MACHINE LEARNING
ASSIGNMENT 8
Principal Component Analysis

INTRODUCTION

We had to implement Principal component analysis to develop a face recognition system. Also, we had to record the change in classification accuracy along with the change in K(Number of feature vectors i.e. Eigen Faces in the case of Face recognition).

Also while testing recognise the impostors(Faces which do not belong to our training Set)

Please note that the two files “IIT2018178.ipynb” and “IIT2018178.py” are exactly the same and have no difference in codes other than small changes like printing a line before finally printing accuracy .etc. These two are included so you can run the code in terminal as well as juptyer notebook i.e. according to your preference.

RESULTS

Our database consisted of 40 folders which consists of 10 pictures of a person. All the 40 folders consisted of different people. Since we had to split our training and test set in 60:40 ratio, we inserted 6 pictures in training set and 4 in test set from each folder until 38 folder. All the pictures in the last 2 folders were inserted in test set and will be used as unknown faces.

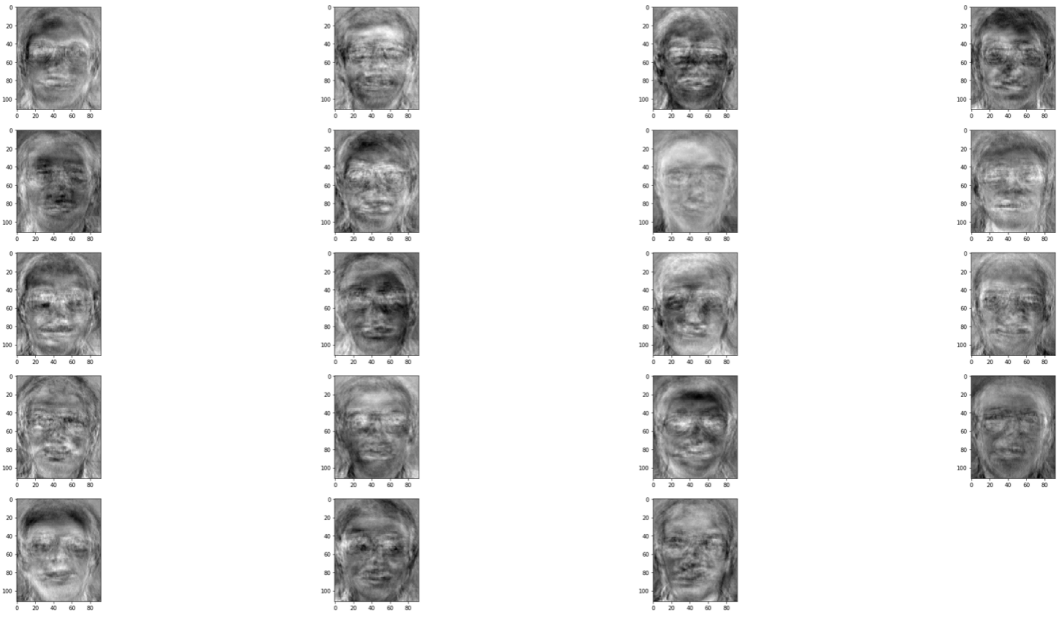
After running the program for multiple ‘k’ values we get 19 feature vectors as the best number of feature vectors to classify our dataset.

As we start increasing K from 5 we can see a sharp increase in our accuracy until 19 after which our accuracy stays saturated until k = 30. And After 30 we can see a constant decrease in the accuracy of our classifier. Also note that at k = 19 we have the highest accuracy.

