Gender Classification Using Discrete Cosine Transformation: A Comparison of Different Classifiers

*Abdul Majid, Asifullah Khan and Anwar M. Mirza**
Faculty of Computer Science & Engineering, GIK Institute, Topi-23460, Swabi, Pakistan
E-mail: mirza@giki.edu.pk

Abstract-- *In this paper, we have investigated the problem of gender classification using a library of four hundred standard frontal facial images employing five classifiers, namely K-means, K-nearest neighbors, Linear Discriminant Analysis (LDA), Mahalanobis Distance Based (MDB) classifiers and our modified KNN classifier. The image data independent discrete cosine transformation (DCT) basis is used for facial feature extraction. Areas under the convex hull (AUCH) of the classifiers are measured by varying the values of threshold for each feature subset in receiver operating characteristics (ROC) curve. The scalar values of AUCH of ROC curve increases with increasing number of features. More features yield a better representation of the gender facial image. The overall performance of classifiers is compared with different values of AUCH versus features under different conditions. It has been observed that when the number of features is increased beyond 5, AUCH starts to saturate. Our experimental results demonstrate that modified-KNN performs better than the rest of conventional classifiers under all conditions. LDA classifier did not perform well in DCT domain, however, it gradually improved its performance with increasing number of features.*

Keywords: *Gender Classification, Discrete Cosine Transformation (DCT), Receiver Operating Characteristics Curve (ROC), Jackknife Technique, Area Under the Convex Hull (AUCH), Karhunen Louve Transformation (KLT).*

1. INTRODUCTION

Gender Classification is a binary classification problem, in which one has to predict an image as that of a man or woman. Gender Classification is an easy task for humans, but it is a challenging task in computer vision. The major difficulty with representing faces as a set of features is that it assumes some priori knowledge about what are the features and/or what are the relationships between them that are essential to the task at hand. Gender classification [2] can be attained using a suitable feature extraction technique from statistical structure of the learned faces. Several techniques for facial feature extraction have been proposed. They include methods based on statistical features [14], geometry features and neural networks [16,17].

High redundancy present in facial images not only degrades the performance of classifiers but also introduce inefficiency when these images are used directly for recognition, identification and classification. A computational model is build to transform pixel images into face features, which generally is robust to variation in illumination, scale and orientation. These features are then used for recognition. Various authors [1, 12,13] have used a set of eigenfeatures based on data dependent KLT transformation for gender classification problem. The major drawback of this transformation is the lack of adaptiveness about the prediction of a new face. Each time new KLT basis is constructed for a new gender, which is computationally inefficient. DCT transformation can find unique discriminate feature vectors accurately between male and female. This technique [3,4] was suggested to reduce the dimension of facial images. But we have used this technique at the preprocessing stage for facial feature extraction to analyze the performance of different classification algorithms. Classifiers are used largely to obtain useful information from large data. Classifiers are an important component of intelligent systems and have wide range of applications. In this paper we have also modified KNN classifier and proposed a new modified-KNN classifier. This classifier performs better not only with respect to conventional KNN, but also the rest of classifiers namely LDA, MBD and K-means.

The remaining part of paper is organized as follows: In section 2, we describe briefly DCT based gender classification system and its modules. Followed by classifiers performance criteria in section 3. In section 4 different experimental results are shown. Discussion and analysis in section 5, finally conclusions in section 6.

2. SYSTEM DESCRIPTION

Gender classification system can be divided into two parts: feature extraction and classification. The main idea is to apply DCT to reduce the information redundancy and to compare the performance of different classifiers in that domain under different conditions. For input face images system first computes and select the limited DCT coefficients, feeds them as input to the chosen classifier. Finally classifier output prediction about gender face. A schematic diagram of DCT based gender classification system is shown in figure 1

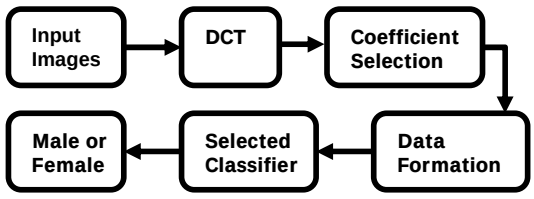

Figure 1: A schematic diagram of DCT based gender classification system

2.1 Face Images Database

Stanford university medical student (SUMS) frontal facial images database has been used as input to the classification [15] This database consists of 200 male and 200 female gray scale images of size 128x128 pixels. .

