SPR HW 3 - Report

Mohammad Ahmadi 9531202
Amir Hossein Ansari 9531888

Contents

1 Part A: PCA

1.1 Visualize the dataset.

Different methods are used for visualizing the dataset. They are described below:

Figure 1 is a list of all images in dataset .

Figure 1: All images in dataset

In figure 2 ,each image in the dataset is vectorized(1*4096) and then put together to form a whole matrix(630*4096) then the matrix is plotted using imshow as an image. You can see the change of value of pixels in gray format and distribution of them.

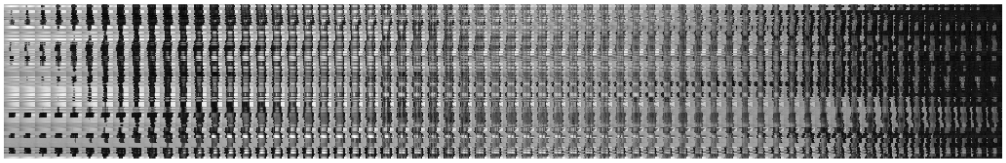

Figure 2: Matrix of the whole dataset

Figure 3 shows histogram of value of pixels of each image in gray format.

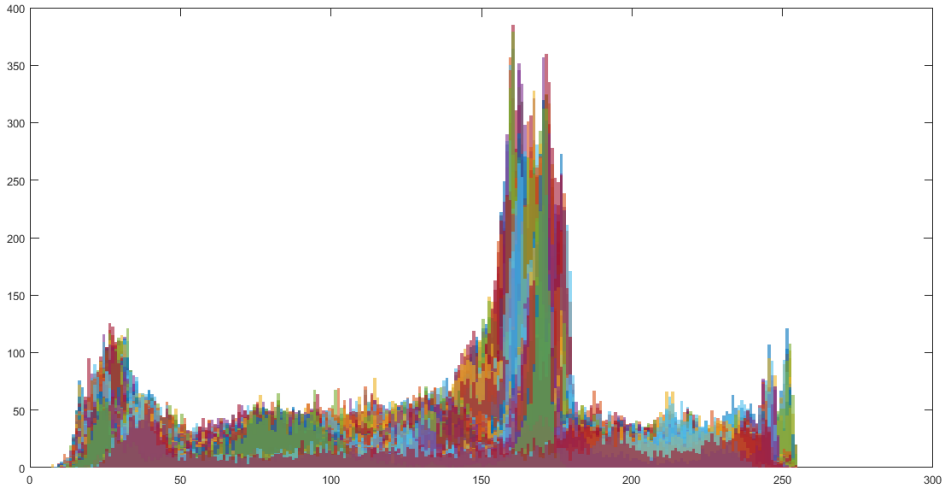

Figure 3: Histogram of each image

Figure 4 is mean of each pixel(total num of pixels is 4096) for all 630 images.

Figure 5 is obtained by averaging each pixel of dataset images.

