Question 4

4.1: Results on the ORL datset

PCA using SVD

In this section, we performed PCA by performing SVD of the X_{dxn} data matrix using the matlab library function.
The maximum test accuracy of 95.31% was obtained at k = 50. There is a subsequent dip in the accuracy as the dataset is really small, and eventually the model starts to overfit.

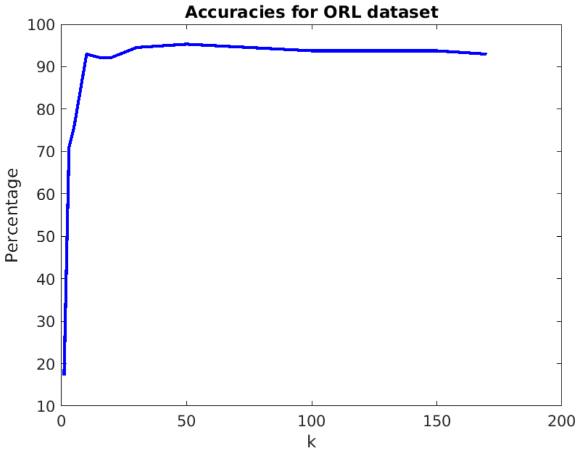

Figure 4.1: Accuracy vs k

PCA using Eigenvector decomposition

In this section, we performed PCA by performing Eigenvector decomposition of the $L = \frac{1}{n-1}X^TX$ matrix using the matlab library function *eig*, where X_{dxn} is the data matrix.
The results were exactly identical to the ones obtained in the previous part, since SVD of a matrix X simply generates the eigenvectors of the matrices X^TX and XX^T (ie the left and right eigenvectors). the The maximum test accuracy of 95.31% was obtained at k = 50.

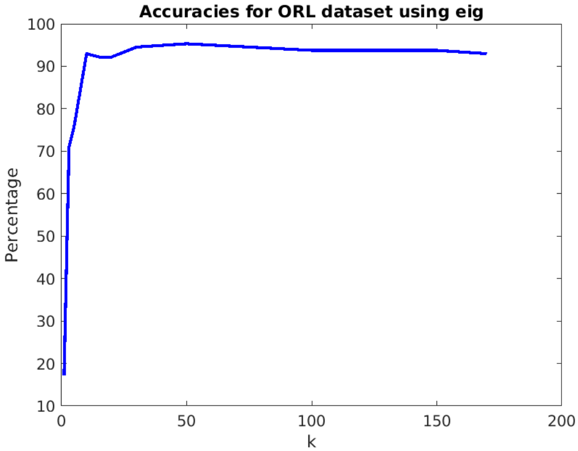

Figure 4.1: Accuracy v/s k

4.2: Results on the Yale dataset

Conventional PCA (Using all the top k eigenvectors)

Using the same method used in Part A of Section 4.1, we perform face recognition on the Yale Dataset. Since there is a huge contribution to the variance in the images arising from the lighting conditions, the eigenvectors corresponding to the highest variance are dominated by a variance in the lighting. Hence the accuracy obtained in this section is suboptimal.
The highest accuracy of 35.65% was recorded for k = 1000.

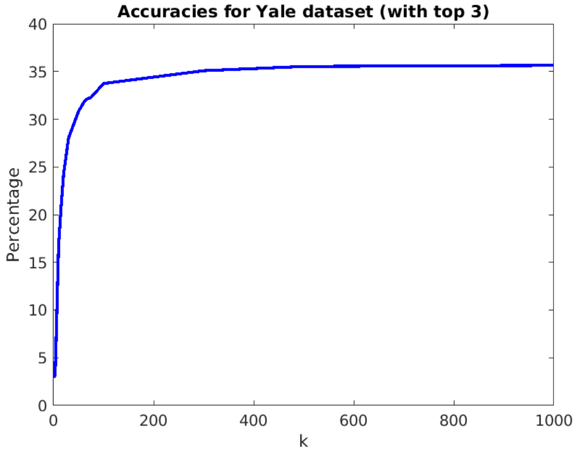

Figure 4.1: Accuracy v/s k

PCA while ignoring the top 3 eigenvectors

As stated in the previous sub section, the first few eigenvectors are dominated by the variance arising due to lighting changes in the images. Hence to counter these effects, we ignore the top 3 eigenvectors, and recompute our accuracies.
The highest accuracy of 60.67% was recorded for k = 1000.

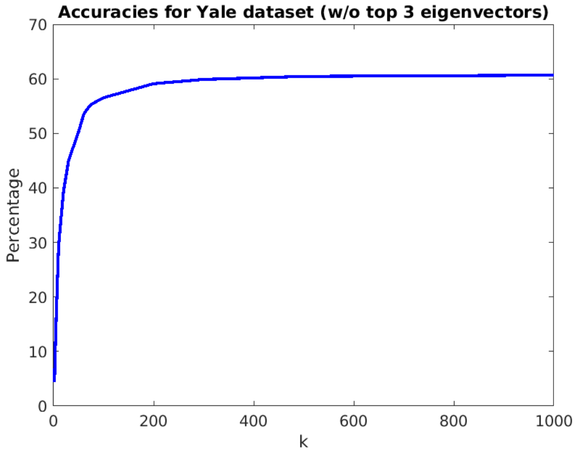

Figure 4.1: Accuracy v/s k

4.3: Usage of code

- Execute the **myMainScript.m** function to display the results and the plots. This takes around 68.22 seconds.
- Since the dataset is being used for Q4, Q5, Q6 in the assignment, we have kept the datasets in a common directory just inside our submission directory. The ORL dataset is expected to be in a relative directory **'../../datasets/'**, and the Yale dataset is expected to be in a relative directory **'../../datasets/CroppedYale/'** That is the per person directories should be as follows:
  '../../datasets/ORL/s*/' and
  '../../datasets/CroppedYale/yaleB/'**
- The **loadOrl.m** and **loadYale.m** functions load the data.
- The **fitPCA.m** and **fitPCAeig.m** functions generate the eigenvectors of the data matrix using the *svd* and *eig* functions respectively.
- The **getPredictor.m** functions returns a struct which stores the relevant eigenvectors, and the subspace transformation operator using the said eigenvectors.
- The **predict.m** function computes the test accuracy for a given data matrix, using the predictor object generated in the previous point.