UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

BIOMETRIC CLASSIFICATION WITH FACTOR ANALYSIS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

NGAO D. MAMUYA
Norman, Oklahoma
2010

BIOMETRIC CLASSIFICATION WITH FACTOR ANALYSIS

A DISSERTATION APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER
ENGINEERING

BY

____
Joseph P. Havlicek, Chair

____
James J. Sluss Jr.

____
Thordur Runolfsson

____
Hong Liu

____
Tomaz Przebinda

Acknowledgment

I would like to thank all my committee members for their work. Especially my advisor professor Joseph P. Havlicek. Finally, I would like to thank my family and friends for all their support.

Table of Contents

List of figures

Abstract

BIOMETRIC CLASSIFICATION WITH FACTOR ANALYSIS

Ngao D. Mamuya
The University of Oklahoma, 2010

Supervisor: Joseph P. Havlicek

This research presents a study on biometrics classification using Factor Analysis (FA). As a multivariate statistical tool, factor analysis is useful for understanding the underlying structure in a dataset. Moreover, in addition to achieving an economy of the variables, the "factors" or hypothetical constructs can provide an alternate yet succinct representation of the data. It is a method of determining, from an observable set of variables, a basic set of components that are common to all the observations. In this study, the loadings (or weights) on the Factors are used to classify the data in alternate representation. In particular, we will examine and group the data according to three biometric features. In the first part, we demonstrate the capabilities of factor analysis to capture the gender of

the individual. This will enable us to use FA as a gender classifier. The next study will show the use of an FA as a facial hair classifier. Given a group of individuals, we will be able to classify them as either having beards or not. Finally, in the last part presented in this work, we will work on classifying the facial expressions of a group of Japanese women. Given all seven universal expressions per subject (two or three of each expression), we will use factor analysis to group each subject according to their expression. Furthermore, given an individual with a particular expression, we will use factor analysis as a biometric measure in the determination of the particular expression exhibited.

Chapter 1.

Introduction

Biometric technologies are methods based on a person's physiological characteristics. These characteristics can include face, speech, fingerprints, voice and iris recognition among others. They can be used in the identification and or verification of individuals for the purpose of controlled access or secured transactions. A biometric system captures and transforms the characteristics of an object into a compact form that is subsequently matched to a stored database of previously processed characteristics. Identification, verification or classification is achieved according to some similarity measure.

Face recognition [1] from still and video images is one such biometric technique which employs the facial characteristics of an individual for use in identification or verification. The potential applications of automated facial recognition systems are numerous. End users of these systems include both public and private sectors. Some applications are mugshot identification, surveillance or screening of crowded areas for known individuals by law enforcement, identity verification for security controlled entry points (airport checkpoints, buildings, accounts access etc…) or secured electronic financial transactions.

Traditionally, facial recognition was solely accomplished by human operators. The human visual system is very sophisticated. It is both accurate and robust to a host of extenuating factors including aging, differing expression, variable illumination, partial occlusion and to a degree some disguises. Some drawbacks are the number of distinct identities a human can accurately and efficiently process and the time required to successfully accomplish each task. A robust and successful automated system can overcome these limitations by utilization of virtually unlimited storage capacity and low computational processing time. In general, a system capable of this is very difficult to develop. Thus, most systems are designed to be either identification, verification or classification systems.

Identification refers to identifying a probe (a user) against a database of known individuals, while verification is only concerned with verifying that a user is indeed who they claim to be. As such, identification is an N-to-N analysis, while verification is a 1-to-N analysis. This difference is critical in the design of an application. In a classification system, the goal is to be able to classify an object accordingly. In this research, I will look at a gender classification system. I am interested in building a gender classifier using Factor Analysis. In addition to this I will also show that an FA is capable of capturing other facial characteristics, specifically facial hair.

Factor Analysis belongs to the family of multivariate statistical methods. As such, in this research, I will examine a few of the multivariate methodologies used towards the goal of biometric analyses. Specifically, I will take a brief look at the methods of: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), and Factor Analysis (FA). We will look at how some of these procedures are used in facial recognition.

What will become evident is how the different methods differ in how they represent the data. Each has strengths and weaknesses. Depending on whether the end goal is representation, classification, or discrimination some of these methods will be more suited to the task than others. In addition to this, we will also discover that within each method, there exist ways of adjusting the procedures to deal with different datasets or at least changes that will enable them to answer some questions more effectively. Some of these are, when doing PCA, how should you choose which eigenvectors for the representation and should some of the leading eigenvectors be discarded. Or when doing a Factor Analysis, what if any rotations should you employ. In the next chapter, I will take a closer look at the above mentioned techniques.

Until recently, the gender classification problem was primarily investigated by the psychology and cognitive communities [1], [2]. Lately, it has come under review by the statistical, and in general, the data processing communities. Generally speaking, there are two methods: geometric and appearance based. The appearance based approaches, which deal with the image as a whole, gained favor after Kirby and Sirovich's successful image representation using principle components [3]. Moghadam and Yang's support vector machine (SVM) method for gender classification [4] achieved relatively good results on the FERET [5] database. SVM's represent the feature vectors as a set of points in a high dimensional space where the boundaries between classes are expressed as hyperplanes. For non linear problems, kernel functions can be used to map the features to a linear solution in the new space. Using a Gaussian kernel, SVM was able to achieve a 3.4% error rate. Abdi *et al*. [6] reported gender classification accuracy of 91.8% for faces with hair information included. Neural networks have also been studied in this area. The Adaboost methods [7] have shown very good results with error rates of less than 5%. A combination of the SVM and Fisher Linear Discriminante (FLD) classifiers [16] show a success rate of 94%. Using a radial base function (RBF) network, Brunelli and Poggio were able to achieve a gender classification rate of 87.5% [9].

A lot of the work done (PCA, LDA, ICA etc…) falls under the general heading of Multivariate Statistical Analysis. In this dissertation, I will take a closer look at some of these methods. I will begin by looking at the method of Principle Component Analysis (PCA), followed by that of the Linear Discriminant Analysis (LDA). The next method to be examined is Independent Component Analysis (ICA). Finally, I will look at Factor Analysis (FA).

From the above, it is clear that as with most multivariate analysis, the goal is that of simplifying the data. Given a large dataset, we would like to be able to get a better feel of the relationship between the many observed or measured variables. These methods will all answer some specific questions about the datasets. Some will lend themselves more intuitively to the problems of classification, while others will be more suited to the task of efficient representation. Furthermore, we will also find that within some procedures, there will be room for customizing the procedure according to the dataset or based on the question. In the coming sections, I will take a closer look at the mechanics of these procedures.

In this research, I will employ a Factor Analysis (FA) on test groups to answer and classify the subjects according to whether they are: male or female, bearded or not, and finally a facial expression classifier.

Chapter 2.

Background

The use of biometrics can be traced back to the 19th century. In 1882, Bertillon presented a biometric system for the aid of identifying criminals [10]. His system consisted of taking several measurements of an individual's head and body. This information was processed into a formula to produce a unique and time invariant identity for each criminal. The system was eventually adopted as a standard in prison systems. Later, fingerprinting became the de facto standard. Today, in addition to finger prints, DNA testing is also an accepted form of identification. All the methods discussed so far share the common trait of requiring the cooperation of the subject. While this can sometimes be achieved, it is not always the case.

Facial recognition, as a biometric, is a very desirable alternative. This is highlighted by the fact that, under certain conditions, the whole process of acquiring the data, processing it and establishing identification can be accomplished with minimal to no cooperation from the subject.

The human system of facial recognition is very sophisticated and robust.

However, a major drawback to this system is in the numbers it can process effectively and accurately. This is where the advantage of a computer is greatest. An automated system capable of facial recognition success comparable to that of humans would find applications in a myriad of cases. Some applications would be for use in the law enforcement, access to ATM machines or computers, screening and surveillance at restricted entry points. Today, many systems use a combination of cards and/or passwords to perform access control. Some even use finger printing or iris scanning to establish identity [1]. The advantage of a facial recognition based system would be the convenience of requiring very little in the way of cooperation.

Early work on facial recognition focused more on measuring the distances between various features on the face. During 1964 and 1965, Bledsoe *et al*. [2] created the first automated system to do this. The system was called a man-machine. It performed recognition with the aid of an operator who manually entered the coordinates of various feature points, such as center of pupils, into the computer by way of a RAND tablet. The computer would calculate various distances between the coordinates, normalize them and store them. A match was made by comparing every new subject to the stored distances in a database, and selecting the closest one.

Another useful biometric task is that of gender classification. Early work on gender classification used a system of point-lights to capture the gait of an individual. It was found that gait was unique to an individual. Johansson [11] was the first researcher to employ the point-lights to identify and analyze an individual's gait. He attached point-lights to the main joints of individuals, and viewed their motion against a dark background. From this, he was able to show that an individual's gait can be used as a identification measure. Using point-lights, Kozlowski and Cutting [12] were able to achieve success rates of about 63% for gender classification. Changing some of the parameters, Barclay [13] showed that structural cues play an important part in the classification. Mather and Murdoch [14] further showed that dynamic motion cues are better suited to the gender classification problem than structural ones. Aside from gait analysis, most of the early research on gender classification was either appearance based or geometry based. The geometry based models utilized various distances between facial features as input. Brunelli [9] used such a system. As the input to a HyperBF neural network, he used 16 such distances. He reported an accuracy of about 79%. Burton *et al.* [15] used 73 facial distances as input to a linear discriminant analysis (LDA) for gender classification. These methods both showed similar error rates of more than 10%.

Appearance based systems treat the whole image as input. They usually start with a training dataset. This is used in training a classifier. Some of the classifiers are neural networks, SVM or LDA systems. Golomb *et al.* [16] used a fully connected, two -layered back propagation network as their classifier (SEXNET). Using a training set of 80 individuals (40 male and 40 female), and a probe or test gallery of 10, they reported accuracy rates of around 91.9%. Yen *et al.* [17] adopted the same method but with a larger dataset. Using 1400 facial images, they report a success rate of 90%. Another neural network system is that of Cottrell [18]. Using a two stage neural network and 64 training images, Cottrell *et al*. obtained accuracy levels of 63%. Utilizing PCA based features as inputs to a perceptron classifier, Abdi *et al*. were able to realize rates of 91.8% [4]. Recently, Moghaddam *et al*. [4] SVM method has shown promise. Using RBF, they have been able to achieve a success rate of 96.6%.

Facial expression analysis is another area that has recently garnered the interest of the scientific community. A system capable of accurate and real time facial expression analysis would be a great step towards a fully automated Human Computer Interaction (HCI) machine. Research in the social sciences [19], [20] has shown that facial expressions help coordinate conversations. Moreover, in a study of non verbal communication, Mehrabian [21] showed that verbal communication contributed to only 7% of the overall message. The rest was attributed to voice intonation and facial expression. Specifically, 38% and 55% respectively.

A complete Facial expression analysis system consists of three separate parts. The first is facial detection, followed by facial expression extraction, and finally classification of the extracted expression [1]. In this work, I will only examine the classification part. To date, the most popular method in use is that of the *Facial Action Coding System (FACS)* [20]. This system, in a sense, is a code book for all the expressions that can be generated by a combination of the contractions of a set of 44 different facial muscles or *Action Units (AU)*. Each set of combinations will result in a subtle change in facial appearance. The FACS system provides an investigator with the corresponding facial expression. Conversely, given a facial expression, an observer can match it, or express it as a set of active AU's. Most studies on facial expression analysis are concerned with the six universal expressions advocated by Ekman [19]. Ekman defined the six basic expressions as *happiness, sadness, surprise, fear, anger, and disgust.* Other researchers include the neutral expression as the seventh. The difficulty encountered by many is the lack of universally accepted definitions or objective facial descriptions of these basic emotions. Another involves the lack of a categorization of blends of these basic expressions [22].

Facial expression systems generally fall under the three categories: template based, rule based, or neural networks. In template based, the facial expression to be examined is compared to previously defined and stored templates of all the expressions. Some of these templates are defined in terms of the 44 AUs listed in the *FACS.* Commonly used methods are PCA, LDA or Elastic graph matching among others. Cohn [23], using discriminant functions on a combination of AUs reported an accuracy of 88% with 100 subjects. Edwards [24] achieved an accuracy rate of 74% with 25 subjects. Their work involved a Mahalonobis distance-based PCA and LDA. Hong's Elastic graph matching method [25] was able to attain a rate of 81% on 25 subjects. Using PCA and LDA on labeled graph vectors, Lyons [26] achieved rates of 75% to 92%.

Neural networks based methods as done by Hara, Padgett, Zhang, and Zhao report rates of 85%, 86%, 90%, and 100% respectively [27]. Pantic's rule based method [28] reported rates of 91% with 8 subjects.

In this research, I will be examining multivariate methods. As such, I will examine more closely the following methods: PCA, LDA, ICA and FA.

2.1 Principle Component Analysis

This method was first described by Karl Pearson back in 1901 [29]. However, it was not until 1933, when a practical description was given by Hotelling [30], that it gained more widespread use.

Today, as a result of the availability of computational hardware, PCA has found numerous applications across a wide range of fields. PCA can be described as a method of dimension reduction. It casts a dataset of many variables (*N*) into one of *P (P<<N)* uncorrelated variables in a manner that is more suitable for representation. This is achieved by transforming the *N* variables $X_1, X_2, \ldots, X_N$ through a linear combination to produce new variables $Y_1, Y_2, \ldots, Y_P$, *(P << N)* that are uncorrelated in order of their magnitudes (variances). As, a result, it can be seen that PCA will not always achieve a more compact form. Clearly, if the original dataset's variables were already uncorrelated, then the new variables, while uncorrelated in order of their importance will not achieve the desired outcome of reduced dimensionality.

As stated earlier, PCA has been found to be useful in varied fields of study. It has recently been widely used in pattern analysis. It is to this end that I will explore its use and contributions.

The Eigenface method for face recognition was introduced by Turk and Pentland

[31]. Their work was based on the earlier work of Sirovich and Kirby [3]. Sirovich and Kirby used the eigenvector representation to approximate the image with only the largest eigenvectors (referred to as "eigenpictures"), thus achieving some compression. By using more of the eigenvectors, the resulting image is an improved approximation of the original. Before examining PCA's direct application towards facial recognition, I will first give a brief mathematical introduction to PCA.

PCA can be described as a subspace projection method which seeks to find a more compact representation or basis of a dataset such that each new axis has maximal variation while being uncorrelated to the other axes. Lets us begin with a dataset ***X*** of k observations $X_1, X_2, \ldots, X_K$, where each observation is X represented by a collection of *N* variables.

$$\boldsymbol{X} = \{X_i, i=1,2,\ldots,k\} \tag{2.1}$$

$$X_j = \{x_i, i=1,2,\ldots,N\}, j=1,\ldots,k. \tag{2.2}$$

The goal now is to find a combination of the *N* x_i's to produce a new set of observations *Y* with reduced variates *L* where *L* < *N*.

i.e.

$$\begin{aligned} Y_1 &= a_{11} X_1 + a_{12} X_2 + \cdots + a_{1\mathrm{L}} X_L + \cdots + a_{1\mathrm{N}} X_N, \\ Y_2 &= a_{21} X_1 + a_{22} X_2 + \cdots + a_{2\mathrm{L}} X_L + \cdots + a_{2\mathrm{N}} X_N, \\ &\vdots \\ Y_N &= a_{NI} X_1 + a_{N2} X_2 + \cdots + a_{NL} X_L + \cdots + a_{NN} X_N. \end{aligned} \tag{2.3}$$

Where the *Y*'s are a linear combination of the *X*'s subject to the condition that the *Y*'s will have zero correlation and the coefficients are such that

$$a_{11}^2+a_{12}^2+\cdots+a_{1\mathrm{N}}^2=1\,,$$

$$a_{21}^2+a_{22}^2+\cdots+a_{2\mathrm{N}}^2=1\,, \tag{2.4}$$

$$\vdots$$

$$a_{N1}^2+a_{N2}^2+\cdots+a_{NN}^2=1\,.$$

The reduction in dimension is achieved by the fact that PCA will order the new observations, *Y*'s, by their variances. Those which have variances below some set threshold will be ignored in the reconstruction. Thus,

$$Y_1=a_{11}X_1+a_{12}X_2+\cdots+a_{1\mathrm{L}}X_L\,,$$

$$Y_i=a_{11}X_1+a_{12}X_2+\cdots+a_{1\mathrm{L}}X_L\,, \tag{2.5}$$

$$\vdots$$

$$Y_L=a_{L1}X_1+a_{L2}X_2+\cdots+a_{LL}X_L\,.$$

Lets us now examine how the coefficients are computed.

The objective of PCA is to find a linear combination of the original variables, *X*, with maximum variance. The variance of *Y* is

$$Var(Y) = \frac{1}{(n-1)} (Xa)^T (Xa). \tag{2.6}$$

We now choose *a* to maximize this variance subject to

$$a^T a = 1. \tag{2.7}$$

This particular constraint is put in place to ensure the maximization is not a result of choosing arbitrary large values of the *a*'s.

We can solve this constrained optimization problem using the Lagrangian multiplier

$$L = (Xa)^T (Xa) - \mu (a^T a - 1). \tag{2.8}$$

Taking the derivative of L with respect to a yields

$$\frac{\delta L}{\delta a} = 2\mathrm{X}^T X - \lambda\, 2\mathrm{a}. \tag{2.9}$$

Setting this to zero, and solving we get the *characteristic equation*

$$(X^T X - \lambda I) a = 0. \tag{2.10}$$

The solution to the above can be obtained by finding the *Eigenvectors and Eigenvalues* of the covariance matrix $X^T X$ (also known as *Spectral Decomposition)*. At this point, we note that the *Eigenvalues* λ , are the variances of Y, and the sought after coefficients are given by their associated *Eigenvector.* The above problem can also be viewed and solved by way of a *Singular Value Decomposition (SVD).*

2.1.1 Facial Recognition using PCA

What makes PCA suitable for facial recognition is its ability for dimension reduction. Instead of computing distances between all the raw images, the distances will be computed between the transformed images. The Images to be considered are arranged as vectors of dimension (*nm*). The mean training image is subtracted from all the training images before solving for the eigenvectors and eigenvalues [31].

The collection of training images is the image space. PCA is used to find the vectors that best describe (in the sense of maximal variation) the distribution of the face images in the entire image space [3]. The projection of data from the original dimension to the reduced dimension or subspace spanned by the principle eigenvectors is optimal in the mean squared error sense[1]. This transformed subspace of vectors is called the face space or sometimes referred to as the *Eigenfaces* [31]. Depending on the application, sometimes not only the eigenvectors associated with the smallest eigenvalues are discarded. Sometimes some of the largest are discarded. This is the case when the most variations within the images are those caused by unwanted elements. If kept, this could introduce variations that would skew the results towards the unwanted characteristics. An example of this is the variance introduced because of variable lighting.

2.1.2 PCA Example

Let us now take a look at a practical example. For the data, I will use the face database provided by Olivetti Research Laboratory (ORL) [32]. There are ten different

[1] The projection of the subspace back to the original space has minimum reconstruction error.

images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position. Fig. 2.1 shows a sample of the raw images.

