braik and S.Bani-Ahmad.2010. Frame work for detection and classification of plant leaf and stem diseases. Signal and Image Processing(ICSIP) international conference. pp.113 – 118

[17] Hongzhi, W., Ying, D,.2008. An Improved Image Segmentation Algorithm Based on Otsu Method. International Symposium on Photo electronic Detection and Imaging SPIE Vol. 6625

[18] Chung, K, L., Liu, Y, W,. and Yan, W, M., 2006. A hybrid grey image representation using spatial – and DCT – based approach with application to moment computation. Journal of Visual Communication and Image Representation Vol. 17, Issue 6.

[1] Dr S. M. Mughal, Professor at Plant Pathalogy Department, PMAS Arid Agriculture University Rawalpindi.