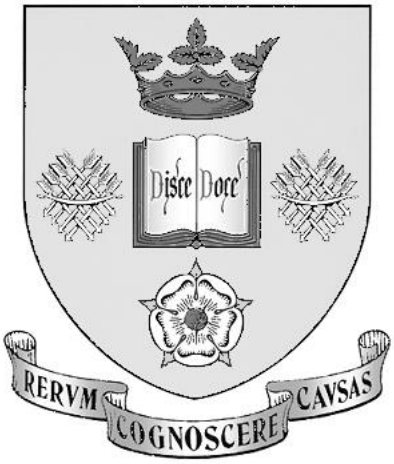
The University Of Sheffield.

Access to Electronic Thesis

Author:	Lucy Morecroft
Thesis title:	A Statistical Approach to Facial Identification
Qualification:	PhD
Date awarded:	19 October 2009

PhD Thesis

A Statistical Approach to Facial Identification



University of Sheffield
Department of Probability & Statistics

Lucy Morecroft

September 2009

Abstract: ***A Statistical Approach to Facial Identification***
Author: Lucy Morecroft

This thesis describes the development of statistical methods for facial identification. The objective is to provide a technique which can provide answers based on probabilities to the question of whether two images of a face are from the same person or whether there could be two different people whose facial images match equally well. The aim would be to contribute to evidence that an image captured, for example, at a crime scene by CCTV, is that of a suspect in custody. The methods developed are based on the underlying mathematics of faces (specifically the *shape* of the configuration of identified landmarks) At present expert witnesses carry out facial comparisons to assess how alike two faces are and their declared expert opinions are inevitably subjective.

To develop the method a large population study was carried out to explore facial variation. Sets of measurements of landmarks were digitally taken from ≈3000 facial images and Procrustes analyses were performed to extract the underlying face shapes and used to estimate the parameters in statistical model for the population of face shapes. This allows pairs of faces to be compared in relation to population variability using a multivariate normal likelihood ratio (MVNLR) procedure. The MVNLR technique is a recognised means for evidence evaluation, and is widely used for example on trace evidence and DNA matching. However, many modifications and adaptations were required because of unique aspects of facial data such as high dimensionality, differential reliabilities of landmark identification and differential distinctiveness within the population of certain facial features.

The thesis describes techniques of selection of appropriate landmarks and novel dimensionality reduction methods to accommodate these aspects involving non-sequential selection of principal components (to avoid ephemeral facial expressions) and balancing of measures of reliability against selectivity and specificity.

List of Tables 8
List of Figures 12
1 Introduction 16
1.1 Motivation, Aims and Overview 16
1.2 Thesis Outline 17
1.3 Existing Methods for Facial Comparison and Identification 19
1.3.1 Background 19
1.3.2 Pattern Recognition 20
1.3.3 Photogrammetry 22
1.3.4 Facial Mapping 23
1.3.5 Overlay Techniques 23
1.3.6 Other Facial Analysis Work 24
1.3.7 Problems with Current Methods 25
1.4 Forensic statistics 26
1.4.1 Admissibility 26
1.4.2 Likelihood Ratios 27
1.4.2.1 Trace Evidence 28
1.4.2.2 DNA Analysis 29
1.4.2.3 Fingerprint Analysis 30
1.5 Historical Development of Shape Theory 31
1.5.1 Background 31
1.5.2 Possible Facial Landmarks 33
1.5.3 Three Dimensional Facial Data 34
1.6 Summary 35
2 The Data 38
2.1 Introduction 38
2.2 Anthropological Facial Landmark Data 39
2.3 Pilot Study - The FBI Catalogue of Facial Types 42
2.4 Main Study - The Geometrix® Facial Image Database 44
2.4.1 Exploratory Data Analyses 45
2.4.2 The Image Collection Procedure 48
2.4.3 The Collection of Landmark Data 51
2.5 Reliability Study 53
2.6 Validation Study 54
2.7 Other Image Data for Testing Facial Matching Techniques 54
2.7.1 Test Data 1 – Matching Data from Two Observers 55
2.7.2 Test Data 2 - FBI Suspects Data 56
2.7.3 Multiple Images of Agent Vorder Bruegge 57
2.7.4 Known Matches and Exclusions for Subset Selection 58
2.7.5 Twins and Controls 58
2.7.6 Other Data from Multiple Images of Like Faces 59
2.8 Key Information for Dataset Variables 60
2.9 Summary 60
3 Statistical Methods 63
3.1 Introduction 63
3.2 Landmark Based Shape Analysis and Procrustes 64
3.3 Who was Procrustes? 64
3.4 Summary of Procrustes Analysis 65
3.4.1 Shape 66
3.4.2 Procrustes Methods 66

3.4.3 Landmarks *67*
3.4.4 Removing Translation *68*
3.4.5 Removing Size *69*
3.4.6 The Pre-shape and Pre-shape Space *70*
3.4.7 Removing Rotation *71*
3.4.8 Removing Reflection *71*
3.4.9 Shape and Shape Space *71*
3.4.10 Full Procrustes Distance *72*
3.5 Ordinary Procrustes Analysis ***73***
3.5.1 Full Procrustes Fit *74*
3.6 Generalized Procrustes Analysis ***75***
3.6.1 Full Procrustes Fit *75*
3.6.2 Full Procrustes Mean *76*
3.7 Principal Components Analysis in the Tangent Space ***77***
3.7.1 Tangent Space *77*
3.7.2 Partial Procrustes Tangent Coordinates *78*
3.8 Evidence Evaluation of Facial Matches using Likelihood Ratios ***78***
3.8.1 Control and Recovered Data *80*
3.8.2 Background Database *80*
3.8.3 Choosing a Method for the Evaluation of Evidence *81*
3.8.4 Method: Likelihood Ratio using a Multivariate Random Effects Model and Assumptions of Normality *82*
3.8.4.1 Model 82
3.8.4.2 Estimating the Model Parameters 84
3.9 Summary ***86***
4 Statistical Shape Analysis for Facial Identification: A Pilot Study .. ***88***
4.1 Introduction ***88***
4.2 Methods ***89***
4.3 Results ***90***
4.3.1 Sources of Facial Variation *90*
4.3.1.1 Variation Attributed to Landmark Placement 90
4.3.1.2 Variation Attributed to Scanning 93
4.3.2 Facial Matching *94*
4.3.2.1 Cluster Analysis 94
4.3.2.2 Likelihood ratios 96
4.4 Summary ***97***
5 Preliminary Examination of Variation Prior to Data Collection ***99***
5.1 Introduction ***99***
5.2 Selection of Landmarks for Data Collection ***100***
5.2.1 The Data and Procrustes Registration *100*
5.2.2 PCA and Consistency between Observers *103*
5.2.3 Excluding the Least Consistent Landmarks *106*
5.2.4 Discrimination between Subjects *109*
5.2.5 Important Landmarks *113*
5.2.6 Landmarks to keep for Further Analysis *117*
5.3 Repeatability of Technique for Data Collection ***119***
5.3.1 The Data and Procrustes Registration *119*
5.3.2 PCA and Exploration of Variability *119*
5.3.3 Cluster Analysis - Groups of Similarity *125*
5.4 Summary ***130***
6 Facial Variation in the Geometrix® Database ***131***
6.1 Introduction ***131***

6.2 Complete Data 132
6.2.1 The Data and Procrustes Registration 132
6.2.2 Differences in Overall Shape and Size 133
6.2.2.1 Age and Sex Differences 133
6.2.2.2 Ethnicity and Sex Differences 135
6.2.3 Size Differences between Males and Females 136
6.2.4 PCA - Facial Shape Variability 140
6.2.5 Examination of Outliers and Data Cleaning 143
6.2.6 Variability of Individual Facial Landmarks 145
6.3 Complete Data with Replicates 150
6.4 The Anterior Facial View for 2D Facial Matching 150
6.4.1 Variability of Anterior Facial Landmarks 151
6.5 A Multivariate Normal Model for Facial Shape 155
6.6 Summary 156
7 Likelihood Ratios for Quantifying Facial Matches 158
7.1 Introduction 158
7.2 LRs for Quantifying Facial Matches 160
7.2.1 Test Data 1 – Matching Data from Two Observers 160
7.2.1.1 Test Data 1 - Results 161
7.2.2 Test Data 2 - FBI Suspects Data 162
7.2.2.1 Test Data 2 - Results 163
7.2.2.2 Checking the Model Fit and Data Cleaning 164
7.2.2.3 Results after Data Cleaning 166
7.2.3 Evidence for Potential Improvements to the Method 167
7.3 Suggested Improvements to the Method 170
7.3.1 Selection of a Subset to Optimise Match Results 170
7.3.2 A New Method for Subset Evaluation – Match: Exclusion Ratio 170
7.3.3 LR Thresholds for Claiming Matches and Exclusions 172
7.4 Applying the New Methods to Facial Data 174
7.4.1 Summary of Facial Matching Performance for all Subsets Investigated 176
7.4.2 Eleven Landmarks 178
7.4.2.1 The Data 178
7.4.2.2 Selecting the ‘best’ subsets 179
7.4.2.3 Subset Performance 180
7.4.2.4 Relating the Results back to the Matching Variables 182
7.4.3 Robustness of the ‘Best’ Subset 184
7.5 Summary 186
8 Performance Evaluation on Selected Subsets and Further Data ... 190
8.1 Introduction 190
8.2 Twins from the Geometrix® database 191
8.3 Other Data from Multiple Images of Like Faces 193
8.3.1 Results 194
8.3.1.1 Averaging Comparison Data Collected by Different Observers 195
8.3.1.2 Extending the Model to Include Observer Error in the Background Data 196
8.4 Multiple Images of Agent Vorder Bruegge 197
8.5 The Affects of Head Orientation on Matching Results 199
8.5.1 The Data 201
8.5.2 Translation to X, Y, Z Axes for Rotation 201
8.5.3 Rotations 202
8.5.4 Inverse Transformation 203

8.5.5 Results *203*
8.5.5.1 x-axis Rotations 203
8.5.5.2 Y-axis Rotations 205
8.5.5.3 Rotation in both x and y directions 207
8.6 Averaging faces ***209***
8.7 Summary ***212***
9 Discussion ***214***
9.1 Introduction ***214***
9.2 Summary by Chapter ***214***
9.3 The Anterior 2D Facial Comparison Method with the Geometrix® Database ***223***
9.3.1 Procedure *223*
9.3.2 Key Points and Suggestions for Improvement *224*
9.4 Conclusions ***224***
9.5 Limitations ***226***
9.6 Future Work ***226***
9.6.1 Ideas for Further Development *226*
10 References ***228***
11 Appendix A – Data Collection Questionnaires ***232***
12 Appendix B – Landmark Placement Manual ***238***
13 Appendix C - Confidential Image Data ***258***
14 Appendix D - Results ***259***
14.1 Twenty-two Anterior Facial Landmarks ***259***
14.1.1 The Data *259*
14.1.2 Results *263*
14.1.3 Subset Performance *264*
14.1.4 Relating the Results back to the Matching Variables *265*
14.2 Fifteen landmarks ***265***
14.2.1 The Data *265*
14.2.2 Results *269*
14.2.3 Subset Performance *269*
14.2.4 Relating the Results back to the Matching Variables *270*
14.3 Ten Landmarks ***270***
14.3.1 The Data *270*
14.3.2 Results *271*
14.3.3 Subset Performance *272*

List of Tables

Table 2.1 – List of anthropological landmark points collected for initial analyses, Farkas (1994). 41
Table 2.2 – Facial characteristics depicted in each catalogue section and the facial landmarks (Figure 2.1) visible in the images. 42
Table 2.3- Summary data, numbers of faces in database by sex and ethnic group 45
Table 2.4 - Summary data, numbers of faces in database by sex and age group 47
Table 2.5 – Number of images captured with the Geometrix® scanner by the ten different photographers (A-J) 50
Table 2.6 – Number of images measured by each observer 53
Table 2.7 – Measurements i and j from observers 1 and 2 respectively. Measurements were taken from the ten faces (subject IDs) used as Test Data 1 56
Table 2.8 – Summary of datasets used throughout this thesis 60
Table 4.1– LR results for pair-wise facial comparisons within the set of forty-eight faces from sections B and C of the facial catalogue. The top three (strongest) matches were between faces 1 and 36 in the dataset, visual assessment suggested these 2 images were an actual match. 97
Table 5.1 - The ranking of landmarks (Table2.1) from subset IV in terms of discriminatory power (-log Λ) between subjects. 115
Table 5.2 - The ranking of landmarks (Table2.1) from dataset IV in terms of consistency 116
Table 5.3 - The reduced list of landmark points, which were collected on the whole Geometrix® database 118
Table 6.1 – Summary of size and shape by sex and age group: number of observations in each age group (n), the mean centroid size of the group (Sbar), standard deviation (sd(S)), the Procrustes shape distance between the means for males and females (Procdist(m,f)) and also the root mean squared shape distance (RMS); subscript _m and _f represent males and females respectively. 133
Table 6.2 - Differences in size and shape between sex and ethnic groups: number of observations in each ethnic group (n), the mean centroid size of the group (Sbar), standard deviation (sd(S)), the Procrustes shape distance between the means for males and females (Procdist(m,f)) and also the root mean squared shape distance (RMS); subscript _m and _f represent males and females respectively. 135
Table 6.3 – Summary of which facial landmarks vary the most on each PC 155
Table 7.1 - Likelihood ratio results for 10 matches (LR>1) found from pair-wise comparisons of test data 1 (10 faces measured by two observers), using the first 5 PC scores 161
Table 7.2 - Likelihood ratio results for 10 matches (LR>1) found from pair-wise comparisons of test data 1 (10 faces measured by two observers), using the first 20 PC scores 162
Table 7.3 – Numbers and percentages of ‘matches’ (LR>1) and ‘exclusions’ (LR<1) from the FBI anterior test data (n = 60 faces) using p = 5 PCs as the number of matching variables. Columns indicate true matches (yes), true exclusions (no), possible, supposed and unverifiable matches – explained meanings in §2.7.2. 163
Table 7.4 – Numbers and percentages of ‘matches’ (LR>1) and ‘exclusions’ (LR<1) from the FBI anterior test data (n = 60 faces) using p = 20 PCs as the number of matching variables. Columns indicate true matches (yes), true exclusions (no), possible, supposed and unverifiable matches. 164
Table 7.5 – Numbers and percentages of ‘matches’ (LR>1) and ‘exclusions’ (LR<1) from the cleaned FBI anterior test data (n = 58 faces) using p = 5 and p = 20 PCs as the

number of matching variables. Columns indicate true matches (yes), true exclusions (no), possible, supposed and unverifiable matches. 166
Table 7.6 – The actual number and percentage of known matches and exclusions in the fifty-eight FBI anterior faces 167
Table 7.7 - 20 facial comparisons (Confidential Appendix, Figures 13.4 – 13.23) used to test LR matching results for different subsets. The LR when using PCs 1-20 from 22 anterior landmarks is given. 175
Table 7.8 - Summary of 'best' subsets tested for performance by matching the 58 anterior FBI faces. Subsets of PCs were obtained by first varying the number of original landmarks taken for analysis. Various LR thresholds for matches and exclusions were explored; percentages of results obtained for each subset and threshold are given (yes, no, possible, supposed and unverifiable matches) and the percentage of false positive and negative results. 177
Table 7.9 - 'Best' subsets in terms of MER and average LRs for known matches and exclusions, 11 landmarks. 180
Table 7.10 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 6 (PCs 1, 3, 4, 7, 9 and 10 from 11 landmarks) 181
Table 7.11 – Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 7 (PCs 3, 6, 7, 8, 9 and 10 from 11 landmarks) 181
Table 7.12 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 8 (PCs 1, 2, 3, 4, 5, 6, 7 and 8 from 11 landmarks) 182
Table 7.13 - PCs included in the 'best' found subset for facial matching and the facial variation represented by each of these PCs 182
Table 7.14 – Sensitivity of the 'best' subset; % of matches and exclusions for fifty-eight FBI test faces when a number of faces were randomly excluded from the background database. 186
Table 8.1 - Facial matching results comparing three sets of twins and age and sex-matched controls 192
Table 8.2 – Number of facial matches evaluated using the 'best' subset of matching variables (§7.4.2.3) with the LR procedure (§3.8.4, §7.3, §7.4). Results are for known facial matches comparing two different photos of the same face, data were collected by one observer. 194
Table 8.3 - Number of facial matches evaluated using the 'best' subset of matching variables (§7.4.2.3) with the LR procedure (§3.8.4, §7.3, §7.4). Results are for known facial matches comparing the data from one photograph collected by two different observers. 195
Table 8.4 - Number of facial matches evaluated using the 'best' subset of matching variables (§7.6.1.2) with the LR matching procedure (§3.8.4, §7.2). Results are for known facial matches comparing the data from one photograph; data were collected by two different observers and then averaged. 196
Table 8.5 - Number of facial matches evaluated using the 'best' subset of matching variables (§7.4.2.3) with the LR matching procedure (§3.8.4, §7.3, §7.4) extended to include observer error in the data model. Results are for known facial matches comparing the data from one photograph collected by two different observers. 197

Table 8.6 – Number of facial matches and exclusions obtained through evaluating LR s using the 'best' subset of matching variables (§7.4.2.3) to compare fourteen images of agent Vorder Bruegge (Confidential Appendix Figure 13.30). 199
Table 8.7 - Number of matches and exclusions when comparing each original face with its downward rotation about the *x*-axis 204
Table 8.8 - Number of matches and exclusions when comparing each original face with its upward rotation about the *x*-axis 205
Table 8.9 - Number of matches and exclusions when comparing each original face with its rotation to the left about the *y*-axis 207
Table 8.10 - Number of matches and exclusions when comparing each original face with its rotation to the right about the *y*-axis 207
Table 8.11 – Facial matches and exclusions obtained when comparing each original face with its rotation when the face had been rotated in both the x and y directions upwards and left. 208
Table 8.12 - Facial matches and exclusions obtained when comparing each original face with its rotation when the face had been rotated in both the x and y directions downwards and right. 208
Table 8.13 – Facial matches and exclusions obtained when comparing each original face with its rotation when the face had been rotated in both the x and y directions downwards and left. 208
Table 8.14 - Facial matches and exclusions obtained when comparing each original face with its rotation when the face had been rotated in both the x and y directions upwards and right. 208
Table 8.15 – Match and exclusion results for the fourteen images of Agent Vorder Bruegge (Appendix C, Fig. 13.30), various different averages were taken of the landmark configurations to see the effect on the number of matches obtained. 211
Table 14.1 - Twenty-two anterior facial landmarks 259
Table 14.2 - The top subsets in terms of MER close to one for twenty-two anterior facial landmarks (§7.5.1). Also included are the average LRs obtained for known matches and known exclusions. 263
Table 14.3 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 9 (PCs 2,3,6,7,8 and 9) 264
Table 14.4 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 12 (PCs 2, 5, 6,7,8,9 and 10) 264
Table 14.5 - The top subsets of PCs in terms of MER close to one for 15 landmarks (§7.5.2). Average LRs obtained for known matches and known exclusions are also given. 269
Table 14.6 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 2 (PCs 1, 2, 3 and 9 from 15 landmarks) 270
Table 14.7 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 3 (PCs 3, 4, 5, 6, 7, 8 and 9 from 15 landmarks) 270
Table 14.8 - The top subsets of PCs in terms of MER close to one for 10 landmarks (§7.5.4). Average LRs obtained for known matches and known exclusions are also given. 271

Table 14.9 - Percentage of true matches (LR>1 and ‘Yes’), true exclusions (LR<1 and ‘No’), false positive (LR>1 and ‘No) and false negative (LR<1 and ‘Yes’) results obtained from quantifying facial matches using LRs calculated from subset 6 (PCs 3, 4, 6, 8, 9 and 10 from 10 landmarks) 272
Table 14.10 - Percentage of true matches (LR>1 and ‘Yes’), true exclusions (LR<1 and ‘No’), false positive (LR>1 and ‘No) and false negative (LR<1 and ‘Yes’) results obtained from quantifying facial matches using LRs calculated from subset 5 (PCs 2, 3, 4, 6, 7 and 10 from 10 landmarks) 272

List of Figures

Figure 1.1 – An illustration of pattern recognition of faces (Hallinan et al, 1999). Algorithms based upon statistical methods are used to identify 'best fits' with data resembling faces under different lighting conditions. 20
Figure 1.3 - Face to face superimposition showing comparison between two images (A and B) of the same person, C shows a vertical wipe image and D shows a horizontal wipe (Yoshino et al, 2000). 24
Figure 1.4– Traditional anthropometric landmarks of the face (Farkas, 1994) 33
Figure 2.2 - Histograms to show distribution of subject age, by sex 47
Figure 2.5 - Screen still of the 3D model produced by the FaceVision802 scanner. 49
Figure 2.6 - Triangulation of the location of the endocanthion landmark point in two 2D images; the Forensic Analzer program obtains the 3D data from two 2D images and scanner calibration information. 52
Figure 4.1 - Plots along the first principal component for the Procrustes rotated coordinates of: a) three sets of measurements taken from the same subject (top); b) the mean measurements taken from four different subjects (bottom). The plots are evaluated at c = -3, -2, -1 (*) standard deviations and c =+ 1, + 2, + 3 (+) standard deviations along the principal component. 91
Figure 4.2 - PC score plots to show that variation between separate measurements taken from the same face (intra-measurement error) is smaller than facial variation (inter-individual variation). Six different faces are represented by different symbols; three different measures were taken of each face. 92
Figure 4.3 - PC score plots showing that the variation attributed to different scans of an image was smaller than inter-facial variation. Six different faces are represented by different symbols; there were three scans per image. 93
Figure 4.4 – Dendogram to show the results of a single-linkage cluster analysis carried out to assess possible facial matches between images in two different sections of the catalogue[2].No. 1–26 represent images from section B and 27–48 from section C. Height refers to the distance between the clusters. 95
Figure 5.1 - Three orthographic projections of the raw landmark coordinate data. The x-y plot shows the anterior facial view (subject forward facing); x-z shows the overhead view (subject nose facing downwards) and y-z shows the profile facial view (subject left facing). 101
Figure 5.2 - Procrustes aligned landmark data, using translation and rotation (preserving scale). The x-y plot shows the anterior facial view (subject forward facing); x-z shows the overhead view (subject nose facing downwards) and y-z shows the profile facial view (subject left facing). 102
Figure 5.3 - The first few PC scores, symbols indicate the data from different observers. The percentage of variation explained by the two plotted PCs is displayed above each plot. 103
Figure 5.4 - The mean landmark configuration for Observer L (grey dashed lines) with vectors drawn to the mean landmark configuration for Observer X (solid black lines). 104
Figure 5.5 - Mahalanobis distances for each observation at each landmark (numbered 1-61), the median distance for each landmark is displayed in grey. 105
Figure 5.6 - The first few PC scores for subset II, the two symbols indicate data from two different observers. PC1 V PC4 still shows some systematic differences between the two observers, indicated by the separation of the symbols. The percentage of variation explained by the two plotted PCs is displayed above each plot. 107
Figure 5.7 - The first few PC scores for subset III, the two symbols indicate data from two different observers. The percentage of variation explained by the two plotted PCs is displayed above each plot. 108

Figure 5.8 - The first few PC scores for subset IV, the two symbols indicate data from two different observers. For this subset there is no clear distinction between the two observers in the first few PCs. The percentage of variation explained by the two plotted PCs is displayed above each plot. 109
Figure 5.9 - Dendrogram displaying the results of a Wards cluster analysis for subset IV. Labels indicate subject face (1-35). Height refers to the square error of the clusters, which are added to those of their lower clusters. 111
Figure 5.10 - Enlarged section of the left hand side of the dendrogram in Figure 5.10; the lowest level clusters in dataset IV group different faces (numbered). Height refers to the square error of the clusters, which are added to those of their lower clusters. 112
Figure 5.11 - Orthogonal views of the mean face for subset IV, the size of the landmark label indicates the discriminatory power between individuals for that landmark, the larger the labels the more discriminatory power. 114
Figure 5.12 – The first few PC scores, observers 1-6 are represented by rings, triangles, crosses, diamonds, solid squares and solid circles respectively. The percentage of variation explained by the two plotted PCs is displayed above each plot. 120
Figure 5.13 - Differences between Observers 1 and 3. Grey dashed lines indicate observer 1 shape, black lines indicate differences of observer 3 from observer 1 (see text) 122
Figure 5.14 - The first few PC scores after correcting for mislabelled points; observers 1-6 are represented by rings, triangles, crosses, diamonds, solid squares and solid circles respectively. The percentage of variation explained by the two plotted PCs is displayed above each plot. 123
Figure 5.15 – The first few PC scores after correcting for mislabelled points, numbers 1-10 represent the 10 different faces under investigation. The percentage of variation explained by the two plotted PCs is displayed above each plot. 124
Figure 5.16 – Dendrogram of Wards' cluster analysis, numbers 1-6 represent the different observers. Height refers to the square error of the clusters, which are added to those of their lower clusters 126
Figure 5.17 - Enlargement of the left hand branch of the dendrogram in Figure 5.16 to examine clusters at the lowest level, numbers represent different observers. Height refers to the square error of the clusters, which are added to those of their lower clusters 127
Figure 5.18 - Wards Cluster dendrogram; numbers represent the faces 1-10. Height refers to the square error of the clusters, which are added to those of their lower clusters. 128
Figure 5.19 - Enlargement of the first branch of the dendrogram in Figure 5.18, numbers represent different faces. Height refers to the square error of the clusters, which are added to those of their lower clusters. 129
Figure 6.1 - Mean centroid size of faces in different age groups, males and female 134
Figure 6.2 - Mean face shape for males (black) and females (grey) for the Procrustes aligned data preserving scale. 136
Figure 6.3 – Mean face shape for males (black) and females (grey) for the Procrustes aligned data preserving scale 139
Figure 6.4 – Scree plot to show cumulative amount of variation explain by PCs of Procrustes aligned data (preserving size); 30 landmark points in 3D (i.e. 90 variables) 140
Figure 6.5 – First few PC score plots for Procrustes aligned (preserving scale) complete configurations, males and females are represented by circles and triangles respectively. 141
Figure 6.6 - PC score plots for Procrustes aligned complete data including the removal of scale; circles represent males, triangles females. 142

Figure 6.7 – Procrustes rotated x and y coordinates of duplicated landmarks (observation number 29 in black and 30 in grey) from one face. Observation 29 has mislabelled landmarks 3 and 13; this can be corrected by swapping over the landmark labels for this configuration 144
Figure 6.8 - Procrustes rotated x and y coordinates of duplicated landmarks (observation number 2052 in black and 2053 in grey) from one face. Landmark 26 has been misplaced for observation 2052, this cannot be corrected for and so the configuration must be excluded from further analyses 145
Figure 6.9 –Loadings for first four PCs for the XY anterior facial view, points represent the mean face shape in the aligned data with solid vectors indicating the direction and magnitude of the loadings. 147
Figure 6.10 - Loadings for first four PCs for the ZY and XZ facial views, points represent the mean face shape in the aligned data with solid vectors indicating the direction and magnitude of the loadings. 149
Figure 6.11- PC plots (PCs 1-4) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows. 152
Figure 6.12 –PC plots (PCs 5-8) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows. 153
Figure 6.13 - PC plots (PCs 9-12) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows. 154
Figure 6.14 – QQ plot to check multivariate normality of complete replicated data configurations 156
Figure 7.1 – QQ plot to check multivariate normality of the background data plus the facial comparison data (control and recovered), three clear outliers are apparent. 165
Figure 7.2 – Graph to show how LR results varied as the number of PCs to use as matching variables in the LR procedure were increased. 168
Figure 7.3 – Ratio of true to false results obtained for each of the thirteen subsets investigated for various LR thresholds used to quantify a 'match'. The corresponding exclusion thresholds were (1/threshold for match). 173
Figure 7.4 – Average strength of evidence for a match for each of the top thirteen subsets. 174
Figure 7.5 - Plots (PCs 1-6 from eleven landmarks) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows. 183
Figure 7.6 - plots (PCs 7-10 from eleven landmarks) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows. 184
Figure 7.7 – Percentage of false positive and negative results for the fifty-eight FBI test faces. Using the 'best' subset of matching variables and excluding faces from the background database to check sensitivity of results. 186
Figure 8.3 - Example of the x-y anterior landmarks of one of the test faces in the original orientation (circles) and generated angles of downwards (triangles) and upwards (squares) tilts by 2 degrees about the x-axis. 204
Figure 8.4 - Example of the x-y anterior landmarks of one of the test faces in the original orientation (circles) and generated angles of left (triangles) and right (squares) tilts by 7 degrees about the y-axis. 206
Figure 12.1. The Glabella (g). 239
Figure 12.2. The Sublabiale (sl). 240
Figure 12.3. The Pogonion (pg). 241

Figure 12.4. Endocanthion (en), Left. 242
Figure 12.5. Endocanthion (en), Right. 242
Figure 12.6. Exocanthion (ex), Left. 243
Figure 12.7. Exocanthion (ex), Right. 243
Figure 12.8. Pupil (p), Left. 244
Figure 12.9. Pupil (p), Right. 244
Figure 12.10. Left Palpebrale Inferius (pi), Left. 245
Figure 12.11. Palpebrale Inferius (pi), Right. 245
Figure 12.12. Sellion (se). 246
Figure 12.13 - Pronasale (prn). 247
Figure 12.14. Alar (al), Left. 248
Figure 14.1 – Twenty-two anterior facial landmarks 260
Figure 14.2- PC plots (PCs 1-4) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows. 261
Figure 14.3 –PC plots (PCs 5-10) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows. 262
Figure 14.4 - PC loadings for the PCs 1-4 from 15 landmarks 267
Figure 14.5 - PC loadings for the PCs 5-10 from 15 landmarks 268

1 Introduction

1.1 Motivation, Aims and Overview

In recent years there have been many advances in the technology of deterrent surveillance (e.g. closed circuit television or CCTV) and facial verification systems to allow access to secure areas. There has been a biometric revolution with identity cards on the increase, and reliable methods of identifying people are sought (in terms of both accurately identifying someone and preventing forgery and stolen identity). There are many crime cases that include CCTV evidence; however there are no recognised, tested scientific ways to assess whether the person pictured in CCTV images is the person you have in custody to any measureable degree of certainty. There are a number of 'experts' in the area of facial comparison, although no general agreement in methodology or whether it can even be justified scientifically in the actual world away from controlled laboratory conditions. These experts can only claim a measure of identification based upon their professional judgement and experience. This project investigates whether the underlying mathematics of faces (based on accurate measurements of specific features) can reliably distinguish different people.