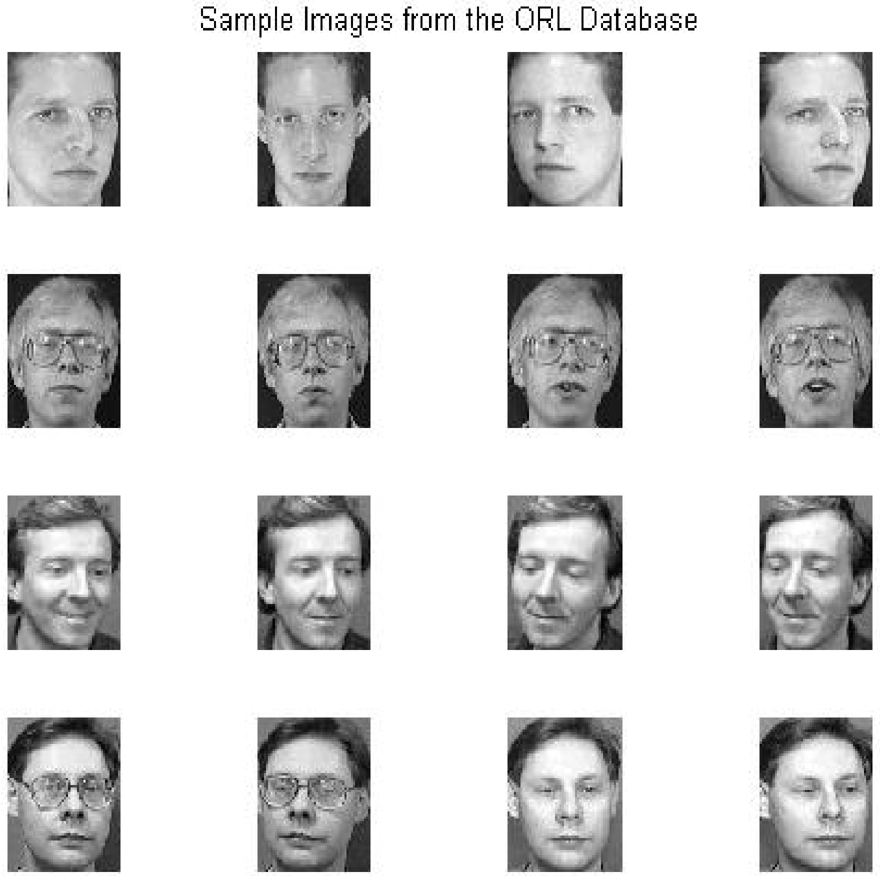
Sample Images from the ORL Database

Figure 2.1 Four subjects in the ORL Database.

In this example, I use the first nine images of every individual for the training set and the last (tenth) image of the individual as the probe set. Below are the steps followed.

1. Create a matrix with the pixel values of all the training images X_t by lining up each image as a row vector.

X_t	Pixel 1	Pixel 2	$\cdots$	Pixel *m*
Image 1	56	56		32
Image 2	45	40		38
$\vdots$				
Image *n*	40	55		35

2. Subtract the mean training image from the training images (rows) of X_t and probe images X_p

$$X_t = X_t - mean\{X_t\}$$

$$X_p = X_p - mean\{X_t\}$$

3. Compute the covariance matrix $X_t^T X_t$. To save computation time and space, we first compute the smaller matrix $(X_t X_t^T)$, and observe that both $(X_t^T X_t)$ and $(X_t X_t^T)$ share the same nonzero eigenvalues. The eigenvectors of

the matrix $(X_t^T X_t)$ are the products of the eigenvectors of $(X_t X_t^T)$ with the training matrix, X_t.

4. Solve the equation $(X_t^T X_t - \lambda I)a = 0$

5. Project the training set X_t onto the subspace[2] spanned by the eigenvectors.

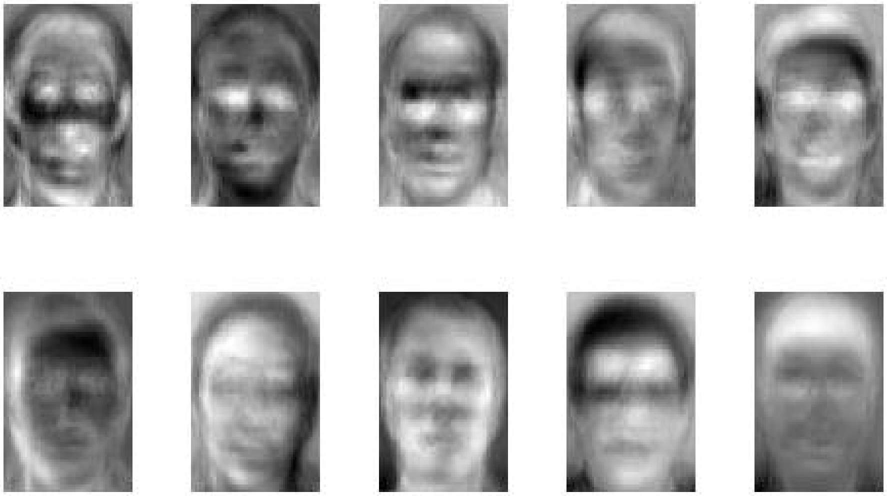

Figure 2.2 Ten Eigenfaces corresponding to the largest eigenvalues.

6. Project the probe set X_p onto the subspace spanned by the eigenvectors whose associated eigenvalues represent the most of the variance[2].

7. Compute the distance from the probe image to all the training images.

8. The probe image with the smallest distance to a training image is selected as a match.

[2] Eigenvectors associated with the eigenvalues accounting for 90% of the total variance.

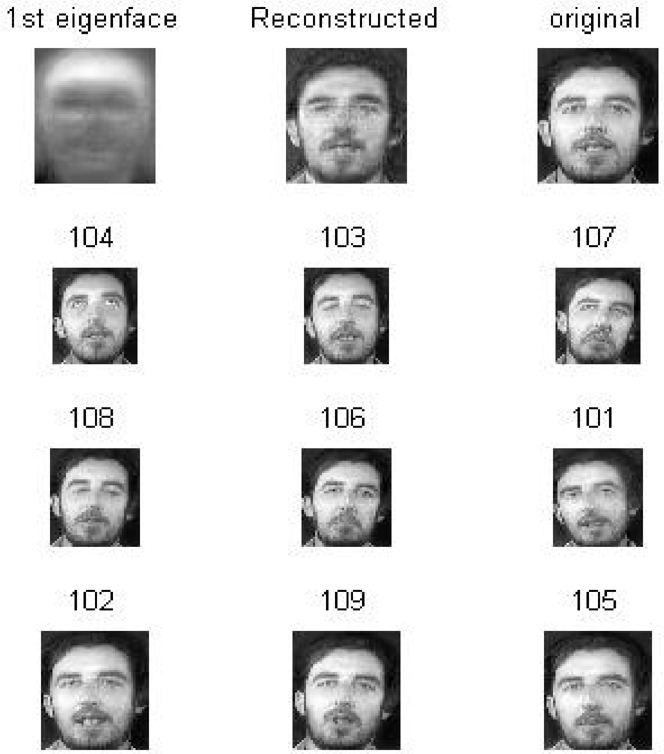

Figure 2.3 PCA Example.

Fig. 2.3 shows the image of the leading eigenvector in the top left position. This is followed by a reconstruction of the actual image using only the retained eigenvectors. The top right image is the original image. Below, starting at the left and moving to the right and down are the images sorted by their respective distances (ascending) to the original image.

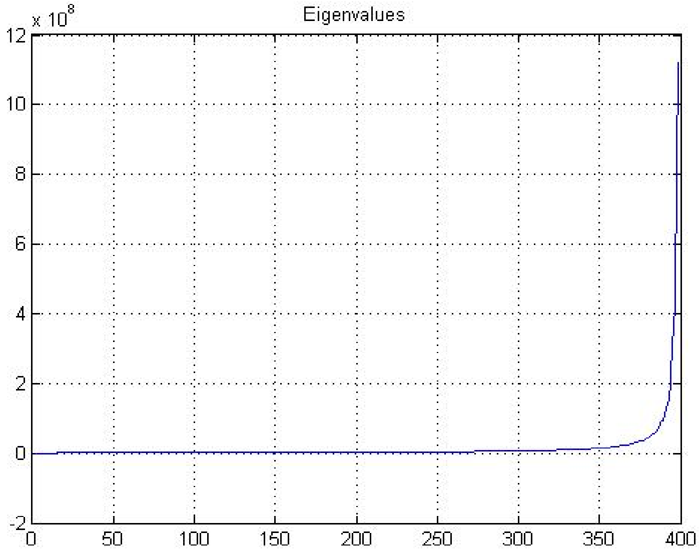

Figure 2.4 Plot of the Eigenvalues.

In Fig. 2.3 we see that the algorithm was successful in matching the probe image to all nine corresponding images in the training dataset. The figure also shows the image of the first eigenvector (eigenface). Fig. 2.4 shows a plot of the eigenvalues. Only a small (less than 50 in this case) number of the eigenvalues are used in the matching process. This is because most of the variation is captured by the few leading eigenvectors. This is can be seen in Fig. 2.4

2.2 Linear Discriminant Analysis.

Similar to PCA, *Linear Discriminant Analysis (LDA)* [33] can also be viewed as a dimension reduction technique. However, while PCA is best suited for representation of data, *i.e*. finding the axes that are most efficient for representation, LDA is a supervised learning algorithm that seeks axes (not necessarily orthogonal) which are efficient for discrimination. For a *C*-class classification problem, LDA finds the *C*-1 basis vectors that maximize the interclass distances while minimizing the intraclass distances.

The LDA classification is achieved by maximizing the ratio of the between class variance to the within class variance. Like PCA, the LDA method projects the raw images onto a subspace before computing distances. The difference lies in the formulation of the subspaces. LDA first calculates the scatter matrices, $\boldsymbol{S_i}$, for all the *i* classes. This scatter matrix, $\boldsymbol{S_i}$, is calculated as the sum of the covariance matrices for images in class *i*. The scatter matrices are given by

$$\boldsymbol{S}_i = \sum_{images\ x \in\ class\ i} (x-\mu_i)(x-\mu_i)^T, \tag{2.11}$$

where *x* are the centered (mean subtracted) images in class *i* and μ_i is the mean of the raw images in class *i*. From this follows the within-class scatter matrix $\boldsymbol{S}_{\boldsymbol{w}}$. This measures the scatter between objects of the same class according to

$$\boldsymbol{S}_w = \sum_{i=1}^{C} \boldsymbol{S}_i. \tag{2.12}$$

Here *C* is the number of classes.

Next, the between-class scatter matrix, $\boldsymbol{S_B}$, is computed. It is calculated as the sum of the weighted covariance matrix of the differences between the overall mean and the class means according to

$$\boldsymbol{S}_{\boldsymbol{B}}=\sum_{i=1}^{C} n_i(\mu_i-\mu)(\mu_i-\mu)^T, \tag{2.13}$$

where n_i is the number of images in class i and μ is the overall mean of the images.

The objective here is to find the transformation vector $\boldsymbol{w}$, such that

$$\max_{w} \frac{\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{B}}\boldsymbol{w}}{\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{W}}\boldsymbol{w}}. \tag{2.14}$$

The solution gives the generalized eigenvectors and eigenvalues of the within-class and between-class scatter matrices. The eigenvectors corresponding to the largest C_i are the basis vectors of the subspace.

In doing LDA, as above, notice that the overall within-class scatter matrix is the mean of the various within-class scatter matrices. This would not be a problem if all the classes were normally distributed, or sufficiently similar to each other [33]. However, this is not always the case. What happens when some classes are outliers? This raises the question of a weighted within-scatter matrix. How do we choose the weights?

2.3 Independent Component Analysis.

Independent Component Analysis (ICA) is a method that seeks to find the statistically independent components of a signal. In general, ICA comes under the class of problems known as blind source separation (BSS) [34]. Similar to other multivariate methods, ICA also projects data onto a different space. However, unlike PCA which uses second-order statistics to find a new set of variables (principle components) that are uncorrelated with maximum variance, ICA uses both second-order and higher-order statistics to find a new set of signals that have minimal dependence. Note that correlation is a weaker property than independence.

Given a mixture of signals ***X*** (training data), ICA finds the matrix ***W*** (the transformation or unmixing matrix) such that

$$\boldsymbol{W}\,\boldsymbol{X}^{\boldsymbol{T}}=\boldsymbol{U}\,. \tag{2.15}$$

Here ***U*** is a matrix with rows that have minimal dependence. Unfortunately, there is no closed form expression for finding ***W***. Instead this is done through iterative search methods [35]. Different methods rely on different search criteria. However, it has been shown that they almost always lead to similar algorithms [35]. A popular algorithm is the InfoMax method by Bell and Sejnowski [36]. InfoMax performs a gradient ascent on the elements of ***W*** to maximize the entropy $H(u)$ where

$$H(u)=\int f_u(u)\log(f_u(u))\,du \tag{2.16}$$

All the algorithms used fall into one of two fundamentally different types: architecture 1 or architecture 2.

In architecture 1, the input images in ***X*** are considered to be a linear mixture of statistically independent basis images ***U*** combined by an unknown mixing matrix A. In this architecture, the face images are variables and the pixel values provide observations for the variables. Projecting the input images onto the learned weight vectors ***W*** produces the independent basis images. Note that this will result in the images being spatially localized. The images are then represented by a linear combination (the coefficients are in the matrix A) of the independent basis image.

In order to control the number of independent components produced by ICA, it has been suggested [37, 38] to reduce the dimension of ***X*** via a PCA transformation. ICA is then applied to the reduced dimension eigenvectors to produce the independent basis images. Below in Fig. 2.5 is an example of eight images and their corresponding independent basis images. In architecture 2, the inputs are the transposed input images of architecture 1. In other words, the pixels are the variables and the images are the observations. The basis images from architecture 2 show more global properties than those from architecture 1. This can be seen below in Fig.2.6.

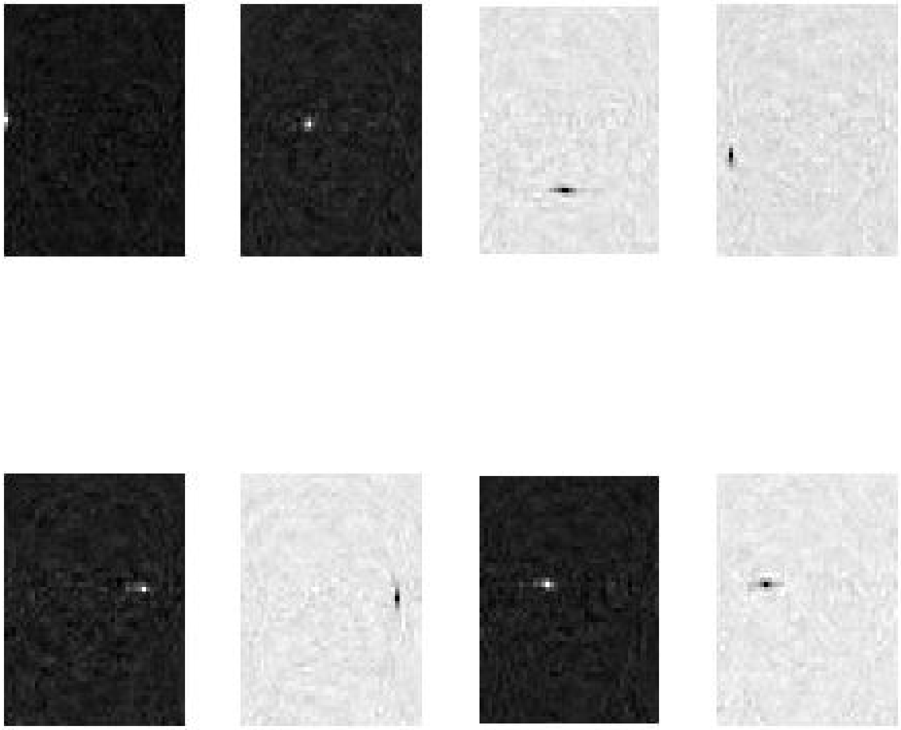

Figure 2.5 Eight feature vectors from ICA Architecture 1.

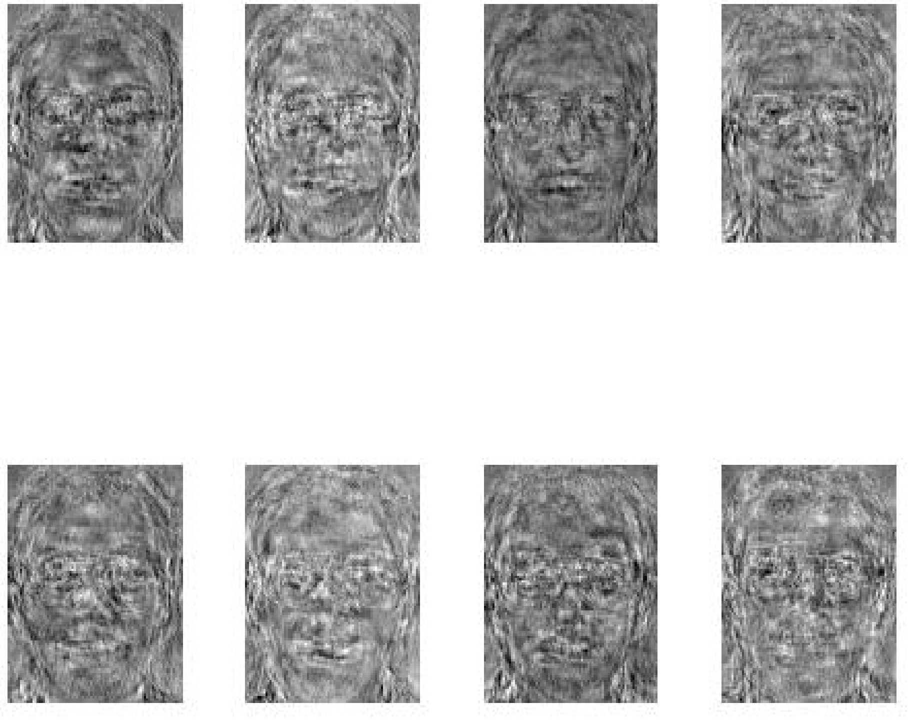

Figure 2.6 Eight feature vectors from ICA Architecture 2.

Chapter 3.

Factor Analysis

In this chapter, I will examine yet another multivariate technique. Furthermore, I will explore the uses of this technique as applied to the FERET facial database.

Factor Analysis can be described as a relatively heuristic methodology. This is because, with a factor analysis, it is the interpretation of the various factors that is important to the researcher. How can we interpret the various factors in a given model to represent the physical data in a meaningful manner? The essential purpose of an FA is to describe the covariance/correlation relationships among many observations in terms of a few underlying but unobservable random quantities called *factors* [39].

3.1 **PCA** Vs **FA**

As a multivariate statistical tool, factor analysis is similar to PCA in that it seeks to replace the *n random variables* with *m (m < n) random variables*. Like PCA, the data for a factor analysis consists of *p* observations, each with *n* variables. Unlike PCA, which is used to find the optimal way of combining variables such that the total variance of the variables is accounted for by fewer new variables or components where each successive component accounts for a maximum variance while being uncorrelated with the other components, factor analysis may be used to identify the underlying structure of the variables and to estimate scores or loadings to measure latent factors.

The factor loadings are a measure of how well the variables agree with the computed factors. Another difference between the two approaches is that with PCA all of the observed variance is mapped onto the uncorrelated components, while in factor analysis the shared variance or correlation is analyzed and explained in terms of common factors and a unique or error factor [39]. Moreover, unlike PCA, factor analysis is based on a postulated model. In this dissertation, I will be concerned with the *Common Factors* model [40].

Factor Analysis can be traced back to the work of Spearman [41]. While studying the correlation between student's test scores, he noted that the relationships could be explained by a simple model suggesting a two-factors model: an overall intelligence factor and a test specific factor. This simple model was later expanded to allow for more factors. Specifically, each test result was postulated to be due to several common factors and a test specific factor.

