Improved Tropical Wood Species Recognition System based on Multi-feature Extractor and Classifier

Marzuki Khalid, RubiyahYusof and AnisSalwaMohdKhairuddin

Abstract—An automated wood recognition system is designed to classify tropical wood species.The wood features are extracted based on two feature extractors: Basic Grey Level Aura Matrix (BGLAM) technique and statistical properties of pores distribution (SPPD) technique. Due to the nonlinearity of the tropical wood species separation boundaries, a pre classification stage is proposed which consists ofKmeans clusteringand kernel discriminant analysis (KDA). Finally, Linear Discriminant Analysis (LDA) classifier and K-Nearest Neighbour (KNN) are implemented for comparison purposes. The study involves comparison of the system with and without pre classification using KNN classifier and LDA classifier.The results show that the inclusion of the pre classification stage has improved the accuracy of both the LDA and KNN classifiers by more than 12%.

Keywords—Tropical wood species, nonlinear data, feature extractors, classification

I. INTRODUCTION

TROPICAL countries are blessed with abundance of wood supply. Together they made up approximately 33% of the world's timber industry in year 2010 [1]. With more demands and more tightly controlled international requirements, many of these countries are required to meet tighter security requirements as well as higher technical demands such as more accurate recognition of the correct species, prevention of fraud, and Environmental Investigation Agency (EIA) requirements, to name a few. In many timber industries one of the major problems is to find good wood graders. Currently, very few certified officers are involved in the traditional wood identification process. The process of training up experienced officers in performing the job is difficult since the job is no longer considered lucrative and rather laborious. Moreover, the possibility of biasness and mistakes in human recognition system has to be considered. Besides that, it is impractical and cost effective for a human to analyze and identify large amount of timber species. Hence automatic wood species recognition system is needed to overcome the errors caused by

traditional wood identification system which based solely on human expertise. Khalid et. al [2] has developed an automatic Tropical Wood Species Recognition System using macroscopic features of timber based on image processing. In the work, the surface texture of timber is captured, and the features are extracted from the images using Grey Level Co-occurrence Matrix (GLCM) and back propagation Neural Network (BPNN) is used to train and classify the timber species.Yusofet. al [3] adopted fusion of two feature sets using multi feature extractor technique. Two feature extraction methods were used in this system: co-occurrence matrix approach (GLCM) and Gabor filters. These two extraction methods produced more variation of features and also improved the accuracy rate of the system. K. Wang et.al [4] proposed a wood surface texture recognition method based on feature level data fusion which uses three types of texture analysis methods. The three texture analysis methods used are gray co-occurrence matrix (GLCM), Gauss-Markov random field (GMRF) and wavelet multi-resolution fractal dimension. The fusion method is based on Simulated Annealing Algorithm with memory. In a later development, Khairuddin et.al [5] used feature selection based on genetic algorithm to improve the accuracy of wood species recognition, in particular to reduce the redundant features which are not considered as discriminatory enough for accurate classification.

Wood samples of the same species might have different colour, tones, density and size of pores. Some of the wood samples also filled with dust, dammar or deposit. This is because some of the wood samples were taken from different part of the same tree or from different trees of the same species. Size of pores of certain trees changes through years. Sometimes the growth location of the tree affects the size of pores. Due to the large variations of features within inter and intra among the species of the tropical wood, and the problems relating to variations in the wood samples, there is a pressing need to improve the methodology of the wood recognition system. The conventional method of using the feature extraction followed by classification in the recognition of wood species is no longer adequate for a larger sample of wood species. Therefore, in this paper, we proposed a two stage classification for the wood recognition system. The first stage is called the pre classification stage that utilizes K-mean clustering as database management and KDA/GSVD as nonlinear dimension reduction technique in order to enhance

Marzuki Khalid is with Centre for Artificial Intelligence and Robotics (CAIRO), UniversitiTeknologi Malaysia, Kuala Lumpur, Malaysia (phone: +607 5530 360; fax:+607 557 1700; e-mail: marzuki@utm.my).

RubiyahYusof is with Centre for Artificial Intelligence and Robotics (CAIRO), UniversitiTeknologi Malaysia, Kuala Lumpur, Malaysia (rubiyah@ic.utm.my).

AnisSalwaMohdKhairuddin is with Department of Electrical Engineering, University of Malaya, Kuala Lumpur, Malaysia (e-mail: anissalwa@um.edu.my).

