# Human Face Recognition and Identifying Ethnicity Using Machine Learning

C. Christy[1], S. Arivalagan[2], P. Sudhakar[3]

[1]*Ph.D. Research Scholar, Department of Computer and Information Science, Annamalai University*

[2]*Assistant Professor, Department of Computer Science & Engineering, Annamalai University*

[3]*Assistant Professor, Department of Computer Science & Engineering, Annamalai University*

*Abstract*—**Data Mining and Knowledge Discovery are used to extract meaningful pieces of information from the source data. Data Mining and Image Analysis area are taking the significant feature by extract Knowledge from an Image. Human face recognition is interesting and challenging problem. It is an important domain such as human-computer interaction and data-driven animation. It is useful for identifying social ethnicity in a large heterogeneous group of people gathering in a common place. A huge set of training sets for personal identification attains good accuracy with the face recognition system. Feature extraction providing good result for one image per person with small training sets. In this Research, three very popular data mining techniques for classification such as Naive Bayes, Decision trees and Random Forest algorithm applied on various images, freely available on the Internet for analysis. The normal image has experimented with the above three algorithms followed by a Gaussian noise. A Kuwahara filtering process is done at the second check. From this, it is to be observed that the particular human face belongs to which origin ethnicity social group. The Ethnicity classification task is done by Linear Discriminant Analysis (LDA) based scheme. The two-class (Asian vs. non-Asian) is check by LDA. The input facial images scaled by Multiscale and ensemble integrated LDA analysis. So the classification performance is improved.**

Keywords— **Image, Data Mining, Gaussian noise, Kuwahara Filter, LDA Ensemble.**

## I. Introduction

Data Mining and Knowledge Discovery attracting many researchers to extract meaningful pieces of information from the dataset. In the field of research, data mining and extracting knowledge is very popular. In both Data Mining and Image Analysis area, Image Analysis and Knowledge Discovery from an Image is obtaining significant role. The various classes of their visual characteristics classify by the image classification. Problem is to classify the human face to find which origin that facial image belongs. Face images and personal identification is done by separation of ethnicity. Face images are one of the representations of the ethnic classification. The demographic information provides ethnicity and gender using human face images. In face related applications ethnicity and gender also play important. Ethnicity identification problem is based on image using of machine learning framework. The face images identified by using classifier algorithms. Identification of images affected by Noise, shadow, and light. So, the Anthropology identification is done by using the facial organs such as eyes, nose, forehead, and mouth. Features extraction is based on the corner points. Noise is reduced by using filters. In many social applications, demographic statistics analyzed by using ethnicity and gender. The ethnicity classification into a two-category classification problem is discussed by this paper.

## II. RELATED WORK

Recognizing faces of their own ethnicity/race by humans [1, 2]. More activity in brain regions linked to face recognition discussed by Golby et al. [3]. The same-race for face identification involves Form Face Area (FFA) is examined by Functional magnetic resonance imaging (FMRI)[4]. Investigate the differences in the way people perceive own- versus other-race faces shown by O'Toole et al. [ 5 ]. People categorize faces of their own-race by sex more efficiently by O'Toole et al. [6]. Database is focused by Identity-related features so; retrieval of information is very effective. [1, 2]. Same-race faces elicit more activity in brain regions linked by face recognition Golby et al. [3]. Important for face recognition identified by the Functional Magnetic Resonance Imaging (FMRI), same-race for face identification involves form FFA, [4]. The way people perceive own- versus other-race faces investigated by O'Toole et al. [5]. The people categorize faces of their own-race by sex more efficiently than another race discussed by O'Toole et al. [6, 8]. The face recognition system of race and gender can help identification to focus more on the identity-related features, and limit the number of entries to be searched in a large database, the search speed and efficiency of the retrieval systems improved. The demographic statistics in many social applications ethnicity and gender are also useful. The ethnic categories are loosely defined classes than the identity. Ethnicity classification into a two-category (Asian and non-Asian) classification problem is discussed in this paper. In [7], two types of features were used, Linear Discriminant Analysis based algebraic features an elastic model based geometric features. They classified images into three minority Chinese groups. An accuracy of 79% was reported using algebraic features and 90.95% with geometric features with K-Nearest Neighbors (k-NN) and C5.0 classifiers.

## III. PREPROCESSING

### *3.1 Feature Extraction.*

Feature Extraction is nothing but transforming the input data into the set of features. The features set will perform the desired task, if the extracted features are carefully chosen. It is expected that using the reduced representation instead of the full-size input. Extracting relevant information from an image is the process of feature extraction.. After detecting a face, some valuable information is extracted from the images which are used in the next step for identifying the image.