One of the reasons for choosing to perform an FA is for the attainment of a parsimonious description of the observed data. A successful FA is able to represent a set of observations through a linear transformation to a smaller set of new variables or factors. A satisfactory resolution will yield *factors* that convey all the essential information in the original variables [41]. This new representation is accomplished through an analysis of the correlation or covariance of the observable variables. These *factors* can be thought of as implicit traits or unobservable patterns buried in the dataset. They afford the researcher an alternative description. Factor analysis, like most methodologies, can take several forms. The one I will be concerned with is the *Common Factors* variety. In essence, it looks to represent the observable set of variables with a set of common factors that are shared by all the observable variables plus a unique factor that is specific to each variable. In this dissertation, I will be looking to find gender, facial hair and facial expression factors in three datasets [5], [42].

In his initial two-factor model, Spearman [41] observed that the pattern exhibited by the correlations followed a simple model. The model he used to explain the data was of the form

$$X_i = a_i F + e_i \quad , \tag{3.1}$$

where, after standardizing the raw observed data (standard deviation of one and a zero mean), X_i is the i*th* test score, *F* is a factor with zero mean and standard deviation of one, a_i is a constant (also known as the factor loading), and e_i is the specificity or the part of X_i that is specific to the ith test (or variable) only. From the above model, the variance of X_i is

$$\begin{aligned} Var(X_i) &= Var(a_i F + e_i) \\ &= Var(a_i F) + Var(e_i) \\ &= a_i^2 Var(F) + Var(e_i) \\ \Rightarrow \quad 1 &= a_i^2 + Var(e_i) \end{aligned} \tag{3.2}$$

From this, we see that the square of the factor loading is the part of the variance that is accounted for by the factor *F*. The general factor analysis model for *m* multiple factors, and a specific factor is given by

$$X_i = a_{i1} F_1 + a_{i2} F_2 + \cdots + a_{im} F_m + e_i \quad , \tag{3.3}$$

where, again, X_i is the standardized and centered i*th* score, the *F*'s are the *m* factors, a_i 's are the factor loadings, and e_i is the specificity which is uncorrelated with any of the common factors. Similarly, the variance can be seen to be

$$\begin{aligned} Var(X_i) &= a_{i1}^2 Var(F_1) + a_{i2}^2 Var(F_2) + \cdots + a_{im}^2 Var(F_m) + Var(e_i) \\ \Rightarrow \quad 1 &= a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + Var(e_i) \end{aligned} \tag{3.4}$$

where

$$a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 , \tag{3.5}$$

is called the communality of X_i is the part of the variance that is related to the factors. These factors F_i are not unique. New factors can be obtained from these by a linear combination. For example, we can create new factors W such that

$$\begin{aligned} W_1 &= d_{11} F_1 + d_{12} F_2 + \cdots + d_{1m} F_m \\ W_2 &= d_{21} F_1 + d_{22} F_2 + \cdots + d_{2m} F_m \\ &\vdots \\ W_m &= d_{m1} F_1 + d_{m2} F_2 + \cdots + d_{mm} F_m \end{aligned} \tag{3.6}$$

These new factors are a result of a rotation of the previous factors. This is often done to help the researcher reach a more meaningful interpretation of the loadings across the various factors [43]. The factor rotation can be orthogonal or oblique. In an orthogonal rotation, the new factors will be uncorrelated. With oblique rotations, the new factors are correlated. A commonly used rotation is the Varimax rotation [44]. This rotation is based on maximizing the variance of the squares of the loadings. By doing so, all variables will have their loadings close to zero or one. This in turn will facilitate an easier and more intuitive interpretation of the factors. The correlation between two variables X_i and X_j is

$$r_{ij} = a_{i1} a_{j1} + a_{i2} a_{j2} + \cdots + a_{im} a_{jm} . \tag{3.7}$$

This shows that two variables are highly correlated if they have high (positive or negative) loadings on the same factors.

3.2 Common Factor Analysis

Common Factor Analysis, as opposed to a full component model, is a special type of factor analysis that seeks to represent the p observations, each consisting of n variables, in terms of m ($m < n$) common factors and p specific factors. In general, FA is similar to PCA. However, unlike PCA, where the new variables are represented by a linear combination of a set of *n* uncorrelated components, the factors in an FA need not be uncorrelated. In fact, in common FA, the *m common factors* can be correlated.

The goal of common factor analysis is to find a set of new variables such that the overall variance of each observation can be explained by way of a common component(s) and a unique one [29]. With a common Factor analysis, it is the covariation among the variables that is of interest. In 1904, Spearman [41] was the first to stress the importance of this model. In carrying out a study to measure the general intelligence, he hypothesized that the observed correlation between variables such as school performance, original thinking and arithmetic reasoning was due to a common intelligence factor and that the unaccounted variance of each variable was it's specific variance. In other words, Spearman's formulation stated that

$$\text{Total Variance} = \text{Common Variance} + \underbrace{\text{Unique Variance}}_{\text{(Specific + Error Variance)}}$$

The general form of an FA model is as in (3.3). Where X_i is the new ith variable, a_i is a constant (also known as a factor loading), the F_i 's are the m common factors and e_i is the variable's specific variance. In matrix notation, we have

$$\boldsymbol{X}-\boldsymbol{\mu}=\boldsymbol{L}\boldsymbol{F}+\boldsymbol{\epsilon} \quad , \tag{3.8}$$

where

$\boldsymbol{X}$ = Observable random vector,

$\boldsymbol{\mu}$ = mean,

$\boldsymbol{L}$ = matrix of factor loadings,

$\boldsymbol{\epsilon}$ = specific variances,

The estimated vectors **F** and $\boldsymbol{\epsilon}$ must be independent [62]. Let

$$E(\mathbf{F}) = E(\boldsymbol{\epsilon}) = 0; \;\; Cov(\mathbf{F}) = \mathbf{I}; \;\; Cov(\boldsymbol{\epsilon}) = \boldsymbol{\psi}, \tag{3.9}$$

where $\boldsymbol{\psi}$ is a diagonal matrix. The basic assumptions of the common factor model stem from the following axioms [45].

1. A variable can be partitioned into two parts: a common and a specific part.

   $X_i = C_i + V_i$ C is the common part component and V is the specific component.

2. $E\{V_i X_k\} = 0 \qquad i \neq k$

   $E\{V_i V_k\} = 0 \qquad i \neq k$

3. $E\{V_i C_k\} = 0 \qquad i \neq k$

4. $E\{V_i C_i\} = 0$

5. $E\{C_i C_k\} = r_{ik} \qquad i \neq k$

The correlation between two variables is due to the common factor portion. This is called the communality. It can be written as:

$$E\{C_i^2\} = h_i^2 \quad . \tag{3.10}$$

or as in (3.5). The part due to it's specificity is the specific variance

$$E\{V_i^2\} = u_i^2 \quad . \tag{3.11}$$

3.3 Procedure for a Factor Analysis

Which extraction method to choose will depend on the available data, and/or purpose of the analysis. Common extraction methods include: Maximum Likelihood (ML), the method of Principal Factors, and the principal component method. I will, primarily, make use of the maximum likelihood method [46].

An important question that must be answered in performing a factor analysis is how many factors should you retain? Again, this is entirely dependent on the researcher and the data. Some general guidelines used are: number of eigenvalues greater than or equal to one (Principal Components Analysis method), scree test/scree plot, percentage of variance, and hypothesis testing [43]. Some methods do not require a postulated number of factors; instead they are concerned with finding a model with enough factors to best reproduce the correlation matrix of the original data.

Finally, in doing a factor rotation, there are two alternatives. There are the orthogonal and the oblique rotations/transformations. Rotations, a geometrical transformation of the axes of factors, are used to increase the interpretability of the results. The orthogonal rotation keeps the factors uncorrelated while trying to increase their interpretability. An oblique rotation allows the new factors to be correlated. A popular orthogonal method is the *Varimax* transformation [44]. It constitutes maximizing the variance of the loadings within factors across the observations. Another method within the orthogonal methods is the *Quartimax* method, which seeks to maximize the variance of the variables across the various factors [39]. Here, I will restrict my attention to the orthogonal methods, specifically the Varimax transformation. I have decided to use this because I am interested in using the loadings on the factors as a classifier.

3.3.1 Principal Component Factor Analysis

The Principal Component Factor Analysis method was developed to facilitate the reduction of a large body of variables to a few [43]. It was originally proposed by Karl Pearson [29], and later developed for use in FA by Hotelling [30]. The model is given by

$$X_i = a_1 F_1 + a_2 F_2 + \cdots + a_N F_N \qquad (i=1,2,\ldots p) \quad . \tag{3.12}$$

This model [43] uses the principal components from PCA as the initial factors. These are rotated until the desired factors are found. Similar to PCA, for *N* variables there will be *N* principal components. These are just linear combinations of the original variables, so that

$$\begin{aligned} Y_1 &= a_{11} X_1 + a_{12} X_2 + \cdots + a_{1\mathrm{L}} X_L + \cdots + a_{1\mathrm{p}} X_p , \\ Y_2 &= a_{21} X_1 + a_{22} X_2 + \cdots + a_{2\mathrm{L}} X_L + \cdots + a_{2\mathrm{p}} X_p , \\ &\vdots \\ Y_p &= a_{pl} X_1 + a_{p2} X_2 + \cdots + a_{pL} X_L + \cdots + a_{pp} X_p , \end{aligned} \tag{3.13}$$

where the a_{ij} 's are obtained from the eigenvectors of the original correlation matrix. For the initial factors (like in PCA), only *L (L<p)* of the principal components are retained. This can be written (inversely) as

$$\begin{aligned} X_1 &= a_{11} Y_1 + a_{21} Y_2 + \cdots + a_{L1} Y_p + \cdots + e_1, \\ X_2 &= a_{12} Y_1 + a_{22} Y_2 + \cdots + a_{L2} Y_p + \cdots + e_2, \\ &\vdots \\ X_L &= a_{1\mathrm{p}} Y_1 + a_{2\mathrm{p}} Y_2 + \cdots + a_{Lp} Y_p + \cdots + e_L, \end{aligned} \tag{3.14}$$

Where the *e*'s are linear combinations of the discarded principal components. The next step is to scale the components to unity. This is followed by the rotation step [43]. In our case, I will use a Varimax rotation, yielding

$$\begin{aligned} X_1 &= d_{11} Z_1 + d_{21} Z_2 + \cdots + d_{L1} Z_L + \cdots + e_1, \\ X_2 &= d_{12} Z_1 + d_{22} Z_2 + \cdots + d_{L2} Z_L + \cdots + e_2, \\ &\vdots \\ X_L &= d_{1\mathrm{p}} Z_1 + d_{2\mathrm{p}} Z_2 + \cdots + d_{Lp} Z_L + \cdots + e_p, \end{aligned} \tag{3.15}$$

where the d_{ij} are the scaled a_{ij} and the Z_i are the rotated factors. For data reduction, only a few of the components are retained [43]. Typically, the retained components will account for most of the variation in the data [39]. However, in order to reproduce the original correlation among the variables, all components will be needed.

3.3.2 Principle Factors Method.

Unlike the method of principal component analysis described in section 3.3.1, the principal factors method requires an estimation of the specific variances a priori [40]. These specific variances allow for the construction of the reduced correlation matrix. The reduced correlation matrix is the original correlation matrix with the specific variances subtracted from the diagonal elements. This reduced correlation matrix will have communality estimates as its diagonal elements . The decomposition of this adjusted (reduced) correlation matrix results in the common factors and unique factors. In fact, when using ones as the diagonal elements, this method reverts to the principal component method. Given the number of factors, the principal factor method will extract factors that account for the maximum variance [40]. These factors are extracted by way of a characteristic roots and vector analysis of the association matrix. In most cases the association matrix is either the correlation or covariance matrix. Like PCA, the first factor extracted will have the maximum variance. The second factor is extracted in such a way that it will be uncorrelated with the first while having the second highest variance [43]. The rest (up to the number of specified factors) of the factors are extracted in a similar manner. Thus, all the extracted principal factors will be uncorrelated. In contrast to the method of principal components, this method seeks to maximally (in a least square sense) reproduce the correlation matrix using *m common factors (m < N)* and a *unique factor*. Thus all the *n* variables will be represented by:

$$Y_j = a_{j1} F_1 + a_{j2} F_2 + \cdots + a_{jm} F_m + d_j U_j \qquad (j = 1, 2, \cdots, p) \tag{3.16}$$

or

$$Y_j = \sum_{k=1}^{m} a_{jk} F_k + d_j U_j \qquad (j = 1, 2, \cdots, p; i = 1, 2, \cdots, n), \tag{3.17}$$

where the common factors account for the correlation among the variables and the unique factor represents the remaining variance of each variable. As before, the factor coefficients are called the loadings. The square of these coefficients or loadings, a^2_{ik}, are an indicator of how much each factor contributes to the communality of a particular variable. The sum of the square of the loadings, $a_{j1}^2+a_{j2}^2+\cdots+a_{jm}^2,$ is the communality of the variable X_i. This is the part of its variance that is related to the common factors. The coefficients are also a measure of the correlation between variables. The correlation between variables can be expressed as the sum of the products of coefficients [39], *i.e* .

$$r_{jk}=\sum_{p=1}^{m} a_{jp}a_{kp}. \qquad (j,k=1,2,\cdots,n). \tag{3.18}$$

Note that $r_{jk} = r_{kj}$ and that the communality is simply r_{jj}. The principal factor method seeks to find factors such that the sum

$$V_1=a_{11}^2+a_{21}^2+\cdots+a_{n1}^2 \tag{3.19}$$

is maximum subject to the condition

$$r_{jk}=\sum_{p=1}^{m} a_{jp}a_{kp}. \tag{3.20}$$

Maximization is achieved by use of Lagrangian multipliers [47].

The first step in the principal factor method is the determination of the coefficients for the first factor under the constraint of maximal communality. This maximization of (3.18) with the constraints of (3.19) results in a system of n equations for the unknowns $a_{j1},(j=1,\cdots,n)$ [40]. A necessary and sufficient condition for a non-trivial solution is the vanishing of the determinant of the reduced correlation matrix [40]. *i.e.,*

$$Det\begin{bmatrix}(h_1^2-\lambda_1) & r_{12} & \cdots & r_{1n}\\ r_{21} & (h_2^2-\lambda_2) & & r_{2n}\\ \vdots & & & \vdots\\ r_{n1} & r_{n2} & \cdots & (h_n^2-\lambda_n)\end{bmatrix}=0. \tag{3.21}$$

The expanded form of (3.21) is the *characteristic* equation. The roots of the characteristic equation are the eigenvalues and their associated solutions are the eigenvectors. The eigenvectors corresponding to the largest eigenvalue, $\lambda_{1,}$ of the reduced correlation matrix R are the coefficients or loadings of the first factor $F_{1.}$ The coefficients of the second factor F_2 are the eigenvectors associated with the largest eigenvalue of the new reduced correlation matrix. This new reduced correlation matrix is the original reduced correlation matrix sans the contributions of the coefficients of the first eigenvectors.

$$R_1=R-a_1 a_1' \tag{3.22}$$

where a_1 is the vector of coefficients for F_1. The largest eigenvalue of the new reduced correlation matrix will have a corresponding eigenvector consisting of the coefficients of the second factor F_2 [47]. In fact, this eigenvalue and its corresponding eigenvector can be obtained directly from the original reduced correlation matrix as the second largest eigenvalue and its associated eigenvector. The rest of the eigenvalues and their associated eigenvectors are obtained in a like manner [40].

When doing a full component analysis, that is starting with unities as the diagonal elements of the correlation matrix, there will be *n* eigenvalues for a full rank matrix. In the case of a principal-factor analysis (with communalities in the diagonal) there will be *m* eigenvalues. All these computations can be achieved by use of electronic computers.

Jacobi's [48] work has been found to be an effective method in the computations of these eigenvalue-eigenvector pairs. The decomposition of the correlation matrix as the product of its eigenvalues and eigenvectors can also be explained by way of the spectral theorem [49], which states that the symmetric matrix R can be diagonalized by means of an orthogonal transformation.

3.3.3 Maximum Likelihood Method.

Unlike the method of principal factors, the Maximum Likelihood method does not require the estimated communality values. Instead, it requires the number of postulated factors. The method of maximum likelihood generally seeks to find the factor matrix that will minimize the residual matrix in a least square sense. In the 1940's, Lawley [50] developed a statistical basis for measuring the effectiveness of a factor analysis. His test was based on the use of the method of maximum likelihood for the estimation of the factors. Starting with an assumption of the number of common factors *m*, and the assumption of a multivariate normal distribution, Lawley's method estimates the universal factor loadings from the test sample. The effectiveness of the model is measured by a *Chi-Squared* test of significance. Under the assumption that the samples are normally distributed, Wishart [51] determined the distribution function of the elements of the covariance matrix as

$$F = K|\Sigma|^{\frac{-1}{2}(p-1)} |\boldsymbol{S}|^{1/2(p-n-1)} e^{-1/2(p-1)\sum_{j,k=1}^{n} \sigma^{jk} s_{jk} \prod_{j<k=1}^{n} ds_{jk}}, \tag{3.23}$$

where *K* is a constant involving the sample size, *N*, and number of observations *n* and where $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ are the sample and population covariance matrices, respectively. This distribution function is the likelihood function. The task is now to find the estimates $\boldsymbol{A}$ (matrix consisting of the common factor loadings) and $\boldsymbol{D}$ (diagonal matrix of the unique variances) such that

$$\boldsymbol{\Sigma}=\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{D}^2 \tag{3.24}$$

will maximize the likelihood function. The maximization is achieved by equating the partial derivatives of (3.21) with respect to the a_{jp} and the diagonal matrix d_j to zero and solving the resulting equations. The solution obtained by Lawley [50] is:

$$\boldsymbol{P}=\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}'+\tilde{\boldsymbol{D}}^2, \tag{3.25}$$

$$\tilde{\boldsymbol{A}}=\tilde{\boldsymbol{P}}\boldsymbol{R}^{-1}\tilde{\boldsymbol{A}}, \tag{3.26}$$

$$\tilde{\boldsymbol{D}}^2=\boldsymbol{I}-diag(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}'), \tag{3.27}$$

where $\boldsymbol{P}$ is the population correlation matrix with the entries of the main diagonal normalized to unity, $\tilde{\boldsymbol{P}}$ is the estimated correlation matrix, and $\boldsymbol{R}$ is the sample correlation matrix. In order to simplify the numerical maximization of (3.23), the population correlation matrix $\boldsymbol{P}$ is assumed to be equal to the sample correlation matrix $\boldsymbol{R}$ [39].

3.4 Preliminary Results Using Factor Analysis

In this dissertation, I performed a factor analysis on the FERET image databases. I ran the analysis several times on different datasets. The aim was to include many different subjects in our datasets. This was done to test the robustness of some of the results and to check for consistency of the results. Initially, the number of factors was varied for each dataset. This was done in a rather heuristic manner in keeping with the exploratory nature of the work at this stage. In fact, a goal of this work is to develop a deeper, and more intuitive understanding of the relationship between the number of factors, the type of dataset, and the desired interpretation of the analysis. Fig. 3.1 shows one such dataset [5].