World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:5, No:11, 2011

the class-discriminatory information. Finally, at the final classification stage, we applied k-NN and LDA classifier to classify the wood species. The experimental results presented in section 3 shows that the implementation of clustering and dimension reduction improves the classification accuracy of the system.

This paper is organized as follows: in Section II, we presented the methodology of the proposed system. In Section III, experimental results are discussed followed by brief conclusion in Section IV.

II. PROPOSED METHODOLOGIES

The methodology of the proposed system consists of image acquisition, image pre-processing, feature extraction, clustering and dimension reduction in the preclassification stage and the final classification. Fig. 1 shows the overview of the proposed system.
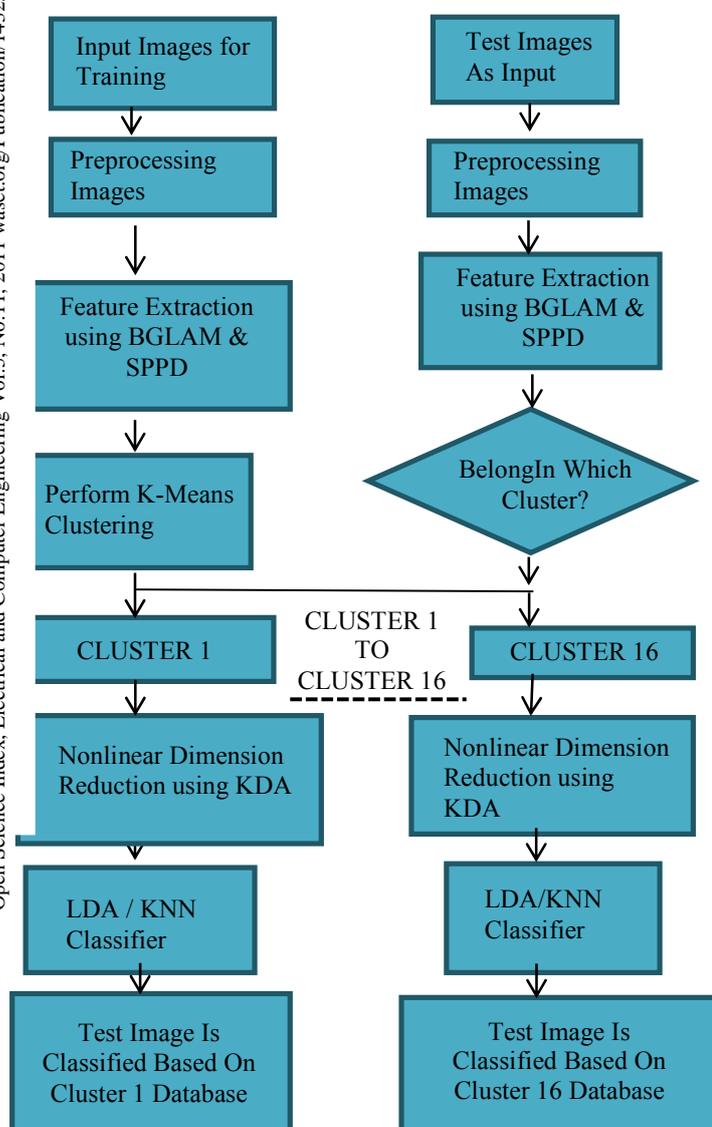
Fig. 1 Overview of the proposed system

A. Image Acquisition

The wood samples for this research are obtained from the Forest Research Institute of Malaysia (FRIM). There are 52 wood species in cubic form (approximately 1 inch by 1 inch in size) where 5 cubes are provided for each species. The images of the wood surface are captured by using a specially designed portable camera with 10 times magnification. The size of the each image is 768x576 pixels. In total, there are 100 images taken from each wood species where 70 images are for training and the other 30 images are for testing.Table I as shown below lists out the wood species included in the system's database.