Such a method of matching facial images from technology such as CCTV would be invaluable to assist in criminal analysis. At present an expert witness usually answers the question 'how alike are these two faces', and they ignore the equally important question 'how many other faces would also be equally alike (or even more so) as these two faces'. To answer this second question a large population study has been carried out to learn more about faces, in a similar way to a DNA database facilitating the calculation of DNA match probabilities. We investigate the feasibility of quantifying how similar two faces are by using sets of empirical measurements taken from facial images. Procrustes analysis is performed to extract the underlying face shapes from the sets of measurements. The shapes of pairs of faces are compared against the large collected sample using a multivariate likelihood ratio (LR) procedure. This compares the likelihood that the pair of faces matches with the likelihood that the queried face matches some other face in the background population. A substantial amount of the research was spent uncovering which facial measurements provide the best quality evidence of matches. A new method for evaluating subsets is proposed that ensures

measured variables are equally good at confirming true and excluding false matches. This involves determining a LR threshold level for confirming matches in facial shape data.

1.2 Thesis Outline

This first chapter critically reviews some of the existing methods used in facial comparison and identification, to explain the motivation behind the current project (§1.3). An investigation into the current techniques used in forensic statistics to evaluate types of evidence is carried out to uncover what is required from a scientific analysis for it to be deemed admissible evidence in a court of law. Statistical shape theory is then explored to investigate the plausibility of using these scientific principals to develop a new method for the analysis and comparison of face shapes.

Chapter 2 explains the diverse range of datasets that were used for this work. The large and complex main database was collected specifically with facial identification in mind. A full summary of the data collection procedure and some exploratory data analyses are given. Many of the features of the main database were not used throughout this research, there are further investigations which could be carried out, some have been suggested in §9.6.1.

Chapter 3 reviews the statistical methods applied and also modified and extended for this project. Given are accepted methods in Shape analysis that were used to extract face shapes from the data; these were a fundamental part of the methodology of the work. Also given is a summary of methods for evaluating evidence. One method which uses likelihood ratios to evaluate multivariate data was chosen as the most appropriate for use with the facial shape data.

Chapters 4 to 8 are my contribution of work to the area of facial matching. Chapter 4 summarises the results of a pilot study carried out on a very small data set to assess the viability of the proposed methods, developed further in chapters 7 and 8. In chapter 5 a sample of face shape data is explored using multivariate techniques to assess different aspects of variation. An evaluation of the facial landmarks is carried out to obtain a subset considered ‘good’ for matching (in terms of discrimination and consistency of

placement). This subset was measured on the main facial database. The method of data collection is validated to ensure measurements collected by multiple observers were suitably comparable.

Chapter 6 looks at the large facial database collected for facial identification purposes (§2.4). Differences in shape and size between different age, sex and ethnic groups are investigated. The variation between faces is examined through principal components analysis (PCA). This uncovered that some data cleaning was necessary. The variation of individual facial landmarks is presented through plots of the loadings of the principal components (PCs). Particular facial features are seen to vary on particular PCs. A multivariate normal model for face shape is applied to the data and found to be an adequate fit, and thus that a multivariate normal model was suitable for modelling Procrustes corrected facial landmark data.

Chapter 7 investigates the method for evaluating multivariate evidential data using likelihood ratios developed in chapter 3. The method was extended to consider more variables and estimate the likelihood of facial matches from the multivariate normal facial landmark data. The inclusion of a higher number of variables brought about new issues concerning the 'best' variables to use in terms of the quality of evidence produced. The chapter outlines the 'best' solution to this optimisation problem and how it was derived. A novel method for the evaluation of subsets is given. Also suggested is a LR threshold by which a match should be confirmed.

Chapter 8 evaluates the performance of the selected subset and examines some of the factors that influence the facial matching results. Three factors found to effect the results are the use of new observers for landmark collection, the position of head to camera during the image capture process and the facial expression of the subject in the image. An improvement to the method is suggested which takes the average of measurements from three different images of both the faces to be compared.

Chapter 9 presents a critical evaluation of the key findings of this research. Steps for the basic procedure of facial comparison using the proposed methods are given. Limitations of the work and suggestions for further improvements are included.

1.3 Existing Methods for Facial Comparison and Identification

1.3.1 Background

There are several techniques already available for comparing and matching facial photographs or video stills from imaging technology. Current practices used in this field are thought to be unreliable, crude and unscientific. This undoubtedly leads to mistaken facial matches or exclusions. Such mistakes can result in miscarriages of justice in both unwarranted convictions and acquittals of guilty suspects, (http://www.innocencenetwork.org.uk/).

A good example of this is the first time in England that an expert witness was called by the prosecution for identification relating to security camera evidence. In 1989 a jury found James Ryan guilty of being one of three raiders involved in an unsuccessful bank robbery. An expert witness had told the court that there were sufficient corresponding facial characteristics between one of the armed robbers on a CCTV film of the incident and Mr Ryan, Ryan was jailed for nine years. The conviction at the time was seen as a breakthrough in the science of facial photographic comparison. However, two years later in 1991 Ryan was freed, partially because of new doubts about the technique used, which was termed "facial mapping". Three appeal judges took notice of fresh defence evidence that stated Mr Ryan was not the same height as the man pictured on the film. Lord Chief Justice, Lord Lane, was quoted as saying this was the first time the appeal court had considered the “arcane” science of the derivation of accurate measurement from photographic evidence, Harley (2004).

The existing methods utilized in facial identification work are explored here to explain the motivation behind this project. The methods can be divided into two main groups – pattern recognition and Photogrammetry. Also explored are some other procedures which, despite never being properly scientifically proven, have appeared in a court room – facial mapping and overlay techniques.

1.3.2 Pattern Recognition

Pattern recognition is the operation and design of systems that recognize patterns in data. Sometimes called statistical pattern recognition it is based on sub disciplines such as discriminant analysis, feature extraction, error estimation, cluster analysis, grammatical inference and parsing (sometimes called syntactical pattern recognition). Important application areas include image analysis and person identification, and also character recognition, speech analysis, man and machine diagnostics and industrial inspection.

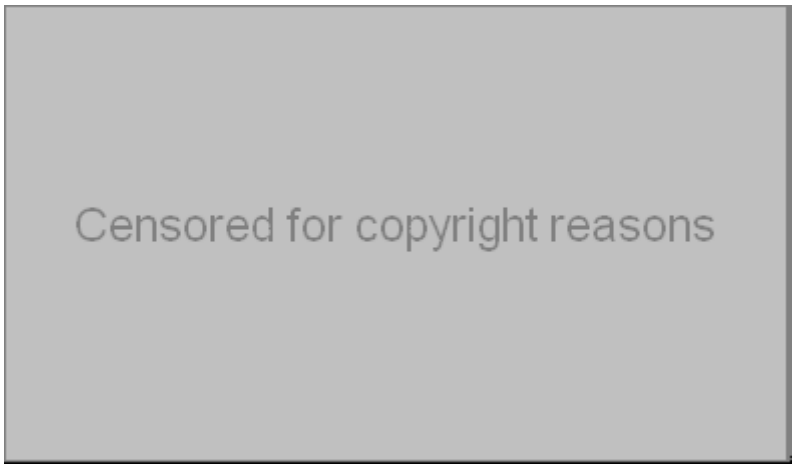


Figure 1.1 – An illustration of pattern recognition of faces (Hallinan et al, 1999). Algorithms based upon statistical methods are used to identify 'best fits' with data resembling faces under different lighting conditions.

Pattern recognition relies on powerful computers that are capable of capturing and processing image data in real time. It compares the shapes, shades and contours of the face of a known individual with photographic matches from an input source such as closed circuit television images. As can be seen in Figure 1.1, the statistical algorithms used in pattern recognition identify 'best fits' with data resembling faces under different lighting and contrast settings. Depending upon the acceptance point, like faces, unlike faces and non-faces (e.g. the dog in Figure 1.1) may be recognised in the data.

Although considerable investments have been made into the area of pattern recognition, the application of the current technology has only been effective in displacing criminal activity somewhere else and not in reliably identifying suspects.

It should be noted that the investigation underway here is specifically on the facial comparison of images and not in the field of facial recognition, the two are somewhat different issues. Facial recognition is a biometric method for the automatic recognition of an individual using unique facial attributes. An example of this is the situation whereby you want to pick out a face in a crowd, such systems for recognising faces are used in airports, for example, to try and recognize known faces of wanted terrorists or illegal immigrants. There are many different methods used in facial recognition to interpret a range of models, however most systems are commercially available and the manufacturing companies are reluctant to make known their techniques, making them unavailable for scientific scrutiny.

Several people have been working on the problem of facial recognition, where an input image is taken and an algorithm is applied to it to find the best match in a finite available library or database of faces. Cootes et al (2001) developed the Active Shape and Active Appearance Models (ASM and AAM respectively) for facial recognition. They use a landmark based approach with a large number of pseudo landmarks around the jaw line, this area of the face is more susceptible to shape changes that occur when the subject changes body weight, so may be less reliable for facial identification if the time between the occurrence of the crime and the identification of the perpetrator is substantial. The ASM or AAM, based on a training set of data, is placed a onto a new image where a coarse to fine iterative search is performed to find the best match in pixel intensity (grey-scale) and texture map at each landmark. These methods will be more affected by image quality than a totally anthropometrical landmark approach would be, as lighting and other factors will affect the grey-scale and pixel intensity, whereas the position of anthropometrical landmarks (e.g. corners of the eyes) should be clearly identifiable, except maybe in extreme cases. The ASM can fail when the starting position of the model is too far from the target; the model is placed in the centre of the new image and so if the face lies far from centre the search may diverge to infinity or converge to an incorrect solution.

The investigation here is into facial identification, which would declare whether two or more facial photographs, videos or other images are more similar than could have arisen by chance with two randomly selected faces. We are not looking to ‘find’ the face in an image or to construct a face from a visual account. We have an image of a face and manually locate this face by placing the anthropometrical landmarks on it; we then want

to compare it to other known faces, so it is more a facial identification through image comparison rather than facial recognition.

1.3.3 Photogrammetry

Photogrammetry is the science of making measurements from photographs. When used for facial identification Photogrammetry is based on the manual comparison of individual characteristics and proportions of faces in photographic images. Figure 1.2 shows an illustration of Photogrammetry where lines are drawn through subjectively identified features of the faces being compared. Congruence in the lines is used to determine similarity or difference.

Figure 1.2 - An illustration of Photogrammetry, Porter and Doran (2000).

Photogrammetry allows the differences in alignment between a set of features in two facial images to be determined. It does this using complex geometric formulas and usually either stereoscopic photography (two pictures taken at the same instant from slightly different vantage points) or anthroscopic visual comparison of morphology and distinguishing characteristics. Features such as scars and moles are identified and compared subjectively. The jaw-line or bridge of the nose can also be used to derive proportions.

The height of a person may be calculated to within an inch using Photogrammetry, Porter and Doran (2000), this is valuable information when it comes to photo–identification, however the methods of Photogrammetry for facial identification are said to be very basic, unscientific, and cause many errors.

1.3.4 Facial Mapping

As with the case of James Ryan (Harley, 2004), facial mapping is carried out by exploring facial characteristics of the accused and of the person on the security film to try to establish identity beyond doubt. Facial mapping is the qualitative examination of two facial images based on corresponding points that are visible in both images. A morphological comparison categorises features of the face according to type or shape, the number of similar categorisations across the comparison images are noted. The method is based upon subjective judgments of facial features. Facial mapping evidence has been used in courts, however there is no general agreement amongst forensic anthropologists and image analysts about what facial mapping is, how it should be done and what protocol should be used.

A key outcome of this project will be to provide a reliable, repeatable and quantitative method for other scientists to follow to carry out a facial comparison.

1.3.5 Overlay Techniques

Facial comparison cases in the courtroom have involved evidence using overlay techniques, where two facial images are laid over one another. Such techniques have been used by Yoshino et al (1997, 2000, 2002, and 2003) in their studies on facial image identification, see Figure 1.3 for examples.

Figure 1.3 - Face to face superimposition showing comparison between two images (A and B) of the same person, C shows a vertical wipe image and D shows a horizontal wipe (Yoshino et al, 2000).

The main problem with overlay techniques is that to get a good match, such as the one seen in Figure 1.3, it is very unlikely that the two or more images to be compared will be exactly the same scale and orientation. In most cases this means that one or both of the images have to be arbitrarily rescaled and manipulated to carry out the comparison. In altering the images in this way the resulting images may not actually be a true likeness of the subject in question at all.

1.3.6 Other Facial Analysis Work

Other approaches to facial analysis include 'building' facial matches from facial composites, for example for improved suspect identification from witness accounts. Solomon et al (2005) can produce compact mathematical representations of the human face suitable for comparison against stored databases of images. They have produced

the EigenFit software (Hancock (2000) developed the EvoFit package which works in similar way). A witness of a crime is asked to remember the sex, race and hairstyle of the offender, an algorithm produces random faces from which the witness opts for the one that seems the closest likeness. This is then fed back into the algorithm which mutates the face into a new set of variants. The cycle continues until the witness is happy with the likeness. Each face is represented by an array of principal components, changing just one of the parameters subtly alters the face, once a feature is correct it can be "locked", and the rest of the face evolved around it.

A major disadvantage of this work is that eye-witness accounts are known to be unreliable (http://www.innocenceproject.org/).

1.3.7 Problems with Current Methods

Both pattern recognition and photogrammetric approaches to facial comparison suffer two primary faults. Firstly, neither method is based on precise empirical measurements that can be shown to be satisfactorily discriminating to confirm identity. Secondly neither approach can rely on previously published studies of facial variation in the general population, to permit the probability of a credible match to be empirically established. There have not been any large enough population studies done to establish whether there is enough facial variability between individuals to be able to statistically distinguish between two faces.

With overlay techniques and facial mapping images are subject to arbitrary rescaling and manipulation, and there are no repeatable guidelines or protocol for analysts to follow. Each case may be carried out in a different way; therefore results from different analysts may be incongruent.

So, although techniques exist to aid in facial identification there are many problems associated with them which contribute to the key purpose of this research. The following section describes what is required from statistical methods that are used in a forensic context. Chapter 4 looks at the viability of using statistical shape analysis to provide a method for facial shape comparison.

1.4 Forensic statistics

1.4.1 Admissibility

Forensic science is the application of science to questions which are of interest to the civil and criminal legal systems. Forensic statistics can therefore be thought of as the application of statistical techniques to questions of interest to the legal system. There are two main divisions of forensic statistics; the first is interpreting laboratory data, similarly to any observational scientist, the second is the interpretation of observations from criminal cases, known as evidence evaluation. Lucy (2005) points out that unless all pieces of evidence in a criminal case point explicitly to an expected conclusion, different pieces of evidence carry different implications with varying degrees of power, therefore evidence evaluation statistics are designed to measure the weight or strength of the available evidence.

For any forensic technique to gain wide scale acceptance it has to be admissible in a court of law. In 1993 the Supreme Court in the United States of America issued detailed guidelines for determining which scientific evidence should be admissible in court. These guidelines are now known as the Daubert standards, as they were launched after the case of Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993), a case in which a family alleged that their children's serious birth defects had been caused by the mothers' prenatal ingestion of Bendectin, a prescription drug marketed by the respondent, Nordberg (2005).

Preceding the Daubert guidelines the admissibility of expert scientific evidence was ruled by Frye v. United States, 54 App. D. C. 46, 47, 293 F. 1013, 1014. The “Frye test” stated that expert scientific evidence was only admissible if the principles on which it was based had gained "general acceptance" in the scientific community, Nordberg (2005). Despite widespread adoption of this rule by the courts, the "general acceptance" standard was viewed as unjustifiably restrictive, because it sometimes provided evidence derived from somewhat novel scientific approaches based on the intellectual credibility of the expert witness.

The Daubert ruling declares that the case judge should determine if evidence is 'reliable', the basis of this determination should incorporate whether they agree that the expert testimony is based upon valid scientific methodology and reasoning, and also whether the testimony is relevant to the matter in question. The Supreme Court also set up a list of factors to form a general framework by which to judge this 'reliability' of evidence. Important things included in this list were whether the basis of evidence is a tested theory or method, if it has been tested and reviewed by peers in the field, any known or potential error rates associated with the techniques, and finally whether the theory is generally accepted within the relevant scientific community.

Many recognised evidence evaluation techniques are accepted in the courtroom, most of these are probabilistic. Lucy (2005) states that "a measure of evidential strength is required to tell us about the strength of evidence in support of a guilty or innocent proposition, without it actually telling us how likely or unlikely the proposition of guilt or innocence itself is" and "it is not the evaluation of the probability of a proposition 'the suspect is guilty' or 'the suspect is innocent', instead evidence evaluation is about the 'probability of evidence' in the light of competing hypotheses."

In other words some kind of impartial measure of the strength of evidence is required, to present to a court of law. Lucy (2005) suggests one way of doing this would be to give the probability of a guilty proposition before the evidence is introduced, and then the probability of the same guilty proposition after the evidence has been revealed to the court.

The following sections detail some methods currently used in forensic statistics, which are thought to be relevant when considering the kind of approach to take with statistical facial identification evidence.

1.4.2 Likelihood Ratios

A rational, intuitive method for placing a simple value on evidence was first suggested by Poincare, Darboux and Appell in the late 19th century, utilising the likelihood ratio measure, Aitken and Taroni (2004). The likelihood ratio (LR) is a statistical method used to directly assess the worth of observations. Lucy (2005) states that the LR is

currently the predominant measure for numerically based evidence. Examples of how the likelihood ratio can be applied to forensic data in the form of trace evidence and DNA evidence are outlined in the following subsections.

1.4.2.1 Trace Evidence

During the process of a crime being committed trace evidence may be transferred from criminal to crime scene or vice versa. Trace evidence might be small glass fragments or fibres (natural or man-made) from garments of clothing or carpet, for example. If a suspect's garment is examined and such trace evidence is found what is required is the probability that this trace evidence could be from the scene of crime.

Statistical methods and kernel density estimates are used to calculate the probability that the trace evidence came from the same source as the control (or crime scene), and the probability that the trace evidence came from some other source in the known population of trace evidence. These two probabilities are then evaluated by means of a LR, Evett et al (1987), Aitken and Lucy (2004).

This approach could be applied to facial identification, calculating the probability that two faces are a 'match', i.e. two facial images are of the same person. The LR with matching faces would be a comparison of the likelihood that the face of the suspect is the same as the face of the person committing the crime (captured on CCTV) and the likelihood that the face of the suspect lies somewhere else in the known population. As the probabilities are epistemic, i.e. there is no inherent knowledge of the system from which to deduce probabilities for outcomes, to enable analysis a large sample of known face shapes would need to be collected. From this sample population knowledge of the face shape system could be obtained, however without examining every member of the population the estimates of probability would always be subject to a quantifiable uncertainty. Lucy (2005) gives an example of "the bloodstain on the carpet may 'match' the suspect in some biochemical way, but was the blood which made the stain derived from the suspect, or one of the other possible individuals who could be described as a match". The same is applied to faces, we want to know if photographs of suspect and perpetrator are a 'match' and also how certain we are of this 'match', the LR provides such a measure.

In order to be able to carry out trace evidence analysis forensic scientists recognise the need for background data collections to assist in the interpretation of evidence. Large collections of information on fibres and other trace evidence exist, as the ability to collect and store data increases. Evett et al (1987) investigated the aspect of fibre colour by examining a collection of 8000 samples; they were able to model how frequently a particular fibre colour occurs based on this sample. Aitken and Lucy (2004) did a study on the elemental composition of 310 different glass fragments. Similarly a background dataset on facial measurements is needed.

1.4.2.2 DNA Analysis

Forensic DNA analysis was a great advance for the criminal justice system. DNA analysis enables large proportions of the population to be excluded as potential contributors of genetic samples (e.g., blood, hair) found at the scene of a crime. A DNA match statistic is expressed as the frequency that the DNA profile occurs in a given population, e.g. a DNA match of 1 in 1,000,000 means that approximately one person out of every one million in a population will match that DNA profile, if the statistic is this low it is unlikely that a match found between a suspect and a recovered genetic sample has occurred by chance.

The conventional method of DNA profiling was based on simple hypothesis tests, known as "match/binning", where firstly it is decided whether or not there is a match between the lengths of DNA found from the suspect and crime samples, if a match does exist then a "match proportion" is carried out, this is the proportion of a database of DNA fragments that would also be a match with the samples, in a given interval or "bin" containing the DNA fragment length of the crime sample.

Berry (1991) and Berry et al (1992) explored a Bayesian approach to DNA analysis, similar to that used with trace evidence, where likelihood ratios of guilty to innocent are calculated. He estimated the population distribution of DNA fragment lengths, attempting to account for measurement error and sampling variability, as he felt the "match/binning" approach had several characteristics that are undesirable in a court of law. For example the yes/no decision on a match is an arbitrary cut off point, meaning

that some fragments deemed “not a match” can be arbitrarily close to other fragments that do match.

As with the trace evidence methods, these sorts of techniques could be applied to faces to estimate the population distribution of face shape, accounting for measurement and sampling error. The concern here is that any particular face shape is unlikely to be as rare as any particular DNA sequence, and so resulting “match probabilities” may not be sufficient to convince a court.

1.4.2.3 Fingerprint Analysis

Fingerprint analysis has been a widely accepted scientific technique since 1900, Galton (1892). A fingerprint, or indeed any partial print that is made with the palm of the hand or a bare foot, is found at the scene of a crime and a comparison is made between that and any suspects in custody.

It is known that several months before a baby is born, ridges develop on the skin of its fingers and thumbs. These ridges arrange themselves in more or less regular patterns, known as ridge patterns, and throughout the lifetime of an individual these patterns remain unchanged, so fingerprint analysis is not probabilistic, it is either a match or not. Experts classify ridge patterns into three main classes: arches, loops, and whorls, and then each of these classes can be further divided into numerous sub-categories. Individuality of a fingerprint is not based upon the general shape or pattern that it forms; it is determined by the ridge structure and specific characteristics (known as minutiae). There are over 150 individual ridge characteristics on the average fingerprint, in legal proceedings a point-by-point comparison must only be demonstrated for at least twelve different points in order to prove or disprove that a fingerprint match is assumed, Bergen County Technical School (2004).

A fingerprint expert carries out the comparison of prints to quantify matches; such experts have been known to make mistakes which have recently cast doubts on the techniques, Mckie (2005), as fingerprint analysis is not an exact science and relies on human judgement mistakes are inevitable.

In the same way that fingerprint analysis is carried out a facial analysis could also be undertaken by experts, looking at the known landmark points on the face to do the comparison. In some ways this seems more plausible than identifying minutiae, as many of the facial landmark points are well-defined it would probably be easier to train a lay person in the methods of landmark placement than it would to explain classifying minutiae. The problem with facial identification is likely to be that the face is subject to changes throughout a person's lifetime, the ageing process means that, unlike fingerprints, the patterns in shape may vary over time.

1.5 Historical Development of Shape Theory

Reviewing the current methods used for comparing faces (§1.3) and a variety of techniques for analysing and presenting different types of forensic evidence (§1.4) it is clear that a stronger scientific basis for facial matching is needed. This section describes how shape theory could be applied to facial data to provide a more quantifiable form of forensic evidence such as DNA (§1.4.2.2) and trace evidence (§1.4.2.1).

1.5.1 Background

In the late 1970's there were several people approaching the problem of how to analyse shapes. Kendall (1977) began by looking at shape in the contexts of archaeology and astronomy. Around the same time Bookstein (1978) began to study shape-theoretical problems associated with zoology. It wasn't long before a general theory for the analysis of shapes obtained from any environment was sought, Kendall (1984).

Kendall (1984) defined the shape of an object as "all the geometric information that remains when location, scale and rotational affects are filtered out from an object". Since this definition was first published it has been used extensively, with vast progress being made towards the applications of shape analysis, Kendall (1984, 1989), Bookstein (1986, 1991), Mardia and Dryden (1989, 1998), Goodall (1991).

