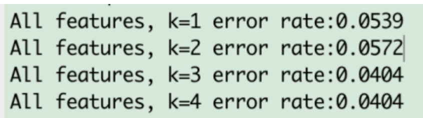
CSCI 5521 Homework 2
Final Report
Chunni Zhao

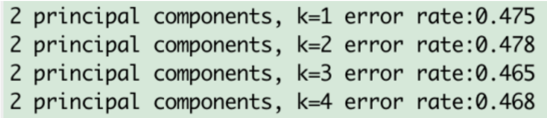
1. (a) When all the original dimensions are used, the error rate obtained with k=1,2,3,4 using KNN is:



```
All features, k=1 error rate:0.0539
All features, k=2 error rate:0.0572
All features, k=3 error rate:0.0404
All features, k=4 error rate:0.0404
```

As k increasing, the error rate is decreasing. And there is no more decrease in error rate when k = 4.

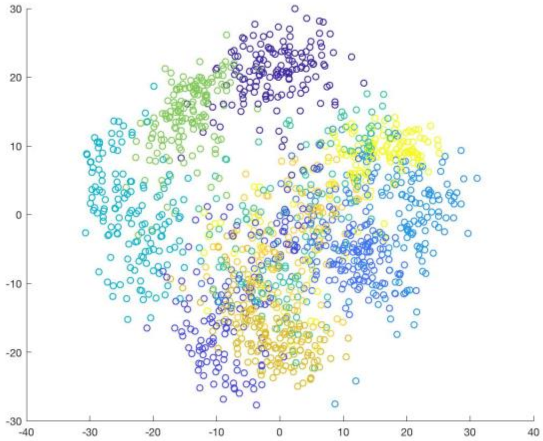
(b) The error rate using KNN with the first two principal components is



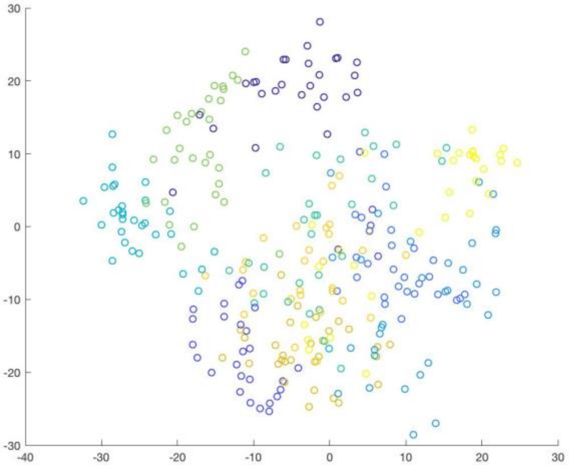
```
2 principal components, k=1 error rate:0.475
2 principal components, k=2 error rate:0.478
2 principal components, k=3 error rate:0.465
2 principal components, k=4 error rate:0.468
```

The results are better than using all the original dimensions. It drops some components that are not that much important, by using only the first two principal components, error rate decreases significantly.

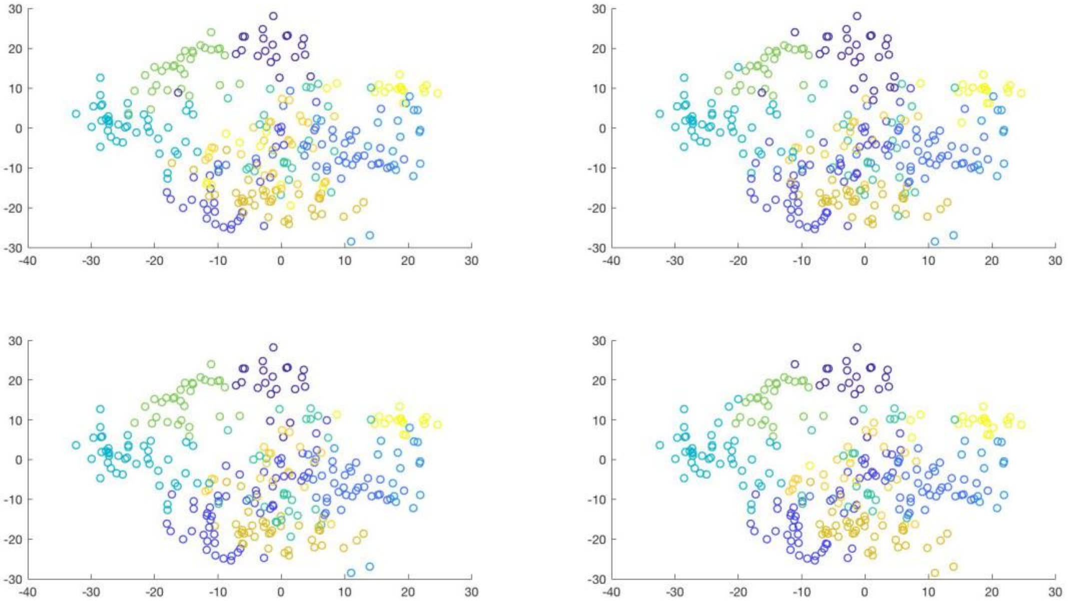
Using the first two principal components, the scatter plot for all the samples in the training set on the projected space is like this.