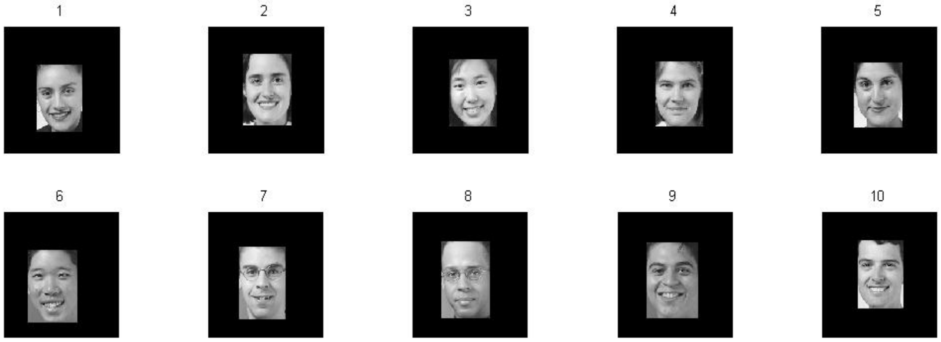

Figure 3.1 Cropped facial images of female (top row), and male (bottom row)

I ran a factor analysis on the dataset shown in Fig. 3.1. Initially, I used two factors. A Varimax rotation was applied to the factors. Fig.3.2 shows the corresponding loadings of the two factors for the ten subjects.

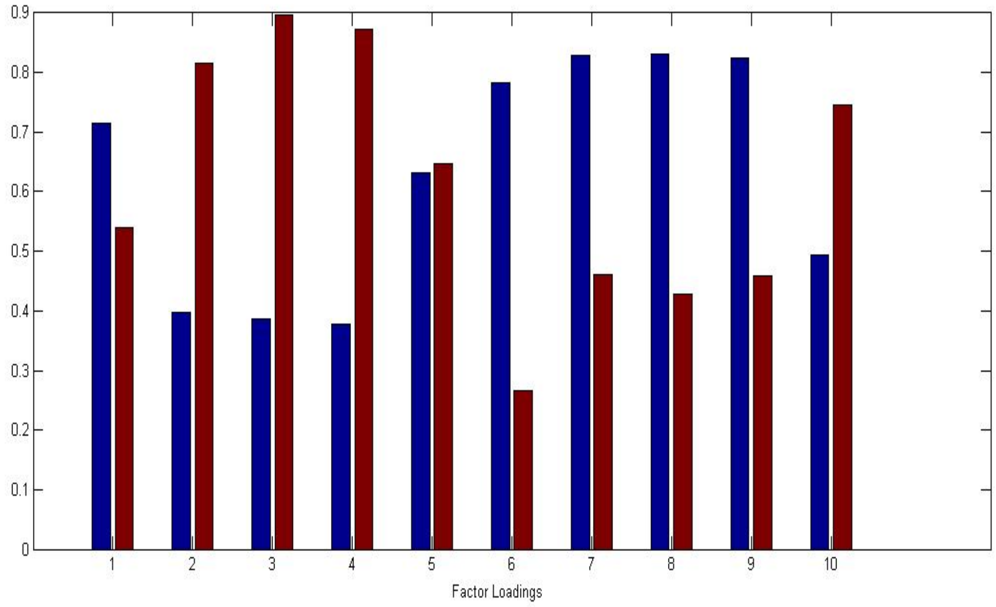

Figure 3.2 Factor Loadings (two factors) for images if Fig 3.1

Looking at the factor loadings, I noticed that these particular factors seemed capable of capturing the gender of the subject. Notice that the first five subjects (with the exception of the first subject) exhibit higher loadings on the second factor than the first factor. This is the opposite with regard to the male subjects (with the exception of the last one. What does this mean? Can we interpret the factors to mean gender? Part of this work is an attempt to be able to answer these questions in a more rigorous manner. I ran another factor analysis on the above dataset using three factors. Fig. 3.3 is a graph of the three loadings for every subject.

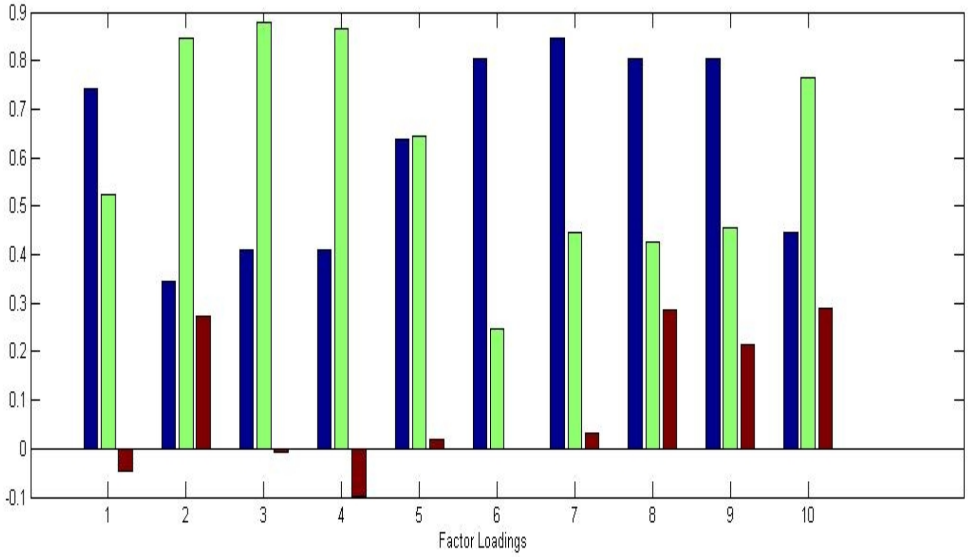

Figure 3.3 Factor Loadings (three factors)

Again, we observe that the first two factors seem to be exhibiting the same behavior as when there were only two factors.

I ran another analysis against a different dataset. Fig. 3.4 is the image dataset, followed by the analysis using two, three, four, and five factors.

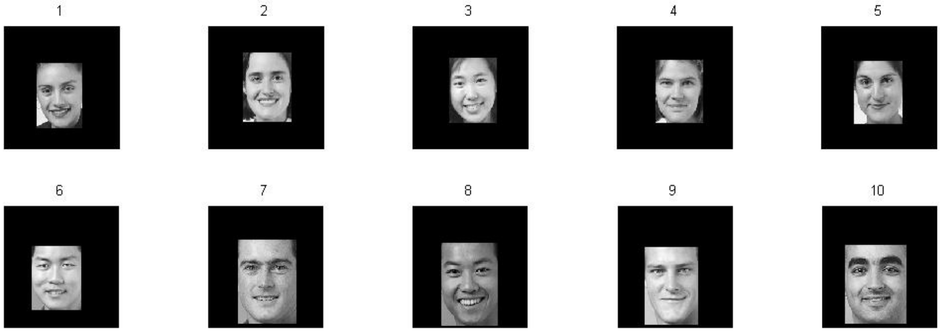

Figure 3.4 Facial Dataset

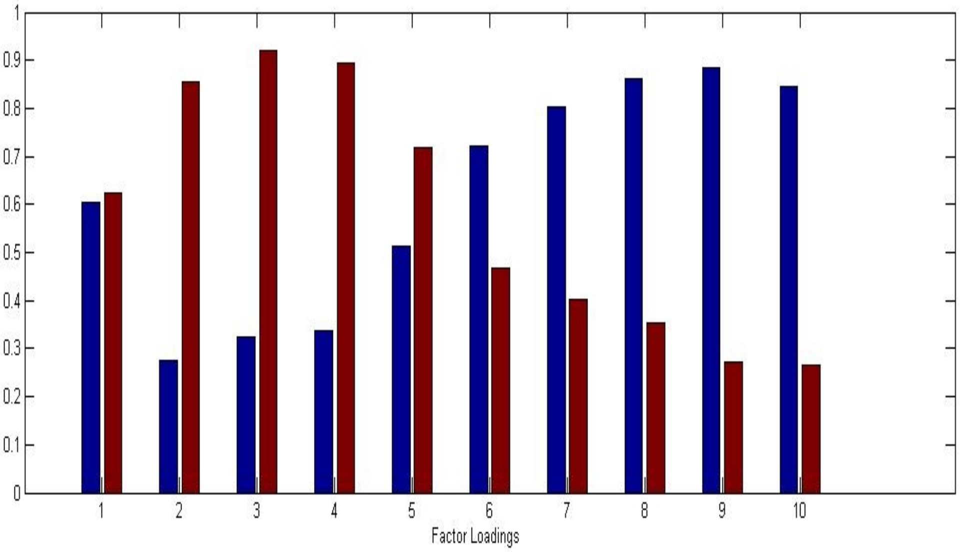

Figure 3.5 Two factors using dataset of Fig. 3.4

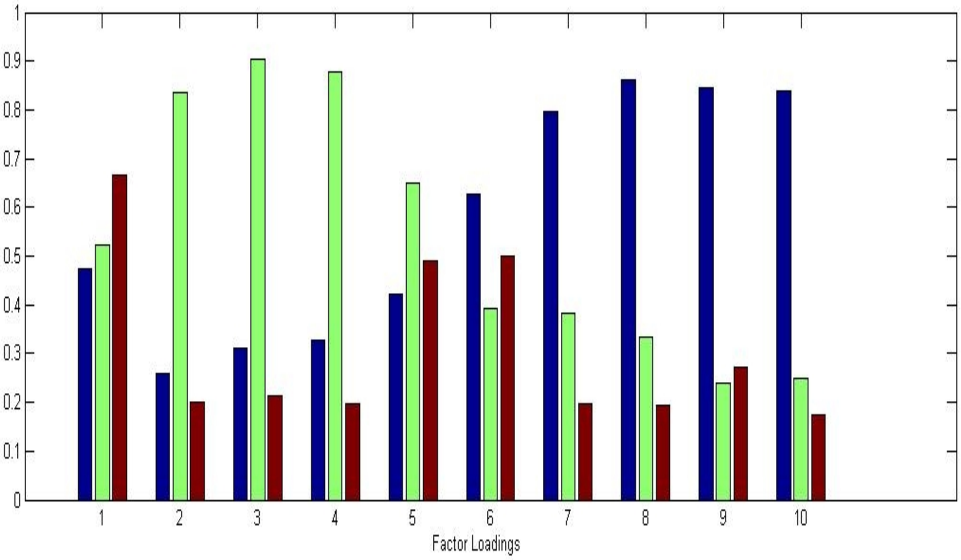

Figure 3.6 Three factors using dataset of Fig. 3.4

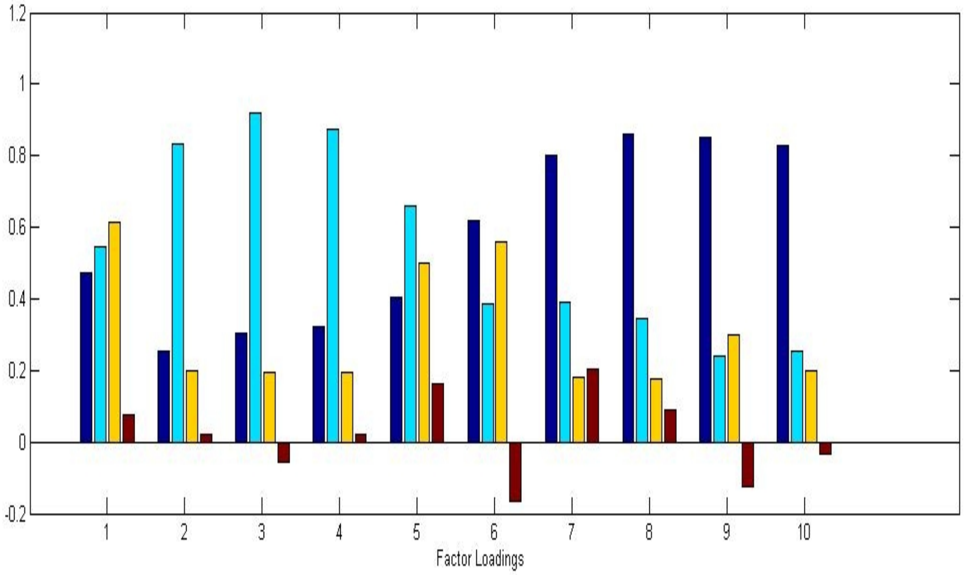

Figure 3.7 Four factors using dataset of Fig. 3.4

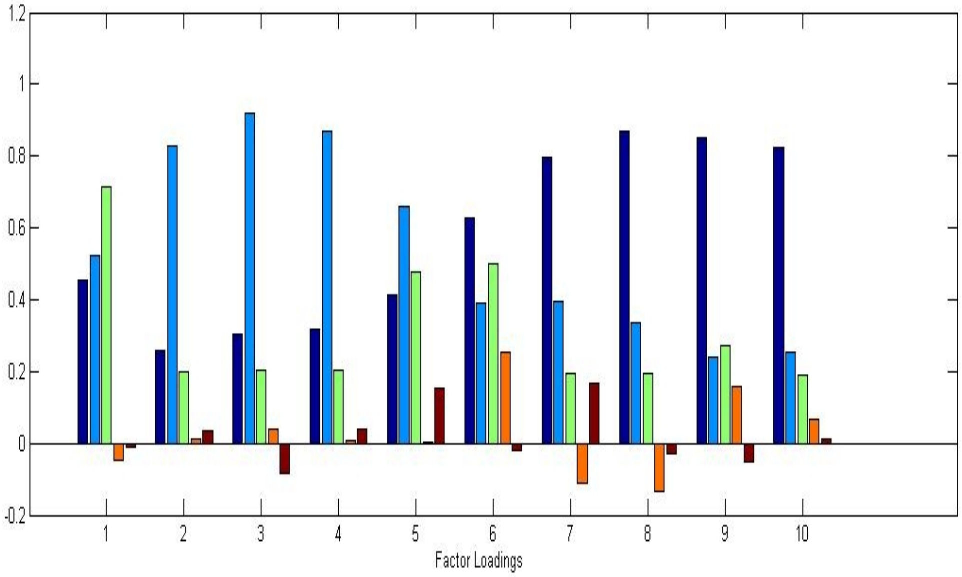

Figure 3.8 Five factors using dataset of Fig. 3.4

All the analysis seems to agree that the first two factors can be interpreted to represent gender. This trend seems to hold even for multiple factors.

I performed another analysis on a different dataset [32]. In addition to gender, this new dataset introduces the element of facial hair. Fig. 3.9 shows the dataset. With this analysis, we see that subjects with facial hair seem to have low loadings on factor 3 in Fig. 3.11, and low loadings on factor 4 in Fig. 3.12. Moreover, we can still see the gender classification coming through the first two factors.

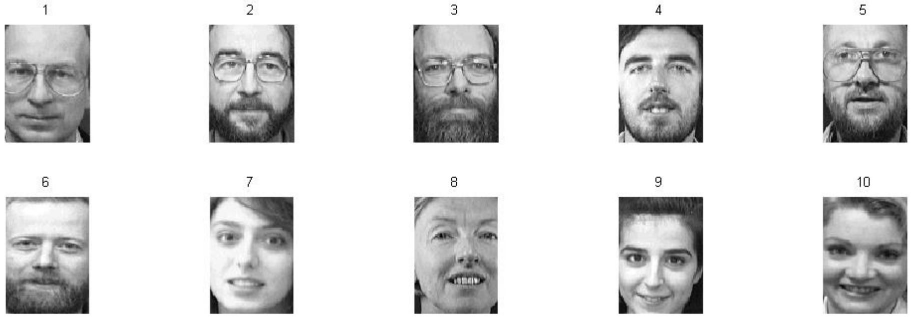

Figure 3.9 Facial Images dataset

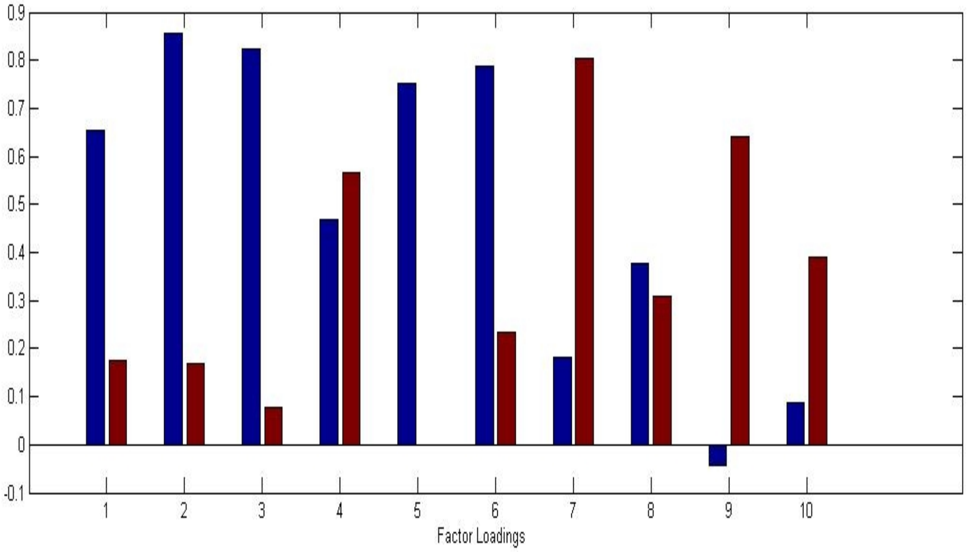

Figure 3.10 Two factors using dataset from Fig. 3.9

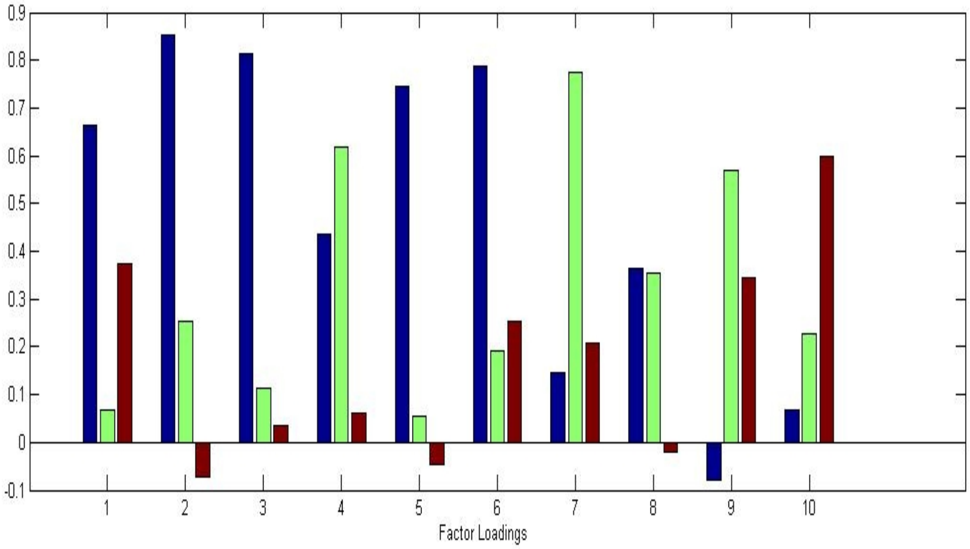

Figure 3.11 Three factors using dataset from Fig.3.9

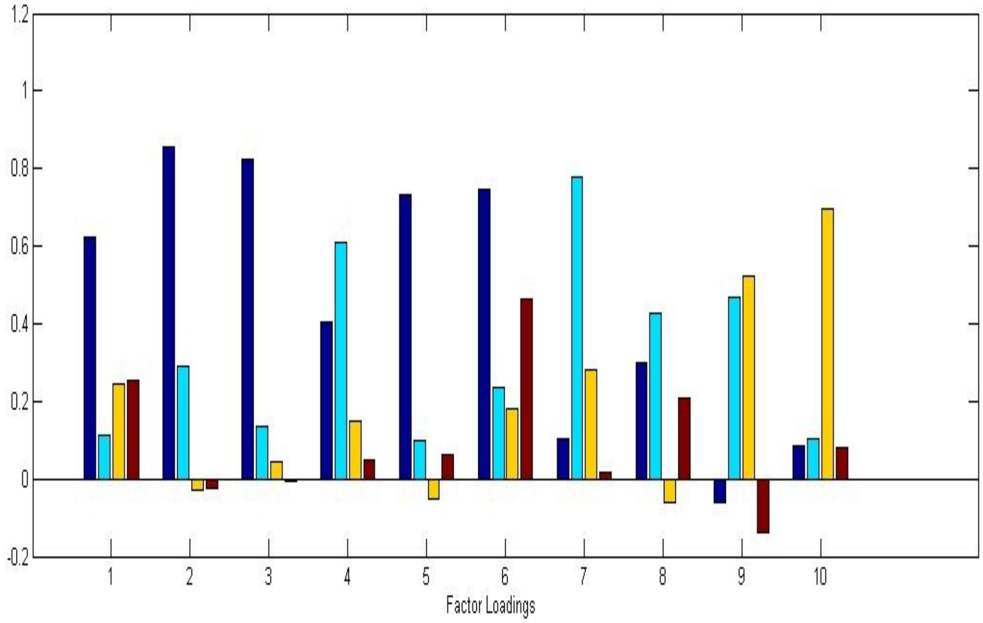

Figure 3.12 Four factors using dataset from Fig. 3.9

Chapter 4.

