EEL5840: Elements of Machine Intelligence

Project One: Principal Component Analysis and K-Means Clustering

Biometrics

Jason Gutel
University of Florida
Gainesville, FL
gutel@hcs.ufl.edu

***Abstract*—This work focuses on using eigenface analysis with the Principal Component Analysis (PCA) method of feature reduction in order to achieve facial recognition with the k-Nearest Neighbors (kNN) algorithm. In addition to this the k-means clustering algorithm is used on the eigenfaces in order to learn a grouping for gender identification. The methods and results of both of these works are documented within.**

***Index Terms*—PCA, eigenface, knn, k-means clustering, biometrics, gender identification, facial recognition, Machine Learning, Machine Intelligence**

I. Introduction

In 1966 Dr. Seymour A. Papert was a professor in the AI Lab at the Massachusetts Institute of Technology. At this time Dr. Papert released a paper called *The Summer Vision Project* which proposed creating a computing system complex enough to perform pattern recognition using an attached camera. The paper was bold enough to assume they could build a system capable of recognizing common objects such as balls, bricks, and cylinders of uniform color/texture with a homogeneous background over a single summer[1].

Throughout the 70's and 80's advancements were made in machine learning computer vision recognition software. Many of the filters used today for images such as the Gaussian blur, Sobel filter, and more were developed during this time and are useful for feature extraction.

In 1991 MIT made a breakthrough with facial recognition with a method referred to as *eigenfaces* outlined in the seminal paper by Turk and Pentland [2]. The method uses a linear combination of *eigenvectors* to generate unique facial images by providing a weighted sum of vectors. The sum is then compared to already computed images of known people to see if a distinction can be made.

This work will focus greatly on the work done by Turk and Pentland. The paper will outline the methodology of creating and using eigenfaces for facial recognition as well as the methodology for using the k-means clustering algorithm to perform gender identification of images. Results will then be shown to this effect.

A. Principal Component Analysis and Generating Eigenfaces

An eigenvector is simply described as a vector whose direction stays consistent when a linear transformation is applied to it. Given a matrix, **A,** an eigenvector can be computed as:

$$\boldsymbol{Av} = \lambda \boldsymbol{v}$$

Where **v** is the eigenvector and λ is the eigenvalue. The eigenvalue allows the vector to be scaled in the same way that a static matrix would scale the vector. When dealing in eigenspace this is extremely useful as the computational complexity is greatly reduced when using scalars opposed to matrices.

An eigenface is an eigenvector computed from an image of a face. In order to create an eigenface an image is first flattened out from a NxN matrix of pixels to a N^2x1 array of pixels as shown in figure 1.

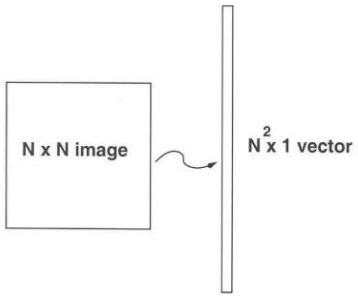

Figure 1: Converting a square image into an array [3]

Once an image has been converted to an array Principal Component Analysis (PCA) is done on the image to reduce the dimensionality. PCA is a method that seeks to maximize the amount of variance between multiple samples in each dimension. The benefit of this is instead of having information spread out greatly among many dimensions it can instead be focused into a set of dimensions that are much smaller than the original set.

In order to reduce the dimensionality, the PCA method first computes the *mean image* out of a set of input training images. The mean image is computed as:

$$\psi = \frac{1}{M}\sum_{n=1}^{M} \Gamma_n$$

Where M are the amount of input training images and Γ_n is a single training image. The mean image is redundant information that every image has in common and is thus not useful for computing characteristic differences between a set of images. In order to continue with PCA the mean is subtracted from every training image.

In biometrics the set of input training images are also referred to as the *gallery images.* This is in contract to *probe images* which are of people of unknown identity and compared to the gallery images for identification.