Following Kendall's intuitive definition of shape Dryden and Mardia (1998) further developed the aspects of shape analysis and provided a text that lays the foundations of

the subject, as well as giving practical guidance and comparing a variety of techniques available. They developed means for statistical shape analysis, which involve computational methods for the geometrical study of random objects where location, rotation and scale information can be removed. Work also included advances in distribution theory for shape, Mardia and Dryden (1989).

The steps involved in statistical shape analysis are firstly to record the coordinates of a common set of points on the objects being investigated, these points are known as landmarks. Bookstein (1991) explained the usefulness of landmark data for the analysis of biological shape change. He defined a landmark as "a discrete point that corresponds across all forms of a dataset", there are lots of such points available on the human face, many of which have been defined by Farkas (1994) and are explored in the following subsection. Landmark coordinates contain information on a shape and hold its position in an *n*-dimensional space.

Once sets of landmark coordinates from a group of different objects are obtained Procrustes analysis is applied (§3.4, §3.5, §3.6). Procrustes analysis brings the configuration of landmarks to a common orientation and size, and preserves the 'shape' of each individual object; this overcomes any randomness in the scale, origin and orientation of the object or coordinate system used.

After Procrustes registration of the data, techniques developed by Dryden and Mardia (1998) can be used to estimate a mean shape of a sample (§3.6.2), to assess whether two groups (for example, males and females) are significantly different in mean shape and to carry out discrimination or clustering on the basis of shape and size information, to examine any structure or groups in the dataset (§3.7). The shape space is non-Euclidean, so careful consideration must be taken when looking for appropriate methods of data analysis. In particular, multivariate statistical procedures cannot be applied directly to non-Euclidean information; however in certain circumstances procedures can be adapted for shape data (§3.7).

1.5.2 Possible Facial Landmarks

We can describe shape by locating a number of 'landmark' points on each specimen. Dryden and Mardia (1998) defined a landmark as "a point of correspondence on each object that matches between and within populations". There are a large number of well-documented anatomical facial landmarks, Farkas (1994). Anatomical landmarks are points that correspond between organisms in some biologically meaningful way, for example the corner of an eye. Other types of landmarks, aside from anatomical, are mathematical landmarks (described by some mathematical or geometrical property, e.g. a high point of curvature) or pseudo landmarks (located in between the anatomical and mathematical points, to increase the number of points in an analysis).

Figure 1.4– Traditional anthropometric landmarks of the face (Farkas, 1994)

Farkas (1994) gave anthropological descriptions and diagrams of facial landmarks, Figure 1.4, directions for measurement of the points (by hand with callipers) and possible sources of error associated. Inevitably some anatomical facial landmarks will be easier to locate than others, points such as the corner of the eye or the lip will be

easier to replicate than points situated on a curve, for example the tip of the nose or the most prominent point on the chin.

When choosing a set of landmark points for investigation a key issue is whether the landmarks have to be assigned by an expert anthropologist, or could they be easily placed by an observer with no previous expertise. This is going to be important in the field of facial identification, where observers in the outside world are likely to be police officers or lawyers, and so any methods used to locate landmarks need to be documented with clear instructions and be easily repeatable by a layperson.

1.5.3 Three Dimensional Facial Data

There are many studies out there analysing photographic images, these are primarily studies examining two-dimensional (2D) facial images, for example the photographs taken of criminals in custody. However, the face is a three-dimensional (3D) object, so information may be lost by simply comparing one 2D view of the face, it would therefore make sense to carry out a study on 3D face shape.

Recent technological advances in computers, camera optics and laser rangefinders have made the production of reliable and accurate 3D depth data possible (http://www.cs.cf.ac.uk/Dave/AI2/node174.html). Consequently many 3D data acquisition systems have been developed, meaning that 3D information on objects can now be obtained easily. 3D data is being used in an increasing number of applications, including facial reconstructive surgery, clothing size measurements using 3D information on the full body, and also in the computer game and film industries.

3D facial data is often used in a medical context, for example to examine images before and after an operation. Bowman and Bock (2005) used 3D facial landmark data to assess the degree to which there was mismatch between a landmark configuration and its relabelled matched reflection (i.e. asymmetry). Their interest was in comparing the degrees of asymmetry in different populations, in particular the extent to which this is larger for children born with unilateral cleft lip (UCL) or cleft lip and palate (UCLP) than it is for the wider population. The data used was from a longitudinal study carried out to track changes in subject-specific asymmetry, investigating the facial development

and growth of children aged three months to two years old with UCL or UCLP conditions.

In the field of facial identification Yoshino et al. (1997, 2000, 2002 and 2003) designed and manufactured their own 3D physiognomic rangefinder to capture 3D facial data, as well as carrying out various small studies into comparing 3D to 2D images. Their motivation behind building a new system instead of using available technology was that the operation time to capture the 3D images was around ten seconds for the current technology, and to scan and capture data from suspects in custody it was thought that this was too slow, so Yoshino et al strived to improve on this in their new system.

Yoshino compared 3D facial models to actual 2D facial images by using superimposition techniques, automatic adjustment of the 3D image was carried out to try and match the orientation of the 2D image, this was done based on aligning seven anthropometrical landmark points. To account for differences in scale Yoshino converted the 3D image into a number of pixels and electronically corrected any perspective distortion by inputting the distance of the face of the criminal from the surveillance camera in the 2D image.

Techniques used by Yoshino et al (1997, 2000, 2002) could be carried out mathematically by using statistical shape analysis. To bring the 3D facial model around to the same scale and orientation as the 2D image a number of matrix transformations could be carried out on the facial landmark configurations. This approach would prevent the need to correct for perspective distortion, as the distance of a criminal's face from a surveillance camera is most probably going to be unknown in a practical situation.

1.6 Summary

This chapter has introduced the topic of facial identification and established that current methods of analyses are rudimentary, unscientific and untested. Other forms of forensic evidence such as DNA and trace evidence are presented as probabilistic measures. To develop a reliable quantitative technique for identifying faces would be a great help to the criminal justice system. The means for carrying out a statistical analysis of face shape based on well-defined anthropometrical landmark points, Farkas (1994), already

exist. The recent advances in technology facilitate the possibility of a 3D facial landmark study to be carried out.

There are many criteria that such a technique for facial identification has to follow to be admissible in court. In order for any developed technique to be accepted and adopted by facial imaging analysts, precise protocols and guidelines for the placement of landmark points and the analysis of the resulting coordinate data should be provided. When choosing a set of landmark points for investigation a key issue is that the landmarks have to be easily placed by an observer with no previous expertise. In the field of facial identification, landmarkers are likely to be police officers or lawyers, and not forensic anthropologists.

Previous attempts at the 3D analysis of facial landmark data for facial identification, Yoshino et al (1997, 2000, 2002, 2003), could be improved by using statistical shape analysis and mathematical methods to bring two images around to the same scale and orientation, rather than arbitrarily rotating and scaling.

Recognized techniques in presenting trace and DNA evidence to courts can be explored to attempt to model face shape in a population, and assess how similar two face shapes are in comparison to a population sample by means of a likelihood ratio. A substantial study looking at population variation in face shape in a large sample of people has to be carried out in order to gain better knowledge of the face shape system.

Some problems to address when trying to develop a facial comparison method are properties that will affect the outcome, for example, facial expression and lighting conditions during image capture. To come up with a method that will be invariant under such properties seems impractical, in that not enough is known about the shape of the face in a controlled environment without bringing in additional factors. What is needed is a large study examining face shape controlling for a ‘natural’, Farkas (1994), facial expression and constant camera and lighting conditions. Once it is known how the face varies in shape between individuals under these controlled conditions the extra factors could then be brought in.

Any statistical methods used to affirm or reject a facial match to aid in facial identification in court have to be easily explainable to the judge and jury, who will most

probably have no expertise in the field of statistics and probability. More importantly the methods must be robustly tested and subjected to peer review in order for them to become accepted within the forensic and statistical scientific communities.

2 The Data

2.1 Introduction

This chapter describes the various sets of data used in this study, including how the facial images were obtained, how measurements were made on these images and what opportunities and limitations this permitted. To permit the likelihood of a facial match, or exclusion, to be empirically established methods need to be based on precise empirical measurements. A substantial study looking at such measurements in a large sample of people is needed to gain better knowledge of facial variation. Two faces are enough to answer the question of how similar two face shapes are, however, population estimates of facial variation are required to answer to the question of how many other face shapes are also likely to be as similar. The following chapter describes the facial landmark data collected (i.e. measured) from various sets of 2D and 3D photographs in order to define face shapes.

An explanation of the anthropological facial landmark data to be collected is given (§2.2). A small subset of size ten of these landmarks (§2.3) was collected for a pilot study (§4) carried out to see whether the methods proposed for facial matching (§3) were suitable for use with landmark data. The large facial image database utilized for the main study of this research is described fully (§2.4), including how images were collected using a Geometrix® 3D scanner (§2.4.2) and how the landmark coordinates were measured on these images (§2.4.3). This main database is used to obtain estimates of facial shape variation in the population (§6) in order to quantify likelihoods of facial matches (§7, §8). Subsets of the large database were selected to carry out some preliminary studies. One subset was used to check the reliability of the landmark data and chose a set of landmark points which were the most appropriate for facial matching (§2.5, §5.2). Another subset was chosen to validate that the data collection procedure was repeatable when multiple observers took the landmark measurements from the facial images (§2.6, §5.3). Details of some additional datasets are also given; these data were used for testing facial matching techniques (§2.7, §7, §8).

2.2 Anthropological Facial Landmark Data

We can describe shape by locating a number of points on each specimen, which are called landmarks. A landmark is a point of correspondence on each object that matches between and within populations. In this study anatomical landmarks were used, these are points that correspond between organisms in some biologically meaningful way, for example the corner of an eye. Farkas (1994) describes many anatomical facial landmarks in his anthropological survey of the head and neck (§1.4.2, Figure 1.4). Such landmarks describe physical characteristics, which are an important aspect of anthropology.

Examination of the facial image database (§2.4) was carried out to see which landmark points would be suitable for collection from the available images. Certain points had to be excluded due to their location being largely determined by the bony structure underneath the face. Previous facial anthropological studies have been carried out on live subjects, as opposed to facial images, where the position of landmarks can be physically determined by the person collecting the data. One landmark point was excluded as it was located on the hairline, for many subjects the hair covered the position of the landmark, or the lack of hair on some subjects made it impossible to judge the location.

Other factors taken into consideration when choosing which landmarks to collect for the main study were the fact that the facial image database (§2.4) was large and the landmark collection procedure was carried out manually, as described in §2.4.3. One observer collecting all these data would have taken a long time; to speed up the data collection multiple observers were available. This then brought in another issue that landmark points chosen had to be easily distinguishable and apt to minimum subjectivity between different observers (§5.2).

Initially sixty-one facial landmarks were considered, Table 2.1. For the pilot study (further details in §2.3 and chapter 4) only ten of these landmarks were investigated, Figure 2.2. For the main study (§5-§8) the list in Table 2.1 was reduced to thirty landmarks for collection on the Geometrix® database (§2.4.3, §5.2.6), these landmarks

were chosen after a preliminary study examined the consistency of the data collected by two observers and the discriminatory power of the landmarks (§2.5, §5.2). The instructions for the placement of landmarks were taken as the detailed descriptions provided by Farkas (1994). When "left" and "right" are discussed, this is standard anatomical siding, therefore refers to the left and right of the subject, not as the pictures are viewed by the observer, e.g. we refer to the subject's left eye even though it appears on the right hand of the page. Clearly this is a potential source of error for novice landmark observers. Two different observers followed Farkas' descriptions to collect data on a number of faces (§2.5). Anything that was unclear in the descriptions was noted down, along with any extra information which was seen as useful.

Following the work of the preliminary study (§5.2) a more extensive landmark placement manual was written by the two observers to assist new users in placing the thirty landmark points (§5.2.6, Table 5.3). The manual included Farkas' descriptions, any additional information from the observers and also detailed images of the landmark location (Appendix B).

Landmark	Name	Landmark	Name
1	Glabella	32	Subalare Right
2	Gonion Left	33	Alare crest Left
3	Gonion Right	34	Alare crest Right
4	Sublabiale	35	Highest point of columella prime Left
5	Pogonion	36	Highest point of columella prime
6	Gnathion	37	Crista philtri Left
7	Endocanthion Left	38	Crista philtri Right
8	Endocanthion Right	39	Labiale superius
9	Exocanthion Left	40	Labiale superius prime Left
10	Exocanthion Right	41	Labiale superius prime Right
11	Centre point of pupil Left	42	Labiale inferius
12	Centre point of pupil Right	43	Stomion
13	Oribitale Left	44	Cheilion Left
14	Oribitale Right	45	Cheilion Right
15	Palpebrale superius Left	46	Superaurale Left
16	Palpebrale superius Right	47	Superaurale Right
17	Palpebrale inferius Left	48	Subaurale Left
18	Palpebrale inferius Right	49	Subaurale Right
19	Orbitale superius Left	50	Postaurale Left
20	Orbitale superius Right	51	Postaurale Right
21	Superciliare Left	52	Otobasion superius Left
22	Superciliare Right	53	Otobasion superius Right
23	Nasion	54	Otobasion inferius Left
24	Subnasion	55	Otobasion inferius Right
25	Maxillofrontale Left	56	Porion Left
26	Maxillofrontale Right	57	Porion Right
27	Alare Left	58	Tragion Left
28	Alare Right	59	Tragion Right
29	Pronasale	60	Preaurale Left
30	Subnasale	61	Preaurale Right
31	Subalare Left		

Table 2.1 – List of anthropological landmark points collected for initial analyses, Farkas (1994).

2.3 Pilot Study - The FBI Catalogue of Facial Types

The image data for a small pilot study (§4) consisted of anterior 'mugshot' photographs of suspects in custody. The data were taken from the FBI Facial Identification Catalogue, anonymous (1988). This catalogue was published during the nineteen eighties and contains around one thousand images of various offenders. It was designed to help witnesses pick the 'best' likeness of a face and is set out in sections containing images with a portion of the face censored; different sections illustrate different characteristics of the face, Table 2.2. Each catalogue page has a specific facial feature to illustrate, e.g. round face, close set eyes, square chin, thick lips etc. This was thought to be an additional benefit, as in real life situations criminal perpetrators could use masking to obscure facial and cranial features, other studies have examined similar data, Yoshino et al (2002). It was known that the catalogue contained repeats of some faces, where the same person was photographed for two or more different sections of the catalogue. It was not known which faces may appear more than once, however facial matches picked out by any matching method could then be confirmed visually.

Catalogue Section	Facial Characteristic	Landmarks Visible
A	Head shape	1,2, 3, 4, 5, 6
B	Eyes	1, 2, 3, 4, 5, 6, 7, 8
C	Eyebrows	1, 2, 3, 4, 5, 6, 7, 8
D	Nose	1, 2, 3, 4, 7, 8
E	Lips	1, 2, 3, 4, 9, 10
F	Chin	1, 2, 3, 4, 9, 10
G	Cheek and cheekbones	1, 2, 3, 4, 9, 10
H	Ears	1, 2, 3, 4, 9, 10
J	Hair	1, 2, 3, 4
K	Mustache and beard	1, 2, 3, 4, 9, 10
L	Facial lines	1, 2, 3, 4, 9, 10
M	Scars	1, 2, 3, 4, 9, 10
N	Forehead	1, 2, 3, 4, 9, 10
O	Skin irregularities	1, 2, 3, 4, 9, 10

Table 2.2 – Facial characteristics depicted in each catalogue section and the facial landmarks (Figure 2.1) visible in the images.

The images in the dataset may have been taken under consistent conditions, however the subject to camera distances seem to vary a great deal indicating the images may have been rescaled for the catalogue. This makes these images more akin to the kind of data

likely to be obtained in real life facial comparison cases. Only one facial match was found in this dataset and this was determined visually, i.e. we have no confirmation that it is a match, however in the opinion of the author the faces look very alike (§4.3.2). The profile views of the face were not given in the catalogue, so landmarks around the nose area could not be included in the pilot study. Ideally an anterior view and two different profile views are needed to accurately locate the tip of the nose, a limitation which was established when the variation of landmark positions were investigated for a repeatability study (§2.5, §5.2.2).

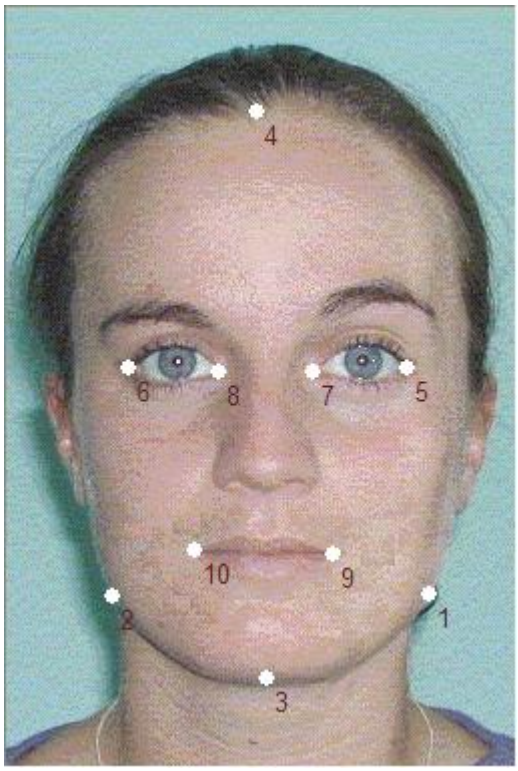


Figure 2.1 – 10 anthropometrical facial landmarks used to represent face shape

To collect landmark measurements for the pilot study 480 facial images from the facial catalogue were digitally scanned using *Adobe Photoshop 7.0* at resolution 720 dpi and stored on a computer. A sample of sixty of these two-dimensional (2D) anterior facial images was used in chapter 4 to illustrate the different sources of variation (§4.3.1) and to try out the likelihood ratio facial matching method (§3.8.4, §4.3.2). Ten anthropometrical landmarks were used to determine facial shape, Figure 2.1. These were the right and left Gonion (labelled 1 and 2), the menton (3) the trichion (4), the right and left Exocanthion (5 and 6), the right and left Endocanthion (7 and 8) and the right and left Cheilion (9 and 10). Descriptions of these landmarks and instructions for locating them were obtained from Farkas (1994). Similar landmark points have been used in studies to look at facial variation in a medical context, Bock and Bowman (2005) investigated facial shape variation and asymmetry in children born with cleft lip or cleft lip and palate when compared to a control group.

Due to the image censoring, each photograph only had a subset of the ten chosen landmarks which were visible, Table 2.2 lists the landmarks (Figure 2.1) which were visible in images from each section of the catalogue. *tpsDIG32.exe* (freely available to download from http://life.bio.sunysb.edu/morph/) was used to capture the 2D

coordinates of the landmarks on the digitised photographs. A visual identification of the landmark was followed by positioning a cross-hair controlled by a mouse on the appropriate location, left-clicking the mouse automatically captured the six-figure coordinates and stored them in a data file. One observer collected the landmark data for this study, the analysis and results of which are in chapter 4.

2.4 Main Study - The Geometrix® Facial Image Database

For the main study on facial identification 2960 different people volunteered to have their faces digitally scanned in three-dimensions (3D) as part of the IDENT research project carried out by the University of Sheffield and sponsored by the United States government on behalf of the FBI (Evison and Vorder Bruegge, 2008). The data used in the current study were collected using a 3D scanner made by Geometrix®, and is hence referred to as the Geometrix® database. The IDENT research project (Evison and Vorder Bruegge, 2008) also collected data using a different 3D scanner made by Cyberware®; these data were not used in the current study. The scanning took place at Magna science adventure centre in Rotherham, England, between 20/12/03 and 19/04/05. Visitors to the science centre were asked if they would like to volunteer to take part in the study. All volunteers were given information describing how their image would be included in a research database, which would be used for research undertaken in the field of crime prevention and detection. Each person photographed filled out a questionnaire (Appendix A) indicating their consent, as well as some biographical information: age, sex and ethnicity. The study passed full ethical approval from the University of Sheffield ethics committee.

The science centre attracted many families; volunteers were also therefore asked whether any of their known relatives had also taken part in the study, since there are thought to be facial similarities in people with the same genetic make up. No information regarding the relationship between relatives was recorded, only an indication of whether or not they were related. Volunteers over the age of fourteen years were kept for use in the facial database and anyone under the age of sixteen was asked to obtain parental consent before being allowed to participate. It is unclear how much the face shape will change from childhood to adulthood, however as facial identification is used in an evidential capacity to solve crimes the majority of criminal perpetrators are

between sixteen and thirty years old, hence the younger ages are of interest in the investigation.

2.4.1 Exploratory Data Analyses

The total number of faces, *n*, in the database was 2960. Tables 2.3, 2.4 and Figure 2.2 display summary information with regards to the ethnicity and age of the faces in the database.

Table 2.1 shows that in general more males than females volunteered to take part in the facial database. The majority of people to volunteer were white British. There are known differences in facial morphology between different ethnic groups, so a predominantly white British database will not give an accurate representation of faces in the general population. However, if facial matching results for this database are promising it would be worth extending the work to look at additional ethnic groups.

Code	Ethnicity	Males	Females
01	White British	1464	1209
02	Other White Background	77	54
03	White and Black Caribbean	2	4
04	White and Black African	2	1
05	White and Asian	6	3
06	Other Mixed Background	3	6
07	Indian	17	14
08	Pakistani	10	4
10	Any Other Asian Background	11	6
11	Caribbean	7	2
12	African	7	7
13	Other Black Background	1	3
14	Chinese	15	18
15	Any Other	3	4
	Total	**1625**	**1335**

Table 2.3- Summary data, numbers of faces in database by sex and ethnic group

Figure 2.2 shows that the majority of all participants were aged between thirty and fifty years, as the science centre attracted many families with parents in this age group. It is thought that the most common age group for criminals is more like 16-30, so this sample may not represent the appropriate population to target for confirming criminal identity through facial identification. If the results prove successful for this database the likelihood is that looking at an averagely younger sample of faces would also be successful.

Out of the 2960 faces in the database, 463 had at least one relative who also appeared in the database. Although not investigated here, it would be of interest to explore whether people who were genetically related had more similar face shapes than people who were unrelated.

Age Group	14-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65+	Total
Males	110	69	103	185	320	358	180	109	65	51	75	**1625**
Females	142	72	98	172	290	243	125	53	55	44	41	**1335**

Table 2.4 - Summary data, numbers of faces in database by sex and age group

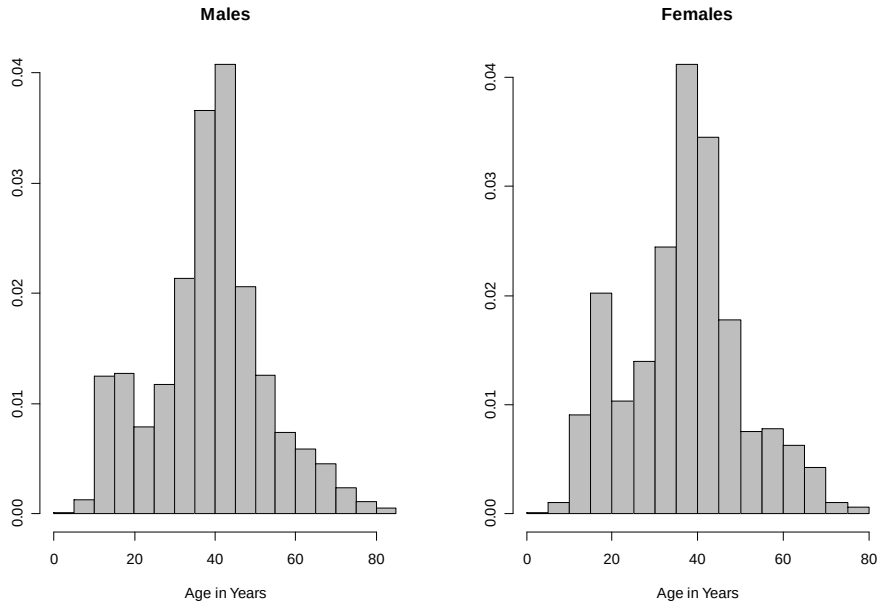


Figure 2.2 - Histograms to show distribution of subject age, by sex

2.4.2 The Image Collection Procedure

The facial scans were taken with a Geometrix® FaceVision802 3D scanner and related software, Figure 2.3. FaceVision802 is a digital stereographic device with eight digital cameras positioned at varying angles along a gantry. It simultaneously takes eight high quality two-dimensional (2D) digital photographs of the subject, resulting in eight different photographic views of the face (Figure 2.4). The scanner is calibrated prior to use, to measure information on the relative camera positions and angles. A computer program then uses the system calibration information to render the set of eight 2D photographs into a 3D facial model. A screen still of a computer-generated 3D facial model can be seen in Figure 2.5, this model was derived from the 2D images in Figure 2.4.

For the purposes of analysing face shape the 3D models were not used in this research. Instead the points on the face were located on the 2D images using software that was included with the Geometrix® scanner to determine the 3D position of points (further details in §2.6).

Figure 2.3 - Geometrix® FaceVision802 3D digital stereographic scanner.

Figure 2.4 - Set of eight 2D digital photographs taken by FaceVision802, these and system calibration information are used to produce the 3D facial model.

Figure 2.5 - Screen still of the 3D model produced by the FaceVision802 scanner.

There were ten different scanner operators (photographers) collecting image data, unfortunately this was not controlled for in the experiment design. The effect of scanner operator could have been measured if the ten different photographers had each taken scans of a set of faces, in a similar way in which the landmark observer error was investigated in §5. 3.

As far as possible the distances from cameras to subject head were kept consistent, as were the lighting conditions and facial expressions of the subjects. Table 2.5 summarizes the number of scans captured by each of the photographers (A-J) who captured images. Some scans were captured by a team of two photographers (e.g. A/D was captured by photographers A and D) with one person helping the subject get into the correct position and the other controlling the computer, it is not clear which photographer performed which task.

Photographer	Number of Scans Captured
A	105
A/B	5
A/D	82
A/E	6
A/H	15
A/J	41
B	149
B/D	17
B/E	15
B/J	1
C	3
C/D	2
D	785
D/E	13
D/F	3
D/H	92
D/I	117
D/J	14
E	151
F	624
G	53
H	73
H/J	21
I	216
I/J	12
J	566
Total	3181

Table 2.5 – Number of images captured with the Geometrix® scanner by the ten different photographers (A-J)

There were a number of rules which the photographers followed to ensure the subject being scanned was in the correct position. These included measuring the distance of the subject to the central camera pair (50cm), asking the subject if they could see their own eyes in a small mirror located between the two cameras in the central pair, ensuring the subject had their hair behind the ears where possible and asking the subject to remove

spectacles if they were wearing them. In addition just before the scan was taken the subject was asked to look straight ahead with a natural facial expression (place their lips together and keep a relaxed jaw) and hold this position for five seconds until the photographer had finished taking the scan.

2.4.3 The Collection of Landmark Data

To obtain the 3D coordinates of facial landmarks the Geometrix© Forensic Analyzer program was used to manually place points onto the eight 2D digital photographs of each face in the database, e.g. Figure 2.4. This software was produced specifically for use with images acquired by the Geometrix© FaceVision802 3D scanner. The 3D model generated by the scanner (Figure 2.5) was not used to measure for landmark points, a separate study (Schofield and Goodwin, 2006) indicates that the 3D surface modelling is inaccurate, as there appears to be a fault in the Geometrix® internal procedures. The Forensic Analyzer program interpolates the 3D landmark positions from the 2D images and there is no reason to doubt the accuracy of this. Both of these results were determined by calliper measurements (Schofield Pers. Comm.).

