Performance Evaluation of Face Recognition System Using Principle Component Analysis

Business Use Case:

Create a face recognition system using the Principle Component Analysis (PCA).

Data Set Used:

AT&T's "Database of Faces" data set for face recognition containing a set of 400 images; 10 different images each for 40 subjects was used. The images were extracted from this data set using the following link made available.

http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

The images extracted have been captured using different facial expressions at different times under varying lighting conditions and other details such as accessories (glasses/contacts).

Process:

The process required to create o build the face recognition system involves the following.

- Training data set is prepared using the individual feature vector of each of the image used by concatenation into a single data set
- Using the training data set, compute the mean of the training data
- Subtract the mean from the training data and compute its covariance matrix
- Eigen values and eigen vectors of the covariance matrix is computed
- Eigen vectors are sorted using the eigen values, resulting in a new feature vector (subspace) of reduced dimensions
- The training and test data sets are now projected into the new subspace; the feature vector with reduced dimensions (NOTE: data sets projected are matrices derived by subtracting the training data mean)
- The projected data is used to compute the Euclidean distance of test data from the corresponding training data
- Appropriate scores are assigned to the distance matrix derived as genuine (0) or imposter (1)
- Using the distance and score matrix, we identify the ROC curve, EER, FAR and FRR at varying threshold ranging from 0 to 1

Data Allocation:

The complete image data set was divided as follows to fit the model of the classifier and evaluate its performance.

- ***Mode 1:***

  - Training data set – first 5 images belonging to each of the 40 subjects
  - Test data set – remaining 5 images belonging to each of the 40 subjects
  - During projection of data into the new subspace (feature vector of reduced dimension), the originally identified training and test data is used

- ***Mode 2:***

  - Training data set – all 10 images belonging to first 25 subjects
  - Test data set – all 10 images belonging to remaining 15 subjects
  - During projection of data into the new subspace (feature vector of reduced dimension), the test data originally identified is divided again training and test data sets further as follows:
    - ✓ Training data set to be projected – first 5 images belonging to each of the 15 test subjects
    - ✓ Test data set to be projected – remaining 5 images belonging to each of the 15 test subjects

Findings:

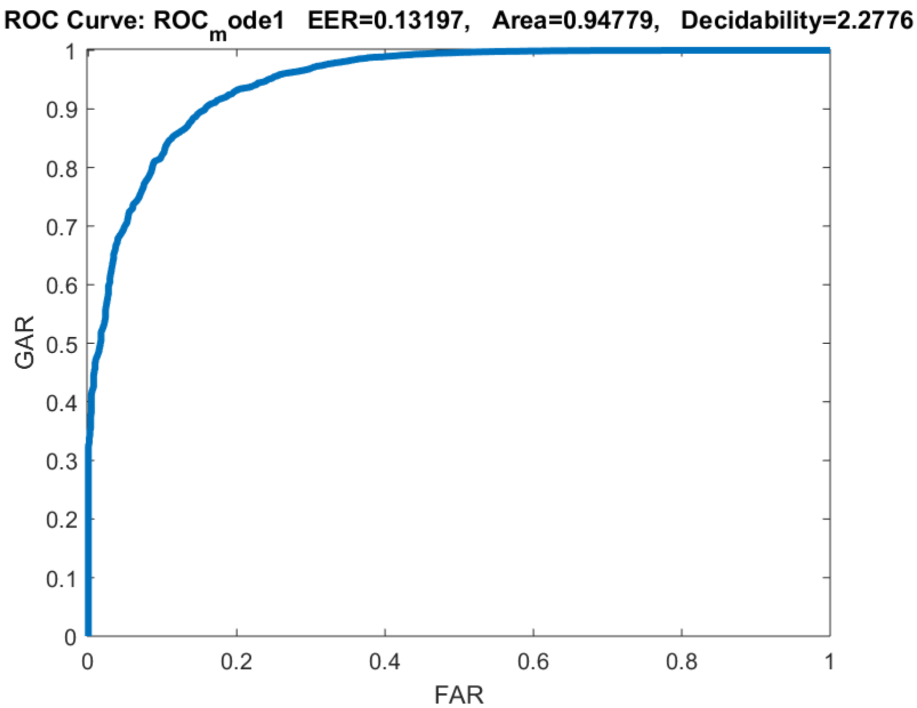
Projection of the test data set on this reduced dimension feature vector resulted in the below False Acceptance Rate (FAR) and False Rejection Rate (FRR).

MODE 1		***MODE 2***	
FAR	***FRR***	***FAR***	***FRR***
0%	0.933259	*0%*	0.8994118
5%	0.297256	*5%*	0.443428571
10%	0.176615	*10%*	0.30847619

Along with the above values, the following ROC curves and corresponding information was also obtained using the “ezroc3” function.