The new training images with the mean subtracted are shown as:

$$\phi_i = \Gamma_i - \psi$$

The set of training images are stored in a matrix **A** such that:

$$A = [\phi_1\ \phi_2\ \phi_3 \dots \phi_M\ ]$$

The **A** matrix is then used to calculate the covariance matrix **C** computed as:

$$\boldsymbol{C = AA^T}$$

The covariance matrix **C** shows how much variance exists between every pair of dimensions. The covariance matrix is used as the basis for generating the eigenvalues and eigenvectors/eigenfaces discussed previously. By computing the covariance matrix on a set of samples that have been normalized via mean subtraction only the extreme differences between the data sets are being highlighted; which is perfect for a system that focuses on the diversity of features.

One issue that can be found with this method is the sheer size of the covariance matrix. If the input images are of size NxN pixels then the covariance matrix will be of size N^2xN^2. As N increases the covariance matrix calculation increases exponentially which could be problematic for calculations.

Fortunately, even though the size of the images are of order N^2 the complexity of the covariance matrix can be reduced to the number of images in the gallery (M) via some clever linear algebra manipulation. By reversing the order of the multiplication of the **A** matrices the M eigenvectors found from this new **C`** correspond to the M largest eigenvalues of **C** [3].

Once the eigenvalues are found the eigenvectors that correlate to the largest eigenvalues can be shown to contain the most variance and are called the principal components. By arranging the dimensions of an image by the principal components the amount of information required to approximate the same image can be greatly reduced.

B. Facial Recognition with eigenfaces

By using a linear projection with the eigenfaces any new image can be projected into *face space* via the following operation:

$$\varpi_k = u_k^T(\Gamma - \psi)$$

Where u_k is the k^{th} eigenface. The set of projected images into eigen space are then shown to be:

$$\Omega_i = [\omega_1, \omega_2, \omega_3, \dots, \omega_k\ ]^T$$

When performing facial recognition, what is most likely occurring is that all of the gallery images are stored in a database already projected into face space. In order to test a new image against the database the probe image is projected into the face space with the operation above and then it is compared to all of the gallery images via some distance metric.

After computing the distance between the probe image and every gallery image in face space a few possibilities exist.

1. The image maybe close to a "face" indicated by being below a certain threshold
2. The image may be in the face space but above a threshold and thus not belonging to any specific class
3. Outside of face space but still "near" a class
4. Out of face space and not near a class (not recognized as a face)

Figure 2 shows a face space composed of two principal axes with three projections from the gallery located on it. The numbers correspond to the four possibilities discussed above [2].

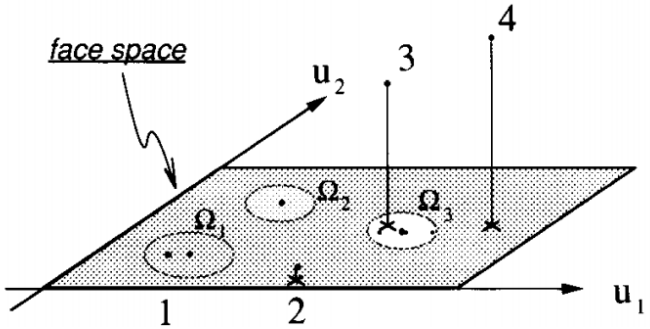

Figure 2: Face Space with two principal components and four possibilities [2]

C. K-Means Clustering

In addition to performing facial recognition it should be possible to identify other attributes when probing the gallery. One such possibility is determining whether the probe is male or female based solely on similarities in the face space of the gallery.

In order to achieve this a naïve approach can be used called k-means clustering. K-means is an unsupervised learning algorithm that seeks to classify a data set as belonging to a *cluster* based on iteratively updating the centroid of the cluster and measuring the distance from the centroid for each data set. In k-means the amount of clusters is fixed *a priori.*

The k-means algorithm will repeatedly loop where each iteration will

1. Calculate the distance between the current data set and all of the clusters
2. Assign the data point to the cluster whose distance is the least
3. Recalculate the cluster's new centroid
4. Check if no data point was reassigned during the entire loop else repeat at one (or break if x iterations were run)

The k-means algorithm performs the best when the data is very distinct; which could be a problem given this data set.