Figure 2: A sample of (SUMS) gender face database

Scaled down jackknife [9,10] scheme is used to utilize the image database efficiently and the performance of the algorithms is tested against the four different selected training to test ratios (1:3, 3:1, 9:1 and 1:9). For example, in case of 1:9 training to test ratio, 20 males and 20 females images are selected out of the total 400 images, for training purposes. Remaining 360 images are used for testing of the classifiers. This process is repeated 10 times, thus allowing all of the images to be in a train set exactly once. The algorithm results were averaged across all test sets to increase their statistical significance. For another training to test ratio of 1:3, 50 male and 50 female images were randomly picked for training and the remaining images were left for testing from 400-face database.

2.2 Discrete Cosine Transformation

Statistical classifiers are trained with a set of known images. It is often much better to reduce the number of dimensions, in an efficient way, allowing accurate training with less observations. Discrete Cosine Transform is a well-known transformation technique used for image compression applications. We use it for dimension reduction in face recognition application to investigate the behavior of statistical classification algorithms. The idea was to use the most significant coefficients to represent the image space, and classify each image based on these coefficients [3,4]. This idea has been applied successfully and we have tested its abilities for gender classification DCT transform. Like Fourier transform, uses sinusoidal basis. The following equation is used for DCT coefficients C (u, v) for an image function f (x, y) of m x n is

$$C(u,v)=\frac{2}{\sqrt{mn}}\left(a(u)a(v)\right)\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}f(x,y)\cos\left[\frac{(2x+1)u\pi}{2m}\right]\cos\left[\frac{(2y+1)v\pi}{2n}\right] \quad (1)$$

For u = 0,1,2,3…m-1 v = 0,1,2,3…n-1

Image basis of DCT are input independent and it has high information packing ability. This property of DCT has proven to be very practical for JPEG image compression

2.3 DCT Coefficients Selection

The most significant DCT coefficients are chosen from the given face image by determining which coefficients have the greatest variance. In the natural images, the DCT coefficients variances drop off for the higher frequency components. DCT coefficient's energy compaction property is used in image data compression. This compaction of variance in the lower frequency components is exploited in the dimension reduction of face image. It is carried out by picking only starting low frequency coefficients and removing the others. These DCT coefficients subsets are used to train and test the classification algorithms.

2.4 Data Formation

We use a method of normalization [4] for computing the most significant DCT coefficient of training and test images in order to improve the learning speed of classifiers. The upper and lower bounds of the facial images data in DCT domain are first determined. Before feeding these to the classification algorithms, the DCT coefficients are normalized within the range [-1, +1] using these upper and lower bounds.

2.5 Classification Algorithms

Classifiers are largely used to obtain useful information from large data. Classifiers are an important component of intelligent systems and have wide range of applications. A classification procedure can be used with some formal methods for making decision in new situations. The output of each classifier is scaled to 0 - 1 range. The selected threshold T ($0 \le T \le 1$) is then applied to the output of classifier to check the performance of classifier in the whole threshold range. The k-means algorithm was chosen as a lower-bound benchmark algorithm. We chose the algorithm because it attempts to divide the data based solely on the natural separation of the data. It does not use any information with respect to the gender of each person to perform the classification. It finds the natural separation in face space based on gender. The algorithm works by picking *K* random data points as center points, where *K* is the number of groups that one would like to find. In our case, *K* was equal to two: one group for males and another for females. Next, the algorithm calculates the Euclidean distance from each training point to each center point. The training point is grouped with the center point that is closest to it. Once all the training points are grouped, a new center point is calculated for each group based on the mean value of all the data points in the group. This process is continued until

the center points do not move anymore. Grouping data based on the Euclidean distance between a test point and two center points is the same as dividing the data with a hyper-plane that splits the two center points.

LDA classifier calculates a set of weights that specify a hyper plane that splits the data into two groups. Finding the optimal value for these weights is straightforward: Given a training data matrix *M*, and a vector *G* that specifies the gender of each member of the training set (for instance: a 0 for males and a 1 for females), the unknown vector of weights W can be found by solving the linear system:

$$M^{T} \cdot W = G \qquad (2)$$

Once W is calculated by multiplying by G an appropriate pseudo-inverse of *M*, determining the gender of a test face X is accomplished by computing the inner-product of X and W. For example, if threshold=0.5, then if $<X, W> <= 0.5$, the test face is classified as a female, otherwise male.

