EEL4930/EEL5840 Fall 2016 – Project 2
Principal Component Analysis and K-Means Clustering – Biometrics

November 11, 2016

Due: November 28, 2016, 11:59 PM

This project will be graded with a letter grade with respect to presentation (25%), methods (35%) and results (40%).

You will find three data set files for this assignment *ProbeSet.rar*, *GallerySet.rar*, and *Gender.txt*. *ProbeSet.rar* contains facial images collected from 100 individuals (two images per person). *GallerySet.rar* contains 100 images collected from the same individuals (one image per person). *Gender.txt* contains information on the gender of each of the individuals in the data set. The purpose of this project is to implement and evaluate the performance of Principal Component Analysis and the K-means algorithms as they apply to the problems of biometric recognition and soft biometric (gender) classification.

The project requires a report explaining the experimental procedures you followed and you must include data to support your conclusions. Please use the format of an IEEE Transactions paper (limited to 7 double column pages). You can download the format from the IEEE website. This means you have to write a brief intro to the theory, explain well the methods and present carefully the results (see below) and conclusions. Remember that any scientific paper should, by definition, contain sufficient information such others can replicate your results. A scientific paper must also contain ORIGINAL material only. If you happen to use text or equations from other source you have to reference what you cut and paste (this is not allowed in a normal publication, but here it is OK provide you reference). Of course, I expect the results to be done by the student alone. I would like to see in the report (at least) the following:

1. Using the gallery set to compute the PCA projection of the data, display the first three principal components as face images. We know that the first three components represent the directions of highest variance in the data but what does the largest principal component represent in terms of facial recognition?

2. Using the eigen-coefficients, Euclidean distance as the distance measure, and varying the number of principal components from 10 to 100 in steps of 10, plot the recognition rate of as a function of the number of principal components used. Discuss the observed trend in recognition performance. Would you expect this trend to continue as the number of principal components increases? Explain your reasoning.

3. Using Euclidean distance as the distance measure and the original images as feature vectors, determine recognition performance. How does the performance obtained compare with that obtained using the PCA projection of the data?

When performing large-scale facial recognition, a way to improve performance is to reduce the search space for matches by first performing soft biometric classification (gender, ethnicity, age).

1. Using the eigen-coefficients, Euclidean distance as the distance measure, and varying the number of principal components from 10 to 100 in steps of 10, perform K-means clustering where $K = 2$. Each cluster should be composed of images from individuals of a single gender.

2. Evaluate the clustering in 1 using an internal criteria and an external criteria. Plot cluster validity as a function of the number principal components. Are the observed trends comparable to those observed in terms of recognition? Would you expect these trends to continue as the number of principal components increases? Explain your reasoning.