The landmarks were positioned manually by an observer who placed a cross-hair over the landmark point and clicked the mouse. This procedure was carried out twice per landmark point, the point being located in two separate 2D images, which portrayed two different facial views, i.e. from different cameras. The Forensic Analyzer program uses 3D geometry along with the scanner calibration information to triangulate between the two different sets of 2D coordinates and produce one 3D location for the point in the form of (x, y, z) coordinates. Once the location of the point had been selected in one 2D image the Forensic Analyzer software draws a line along which the point should be in the second image to help the observer find the location, Figure 2.6.

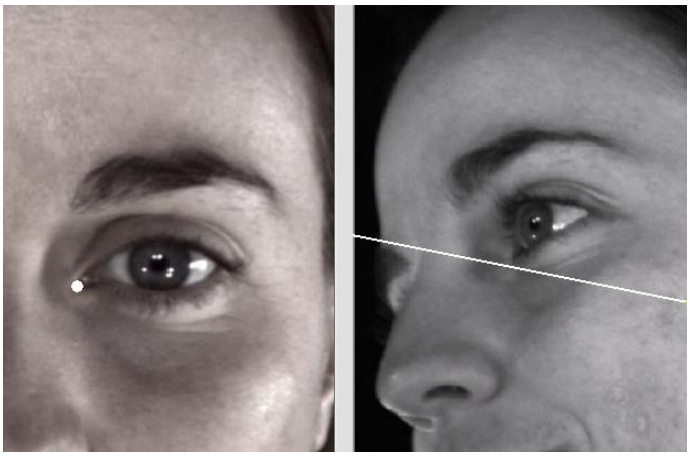


Figure 2.6 - Triangulation of the location of the endocanthion landmark point in two 2D images; the Forensic Analzer program obtains the 3D data from two 2D images and scanner calibration information.

Initially sixty-one landmarks (Table 2.1, §2.5, §5.2) were investigated as potential measurements to take for comparing faces, a subset of thirty of these was chosen (§5.2.6, Table 5.3) for collection from all 2960 faces in the Geometrix® facial database. For each of the images the landmark locations were obtained twice by either the same or different observers, there were six different observers. The multiple sets of measurements enabled an assessment of the variability between observers and facilitated the use of the likelihood ratio statistic to carry out facial evidence evaluation (§3.8).

There were ten photographers capturing the images using the scanner as described in §2.4.2. There were six observers (1 – 6) collecting the landmark points from the images, four of these had also captured some 3D scans, the other two had not. Prior to data collection a small study was carried out to validate the technique and ensure all observers were collecting comparable landmark data, this study is outlined in §2.6 and the results are in §5.3.

Observer	Number of Images Measured	
	1st measurements	2nd measurements
1	288	67
2	1017	579
3	459	768
4	72	0
5	733	899
6	601	857
Total	3170	3170

Table 2.6 – Number of images measured by each observer

Table 2.6 summarizes the number of images that each observer placed the landmark points on; a key limitation is the imbalance in the experimental design. The allocation of images to observers was arbitrary based on the availability of the photographers; this was due to pressures of the study and demands from the clients to meet monthly deadlines. Although the necessary deadlines were met assessments of variability could have been more efficient in terms of duplications had the design been more orthogonal. Each image had the landmarks collected in duplicate, some images had the same observer collect both duplicate measurements and some images had two different observers collecting the two duplicate measurements.

2.5 Reliability Study

The study presented in chapter 5, §5.2, was performed to assess the reliability of potential landmarks. The set of $k = 61$ facial landmarks from Table 2.1 were obtained in 3D from scans of a small sample of thirty five different faces taken from the Geometrix® facial image database collected for the main study (described in §2.4.3). The landmark locations were placed on each image three times by each of two observers (L and X). So, in total there were $n = 210$ landmark configurations.

The reliability of the sixty-one landmarks was assessed in terms of the consistency of which the observers placed the points and also the ability of each point to discriminate between different faces. From the results the landmark list was reduced to thirty (§5.2.6, Table 5.3) for collection from the main database of facial images (§2.4). A detailed landmark placement manual (Appendix B) was produced for the thirty chosen landmarks to ensure all available observers followed the same procedures.

2.6 Validation Study

As described in §2.4.3 there were six different observers available to carry out the placement of 3D landmark points on the images in the Geometrix® facial database. Prior to data collection a small subset of faces were used as a test dataset to check the repeatability of the landmark collection technique (§5.3). The six observers were each given a copy of the landmark placement manual (Appendix B). The manual contained clear instructions and illustrations describing the location of each of the thirty landmarks (§5.2.6, Table 5.3) to be collected from the main database of images (§2.4). Each observer was also given a tutorial on how to use the Forensic Analyzer (Geometrix©) software for placing landmarks (§2.4.3). Of the six observers, two had more experience (around six months) in the landmark placement procedure through previous work. The other four were of the same 'beginner' standard.

The data were collected from a subset of ten facial images from the Geometrix® facial database (§2.4). Each observer placed thirty landmark points on each image; they repeated the process three times to enable the assessment of consistency of each observer. All six observers placed points on the same ten faces, so inter-observer consistency could also be assessed. The total landmark dataset consisted of $n = 180$ configurations (thirty from each observer) on $k = 30$ landmark points in $m = 3$ dimensions. It was found that all six observers were producing comparable data (§5.3).

2.7 Other Image Data for Testing Facial Matching Techniques

The Geometrix® facial database (§2.4) is what makes up a population sample of face shapes from which a suitable statistical model (§6.4) can be found and the population variation can be explored. To use this database to carry out the comparisons of different faces some data are required where certain pairs of images are of the same people (known facial matches) and certain pairs of images are of different people (known facial exclusions).

Various data were available to test facial comparison methods (§2.7.1 - §2.7.6), some of which was simulated using the collected main database (§2.7.1, §2.7.5) and some of which was obtained externally from different sources (§2.7.2, §2.7.3, §2.7.4, §2.7.6). These data sets are:

- Test Data 1 – Matching Data from Two Observers (§2.7.1)
- Test Data 2 - FBI Suspects Data (§2.7.2)
- Multiple Images of Agent Vorder Bruegge (§2.7.3)
- Known Matches and Exclusions for Subset Selection (§2.7.4)
- Twins and Controls (§2.7.5)
- Other Data from Multiple Images of Like Faces (§2.7.6)

Key information regarding the number of faces, observers, replicated measures and landmarks used in each dataset are given in §2.8.

2.7.1 Test Data 1 – Matching Data from Two Observers

Several particular small data sets were used for specific aspects of the analysis described later in §7 - §8. The first, Test Data 1, was considered to imitate some data for facial matching a sample of ten different facial images were taken from the Geometrix® database (§2.4), Table 2.7 lists the subject IDs for these images. Twenty-two anterior landmark points (§7.5.1, Table 7.8, Figure 7.3) were placed on each of these images by two different observers, the measurements were taken twice giving in total four measurements taken from each facial image. Instead of treating each image as having four measurements the data from each observer was treated as being from a different source, so in total there were twenty test faces, i.e. observer one placed two measurements on faces 1-10 and observer two placed two measurements on faces 11-20. Table 2.7 links faces 1-20 (Faces) with subject (Subject ID). To perform the multivariate normal likelihood ratio (MVNLR) procedure (§3.8.4) to compare these data and quantify facial matches meant that there were ten known matches, as faces 1 and 11; 2 and 12; ….. ; 10 and 20 were the same individual.

The results of the facial comparisons of Test Data 1 using the MVNLR procedure (§3.8.4) are given in §7.2.1.

Faces	i	j	Subject ID
1, 11	1287	1297	110104_00018
2, 12	1288	1298	110104_00019
3, 13	1289	1299	110104_00020
4, 14	1290	1300	110104_00022
5, 15	1291	1301	110104_00024
6, 16	1292	1302	110104_00026
7, 17	1293	1303	170104_00002
8, 18	1294	1304	170104_00007
9, 19	1295	1305	170104_00009
10, 20	1296	1306	170104_00010

Table 2.7 – Measurements i and j from observers 1 and 2 respectively. Measurements were taken from the ten faces (subject IDs) used as Test Data 1.

2.7.2 Test Data 2 - FBI Suspects Data

Test Data 2 consists of 2D measurements of anterior facial images (Confidential Appendix C, Figures 13.1 – 13.2). It was sent by the FBI for the purposes of testing out facial matching techniques on real-life images. Included in the dataset were multiple images of FBI agents, actual criminal case photographs from suspects in custody and driving license images, so images were not obtained under controlled conditions. The images were captured both in the past and fairly recently. Notes accompanied each image to state which other images in the dataset it was known to match with. It was presumed that if a match was not stated then two images were a known exclusion. There were also some 'supposed' and 'possible' matches, where there was not enough evidence to convict the person either way, yet a match was thought supposed or possible.

There were sixty-seven images in the FBI anterior dataset, each had twenty-two anterior landmark points (Results Appendix D §14.1.1, Table 14.1, Figure 14.1) placed onto them if the location was able to be determined from the photograph. All images in this test data had the landmark points positioned in duplicate by one observer. On inspection of the landmark data there were many missing values, where for one reason or another certain landmarks could not be clearly determined in the facial images, this was one of the issues when using images captured in real life as opposed to under controlled conditions of facial expression, lighting, subject distance from camera, making sure hair did not obstruct landmarks etc. It was important with this dataset that scale was removed during Procrustes alignment (§3.4, §3.6), as when extending to actual 'live' photographs the distances of subject to camera are unknown.

The prerequisite for Procrustes analysis (§3.4) that all landmark configurations must be complete meant further inspection of incomplete values was required. Any subjects that had incomplete values in their configuration were removed from the dataset analysed for facial matches. The remaining data consisted of configurations of twenty-two 2D landmark points (Appendix D §14.1.1, Table 14.1, Figure 14.1) for sixty different faces (of which there was a list provided by the FBI detailing where there were known or supposed matches between images), with two sets of landmark measurements for each face. A few anomalies were found in the data for these sixty faces (§7.2.2.2), two of which could not be corrected for. Thus two faces had to be dropped leaving fifty-eight FBI anterior faces to test the MVNLR procedure for facial matching (§7.2.2.3) and assess how well different subsets of matching variables (§7.4.2, Appendix D) performed in terms of the true and false rates (positive and negative) quantified by the likelihood ratio (LR) results for the known matches and exclusions in the dataset (§7.4.2.3, Appendix D).

A number of different comments were given to each of the found matches after examination of the source images, these were: 'possible' match, given when the image was a case comparison with another image, but it was not known from the FBI notes whether there was a definite match; 'yes' was given when the images were known to be a definite match; 'supposed' was given when it was not definitely known the images match, though it was supposed that they do; 'no' was given when there was a false positive result, i.e. when it was known that the images did not match, finally 'unverifiable' was given when, from the available notes, it was thought that the images wouldn't be matches and the original sources (images) were unavailable to quantify the results, as data from these images were added to the landmark database and were additional images not supplied by the FBI.

2.7.3 Multiple Images of Agent Vorder Bruegge

A third subset of the FBI anterior test data (§2.7.2) consisted of multiple images of an FBI agent who was involved with the IDENT project (Evison and Vorder Bruegge, 2008) and the collection of the Geometrix® data. The images were taken at different times, in particular one was taken from the agents driving license and was several years older than the other image. The images had a variety of facial positions, where slight rotations of the head or shoulders were apparent. There were also some images which

showed a different facial expression to the 'natural' look that the volunteers for the Geometrix® database were asked to maintain; two faces were smiling showing the teeth. In addition to this there were two other photos where agent Vorder Bruegge wore glasses.

A selection of ten images of agent Vorder Bruegge (Appendix C, Figure 13.3) that were known facial matches were selected from the sample of FBI anterior images (§2.7.2) to illustrate the effect of increasing the number of PCs used as the *p* matching variables on the resulting LRs (§7.2.3).

A selection of fourteen images of agent Vorder Bruegge, (Appendix C Figure 13.30) were also used in Chapter 8 to investigate how well the matching method performed when looking at different images of the same face (§8.4).

2.7.4 Known Matches and Exclusions for Subset Selection

To explore potential subsets of matching variables (§7.4) a small subset of the FBI anterior images (§2.7.2) was selected. This subset consisted of ten known pairs of facial matches and ten known pairs of exclusions (§7.4, Table 7.7, Confidential Appendix C Figures 13.4 – 13.23). 'Good' subsets were chosen as ones that correctly identified the known matches and exclusions with a good strength of evidence, as indicated by the LR. The analyses and results for all subsets investigated are described in §7.4.2 and Appendix D.

2.7.5 Twins and Controls

The main facial image database collected with the Geometrix® scanner (§2.4) was known to contain three pairs of twins (two identical and one non-identical, this was confirmed visually from the images). Obviously twins are well known to look similar to one another and these data were used to test whether the facial matching procedure was able to distinguish between twins (§8.2). Additionally three pairs of non-related controls were also taken to match the sex and age of the twins. All images of twins and controls are displayed in the Appendix C, Figures 13.24 – 13.29. The interest lay in how strong the evidence for a facial match was for the identical twins in comparison to the non-

identical twins and also the controls using the LR facial matching method (§3.8.4) with 'best' found subset of matching variables (§7.4.2.3). Results of these analyses are found in §8.2.

2.7.6 Other Data from Multiple Images of Like Faces

Some alternative data obtained from a different source was used to test the matching methods developed (§8.4) by comparing multiple images of five faces. The photographs were taken, as part of the IDENT project (Evison and Vorder Bruegge, 2008), at different time periods across one day. The location of subject to camera was kept consistent, as were the lighting conditions. The data consisted of five different faces (A, B, C, D and H); each face had been photographed, in the 2D anterior view, three times at different intervals over a period of a day (photos 0, 1 and 2), this gave a total of fifteen facial images: Appendix C, Figure 13.31.

Some different software was used to collect the landmark data; this was specifically written by the IDENT project (Evison and Vorder Bruegge, 2008) to deal with 2D digital images. Three different observers each placed eleven anterior landmark points (§7.4.2.1) on each of the fifteen images; they repeated this process three times for each photo (reps 0, 1 and 2), giving a total of forty-five configurations per observer and a grand total of 135 configurations for analysis. It should be noted that the three observers who collected this data had not collected any of the data for the main study (§2.4).

The data were used to examine various different things (§8.4). The performance of the LR matching procedure was assessed in terms of how well different images of the same face matched and also how well configurations from different observers placing landmarks on the same photograph matched. The source of the data meant that the developed facial matching techniques could be checked to see whether bringing in data collected from different software by additional observers affected the matching results. The results of these analyses are found in §8.4.

2.8 Key Information for Dataset Variables

The following table summarizes the key information for each data sample used.

Dataset	Facial Images	Replicates	Observations	Observers	Landmarks	Dimensions
Main Background Data	3170	2	6340	6	30	3
Complete Background Data		2	3254	6	30	3
Complete Replicated Background Data	1286	2	2572	6	30	3
Reliability Study	35	3	210	2	61	3
Validation Study	10	3	180	6	30	3
Test Data 1	10	2	40	2	22	3
Test Data 2	67	2	134	1	22	2
Multiple Images of Agent Vorderbrugge	14	2	28	1	22	2
Known Matches and exclusions for Subset Selection	40	2	80	1	22	2
Twins and Controls	12	2	24	6	22	2
Other multiple images of like faces	15	3	135	3	11	2

Table 2.8 – Summary of datasets used throughout this thesis

2.9 Summary

This chapter has fully described the wide range of facial image and landmark data available for use throughout this research. Descriptions of anthropological facial landmarks which may be suitable for the comparison of face shapes have been given (§2.2). Chapter 3 outlines the statistical theory needed to extract facial shape data from landmark coordinate data (§3.4 - §3.6). Methods for modelling the extracted shape data as a multivariate normal distribution and using the associated model parameters to estimate the likelihood of two facial shapes being quantified a ‘match’ or ‘exclusion’ are also given (§3.8.4).

An initial set of sixty-one facial landmarks to explore for facial matching have been described (§2.2). Chapter 5 explores the variation in these landmark points (§5.2) and chooses a subset of thirty points for data collection from the main database (§2.4.2). A manual (Appendix B) to assist observers in locating the points effectively was written. The collection technique for the thirty chosen points was validated to ensure that data collected from different observers was comparable (§5.3).

A small subset of ten facial landmarks were collected for a pilot study (§2.3, §4) carried out to confirm that the methods proposed for facial matching (§3) were suitable for use with landmark data.

The large Geometrix® facial database (§2.4) facilitates the calculation of population estimates of facial variation (§6), which can be used to quantify the likelihood that two faces 'match' (§7, §8). Descriptions of how the images were collected (§2.4.2) and how the landmark coordinates were measured on these images (§2.4.3) have been provided. There is some bias in the sample in terms of the ethnic distribution of the data (§2.4.1). The majority of faces are of white British ethnicity, therefore not an accurate cross representation of the general population. This is important here because the data used to test the methods was obtained from images sent from the USA. Also the average age of the faces in the sample may not represent the target criminal population for facial identification cases (§2.4.1). As long as we are conscious of these limitations, if the results for the available data are reasonable then further exploration of different ethnic and age groups could be carried out at a later stage.

There are some observed flaws in the design of experiment for the collection of the main facial database. There were ten different photographers collecting image data and this was not controlled for. There was also imbalance in the numbers of images measured by each landmark observer (§2.4.3). The pressures of the study in terms of client demands and deadlines meant that the data collection was done as quickly as possible using any available photographers or landmarkers. Ideally two observers collecting the points on all of the images would be the best option. Alternatively if each observer were given a set of faces and they were responsible for taking both sets of the two landmark measurements from the set then inter-measurement variability for each observer could have been measured. Another flaw in the design of the whole study is that the facial matching has only been carried out in the 2D anterior view. On

examination of images from different views of the face it is much harder to visually confirm alike faces when two images have been taken at different angles.

Other image and facial landmark data have also been described (§2.5, §2.6, §2.7). A subset of the main Geometrix® database was selected to check the reliability of the landmark data and chose a set of landmark points which would be the most appropriate for facial matching (§2.5, §5.2). Another subset was chosen to validate that the data collection procedure was repeatable when multiple observers took the landmark measurements from the facial images (§2.6, §5.3). The remaining data in the chapter (§2.7) have been described for testing the performance and accuracy of the developed facial matching methods (§7, §8). These data are from other sources external to the main Geometrix® data and consist of multiple images or landmark measurements of the same face, i.e. are known facial matches or known facial exclusions.

Another limitation with the landmark data collection is that the Forensic Analyzer® program could only be used with Geometrix® image data (§2.4.3). Therefore landmark data from the other images (§2.7) were collected using different software developed by the IDENT project (Evison and Vorder Bruegge, 2008), 2D images were imported and the mouse positioned and clicked over the landmark location as in Forensic Analyzer®. This could bring in additional error; however it is necessary for developed techniques to be able to handle facial landmark data acquired from different sources and the proposed methods for the analysis and comparison of shapes (§3) should be able to contend with such data.

3 Statistical Methods

3.1 Introduction

This chapter outlines some of the statistical methods relevant for use with the facial landmark data described in the previous chapter. The first part of the chapter summarizes the substantial theory behind statistical shape analysis, highlighting techniques that are useful for analysing landmark data. In particular various methods based on Procrustes superimposition for the extraction of shapes are examined (§3.4, §3.5, §3.6); the theory is for the most part taken from Dryden and Mardia (1998). The latter part of the chapter describes a technique for evaluating multivariate evidence (§3.8); this theory is taken from Aitken and Lucy (2004).

Some more traditional methods used for the analysis of shape are briefly discussed. An explanation of how Procrustes analysis came to get its name (§3.2) is followed by the theory behind the methods it uses (§3.3). The concept of defining a shape in terms of landmark points is described (§3.4.3). A formal definition of shape is given and a summary of the steps involved in Procrustes methods (§3.4.2 - §3.4.9). Ordinary Procrustes analysis (suitable for use when matching two shape configurations) is explained (§3.5) and the direct generalization of this to generalized Procrustes analysis, which deals with matching many configuration matrices (§3.6). Details of how to derive the full Procrustes coordinates (§3.5.1, §3.6.1) and an estimate of mean shape (§3.6.2) are also given, along with an explanation of how the data are projected onto a tangent plane of the shape space (§3.7). Tangent space is a linearization of the shape space and so standard multivariate methods can be used with tangent data as a good approximation to the shape data.

After the extraction of the facial shape data a method is required to compare different shapes. A description of five different techniques for evaluating multivariate evidence such as the tangent space coordinates of the facial data is given (§3.8.3). One of these methods was chosen as the most appropriate for facial matching; this involved the evaluation of likelihood ratios using a multivariate normal model to obtain the likelihood that two faces were more similar to each other than they were to all other faces in the known population sample (§3.8.4). The theory is based on Aitken and Lucy

(2004), where they used an example involving evidence related to the composition of glass fragments. For this research we will adapt and extend these methods to apply them to facial identification evidence, chapter 7. The extension of the theory was required to accommodate the large number of variables available in the facial data and the low dispersion in the shape data we were trying to match.

3.2 Landmark Based Shape Analysis and Procrustes

Previous studies, and indeed less statistical current studies, use multivariate morphometrics to examine shape. Traditionally applied to biological data, many distances and angles between points are measured. Ratios of these distances and angles are calculated and subjected to a standard analysis, e.g. t-tests, ANOVA and MANOVA. A limitation of these methods is that a greater workload is required to measure all the distances and angles and calculate ratios. Multivariate morphometrics usually only deals with positive variables (distances, angles and ratios), which can discard the geometry of a shape. The benefit of a landmark coordinate system is that the relative location of points can be described, and the distances and angles can still be derived from the landmark coordinates. The key objective for this study is to assess population variability and therefore estimate a population distribution; landmarks offer a convenient route to this via a multivariate normal distribution. In principal distances between points could be used however the configuration of the positions of recognisable facial features captures the idea of a face more directly.

Other previously adopted approaches include geometrical methods, which provide the ability to work with landmark coordinates directly. The idea is that work is carried out on the complete geometrical object itself (up to similarity transformation), as opposed to working with quantities derived from an object.

3.3 Who was Procrustes?

"In Greek mythology Procrustes was the nickname of a robber Damastes, who lived by the road from Eleusis to Athens. He would offer travellers a room for the night and fit them to the bed by stretching them if they were too short or chopping off their limbs if

they were too tall. We can regard one configuration as the bed and the other as the person being ‘translated’, ‘rotated’ and possibly ‘rescaled’ so as to fit as close as possible to the bed.”

Dryden and Mardia (1998)

3.4 Summary of Procrustes Analysis

Once coordinates for a common set of landmarks are recorded on the objects under investigation (§2.2, §2.4.3), Procrustes analysis can be applied to configurations of coordinates obtained from different objects. Procrustes analysis brings the configurations of landmarks to a common orientation and size and yet preserves the ‘shape’ of each individual object; this overcomes any mismatch in the scale, origin and orientation of the object or coordinate system used. Figure 3.1 describes the basic stages in Procrustes analysis using some artificial faces as an example.

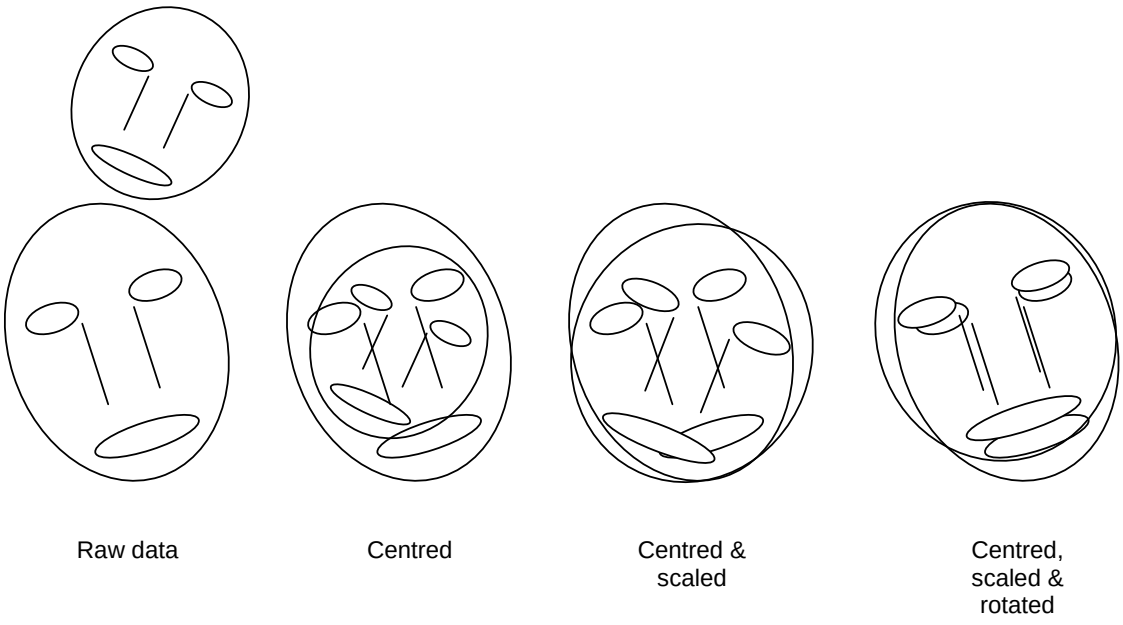


Figure 3.1 - The stages of Procrustes superimposition, translation to a common centre, scaling to the same unit centroid size and rotation to minimize the sum of squared distances between corresponding landmarks.

3.4.1 Shape

Dryden and Mardia (1998) define ‘shape’ as “the geometrical information that is invariant under translation, rotation and scaling”. They developed means for statistical shape analysis, which involve methods for the geometrical study of random objects where location, rotation and scale information can be removed.

The shape space is a non-Euclidean manifold and careful consideration must be used when looking for appropriate methods of data analysis. In particular, multivariate statistical procedures cannot be applied directly to non-Euclidean information, but in certain circumstances can be adapted for shape data (§3.7).

This project involves the statistical shape analysis of landmark data, where landmarks correspond to identifiable features on a subject (§2.2, §2.4.3). Here we are interested in comparing photographs of faces to see if there are significant differences in the shape of each face. We require a way of measuring shape, some notion of the distance between two shapes (faces) and methods for statistical analysis of shape.

Techniques developed by Dryden and Mardia (1998) can be used to estimate a mean shape, to assess whether two groups are significantly different in mean shape and to carry out discrimination or clustering on the basis of shape and size information.

3.4.2 Procrustes Methods

Procrustes analysis uses linear transformations to remove location and scale information and orthogonal matrices to remove reflection and rotation information on data. It involves matching configurations with similarity transformations to be as close as possible according to Euclidean distance, using least squares techniques. The steps of the similarity transformations are outlined in §3.4.4 to §3.4.7. After the transformations what is left of the data is the underlying shape information, which is described in §3.4.8.

Procrustes methods can be used to look at the distance between two different shapes (3.4.9) and to estimate an average shape (3.5.2 and 3.6.2). The structure of shape variability in a dataset can also be explored through principal components analysis of

the tangent shape coordinates using the Procrustes mean as the pole to project shape coordinates onto a tangent plane to the shape space (3.7).