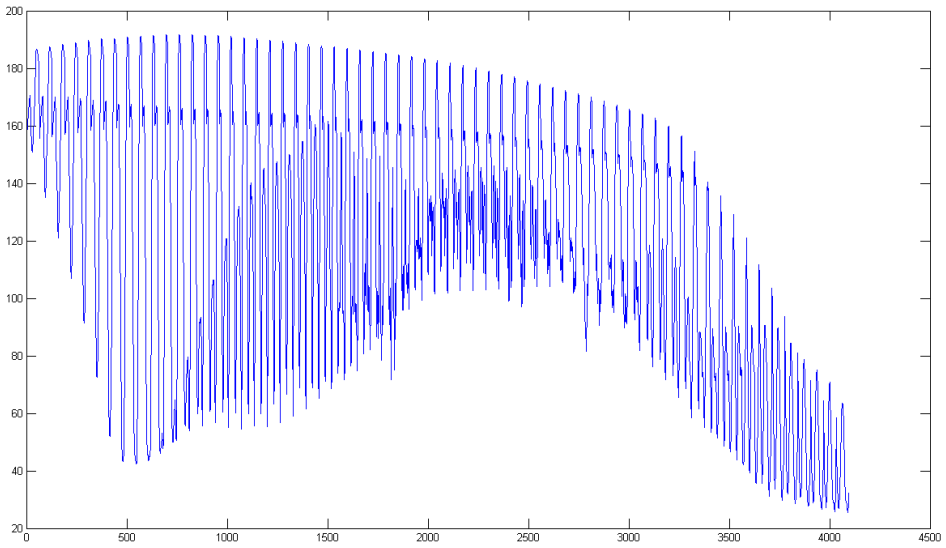

Figure 4: Mean of each pixel in all images

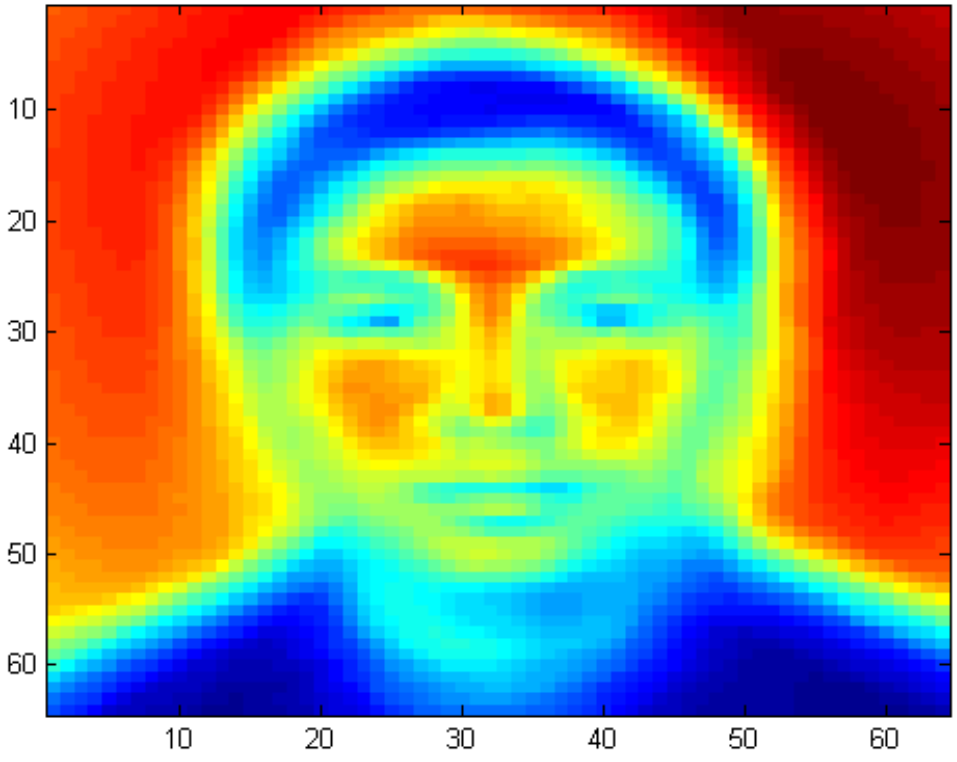

Figure 5: image of averaging(mean) of whole dataset average

1.2 Preprocess and normalize the dataset.

No report.

1.3 Implement the PCA function, then apply it on the dataset.

No report.