So we choose k as 19. You can see the screenshot of the trend in the image attached below.

```
The accuracy of our model for k :  5  is :  59.30232558139535 %
The accuracy of our model for k :  6  is :  62.7906976744186 %
The accuracy of our model for k :  7  is :  66.86046511627907 %
The accuracy of our model for k :  8  is :  69.76744186046511 %
The accuracy of our model for k :  9  is :  68.6046511627907 %
The accuracy of our model for k :  10  is :  67.44186046511628 %
The accuracy of our model for k :  11  is :  71.51162790697676 %
The accuracy of our model for k :  12  is :  72.67441860465115 %
The accuracy of our model for k :  13  is :  73.83720930232558 %
The accuracy of our model for k :  14  is :  75.0 %
The accuracy of our model for k :  15  is :  75.0 %
The accuracy of our model for k :  16  is :  77.32558139534885 %
The accuracy of our model for k :  17  is :  79.65116279069767 %
The accuracy of our model for k :  18  is :  80.23255813953489 %
The accuracy of our model for k :  19  is :  81.3953488372093 %
The accuracy of our model for k :  20  is :  77.90697674418605 %
The accuracy of our model for k :  21  is :  79.65116279069767 %
The accuracy of our model for k :  22  is :  80.23255813953489 %
The accuracy of our model for k :  23  is :  79.06976744186046 %
The accuracy of our model for k :  24  is :  79.06976744186046 %
The accuracy of our model for k :  25  is :  80.81395348837209 %
The accuracy of our model for k :  26  is :  80.23255813953489 %
The accuracy of our model for k :  27  is :  79.65116279069767 %
The accuracy of our model for k :  28  is :  79.06976744186046 %
The accuracy of our model for k :  29  is :  80.23255813953489 %
The accuracy of our model for k :  30  is :  80.23255813953489 %
The accuracy of our model for k :  31  is :  79.06976744186046 %
The accuracy of our model for k :  32  is :  78.48837209302324 %
The accuracy of our model for k :  33  is :  79.65116279069767 %
The accuracy of our model for k :  34  is :  77.32558139534885 %
The accuracy of our model for k :  35  is :  77.90697674418605 %
The accuracy of our model for k :  36  is :  76.16279069767442 %
The accuracy of our model for k :  37  is :  75.5813953488372 %
The accuracy of our model for k :  38  is :  75.0 %
The accuracy of our model for k :  39  is :  74.4186046511628 %
The accuracy of our model for k :  40  is :  73.25581395348837 %
The accuracy of our model for k :  41  is :  71.51162790697676 %
The accuracy of our model for k :  42  is :  71.51162790697676 %
The accuracy of our model for k :  43  is :  71.51162790697676 %
The accuracy of our model for k :  44  is :  70.93023255813954 %
The accuracy of our model for k :  45  is :  69.76744186046511 %
The accuracy of our model for k :  46  is :  69.18604651162791 %
The accuracy of our model for k :  47  is :  68.6046511627907 %
The accuracy of our model for k :  48  is :  65.69767441860465 %
The accuracy of our model for k :  49  is :  65.11627906976744 %
The accuracy of our model for k :  50  is :  63.95348837209303 %
The accuracy of our model for k :  51  is :  63.95348837209303 %
The accuracy of our model for k :  52  is :  62.7906976744186 %
The accuracy of our model for k :  53  is :  62.2093023255814 %
The accuracy of our model for k :  54  is :  61.627906976744185 %
The accuracy of our model for k :  55  is :  59.883720930232556 %
The accuracy of our model for k :  56  is :  59.30232558139535 %
The accuracy of our model for k :  57  is :  58.720930232558146 %
The accuracy of our model for k :  58  is :  58.720930232558146 %
The accuracy of our model for k :  59  is :  58.720930232558146 %
The accuracy of our model for k :  60  is :  57.55813953488372 %
```

After we choose ‘k’ as 19, we displayed the 19 Eigen faces. The screenshot of it is attached below.



Finally after finding our signature matrix we find the accuracy of our model. Also, for recognising unknown faces and predicting it as unknown(Impostor) we require a threshold value. While recognising the face we calculate the euclidean distance between the projected test face(Projecting the mean aligned face on Eigen faces) and all the vectors in the signature matrix and select the vector of the signature matrix with the least euclidean distance between it and the projected test face. However, if even this distance happens to be higher than the threshold value(tolerance), then we classify it as an unknown face.

Note that the face might be a known face and present in the dataset, however anomalies our bound to be present in all databases.

After trial and error and going through the distances it was noted that the minimum euclidean distance noted was approximately 800,000 and maximum euclidean distance for known faces was 10,000,000. However, for unknown faces, the minimum euclidean distance started from 5,000,000 and the average was 7,800,000. Also, the number of known faces having euclidean distance above 7,800,000 were found to be around 20 out of 152. So we used this average as our tolerance(threshold value) to classify image as unknown or a known face.

Finally, The accuracy of Model is :

```
The accuracy of our model for k :  19  is :  81.3953488372093 %
```