TABLE I
LIST OF 52 WOOD SPECIES (SCIENTIFIC NAME) IN THE DATABASE

1. Shorealaevis	27. Dialumplatysepalum
2. Shoreamaxwelliana	28. Dipterocarpussp
3. Shoreaexelliptica	29. Dipterocarpuscrinitus
4. Calophyllumkunstleri	30. Dipterocarpusbaudii
5. Calophyllum Flavo-ramulum	31. Dipterocarpus hasseltii
6. Calophyllum bornecusis	32. Scorodocarpus borneensis
7. Palaquiumstellatum	33. M quadrifida
8. Palaquiumridleyi	34. M microphylla
9. Neobalanocarpasheimii	35. M caesia
10. Duriospp	36. M foetida
11. Parashoreaglobosa	37. Kokoonasessilis
12. Parashoreastellata	38. Kokoonareflexa
13. Hopeaspp	39. Kokoonalittoralis
14. Hopeapachycarpa	40. Dehaasiacurtisii
15. Hopeaferrea	41. Sp of lauraceacae
16. Hopeaapiculata	42. Pentaceadenophora
17. Dyeracostulata	43. Pentacecurtisii
18. Dryobalanopsaromatica	44. Pentace floribunda
19. Dryobalanopsoblongifolia	45. Shoreauliginosa
20. Pometiaridleyi	46. Lophopetalumsp
21. Cynometramalaccensis	47. Lophopetalum beccarianum
22. Artocarpusdadah	48. Lophopetalum floribundum
23. Artocarpuslanceifolia	49. Shoreaguiso
24. Artocarpuskemando	50. Shoreacollina
25. Dialumkunstleri	51. Heveabrasiliensis
26. Dialumkingii	52. Endospermumdiadenum

B. Image Pre-Processing

The wood images are pre-processed using homomorphic filters in order to enhance the image presentation. Homomorphic filtering sharpens features and flattens lighting variations in an image. Hence, illumination and reflectance on the image can be removed. Fig. 2 as shown below are the images of wood species 'Shorealaevis'. The homomorphic image is obtained by performing homomorphic filtering on the original image.
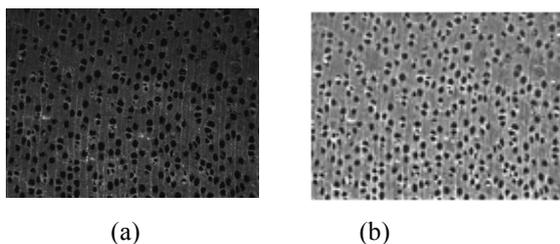
World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:5, No:11, 2011

(a)   (b)

Fig. 2 (a) Original image of wood, (b) Homomorphic image of wood

C. Feature Extraction

In this paper, we use the same features that have been used by Khairuddin et.al [5]. In their work, the wood features were extracted based on the fusion of Basic Grey Level Aura Matrix (BGLAM) & Statistical Properties of Pores Distribution (SPPD). The total number of features extracted from each wood sample is 157 features (136 features from BGLAM and 21 features from SPPD).

An image can be uniquely represented by its BGLAMs. BGLAM technique is applied on the homomorphic image of wood since the BGLAMs of an image characterize the co-occurrence probability distributions of gray levels at all possible displacements configurations. It is proven that BGLAMs can give the necessary and sufficient information to differentiate between images. In addition, BGLAM features are directly calculated from the images, thus, providing a more accurate representation of an image. The framework of BGLAM feature extractor is explained in detail in [6].
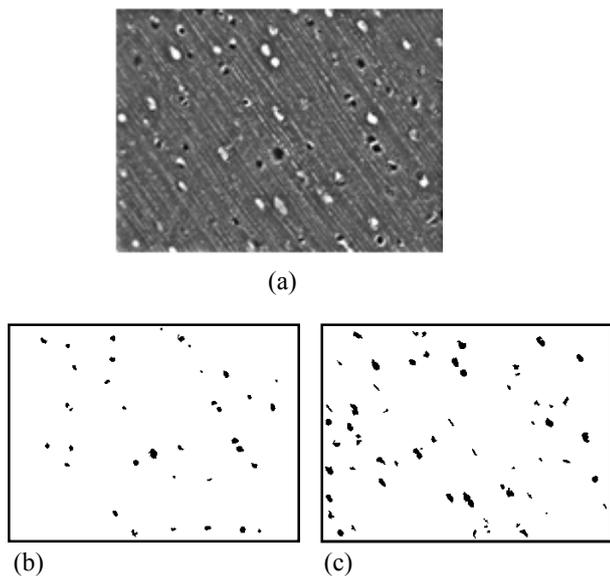

(a)

(b)   (c)

Fig. 3 (a)Homomorphic image of wood, (b) Black pores image of wood, (b) White pores image of wood