1.4 Visualize the reduced dataset using 2D and 3D plots.

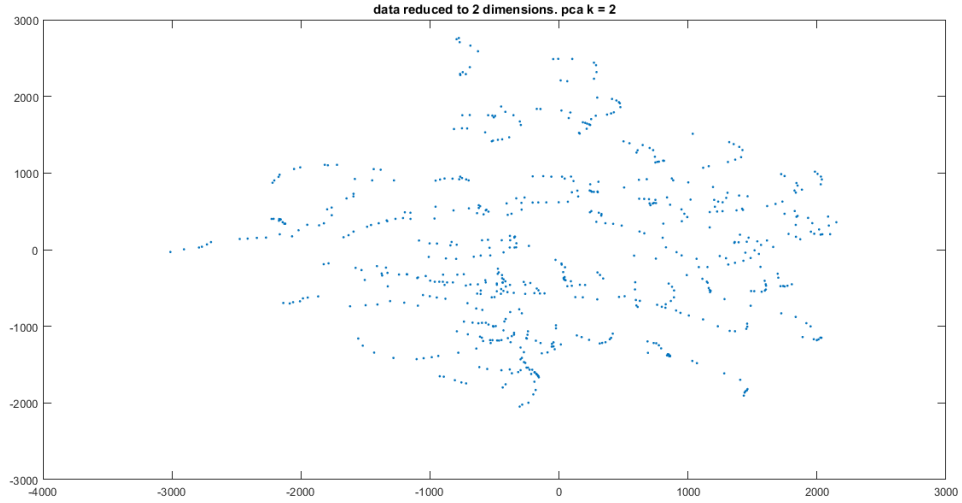

Figure 6: data reduced to 2 dimensions using pca

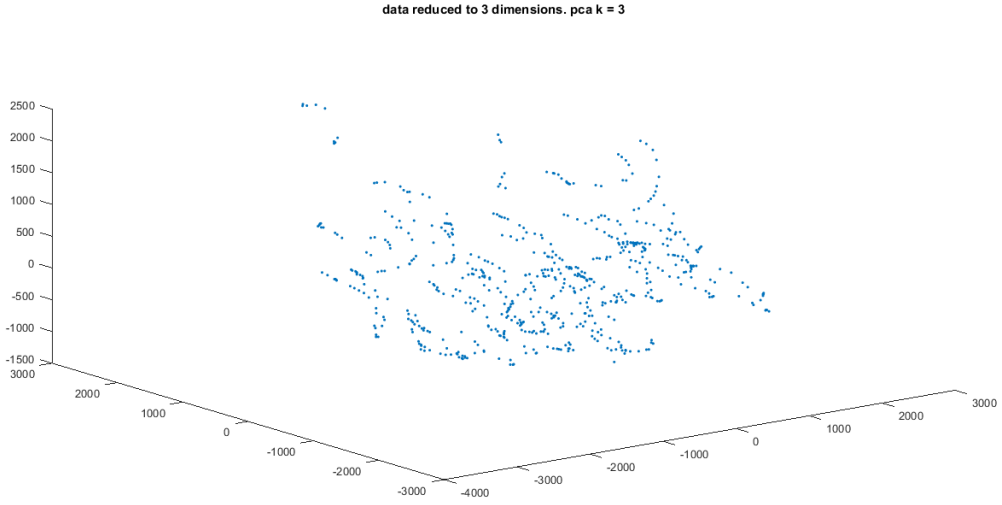

Figure 7: data reduced to 3 dimensions using pca

1.5 Reconstruct the original data by using K principle components (Show reconstructed images of each individual for K=1,30,120).