A different way to classify data within two groups is by using the Mahalanobis distance instead of Euclidean distance. The Mahalanobis metric represents the distance from the mean group value that has a constant covariance; so in two-dimensions, this distance is given by an ellipsoid. As an example, a cut at a certain height through a two-dimensional Gaussian distribution represents a Mahalanobis curve. The equation for this metric is:

$$r^2 = (x - m_x)^T C_x^{-1} (x - m_x) \qquad (3)$$

Where $\mathbf{m}_x$ and Cx represents the mean and the covariance matrix respectively. The classification process consists of calculating the Mahalanobis distance of a test point to the mean of the two groups and then deciding which mean is the closest one.

In the KNN classifier, the nearest neighbor approach is used to measure the distance of the test sample to every training sample, rather then just the distance to the mean training sample. The test sample is then assigned to the class, which has the shortest distance. This is the special case of shortest K-nearest neighbor approach where K =1. Where K>1 the method is made sensitive to outliers. Rather then choosing the class with the shortest distance to a training sample, the class with the K-nearest neighbor is chosen. This help smooth out the distribution, lessening the effect of outliers. Both nearest neighbor approach have the practical disadvantage that all the projected training samples must be stored and search during the testing phase.

We modified the K-nearest neighbor approach by considering neighbors in annular strips instead of keeping K prefixed. It gives more weightage to the smaller strips, i.e. closest neighbors. As in K-nearest neighbor approach first the Euclidian distance of the test sample to every training sample is computed. The mean of these distances is then used as a distance measuring metric in this Euclidian space. We have used two annular strips of radii 1/3 and 1/5 of mean distance. The strips were given weightage as follows:

$$W(r) = e^{-2r} \qquad (4)$$

Where 'r' is the strip radius in terms of the mean distance. This was obtained through heuristics. This modified KNN (although computationally expansive compared to KNN) outperforms the KNN algorithm, as it has more confidence in decision and further lessens the effect of outliers. Large weightage is given to small radii strips to incorporate greater effect of the nearby samples.

3. PERFORMANCE MEASURE CRITERIA

The performance of a classifier can be measured through Receiver Operating Characteristics (ROC) curve. ROC curve is a characteristics curve of a classifier for a particular problem. It summarizes how well the classifier has performed for that problem at different thresholds. It allows us to show graphically the trade off of each classifier between its true positive rate (the number of correct positive cases divided by the total number of positive cases) and its false positive rate (the number of incorrect positive cases divided by the total number of negative cases) [5]. Although single figures of merits are useful when comparing a classification system under a number of different conditions or settings, one of the greatest assets of testing is lost because they don't characterize the system over its entire operating range [6]. Hence for adjusting the classification threshold we should plot ROC curves. The selection of operating threshold is then application-specific, depending on the maximum acceptance of false and true positives. The ROC curve allows the operator to select the operating point, which best fulfills their requirements, or simply reject the system outright if it is unable to meet their needs. In the absence of an application-specific operating point, the equal error rate (ERR) can be used to provide a single figure of merit. This is the point on the ROC curve where the likelihood of a false positive and true positive are equal. To obtain an ROC curve, each classifier is treated, as though it has a sensitivity level (threshold), which allows the classifier to be tuned. When the sensitivity is at the lowest level, the classifier produces no false alarms, but detects no positive cases, i.e. the origin of the ROC. As the sensitivity is increased, the classifier detects more positive examples but may also start generating false alarms (false positives). Eventually the sensitivity may become so high that the classifier always claims each case is positive [7]. This corresponds to the top right hand corner of the ROC. On average a classifier which simply makes random guesses will have an operating point somewhere on the diagonal line between the origin and top right hand corner (1, 1).

Naturally one wants the true positive rate to be as high as 1, and the false positive rate to be 0, i.e. points at the top left corner of the ROC curve. Both the Y-axis and X-axis are normalized (range 0-1); hence the area under the

ideal ROC curve will be 1. Hence a given classifier is said to be an optimal one for a given problem, if the area under its ROC is near to 1. An ROC curve for six features subset as shown in figure 3.

A "maximum realizable" ROC [8] is the convex hull of the classifier's ROC. Therefore the Area Under the Convex Hull (AUCH) of an ROC curve is taken as a measure of the performance of a classifier. AUCH is obtained for each of the resulting ROC curve. In other words varying feature subset give us different ROC curves for the same classifier. These different ROC curves, obtained for the same classifier, have different values of AUCH. When these values of AUCH are plotted against number of features, we obtain curves like shown in figure 4 to figure 11.