Fig. 3 shows the images of wood species 'Palaquiumridleyi'.SPPD extracts statistical properties of pores of distribution on the image sample. Firstly, the black pores and white pores of the wood image were extracted from the homomorphic image. Then, SPPD technique is performed on both black pores and white pores image. The SPPD features will allow only distinct pores to be acknowledged as characteristics of a species. The SPPD features extracted from each wood sample are:

1. Mean size of the pores and the corresponding standard deviation (2 features)
2. Mean distance between pores and the corresponding standard deviation (2 features)
3. Number of small, medium and large pores (3 features)
4. Number of pairs and solitary pores (2 features)
5. Number of pores per square millimeters (1 feature)

By using the SPPD feature extractor, 10 features are obtained from black pores image and 10 features are obtained from white pores image. Another 1 feature is the mean grey level feature which is obtained from the original image. Hence, total features extracted from each image by using the SPPD technique is 21 features.

D. Pre-Classification Stage

Pre-classification stage includes clustering using k-means clustering technique and nonlinear dimension reduction using the KDA technique. The goal of pre-classification stage is to enhance the class-discriminatory information in a lower-dimensional space.

Clustering

Based on the features extracted from each wood sample, the wood database is clustered into 16 clusters using k-means clustering. The number of cluster is chosen based on results presented in Fig. 5 and Fig. 6 in Section III. Each cluster will have its own wood database and classification stage.

Given a set of training wood samples n (x_1, x_2, ..., x_n), where n represents the number of training wood samples in the database. Each wood sample, x is a real vector of wood features where $x = (x_1,...,x_{157})$ since there are 157 features extracted from each wood sample. K-means clustering aims to partition n wood samples into k clusters ($k \leq n$).

$$S = \{S_1, S_2, ..., S_k\} \quad (1)$$

so as to minimize the within-cluster sum of squares as in (1):

$$\underset{S}{\arg\min} \sum_{i=1}^{k} \sum_{x_n \in S_i} \|x_n - \mu_i\|^2 \quad (2)$$

μ_i is the mean of points in S_i.

World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:5, No:11, 2011

The algorithm of the k-means clustering is shown below [15]:

1) Determine number of cluster, k.
2) Determine initial set of k means $\boldsymbol{\mu}_1,...,\boldsymbol{\mu}_k$ as cluster centers using random permutation.
3) Assign each wood sample to the nearest cluster center based on minimum-distance classifier. That is, we can say that $\boldsymbol{x}_n$ is in cluster k if $\|\boldsymbol{x}_n - \boldsymbol{\mu}_i\|$ is the minimum of all the i distances based on (2).

$$S_i = \{\boldsymbol{x}_n : \|\boldsymbol{x}_n - \boldsymbol{\mu}_i\| \leq \|\boldsymbol{x}_n - \boldsymbol{\mu}_{i^*}\| \qquad (3)$$
for all $i^* = 1,...,k\}$

4) Based on the clustering results from Step 3, we will update the cluster centers, $\boldsymbol{\mu}_i$ by calculating new means to be the centroid of the wood samples in the cluster i. Where $\boldsymbol{x}_i$ represents the wood samples belong to cluster i. The centroid is updated based on (4).

$$\boldsymbol{\mu}_i = \frac{1}{|S_i|}\sum_{x_i \in S_i} \boldsymbol{x}_i \qquad (4)$$

5) Repeat Steps 3 and 4 until the centroid for every clusters no longer move which means convergence has been reached.

There are various techniques that can beimplementedto determine cluster validity where the main goal is to search for clustering that is well separated [10]. Since the k-means clustering aims to minimize the sum of squared distances from all points to their cluster centers, this should result in compact clusters.The sum of the squared error (SSE) is used to measure the quality of clustering. We prefer the one with the smallest SSE since this means that the centers of this clustering are a better representation of the points in their cluster. The values of SSE for different number of clusters are shown in Fig.5. The SSE is formally defined as:

$$\text{SSE} = \sum_{i=1}^{k}\sum_{x \in C_i} dist(c_i, x)^2 \qquad (5)$$

whereC_i is the ith cluster, c_i is the center of cluster C_i, *dist* is the standard Euclidean distance between two objects in Euclidean space, and x is a point in C_i.

Nonlinear dimension reduction using KDA

Linear feature extraction methods have been widely used to remove redundant features and to speed up data processing. A linear feature extraction method cannot prevent a loss of class separability during dimension reduction. This limitation can be observed in classes that have nonlinear decision boundaries. Various nonlinear feature extraction methods based on kernel tricks have been implemented to overcome the

limitations associated with linear feature extraction [11,12,13,14].

