Comp Methods Project I

Joe Sinotte

October 20, 2014

Contents

1 Problem Statement

I have been given two sets of data, Ytrain and Ytest. Each data set contains 200 images of 40 individuals, with 5 images of each individual.

Problem: With the given data, is it possible to code a matlab program to match images from Ytrain to an individual in Ytest?

2 Methodology

2.1 Dimension Reduction

Since image data are usually very large, it can be beneficial to reduce the dimensionality of the data to make the program less computationally expensive. There are various methods to reduce dimensionality. This paper shall use two of these methods, which will be described in greater below.

The first approach to dimension reduction that will be discussed is Principal Component Analysis (PCA). The steps of PCA are as follows:

1. Find the sample covariance matrix, $C \in \mathbb{R}^{n \times n}$, of the data.
2. Compute the singular value decomposition of C, which yields the orthogonal matrix $UPCA \in \mathbb{R}^{n \times n}$.
3. Select the first k columns of U and denote them by the matrix $UPCA$.
4. Define a transformation of the data as $Z = (UPCA)'^{*}Ytrain$. Where, $Z \in \mathbb{R}^{m \times k}$

In addition, a feature vector is formed by applying the same transformation to a randomly chosen column vector from Ytest, denoted as I. Hence,

$$Z1 = (UPCA)'^{*}I$$

This method provides the benefit of only retaining the k dimensions with the most explanatory power, and all of the superfluous dimensions are omitted from the rest of the process, decreasing the computational expense of the entire procedure.

The second method that will be examined is the Simple Projection (SP) method. In this method, a transformation is derived by selecting the first k columns of an identity matrix which, in this example, is of dimension $(s_1 \times s_2) = 644$, and defining the newly acquired matrix as USP. Thus, the transformation can be expressed as:

$$Y1 = (USP)'^{*}Ytrain$$

With the feature vector,

$$I1 = (UPCA)'^{*}I$$

The SP method carries benefits and costs of its use. This method is simple and fast, making it a computationally cheap way to reduce data dimensionality. On the other hand, this method, in contrast to the PCA method, does not allow the user to only include the most descriptive dimensions. This means that in the process of reducing dimensionality, the transformation may omit the most informative portions of the data.

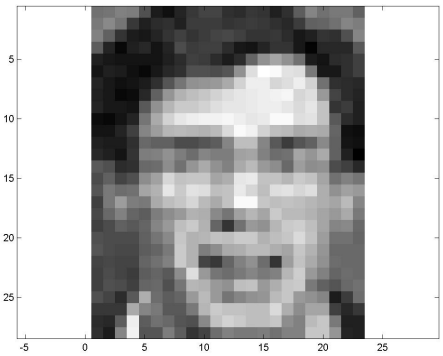
2.2 Image Recognition

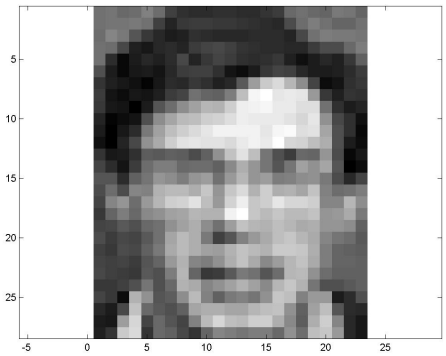
Having successfully reduced the dimensionality of the images under scrutiny, I will now turn my attention to the method used to match a randomly selected column from I1 and Z1 to columns of Y1 and Z, respectively. Here, I use the Nearest Neighbor Classifier to find the column of the Y1 that minimizes the normed distance between Y1 and I1, and likewise the column of Z that minimizes the normed distance between Z and Z1. The column that minimizes this distance in each case is called the nearest neighbor, and if that column corresponds to an image of the same individual

The process described above was repeated for k equals one to forty dimensions. For each k, one hundred loops are run, and the proportion on correct identifications is calculated and defined as F(K).