II. Problem and Methodology

A set of 200 gallery images of 100 unique individuals (two per person) were provided along with 100 probe images (one per

person) of the same set of people. The gallery (and probe) images are all of the same size with all of the faces centered in the image. This is critical for proper facial recognition and analysis. With this data there were two tasks to be accomplished:

1. Facial Recognition: Use the eigenface method to analyze and create the eigen space for the gallery images and then probe while varying the value, k, of principal components used
2. Gender Identification: Use the k-means clustering algorithm to partition the data into two classes, male and female, and then proceed to attempt to identify based on the trained system whether or not the probe images were being properly inserted into the correct class

A. Problem One: Facial Recognition

Following the steps in section I part A, *Principal Component Analysis and Generating Eigenfaces*, a python script was written to map the 200 gallery images into the face space. The eigenfaces generated from the 200 gallery images can be seen in figure 3. Each of the images in figure 3 reveal information about the gallery images. The mean image's most prominent feature are the eyes and eyebrows which show the least deviation amongst all of the photos and as such are removed prior to calculating the eigenfaces.

The 1st eigenface displays a very plain face that tends to show more information about the lighting in the image than anything about the face itself. The following eigenfaces tend to focus on a set of features of the face. Discerning information from them directly by looking at them is not an easy task.

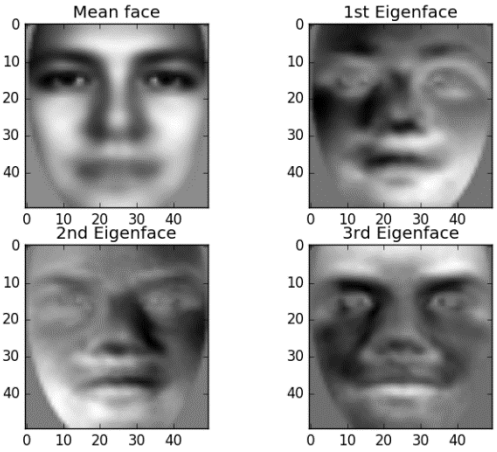

Figure 3: The mean face and top three principal component eigenfaces generated from the gallery images (X,Y) axes represent pixel values

Once the face space was generated and populated by the gallery face projections, each face in the probe image set was matched (to some success) to the gallery images. This was done by computing the Euclidean distance between each probe projection against each gallery projection. The solution was given by selecting the face form the gallery with the smallest distance.

This problem was repeated over a range of eigenfaces generated from a different number of principal components. The number of principal components was ranged from [10,100] with a step size of 10.

The amount of variance that the number of principal components accounted for over the step size (using the entire gallery set) was calculated as shown in table 1. This table shows just how much variance is contained in the first few principal components; in fact three quarters of all of the variance is within the first 20 principal components. As the number of principal components is increased a drastic improvement in recognition should not be expected.

Number of Principal Components	Total Variance of images accounted for
1	23%
10	62%
20	75%
30	81%
40	86%
50	89%
60	91%
70	93%
80	94%
90	95%
100	96%

Table 1: Total variance accounted for as a function of the number of principal components

Finally, recognition was re-performed without using the eigenface projections. To do this each raw probe image was compared to the raw gallery images until the minimum distance between the images was found to be the smallest. The success rate for this method along with the speed to compute it was compared to the previous methods using the face space.

B. Problem Two: Gender Identification

Next the k-means clustering was used to try and gain information about the sex of the images in the probe data based on clusters generated from the gallery images in the face space. In order to attempt to cluster the probe data based on the data received from the gallery two unique approaches were used.

The first approach involved splitting the gallery image based on the *a priori* knowledge of whether a person was male or female. The two groups were then averaged together to create the centroids for their respective class. The k-means algorithm was then run on the probe data attempting to adjust and place the images into the correct class. This was done over a range of principal components - [10,100] in steps of 10.

The second approach naively passed all the gallery data projected into face space and used the k-means algorithm to find the decision boundary between the two classes automatically. Once the cluster was trained the probe set (also in face space) was then fitted to the classes. This was done over a range of principal components – [10,100] in steps of 10.