Fig.4 shows the nonlinear distribution of wood species where each symbol represents different type of wood species.The scatter plot of wood training database is highly nonlinear and complex due to variation in viewpoint, illumination, and texture features within inter and intra among the species as mentioned in the introduction part. Hence kernel-based method which is known as kernel discriminant analysis (KDA) has been applied to deal with nonlinearity of the wood patterns' distribution. In KDA, Radial basis function (RBF) kernel function was employed into LDA to handle nonlinearity in a computationally attractive manner. However, due to the nonlinear map by the kernel function, the dimension of the feature space often becomes much larger than that of the original data space, and as a result, the scatter matrices become singular, which is referred to as "small sample size" (SSS) problem. Hence, KDA has been proposed, which is a generalization of LDA based on the generalized singular value decomposition (GSVD). It overcomes the SSS problem.
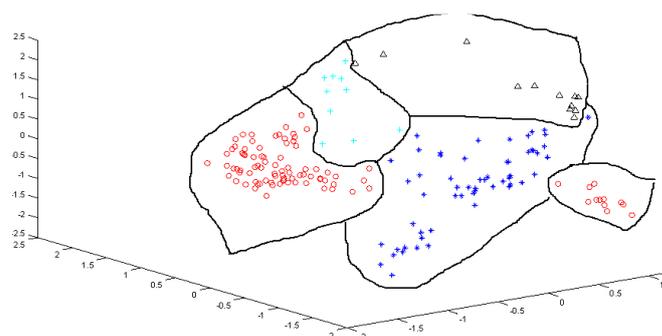

Fig. 4 Scatter plot of 5 wood species intraining databasefor Cluster 1 with nonlinear boundaries

The main idea of the kernel method is that without knowing the nonlinear feature mapping or the mapped feature space explicitly, we can work on the feature space through kernel functions, as long as the problem formulation depends only on the inner products between data points. This is based on the fact that for any kernel function k satisfying Mercer's condition, there exists a mapping φ such that,

$$< \varphi(a), \varphi(b) > = k(a,b) \qquad (6)$$

where$<,>$ is an inner product in the feature space transformed by φ.

We apply the kernel method to perform LDA in the feature space instead of the original input space. Given a kernel function k, let φ be a mapping satisfying (6) and define $F \subset \mathbb{R}^N$ to be the feature space from the mapping φ. A Gaussian is used as the RBF kernel based on (7).

$$k(A,B) = K_{rbf2D}(A,B) = exp(-\frac{d^2(A,B)}{\sigma^2}) \qquad (7)$$
where $\sigma \in R$

World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:5, No:11, 2011

The RBF kernel returns an RBF kernel matrix from the input coordinates. The inputs of the RBF kernel are a matrix containing all wood features and the kernel width. The output of the kernel is dependent on the Euclidean distance of B from A (one of these will be the support vector and the other will be the testing data point. The support vector will be the center of the RBF and σ will determine the area of influence this support vector has over the data space.

The algorithm of KDA in [7,8] is summarized as shown below

Given a data matrix $A = [a_1, ..., a_n] \in R^{mxn}$ with r classes and a kernel function K, it computes the $r-1$ dimensional representation of any input vector $z \in R^m$ by applying GSVD in the feature space defined by the feature mapping φ such that $K(a_i, a_j) = <\varphi(a_i), \varphi(a_j)>$

Where :
A = matrix of m x n training data samples
m = number of features
n = number of training samples
r = number of wood species

1) Compute between-class scatter matrix, K_b and within-class scatter matrix, K_w.

   c = number of wood species (class) in the training samples,
   M_i = the number of training samples in class i,
   x_i = set of wood images belonging to class i
   μ_i is the mean vector of samples in class i,
   x_k is the kth image of that class

$$K_b = \sum_{i=1}^{c} M_i (x_i - \mu)(x_i - \mu)^T \quad (8)$$
$$K_w = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T c \quad (9)$$

2) Apply GSVD to the pair K_b and K_w. And we have vector of singular matrix,

$$X^T K_b K_b{}^T X = \begin{bmatrix} \Gamma_b^T \Gamma_b & 0 \\ 0 & 0 \end{bmatrix} (10)$$

$$X^T K_w K_w{}^T X = \begin{bmatrix} \Gamma_w^T \Gamma_w & 0 \\ 0 & 0 \end{bmatrix} \quad (11)$$