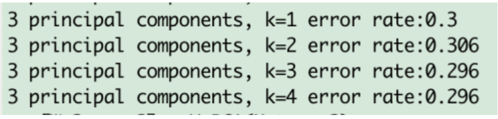
Using the first two principal components, the scatter plot for all the samples in the test set on the projected space is like this.



Classify the test set (in the projected dimension) using k nearest neighbors, (k = 1, 2, 3, 4)



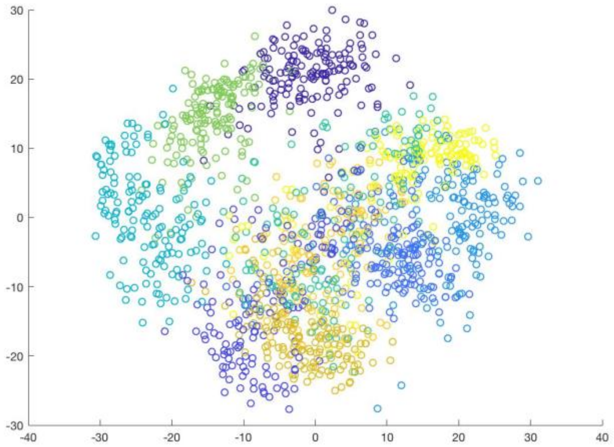
(c) The error rate using KNN with the first two principal components is



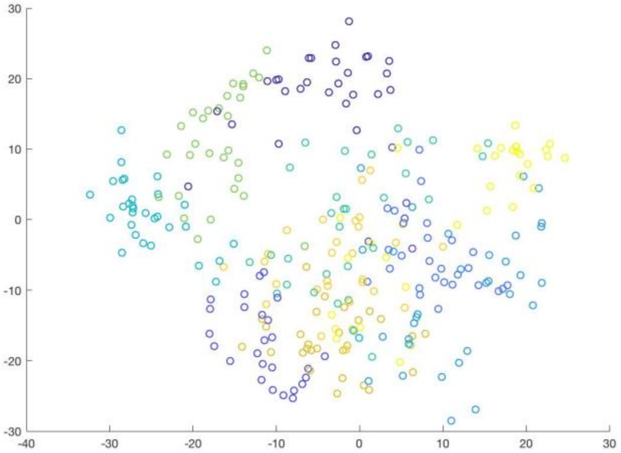
```
3 principal components, k=1 error rate:0.3
3 principal components, k=2 error rate:0.306
3 principal components, k=3 error rate:0.296
3 principal components, k=4 error rate:0.296
```

The results are much better than that using only two principal components. This time, we use one more component to represent the data, there are more information left in our projected training data, which could explain test data much better than the one with only first two principal components, so, the error rate is smaller than that with two principals.