Figure 8: PCA with k=1

Figure 9: PCA with k=30

Figure 10: PCA with k=120

1.6 Plot the MSE between the original and reconstructed images in terms of number of eigenvectors.

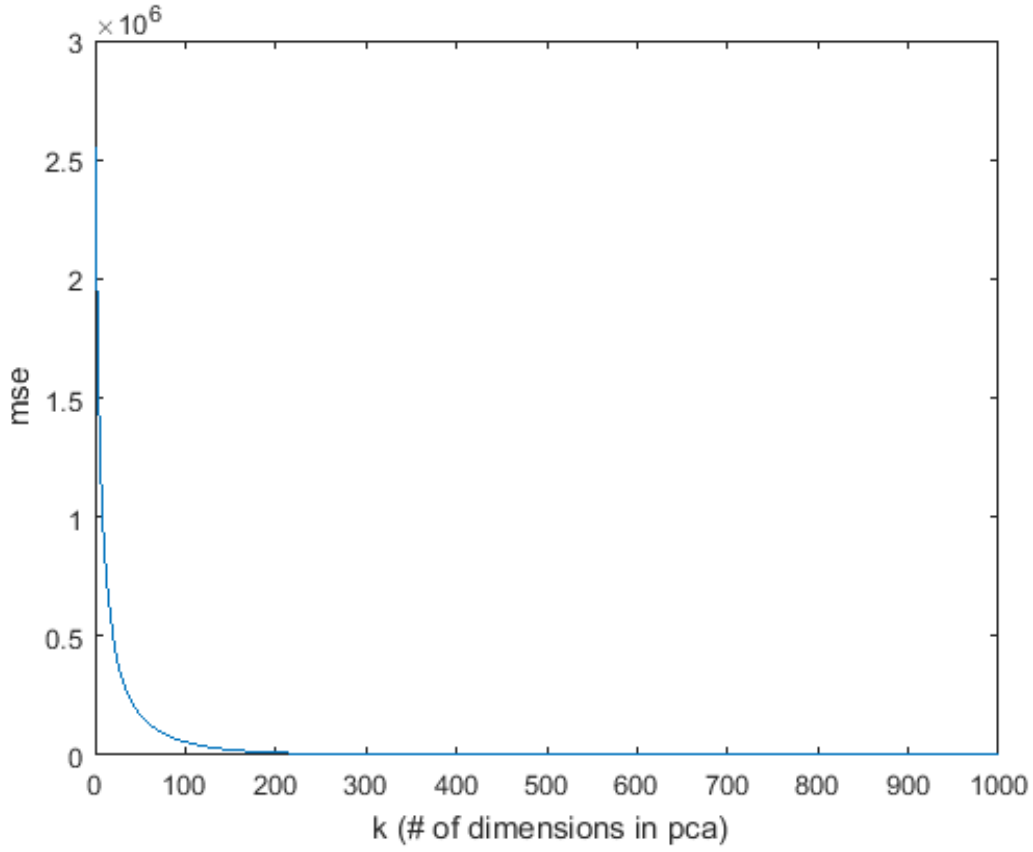

Figure 11: mse for different k s in pca

1.7 Visualize some of the first principal components.

Figure 12: 1st eigen vector

Figure 13: 2nd eigen vector

Figure 14: 3rd eigen vector

Figure 15: 4th eigen vector

Figure 16: 5th eigen vector

1.8 How many principle components are enough so that you have acceptable reconstruction? How do you select them?

According to mse figure (figure 11) about 220 is a good k for pca. Because as you see mse is almost zero.

2 Part B: Fisher LDA

2.1 Implement and apply the Fisher LDA function for multi-class problem.

No report.

2.2 What is the problem of applying Fisher LDA on the dataset?

This warning was noticed during running the code: Warning: Matrix is close to singular or badly scaled. Results may be inaccurate. RCOND = 3.373158e-23.

This means that because Sw is not invertible we can not use fisher LDA method.

As was mentioned in the paper decribing DLDA there are two main problems with Fisher's method.First, it is computationally challenging to handle big matrices (such as computing eigenvalues). Second, those matrices are almost always singular, as the number of training images needs to be at least 16M for them to be non-degenerate.

2.3 Propose a new solution to solve the problem and use your proposed method instead of Fisher LDA (You can use Direct LDA method to solve your problem).

No report.

2.4 Reconstruct the original data by using K basis vectors obtained from LDA. (Show reconstructed images of one person for k=1, 6, 29).

Figure 17: original image

Figure 18: reconstructed image using DLDA with k=1

Figure 19: reconstructed image using DLDA with k=6

Figure 20: reconstructed image using DLDA with k=29

2.5 What would happen if we have a large amount of outliers in the dataset?

The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. For this LDA tries to maximize between class scatter and minimize within class scatter.If we have a large amount of outliers then mean of each class is affected and also within class scatter and between class scatter. As these values are affected by outliers then DLDA can not perform well.

2.6 Plot the MSE between the original and reconstructed images in terms of number of eigenvectors.

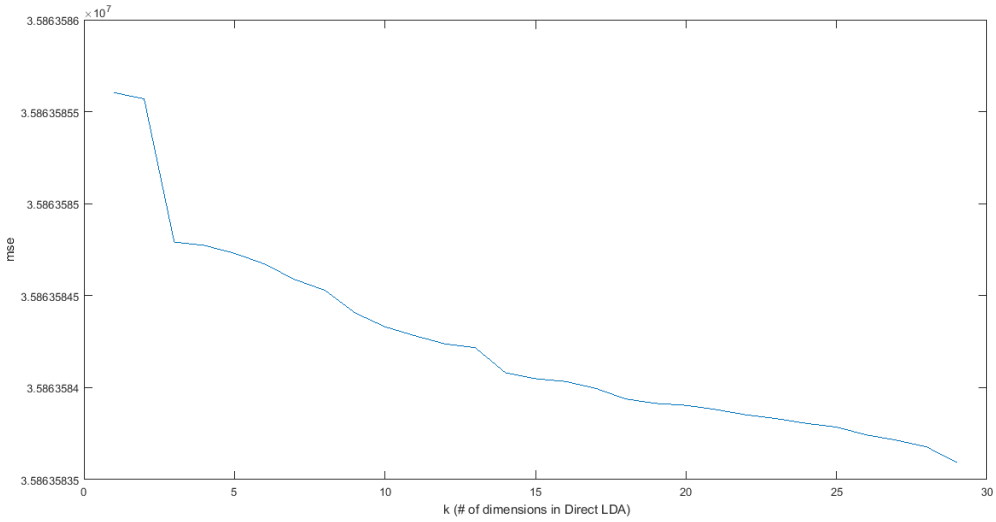

Figure 21: mse of applying DLDA based on different values of k (number of eigen vectors)

3 Part C: Classification

3.1 Design a KNN classifier.

No report.

3.2 Classify the projected data of parts A and B (the data projected by PCA and LDA) by using the designed classifier with K=1, i.e., 1NN.

For the next part 1nn classifier is runned on data with different values of k for bth PCA and DLDA.

3.3 What is the minimum number of basis vectors you need to achieve 100 accuracy on the dataset?

For PCA with k = 17 accuracy of 100 percent was achieved. For DLDA with k = 12 accuracy of 100 percent was achieved.