3) The LDA in feature space finds transformation matrix

$$G = [\varphi^{(1)}, ..., \varphi^{(r-1)}] = \begin{bmatrix} \varphi_{1,1} & \cdots & \varphi_{1,r-1} \\ \vdots & \vdots & \vdots \\ \varphi_{n,1} & \cdots & \varphi_{n,r-1} \end{bmatrix} \quad (12)$$

Where the columns of G are the eigenvectors corresponding to $r-1$ largest eigenvalues of

$$K_b \varphi = \lambda K_w \varphi \quad (13)$$

4) For any input vector, $z \in R^{mxl}$, a dimension reduced representation is computed as

$$G^T \begin{bmatrix} K(a_{1,z}) \\ \vdots \\ K(a_{n,z}) \end{bmatrix} \in R^{(r-1)x1} \quad (14)$$

E. Final Classification Stage

The two classifiers used in this system are LDA and k-NN. Both classifiers are used to compare the classification accuracy based on (15).

$$\text{ClassificationAccuracy} = \frac{\text{number of correctly classified test images}}{\text{number of test images}} \times 100 \quad (15)$$

(i)   KNN classifier

KNN classifier is used to classify the wood species. The test data is classified by calculating the Euclidean distance, $d(x,y)$ to the nearest training data as in (16). Then, use simple majority of the category of nearest neighbors to determine the wood species of the test data.

$$d(x,y) = \sqrt{\sum_{i=1}^{157}(x_i - y_i)^2} \quad (16)$$

where x = training data and y = test data.

(ii)   LDA classifier

The algorithm of LDA is as follows [14]:

(a) Find mean μ_i for each features of a species i (μ_i is a matrix with number of column equal to number of features)

(b) Construct covariance matrix c_i for each species

(c) Construct pooled within group covariance matrix,

$$C(r,s) = \frac{1}{n} \sum_{i=1}^{g} n_i \cdot c_i(r,s) \quad (17)$$

Where   which is calculated for each (r, s) entry in the matrix.

(d) Construct prior probabilities vector, p. Prior probabilities for species i is equal number of samples in species i divided by number of total samples in database.

$$p_i = \frac{n_i}{N} \quad (18)$$

(e) After discriminant analysis is completed for the training set, a linear coefficient for each class is obtained. To make classification for testing set, the following discrimination function is used:

$$f_i = \mu_i C^{-1} x_k^T - \frac{1}{2} \mu_i C^{-1} \mu_i^T + \ln(p_i)$$

World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:5, No:11, 2011

Where x_k is the feature array of test sample k and sample k will be assigned to Species i that has maximum f_i.

III. EXPERIMENTAL RESULTS

For clustering method, it is important to select the correct number of clusters for good recognition rate. In order to determine the right number of clusters for our system, we calculated the sum of squared error (SSE) for different number of cluster (from 1 cluster to 52 clusters) as shown in Fig. 5 below based on (5). The wood species database is clustered based on 157 features obtained from fused BGLAM and SPPD feature extractors. Based on Fig. 5, it can be seen that SSE is low after k=16 onwards. Hence, the right number of cluster should be between k=16 until k=52.
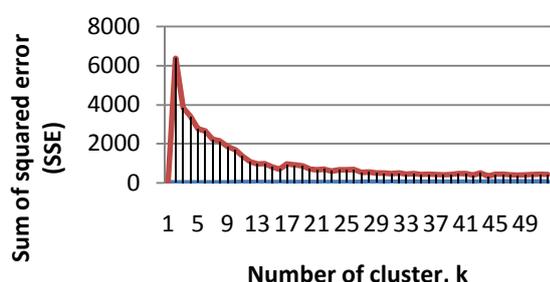

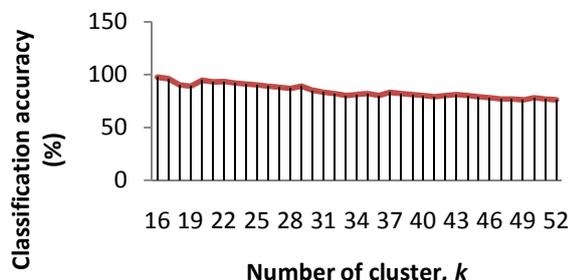Fig. 5 Sum of Squared Error for different number of clusters

Fig. 6 Classification accuracy based on different number of clusters

We performed several experiments to investigate the effectiveness of the proposed method. Based on results shown in Fig. 6, the classification accuracy rate is highest at number of cluster, k= 16. Hence, the right number of cluster that we chose to cluster the wood database is 16 clusters. As shown in Table II and Table III, the implementation of clustering and nonlinear dimension reduction has improved the classification accuracy from 84% to 96.92% for LDA classifier. In this paper, LDA classifier gives higher classification accuracy compared to KNN classifier.