4. EXPERIMENTAL RESULTS

Each classifier is treated, as though it has a sensitivity level. This sensitivity level (threshold) is varied from 0 to 1 with step 0.1 for each feature subset. AUCH is obtained for the resulting ROC curve. The following figures (see figure 4 to figure 8) are considered for four different training to test ratios against the varying number of features. The classifiers are compared for 20 features, starting from the first feature, the feature subset is increased. The values of AUCH of ROC curve obtained are shown in figures 4,6,8 and 10, while for 50 features in figures 5,7,9 and 10 for different training to test ratios.

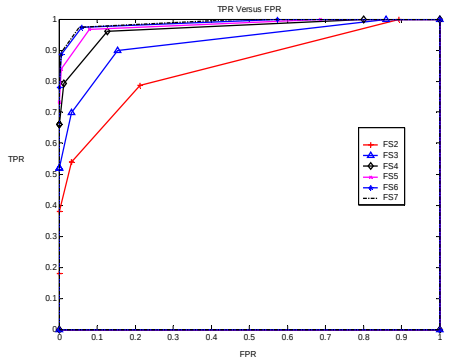

Figure 3: ROC curves of KNN at different features subset

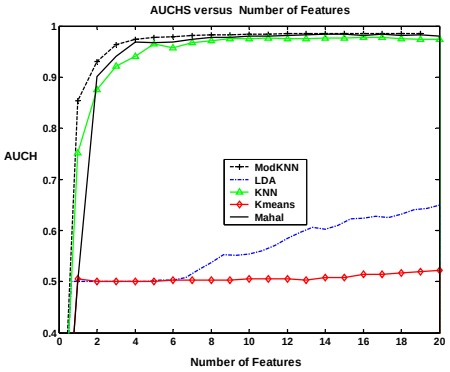

Figure 4: Comparison AUCHs of ROC curve of classifiers for first 20 features with 1:9 training to test ratio

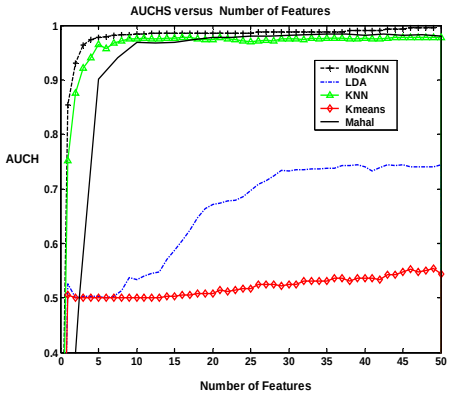

Figure 5: Comparison AUCHs of ROC curve of classifiers for first 50 features with 1:9 training to test ratio.

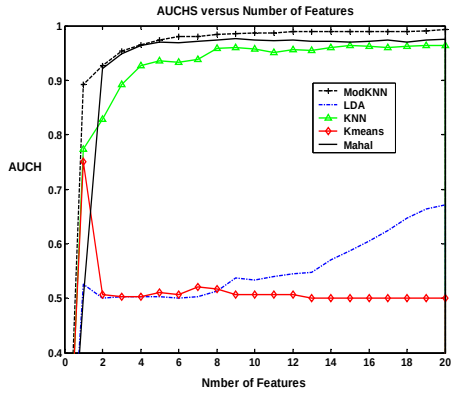

Figure 6: Comparison AUCHs of ROC curve of classifiers for first 20 features with 1:3 training to test ratio.

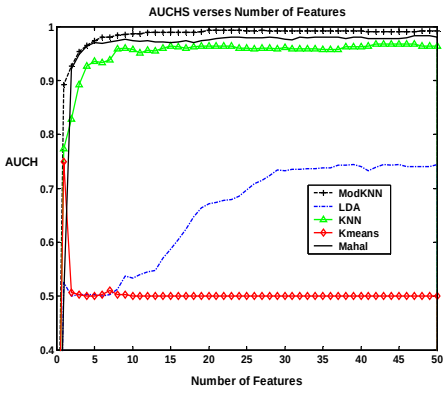

Figure 7: Comparison AUCHs of ROC curve of classifiers for first 50 features with 1:3 training to test ratio.