The amount of recognition achieved in these methods were recorded along with a validation based on two criteria: the Dunn internal index and the F-measure external index. Internal indices are criteria that use information inherent in the data while external indices require information about the data itself [4]. This was done over a range of principal components – [10,100] in steps of 10.

The Dunn index is an internal index that evaluates the clustering based off criteria from within the cluster. The way it operates is by finding the smallest distance between two clusters and dividing it by the maximum distance found within a cluster. It is attempting to compare the dissimilarity between multiple clusters along with their diameters. The Dunn index is attempted to be maximized and exists between [0,∞) [4].

The F-measure external index is a weighted average of the ratio of precision and recall from information theory. The F-measure exists between [0,1] and is best at one. Precision is the ratio of the number of correct classifications to the number of classifications made while recall compares the number of correct classifications to the true amount of data in that class [4].

III. Results

A. Problem One: Facial Recognition

Similarly as to what was shown in table 1, it was observed that the amount of recognition tapered off quite early in the recognition process. As the range of principal components increased the recognition rate increased as well, though much more slowly than at first. Table 2 shows the recognition rate along with the time for each iteration (run on a Core i5 2.6GHz 4-core with 8GB RAM and no dedicated GPU). Figure 4 shows the corresponding recognition plot as a function of the amount of principal components.

# Principal Components	Time for algorithm (s)	Recognition Rate (%)
10	0.4301	74
20	0.4300	82
30	0.4327	83
40	0.4306	84
50	0.4397	87
60	0.4378	87
70	0.4363	87
80	0.4372	88
90	0.4406	88
100	0.4415	89

Table 2: Recognition Rate as a function of the number of principal components

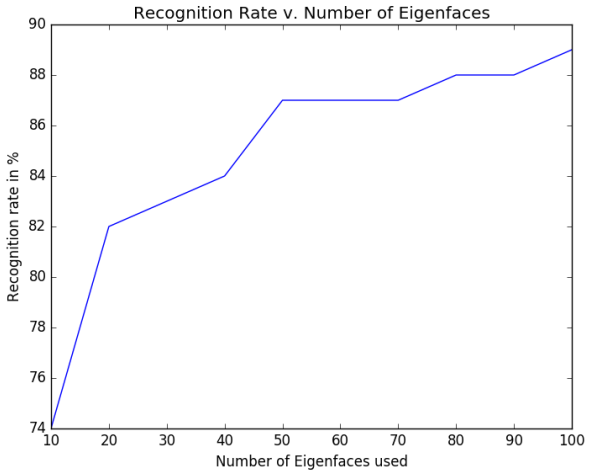

Figure 4: Recognition Rate as a function of the number of principal component

After removing the first principal component (the one most dependent on lighting) and rerunning the same procedure the recognition rate started out at a lower value then ended up at a maximum of 91% when 100 principal components were used at 0.4406 s. Figure 5 shows the recognition rate with the removed first principal component with a smoother curve due to starting at a lower variability set of principle components. The curve also displays a curious drop at 80 components compared to the monotonically non-decreasing curve shown in figure 4.

The recognition is lower in figure 5 than in figure 4 for a small number of principal components because a vast amount of the image variance is contained within the first principal component.

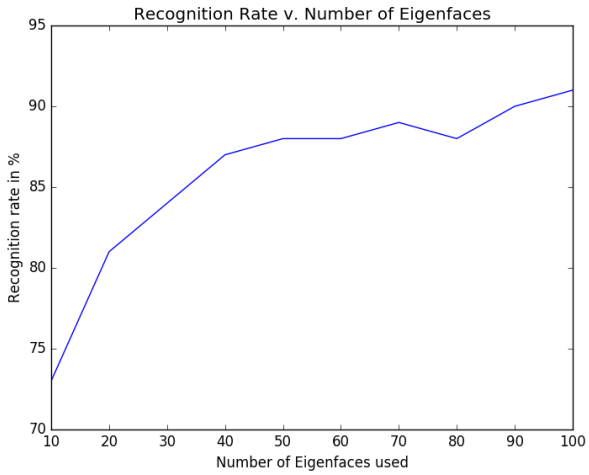

Figure 5: Recognition Rate as a function of the number of principal components with the first component removed

When using the raw images without face space projections it was observed that the recognition rate was also at 91%. However the classification time for this was 0.7974 seconds which is nearly double the amount of time required for the improved PCA

method to run. The PCA method reduced the gallery feature size from 500,000 data points to 20,000.