Gender Classification with Factor Analysis

In this research, I will perform a common factor analysis on a set of variable images [5],[42]. Each image will be considered as an independent composite variable. In the analysis, I used a total of ten images. Five male and five female subjects. The first step is to construct a correlation or covariance matrix. Each n by m image will be represented by a vector of length nm. In this dissertation, I will work with standardized data, and therefore the correlation matrix. Using images from the FERET database [5], I randomly selected ten images shown in Fig. 4.1 and 4.2 to construct our correlation matrix. All images were cropped and resized. Each of these images will be regarded as an observation with nm variables. This will result in a correlation matrix of size 10x10.

After computing the correlation among the ten facial images, an FA, using the method of maximum likelihood with two common factors was utilized in the estimation of the factors and the loadings.

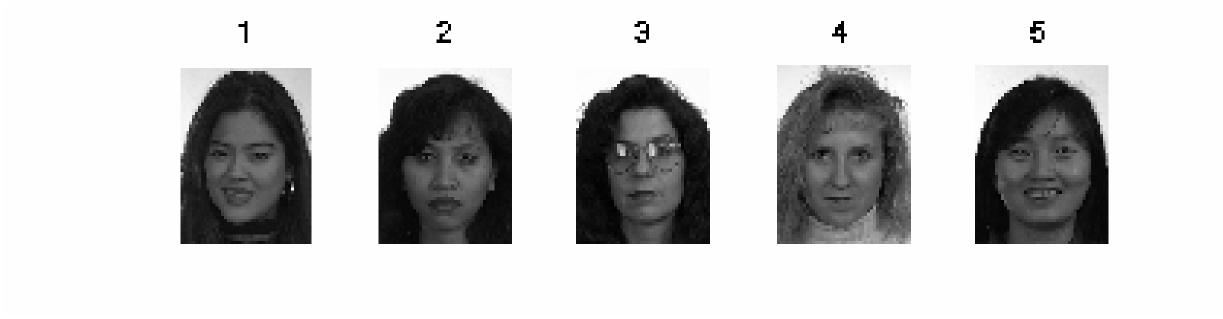

Figure 4.1 Female images

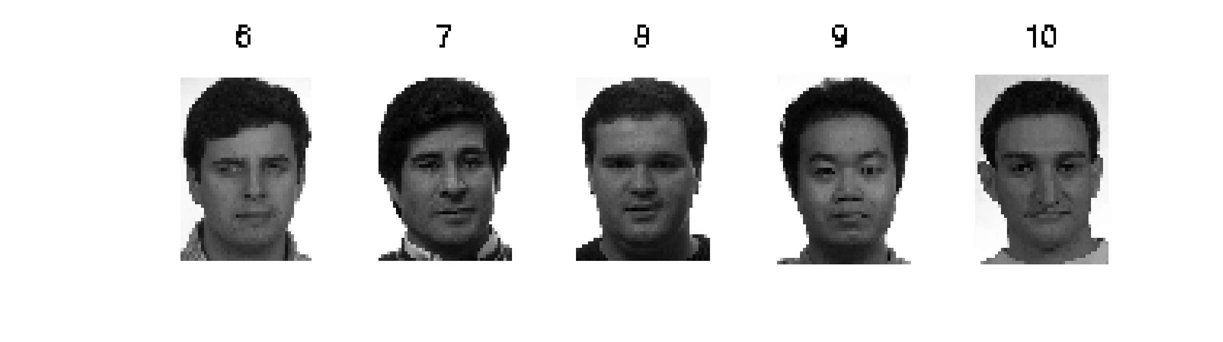

Figure 4.2 Male images

The maximum likelihood estimates (for *p* observations of *n* x *1* vectors) are obtained by a numerical maximization of the joint likelihood function:

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma})=\prod_{i=1}^{p}\frac{\left|\boldsymbol{\Sigma}^{(-1/2)}\right|}{(2\pi)^{(n/2)}}\,\mathrm{e}^{[(-1/2)(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})]}. \qquad (4.1)$$

Subject to the uniqueness condition (to overcome the many possible transformations/solutions) that (4.2) be a diagonal matrix.

$$L^T \Psi^{-1} L = \Delta. \qquad \textit{(Diagonal matrix)} \qquad (4.2)$$

Using an orthogonal rotation (Varimax method [44]) of the estimated factors, we see in Fig 4.4 how the first five (female) variables load high on the first factor and low on the second. The opposite is true for the male images. They load higher on the first factor than on the second factor. It is this separation of the factors along with how the two groups load up on them that will enable us to create a linear classification rule. Fig.4.3 shows the two factors as images.

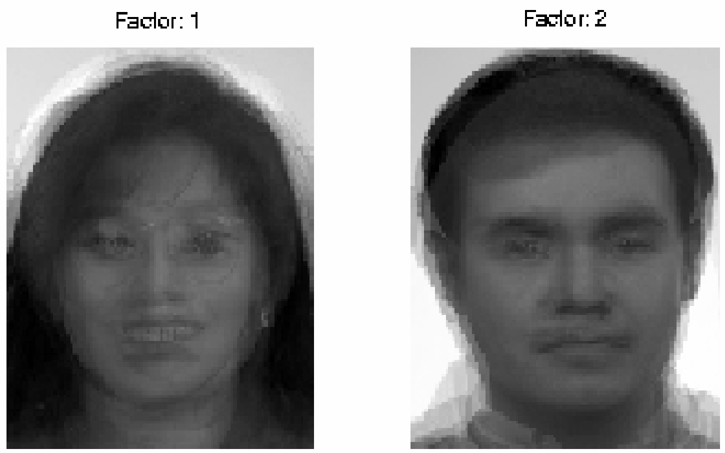

Figure 4.3 Images of the factors

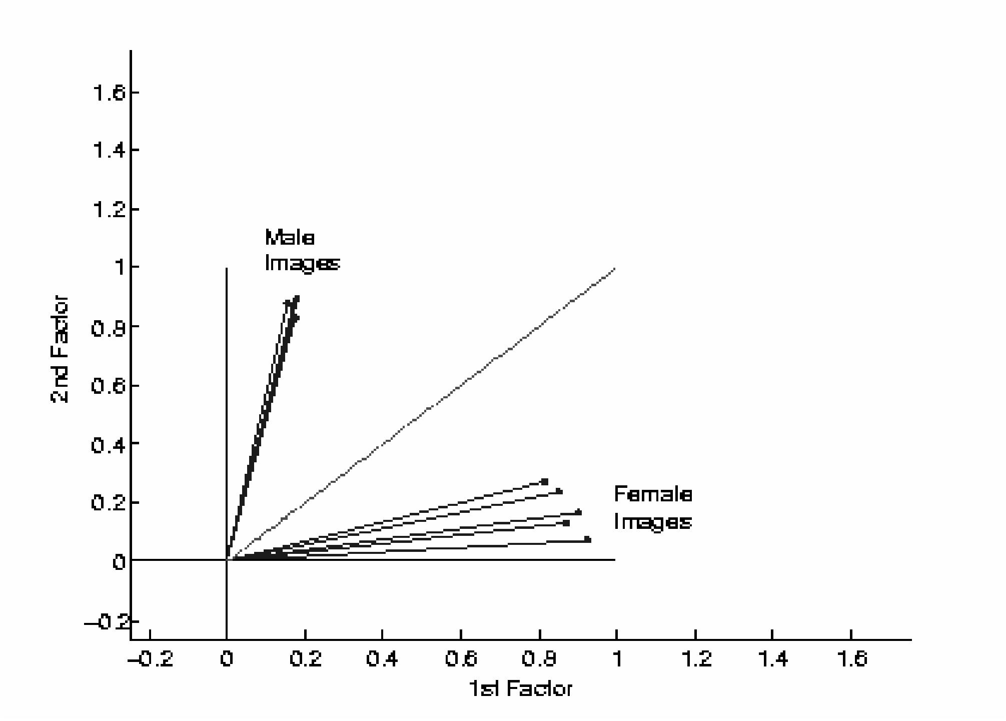

Figure 4.4 Loadings on the two factors

The next step is to estimate the factor scores from our analysis. These factor scores will be used to compute the new representation of our probe dataset in the factor space. The probe dataset or test cases consists of the images not used in the initial factor analysis. They will be tested against our linear classification rule. There are several methods of estimating these factor scores. I will use a W*eighted Least Squares (WLS)* method. To use the weighted least squares method, we must first know all the variables in the model

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon}. \tag{4.3}$$

Given that the loadings, ***L*****,** and specific variances, $\boldsymbol{\psi}$, are themselves estimates, we will treat them as though they are the true population parameters in the estimation of the factor scores. Bartlett [38], advocated that since the specific variances, Cov($\boldsymbol{\epsilon}$) = $\boldsymbol{\psi}$**,** need not be equal, the sum of the squares should be weighted by the reciprocal of their variances.

From the above model, we have:

$$\boldsymbol{\epsilon}=\boldsymbol{X}-\boldsymbol{\mu}-\boldsymbol{LF}\,. \tag{4.4}$$

The weighted sum of squares is

$$\sum_{i=1}^{p}\frac{\boldsymbol{\varepsilon}_i^2}{\boldsymbol{\Psi}_i}=\boldsymbol{\varepsilon}^T\,\boldsymbol{\Psi}^{-1}\,\boldsymbol{\varepsilon}=(\boldsymbol{x}-\boldsymbol{\mu}-\boldsymbol{Lf})^T\,\boldsymbol{\Psi}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}-\boldsymbol{Lf})\,. \tag{4.5}$$

To find the estimated factor scores, we find the estimates, $\hat{\boldsymbol{f}}$**,**that minimize the above. The solution is [45]

$$\hat{\boldsymbol{f}}=(\hat{\boldsymbol{L}}^T\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{L}})^{-1}\hat{\boldsymbol{L}}^T\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})\,. \tag{4.6}$$

Using these factor score estimates, the new representation of the probe images in terms of the two factors is computed. The loadings are used in the classification. Fig.4.5 is a plot of the loadings of the probe dataset with respect to the previously found factors. From the plot, we can see how the same gender images identify with their respective factors. It is this separation that allows us to create an identification boundary. In general, depending on how we choose to transform the factors (orthogonal vs oblique transformations), we will have different classification rules.

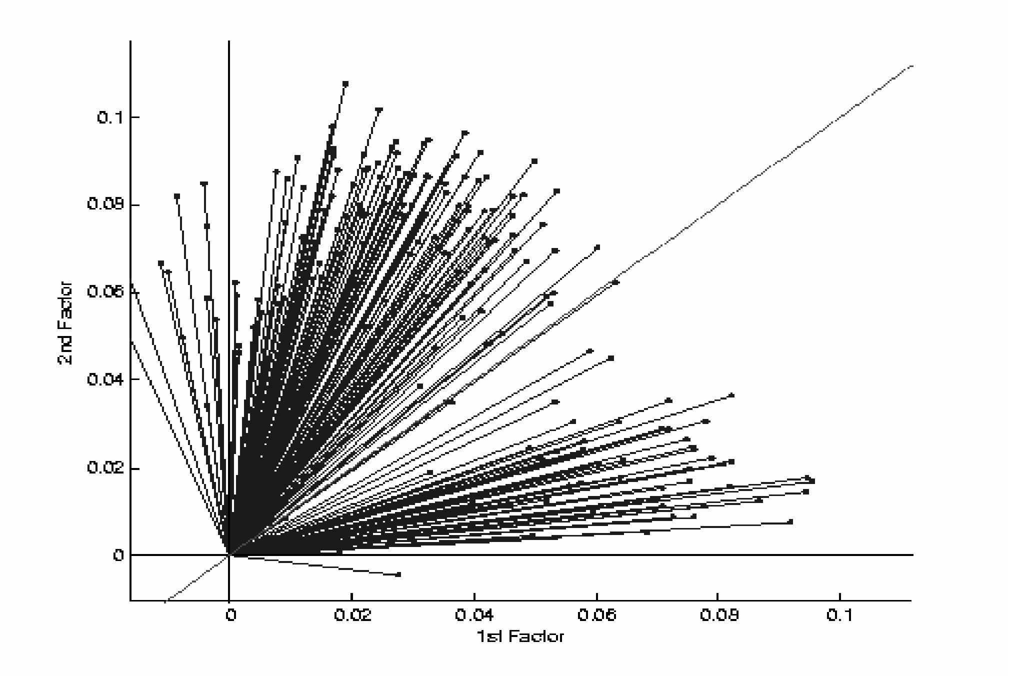

Figure 4.5 Loadings of probe images on the two factors

In this dissertation, I used images from the FERET [5] database. All the images were cropped and resized to 112x92 pixels. This was done to keep the sizes of the images from both databases [32], [5] consistent. Other than that, no further processing was applied. Using the $y=x$ line as the classification boundary and a test gallery of 200 images, I was able to achieve average classification rates of 90%.

An alternate method of using factor analysis for gender classification is to divide the facial images database into two groups: a known training group and a testing group. The training group will comprise an equal number of images of both genders. These individuals are chosen according to how well they set up the correlation matrix. We want a group such that the correlation matrix clearly exhibits a separation between the genders. An example of such a group is shown in Fig.4.6. The associated correlation matrix is shown below. Note that, with this particular group, we can clearly see that both genders correlate significantly with members of their respective gender class. The shaded areas in the matrix below show the high correlation among members of the same sex.

	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	$\mathbf{I_4}$	$\mathbf{I_5}$	$\mathbf{I_6}$	$\mathbf{I_7}$	$\mathbf{I_8}$	$\mathbf{I_9}$	$\mathbf{I_{10}}$
$\mathbf{I_1}$	1.000	0.781	0.742	0.751	0.843	0.338	0.328	0.390	0.343	0.367
$\mathbf{I_2}$	0.781	1.000	0.833	0.782	0.836	0.281	0.298	0.274	0.335	0.263
$\mathbf{I_3}$	0.742	0.833	1.000	0.763	0.791	0.236	0.276	0.254	0.284	0.228
$\mathbf{I_4}$	0.751	0.782	0.763	1.000	0.755	0.379	0.371	0.325	0.373	0.384
$\mathbf{I_5}$	0.843	0.836	0.791	0.755	1.000	0.211	0.214	0.228	0.213	0.223
$\mathbf{I_6}$	0.338	0.281	0.236	0.379	0.211	1.000	0.797	0.772	0.806	0.721
$\mathbf{I_7}$	0.328	0.298	0.276	0.371	0.214	0.797	1.000	0.754	0.792	0.740
$\mathbf{I_8}$	0.390	0.274	0.254	0.325	0.228	0.772	0.754	1.000	0.748	0.709
$\mathbf{I_9}$	0.343	0.335	0.284	0.373	0.213	0.806	0.792	0.748	1.000	0.805
$\mathbf{I_{10}}$	0.367	0.263	0.228	0.384	0.223	0.721	0.740	0.709	0.805	1.000

Figure 4.6 Correlation Matrix for the ten training Images.

Testing of an individual is performed by first creating an association matrix of the test image with the ten preselected images. An FA is carried out on the new correlation matrix. Determination of the gender of the test image is accomplished by examining the loadings of the new image on the two factors. The success rates are comparable to the previous method. Using a test bed of 168 facial images, I was are able to achieve a success rate of 89.88%.

Chapter 5.

Facial hair Classification with Factor Analysis

Notice in Fig. 3.12, that the factor analysis was able to capture and group the individuals according to gender. Moreover, it was also capable of grouping the individuals by facial hair. To further study this phenomena, a factor analysis was setup to investigate this behavior. Using the FERET database [5], male subjects, with and without facial hair, were chosen. Of the chosen subjects, some were selected to be the training set. The remainder were used as the test bed. Using forty individuals, twenty with facial hair and twenty without facial hair a correlation matrix was constructed. The forty training set individuals are shown below in Fig.5.1 and the test images gallery is shown in Fig.5.2.

After the correlation of the forty subjects was computed, two factors were extracted by means of a maximum likelihood factor analysis. The two extracted factors are shown in Fig.5.3.

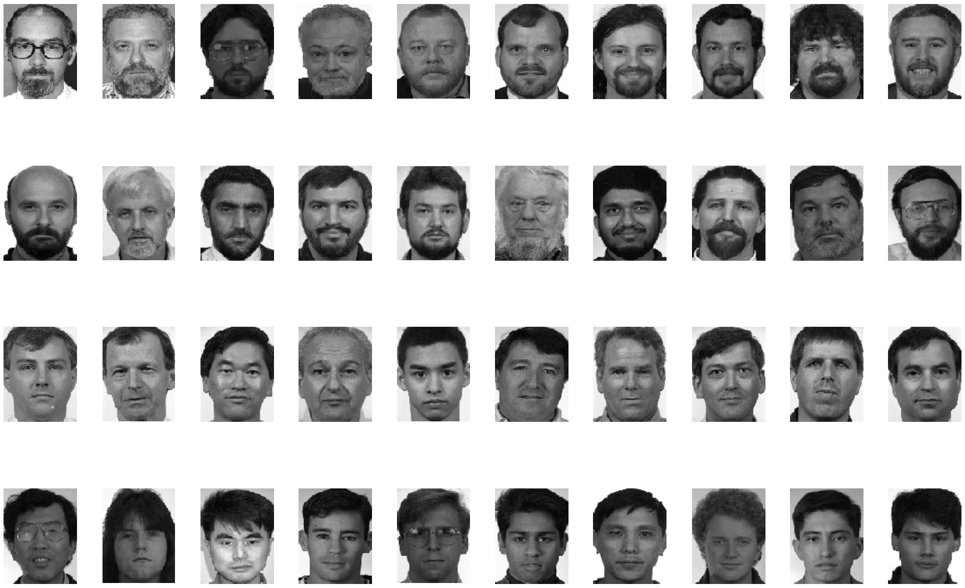

Figure 5.1 Training set images

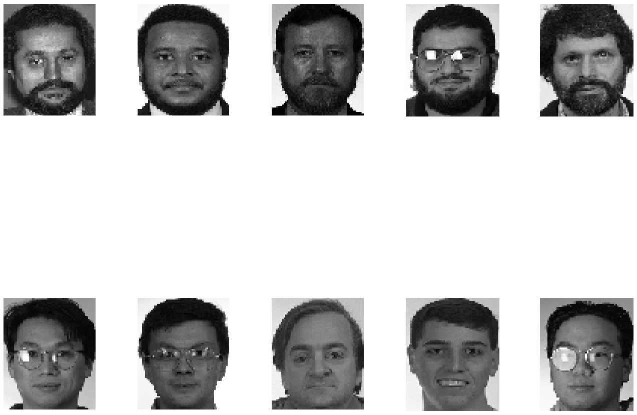

Figure 5.2 Images of the test database

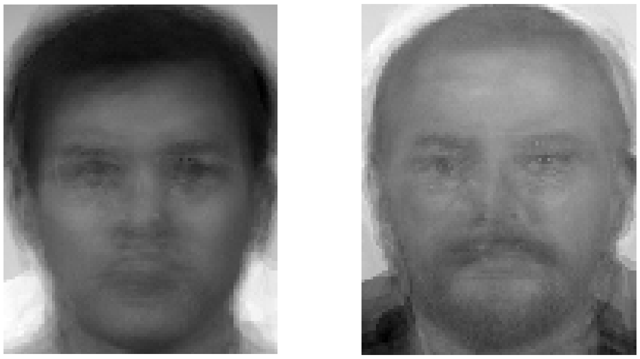

Figure 5.3. Image of the two Factors.

