Chapter 5

Classification of Diseased Arecanut

5.1 Preamble

In this chapter, classification of diseased and undiseased arecanut has been addressed using texture features of Local Binary Pattern (LBP), Haar Wavelets, GLCM and Gabor wavelets. Four models for classification of diseased and undiseased arecanut are proposed. In first model, an RGB arecanut image has been transformed to HSV image. The saturation component of HSV color space is used for segmentation. Segmentation of arecanut has been done using threshold based Otsu segmentation method. A modified LPB has been applied on the value component of

Some parts of the material in this chapter have been published in the following Journals:

HSV color space. A histogram is generated from LBP image. Classification has been done based on correlation between histogram of training set and query sample. In second model, arecanut is segmented from a given image using threshold based Otsu method. The Haar Wavelets, GLCM and Gabor features are extracted from the components of HSV and YCbCr color models. The best discriminative subsets of features from the components of HSI and YCbCr color models are selected empirically based on combination of features.

In third model, symbolic data analysis has been presented. The symbolic models range method and three sigma control limits have been used for classification of diseased and undiseased arecanut. These symbolic data models give interval type of data. Features of the query sample are crisp values. These crisp values are compared with interval values of training set and query sample will be labeled based on majority of voting.

5.2 Related Work

Related work on color based histograms has been discussed in the section 2.2.2. Related work on feature selection and symbolic data analysis have been discussed in the sub sections 5.2.1 and 5.2.2.

5.2.1 Features Selection

Selection of information is one of the most challenging problems facing by the real-life applications of computational intelligence methods. Classification algorithms like Support Vector Machines, neural networks etc. need tools for feature selection and training vector selection to work efficiently. There are many advantages of applying such techniques: improve overall accuracy, speeding up the training process, and reducing computational complexity of the data model, making it more comprehensible. Real world problems may have thousands of irrelevant features and in such situations feature selection should provide high reduction rate preserving important information hidden in the full dataset. Selection methods should also be fast, especially if they are used in iterative wrapper schemes.
Feature selection methods may be of supervised or unsupervised type. Search for the best subset of $m$ features out of all $n$ features is NP hard (Guyon I. et al., 2006). For large $n$ exhaustive search testing all possible feature subsets is prohibitively expensive, and many global search techniques, stochastic methods and sequential search techniques have been developed and implemented in this area (Somol P. et al., 2002).

Feature selection methods may be classified into filters, wrappers and embedded approaches. Embedded methods are an integral part of specific predictors, for example decision trees or neural networks with regularization. Filters are completely independent of predictors, ranking features or selecting feature subset using some indices of relevance to measure the quality of selected feature subsets. Many statistical correlations, probability distributions, information theory and other types of measures are used to define filters (Duch W., 2006). These methods have low complexity and are universal, but are not always matched in an optimal way to a given predictor. Methods that check performance on various subsets wrapping themselves around particular predictors are called wrappers (Guyon I. et al., 2006). They may use incremental, decremental or mixed strategy to select feature subsets, but the necessity of repeatedly training and testing predictors each time a new subset is defined can be very time consuming. A combination of filters and wrappers, called frappers (Duch W., 2006), is also possible, using filters first for feature ranking and in the second stage adding new features in their ranking order and rejecting those that do not improve results of a given predictor.

The number of features captured in the data is very large. Authors used entropy-based (Fayyad U., 1993), a $\chi^2$-statistics, a correlation-based (Hall M.A., 1998), a t-statistics, and a correlation-based (Golub T. R. et al., 1999) feature selection method to filter out irrelevant features. The first three methods are commonly used in the machine learning community, while the last two are favored by works on gene expression analysis. The $\chi^2$-statistics and the correlation-based method are two refinements of the entropy method. The correlation method can be considered as a variant of the t-statistics. The only way to assure that the highest accuracy is obtained in practical problems is by testing a given classifier on a number of feature subsets.
5.2.2 Symbolic Data Analysis

An interesting approach would be to use symbolic data analysis (SDA) popularized by Bock and Diday (2000). Within this framework, interval data representation can be used to take into account the uncertainty and noise inherent to measurements (Billard, 2008).