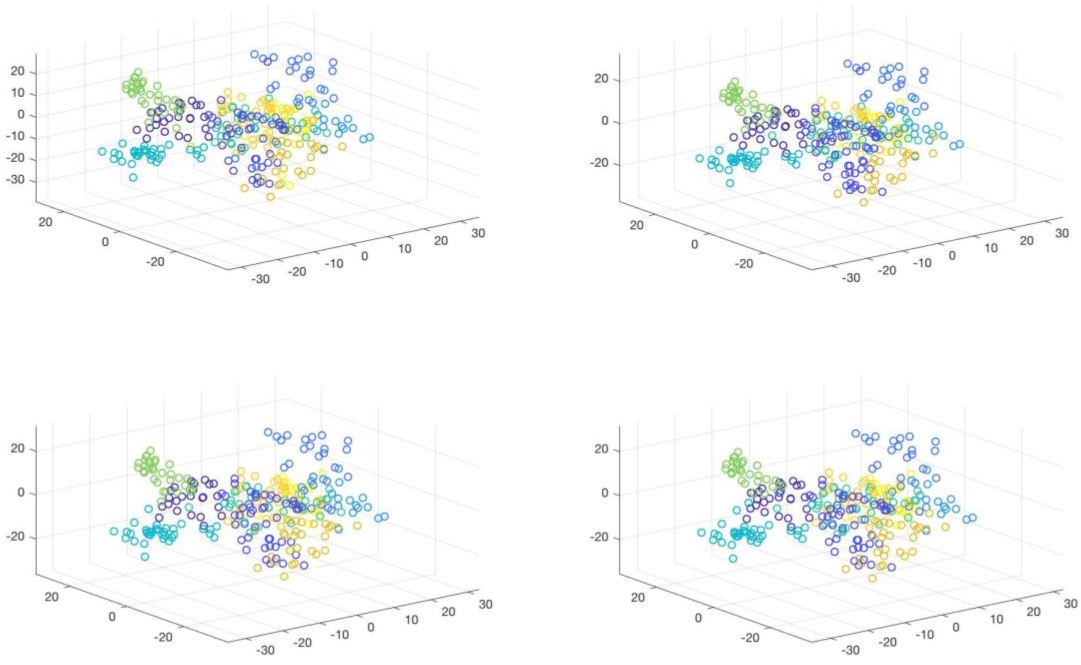
Using the first three principal components, the scatter plot for all the samples in the training set on the projected space is like this.



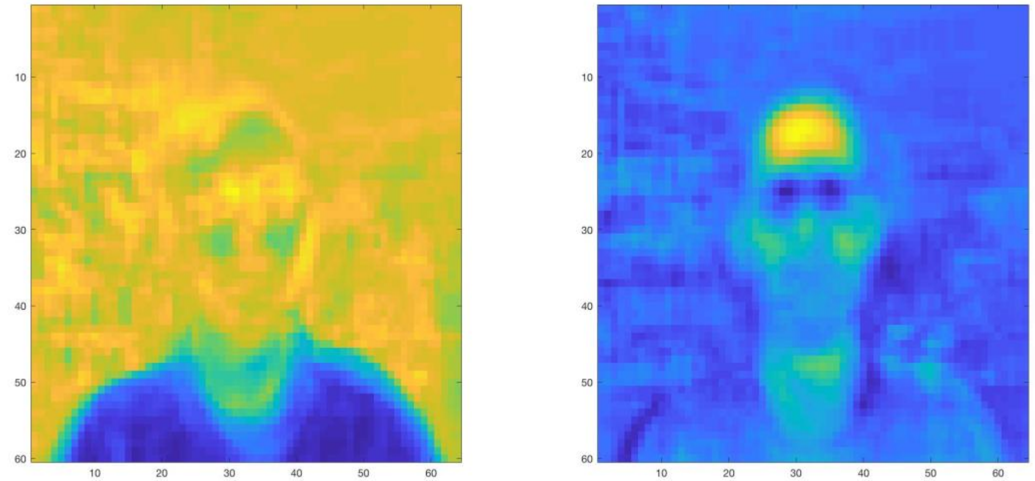
Using the first three principal components, the scatter plot for all the samples in the test set on the projected space is like this.



Classify the test set (in the projected dimension) using k nearest neighbors, (k = 1, 2, 3, 4)