Subsequent to the analysis, the images are arranged by the magnitude of their loadings on the two factors. Fig.5.4 shows the top ten (by magnitude) images of the two factors. The first two rows are the ten images with the largest magnitudes on factor one, and the last two rows are the images with the largest loadings on factor two. Clearly, we can see that the first factor has images of non-facial hair subjects while the second factor is dominated with images of individuals with facial hair. Using this characteristic of the two factors, I was are able to to classify the images of the test gallery according to the magnitude of their loadings on the two factors. To test an image, the test image was appended to the training gallery prior to running an FA.

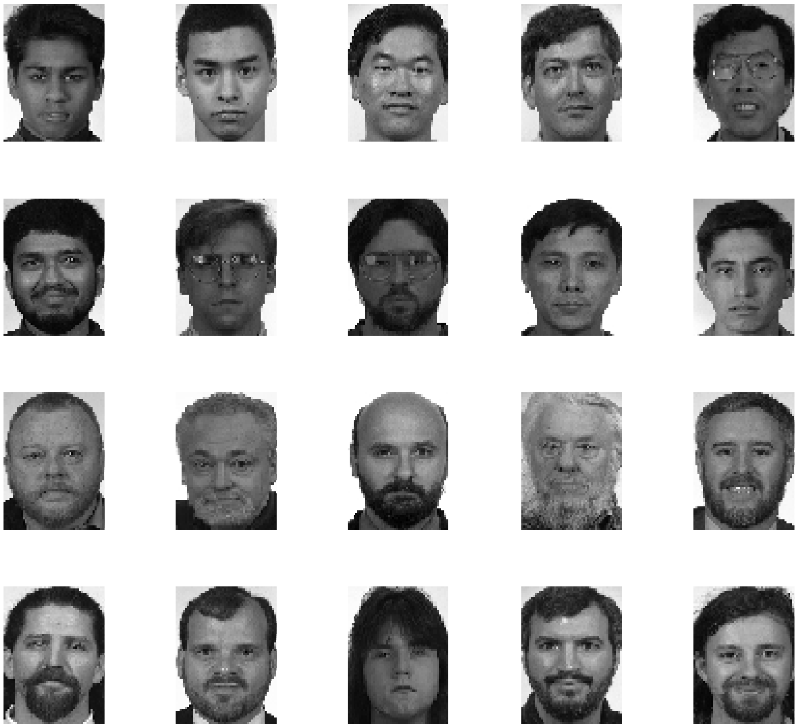

Figure 5.4 Images of the ten subjects with largest magnitudes on the two loadings

Depending on the magnitude of its loading on the two factors, a classification was reached. Fig. 5.5 shows our test gallery ordered as per such a classification rule.

A success rate of 90% was achieved with this particular test gallery. Another test gallery was constructed as shown in Fig. 5.6. The results of running this test gallery against the training images of Fig. 5.7 are shown in Fig. 5.8. A success rate of 80% is obtained. Fig.5.9 shows another example of a test gallery, followed by the classification results in Fig.5.10.

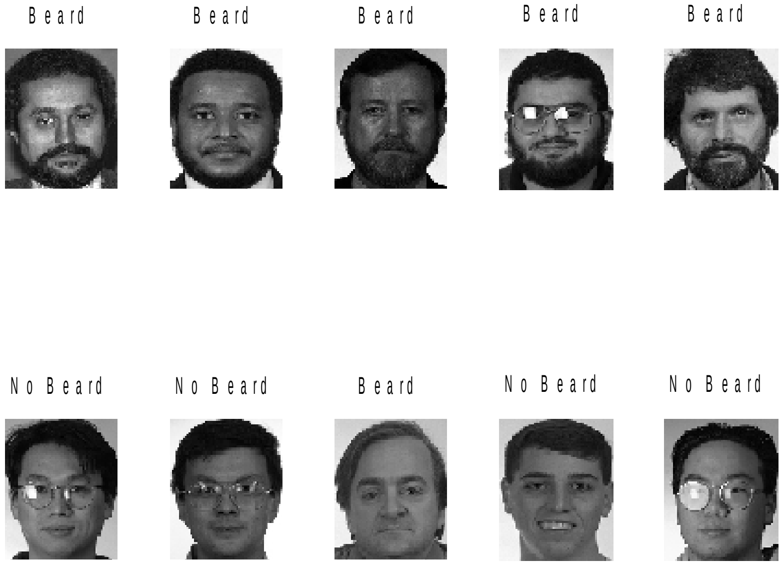

Figure 5.5 Results of the classified images

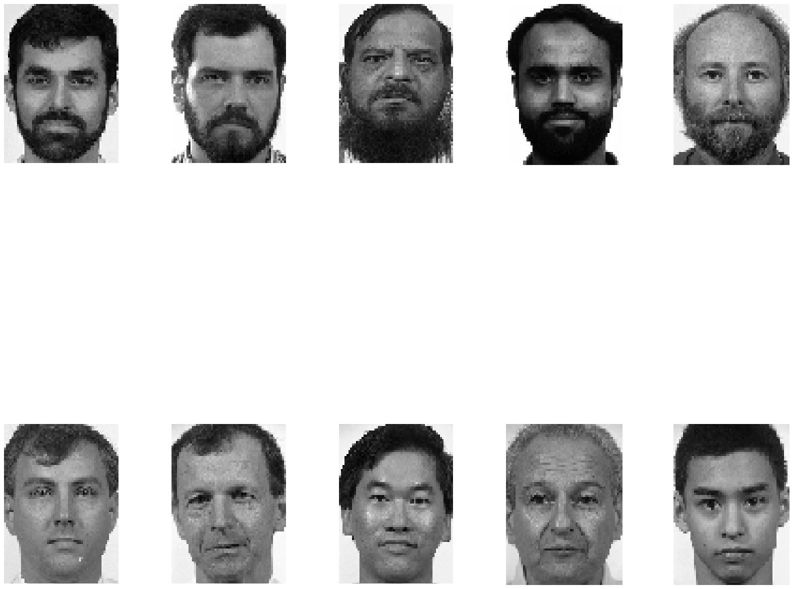

Figure 5.6 Images of test gallery

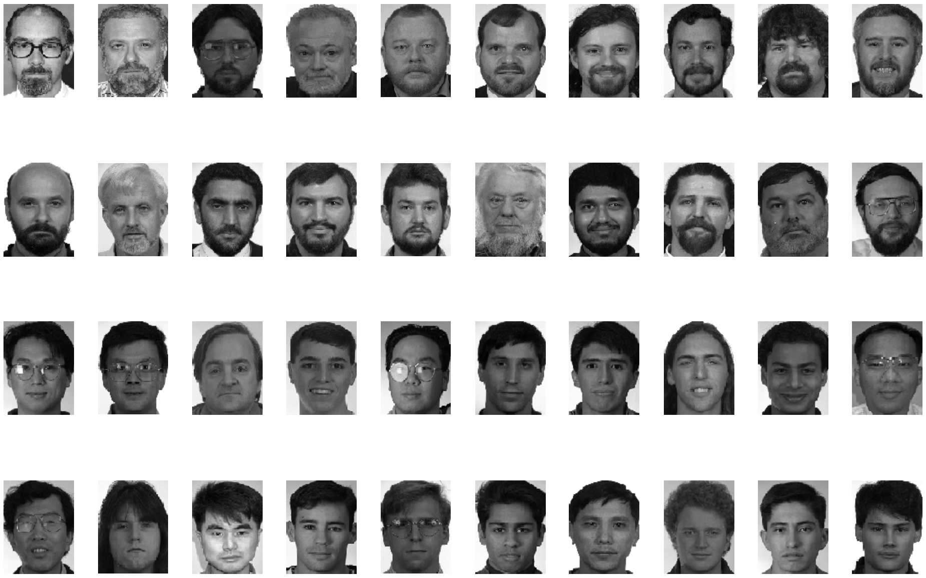

Figure 5.7 Training set images

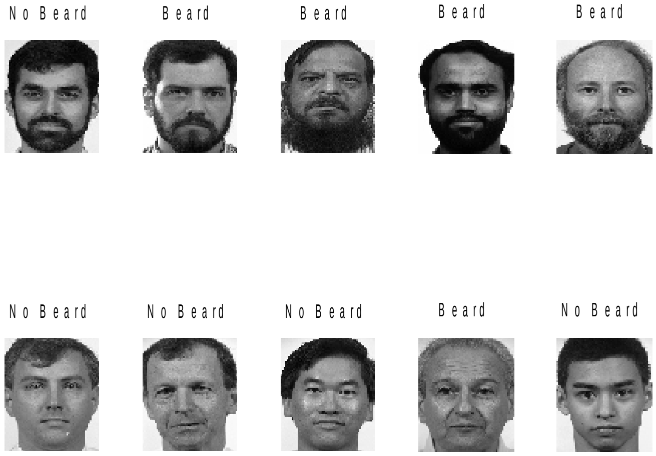

Figure 5.8 Classification of test gallery images

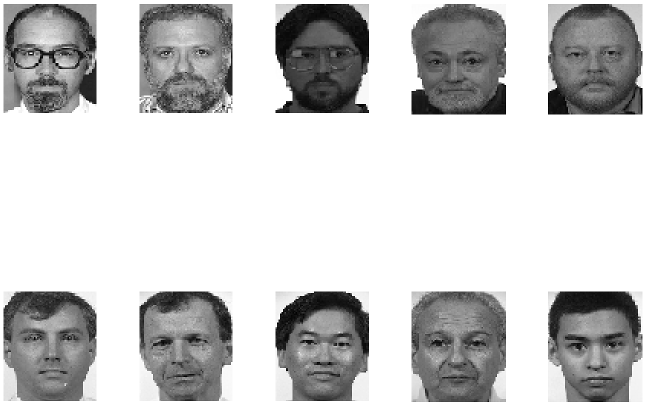

Figure 5.9 Another test gallery

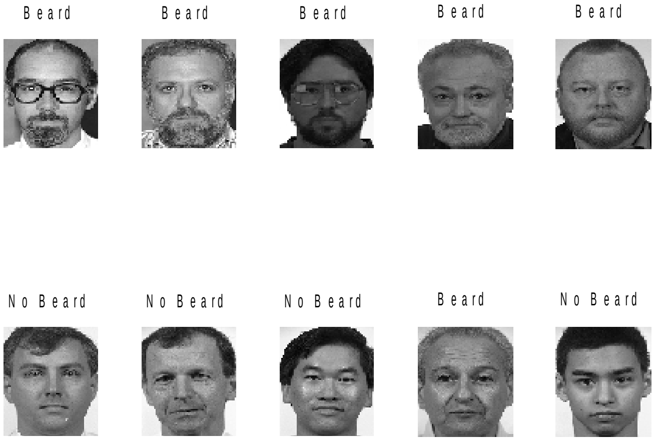

Figure 5.10 Classification results of images in Fig. 5.9

The success rate of Fig.5.10 is 90%. The average success rate for the three is 86.67%.

Chapter 6.

Facial Expression Classification with Factor Analysis

In this chapter, I will use factor analysis to analyze and classify the facial expressions of a group of Japanese women from the Japanese Female Facial Expression Database (JAFFE) [52,53]. Each of the ten subjects in the database will have two or three images of the seven universal expressions: anger, disgust, fear, happiness, neutral, sad and surprise. Using factor analysis, I will first show the procedure's ability to capture the common variation among a subject's facial expression gallery. That is, it will be shown that a factor analysis is able to identify, and group an individual's different facial expressions.

Facial expression analysis can be traced back to the nineteenth century work of Darwin [54]. Darwin showed that facial expressions were universal to both man and animal. He postulated that man had certain inborn emotions. A century later, Ekman and Friesen [55] laid the ground work for the existence of the six primary or basic emotions: anger, sadness, happiness, disgust, surprise and fear. In [20], Ekman and Friesen introduced the most wildly used system for facial expression analysis, FACS. Their system, *Facial Action Coding System,* or FACS is based on characterizing visually distinguishable facial movements by *Action Units,* or AU. They showed that 46 AUs are needed to account for the changes in facial expressions with different expressions corresponding to different combinations of these AUs.

Given a gallery comprising of an individual's images of the seven universal expressions, a factor analysis will be used to group the images according to the expression displayed. Given that every subject's gallery exhibits all the seven facial expressions (the six basic and a neutral), I will run our factor analysis using seven as the number of common factors. Each common factor will account for one of the six facial expressions and the neutral expression. Fig. 6.1 shows one of the subject's gallery where the following abbreviation is used:

AN- Angry; DI- Disgust; FE- Fear; HA- Happy; NE- Neutral; SA- Sad; SU- Surprise.

Fig. 6.2 shows the image of the seven factors obtained from the factor analysis with seven common factors of images of Fig. 6.1. Fig. 6.3 shows the results of running a factor analysis with seven common factors. Each image is grouped, column wise, according to the value of its loading on the seven common factors. The three images with the highest loadings on each common factor are displayed. Fig. 6.4 is another individual's gallery from the database [52]. After running a FA with seven common factors, images of the seven factors obtained are shown below in Fig.6.5. Fig. 6.6 shows a grouping of the images of Fig 6.4 according to their loadings on the seven common factors.

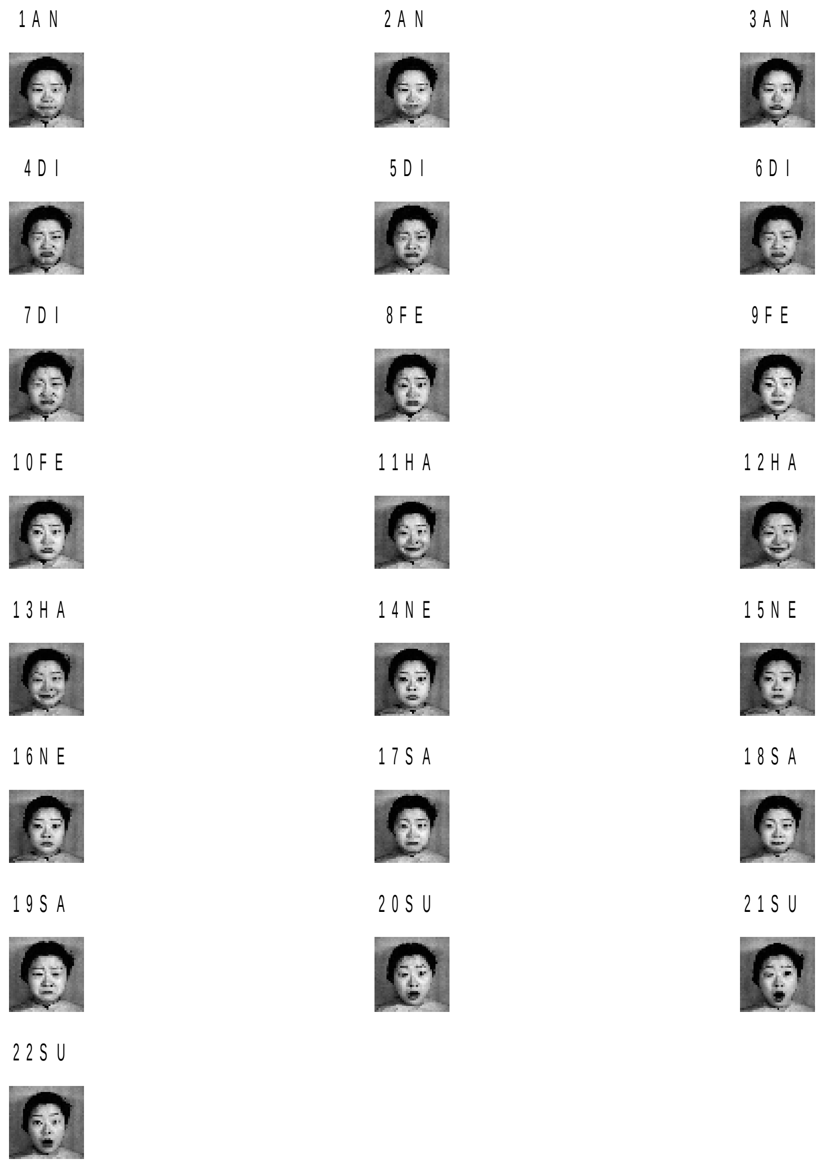

Figure 6.1 Seven Facial Expressions of a subject

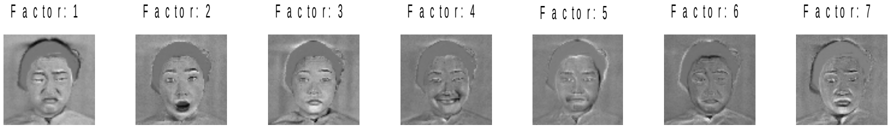

Figure 6.2 The seven common factors obtained from the FA.

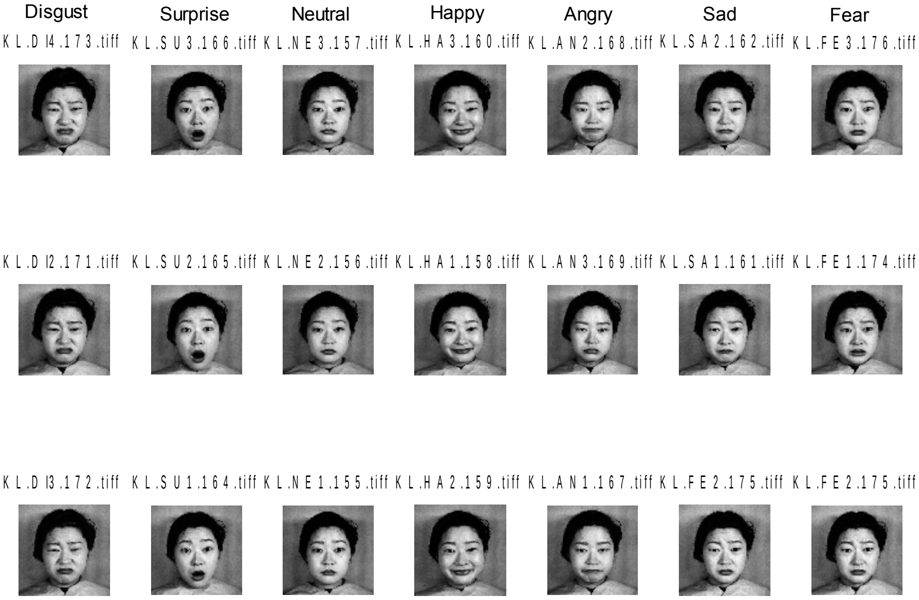

Figure 6.3 Subject grouped (column wise) according to facial expression.