Before going straight into the theory behind Procrustes analysis a definition of what is meant by landmark data is given in the following subsection.

3.4.3 Landmarks

Shape can be described by locating a number of points on a specimen, which are called landmarks. A landmark is a point of correspondence on an object that matches between and within populations. In this study of face shapes anatomical landmarks were used (§2.2, $2.4.3), these are points that correspond between organisms in some biologically meaningful way, for example the corner of an eye. Landmarks can also be mathematical, for example the maximum point of curvature on an arc. Pseudo landmarks can also be used to describe shape; these are constructed points, for example at regular intervals around the outline of a shape, or the mid-point between two anatomical or mathematical points.

The configuration of landmarks of an object X is typically represented by a $k \times m$ matrix of coordinates, where k is the number of landmark points in m dimensions. For example in three dimensions:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ \vdots & \vdots & \vdots \\ x_{k1} & x_{k2} & x_{k3} \end{bmatrix} \quad \ldots.(1)$$

The order of the landmark points is arbitrary, however when comparing configurations the landmarks must correspond between objects. The configuration space is the space of all possible landmark coordinates, typically $\mathbf{R}^{km}$.

Now, we have already defined shape as the remaining information after size, location and rotation has been removed from an object. The following subsections describe how each of these properties is removed from one configuration X. In order to represent shape it is convenient to remove these similarity transformations one at a time. §3.5

extends this theory to look at aligning two shape configurations and §3.6 extends further to provide a general case for aligning many configurations, like the faces database available here.

3.4.4 Removing Translation

Translation is the easiest to filter from a configuration matrix, *X*, and is done so by considering contrasts of the data by pre-multiplying by a suitable matrix. We can make a specific choice of contrast by pre-multiplying *X* with the Helmert sub-matrix equation.

The j^{th} row of the Helmert sub-matrix, a $(k\text{-}1) \times k$ matrix H, is given by

$$(h_j, \ldots, h_j, -jh_j, 0, \ldots, 0), \qquad h_j = -\{j(j+1)\}^{-1/2}$$

and so the j^{th} row consists of h_j repeated j times, followed by $-jh_j$ and then $k-j-1$ zeros, $j = 1, \ldots, k-1$.

So, we write

$$X_H = HX \in \mathbb{R}^{(k-1)m} \setminus \{0\}$$

The origin is removed because coincident landmarks are not allowed, and we refer to X_H as the Helmertized landmarks.

An alternative choice of contrast uses the centred landmarks for removing location, these are given by

$$X_C = CX$$

We can get to the centred landmarks from the Helmertized landmarks by pre-multiplying by H^T, as

$$H^T H = I_k - \frac{1}{k} 1_k 1_k^T = C$$

so

$$H^T X_H = H^T HX = CX$$

When the facial images were collected for the current study a set of rules were applied to ensure that the same procedure for image capture was carried out for every subject

(§2.4.2). One of these rules was that the subject should be able to see a reflection of their own eyes in a small mirror which was located on the central camera pair of the scanner. The seat where the subject sat was height adjustable. So, the translation error should only be small, although still exists in terms of whether the subject could see the reflection of their eyes directly in the centre of the mirror or whether they were positioned slightly to the left or right or above or below the middle point, yet were still visible.

3.4.5 Removing Size

To allow size to be removed from a configuration a definition of what is meant by the size is required. The centroid size *S(X)* is the square root of the sum of squared Euclidean distances from each landmark to the centroid. It should be noted that an alternative method of shape alignment is partial Procrustes analysis, which is Procrustes without scaling. This method was applied in §6.2.4 to show that size differences exist between male and female faces, however in terms of facial identifications it is thought that presuppositions of the sex of perpetrator should not be made. The nature of the available facial data is such that images have varying scales, as they were obtained from different sources; therefore it is necessary to remove the scale before any comparisons of shapes can be carried out. For the collection of the facial database the distance of subject to camera was measured to be fifty centimetres so scale differences should be minimal. Differences in scale which need to be removed from other sources of data can be thought of in terms of standardising the distance of the subject to camera.

The centroid size is given by

$$S(X) = \|CX\| = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{m}\left(X_{ij} - \bar{X}_j\right)^2} \qquad X \in \mathbf{R}^{km}$$

where X_{ij} is the (i, j)th entry of X, $\hat{X}_j = \frac{1}{k}\sum_{i=1}^{k} X_{ij}$ is the arithmetic mean of the jth dimension, $C = I_k - \frac{1}{k}1_k 1_k^{\mathrm{T}}$ is the centring matrix, $\|X\| = \sqrt{\mathrm{trace}(X^{\mathrm{T}}X)}$ is the Euclidean norm, I_k is the $k \times k$ identity matrix, and 1_k is the $k \times 1$ vector of ones.

In computing a distance between shapes it is necessary to standardise for size, this is done by dividing through by the centroid size, which is also given as:

$$\begin{aligned} \|X_H\| &= \sqrt{\text{trace}\left(X^{\mathrm{T}} H^{\mathrm{T}} HX\right)} \\ &= \sqrt{\text{trace}\left(X^{\mathrm{T}} CX\right)} \\ &= \|CX\| \\ &= S(X) \end{aligned} \qquad \ldots(2)$$

since $H^{\mathrm{T}}H$ = C is idempotent.

3.4.6 The Pre-shape and Pre-shape Space

After the location and scale information have been filtered out from the configuration matrix, what remains is known as the pre-shape. The pre-shape is one step away from shape, since rotation still has to be removed. The term pre-shape was coined by Kendall (1984).

The pre-shape of a configuration matrix X is given by

$$Z = \frac{X_H}{\|X_H\|} = \frac{HX}{\|HX\|}$$

and this is invariant under the translation and scaling of the original configuration.

The pre-shape space is the space of all possible pre-shapes. Formally the pre-shape space S_m^k is the orbit space of the non-coincident k point set configurations in $\mathbf{R}^m$ under the action of translation and isotropic scaling. The pre-shape space $S_m^k \equiv S^{(k-1)m-1}$ is a hypersphere† of unit radius $(k-1)m$ real dimensions, since $\|Z\| = 1$.

† A hypersphere is a shape in four dimensions that is analogous to a sphere.
Similarly, a sphere is a shape in three dimensions that is analogous to a circle.

3.4.7 Removing Rotation

In order to remove the rotation information from a configuration, all rotated versions of the pre-shape must be identified. This set or equivalence class is denoted as the shape of *X*. An alternative definition of the shape of *X* is as follows.

3.4.8 Removing Reflection

The reflection information can also be removed from data, however the facial data does not require this as no information on reflection was collected.

3.4.9 Shape and Shape Space

The shape of a configuration matrix *X* is all the geometrical information about *X* that is invariant under location, rotation and isotropic scaling (Euclidean similarity transformations). Shape can be represented by the set [*X*] given by

$$[X] = \{Z\Gamma : \Gamma \in SO(m)\}$$

where Γ is the rotation matrix, $SO(m)$ is the special orthogonal group of rotations and Z is the pre-shape of X.

The shape space is the set of all possible shapes. Formally, the shape space $\sum_m^k$ is the orbit space of the non-coincident k point set configurations in $\mathbf{R}^m$ under the action of Euclidean similarity transformations.

The dimension of the shape space is

$$M = km - m - 1 - \frac{m(m-1)}{2}$$

since we initially have km co-ordinates, from which one dimension is removed for uniform scale, m dimensions are removed for location, and $\frac{1}{2}m(m-1)$ dimensions are removed for rotation.

In the standard formulation it is presumed that all of the landmarks are required to define the shape of an object, e.g. the vertices of a polygon. In the application here it will be seen (§6.2.6, §6.4.1, §7.2.3, §7.3) that it may be preferable to select only certain features from the full set, so that they are more useful in various senses that will become apparent later.

3.4.10 Full Procrustes Distance

Now that we have defined shape and shape space we can now think about the idea of the distance between two shapes. A concept of distance is required to fully define the non-Euclidean shape metric space. Here the full Procrustes distance will be used as the distance between two shapes.

Consider two configuration matrices X_1 and X_2 for k points in m dimensions, with pre-shapes Z_1 and Z_2. We minimise over rotations and scale to find the closest Euclidean distance between Z_1 and Z_2. The full Procrustes distance between X_1 and X_2 is therefore:

$$d_F(X_1, X_2) = \inf_{\Gamma \in SO(m), \beta \in R} \left\| Z_2 - \beta Z_1 \Gamma \right\|$$

where $Z_r = HX_r / \left\| HX_r \right\|$, $r = 1,2$, Γ is the minimizing rotation matrix and β is the minimizing scale. Inf refers to the infimum, or informally the greatest lower bound.

The full Procrustes distance between X_1 and X_2 can be written as,

$$d_F(X_1, X_2) = \left\{ 1 - \left(\sum_{i=1}^{m} \lambda_i \right)^2 \right\}^{1/2} \qquad \ldots(3)$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{m-1} \geq \left| \lambda_m \right|$ are the square roots of the eigenvalues of $Z_1^{\mathrm{T}} Z_2 Z_2^{\mathrm{T}} Z_1$, and the smallest value of λ_m is the negative square root if and only if $\det\left(Z_1^{\mathrm{T}} Z_2 \right) < 0$.

The minimising rotation matrix Γ and scaling factor β can be easily obtained from singular value decomposition (SVD) of $Z_2^{\mathrm{T}}Z_1$.

$$Z_2^{\mathrm{T}} Z_1 = V \Lambda U^{\mathrm{T}}$$

where $\Lambda=\mathrm{diag}(\lambda_1, \lambda_2, \ldots \lambda_m)$. Then the minimising rotation and scale are given by,

$$\hat{\Gamma} = UV^{\mathrm{T}} \text{ and } \hat{\beta} = \sum_{i=1}^{m} \lambda_i$$

3.5 Ordinary Procrustes Analysis

Consider the case where two configuration matrices X_1 and X_2 are available (with both $k \times m$ matrices of coordinates from k points in m dimensions), and we want to get the two configurations to match as close as possible, up to similarity transformations (assuming without loss of generality that the configuration matrices X_1 and X_2 have already been centred by Equation 2).

The method of full ordinary Procrustes analysis involves the least squares matching of two configurations using the similarity transformations. Minimising the squared Euclidean distance carries out estimation of the similarity parameters γ, Γ and β:

$$D^2_{OPA}(X_1, X_2) = \left\| X_2 - \beta X_2 \Gamma - 1_k \gamma^{\mathrm{T}} \right\|^2 \qquad \ldots(4)$$

where $\|X\| = \{\mathrm{trace}(X^{\mathrm{T}} X)\}^{1/2}$ is the Euclidean norm, Γ is an $(m \times m)$ rotation matrix ($\Gamma \in SO(m)$), $\beta > 0$ is a scale parameter, and γ is an $(m \times 1)$ location vector. The minimum of this equation is written as $OSS(X_1, X_2)$, which stands for Ordinary Sum of Squares.

We can then solve Equation 4 to find the minimum, which is given by $\left(\hat{\gamma}, \hat{\beta}, \hat{\Gamma}\right)$ where

$$\hat{\gamma} = 0$$
$$\hat{\Gamma} = UV^{\mathrm{T}}$$
$$\hat{\beta} = \frac{\operatorname{trace}\left(X_2^{\mathrm{T}} X_1 \hat{\Gamma}\right)}{\operatorname{trace}\left(X_1^{\mathrm{T}} X_1\right)}$$

where $X_2^{\mathrm{T}} X_1 = \|X_1\| \|X_2\| V \Lambda U^{\mathrm{T}}$, $U, V \in SO(m)$.

Therefore this gives

$$OSS\left(X_1, X_2\right) = \|X_2\|^2 \sin^2 \rho\left(X_1, X_2\right)$$

where $\rho(X_1, X_2)$ is the Procrustes distance.

Using this we can then calculate the full Procrustes fit.

3.5.1 Full Procrustes Fit

The full Procrustes fit (or full Procrustes coordinates) of X_1 onto is X_2 given by

$$X_1^P = \hat{\beta} X_1 \hat{\Gamma}_1 + 1_k \hat{\gamma}^{\mathrm{T}}$$

where $\hat{\Gamma}_1$ is the rotation matrix, $\hat{\beta} > 0$ is the scale parameter and $\hat{\gamma}^{\mathrm{T}}$ is the location parameter. The superscript '*P*' denotes the Procrustes superimposition and this then allows us to calculate the full Procrustes mean as follows.

3.6 Generalized Procrustes Analysis

The generalization of ordinary Procrustes analysis to cope with problems where $n \geq 2$ configuration matrices are available is simply termed generalized Procrustes analysis (GPA). GPA is required to align the large database of facial landmarks in the current study.

Full GPA involves rotation, rescaling and rotating all configurations relative to each other in order to minimise a total sum of squares. A quantity proportional to the sum of squared norms of pair wise differences is minimized, and is called the generalized (Procrustes) sum of squares:

$$G(X_1,...,X_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left\|\left(\beta_i X_i \Gamma_i + 1_k \gamma_i^{\mathrm{T}}\right) - \left(\beta_j X_j \Gamma_j + 1_k \gamma_j^{\mathrm{T}}\right)\right\|^2$$

subject to a constraint on the size of the average, $S(\bar{X}) = 1$, where *S(X)* is the centroid size, $\Gamma_i \in SO(m), \beta_i > 0, \|X\| = \sqrt{\mathrm{trace}(X^{\mathrm{T}}\mathrm{X})}$ and the average configuration is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\left(\beta_i X_i \Gamma_i + 1_k \gamma_i^{\mathrm{T}}\right)$$

Full GPA matching involves the superimposition of all the configurations in optimal positions by translating, rotating and rescaling each configuration to minimize the sum of Euclidean distances.

3.6.1 Full Procrustes Fit

The full Procrustes coordinates (or fit) of each of the configurations X_i *i*s given by,

$$X_i^P = \hat{\beta}_i X_i \hat{\Gamma}_i + 1_k \hat{\gamma}_k^{\mathrm{T}}, i = 1,...,n$$

where rotation matrix $\hat{\Gamma}_i \in SO(m)$, scale parameter $\hat{\beta}_i > 0$ and location parameters $\hat{\gamma}_k^{\mathrm{T}}$ are the minimising parameters.

3.6.2 Full Procrustes Mean

The full Procrustes mean (full Procrustes estimate of the mean shape) is given by $[\hat{\mu}]$, where

$$\hat{\mu} = \arg \inf_{\mu: S(\mu)=1} \sum_{i=1}^{n} \sin^2 \rho(X_i, \mu)$$

$$= \arg \inf_{\mu: S(\mu)=1} \sum_{i=1}^{n} d_F^2(X_i, \mu)$$

If we note that

$$G(X_1, \ldots, X_n) = \inf_{\mu: S(\mu)=1} \sum_{i=1}^{n} \sin^2 \rho(X_i, \mu),$$

the point in shape space corresponding to the arithmetic mean of the Procrustes fits:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i^P$$

has the same shape as the full Procrustes mean. Therefore, once a group of objects have been matched into full Procrustes position, the full Procrustes mean shape can be calculated by taking the arithmetic means of each coordinate. This is equivalent to minimising the sums of squared distances in the shape space d_F^2 (defined in Equation 3).

For data in two dimensions an explicit eigenvector solution is available to obtain the full Procrustes mean. For $m = 3$ dimensions and higher and the full Procrustes mean shape has to be found iteratively using the following algorithm.

1. Choose an initial estimate of the mean shape (e.g. the first shape in the dataset)
2. Align all the remaining shapes to the initial estimate of mean shape
3. Re-calculate a new estimate of mean shape from the aligned shapes
4. If the estimated mean has changed return to step 2 until both steps converge

The above procedure is also used to obtain the rotation, scale and location parameters.

3.7 Principal Components Analysis in the Tangent Space

Derived from the shape space, there exists an additional coordinate system of tangent space, which is very useful in shape analysis. The tangent space allows linear multivariate statistical methods to be applied to shape data. Principal components analysis (PCA) of the tangent shape coordinates using the Procrustes mean as the pole provides a suitable method of exploring the structure of shape variability in a dataset. An advantage of the application of PCA to shape data is that the results can be visualised as shapes or shape changes. As with ordinary multivariate analysis PCA transforms the shape data to reduce the dimensionality of the problem and examine the main patterns of variability. With shape analysis PCA searches for the orthogonal axes of shape variation that summarize large percentages of shape variability.

3.7.1 Tangent Space

The tangent space is a linearised version of the shape space close to a particular point in shape space, called the pole of the tangent projection. The pole is usually chosen to be a mean shape obtained from the dataset of interest, so that the choice of co-ordinates depends on that dataset. Here a tangent projection to the pre-shape sphere is considered, which does not depend on the original rotation of the figure, and is therefore a suitable tangent co-ordinate system for the shape.

The Euclidean distance in the tangent space is a good approximation to the Procrustes distances in shape space for points close to the pole. Therefore, if the majority of the objects in a dataset are quite close in shape, then using the Euclidean distance in the tangent space will be a good approximation to the shape distances in the shape space. Hence standard multivariate statistical methods in tangent space will be good approximations to non-Euclidean shape methods, provided the data are not too highly dispersed. With the facial landmark data the relative locations of the landmarks are the same for all configurations, e.g. the eyes are always above the nose, so shapes are expected to show very little dispersion.

Let us consider a set of complex landmark points, $z^o = \left(z_1^o, \ldots, z_k^o\right)^{\mathrm{T}}$, with pre-shape,

$$\begin{aligned} z &= \left(z_1, \ldots, z_{k-1}\right)^{\mathrm{T}} \\ &= Hz^o / \left\| Hz^o \right\| \end{aligned}$$

Take γ to be a complex pole on the complex pre-shape sphere usually chosen as a mean shape. Then, rotate the configuration by an angle θ so as to be as close as possible to the pole. Then project onto the tangent place at γ, denoted by $T(\gamma)$. Now the partial Procrustes tangent coordinates can be defined as follows.

3.7.2 Partial Procrustes Tangent Coordinates

The partial Procrustes tangent co-ordinates for a planer shape are given by

$$v = e^{i\hat{\theta}}\left[I_{k-1} - \gamma\gamma^*\right]z, \; v \in T(\gamma)$$

where γ^* indicates the transpose of the complex conjugate of γ and $\hat{\theta} = \arg\left(-\gamma^* z\right)$. The partial Procrustes tangent co-ordinates involve only rotation (and not scaling) to match the pre-shapes.

Note that $v^*\gamma = 0$, so the complex constraint means that we can then regard the tangent space as a real subspace of $\mathbb{R}^{2k-2}$ of dimension $2k - 4$. We can therefore use partial Procrustes tangent co-ordinates in real space, and hence utilise these in the analysis of shapes in real space, with tools such as principal components analysis.

3.8 Evidence Evaluation of Facial Matches using Likelihood Ratios

§3.2 to §3.7 have described the means for extracting facial shapes from the facial landmark data that is available for this project. The following section will now discuss techniques for comparing these facial shapes in order to determine whether two images could be declared a 'match' or not. Essentially when working in tangent space (§3.7.1)

the facial shape data can be considered as a multivariate dataset, of which we want to measure the similarity or distance between observations.

The likelihood ratio (LR) is described as a statistical method used to directly evaluate the strength of evidential observations. The LR is a rational, intuitive method for placing a simple value on evidence that was first suggested by Poincare, Darboux and Appell in the late 19th century, Aitken and Taroni (2004). Lucy (2005) states that the LR is currently the predominant measure for numerically based forensic evidence. He gives an example of the use of the LR saying "the bloodstain on the carpet may 'match' the suspect in some biochemical way, but was the blood which made the stain derived from the suspect, or one of the other possible individuals who could be described as a match". The LR can be similarly applied to facial identification, stating whether photographs of suspect and perpetrator are a 'match' and also giving an estimated measure of the strength of this 'match' i.e. how certain we are that two facial images depict the same person.

As there is no inherent knowledge of the system from which to deduce probabilities for outcomes, without examining every member of the population the estimates of probability will always be subject to a quantifiable uncertainty. The available large multivariate sample of known face shapes (§2.4), based on corresponding anthropometrical landmark points on the face (§2.2, §2.4.3), can be used to estimate probabilities. Loosely we can think of a LR for matching faces as a comparison of the probability that the (recovered) face of a suspect is the same as the (control) face of a person committing a crime (captured on CCTV for example) and the probability that the face of the suspect lies somewhere else in the known population (§3.8.1).

Using a statistical model for the numerical facial data, estimates of model parameters can be used to estimate the likelihood that a piece of recovered evidence came from the same source as the crime scene (or control), §3.8.4.1. In a similar way the probability that the same piece of evidence came from some other source in the known population of evidence can also be estimated. These two probabilities are then compared by a LR, evaluating the ratio of the first probability to the second. A LR greater than one indicates more evidence to support the hypothesis for the prosecution (H_p) that the control and recovered data both come from the same source, i.e. the crime scene. A LR

less than one indicates evidence is more in favour of the defence hypothesis (H_d) that the control data comes from some other source, i.e. not the recovered.

This section extends theory developed by Aitken and Lucy (2004) for matching multivariate data on fragments of glass, which consisted of just three variables to evaluate. The following formulae outline how to deal with the much larger dataset of faces, which has up to ninety variables, i.e. thirty landmark points in three dimensions.

3.8.1 Control and Recovered Data

From Aitken and Lucy (2004) a number, n_1 (≥ 1), of replicate measurements are taken from a crime scene, these measurements are referred to as control data, as the source, P_1, of the measurements is known. A number, n_2 (≥ 1 and not necessarily equal to n_1), of replicate measurements are also taken from a suspect, these measurements are referred to as recovered data assumed to have come from a source P_2. The prosecution proposition, H_p, is that the sources P_1 and P_2 are the same. The defence proposition, H_d, is that they are not. One of the prerequisites of the LR test described by Aitken and Lucy (2004) is that there must be replications in the data.

Aitken and Lucy (2004) use replicate measurements on glass fragments taken from a crime scene and glass fragments found on a suspects clothing, the test is to determine whether these two sets of fragments could have come from the same glass window. In the present context the replicate measurements n_1 were taken of facial landmark points from one facial image assumed to come from a person P_1, e.g. which was captured of a perpetrator at a crime scene. The replicate measurements n_2 were taken of facial landmark points from a second facial image assumed to have come from a person P_2, e.g. which was captured of a suspect in custody. The prosecution proposition, H_p, is that P_1 and P_2 is the same person. The defence proposition, H_d, is assumed to be for the purposes of this thesis that they are not.

3.8.2 Background Database

The background population available for use with LR calculations largely consists of the Geometrix® facial database (§2.4). These data are in the form of sets of coordinates

of landmark points taken from faces collected from the Geometrix FaceVision802 3D scanner and Forensic Analyzer software, as described in §2.4.2 and §2.4.3. The main database contains landmark coordinate data on thirty different 3D points of the face (§5.2.6, Table 5.3). There are different numbers of faces with complete data for different subsets of landmark points. The more landmark points that are used in analyses the fewer faces that encompass all the data. Some landmark points are likely to be more useful than others in terms of providing good variables for matching faces (§6.2.6, §6.4.1, §7.2.3, §7.3, §7.5, §7.6), however to begin with the full set of thirty landmark points is utilised.

For all thirty landmarks there are $n = 1286$ different faces, where for each face there are $r = 2$ replicate measurements of the landmarks made by either the same or different observers. The total number of observations ($N = nr$) in the background database for thirty landmark points is $N = 2572$ configurations of landmark coordinates. Also, the configurations for the faces under comparison have to be included in the background database to ensure that they fit into the model for the data.

3.8.3 Choosing a Method for the Evaluation of Evidence

Aitken and Lucy (2004) describe five different methods for the assessment of evidence for multivariate data. Two of these are based on significance tests and the other three evaluate likelihood ratios. Although significance tests could assess whether there was a significant difference between the mean of the control face and the mean of the recovered face, what could not be assessed is the degree of difference between the two faces in question. Also the significance test methods assume that the within-source variability is constant, with the facial study different observers collected the data. It was shown (chapter 5, §5.3) that although the within-source variability was small enough to validate the data collection method for multiple observers, it was not in fact constant. For these reasons the significance test methods were thought inappropriate for evaluating facial shape evidence.

Of the three likelihood ratio methods, Aitken and Lucy (2004), one describes a likelihood ratio where the numerator is the density of the Hotelling's T-squared statistic and the denominator a kernel density estimate of the distribution over a transformation

to one dimension of the data from the background population database, Curran et al (1997). With the facial data there are many more dimensions than with the glass fragment data used in Aitken and Lucy (2004), i.e. ninety variables as opposed to just three. It was thought that to transform onto only one of the ninety dimensions of the shape space of facial landmarks would be insufficient to accurately model the data. In fact in chapters 6 and 7 it is seen that the information in the different dimensions of the background data varied considerably (§6.2.6, §6.4.1, §7.2.3). Hence it was thought that rather than a univariate projection approach a multivariate model would be preferable in this situation.

The remaining two likelihood ratio methods are multivariate approaches modelling for two levels of variation (within-source and between-source). One method assumes normality in the between-source variability, the other models between-source variability with a multivariate kernel density estimate. Chapter 6 shows how a multivariate normal distribution adequately models the background facial data (§6.5), so the former of these two methods seems the most reasonable to apply to the face data and is outlined in the subsequent section.

3.8.4 Method: Likelihood Ratio using a Multivariate Random Effects Model and Assumptions of Normality

The following method, based on Aitken and Lucy (2004), is referred to as the multivariate normal likelihood ratio (MVNLR) procedure. The method is considered for the valuation of control and recovered facial landmark data (§3.8.1) to compare propositions that the two sets of data have come from the same (H_p) or from different sources (H_d).

3.8.4.1 Model

Let Ω denote the population of p variables to be used for facial matching, e.g. if coordinates of facial landmarks are to be used then p = landmarks (k) * dimensions (m). So, thirty landmark points in a 3D analysis would have p = 90 variables. A 2D analysis, which is more likely to occur in facial matching using real life data from CCTV for example, would have p = 60 variables. The background data (§3.8.2) are measurements

of these p variables on a random sample of n faces from Ω with r (≥ 2) replicate measurements on each of the n faces. Denote the background data as

$$\mathrm{x}_{ij} = (x_{ij1},\ldots, x_{ijp})^{\mathrm{T}} ;\ i = 1,\ldots, n;\ \ j = 1,\ldots, r,$$

The average of the multiple measures for each face n can be written as

$$\bar{\mathrm{x}}_i = \frac{1}{n}\sum_{j=1}^{n} \mathrm{x}_{ij}$$

The measurements for the control and recovered faces to be compared are denoted

$$\{\mathrm{y}_1\} = (\mathrm{y}_{lj},\ j = 1,\ldots, n_l;\ l = 1, 2), \text{ where } \mathrm{y}_{lj} = (y_{lj1},\ldots, y_{ljp})^{\mathrm{T}}.$$

$$\text{Let } \bar{\mathrm{y}}_l = \frac{1}{n_l}\sum_{j=1}^{n_l} \mathrm{y}_{lj}$$

Denote the individual variable means over n_l measurements as

$$\bar{\mathrm{y}}_{l.k} \text{ for } k = (1,\ldots,k)$$

The model assumes two sources of variation, that between replicated measures taken within the same face (within-source variation) and that between different faces (between-source variation). It is assumed that both the variation within-source and between-source is constant and normally distributed. As shown in §6.5 the multivariate normal distribution provides a good fit for the facial landmark data. If this were to be a test then twice the log likelihood could be compared to a Chi-squared distribution with (km-1) degrees of freedom. However, the danger with doing this is that differences we are not interested in could be picked out, or a failed test could mean there was not enough data. Additionally, because the responsibility for choosing the size of the test would rest with the court, it seems more useful to proceed by looking at just the scale of the LRs.