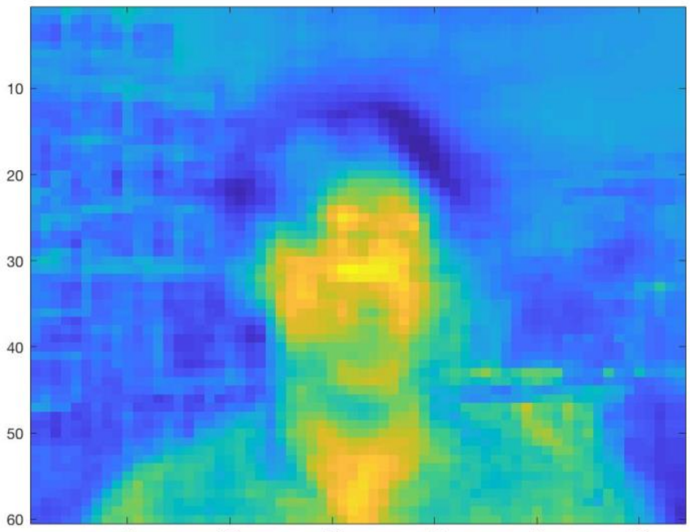


(d) The first two principal components are displayed below, they presents the first two significant components of this image, and a large part of the information of the original image.

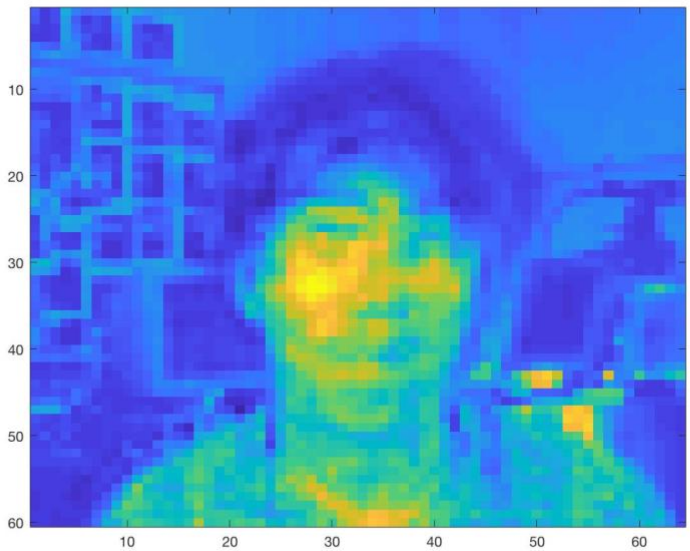


(e) The reconstructed image using first k principal components are displayed below, as the number of principal components used become increases, the reconstructed image is becoming more and more clear and more similar to the original imagre.