B. Problem Two: Gender Identification

For both of these problems the primary component eigen face which deals with luminosity was removed from the set before testing.

a) Using a priori gender information

When using the pre-generated centroids from the male and female classes respectively an extremely low recognition rate was observed. After iterating over the range [10,100] principal components, the recognition rate was noisy and consistently low as shown in figure 6. This is consistent with a poor and unstable F score shown in figure 7.

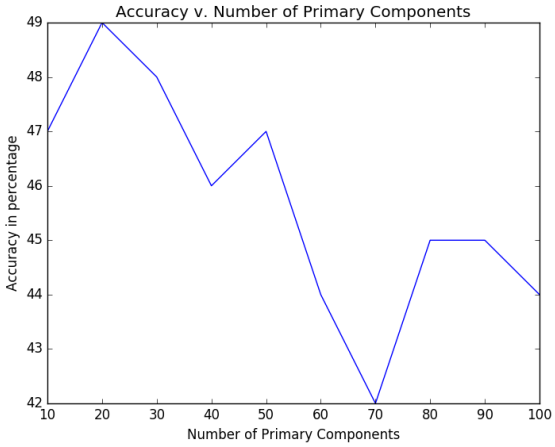

Figure 6: Accuracy v. number of Primary Components for the pre-separated data

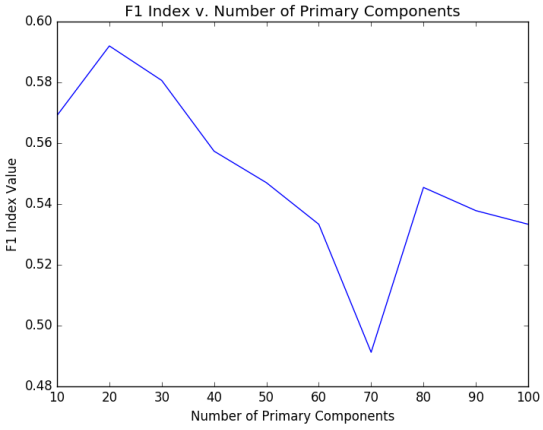

Figure 7: The decreasing F score for the pre-separated data

The Dunn score as shown in figure 8 showed an improved increase which is not surprising considering the goal of the metric is to show strong correlation between inter cluster data. When looking at figure 9 which shows the groupings for the two classes it is obvious that the groupings were made along the primary component axes for 10 principal components; which does not likely relay information about the gender. The plot shows the primary component on the x-axis, the secondary component on the y-axis, and the size of the data points represents the third primary component; the color is indicative of class.