Guru et al. (2004) suggested a novel similarity measure for estimating the degree of similarity between two patterns, described by interval type data. In particular, this measure computes the degree of similarity between two patterns and approximated the calculated similarity value by a multi-valued type data. Then, based on this similarity, the authors modified the agglomerative method by proposing the concept of mutual similarity value for clustering symbolic pattern. Yang et al. (2004) proposed fuzzy clustering algorithms for mixed features of symbolic and fuzzy data, by modifying dissimilarity measure for symbolic data (Gowda & Diday, 1991, 1992) and also changing the parametric approach for fuzzy data suggested by Hathaway et al. (1996). de Souza & de Carvalho (2004) introduced adaptive and non-adaptive clustering methods for interval valued data based on city-block distances. They suggested two dynamic clustering methods for partitioning a set of symbolic objects where each object is represented by a vector of intervals. The first method utilizes a suitable extension of the city-block distance which compares a pair of vector of intervals. The latter method uses two adaptive versions of this extended city-block distance for interval valued data. In first version, the adaptive distance has only a single component, whereas it has two components in the second version. In both methods, the prototype of each cluster is also represented by a vector of intervals whose bounds, for each interval, are the median of the set of lower bounds and the median of the set of upper bounds of the intervals of the objects belonging to the cluster.
5.3 Proposed Model

In the proposed model, arecanut is segmented from a given image using Otsu method. Texture features and a simple kNN classifier have been used for classification. The texture features used in this work are, LBP, Haar Wavelets, GLCM and Gabor features. Results obtained from each texture feature set and best discriminative subsets of features from the components of HSV and YCbCr color models are selected empirically based on combination of features.

5.3.1 Histogram based Approach

In the histogram based approach, an RGB arecanut image has been transformed to HSV image. The saturation component of HSV color space is used for segmentation. Segmentation of arecanut has been done using threshold based Otsu segmentation method. LBP has been applied on the value component of HSV color space. Basic LBP operator used in the section 3.3.3 has been modified in this method. Each neighbor pixel is compared with the average of the pixels encompass by a mask and this average value is considered as a threshold. The obtained threshold is insensitive to noise. The working principle of LBP used in this method is shown in Figure 5.1.

The ones whose intensities exceed the threshold are marked as 1, otherwise 0. In this way we get a simple circular point features consisting of only binary bits. Typically the feature ring is considered as a row vector, and then with a binomial weight assigned to each bit, the row vector is transformed into decimal code for further use. LBP using circular neighborhoods and linearly interpolating the pixel values allows the choice of any radius $R$ and number of pixel in the neighborhood $P$ to form an operator, which can model large scale structure.

Threshold value 143 is considered for this instance. The equation for LBP for this method is shown in equation (5.1).

$$LBP_{p,R}(x,y) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p$$

where $g_c$ is the average value of pixels encompass by a mask, $g_p$ is the value of its neighbors.
A descriptor for texture analysis is a histogram, \( h(i) \), of the local binary pattern shown in equation (5.2) and its advantage is that it is invariant to image translation.

\[
h(i) = \sum_{x,y} B(LBP, R(x, y) = i) \mid i \in [0, 2^{p-1}], B(v) = \begin{cases} 
1 & v > T \\
0 & otherwise 
\end{cases} 
\]  \( \ldots (5.2) \)

In order to perform classification of arecanut, each arecanut image in the training and test sets is converted to a spatially enhanced histogram via the process described above. Then ordinary nearest neighbor classification is performed with a histogram distance measure. The statistical measure called correlation is used to find a distance between histograms and is given in equation (5.3).

\[
Corr(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}} 
\]  \( \ldots (5.3) \)

where \( \bar{x} \) and \( \bar{y} \) are the average intensity values of the histogram \( h \). The histogram vectors \( x_i \) and \( y_i \) (\( 0 \leq i \leq n \)) are the histogram values of training set and test set respectively.