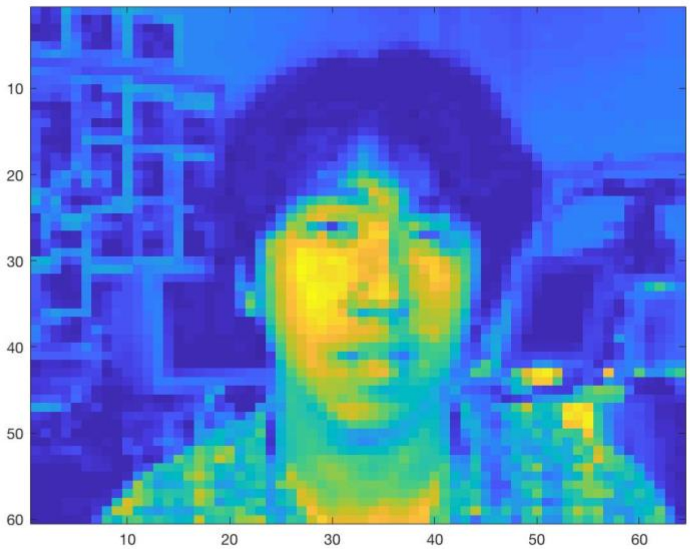
$k = 10$



$k = 50$

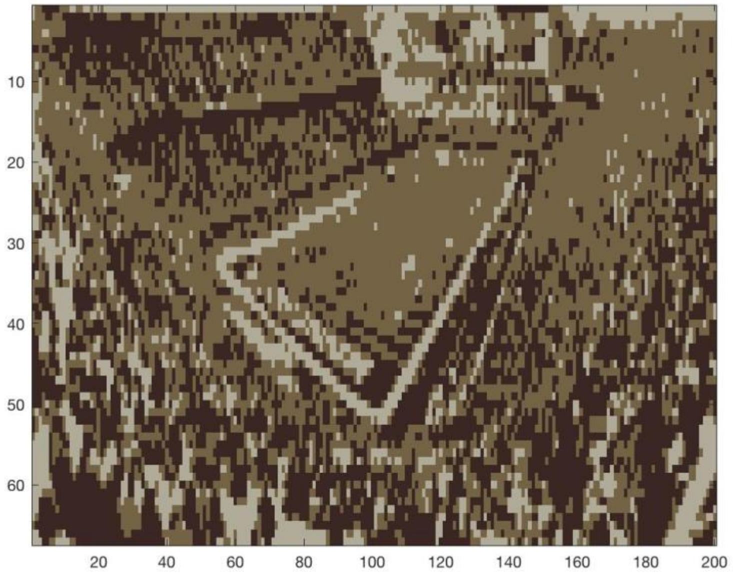


$k = 100$

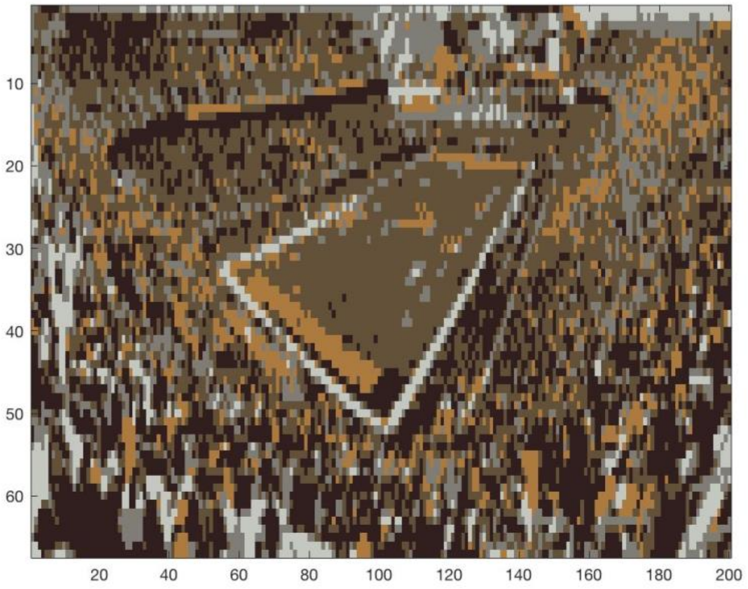


2. (a) The resulting images using **K-Means**(k = 3,5,7) are displayed below, as the k increasing, the compressed image will become more similar to the original image.