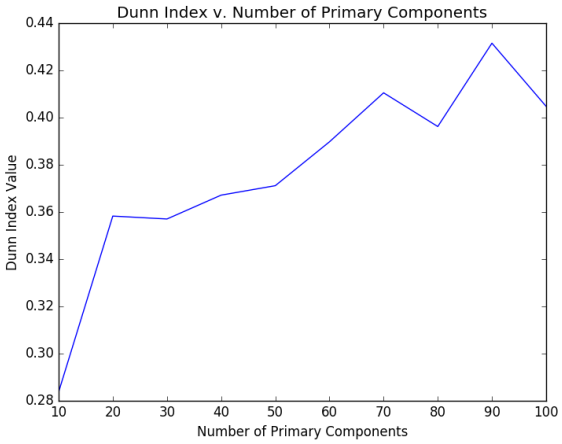

Figure 8: The increasing Dunn index for the pre-separated data

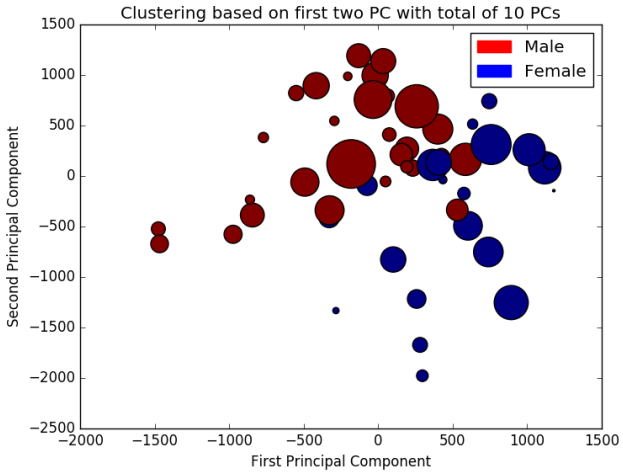

Figure 9: The separation of the classes based on the first three primary components with a total of 10 PCs

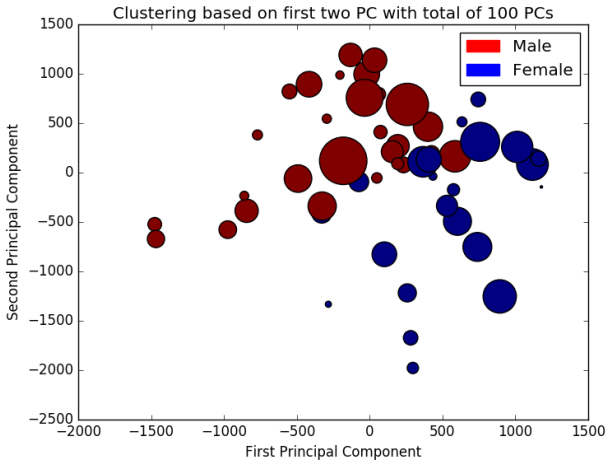

Figure 10: The separation of the classes based on the first three primary components with a total of 100 PCs

Figure 10 shows the same plot used with 100 principal components. Between figures 9 and 10 there is only one data point that has switched from male to female. This implies that increasing the number of principal components did not affect the recognition rate.

b) Training the cluster without a priori knowledge

When using the gallery data to train the cluster and generate the centroids without manually selecting centroids the performance overall was improved though still far below what was useful. The recognition rate was still nearly as bad as flipping a coin as shown in figure 11. The F index, while having different values than the recognition rate, followed the trend almost identically as shown in figure 12. Due to the dependence on the recognition the F-index does not provide useful information about the performance of this clustering technique.