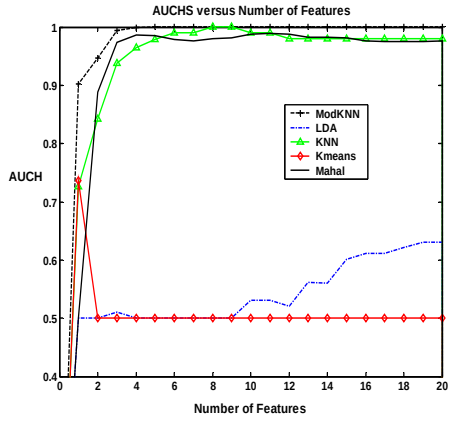

Figure 8: Comparison AUCHs of ROC curve of classifiers for first 20 features with 3:1 training to test ratio

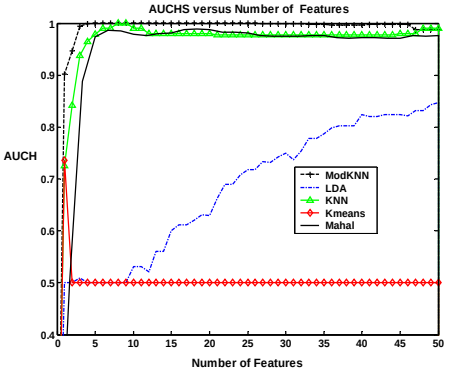

Figure 9: Comparison AUCHs of ROC curve of classifiers for first 50 features with 3:1 training to test ratio.

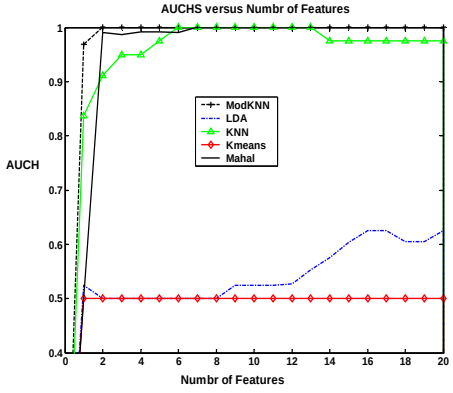

Figure 10: Comparison AUCHs of ROC curve of classifiers for first 20 features with 9:1 training to test ratio.

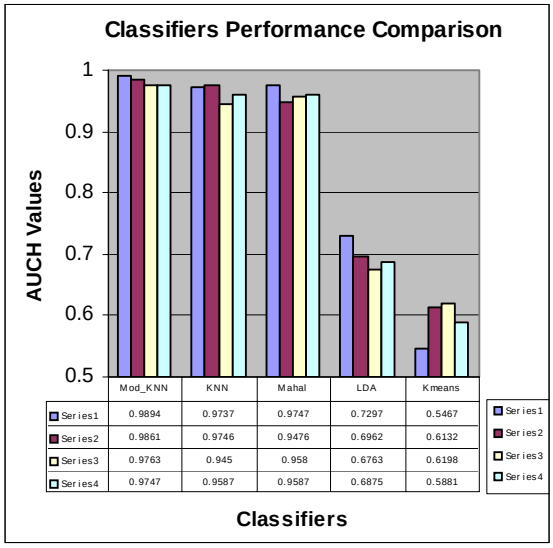

Figure 11: Bar Chart showing the performance of classifiers in terms of AUCHS of ROC curve
Series 1 training: test ratio = 9:1, features = 50
Series 2 training: test ratio = 3:1, features = 50
Series 3 training: test ratio = 1:3, features = 50
Series 4 training: test ratio = 1:9, features = 50

5. DISCUSSIONS

The general trend for all the classifiers from the figure 4 to figure 10 is that the values of AUCH start to saturate more slowly for low training to test ratios 1:3 and 1:9, while values of AUCH saturate sharply for high training to test ratios of 3:1 and 9:1. In case of Mod-KNN classifier the values of AUCH saturate even after one feature subset due to large availability of training data. Mod-KNN classifier performs better then the rest of classifiers even for less number of features.

LDA classifiers are not performing well in all training to test ratios. LDA classifier gradually improve its performance with increasing number of features due to the reason that face database independent basis are used for facial feature extraction. The AUCH values of LDA limited at approximately 0.70 in all training to test ratios.