*3.1.1) Gaussian Noise:*Gaussian noise occurs in images during acquisition. This Gaussian noise can be reduced using of kuwahara filter.When smoothing an image in a spatial filter, an undesirable outcome may result in the blurring of fine-scaled image edges and details because they also correspond to blocked high frequencies.

*3.1.2)Kuwahara filtering:* The kuwahara filter is a non-linear smoothing filter. The advantage of kuwahara filter is, it's used for adaptive noise reduction and able to apply to smooth on the image while preserving the edges. The other filters also reduce noise but also blur out the edges.

### *3.2 Classification.*

In the Classification system database is very important that contains predefined sample patterns of an object under consideration that compare with the test object to classify it, appropriate class.Image Classification is an important task in various fields [8].

*3.2.1) Naive Bayes:*This Classifier is based on Bayes Theorem with independence assumptions between predictors. It is easy to build. No complicated iterative parameter Estimation. Useful for very large datasets. The condition probability for this large data set is P(C/X)=P(X/C)P(C) /P(X).Bayes theorem provides a way of calculating the posterior probability, *P*(C/X), from P(C), P(X), and P(X/C).The effect of the value of a predictor (*X*) on a given class (*C*) is independent of the values of other predictors assumed by Naive Bayes classifier. This assumption is called class conditional independence.

*3.2.2) Decision Tree:* A classification or regression model builds by Decision tree in the form of a tree structure. The data set is broken by smaller subsets, at the same time an associated decision tree is incrementally developed. A tree with decision nodes and leaf nodes are the final result. A decision node has two or more branches. Leaf node represents a classification or decision. The root node is a decision node in a tree and it is called the best predictor. Both categorical and numerical data handled by Decision trees.

A decision tree is partitioning the data into subsets that contain instances with similar values (homogenous). The homogeneity of a sample calculated by the use of ID3algorithm .The entropy is zero when the sample is completely homogeneous. The entropy is one when the sample is equally divided. Entropy = $-p\log_2 p - q \log_2 q$. Calculate two types of entropy using frequency tables. Entropy using the frequency table of one attributes

$$E(S) = \sum - p_i \log_2 p_i. \text{ ---------------------- } (1)$$

Where, ∑ varies i-1 to c.

Entropy using the frequency table of two attributes.

$$E(T, X) = \sum P(C)\, E(C) \text{ ----------------- } (2)$$

Where, C € X.

*3.2.3) Random Forest:*

Random Forests grows many classification trees. Each tree is grown as follows:

- In the training set, if the number of cases is N and at the random sample N cases, replace the original data. The training set for growing the tree is the sample.
- At each node, if there is M an input variable, m variable is selected at random out of the M and is used to split of the node. The best split on this m.
- If there is no pruning, Each tree is grown at the largest possible extent.

*3.2.4)K-NN Classifier:*

The K-NN is also the classifier of the category of the supervised learning algorithm. K-nearest neighbors are easy and best algorithm that has a record of all available classes can perfectly put the new item into the class on the basis of the largest number of vote for k neighbors. In this way, KNN is one of the alternates to classify an unlabeled item into identified class. Selecting the no. of nearest neighbors or in other words calculating k value plays important role in determining the efficiency of the designed model. The accuracy and efficiency of the k-NN algorithm basically evaluated by the K value determined. A larger number for k value has an advantage in reducing the variance because of noisy data.

*3.3 Process Methodology.*

Input is the Normal facial image. This is Training Data. Preprocessing the image by applying Gaussian Noise and Kuwahara filter. Preprocessing and Extract the features and testing the data by applying classifications Naive Bayes, Decision trees, Random forest, and K-NN . Preprocessing use the very popular data mining techniques such as Naive Bayes, Decision trees and Random Forest algorithm and K-NN on various images, freely available on the Internet for analysis. Experiments are conducted for a Normal image at first then the three algorithms followed by a noisy one by applying Gaussian noise to it and then a Kuwahara filtering process at the second to check the effective classification of the model. LDA Linear Discriminant Analysis is evaluating the test result. From the experimental results, it is observed that the origin ethnicity belong the social group.

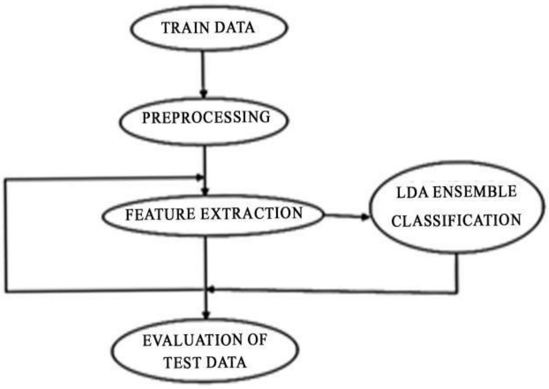

Fig-1 Process diagram

## IV. EXPERIMENTAL RESULT

*4.1 LDA Ensembles at Multiple Scales.*

Used for 2 class (Asia vs. Non-Asia) ethnicity classification task. Facial images are analyzed by Multiscale. An ensemble framework integrates the LDA analysis for the input face images at different scales. (Improve the classification performance).the ratio of between-class scatter to within-class scatter is maximized. Multidimensional data converted into a lower dimension by a statistical method. Consider a 2-D face image into an1-D vector, by combining each row (or column) of the image.