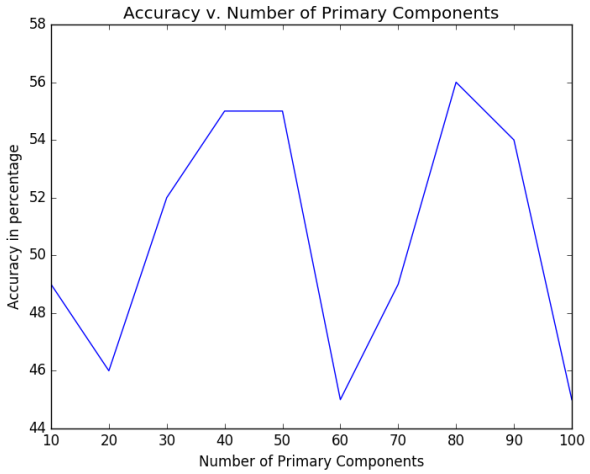

Figure 11: The recognition rate of the adapted weights method

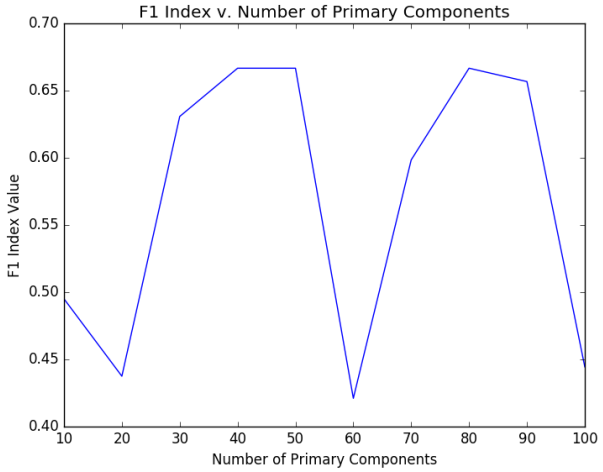

Figure 12: The F-Index following the recognition trend

The grouping of the data using the adapted weight method shows a different partition than the previous method of using the *a priori* gender information. Figure 13 shows the separation region is more sensitive to the second principal component than was seen in figure 9. This is likely the result of the clustering method using more information overall about the principal components by adapting over many iterations.

Figure 14 shows the Dunn index for the adapted weights method. The Dunn index is favorably increasing and doing so in a much more smooth manner than the *a priori* method. The Dunn index is also increasing monotonically suggesting a better solution exists with the adapted weights method.

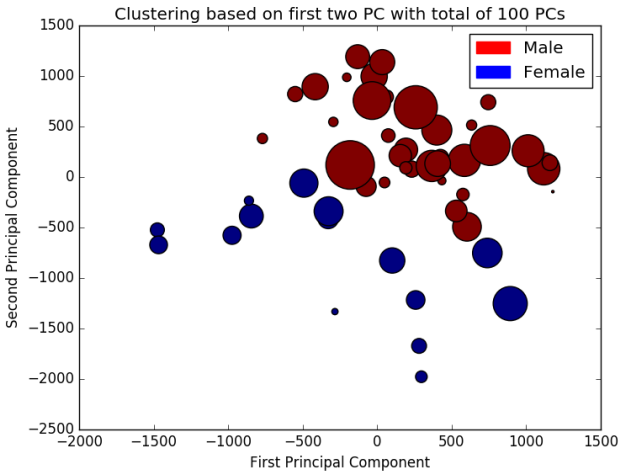

Figure 13: The class clustering without using the *a priori* information for 100 PCs

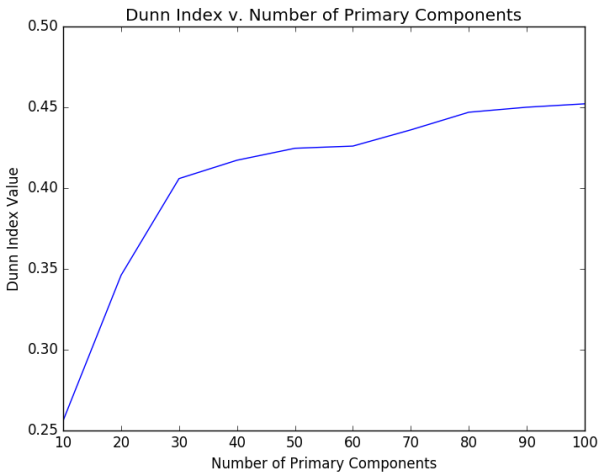

Figure 14: Dunn index for the adaptive weights method

When comparing the two clustering techniques another important insight can be made when observing the difference in the centroids over all the dimensions. Figure 15 shows the value of the centroid with respect to the dimension of the principal component for the *a priori* method while figure 16 shows the same plot for the adapted weights method.

It was seen that there was a greater magnitude of difference in general between centroids for the *a priori* method. This would imply that the clustering has more information to base decisions on. By observing the clustering in figure 13 for the adapted weights method the class boundaries for the first two classes can be seen to fall more along the positive and negative bound for each component's axis. In figure 10 for the *a priori* method this is not the case which is most likely a result of the turbulence observed in figure 15.