The exhaustive experiment is conducted on dataset with 250 images, among these 60% of images has been used for training and 40% of images has been used for testing. Images were resized to 300 X 300 pixel resolution to improve the
computation speed. In this method, LBP has been applied on value component of HSV color space and then histogram is generated. Further, distance between the training set and query sample is measured using correlation of histogram and obtained a success rate of 92.00%.

5.3.2 Features Selection

Diverse feature selection techniques have been proposed in the machine learning literature. The purpose of this techniques is to discard irrelevant or redundant features from a given feature vector. For the purpose of this experiment, selection method is used with the help of empirical results.

In the proposed method, arecanut is segmented from a given image using threshold based Otsu segmentation method. Texture features and kNN classifier have been used for classification of diseased and undiseased arecanut. The textures used in this work are Haar Wavelets, GLCM and Gabor features. The best discriminative subsets of features from HSV and YCbCr color models are selected empirically based on combination of features. An algorithm based on proposed methodology is as given below.

Algorithm (Input: RGB image of arecanut, Output: Classification Label)

1. Convert RGB image of arecanut to YCbCr and HSV color spaces.
2. Extract Saturation component from HSV color space for segmentation of arecanut.
3. Global Threshold based segmentation using Otsu method is performed.
4. Hue, Saturation, Intensity components are separated from HSV similarly, Yellow, Chromatic Blue and Chromatic red are seperated from YCbCr.
5. From each color component Haar Wavelets, GLCM, Gabor texture features are extracted.
6. Diseased and undiseased arecanut classification has been done using kNN classifier using texture feature set mentioned in step 5.
7. Experimentally, subsets of texture features with high degree of discrimination power based on different combinations of features are determined.
8. Classification has been done using kNN classifier with subsets of texture features.

![Figure 5.2. Sample Experimental Results of HSV Color Model](image)

The segmented saturation component of HSV color space is converted to binary image and this is called the mask and shown in Figures 5.2(e), 5.2(k), 5.2(q) and Figures 5.3(e), 5.3(k), 5.3(q). The mask is multiplied with color components of YCbCr and HSV color models as discussed in the algorithm. The equation for multiplying mask with color components hue, saturation and value of HSV color space is given in equations (5.4) - (5.6) respectively and equations for multiplying with color components yellow, chromatic blue and chromatic red is given in equations (5.7) - (5.9). The segmented results of hue, saturation and value of HSV color space
are shown as in Figures 5.2(f), 5.2(l), 5.2(r), and the segmented results of yellow, blue chromatic blue and chromatic red are shown in Figure 5.3(f), 5.3(l), 5.3(r).

\[ f_{sH}(x, y) = \sum_{x=1}^{M} \sum_{y=1}^{N} m(x, y) H(x, y) \] \hspace{1cm} \ldots (5.4)

where \( H(x, y) \) and \( f_{sH}(x, y) \) are the hue and segmented hue components of the HSV color space, \( m(x, y) \) is the mask. Variables \( M \) and \( N \) are spatial sizes of images.

\[ f_{sS}(x, y) = \sum_{x=1}^{M} \sum_{y=1}^{N} m(x, y) S(x, y) \] \hspace{1cm} \ldots (5.5)

where \( S(x, y) \) and \( f_{sS}(x, y) \) are the saturation and segmented saturation components of the HSV color space, \( m(x, y) \) is the mask. Variables \( M \) and \( N \) are spatial sizes of images.

\[ f_{sV}(x, y) = \sum_{x=1}^{M} \sum_{y=1}^{N} m(x, y) V(x, y) \] \hspace{1cm} \ldots (5.6)

where \( V(x, y) \) and \( f_{sV}(x, y) \) are the value and segmented value components of the HSV color space, \( m(x, y) \) is the mask. Variables \( M \) and \( N \) are spatial sizes of images.