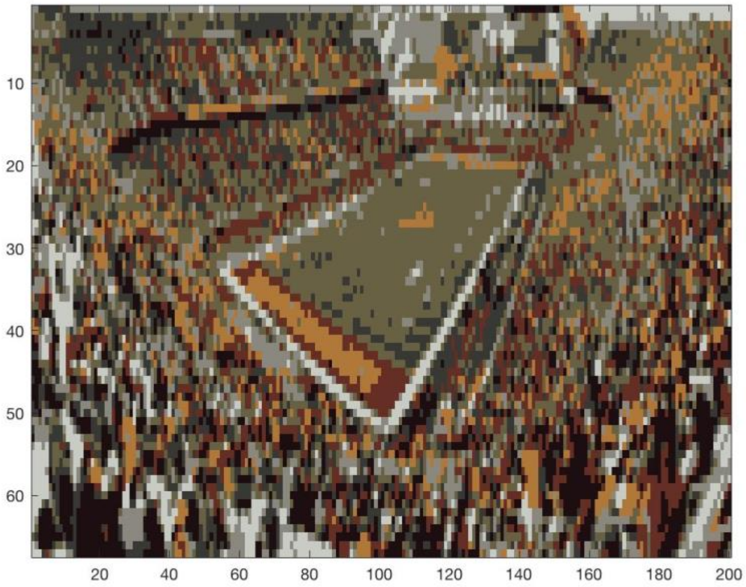
$k = 3$



$k = 5$

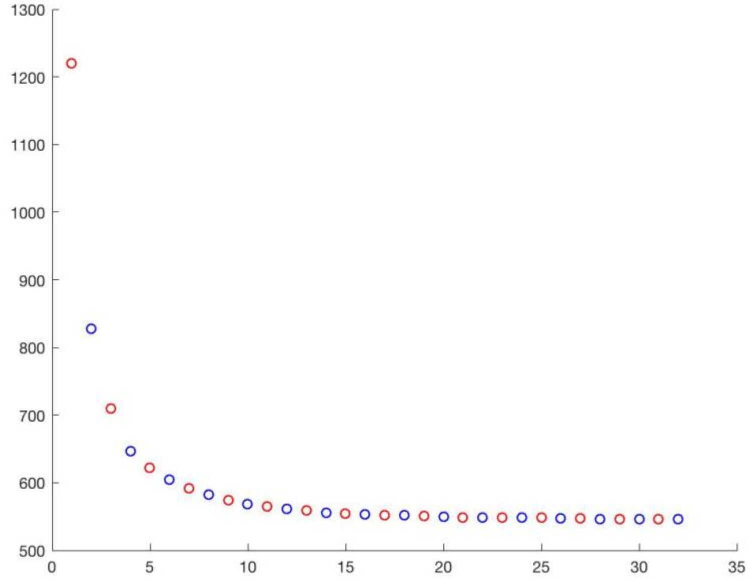


$k = 7$

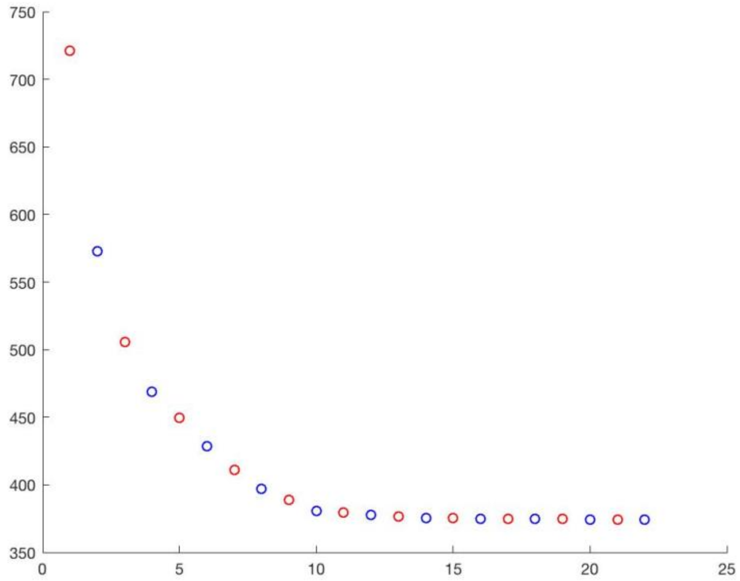


(b) The reconstruction error after every updating is displayed below, the plot shape is just like what I expected, the reconstruction error will decrease as the updating, because after every update, the mean of every cluster will be more accurate and more close to the true center of each cluster.

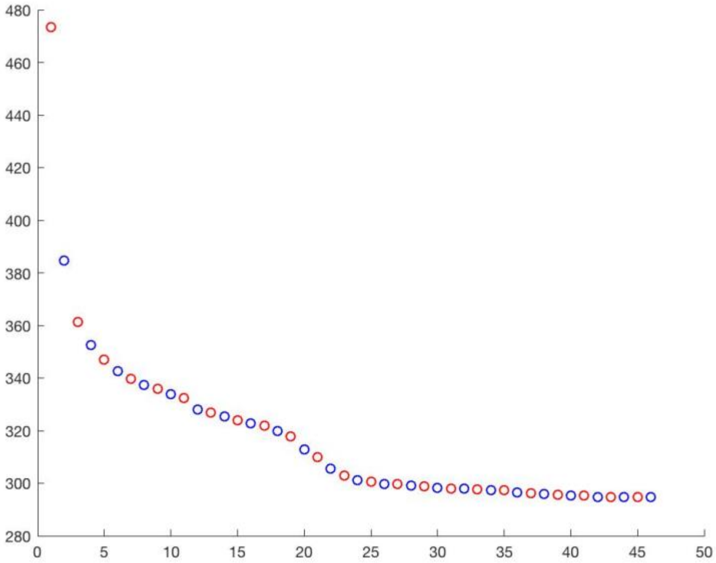
$k = 3$


1300
1200
1100
1000
900
800
700
600
500
0
5
10
15
20
25
30
35

$k = 5$


750
700
650
600
550
500
450
400
350
0
5
10
15
20
25

$k = 7$


480
460
440
420
400
380
360
340
320
300
280
0
5
10
15
20
25
30
35
40
45
50