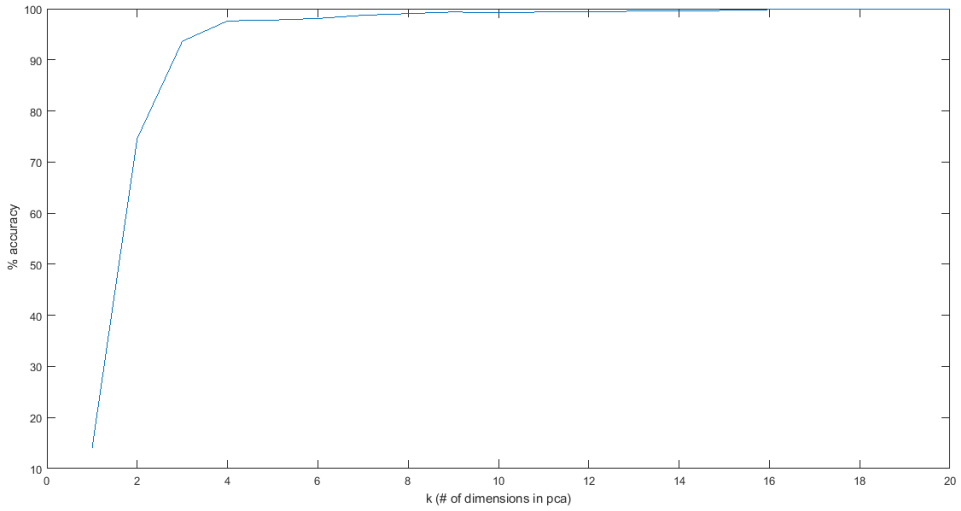

Figure 22: percent of accuracy of 1nn classifier for different values of k for PCA

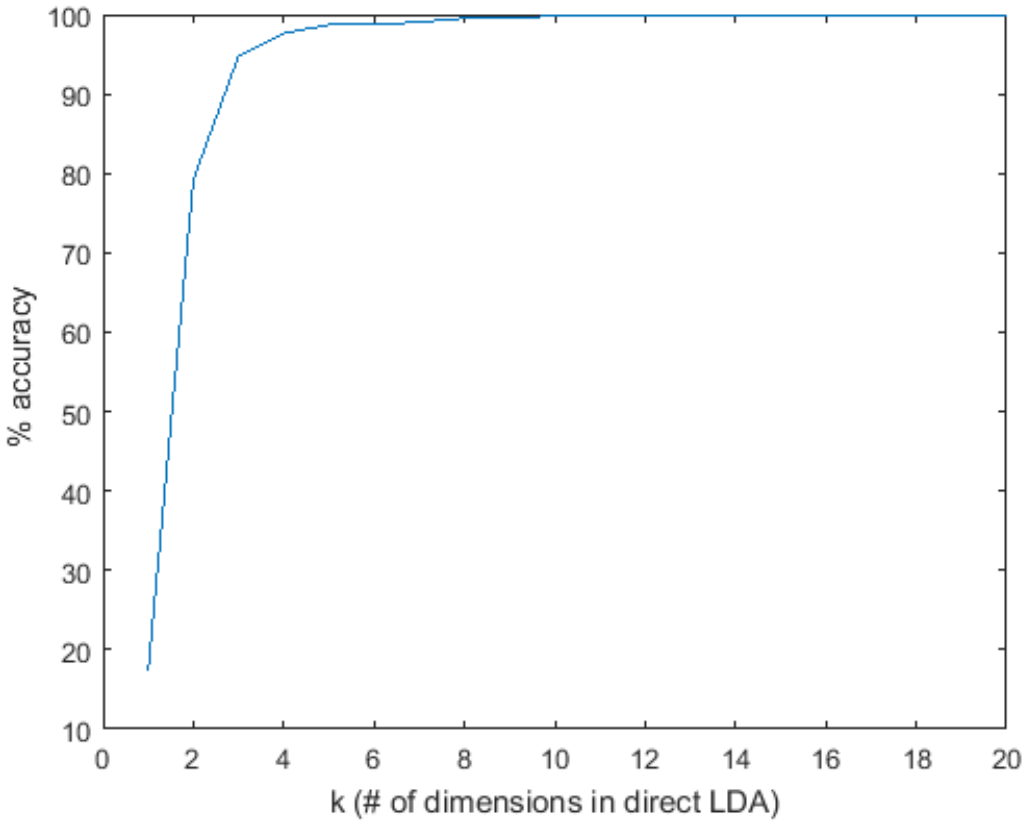

Figure 23: percent of accuracy of 1nn classifier for different values of k for DLDA

3.4 Plot the dataset and the models decision boundary.

Because only 6 colors could be used some classes are assigned repeated colors(only neighbour classes have different colors).