\[ f_{sY}(x, y) = \sum_{x=1}^{M} \sum_{y=1}^{N} m(x, y) Y(x, y) \] \hspace{1cm} \ldots (5.7)

where \( Y(x, y) \) and \( f_{sY}(x, y) \) are the yellow and segmented yellow components of the YCbCr color space, \( m(x, y) \) is the mask. Variables \( M \) and \( N \) are spatial sizes of images.

\[ f_{sCb}(x, y) = \sum_{x=1}^{M} \sum_{y=1}^{N} m(x, y) Cb(x, y) \] \hspace{1cm} \ldots (5.8)

where \( Cb(x, y) \) and \( f_{sCb}(x, y) \) are the chromatic blue and segmented chromatic blue components of the YCbCr color space, \( m(x, y) \) is the mask. Variables \( M \) and \( N \) are spatial sizes of images.

\[ f_{sCr}(x, y) = \sum_{x=1}^{M} \sum_{y=1}^{N} m(x, y) Cr(x, y) \] \hspace{1cm} \ldots (5.9)
-where $Cr(x, y)$ and $f_{Cr}(x, y)$ are the chromatic red and segmented chromatic red components of the YCbCr color space, $m(x, y)$ is the mask. Variables $M$ and $N$ are spatial sizes of images.

The GLCM features used in this model is given in Table 5.1. The Haar wavelets features used are given in Table 5.2. A filter bank of Gabor filters with three scales and four orientations are created. Totally twelve feature vectors were extracted for the above mentioned three scales and four orientations. The corresponding mean ($\mu$) value is determined for each feature vector using equation (5.10).

-where $x$ is the feature vector and $n$ is the size of feature vector.
\[ \mu = \frac{\sum_{i=1}^{n} x_i}{n} \]

\[
\text{Contrast} \quad \sum_{i,j} |i-j|^2 p(i,j) \\
\text{Correlation} \quad \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j} \\
\text{Energy} \quad \sum_{i,j} p(i,j)^2 \\
\text{Homogeneity} \quad \sum_{i,j} \frac{p(i,j)}{|i-j|} 
\]

Table 5.1. Different GLCM Features used in this work
Table 5.2. Haar Wavelets Features.

| Mean Value of approximation, horizontal, vertical and diagonal energy | $\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}$ |
| Mean of Decomposition Vector | $\frac{1}{n} \sum_{i=1}^{n} x_{i}$ |
| Variance of Decomposition Vector | $\frac{1}{n} \sum_{i=1}^{n} (x_{i} - \mu)^2$ |
| Entropy of Decomposition Vector | $-\sum_{i=1}^{n} p(x_{i}) \log_2 p(x_{i})$ |

Table 5.3. Results obtained for Different Texture Features

<table>
<thead>
<tr>
<th>Features</th>
<th>Success Rate in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haar Wavelets</td>
<td>92.30</td>
</tr>
<tr>
<td>GLCM</td>
<td>84.61</td>
</tr>
<tr>
<td>Gabor</td>
<td>98.00</td>
</tr>
<tr>
<td>Combinations of Best Discriminative Subsets of Features</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Table 5.4. Texture features with high degree of discrimination power

<table>
<thead>
<tr>
<th>Image Component</th>
<th>Texture Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity component from HSI color model</td>
<td>Contrast, Correlation, Energy and Entropy of GLCM</td>
</tr>
<tr>
<td>Yellow Component from YCbCr color model</td>
<td>Vertical and diagonal approximation features of Haar Wavelets</td>
</tr>
</tbody>
</table>
The experiment is conducted on dataset of 250 images in which 60% of images have been used for training and 40% of images have been used for testing. Texture features of Haar Wavelets, GLCM and Gabor Wavelet have been obtained from each color component of HSV and YCbCr color models and the results are shown in Table 5.3 and the same are plotted in Figure 5.4. Best discriminative subsets of features have been determined from Haar Wavelets, GLCM and Gabor features empirically based on combination of features using kNN classifier. Texture features with high degree of discrimination power are mentioned in Table 5.4.