Within-source: Denote the mean vector within source *i* by θ_i and the matrix of within-source variances and covariance by *U*. Then, given θ_i and *U*, the distribution of X_{ij} is taken to be normal:

$$(X_{ij} | \theta_i, U) \sim N(\theta_i, U),\ i = 1, \ldots, n;\ j = 1, \ldots, r.$$

Between-source: Denote the mean vector between sources by μ and the matrix of between-source variances and covariance by *C*. The distribution of the θ_i, as measures of between-source variability, is taken to be normal:

$$(\theta_i | \mu, C) \sim N(\mu, C),\ i = 1, \ldots, n.$$

The distributions of the measurements y_1, y_2 on the control and recovered data, conditional on the source (crime or suspect), are also taken to be normal. The means, $\bar{Y}_l$, have normal distributions with mean θ_l and variance-covariance matrix D_l where

$$D_1 = n_1^{-1}U \text{ and } D_2 = n_2^{-1}U:$$
$$(\bar{Y}_l | \theta_l, D_l) \sim N(\theta_l, D_l)\ ;\ l = 1, 2.$$

Then for the assumption of between-source normality,

$$(\bar{Y}_l | \mu, C, D_l) \sim N(\mu, C + D_l)\ ;\ l = 1, 2.$$

3.8.4.2 Estimating the Model Parameters

The mean μ is estimated by $\bar{x}$, the mean vector over all groups.
The within-group covariance matrix *U* is estimated from the background data $\{x_{ij}\}$ by

$$\hat{U} = \frac{S_w}{(N - r)} \quad (1)$$

$$\text{Where } S_w = \sum_{i=1}^{r} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T$$

The between-group covariance matrix C is estimated from the background data $\{x_{ij}\}$ by

$$\hat{C} = \frac{S^*}{(r-1)} - \frac{s_w}{n(N-r)} \quad (2)$$

$$\text{Where } S^* = \sum_{i=1}^{r} \left(\bar{x}_i - \bar{\bar{x}}\right)\left(\bar{x}_i - \bar{\bar{x}}\right)^T$$

The value of the evidence y_1 and y_2 is the ratio of two probability density functions of the form $f(y_1, y_2|\mu, C, U)$, one for the numerator of the LR, where H_p is assumed true, and one for the denominator, where H_d is assumed true. In the numerator the source means θ_1 and θ_2 and assumed equal (to θ, say) but unknown. In the denominator it is assumed that θ_1 and θ_2 are not equal.

Numerator: Denote the probability density function by $f_0(y1, y2 \mid \mu, U, C)$. It is given by

$$\int_\theta f(y_1 \mid \theta, U) f(y_2 \mid \theta, U) f(\theta \mid \mu, C) d\theta \qquad (3)$$

where the three probability density functions are multivariate normal.

Denominator: Denote the probability density function by $f_1(y1, y2 \mid \mu, U, C)$, which is given by

$$\int_\theta \{f(y_1 \mid \theta, U) \times f(\theta \mid \mu, C)\} d\theta \times \int_\theta \{f(y_2 \mid \theta, U) \times f(\theta \mid \mu, C)\} d\theta \qquad (4)$$

where y_1 and y_2 are taken to be independent as the data are assumed to be from different sources.

The value of evidence is the ratio of (3) to (4), Aitken and Lucy (2004) show that this is equal to the ratio of

$$\left|2\pi\left[\left(n_1+n_2\right)U^{-1}+C^{-1}\right]^{-1}\right|^{\frac{1}{2}} exp\left\{-\frac{1}{2}\left(H_2+H_3\right)\right\} \quad (5)$$

to

$$\left|2\pi C\right|^{-\frac{1}{2}}\left|2\pi\left(n_1U^{-1}+C^{-1}\right)^{-1}\right|^{\frac{1}{2}}\left|2\pi\left(n_2U^{-1}+C^{-1}\right)^{-1}\right|^{\frac{1}{2}} exp\left\{-\frac{1}{2}\left(H_4+H_5\right)\right\} \quad (6)$$

Where

$$H_2=\left(\mathrm{y}^*-\mu\right)^T\left(\frac{U}{n_1+n_2}+C\right)^{-1}\left(\mathrm{y}^*-\mu\right)$$

$$\mathrm{y}^*=\frac{\left(n_1\bar{\mathrm{y}}_1+n_2\bar{\mathrm{y}}_2\right)}{\left(n_1+n_2\right)}$$

$$H_3=\left(\bar{\mathrm{y}}_1-\bar{\mathrm{y}}_2\right)^T\left(D_1+D_2\right)^{-1}\left(\bar{\mathrm{y}}_1-\bar{\mathrm{y}}_2\right)$$

$$H_4=\left(\mu-\mu^*\right)^T\left[\left(D_1+C\right)^{-1}+\left(D_2+C\right)^{-1}\right]\left(\mu-\mu^*\right)$$

$$H_5=\left(\bar{\mathrm{y}}_1-\bar{\mathrm{y}}_2\right)^T\left(D_1+D_2+2C\right)^{-1}\left(\bar{\mathrm{y}}_1-\bar{\mathrm{y}}_2\right)$$

$$\mu^*=\left\{\left(D_1+C\right)^{-1}+\left(D_2+C\right)^{-1}\right\}^{-1}\left[\left(D_1+C\right)^{-1}\bar{\mathrm{y}}_1+\left(D_2+C\right)^{-1}\bar{\mathrm{y}}_2\right]$$

3.9 Summary

The chapter has summarised the theory behind the Procrustes methods (§3.4 - §3.7), which will be applied in the subsequent chapters on preliminary and main analyses in facial comparison. Definitions for shape (§3.4.1), space (§3.4.9), distance (§3.4.10) and useful co-ordinate systems (§3.7.1) have been adequately defined. The transformations for filtering out the location (§3.4.4), scale (§3.4.5) and rotation (§3.4.7) information from one configuration have been described, along with how this theory is extended to

deal with two and many configurations (§3.5, §3.6). The basic theory behind using LRs for evaluating multivariate evidence has been explained and the technique seems to be appropriate to use for facial matching (§3.8).

Generalized Procrustes Analysis is a useful tool for the analysis of two and three-dimensional shape data. We have shown how estimates of mean shapes can be obtained (§3.6.2), and how the structure of shape variability in a data set can be explored by using multivariate methods such as principal components analysis to examine the tangent space coordinates (§3.7).

We have also summarised techniques for evaluating multivariate evidence such as the tangent space coordinates of the facial data (§3.8). Five methods were reviewed (§3.8.3) and one was chosen as the most appropriate for facial matching. This method evaluates likelihood ratios using a multivariate normal model to obtain the likelihood that landmark configurations from two faces are more similar to each other than they are to all other faces in the known population sample (§3.8.4). Here the known population sample is the main Geometrix® database (§2.4).

The Procrustes methods described are used throughout chapters 4, 5 and 6 to examine the facial variation in the available population sample (§2.4, §6) and carry out various preliminary data checks (§4, §5). Chapter 7 applies the likelihood ratio method (§3.8.4) to compare faces in various datasets containing known facial matches (§2.7). The method models the tangent coordinates of the main facial data (§2.4) to obtain likelihoods of facial matches or exclusions (§7.2). Certain extensions to the method had to be carried out to handle the large complex dataset; the data were transformed onto principal components (PCs) to overcome the high correlation in the data. Subsets of different landmark points (§7.3.1) and then subsets of the PCs were investigated to find a set of variables that optimised the results of facial matching for some known matches (§7.3, §7.4). Chapter 8 evaluates the method and ‘best’ found subset of variables for a range of datasets containing known facial matches, factors which affected the results were explored and suggestions for improving the method were made.

4 Statistical Shape Analysis for Facial Identification: A Pilot Study

4.1 Introduction

This chapter summarises some preliminary work that was carried out on two-dimensional (2D) facial landmark data from a set of sixty faces. This is an extended version of a study undertaken in 2002 with new data and a more extensive analysis. Data were collected from images used in a pilot study that was carried out to test the viability of the techniques in Shape analysis for facial matching, Morecroft (2002). The data, described fully in §2.3, were taken from the FBI Facial Identification Catalogue, anonymous (1988) and consisted of partially covered anterior facial images. This chapter illustrates the main findings of the pilot study and extends the analysis to investigate the feasibility of using the likelihood ratio method for matching facial shape data, §3.8.4.

Current techniques employed in facial image comparison include manual evaluation of the individual characteristics of a subject (e.g. moles, scars and dimples), the form, size, symmetry and shape of the facial features and anthropometric measurement. The flaw in these methods is that they are based subjective expert judgement and not precise empirical measurements. There are many expert witnesses who claim to be able to quantify facial matches however they do not disclose their methods, their techniques are most probably are based on opinion. They also only compare the two faces in question and do not address the separate issue of when two faces do appear to be similar how many other faces in the population could also be classed as similar in their opinion. In order to assess the quality of a facial match in this way an examination of the population variation is required. If population variation was recorded then future facial comparison cases could rely on other published studies of facial variation to permit the likelihood of a match, or exclusion, to be empirically established.

The main aim of the pilot study, Morecroft (2002), was to provide proof of concept for a method of confirming or excluding an identity, which is based on precise empirical measurements of facial attributes. In order to achieve this it was first necessary to establish whether there was sufficient population variation between facial shapes to permit comparison leading to the exclusion of similar, but unmatched faces.

There are various related objectives in facial recognition and analysis which are distinct from facial identification. For example Cootes et al (2001) apply an algorithm to an input image to find the 'best' match of a face in the image, based on a model obtained from a 'training set' of similar data. Other approaches to facial analysis include 'building' facial matches from facial composites, for example for improved suspect identification from witness accounts, Solomon et al (2005), Hancock (2000). By contrast our objective of facial identification is to provide a measure of how certain we are that two different facial images depict the same person based on the facial variation in a sample of other measured faces. We do this by locating a number of predefined anthropometrical landmark points (§2.5, §3.2) on the facial images; ideally this is carried out automatically and computer-aided (§2.6, §4.2.1). Shape analysis (§3.4) is then used to extract the shape data from the collected points and shapes are compared by means of a likelihood ratio (§3.8).

4.2 Methods

The facial landmarks from sixty images from the pilot study, described in §2.3, were transformed using generalized Procrustes analysis (§3.6) to extract the facial shape information. The resulting Procrustes registered data were analysed, firstly to assess whether the variation in the face shape data was enough to be able to distinguish different faces, §4.3.1. To carry out this assessment it was first necessary to account for the variation attributed to taking different scans of images, §4.3.1.1, or capturing facial landmark points on different occasions, §4.3.1.2. To explore the structure of shape variability the partial Procrustes tangent shape coordinates were transformed onto principal components (as described in §3.7.2).

Secondly the shape data was analysed to determine whether the likelihood ratio matching technique, described in §3.8.4, was appropriate for use with matching facial shape data, §4.3.2. A cluster analysis was carried out on the facial shape information to visually look for any groups in the data, §4.3.2.1. The more formal method of evaluating likelihood ratios (§3.8.4) was then applied to search for possible facial matches in the dataset, §4.3.2.2. The results for both the assessment of facial variation and the facial shape matching are summarized in the subsequent section.

4.3 Results

4.3.1 Sources of Facial Variation

4.3.1.1 Variation Attributed to Landmark Placement

The intra-measurement (or within-face) variation between different sets of measurements taken from the same image was visually examined in a small subset of six facial images. Each image was measured three times by the same observer and the landmarks were captured. The data were Procrustes aligned and the variation was assessed by plotting the first principal component of the partial Procrustes tangent shape coordinates (§3.7.2). In terms of the covariance matrix, *V*, of Procrustes fits (§3.6.1) of the configurations to the Procrustes mean (§3.6.2) the following components of variation exist:

$$V = \frac{1}{nq-1}\sum_{i=1}^{n}\sum_{m=1}^{q}(vec(X_{im}^{P}) - vec(\overline{X}^{P}))(vec(X_{im}^{P}) - vec(\overline{X}^{P}))^{T}$$
$$= \frac{1}{nq-1} SSP(vec(X_{im}^{P})),$$

where X_{im}^{P} denotes the Procrustes fit of measurement *m*, $m = 1, \ldots, q$, from face *i*, $i = 1, \ldots, n$. $\overline{X}^{P} = \frac{1}{nq}\sum_{i=1}^{n}\sum_{m=1}^{q} X_{im}^{P}$ and $SSP(vec(X_{im}^{P}))$ denotes the sum of squares and products matrix of vectorised Procrustes fits.

The upper image in Figure 4.1 shows an example of the within-face variation observed, displayed are three triplicate measures of six landmarks from one image of one particular face. A similar examination of between-face variation is displayed in the lower image in Figure 4.1. The between-face variation shown in the means of the triplicate measures of landmark coordinates for four different faces. Clearly comparing the two plots in Figure 4.1 the variation between different faces is greater than that for triplicate measures of the same face.

PC score plots were also examined; Figure 4.2 displays data for six different faces represented by different symbols. It can be seen that the three triplicate measures taken for each face are clustered together, there is some overlap in these clusters however most are distinct.

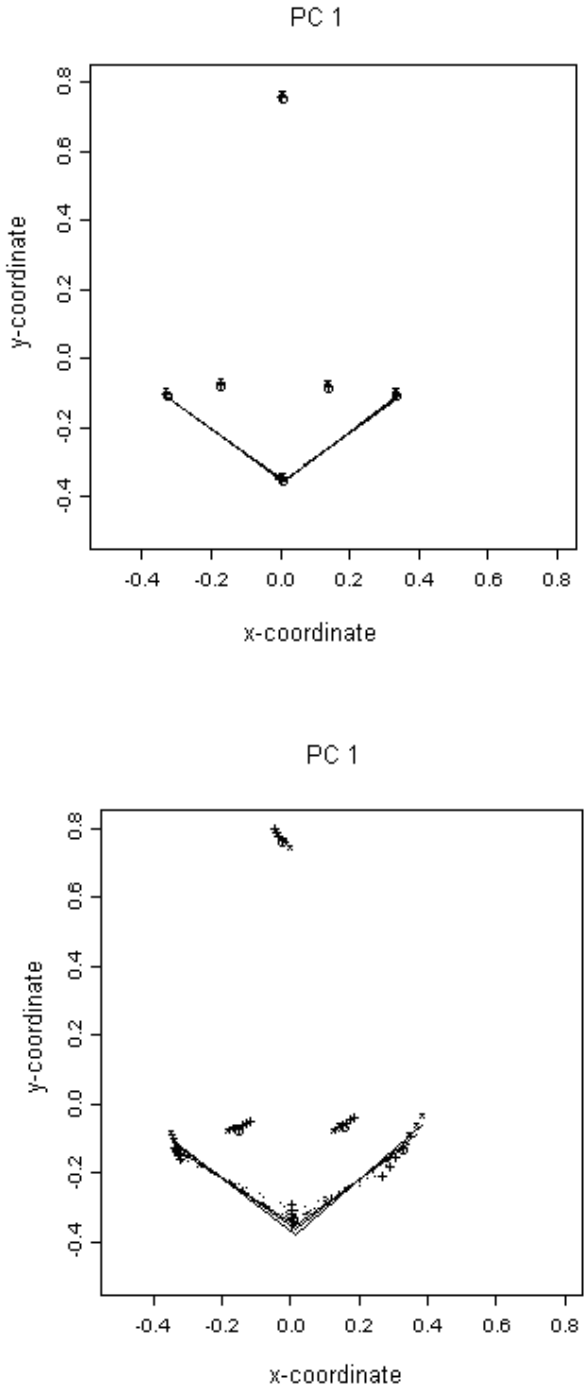


Figure 4.1 - Plots along the first principal component for the Procrustes rotated coordinates of: a) three sets of measurements taken from the same subject (top); b) the mean measurements taken from four different subjects (bottom). The plots are evaluated at *c* = -3, -2, -1 (*) standard deviations and *c* =+ 1, + 2, + 3 (+) standard deviations along the principal component.

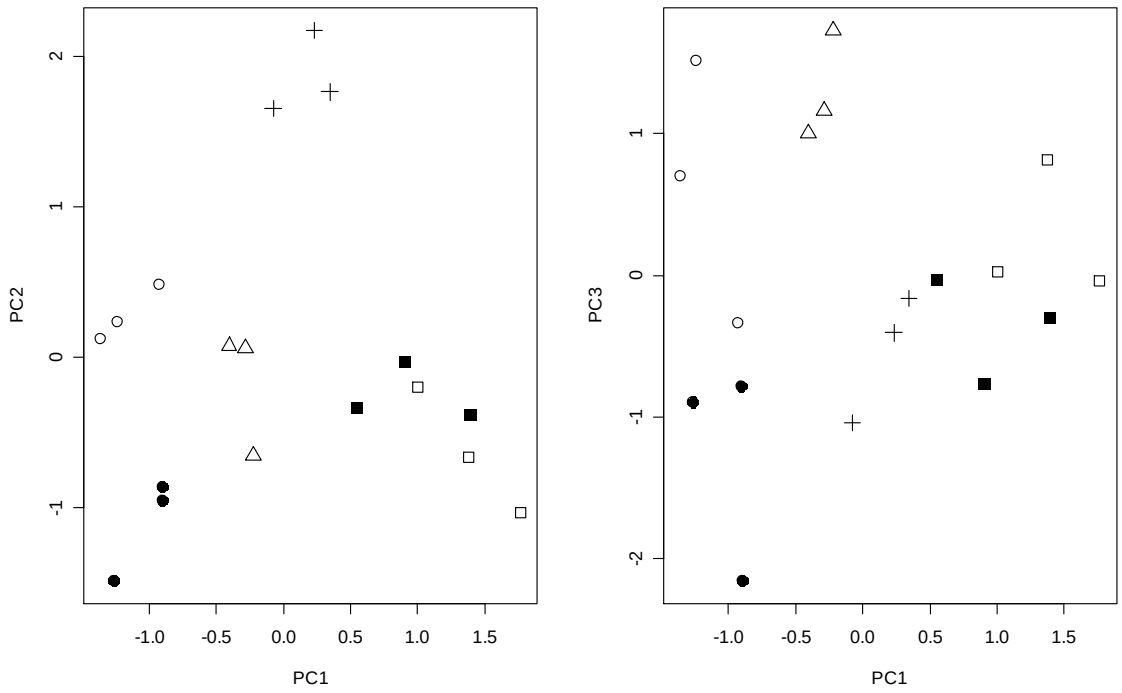


Figure 4.2 - PC score plots to show that variation between separate measurements taken from the same face (intra-measurement error) is smaller than facial variation (inter-individual variation). Six different faces are represented by different symbols; three different measures were taken of each face.

Using a larger dataset to formally assess if differences between mean shapes of faces were significantly more than differences attributed to repeated landmark coordinates taken on the same face a multivariate analysis of variance (MANOVA) was performed, Venables and Ripley (2001). The MANOVA formally assessed whether the within-face variation was significantly less than the between-face variation.

The data consisted of $n = 48$ different images taken from sections B and C (Table 2.2) of the facial identification catalogue. Eight corresponding facial landmark points were placed in all images resulting in $k = 16$ variables (eight landmarks in 2D). The measurements were taken $p = 3$ times. Procrustes analysis was applied to the 144 observations in the data set (i.e. three separate measurements of sixteen coordinates for forty-eight images) and the full Procrustes coordinates were obtained. These coordinates were found to be very highly correlated and so were transformed onto principal components to overcome the problem of singularity in the MANOVA calculations. The scores of the principal component analysis were used for the MANOVA. The total sum

of squares for within-measurement variation, $SS_w = 474.92$, when compared to the total sum of squares for between-face variation, $SS_B = 22145.29$ was small. The p-value for the MANOVA test with *(n-p)-k+1* degrees of freedom was very small ($p<0.001$) showing strong evidence that the between-face variation was greater than the within-measurement variation.

4.3.1.2 Variation Attributed to Scanning

To examine the variation attributed to taking different scans of the same face twelve photographs were selected from the facial catalogue and digitally scanned three times. In a similar way to intra-observer error it was required that intra-scan error be sufficiently smaller than inter-individual or between-face variation. Three landmarks from the thirty-six scans (three of each of the twelve images) were Procrustes aligned and the PC scores of the tangent coordinates were examined. Figure 4.3 shows six different faces represented by different symbols, as we saw with landmark placement error, measurements taken from different scans of the same face were clustered together in distinct groups for each face with marginal overlapping between different faces.

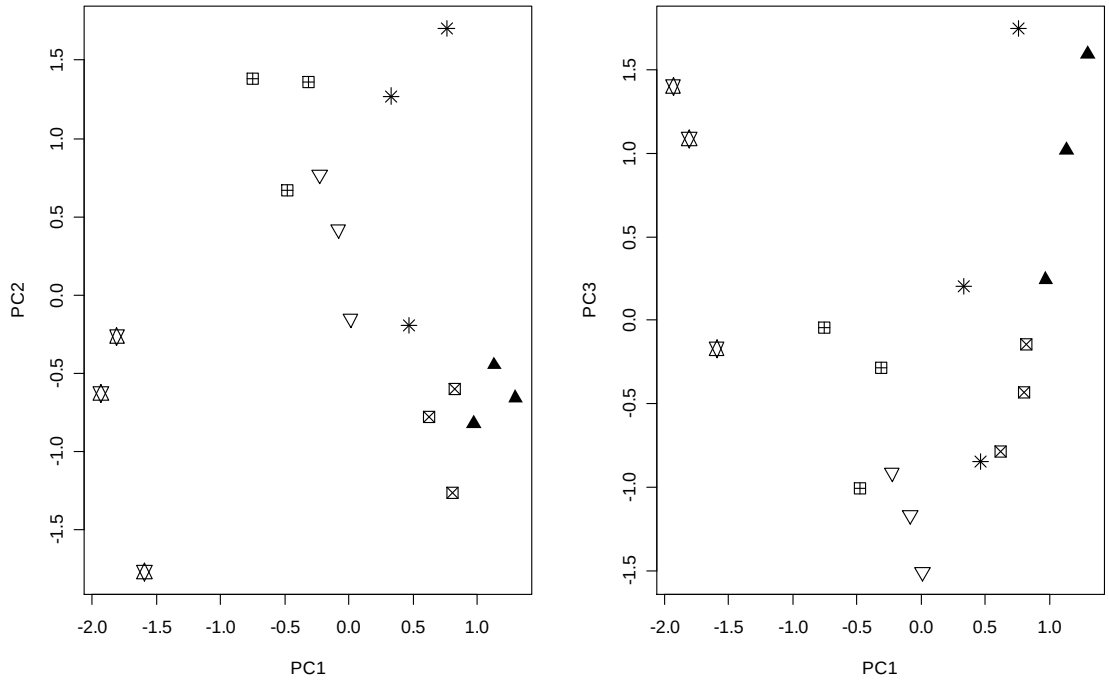


Figure 4.3 - PC score plots showing that the variation attributed to different scans of an image was smaller than inter-facial variation. Six different faces are represented by different symbols; there were three scans per image.

A MANOVA on the PC scores was carried out to formally assess whether between-scan error was smaller than between face error, the test p-value was very small ($p<0.001$) indicating lack of support for the null hypothesis.

4.3.2 Facial Matching

Two methods were examined for finding matching faces in the data collected; Cluster analysis, Venables and Ripley (2001), and likelihood ratios, Aitken and Lucy (2004). For each set of comparisons two or more sections of the facial catalogue were taken and searched for matches; using the largest possible subset of landmarks visible in all images from all catalogue sections, Table 2.2. Mean coordinates for each landmark (calculated from three separate measurements) for each image were Procrustes aligned and the Procrustes tangent coordinates were obtained. The following results are for comparing the forty-eight faces described above from sections B and C of the facial catalogue, these faces had in common eight landmarks positioned around the eye area.

4.3.2.1 Cluster Analysis

The Mahalanobis distance matrix of the Procrustes tangent coordinates was used in a cluster analysis to search for groups or matches of similar facial shape in the data. For this investigation the method of hierarchical clustering chosen for the analysis of the facial data was single-linkage cluster analysis, Venables and Ripley (2001). This method was chosen as we were looking for matches in the data, i.e. pairs of the observations with closest shape (and therefore the least dissimilarity).

A dendrogram or classification tree gives an informative view of the results of clustering and represents the minimum variance hierarchical classification. Dendrograms were drawn for each set of catalogue sections being compared. They were examined for clusters containing pairs of observations and each pair was manually checked at the data source to determine visually if the two images could be the same person.

A dendogram displaying clusters of similarity in the data can be seen in Figure 4.4; here one match (circled) was found between two different images. There existed other pairs in the hierarchy which were found not to be facial matches, so the clustering method may not be an appropriate one to use, although increasing the number of landmarks and amount of facial coverage they demonstrate could improve on results.

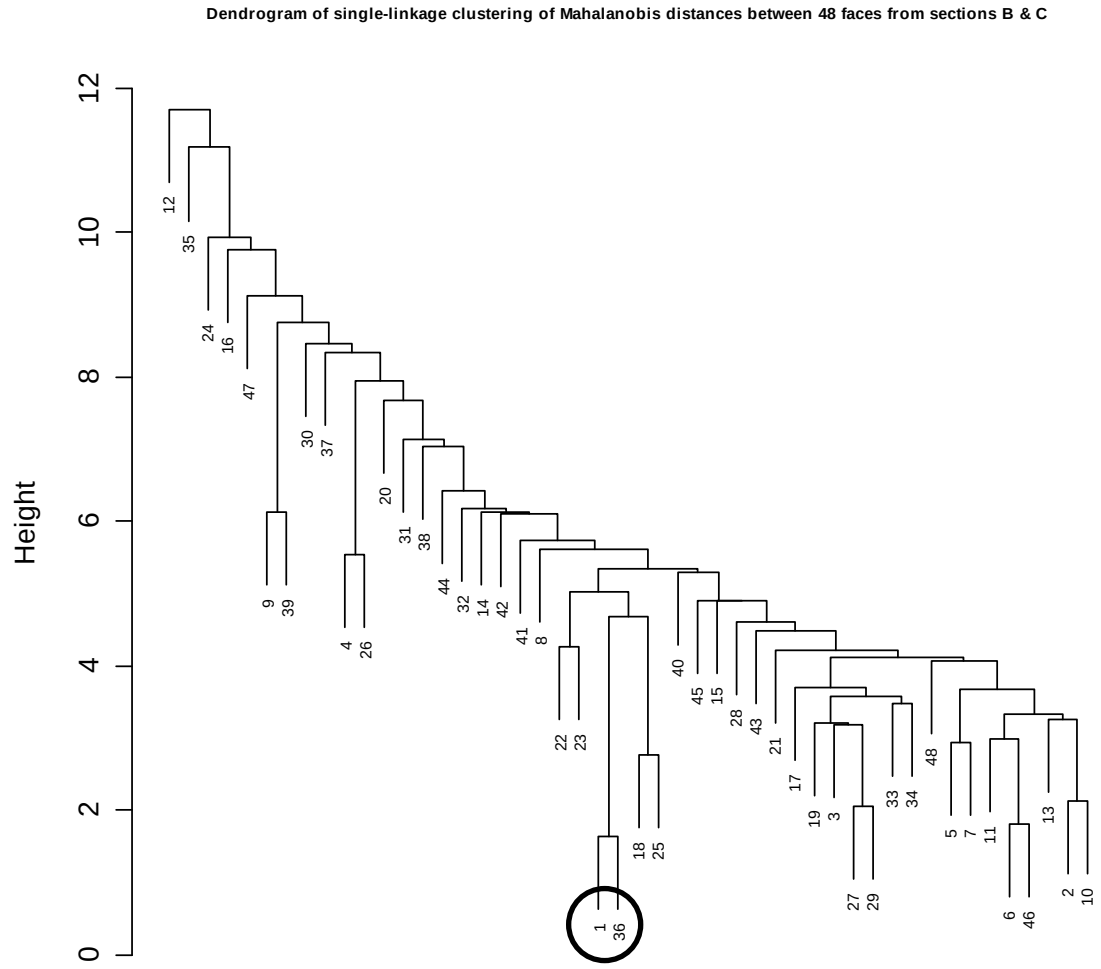


Figure 4.4 – Dendogram to show the results of a single-linkage cluster analysis carried out to assess possible facial matches between images in two different sections of the catalogue[2].No. 1–26 represent images from section B and 27–48 from section C. Height refers to the distance between the clusters.