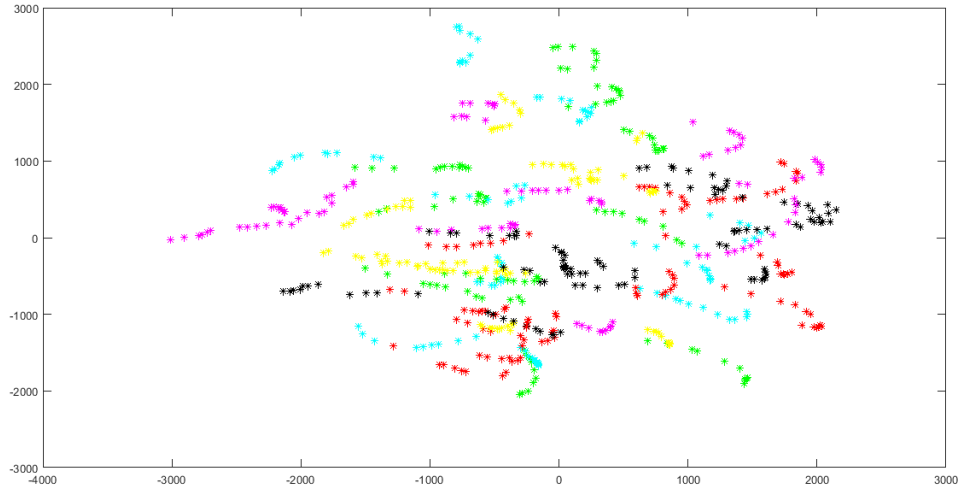

Figure 24: dataset after PCA reduction applied with k = 2. Different classes are with different colors.

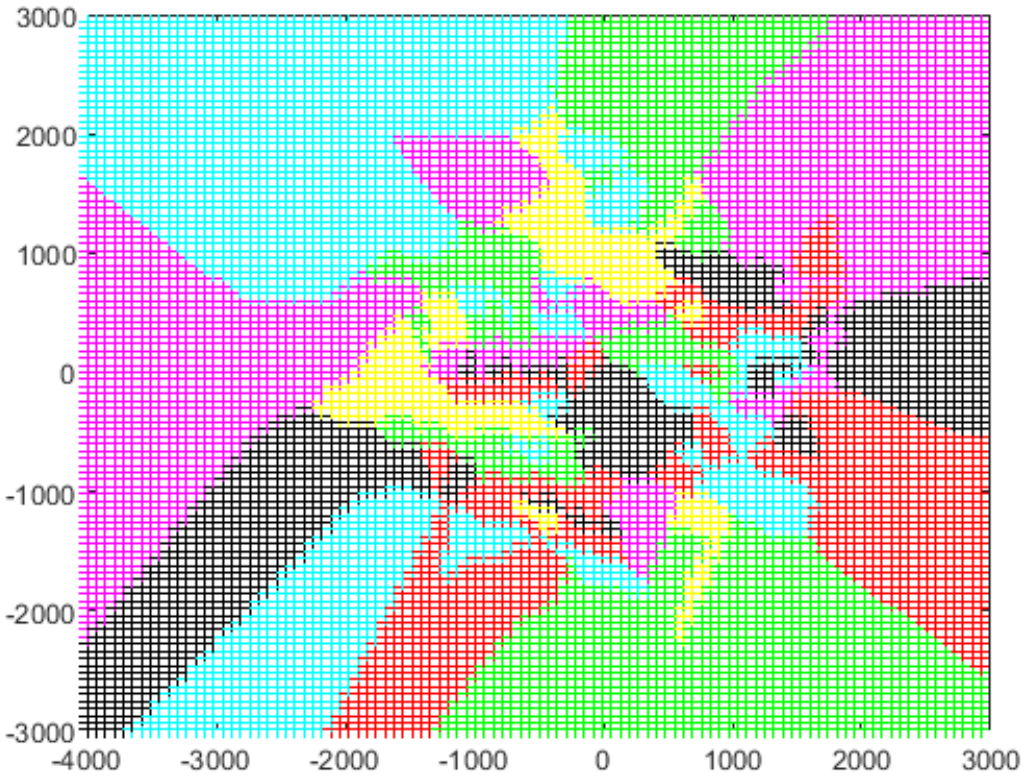

Figure 25: Decision boundary for 1nn classifier after PCA with k=2 is applied on data