In this work, disease of arecanut that occurs only after processing is considered. Diseases that occur before processing can be considered in future work. This can be a base paper to determine some other diseases of arecanut and early detection of disease that comes when nut is growing in areca tree.

5.3.3 Symbolic Data

Conventional data analysis can handle scalars, vectors and matrices. However, lately, some datasets have grown beyond the framework of conventional data analysis. Most statistical methods do not have sufficient power to analyze these datasets. Symbolic
data analysis proposed by Diday (2006, 1988) is an approach for analyzing new types of datasets. Symbolic data consist of a concept that is described by intervals, distributions, etc. as well as by numerical values. The use of symbolic data analysis enriches data description, and it can handle highly complex datasets. This implies that complex data can be formally handled in the framework of symbolic data analysis. However, most symbolic data analysis works have dealt with only intervals as the descriptions and are very few studies based on this simple idea.

In this model, symbolic data analysis has been done using the texture features of GLCM, Gabor and Haar wavelets.

### 5.3.3.1 Range Method

Let us consider a feature set \( F = \{f_1, f_2, f_3, \ldots, f_n\} \) of \( N \) feature vectors. The interval value is defined by the set \( I \) using equation (5.11).

\[
I = \{ (f_1^-, f_1^+), (f_2^-, f_2^+), (f_3^-, f_3^+), \ldots, (f_N^-, f_N^+) \} \quad \ldots (5.11)
\]

-where \( f_i^- \) and \( f_i^+ \) (\( i = 1 \) to \( n \)) are the lower limit and upper limit of feature vectors respectively which gives interval data. The value of \( f^- \) and \( f^+ \) are defined using equations (5.12) and (5.13).

\[
f^- = \mu - \sigma \quad \ldots (5.12)
\]

\[
f^+ = \mu + \sigma \quad \ldots (5.13)
\]

-where \( \mu \) and \( \sigma \) mean are standard deviation.

The mean and standard deviation of a feature vector \( F \) of size \( n \) is defined using equations (5.14) and (5.15) respectively.
5.3.3.2 Sigma Controls

In this model, three sigma control limits are used to determine symbolic data. Symbolic data for the set mentioned in equation (5.11) is obtained using equations (5.16) and (5.17).

\[
f^− = \mu - 3\sigma \quad \ldots (5.16)
\]

\[
f^+ = \mu + 3\sigma \quad \ldots (5.17)
\]

-where \(\mu\) and \(\sigma\) are the mean and standard deviation of a feature vector. The mean and standard deviation of a feature vector \(F\) of size \(n\) is determined using equations (5.14) and (5.15) respectively.

Proposed model considered features of GLCM (Table 5.1), Haar wavelet features (Table 5.2) and Gabor features. Experimental results have given success rate of 92.00%.
5.4 Conclusion

Three models are proposed in this chapter. In first model, LBP has been applied on arecanut image. In the basic version of LBP, a center pixel is the threshold value to compare neighbors of pixels encompassed by a mask. In the proposed method, threshold value is the average of pixels encompassed by a mask and it is insensitive to noise. Histogram has been generated from LBP image and correlation is used to measure the distance between training samples and testing query. For the same features, kNN classifier has given a success rate of 92%.

In second model texture features of GLCM, Gabor and Haar wavelets have been extracted from components of HSV and YCbCr color spaces. Experiments reveals that a least success rate of 84.61% for GLCM features, highest success rate of 100% for discriminative subset of texture features are obtained.

In third model, Symbolic data has been employed for classification of diseased and undiseased arecanut. The texture features of GLCM, Gabor and Haar wavelets have been extracted from the components of HSV and YCbCr color spaces. A symbolic data is determined for the texture features of GLCM, Gabor and Haar wavelets. A range method is used for determining symbolic data for the GLCM, Gabor and Haar wavelets features and obtained the success rate of 84%. The three sigma control limits are also used to define a symbolic data for GLCM, Gabor and Haar wavelets and obtained success rate of 92%.