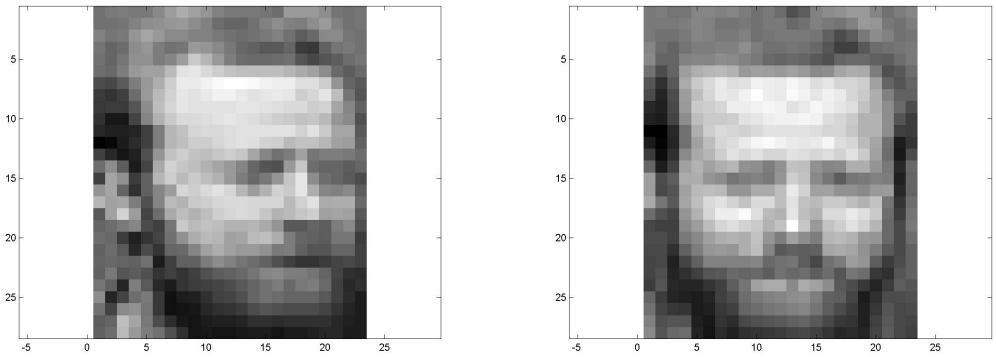
3 Results

Here, I will give a brief overview of the results of this exercise. Overall, the program was a success. Below, I have included examples of successful matches, both with SP and PCA.





This is an example of a match with a simple projection.



This is an example of a match with PCA.

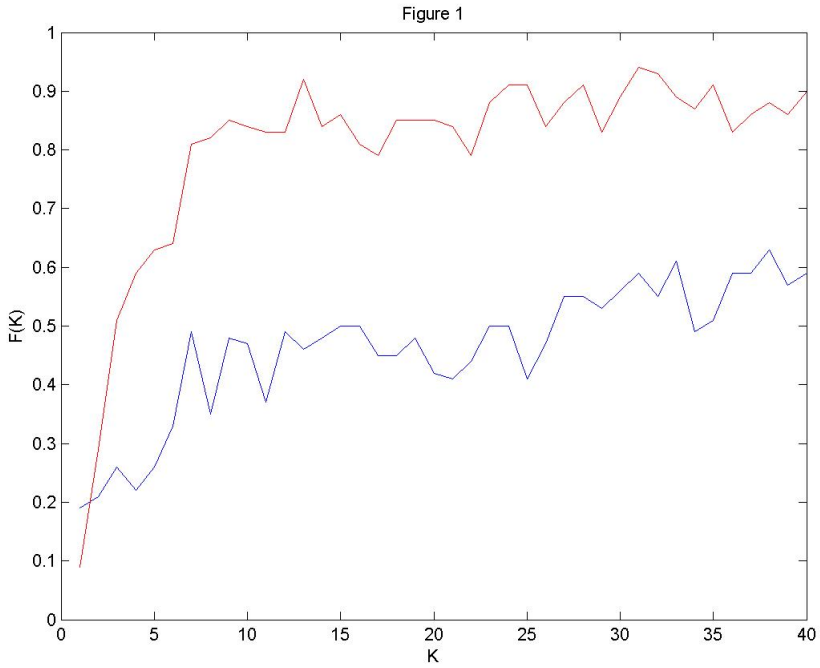


Figure 1 displays the results of the exercise under discussion. The horizontal axis denotes the number of dimensions to which the data sets are being reduced, while the vertical axis displays the proportion of correct identifications for each value of k. As the figure shows, the simple projection method is more accurate for very small dimensional problems. As the number of dimensions increases, PCA becomes significantly more efficient, and consistently yield high match proportions. Thus, it can be concluded that PCA gives the most accurate results for higher dimensional problems.

It is also worth noting that the variation in F(K) is quite high with smaller K values.

However, as we allow K to increase, F(K) becomes much more stable for both the simple projection and principal component methods.

In conclusion, I find the exercise a success, with a high proportion of matches under PCA and a lower proportion under SP.

4 Matlab Programs

All programs are attached on a separate sheet.