TABLE II
CLASSIFICATION ACCURACY OF WOOD RECOGNITION SYSTEM WITHOUT INCLUDING THE PRE CLASSIFICATION STAGE

Recognition system condition	Classification accuracy
Complete proposed system using KNN classifier	84.0%
Complete proposed system using LDA classifier	84.6%

TABLE III
CLASSIFICATION ACCURACY OF THE PROPOSED WOOD RECOGNITION SYSTEM INCLUDING THE PRE CLASSIFICATION STAGE

Recognition system condition	Classification accuracy
Complete proposed system using KNN classifier	96.35%
Complete proposed system using LDA classifier	96.92%

IV. CONCLUSION

This paper concludes that the implementation of k-means clustering and nonlinear dimension reduction as the pre-classification stage have reduced the computing time and improved the performance of the tropical wood recognition system. The k-means clustering enables the system to only compute the wood database in a respective cluster instead of computing the entire database to classify a wood species while nonlinear dimension reduction enables the wood samples to be represented more accurately in a lower-dimensional space. BGLAM and SPPD have been implemented as the feature extractors while KNN and LDA classifiers were used to measure the classification accuracy of the proposed system.

REFERENCES

[1] *State of the World's Forests,* FAO, 1995-2010
[2] M. Khalid, Lew Y. L., R. Yusof and M. Nadaraj, "Design of An Intelligent Wood Species Recognition System,"*IJSSST*, Vol. 9, No. 3, September 2008
[3] R. Yusof, Nenny R. Rosli, M. Khalid,"Using Gabor Filters as Image Multiplier for Tropical Wood Species Recognition System,*"12th International Conference on Computer Modelling and Simulation*, 2010, pp.289-294.
[4] Keqi Wang and XuebingBai, "Research on classification of wood surface texture based on feature level data fusion,"*2nd International Conference on Industrial Electronics and Applications*, 2007
[5] UswahKhairuddin, RubiyahYusof, Marzuki Khalid and Florian Cordova, "Optimized Feature Selection for Improved Tropical Wood Species Recognition System,"*ICIC Express letters, Part B: Applications, An International Journal of Research and Surveys,* Volume 2, Number 2, April 2011, pp 441-446.

World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:5, No:11, 2011

[6] Xuejie Qin; Yee-Hong Yang, *"Similarity measure and learning with gray level aura matrices (GLAM)for texture image retrieval"Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR,* 2004. Vol. 1, pp. 326 - I-333

[7] Park, C.H. and Park, H,"Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition,"*SIAM Journal on Matrix Analysis and Applications*, 2005

[8] Howland,P and Park, H. "Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition,"*IEEE Transaction, PAM,* 2004

[9] Shi Na; Liu Xumin; Guan Yong,"Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorith*,"3rd International Conference on Intelligent Information Technology and Security Informatics (IITSI),* 2010, pp.63-67,

[10] Ming-Hseng Tseng, Chang-Yun Chiang, Ping-Hung Tang, Hui-Ching Wu, "A study on cluster validity using intelligent evolutionary k-means approach", *Proceedings of the ninth international conference on machine learning and cybernatics,* July 2010.

[11] Zhu XiaoKai and Li Xiang, "Image kernel for recognition,"*Proceedings ICSP2008,* 2008.

[12] Fei Ye, Zhiping Shi and Zhongzhi Shi, "A comparative study of PCA, LDA and Kernel LDA for image classification",*International Symposium on ubiquitous virtual reality,* 2009.

[13] Wen-Sheng Chen, Pong C Yuen, and Zhen Ji, "Kernel subspace LDA with convolution kernel function for face recognition,"*International Conference on Wavelet Analysis and Pattern Recognition,* 2010.

[14] Yunfei Jiang, Xinyu Chen, Ping Guo and Hanqing Lu, "An improved random sampling LDA for face recognition,"*Congress on Image and Signal Processing,* 2008.

[15] Jin HuaXu, Hong Liu, "Web user clustering analysis based on k-means algorithm,"*International conference in information, networking and automation (ICINA),* 2010.