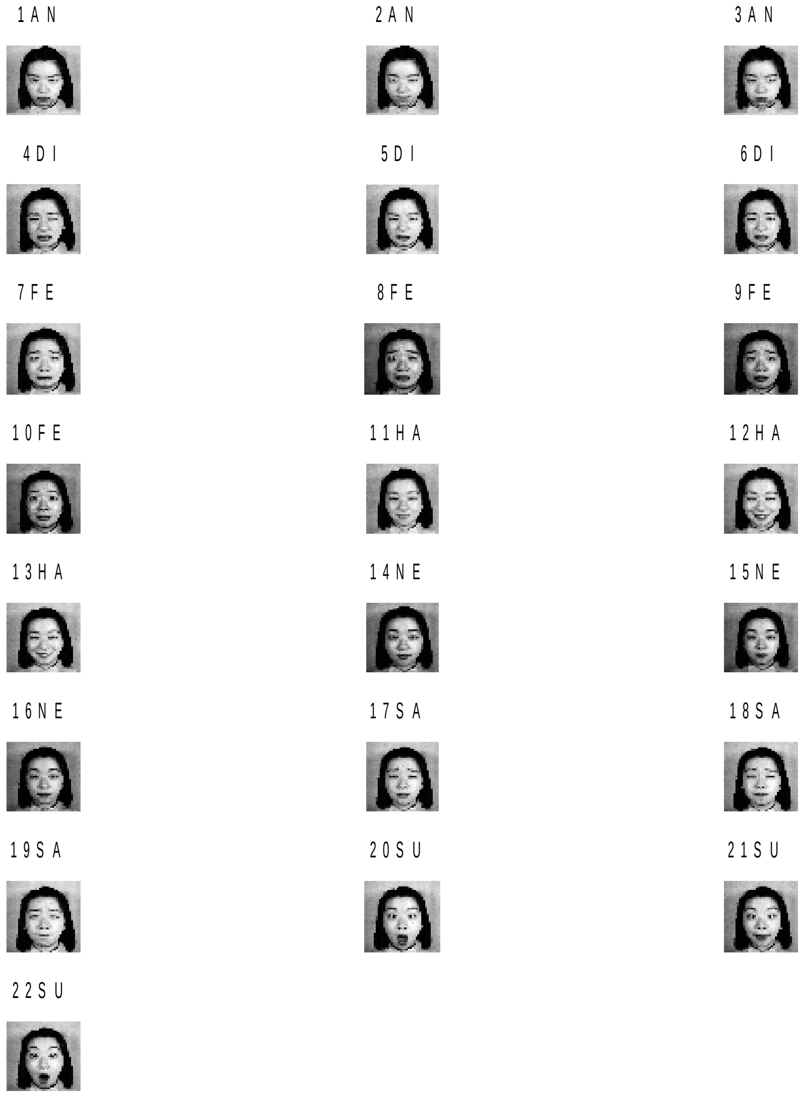

Figure 6.4 A subject with all seven facial expressions

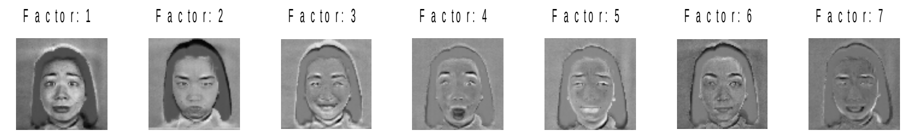

Figure 6.5 Images of the seven common factors from an FA of images of Fig 6.4

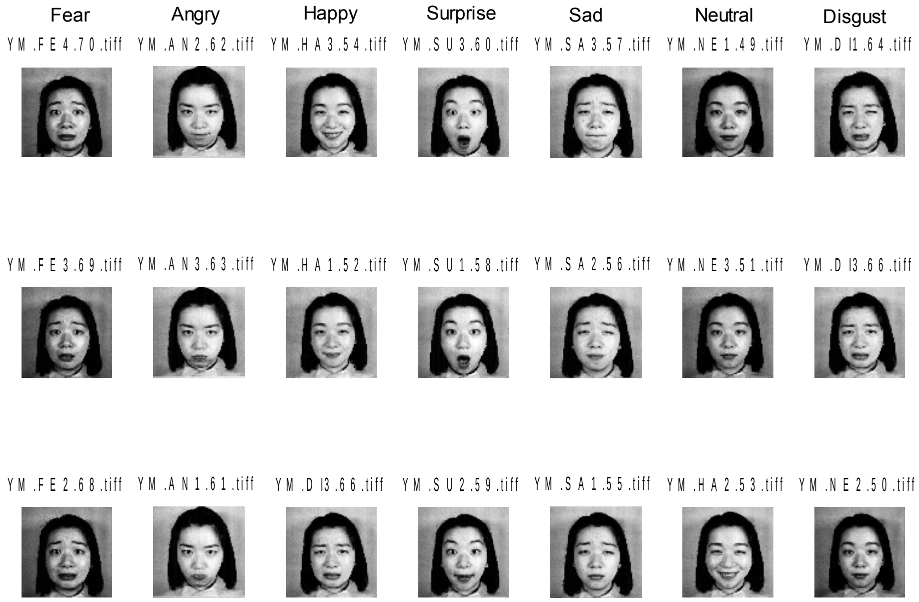

Figure 6.6 Subject grouped according to facial expression.

As evidenced from the above, factor analysis is able to capture images of similar expressions. These images of like expressions have high loadings on the common factors exhibiting a particular expression. The next question is whether we can use this to determine a particular expression. To this end, I ran an FA using seven common factors on a given dataset consisting of a subject with several images per expression. The resultant grouping of the images as per the seven common factors was then used to determine the expression of an unknown facial expression image. This was done by examining the loading of the unknown expression on the seven common factors.

To run the above classification, I first select an individual's gallery. An image of the subject with one of the exhibited expression is removed from the gallery. This will be the expression that is tested in the classification. Using the remaining images in the gallery, an FA with seven common factors is run. This results in seven common factors. An additional FA is then performed with the unknown image as part of the gallery. The classification is made according to the magnitude of the loading of the unknown expression on the established seven factors. Using a leave one out strategy, each image in an individual's gallery is run with the remaining images. As an example, the classification results for the images of Fig. 6.4 are given in Fig. 6.7. In order to quantify the success of the classification, the classifications obtained are compared to the baseline of the expressions in Fig. 6.4. So, looking at the first three expression images (image numbers 1, 2, 3), we see that the FA was able to identify the correct expression of anger. In such a manner, all the image expressions in Fig. 6.7 are checked against the baseline image expressions in Fig. 6.4. An overall success rate is determined as the percentage of the number of correctly identified expressions. Looking at Fig. 6.7, we see that with the exception of image numbers 6, 7 and 17, FA was able to correctly classify all the expressions of this particular individual.

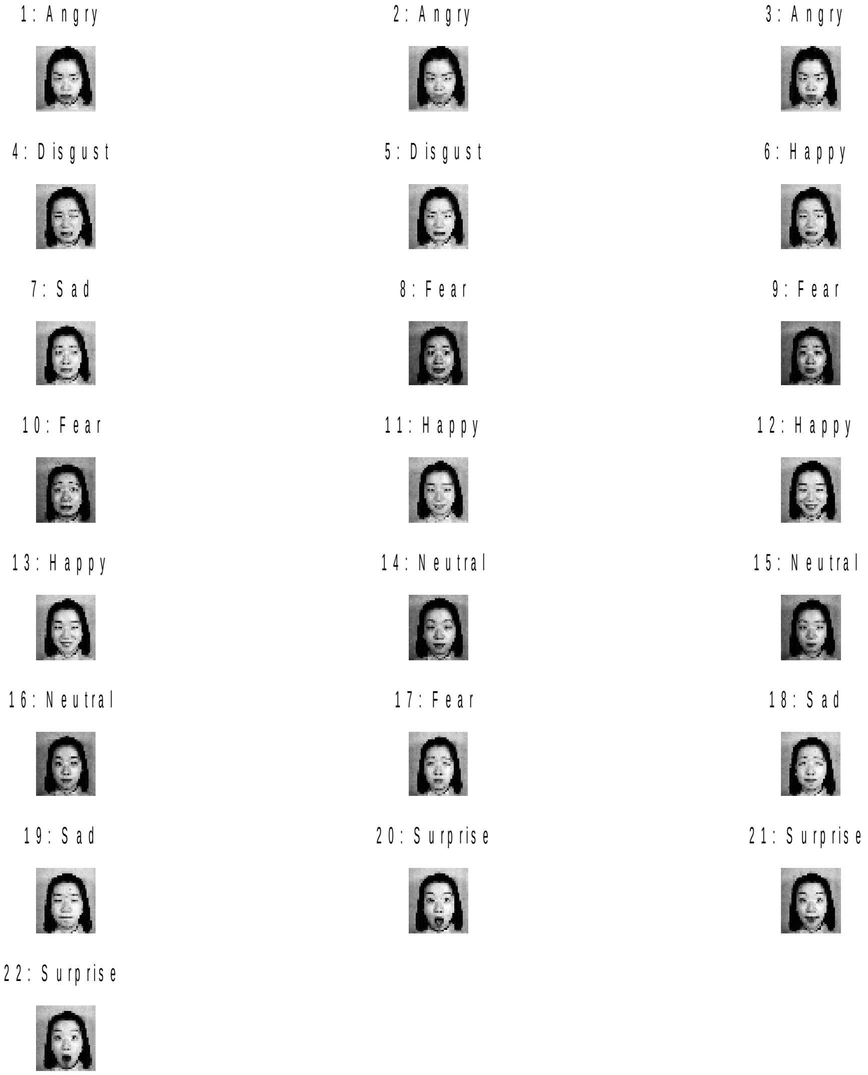

Figure 6.7 Classification of each image

That is a success rate of 86.4% for this subject. Another test is shown below in Fig. 6.8

corresponding to the subject in Fig. 6.1.

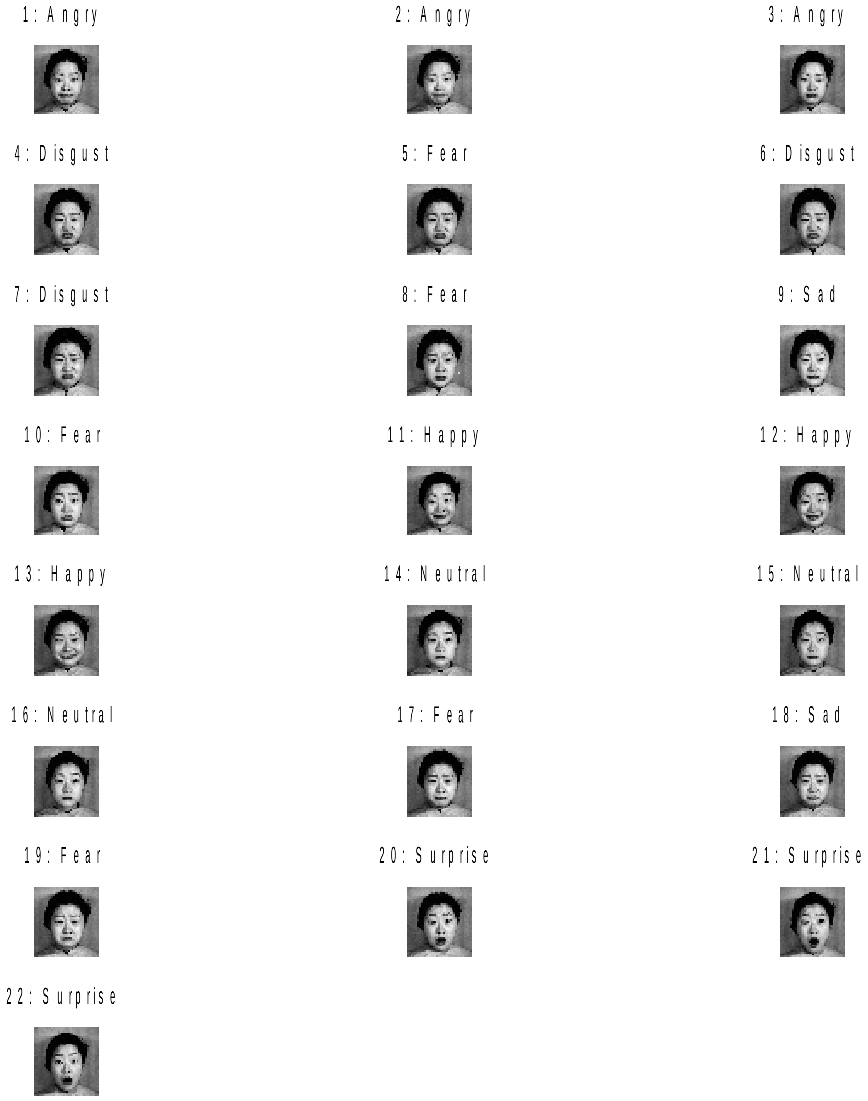

Figure 6.8 Classification result of gallery in Fig. 6.1

Comparing Fig.6.8 and Fig 6.1 we see that with the exception of image numbers 9, 17 and 19 all expressions are correctly classified. The above success rate is again 86.4%.

Fig. 6.9 shows another subject's gallery.

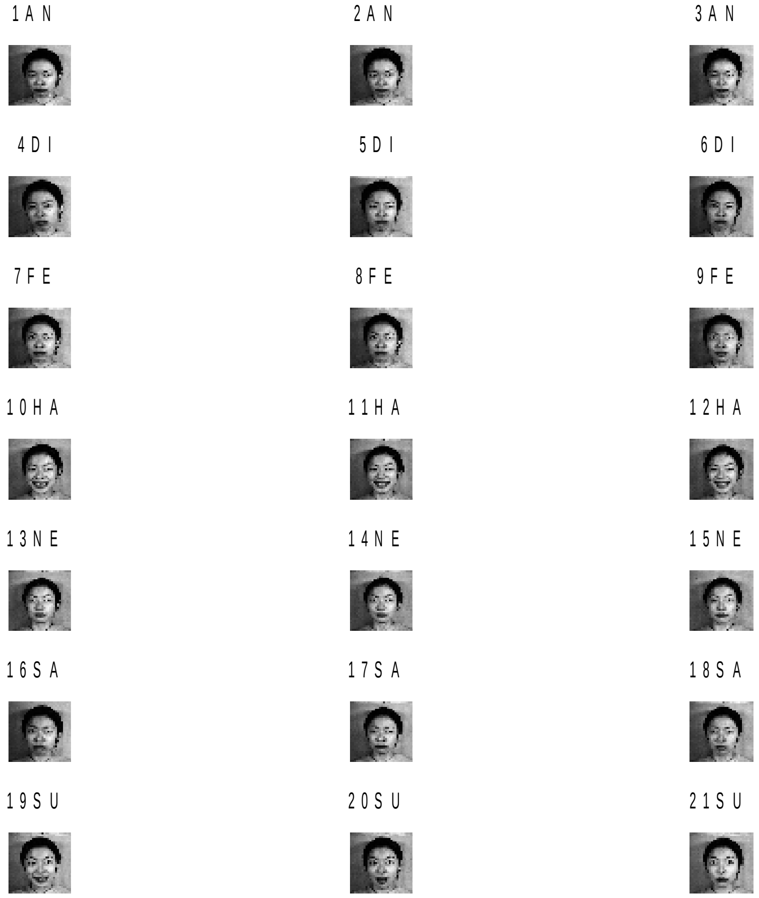

Figure 6.9 Facial images of a subject in a gallery.

After running a FA with seven common factors, an image of the seven factors obtained are shown below in Fig. 6.10.

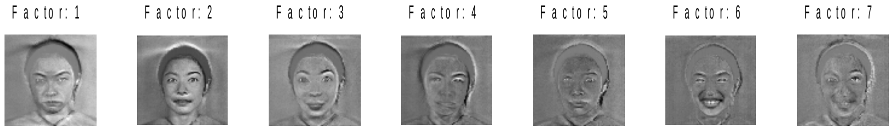

Figure 6.10 Images of the seven common factors of gallery in Fig.6.9

Fig. 6.11 shows the grouping of facial expressions of images in Fig.6.9.

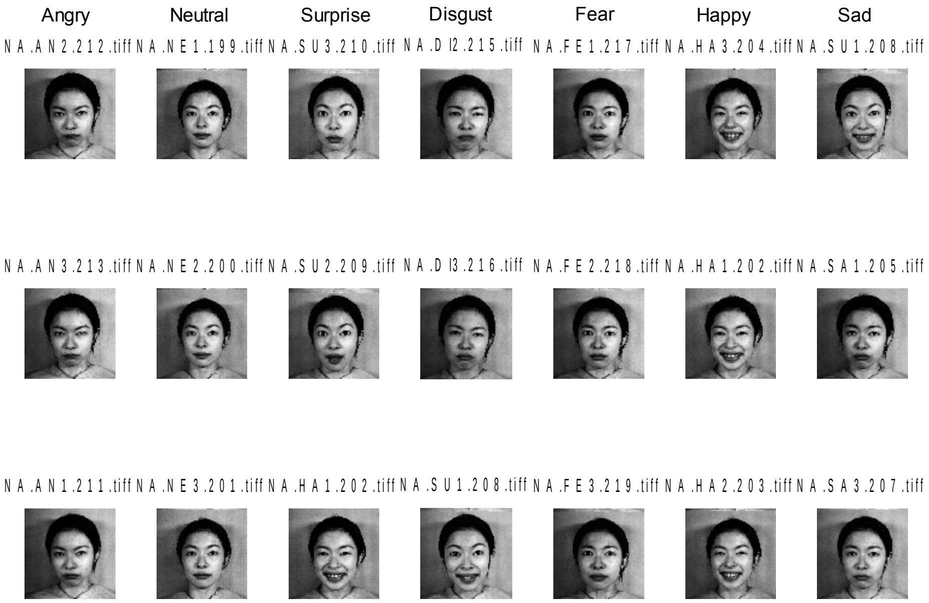

Figure 6.11 Classification if images in Fig. 6.9

The results of the classification of images in Fig. 6.9 are shown in Fig. 6.12

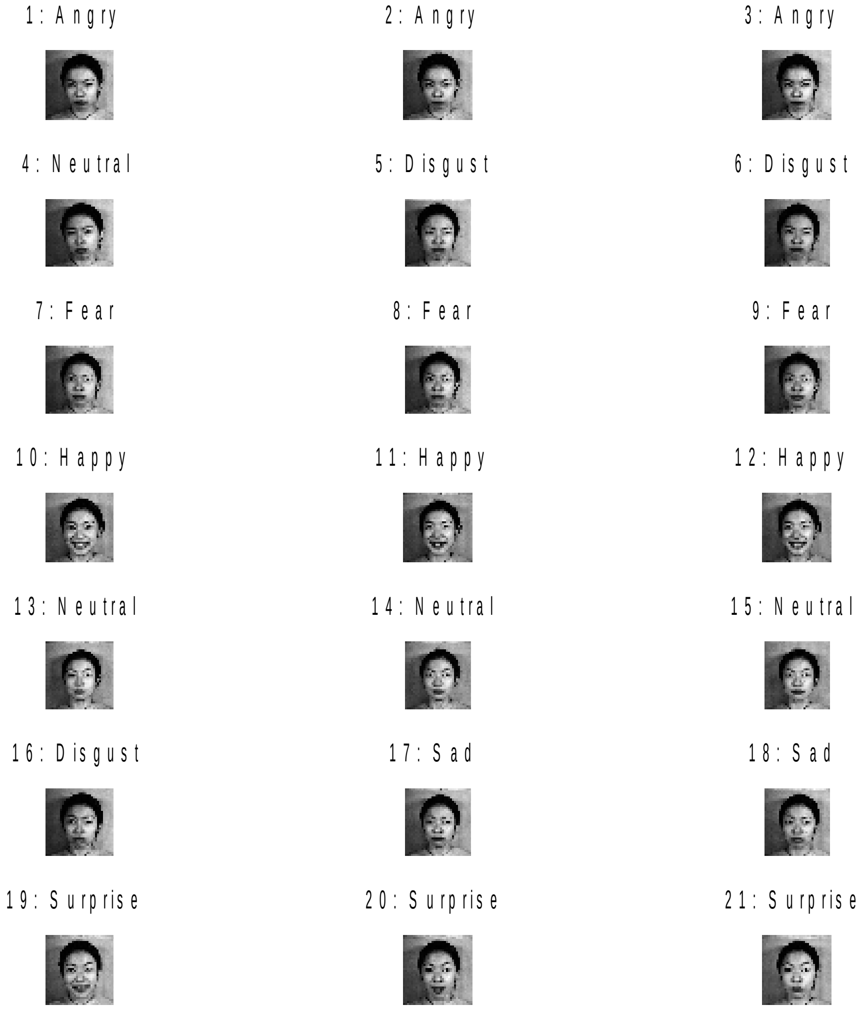

Figure 6.12 Classification results of images of Fig. 6.9

Comparing Fig. 6.12 and Fig. 6.9 we see that other than image number 4 and number 16, all are correctly identified. Giving a success rate of 90.5%. Fig. 6.13 shows a table with the success rates for all ten subjects. The average overall success rate is 85.02%.

Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10
95.2%	69.6%	90.9%	90%	85.7%	65%	90.5%	90.5%	86.4%	86.4%

Figure 6.13 Success rate for all ten subjects

Chapter 7.

Conclusion

In the course of this dissertation, I examined the different multivariate analysis tools used for the purpose of facial images. I first investigated the use of principal component analysis as used for facial recognition. I also studied both the linear discriminate and independent components analyses. While both of these techniques have been shown to be effective against the classical face recognition problem, I found the factor analysis method to be a powerful tool for classifying face images into groups based on traits such as gender. This is because, unlike a principal components analysis, a factor analysis seeks to explain a number of variables, or images in this case, in terms of a few underlaying traits. These traits or categories take the form of factors. This is done by an analysis of the correlations among the variables.

Prior to applying the FA algorithm, the images used were all manually cropped and resized. This, I realize, is somewhat empirical, and almost certainly has some impact on the performance of the approach. In future work, I would like to automate this step, or at the very least use a more systematic and controlled procedure. In this dissertation, I have shown that Factor analysis can be a powerful tool for use as a biometric classifier. Starting with the application presented in Chapter 4, I was able to use FA for the purpose of gender classification with an overall success rate of 90%. In Chapter 5, I used FA to classify subjects as having facial hair or not. The success rates for the three galleries considered were 90%, 80% and 90% for an overall average correct classification rate of 86.67%. Finally, in Chapter 6, I used a FA to classify human facial expressions. In the test, I grouped (column wise) each expression image with images of the same expression. The next experiment involved is the use of the FA to classify a new probe image (one not used in the initial FA) as one of the seven expressions. Fig. 6.13 shows a table with the success rates for the ten subjects. An overall success rate of 85.02% was achieved. This compares well with the rate of the other two multivariate statistical analyses used in the analysis of static facial images. Edwards [60] reports a rate of 74% when using PCA based on Mahalonobis distance and LDA, while Huang [61], reports a rate of 84.5% when using PCA with a minimum distance classifier on a 2D emotion space.

The main original contribution of this dissertation has been to show that factor analysis is a powerful method for treating several important biometric classification problems that fall outside the scope of the classical face recognition problem. An important aspect of future research will be to determine if this generalizes broadly to biometric classification problems in general or is specific to the three applications considered in this dissertation.

While the gender discrimination problem has been treated elsewhere, the performance obtained here is excellent and FA is distinctly parsimonious compared to the competing methods. To the best of my knowledge, classification based on facial hair has not been treated previously.

Both of these classification applications are of great practical interest for the automated surveillance, market analysis, and in the analysis of shopper behavior. For example, with a gender classifier, department stores could easily and automatically generate data to analyze how certain types of displays influence shopping habits based on gender.

Perhaps even more significant are the good success rates I obtained using FA to classify emotions in Chapter 6. An application capable of classifying human emotions from facial images could be integrated in the design of a smart house or work environment. It could be used in setting a more appropriate ambiance. Another use would be in the area of determining mood or intent for security applications. Such an application can more accurately gauge individuals by reading their exhibited emotions. This would also allow stores to gain valuable feedback from shoppers toward a host of products, or even placements of certain items.

Bibliography

1. R. Chellappa, C. Wilson, and S. Sirohey, “Human and Machine recognition of faces: A survey.” *Proceedings of IEEE*, vol. 83, no.5, pp705-740, May 1995.
2. W.W. Bledsoe, “Man machine facial recognition”. Technical Report PRI 22, Panoramic Research Inc., Palo Alto, CA, 1966.
3. M. Kirby and L. Sirovich, “Application of the Karhunen-Loeve procedure for the characterization of human faces,” *IEEE Transaction on Pattern Analysis and Machine Intelligence,* Vol. 12, no. 1, pp.103-108, January 1990.
4. Baback Moghaddam , Ming-Hsuan Yang, Learning Gender with Support Faces, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v.24, no.5, p.707-711, May 2002.
5. P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The FERET Evaluation”, in Face Recognitioin: From Theory to Applications, H. Wechsler, P. J. Phillips, V. Bruce, F.F. Soulie, and T.S. Huang, Eds., Berlin: Springer-Verlag, 1998.
6. Abdi *et al*., More about the difference between men and women: Evidence from linear neural network and principal component approach. Neural Computation. Vol. 7, pp. 1160-1164, 1995.
7. Wu, B., Ai, H., Huang, C.: Lut-based adaboost for gender classification. In

AVBPA, Vol. 2688, pp. 104–110, June 2003.

8. J. Yang, and J. Yang, "From Image Vector to Matrix: A Straightforward Image Projection Technique-IMPCA vs. PCA", *Pattern Recognition*, Vol. 35, no.9, pp1997-1999, 2002.

9. R. Brunelli and T. Poggio, “Face recognition: features versus templates”, *IEEE Trans. on PAMI*, Vol. 15, No. 10, pp. 1042-1052, Oct. 1993.

10. “Bertillon system.” Encyclopedia Britannica Online. <*http://www.britannica.com/EBchecked/topic/62832/Bertillon-system*>.

11. G. Johansson, “Visual perception of biological motion and a model for its analysis,” *Perception and Psychophysics* 1973, Vol. 14, pp. 201-211.

12. I. Kozlowski & J. Cutting, “Recognizing the sex of a walker from a dynamic point-light display,” *Perception and Psychophysics* 1977*,* Vol. 21, No. 6, pp. 575-580.

13. Barclay, C.D., Cutting, J.E. & Kozlowski, L.T. Temporal and spatial factors in gait perception that influence gender recognition. *Perception & Psychophysics,* 1978, 23, pp. 145-152.

14. G. Mather & L. Murdoch, “Gender discrimination in biological motion displays based on dynamic cues,” *Proceedings of the Royal Society of London Series B: Biological Sciences* 1994*,* Vol. 258, pp. 273-279.

15. A. Burton, V. Bruce and N. Dench, “What's the difference between Men and Women? Evidence from facial measurements”, *Perception*, vol. 22, pp. 153-176,

1993.

16. Golomb B.A., Lawrence D.T and Sejnowski T.J. “SEXNET: A neural network that recognizes sex from human faces.” *Advances in Neural Information Processing Systems.* pp. 572-577, 1991.
17. S. Yen, P. Sajda and L.Finkel, “Comparison of gender recognition by pdp and radial basis function networks,” *The neural biology of Computation,* pp. 433-438, 1994.
18. G. W. Cottrell, J. Metcalfe, EMPATH: face, emotion, and gender recognition using holons, *Proceedings of the 1990 conference on the advances in neural information processing systems* vol.3, pg. 564-571, October 1990, Denver, CO.
19. P. Ekman and W.V. Friesen. Constants Across Cultures in the Face and Emotions. *Journal of Personality and Social Psychology,* pages 124-129, 1972.
20. P. Ekman and W.V. Friesen. *Facial Action Coding System.* Palo Alto, CA. Consulting Psychologist Press, Inc. 1978.
21. A. Mehrabian, “Communication without Words,” Psychology Today, vol. 2, no. 4, pp. 53-56, 1968.
22. Ekman, P. (1999). "Basic Emotions". In: T. Dalgleish and M. Power (Eds.). *Handbook of Cognition and Emotion*. John Wiley & Sons Ltd, Sussex, UK.
23. J.F. Cohn, A.J. Zlochower, J.J. Lien, and T. Kanade, Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression, *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 396-401, 1998.

24. T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *IEEE Trans. on PAMI*, Vol. 23, No. 6, pp. 681-685, June, 2001.

25. H. Hong, H. Neven, and C. von der Malsburg, Online Facial Expression Recognition Based on Personalized Galleries, *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, pp. 354-359, 1998.

26. M.J. Lyons, J. Budynek, and S. Akamatsu, Automatic Classification of Single Facial Images, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1,357-1,362, 1999.

27. W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips,(2000) "Face Recognition: A Literature Survey", *UMD CfAR Technical Report CAR-TR948.*

28. M. Pantic and J.M. Rothkrantz. “Automatic Analysis of Facial Expressions: The State of the Art”. *IEEE Transaction on Pattern analysis and Machine Intelligence. Vol 22, No 12, December 2000.*

29. Pearson, K. “On lines and planes of closest fit to a system of points in space” *Philos. Mag., vol.*2, pp.557-572, 1901.

30. Hotelling, H. “Analysis of a Complex of statistical variables into principle components*”, J. Educational Psychol., vol.* 24, pp.417-441; 498-520, 1933.

31. M. Turk and A. Pentland, “Face Recognition using Eigenfaces,” *Proceedings of International Conference on Pattern Recognition*, pp. 586-591, 1991.

32. AT&T Laboratories Cambridge. http://www.orl.co.uk/facedatabase.html.

33. W. Zhao, R. Chellapa, and P. Philips, “Subspace linear discriminant analysis for

face recognition", Technical Report CAR-TR-914, 1996.

34. Hyvärinen, A. and E. Oja, "Independent Component Analysis: Algorithms and Application*"*, *Neural Networks*, 2000.
35. Stone, James V. "*Independent Component Analysis: a tutorial introduction*". 2004.
36. Bell, A.J. and T.J. Sejnowski, "An information-maximization Approach to Blind Separation and Blind Deconvolution". *Neural Computation,* 1995.
37. M.S. Bartlett, H.M. Lades and T.J. Sejnowski, "Independent component representations for face recognition," Presented at *SPIE Symposium on Electronic Imaging*: Science and Technology; Conference on Human Vision and Electronic Imaging III, San Jose, CA 1998.
38. M.S. Bartlett, J.R. Movellan and T.J. Sejnowski, "Face Recognition by Independent Component Analysis," *IEEE Transaction on Neural Networks*, Vol. 13, pp1450-1464, 2002.
39. S.A. Mulaik. The Foundation of Factor Analysis. McGraw-Hill Book Company.
40. H.H Harman. Modern Factor Analysis. 2nd ed. University of Chicago press.
41. Spearman, C. "General Intelligence, objectively determined and measured". *Am. J. Psychol., vol.*15, pp201-293.
42. S. Rizvi, P. J. Phillips and H. Moon, "A Verification Testing Protocol and statistical performance analysis for face recognition algorithms". *Proceedings, IEEE Conference on ComputerVision and Pattern Recognition*. Pp833-838.

43. Bryan F. J. Manly. "*Multivariate Statistical Methods"*. Chapman and Hall, 1944.

44. Kaiser, H.F. The Varimax criterion for analytic rotation in factor analysis.

45. R.A. Johnson, D.W.Wichern. Applied Multivariate Statistical Analysis. 5th ed. Prentice Hall, NJ, 2002.

46. Härdle, W. and Simar, L. "Applied Multivariate Statistical Analysis". Springer. 2003.

47. Apostal, T.M. *Mathematical analysis: A modern approach to advanced calculus.* Reading, Massachusetts. Addison-Wesley Publishing Co., 1957.

48. Jacobi,C.G.J. Ueber ein leichtes Verfahren die in der Theorie der Saecular-stoerungen vorkommeden Gleichungen numerisch aufzuloeson. *J. Reine Angewandte Mathematik* (1846).

49. Stewart, M. *Introduction to linear Algebra.* New York: D. Van Nostrand Company, 1963.

50. Lawley, D.N. The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Edin.,*(1940).

51. B. Moghaddam, W. Wahid, and A. Pentland, "Beyond eigenface: probabilistic matching for face recognition", *Proceedings of Int'l Conf. on Automatic Face and Gesture Recognition (FG'98)*, pp. 30-35, April, 1998.

52. M.J.Lyons, S.Akamatsu, M.Kamachi, J.Gyoba. Coding Facial Expressions with Gabor Wavelets. *Proceedings, third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE Computer Society,* pp200-205, April 14-16

1998, Nara Japan.

53. M.Lyons J.Budynek, S.Akamatsu. Automatic Classification of single facial images. *IEEE Transactions on Pattern and Machine Intelligence,* vol . 21, no.12, pp. 1357-1462, 1999.
54. C. Darwin. *The Expression of the Emotions in Man and Animal.* J. Murray, London, 1872.
55. P. Ekman and W.V. Friesen, *Unmasking the Face. New Jersey:* Prentice Hall, 1975.
56. Wishart, J. The generalized product-moment distribution in samples from a normal multivariate population. *Biom.* A, (1928).
57. Y. Adini, Y. Moses and S. Ullman, "Face Recognition: the Problem of Compensating for Changes in Illumination Direction", *IEEE Trans. on PAMI*, Vol. 19, no. 7, pp. 721-732, July 1997.
58. L. Wiskott, J.M. Fellous, N. Kruger and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," Proceedings of ICIP, Vol. 1, pp. 129-132, 1997.
59. Micheal J. Lyons, Shigeru Akamatsu. Miyuki Kamachi, Jiro Gyoba. "Coding Facial Expressions with Gabor Wavelets" *Proceedings, Third IEEE International Conference on Automatic Face and Gesture Recognition*, April 1998. Nara Japan, IEEE Computer Society.
60. G.J. Edwards, T.F. Cootes, and C.J. Taylor, "Face Recognition Using Active

Appearance Models," *Proc. European Conf. Computer Vision*, vol. 2, pp. 581-695, 1998.

61. C.L. Huang and Y.M. Huang, "Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification," *J. Visual Comm. and Image Representation*, vol. 8, no. 3, pp. 278-290, 1997.
62. M. Turk, "A Random Walk Through Eigenface", *IEICE Trans. INF. & SYST.*, Vol. E84-D, no. 12, Dec. 2001.
63. M. Kass, A. Wttkin, and D. Terzopoulos, "Sankes: Active Contour Models", International Journal of Computer Vision, pp. 321-331, 1998.
64. D. S. Turaga, and T. Chen, "Face Recognition Using Mixtures of Principal Components," *Proceedings of ICIP*, pp. 101-104, 2002.
65. P. Navarrete and J. Ruiz-del-Solar, "Eigenspace-based Recognition of Faces: Comparisons and A New Approach", *Proceedings of Image Analysis and Processing*, pp. 42-47, 2001.
66. A. J. Bell, and T. J. Sejnowski, "An Information-maximisation Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129--1159, 1995.
67. W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swet, and J. Weng, "Discriminant Analysis of Principal Components for Face Recognition", Springer-Verlag, 1998.
68. B. Moghaddam, "Principal Manifolds and Probabilistic Subspaces for Visual Recognition", *IEEE Trans. on PAMI*, Vol. 24, No. 6, pp. 780-788, June 2002.

69. J. W. Davis & Hui Gao, “Gender Recognition from Walking Movements using Adaptive Three-Mode PCA,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 2004, 1063-6919/04.

70. K. Etemad, and R. Chellapa, “Discriminant analysis for recognition of human faces”, *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 2148-2151, 1996.

71. W. Zhao, R. Chellapa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition", *Proceedings of Automatic Face and Gesture Recognition*, pp. 336-341, 1998.

72. E. Boyle, A.H. Anderson, and A. Newlands, “The Effects of Visibility on Dialogue in a Cooperative Problem Solving Task,” *Language and Speech*, vol. 37, no. 1, pp. 1-20, 1994.

73. K. Back, B. A. Draper, J. R. Beveridge, and K. She, "PCA vs. ICA: A Comparative on the Feret Data Set" presented at joint Conference of Information Sciences, NC, 2002.

74. G.M. Stephenson, K. Ayling, D.R. Rutter, “The Role of Visual Communication in Social Exchange,” *Britain J. Social Clinical. Psychology*, vol. 15, pp. 113-120, 1976.

75. B. Scholkopf, A. Smola, and K. Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, 10:1299-1310, 1998.

76. M. Turk and A. Pentland, "Eigenfaces for recognition", *J. of Cognitive*

Neuroscience, Vol. 3, No. 1, pp. 71-86, 1991.

77. Kelly, Truman L. Essential traits of mental life. *Harvard Studies in* Education, (1935). Cambridge, Mass. Harvard University Press.
78. E. K. Tang, P. N. Suganthan and X. Yao, “Generalized LDA Using Relevance Weighting and Evolution Strategy”. *IEEE World Conference on Computational Intelligence,* 2004.
79. G.J. Edwards, T.F. Cootes, and C.J. Taylor, Face Recognition Using Active Appearance Models, *Proc. European Conf. Computer Vision*, vol. 2, pp. 581-695, 1998.
80. A. Pentland, B. Moghaddam, T. Starner, “View-based and modular eigenspaces for face recognition”, *Proceedings of IEEE, CVPR*, 1994.
81. Baek, K., Draper, B., Beveridge, J. and She, K. “PCA vs. ICA: A comparison on the FERET dataset”. *Proceedings, IEEE Conference on Computer vision and Pattern Recognition.* 2001.
82. H. Kobayashi and F. Hara, Facial Interaction between Animated 3D Face Robot and Human Beings, *Proc. Int'l Conf. Systems, Man, Cybernetics,*, pp. 3,732-3,737, 1997.
83. H. Peng and D. Zhang, "Dual EIgenspace Method for Human Face recognition", *IEEE Electronics Letters,* Vol. 33, No. 4, pp.283-284, 1997.
84. Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, Comparison between Geometry-Based and Gabor Wavelets-Based Facial Expression Recognition

Using Multi-Layer Perceptron, *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, pp. 454-459, 1998.

85. J. Zhao and G. Kearney, Classifying Facial Emotions by Back propagation Neural Networks with Fuzzy Inputs, *Proc. Conf. Neural Information Processing,* vol. 1, pp. 454-457, 1996.
86. J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face Recognition Using Kernel Direct Discriminant Analysis," *IEEE Trans. on Neural Networks*, pp. 117-126, Vol. 14, No. 1, Jan. 2003.
87. P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection", *IEEE Transaction. Pattern Analysis and Machine Intelligence*, vol. 19, No 7, 1997, pp711-720.
88. M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience,* Vol. 3, pp. 72-86, March 1991.
89. S. Lee, S. Jung, J. Kwon, and S. Hong, "Face Detection and Recognition Using PCA", *Proceedings of the IEEE Region 10 Conference*, Vol. 1, pp. 84-87, 1999.
90. Cardoso, J. "Infomax and Maximum Likelihood for source separation", *IEEE Letters on Signal Processing*, 1997.
91. L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America*, vol. 4, March 1987.
92. B. Moghaddam and A. Pentland, "Face Recognition using View-Based and Modular Eigenspace", In Proceedings, *IEEE Conference on Computer Vision and*

Pattern Recognition, 1994.