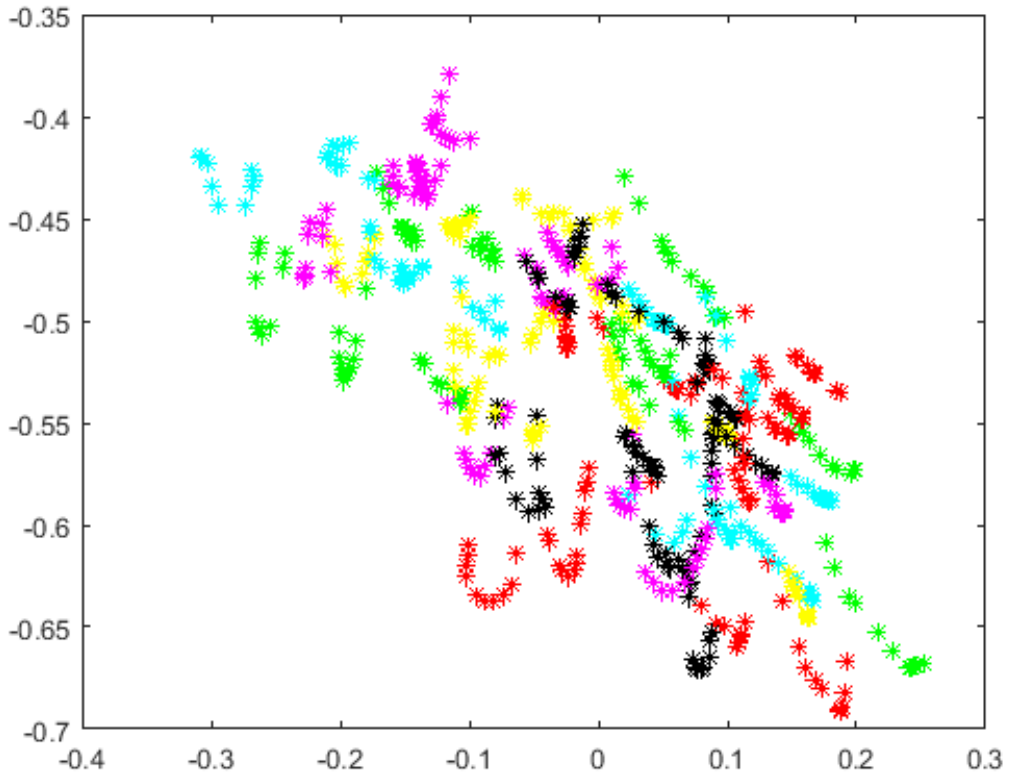

Figure 26: dataset after DLDA reduction applied with k = 2. Different classes are with different colors.

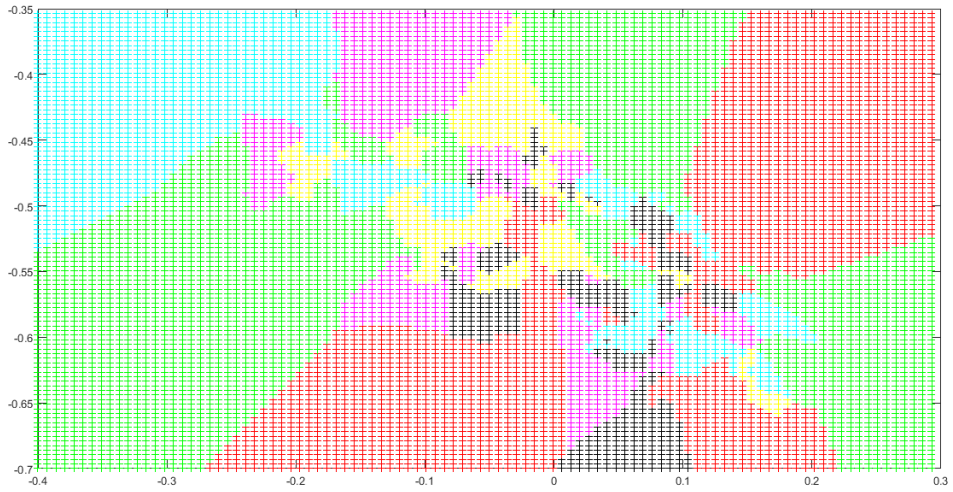

Figure 27: Decision boundary for 1nn classifier after DLDA with k=2 is applied on data

4 Part D: K-Means

4.1 Implement K-means algorithm.

No report.