Data matrix $X = (x_1; x_2; : : : ; x_i; : : : ; x_N)$, where $n$ - total number of pixel in the image. N - Number of training set image. Original data matrix X is converted into data matrix projection Y.

CONVERSION OF MATRIX TO PROJECTED MATRIX

$$Y=W^TX;$$

$$W_{LDA} = \arg\max \frac{W^TS_BW}{{}_{W}W^TS_WW} \; ;$$

$S_B$ is the between-class scatter matrix and $S_W$ is the within-class scatter matrix,

*4.2 LDA Classifier:*

A low dimensional representation of a high dimensional face feature vector space derived by LDA. The transformation matrix is projected bythe face vector *W*. The feature representation of each face image is the projection coefficients.

*4.3* Methodology: (LDA Based classification)

- The Image is resized by three different scales.
- An LDA based classifier is constructed by each scale.
- The number of classifiers and number of scales is the same in ensemble.
- LDA based classifiers at different scales are the resulting ensemble.
- Training database is divided into two parts,( Two-thirds - training set , one-third -test set)

Same subject Images are grouped by eliminating the identity factor.

Data Set For Training

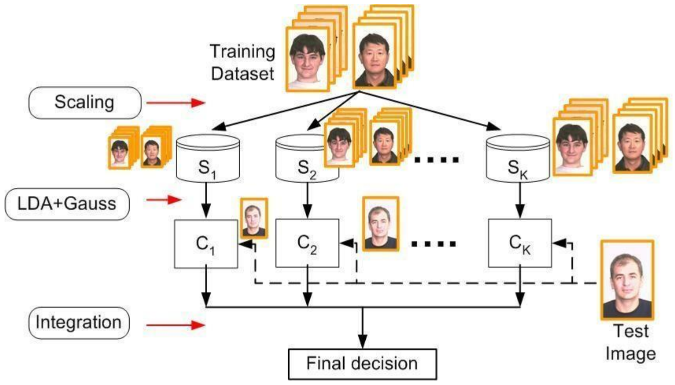

Fig-2 Data Set for Training

## V. CONCLUSION

This paperhasdiscussedan Efficient Image Mining Study Analysisbasedonfacialimages. Preprocessing is performed by the use of machine learning algorithms. Forthe development of two-class(Asian vs. Non-Asian) ethnicityclassification task, the Linear DiscriminantAnalysisbased schemehasbeen used. Classification performance is improved by an ensemble framework at different scales. Infuture Region of interest (ROI) histogram is also combined for finding the efficient identification of the social group.

## REFERENCES

[1] R. Malpass , J. Kravitz, J. Perc,*Recognition for faces of own Ethnicity/faces*, soc.Psychol.13,pp.330-334,1969.

[2] J.Brigham, P. Barkowitz, *Do they all look like? The effect of race, sex, experience, "Differential responses in the fusiform attitudes onthe ability to recognize faces*, Appl Soc.Psychol,pp.306-318, 1978.

[3] A.Golby, J.Gabrieli et al., *Different responses in the fusiform region to same race, other different race faces,*, Nature Neuroscience 4(8), pp. 845–850, 2001.

[4] A.Puce, T. Allison et al., *Face-sensitive regions in human extra striate cortex study by functional MRI*, Neuropsychology , vol.74, pp. 1192–1199, 1995.

[5] O'Toole,K.Deffenbacher, et al.,*Structural aspects of face recognition and another race effect*, Memory & Cognition 22, pp. 208–224, 1994.

[6] O'Toole, A. Peterson, et al.,*The race effect for classifying faces by sex*,Perception 25, pp. 669–676, 1996.

[7] X. D Duan, C. R. Wang, Z. J. Li, J. Perc etal.,*Ethnic feature extraction and Recognition of human faces*,

In Proceedings of the 2nd International Conference on Advanced Computer Control (ICACC), pages 125–130, 2010.

[8] Rama Gaur, Dr. V.S. Chouhan,”Classifiers in Image processing”, International Journal on Future Revolution in Computer Science & Communication Engineering, Volume: 3 Issue: 6 pages: 22 – 24. 2017

[9] Christy,S. Arivalagan, *Image Encryption Techniques based on Hash Algorithm*, International journal of pure and applied mathematics, Volume.119, NO.16. PP. 343-352, 2018.