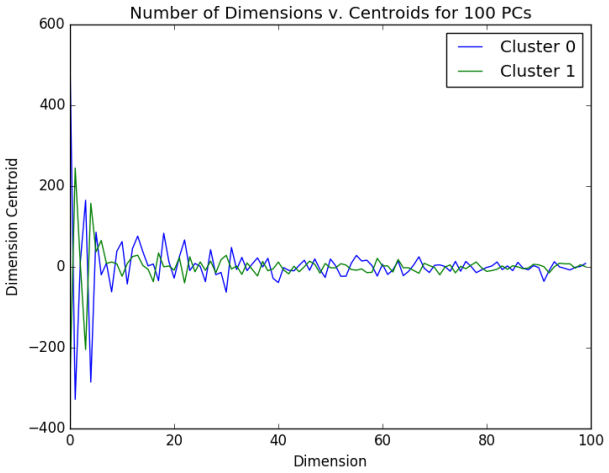

Figure 15: The centroid centers on a per dimension basis for the *a priori* method

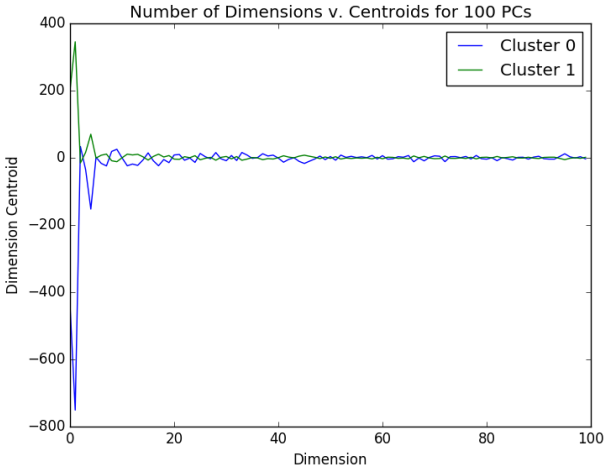

Figure 16: The centroid centers on a per dimension basis for the non *a priori* method

IV. Conclusions

With respect to facial recognition it was shown that using the eigenface method is a very powerful tool. The eigenface method could both achieve the same level of recognition that using raw uncompressed images could and it was also able to speed up the process while simultaneously compressing the data.

By intelligently selecting the proper number of principal components, and recognizing that the first principal component may contain misleading information, the problem of facial recognition for this class of photographs (greyscale images of the same dimensions, of the same angle, and with the same facial centering) becomes a trivial problem.

The soft biometric classification problem of determining gender from the gallery face space data was not as easily solved. By using two different techniques to train the k-means clustering algorithm the recognition rate was not much better than a coin flip for either.

Since the clustering failed to solve the binary class problem with this data set the criteria used to classify the data was not extremely useful. To assess why the clustering failed it would be useful to include more internal and external metrics. By increasing the number of principal components used in the clustering process the recognition rate did not increase indicating that the problem is more significant than just providing more information about the images themselves.

As an alternative, it would be worthwhile to investigate whether Linear Discriminant Analysis (LDA) opposed to PCA might be a better tool to solve this problem. LDA works by maximizing the ratio of between classes and within classes instead of the overall scatter which PCA uses. By applying LDA to the *a priori* information of gender in the gallery space more useful cluster boundaries may become available [5].

This method is also known as taking the Fischer Faces when using LDA compared to the Eigen Faces in PCA.

The source code and results for this project are available at https://github.com/gutelfuldead/eigenface_project.

V. References

[1] S. A. Papert, "The Summer Vision Project," MIT, Masachussets, 1966.

[2] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

[3] John Hopkins University, "Eigenfaces for Face Detection/Recognition," 2008. [Online]. Available: http://www.vision.jhu.edu/teaching/vision08/Handouts/case_study_pca1.pdf. [Accessed 26 November 2016].

[4] E. Rendon, I. Abundez and e. al., "Internal versus External cluster validation indexes," *International Journal of Computers and Communications*, vol. 5, no. 1, pp. 27-34, 2011.

[5] OpenCV, "Face Recognition with OpenCV," 2014. [Online]. Available: http://docs.opencv.org/2.4/modules/contrib/doc/facerec/facerec_tutorial.html#fisherfaces. [Accessed 26 November 2016].