4.3.2.2 Likelihood ratios

Aitken and Lucy (2004) describe an approach for evaluating forensic evidence using likelihood ratios (LRs), the example they use is for matching multivariate glass fragment data and is described further in §3.8. The facial data can be analysed in a similar way, taking PC scores of the Procrustes registered tangent coordinates as multivariate variables for input into the LRT, using a multivariate normal model. This method of matching was carried out on the triplicate measurements for the forty-eight faces.

Increasing numbers of PCs (from two up to ten) were explored as the facial matching variables; the effect on the LR matching results was examined. The top ten 'match' results in terms of largest LRs obtained are displayed in Table 4.1. The top three strongest results were a true facial match, the same one as found and circled in the dendrogram in Figure 4.4. The next best matches (4 to 10, Table 4.1) were found to be false positive results, however the magnitude of LRs for the true matches was greater than for the false matches indicating that perhaps some kind of threshold for confirming a match (e.g. LR>200) was required. One limitation here is that the eight landmark points used throughout the analyses are from the extremes of the face (chin and forehead) and around the eye area. A more substantial study which looks at the whole face is required to get a more comprehensive representation of the shape and all the features of the face.

The preliminary conclusion that can be drawn from Table 4.1 is that it may not necessarily be better to use more PCs to search for matches. It is seen that the strongest match result in terms of the magnitude of the LR was for eight PCs, then the second and third strongest were for nine and ten PCs respectively. There was further evidence of this in the main study, see §7.2.3.

Face i	Face j	N	LR	No. PCs
1	36	48	415.8064	8
1	36	48	306.5394	9
1	36	48	303.985	10
2	10	48	190.0974	10
7	10	48	159.4036	8
22	32	48	147.1404	6
1	36	48	143.386	7
5	7	48	138.8363	8
5	7	48	133.2008	9
22	42	48	132.6324	7

Table 4.1– LR results for pair-wise facial comparisons within the set of forty-eight faces from sections B and C of the facial catalogue. The top three (strongest) matches were between faces 1 and 36 in the dataset, visual assessment suggested these 2 images were an actual match.

4.4 Summary

Using statistical shape analysis with the quantitative measurements of facial landmarks has shown potential for identification and facial matching. MANOVA results have established there is lack of support for the null hypotheses that aligned and transformed facial landmark data show no difference in mean face shape between repeated landmark measures and repeated scans of the same image. The results were promising even though only a proportion of facial features were examined due to the censoring in the images; a further study giving a more complete representation of the face is required. Chapters 5 to 8 describe a much larger study.

Cluster analysis was applied to find possible facial matches in the data. A single-linkage cluster analysis found two out of the forty-eight images analysed were a visual facial match. Although a dendrogram is a nice way to visualize results there were several pairs found in the data that were false matches, indicating that clustering may not be a statistically appropriate technique for matching the facial shape data. Also when looking at displaying much larger datasets dendrograms are inappropriate.

LRs for evaluating the strength of a facial match are another way of quantifying results, modelling the data with a multivariate normal distribution. The top three matches found using this method were true matches, there were also false positive results here, though these could be overcome by applying a threshold to the LR results, here LR>200 would have been an appropriate level to only select the true matches. A substantially larger study of faces is carried out in Chapters 5 to 8 for a more comprehensive investigation.

We have explored the situation where a partial facial matching could be done successfully; this is useful for real life crimes where the perpetrators mask their facial features in someway. The images examined in this study were all taken of faces in the anterior position (i.e. the subject was looking towards the camera). In real life criminal situations, for example where we wish to identify someone from some CCTV footage, it is unlikely that a criminal will look directly at the camera in this way. Therefore further investigations could be carried out to see if the techniques applied in this study could also be applied to non-anterior facial images. This has not been investigated here, although §8.4 uses three dimensional facial data to simulate what happens to matching results when the face is rotated a few degrees away from the anterior view.

In summary, the results of the pilot study have proven that statistical shape analysis and likelihood ratios are effective methods to use in quantifying facial matches. Precise empirical measurements of coordinates of attributes of the face are used, which gives these methods a clear advantage over the current techniques used for facial identification. It also means previous studies of facial variation could be used to permit the probability of a credible match to be empirically established; a database of measurements could be expanded as more facial data is collected to improve the model used in the matching.

5 Preliminary Examination of Variation Prior to Data Collection

5.1 Introduction

This chapter outlines some preliminary studies that were carried out at an early stage before all the facial scan data (§2.4.2) were collected. There were two main aims of the work, the first was to determine the most appropriate facial landmark data (§2.2) to collect from the Geometrix® facial database, described in §2.4, for use with facial comparisons (§5.2.6). The second main aim was to check the repeatability of the data collection method for the chosen facial landmarks, using multiple observers to collect the landmark data (§5.3). A key objective of the facial comparison method being developed here was that the process did not depend on the observer, therefore improving on the expert witness opinion (as described in §4.1). It was necessary to select a subset of landmark variables which were consistent to place, so that different observers could produce comparable data.

Section 5.2 describes how sixty one points (Table 2.3) were investigated on a small subset of faces (§2.5) to determine the best points to use for facial comparison. Two different observers collected data on the subset of faces, an examination of intra and inter-observer error was carried out. Principal components analysis (PCA) was performed on the tangent shape coordinates (§3.7, §4.3) to check the variation in the shape data from the two observers (§5.2.2). There were seen to be some differences, these were investigated further by looking at differences in the landmark locations for the two observers to check the consistency of placement (§5.2.2). The Mahalanobis distance between the two observers data was used to investigate statistical significance between the data for each landmark, the worst performers in terms of consistency were excluded from the analyses (§5.2.3). The ability of the landmarks to discriminate between different subjects was explored by looking at distance matrices and cluster analyses to observe groups of similarity in the data (§5.2.4). The most important landmarks in terms of consistency and discrimination were measured using Mahalanobis distances and Wilks' lambda respectively (§5.2.5). Examining all the results a set of thirty landmarks (§5.2.6) to collect from the main database (§2.4) were selected as the most appropriate of the sixty one investigated for facial comparison and matching.

Section 5.3 describes how after choosing thirty landmark points the repeatability of the method for data collection using six different observers for landmark placement was checked. The landmark placement manual (Appendix B) was used as a guide for six different observers who collected the 3D landmarks for a small subset of faces (§2.6). Some initial training and experience was necessary for capturing reliable landmark measurements. PCA on the tangent shape coordinates (§3.7, §4.3, §5.2.2, §5.2.3) explored the variation in the shape data obtained from the different observers (§5.3.2). Plotting the mean face shapes for each pair of observers highlighted differences in the location of particular landmarks (§5.3.2) and some key items of guidance needed to be clarified. Cluster analyses were carried out to look for groups of similarity in the data (§5.3.3). The ten faces under investigation were separated into distinct clusters. Therefore the data collected from multiple observers was deemed comparable for the subset of faces investigated, indicating that the technique for data collection using the thirty chosen landmarks was repeatable.

5.2 Selection of Landmarks for Data Collection

This section benefited from the expertise of Professor Ian Dryden in the modification of the R 'Shapes' package to extend 3D analyses capabilities. Initial work was carried out under IDENT (Evison and Vorder Bruegge, 2008), this was repeated later for the purposes of this thesis, as described in §2.5, §5.2.1-§5.2.6.

5.2.1 The Data and Procrustes Registration

The data for this reliability study were described in detail in §2.5. Orthographic projections of the raw data are displayed in Figure 5.1. Generalized Procrustes analysis (GPA) without scaling was carried out to extract the facial shape information from the data, §3.6. Scaling was excluded because during the image collection procedure the subjects were placed at the same (measured) distance from the 3D scanner, so all images and landmark data extracted from them should be equivalent in scale. The arbitrary units of scale are omitted in all figures. The orthographic projections of the Procrustes registered data are displayed in Figure 5.2. Comparing Figures 5.1 and 5.2 it can be seen that after the Procrustes alignment the data are split into clear clusters representing the different landmark points.

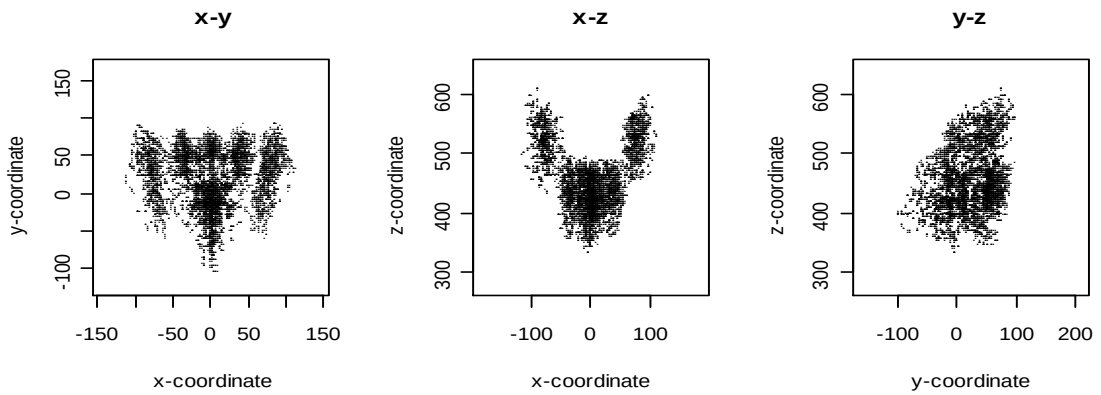


Figure 5.1 - Three orthographic projections of the raw landmark coordinate data. The x-y plot shows the anterior facial view (subject forward facing); x-z shows the overhead view (subject nose facing downwards) and y-z shows the profile facial view (subject left facing).

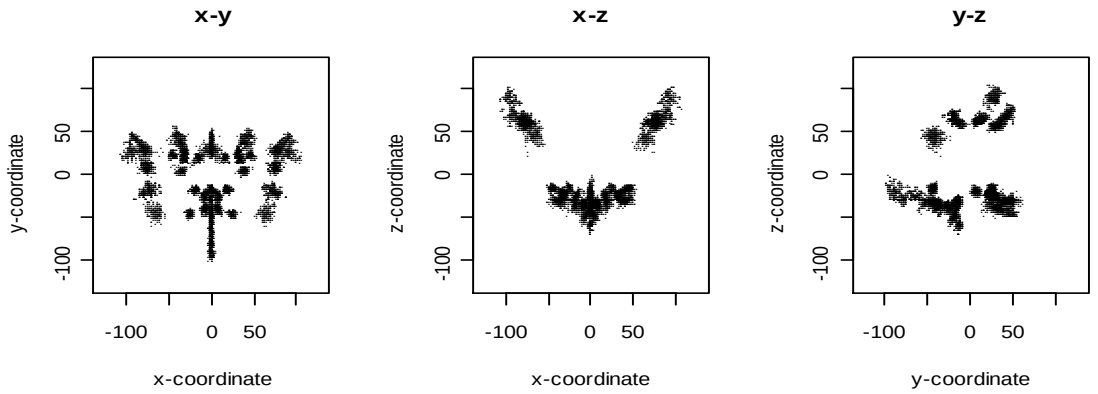


Figure 5.2 - Procrustes aligned landmark data, using translation and rotation (preserving scale). The x-y plot shows the anterior facial view (subject forward facing); x-z shows the overhead view (subject nose facing downwards) and y-z shows the profile facial view (subject left facing).

5.2.2 PCA and Consistency between Observers

Principal components analysis (PCA) was carried out on the Procrustes tangent coordinates (see §3.7, §4.3) after centring the aligned data. Plots of pairs of the first few principal component (PC) scores for the data from two observers are shown in Figure 5.3; the percentage of the data variation contained on each pair of PCs is displayed above the plots. The second and third plots in Figure 5.3 pick up some systematic differences in the data for the two observers, seen in the separation of the two different symbols.

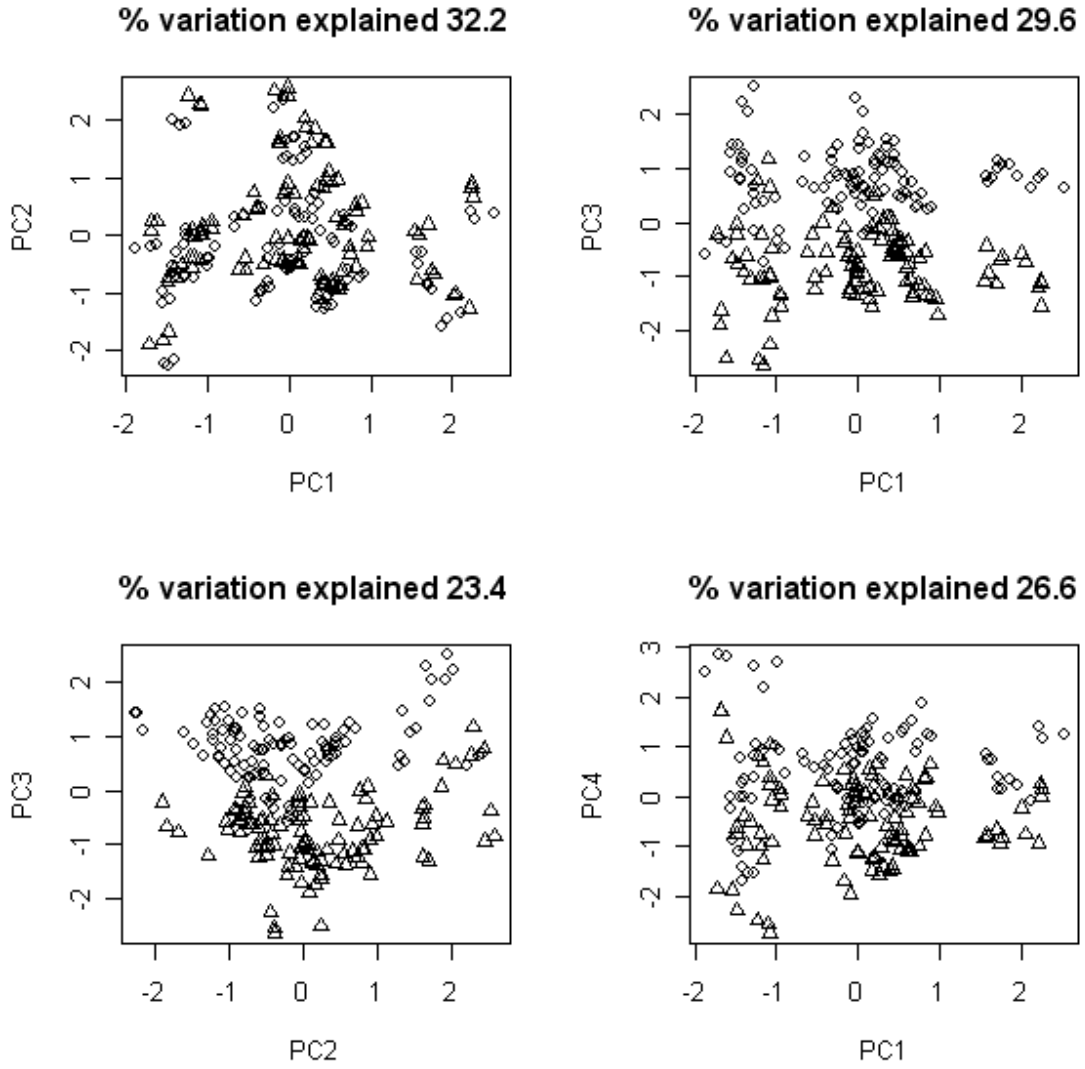


Figure 5.3 - The first few PC scores, symbols indicate the data from different observers. The percentage of variation explained by the two plotted PCs is displayed above each plot.

The mean shape configurations for each of the two observers (without Procrustes registering the data) were examined to investigate where the differences between observers lay, in terms of the landmark points on the face. The x-y orthogonal view of the face was plotted, Figure 5.4. The dashed lines join the mean landmarks from observer L; black solid vectors are drawn to the mean landmarks of observer X. The longer the vectors the more divergent the landmark position was between observers, hence the most variable landmark points.

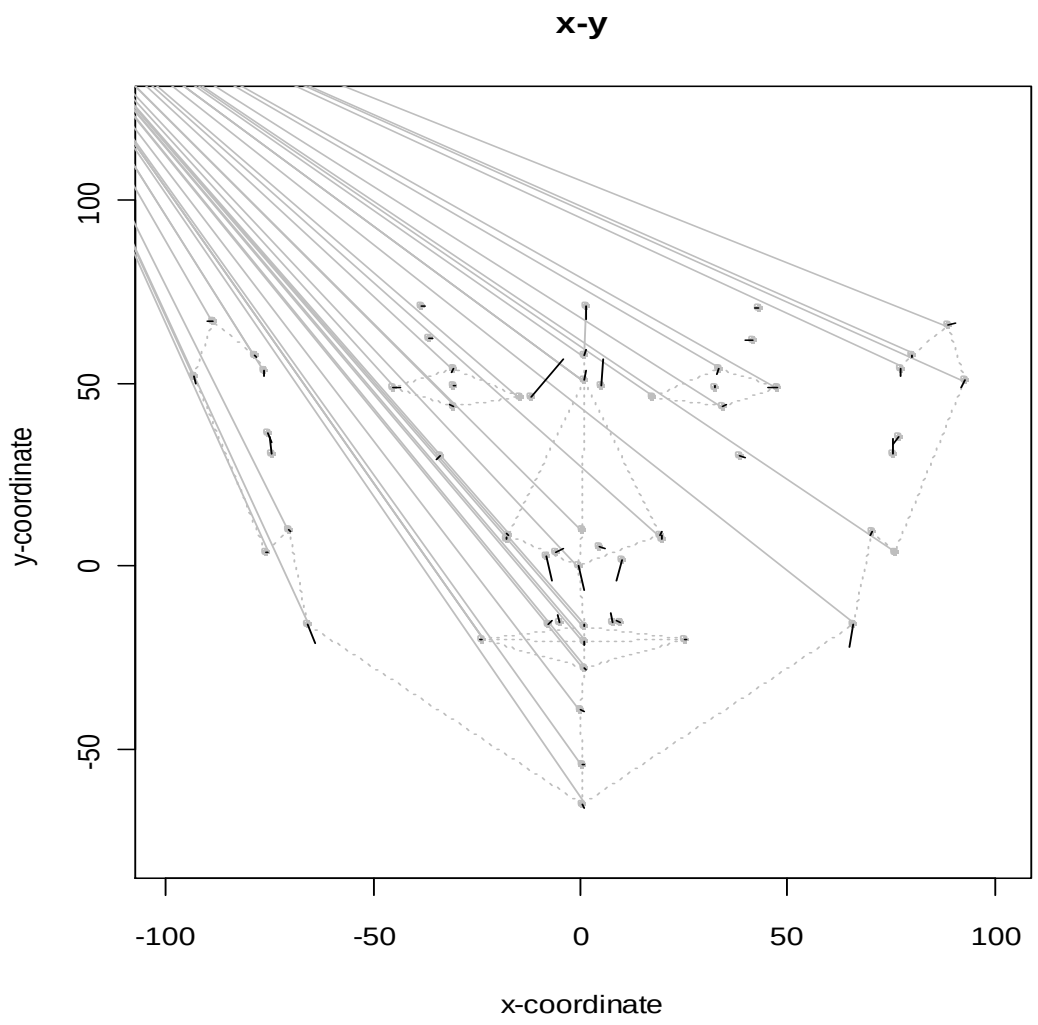


Figure 5.4 - The mean landmark configuration for Observer L (grey dashed lines) with vectors drawn to the mean landmark configuration for Observer X (solid black lines).

The mean locations of the majority of landmarks as measured by the two observers were very similar. Certain points around the ears, between the eyes and around the tip of the nose showed some larger differences. To investigate statistical significance between the observers the two sample Mahalanobis distances were calculated between the non-

registered data at each landmark and subject. As there were only three observations in each sample 1 was added to the variance in each case to make distances more stable, this had the effect that only larger mean distances contributed to a large distance, Dryden (Pers. Comm.). The Mahalanobis distances and the median at each landmark are displayed in Figure 5.5.

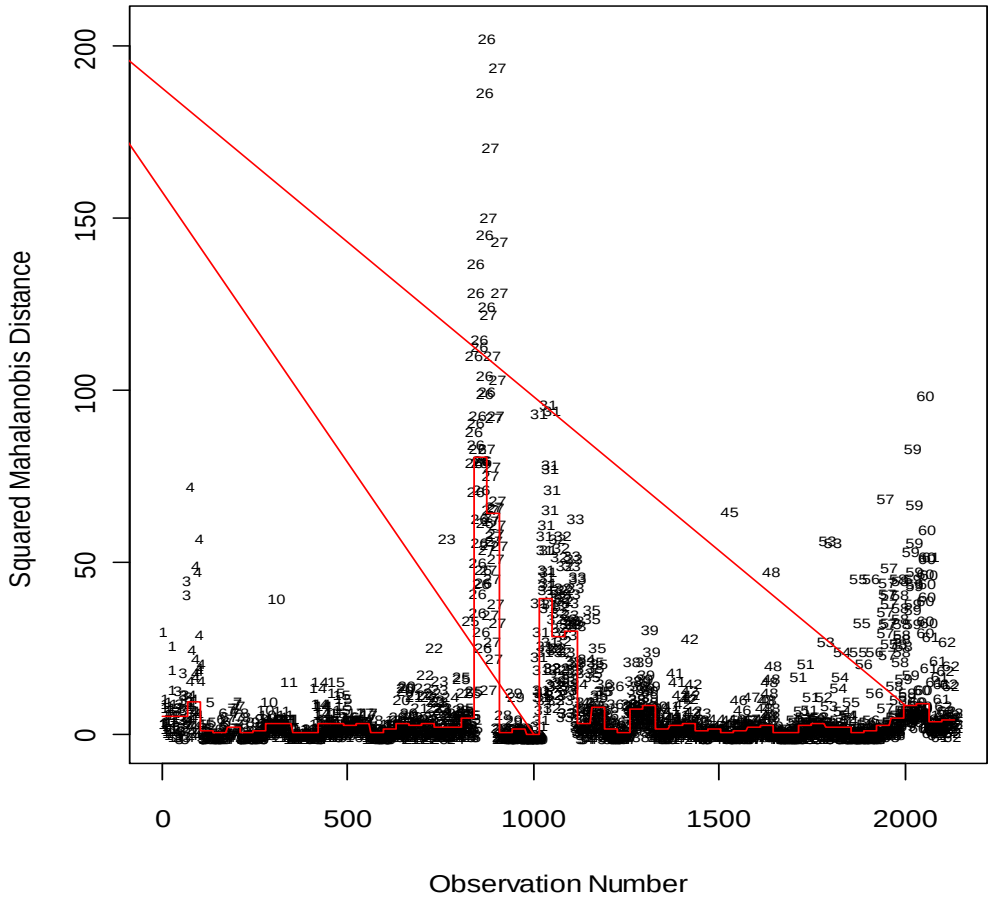


Figure 5.5 - Mahalanobis distances for each observation at each landmark (numbered 1-61), the median distance for each landmark is displayed in grey.

Landmarks with the most different values (i.e. highest distances) between observers were numbers 25-26, 30-32 and 56-59, Table 2.1. In terms of consistency these were the worst performers and would not be good choices for facial comparisons and matching. These landmarks were excluded from the data to see the effect this had on the PCA and consistency, see following section.

5.2.3 Excluding the Least Consistent Landmarks

Based on the Mahalanobis distances calculated in §5.2.2 some or all of the worst performing landmarks (in terms of consistency to place) were excluded from the data for the subsequent analysis. The following four nested subsets in decreasing order of size were considered:

Subset I: the initial dataset (from §5.2.1)
Subset II: the initial dataset excluding landmarks 25 and 26
Subset III: the initial dataset excluding landmarks 25, 26, 30-32
Subset IV: the initial dataset excluding landmarks 25, 26, 30-32 and 56-59

Scores for the first few PCs for subsets II, III and IV are displayed in Figures 5.6, 5.7 and 5.8 respectively. Subset II showed there were still some systematic differences between the data from the two different observers, Figure 5.6. Subsets III and IV showed no obvious differences between observers, Figures 5.7 and 5.8.