4.2 How to select initial center points in K-Means?

The best way is to randomly select initial centers among dataset points. Because if we a point that is far from the dataset elements is chosen randomly then their value woudldn't be updated(because if we update that center the total number of that class is 0 so the result is 0/0 NAN).

4.3 Apply K-means to image compression.

No report.

4.4 Plot your original image alongside to the reconstructed one.

Figure 28: Original image

Figure 29: Compressed image to 16 colors. Using kMeans with 16 centroids

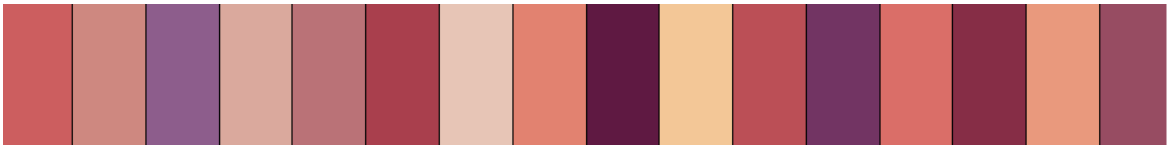

Figure 30: 16 colors achieved using kmeans classifier

4.5 Choose the best K and explain why you choose that specific K?

In order to find best k for kmeans we have plotted mse between reconstructed image and original image for different values. According to this figure for k about 110 algorithm has a good value of mse and seems to have converged.

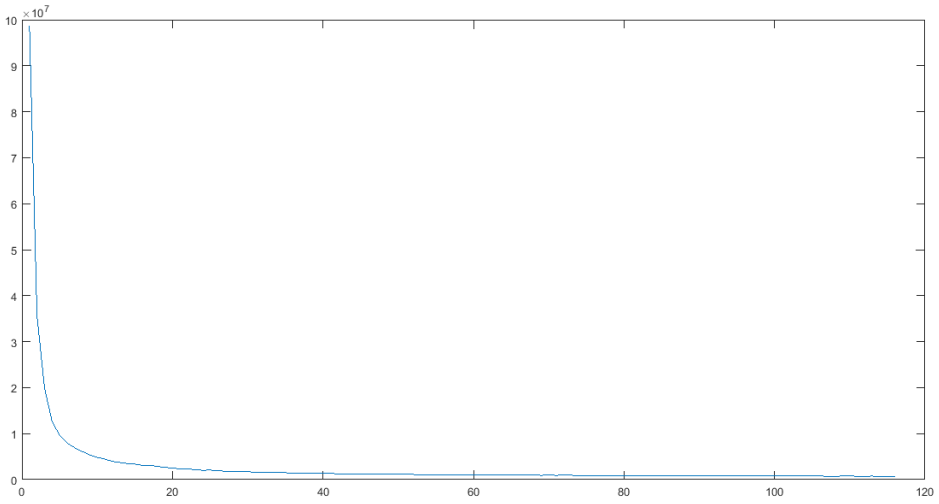

Figure 31: MSE of applying kmeans to compresss an image and originnal image for different ks

5 Part E: Extra

5.1 Use your K-means algorithm and apply it on one of your images (any image you like).

5.1.1 First image

Figure 32: Original image

Figure 33: Compressed image to 16 colors. Using kMeans with 16 centroids

Figure 34: 16 colors achieved using kmeans classifier

5.1.2 Second image

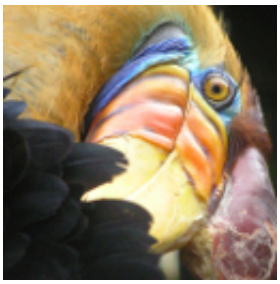

Figure 35: Original image

Figure 36: Compressed image to 16 colors. Using kMeans with 16 centroids

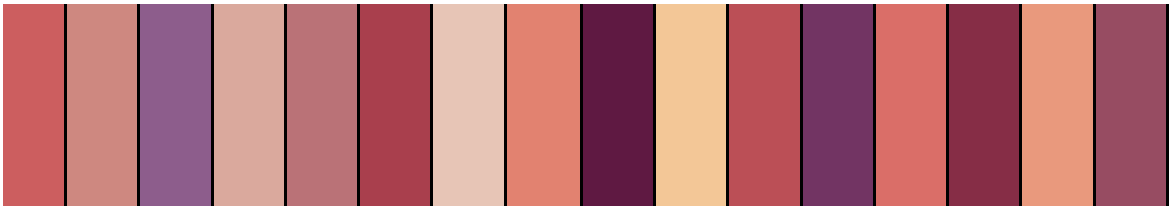

Figure 37: 16 colors achieved using kmeans classifier