It is obvious from the figure 4 to figure 11 that the AUCH values for K-means do not vary appreciably with the increase in feature subset. This is due to the fact that gender is not the main separation factor in data. Skin color, glasses or no glasses, close picture or far picture may have the stronger separation effect than gender. The K-means performance degrades with increase in feature subset. This happened because AUCH values are greatly affected by the initial abscissa values and this effect decreases with increase in number of feature subset. Hence AUCH values of ROC curve do not depict the decrease in performance of K-means. Our modified-KNN performs better than the rest of classifiers for all training to test ratios (see figure 11: Comparison of Classifiers). For first 50 features Mod-KNN has higher AUCH values than its closest companion

KNN. KNN has slightly better performance than Mahalanobis classifiers. In figure 11, it is observed that the classifiers AUCH values are improved with increasing training samples due to the availability of more information for better decision. When the feature subset is increased beyond 20 up to 50, AUCH values increases for all of the classifiers.

6. CONCLUSIONS

We have improved the discrimination power of conventional KNN classifier by dividing the mean of the nearest decision region into two annular strips and more weight-age is given to the closer strip and have proposed a new Mod-KNN classifier. Modified KNN can further be improved by following Triangle inequality [11]. Using only those feature vectors that are most important for gender classification instead of the first high-energy feature vectors could further improve the performance of a classifier. This might be carried out [3,4.] for feature selection and then sorting them according to their relevance, before being presented to the classifier. We have proposed a new technique, graphical representation of AUCH scalar value of ROC curve, which summarizes the performance of a classifier in an efficient way under different features subset**.**

Sponsorship:

This work is sponsored by the Ministry of Science and Technology and Higher Education Commission, Government of Pakistan.

REFERENCES

[1]. M. Burton, V. Bruce and N. Dench (1993), "What is difference between men and women? Evidence from facial Measurement." Perception, 22,153-176.

[2]. H. Abdi, D. Valentin, B. Eldmen and J.A.O'Toole (1995). More about difference between men and women: evidence from linear neural network and principal component approach", Perception, 24. 539-562.

[3]. Z. Pan, R. Adams and H. Bolouri, "Image Recognition using Discrete Cosine Transformation as Dimensionality Reduction" IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP01), 3 June - 6 June, 2001

[4]. Z. Pan, R. Adams and H. Bolouri, "Dimensionality Reduction for face Images Using Discrete Cosine Transformation for Recognition" Technical Report, Science and Technology Research Centre (STRC)

[5]. J. A. Swets, R. M. Dawes and J. Monahan. "Better decisions through science." Scientic American, pages 70-75, October 2000

[6]. C. Michel Lincoln, "Pose independent face recognition", PhD thesis, Dec. 2002, Department of Electronic System Engineering, University of Essex.

[7]. W. B. Langdon and B. F Buxton, "Genetic programming for combining classifiers", In GECCO'2001. Morgan Kaufmann

[8]. M. J. J. Scott, M. Niranjan and R. W. Prager, "Realizable classifiers: Improving operating performance on variable cost problems" Ninth British Machine Vision Conference, Volume 1, pages 304-315,University of Southampton, UK.

[9]. M. Kirby and L. Sirvich "Application of Karhunen Louve procedure for the application of human " IEEE, 12:103-108, 1990.

[10]. L. Sirovich and M. Kirby," Low dimensional procedure for the characterization of human face."J,Optical Society Am 4 p.no.519-524, 1987.

[11]. M.A. Greenspan, G. Godin, and J. Talbot. "Acceleration of Binning Nearest Neighbor Methods.", Proceedings of Vision Interface 2000, Montreal, PP.337-344. NRC 44167, 2000

[12]. H. Abdi. (1988), "A general approach for connectionist auto-associative memory: interpretation, implication an illustration for face processing", In J. Demongeot (Ed). Artificial intelligence and cognitive sciences. Manchester University Press.

[13]. D. Valentin, H. Abdi, B.E.Elderman and A. J.O'Tooli, "Principal Component and Neural Net Analysis of Face Images: What can be generalized in gender classification?" Journal of Mathematical Psychology, 41,398-412. 1997.

[14]. A. Lanitis and C. Talor, "automatic interpretation and coding of face images using flexible models" IEEE Transaction on Pattern Analysis and Medicine Intelligence" vol. 19,no. 7, 1997.

[15]. http:/iseo.standford.edu / class / ee368-project00 // project15.

[16]. C. Nebaur "Evaluation of convolutional neural network for visual recognition" IEEE Transaction on Neural Network vol. 9 No. 4, 1998.

[17]. J. Zhang and Y. Yan "Face recognition: Eigenface, elastic matching and neural net" IEEE proceeding vol.85, no.9, 1997.