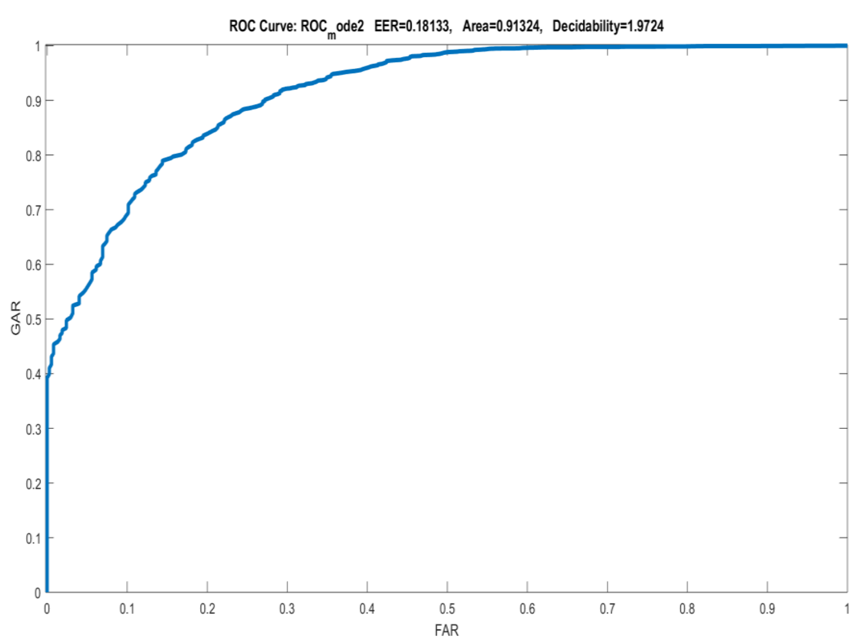
- ***ROC for Mode1:***

  - EER = 0.13197
  - Area = 0.94779
  - Decidability = 2.2776



- ***ROC for Mode 2:***

  - EER = 0.18133
  - Area = 0.91324
  - Decidability = 1.9724



Key Factor of Analysis:

The key factor resulting in different values of FAR and FRR is the inclusion of some of the images of the subject belonging to the test data set during training stage for mode 1, while the test data set is completely new (independent/unknown) for mode 2 as none of these subjects where involved during training.

Accuracy of Classifier:

Computation of the accuracy of the classifier using the findings above using formula:

Accuracy of Classifier = 1 – ((FAR + FRR)/2)

The accuracy of classifier when the threshold is set to 0 is

- Accuracy for Mode 1 = 1 – ((0 + 0.93)/2) = 1 – 0.465 = 0.535 ~ 54%
- Accuracy for Mode 2 = 1 – ((0 + 0.44)/2) = 1 – 0.22 = 0.78 ~ 78%

The accuracy of classifier when the threshold is set to 0.05 is

- Accuracy for Mode 1 = 1 – ((0.05 + 0.89)/2) = 1 – 0.17 = 0.83 ~ 83%
- Accuracy for Mode 2 = 1 – ((0.05 + 0.44)/2) = 1 – 0.245 = 0.755 ~ 76%

The accuracy of classifier when the threshold is set to 0.10 is

- Accuracy for Mode 1 = 1 – ((0.10 + 0.17)/2) = 1 – 0.135 = 0.865 ~ 87%
- Accuracy for Mode 2 = 1 – ((0.10 + 0.30)/2) = 1 – 0.2 = 0.8 ~ 80%

ACCURACY OF CLASSIFIER	***MODE 1***	***MODE 2***
@ 0% Threshold	54%	78%
@ 5% Threshold	83%	76%
@ 10% Threshold	87%	80%

Observation & Analysis:

When the training set involves part of the test data set (mode 1) as well, the model is trained well leading to better performance of the classifier during the testing. It results in the model fit being more robust and accurate. This means the reduced feature vector identified uses more robust and important features to recognize the images. It accommodates for the different aspects within the images viz. lighting, alignment, background etc. in a better fashion resulting in more accurate recognition.

On the other hand, when the test data set is unknown during the training stage (mode 2), the classifier does not perform as well as it does during mode 1. This is resulting from a more generalized hypothesis. Thus, there is room to improve the performance of the classifier by reducing the dimensions further incorporating the most relevant and important features for classification (image recognition in this scenario).

Learning Outcomes:

The performance or accuracy of the classifier is highly impacted by the data used for training the model and testing the same. When there is overlap of subjects' data in training and testing stages, the performance of the model/classifier is better as opposed to when the test data is completely unknown. Thus, the model using the former case is more robust. However, there may also be a chance of overfitting of the model resulting in inaccuracy in the performance of the classifier in the long run. Thus, the approach always depends on the use case in terms of purpose, application, type of data etc. associated with the same.