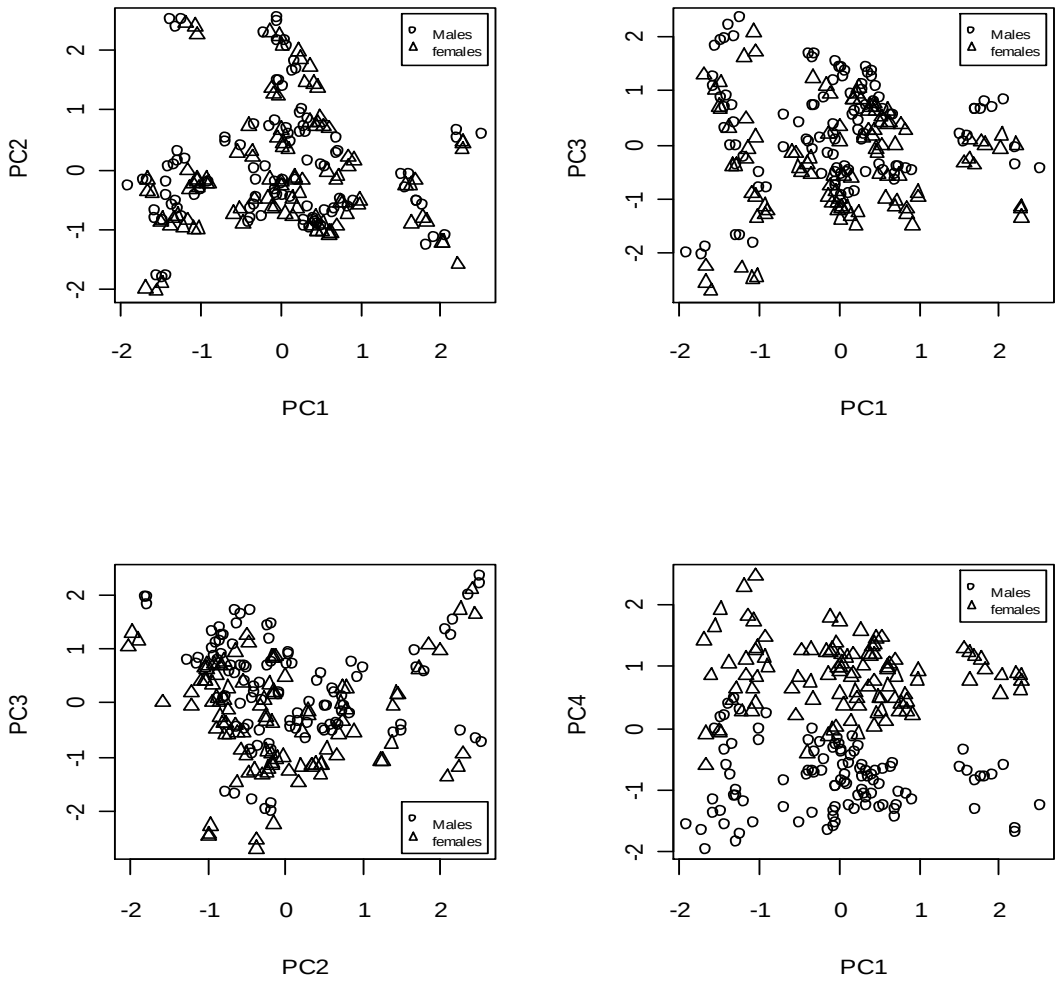


Figure 5.6 - The first few PC scores for subset II, the two symbols indicate data from two different observers. PC1 V PC4 still shows some systematic differences between the two observers, indicated by the separation of the symbols. The percentage of variation explained by the two plotted PCs is displayed above each plot.

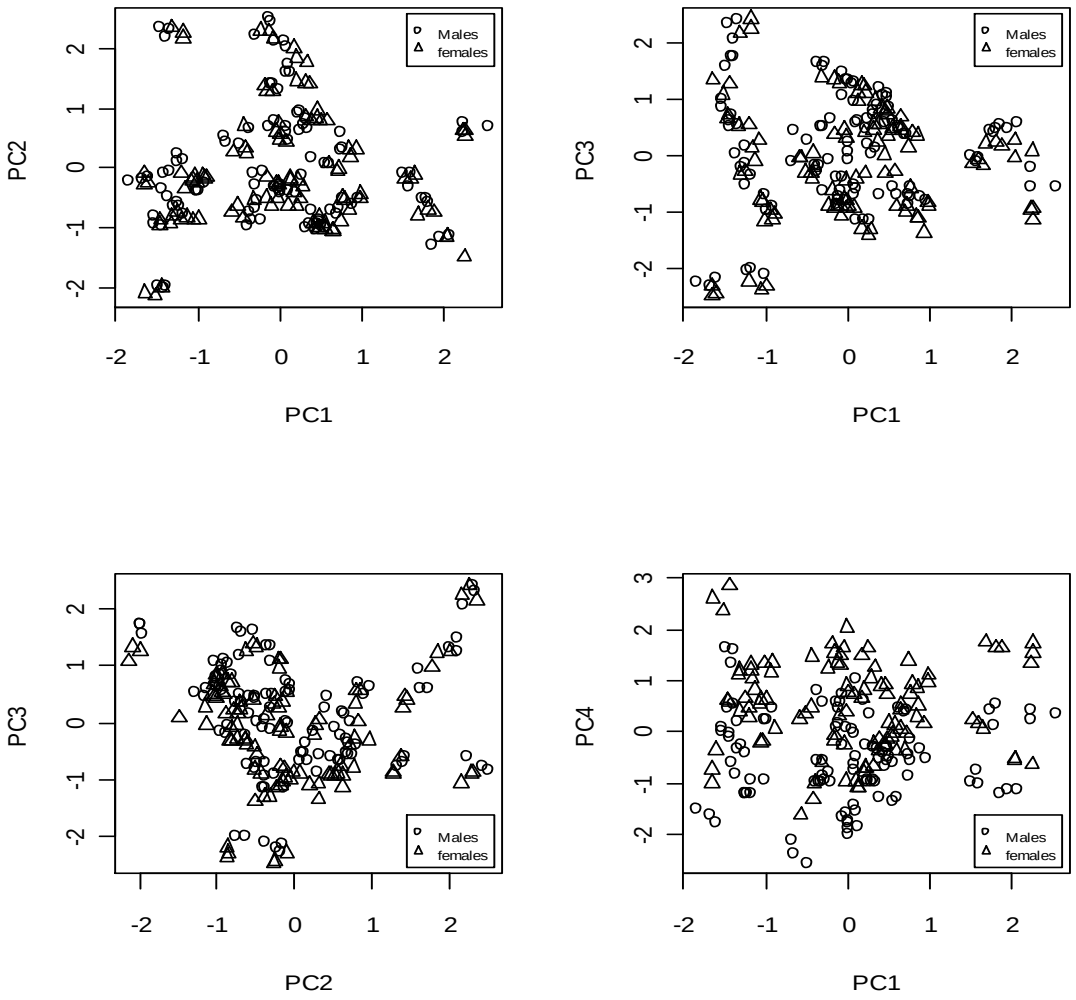


Figure 5.7 - The first few PC scores for subset III, the two symbols indicate data from two different observers. The percentage of variation explained by the two plotted PCs is displayed above each plot.

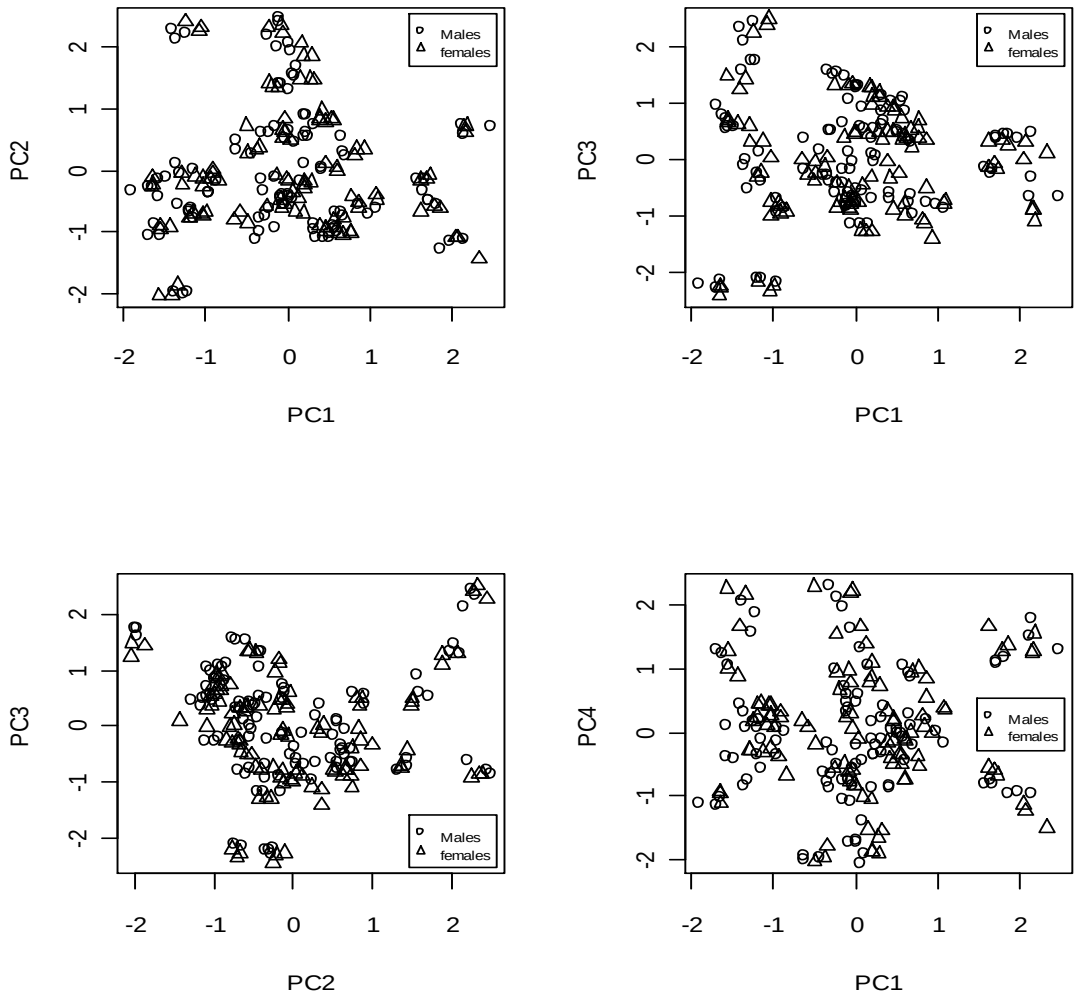


Figure 5.8 - The first few PC scores for subset IV, the two symbols indicate data from two different observers. For this subset there is no clear distinction between the two observers in the first few PCs. The percentage of variation explained by the two plotted PCs is displayed above each plot.

5.2.4 Discrimination between Subjects

An additional requirement of the landmark points was that they had to show sufficient discriminatory power, i.e. have enough remaining variation (after accounting for observer differences) to be able to discriminate between different faces. To examine the potential for discrimination between different configurations, which would permit facial matching and identification, cluster analyses were applied to the subsets (§5.2.3). A Wards cluster analysis was carried out on each subset to look for groups of similar configurations. Ward clustering uses a weighted group average calculation to merge related clusters, taking into account the variances rather than simply distances between

clusters. The results for a Wards analysis on subset IV are displayed in Figure 5.9 and an enlarged section in Figure 5.10. It was seen that all four subsets had clusters of size three at the lowest level, which contained the three triplicate measurements of the same face (Figure 5.10). Subsets I and II displayed some clusters of size six that contained more than one face; however in subsets III and IV all clusters of size six represented one face (Figure 5.10). So, triplicate measurements taken from each face were more similar to each other than they were to measurements from different faces even when two different observers collected the data. All faces were clearly distinguishable on the dendrogram. These results demonstrated the potential for using data from multiple observers for identification and facial matching purposes, further validation of this is summarized in §5.3.

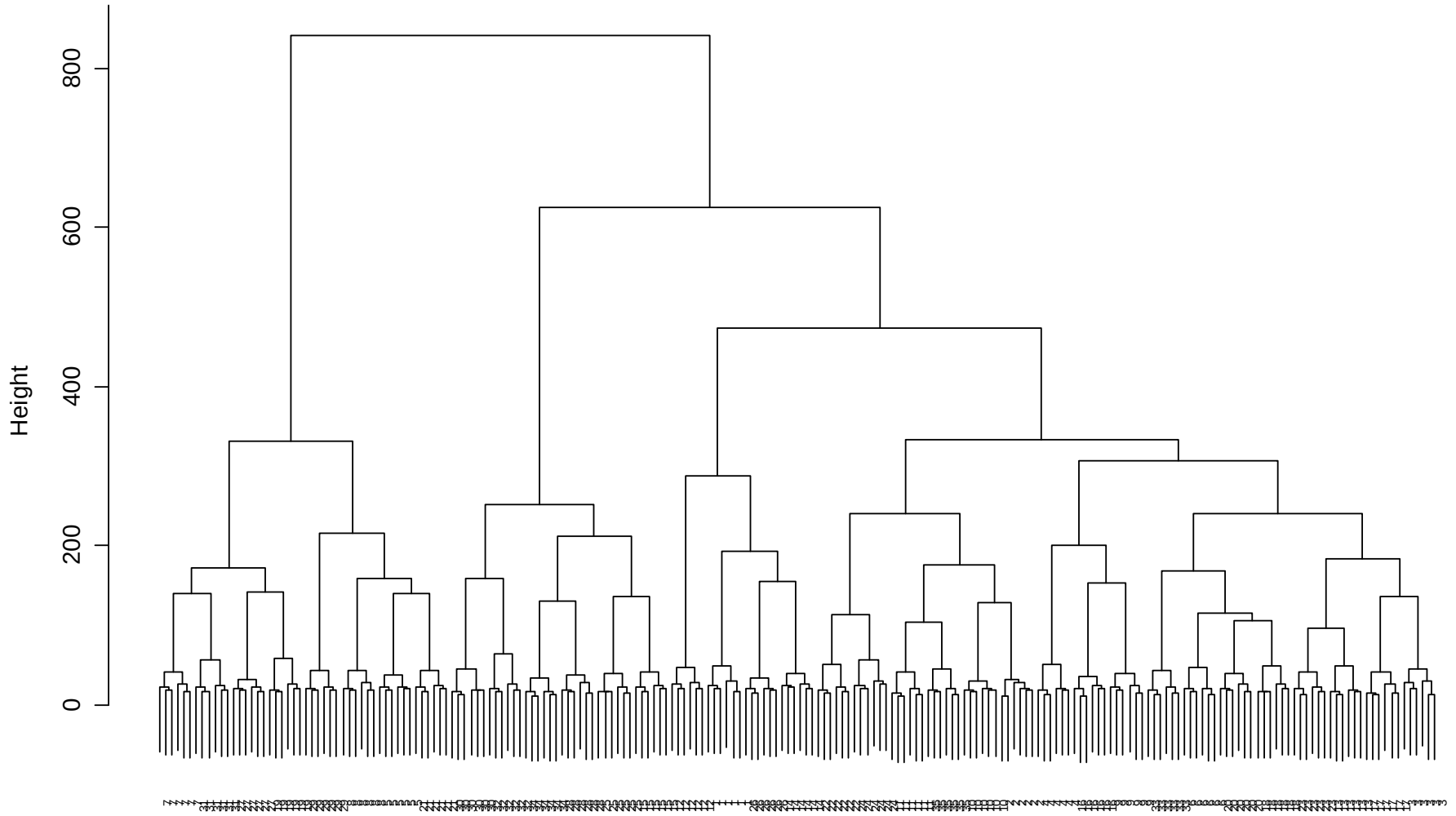


Figure 5.9 - Dendrogram displaying the results of a Wards cluster analysis for subset IV. Labels indicate subject face (1-35). Height refers to the square error of the clusters, which are added to those of their lower clusters.

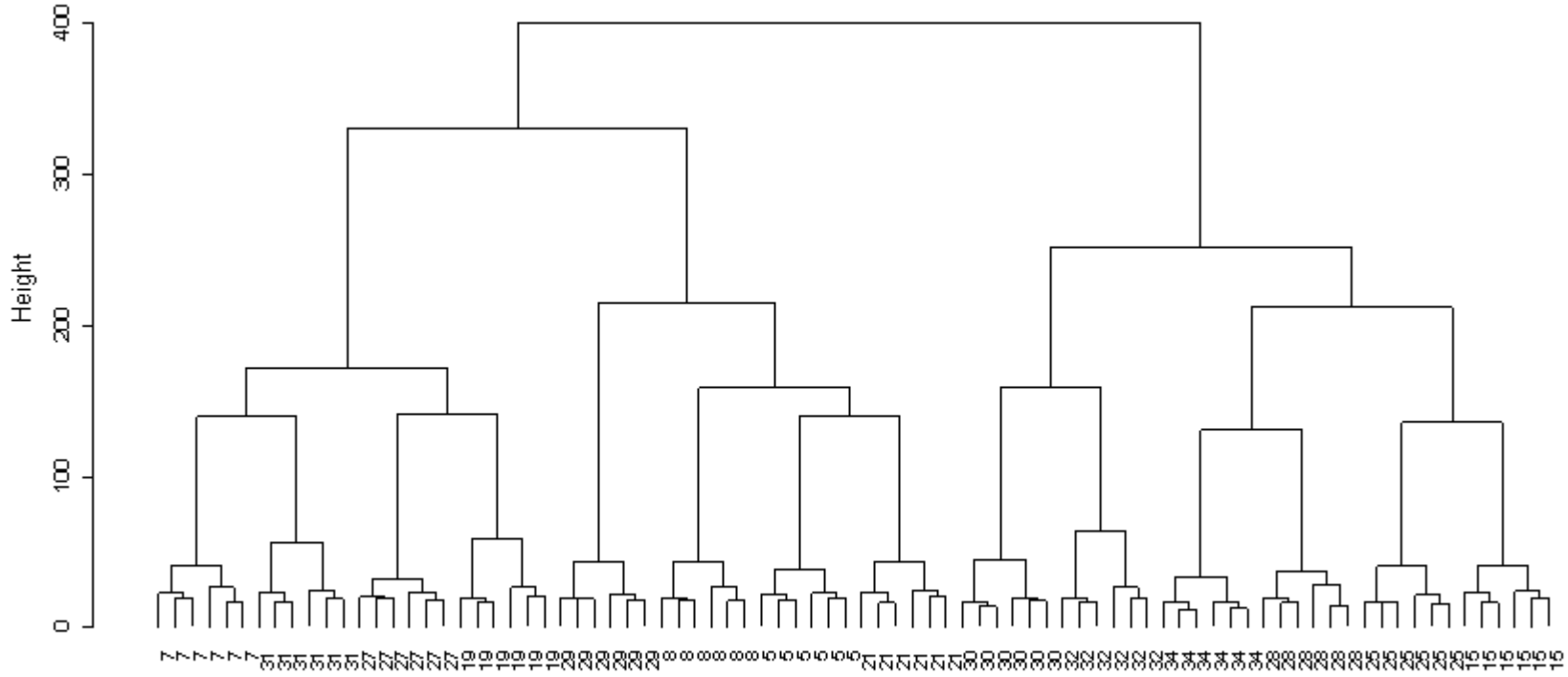


Figure 5.10 - Enlarged section of the left hand side of the dendrogram in Figure 5.10; the lowest level clusters in dataset IV group different faces (numbered). Height refers to the square error of the clusters, which are added to those of their lower clusters.

5.2.5 Important Landmarks

The following section explores which particular facial landmarks are good for discriminating between different faces and checks the consistency of the landmarks in subset IV. To examine formally which landmarks were the most important for discrimination Wilks' lambda (Λ) was calculated, using the within-subject variance (W) and between-subject variance (B) of the Procrustes registered data for each landmark point. Each landmark point is represented by a vector of length three, so W and B are 3 x 3 matrices and

$$\Lambda = \frac{\|W\|}{\|W + B\|}$$

Wilks' lambda is a general test statistic used in multivariate tests of mean differences among more than two groups, here the faces were the group variable. Low values of Wilks' Λ indicate landmarks that varied between subjects after accounting for the observer variability in W. Figure 5.11 shows the landmark positions (labels correspond to those in Table 2.1) for the mean face for subset IV for three orthogonal views, the size of the landmark label number is proportional to $-\log \Lambda$. Acknowledgement should be given to Ian Dryden for the provision of the R routine to draw the plots in Figure 5.11. The large numbers in Figure 5.11 indicate the least important landmark points in terms of discrimination, these points were therefore not a good choice for facial matching. The small numbers indicate landmarks which were good at discriminating between the faces in the sample after accounting for observer variability.

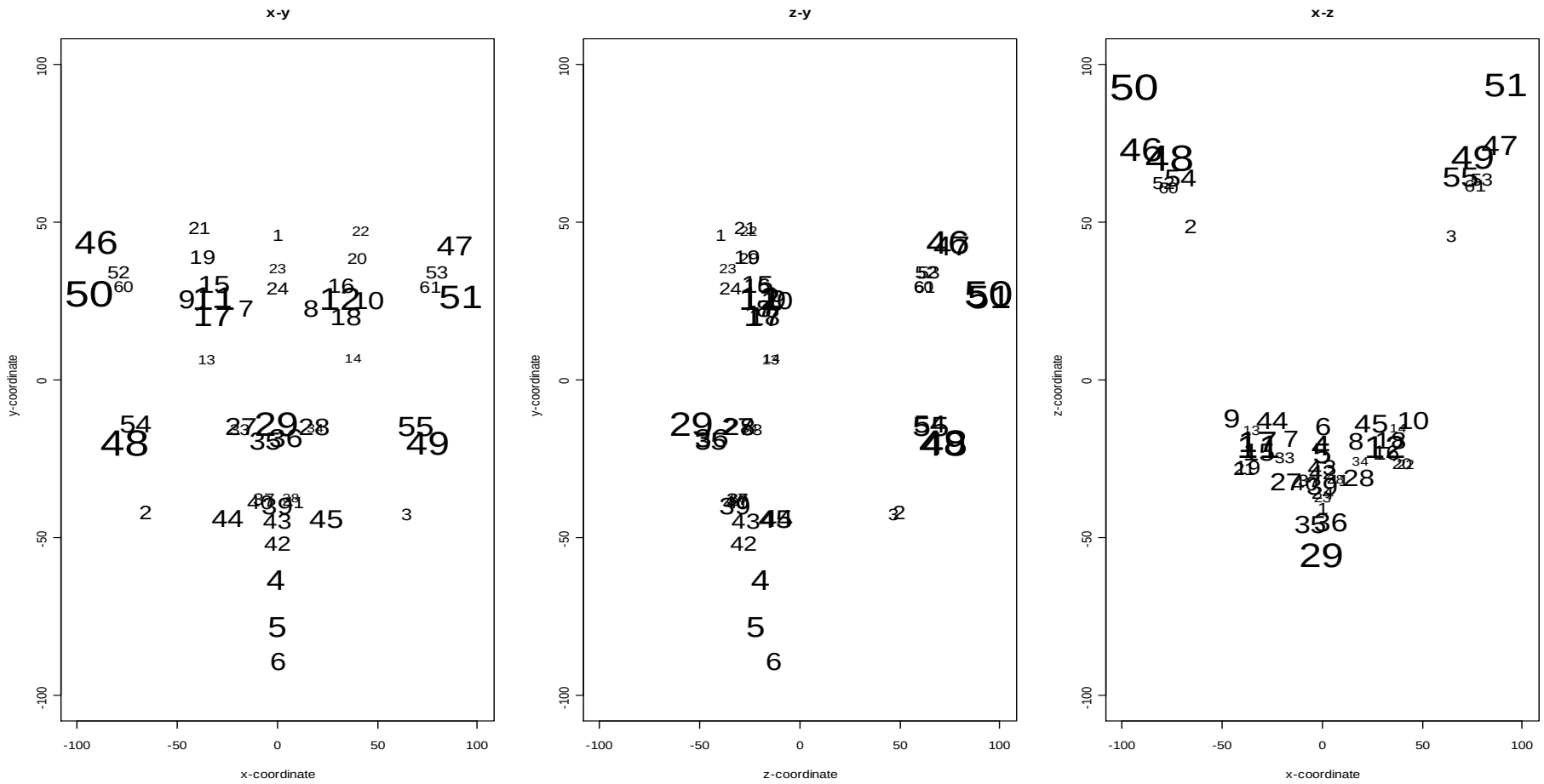


Figure 5.11 - Orthogonal views of the mean face for subset IV, the size of the landmark label indicates the discriminatory power between individuals for that landmark, the larger the labels the more discriminatory power.

Table 5.1 shows the numerical values for – log Λ, the landmarks are ranked in terms of decreasing –logΛ. The best landmarks, in terms of how well they discriminate between different faces, come early in the ranking in the list. Though the differences in the –log Λ values are small and results are likely to be sensitive if an alternative set of faces were analysed.

Discriminatory rank	Landmark number	-log Λ		Discriminatory rank	Landmark number	-log Λ
1	50	4.46		27	8	2.75
2	48	4.44		28	39	2.74
3	29	4.14		29	43	2.56
4	11	4.01		30	16	2.53
5	49	3.98		31	42	2.48
6	51	3.98		32	2	2.38
7	46	3.91		33	19	2.36
8	12	3.79		34	40	2.29
9	17	3.64		35	21	2.18
10	5	3.45		36	53	2.15
11	4	3.42		37	52	2.11
12	55	3.29		38	24	2.07
13	47	3.22		39	3	2.03
14	9	3.15		40	1	2.02
15	36	3.09		41	61	1.94
16	45	3.07		42	41	1.93
17	44	3.00		43	37	1.90
18	54	2.97		44	33	1.87
19	35	2.96		45	20	1.84
20	27	2.92		46	60	1.82
21	6	2.92		47	23	1.72
22	10	2.91		48	22	1.64
23	15	2.90		49	13	1.57
24	18	2.84		50	38	1.54
25	7	2.82		51	34	1.49
26	28	2.81		52	14	1.45

Table 5.1 - The ranking of landmarks (Table2.1) from subset IV in terms of discriminatory power (-log Λ) between subjects.

To look at the consistency between observers for each landmark in subset IV the Mahalanobis distances were calculated adding 1 to the variance for each group for stability (as in §5.22). Table 5.2 shows the order of landmarks in terms of the median Mahalanobis distance; the most consistently placed landmarks come early in the ranking. The analysis shows that the features are clearly sensitive; the differences between the median Mahalanobis distances for two consecutive landmarks in the ranked lists are small.

Consistency rank	Landmark number	Median Mahalanobis distance		Consistency rank	Landmark number	Median Mahalanobis distance
1	29	0.33		27	23	2.38
2	7	0.58		28	46	2.48
3	17	0.69		29	40	2.70
4	36	0.70		30	20	2.85
5	27	0.71		31	47	2.93
6	5	0.72		32	15	2.94
7	11	0.73		33	50	2.99
8	54	0.76		34	21	3.12
9	49	0.76		35	16	3.13
10	12	0.78		36	19	3.17
11	48	0.85		37	33	3.18
12	44	0.89		38	13	3.20
13	42	1.01		39	51	3.24
14	55	1.03		40	10	3.28
15	4	1.04		41	14	3.32
16	8	1.13		42	9	3.36
17	45	1.15		43	41	3.38
18	39	1.45		44	60	3.90
19	28	1.46		45	61	4.52
20	35	1.54		46	24	4.83
21	18	1.65		47	2	5.34
22	43	1.83		48	1	5.46
23	6	1.99		49	37	7.35
24	53	2.09		50	34	8.15
25	52	2.24		51	38	8.47
26	22	2.26		52	3	9.53

Table 5.2 - The ranking of landmarks (Table2.1) from dataset IV in terms of consistency

5.2.6 Landmarks to keep for Further Analysis

Using the consistency and discriminatory results in Tables 5.1 and 5.2, along with verbal feedback from the two observers, the list of landmark points for collection from the main facial database was reduced from sixty-one (Table 2.1) to thirty (Table 5.3). Some landmarks were excluded as they were thought to be particularly bad to place accurately, quickly and consistently in the opinion of the observers.

The first nine landmarks to be excluded were the ones with high median Mahalanobis distances in subset I, which were the source of the systematic differences seen between the two observers in the PCA score plots (Figure 5.3). These points were:

25, 26 - Maxillofrontale (Left and Right)
30 - Subnasale
31, 32 - Subalare (Left and Right)
56, 57 - Porion (Left and Right)
58, 59 - Tragion (Left and Right)

Other exclusions were based on how consistently the landmarks were placed by the observers (median Mahalanobis distance), how discriminant the landmarks were (Wilks' Λ), and whether the landmark point was visible in the majority of images in the facial database. After taking all these things into account the following landmarks were also excluded:

2, 3 - Gonion (Left and Right)
6 - Gnathion
13, 14 - Oribitale (Left and Right)
15, 16 - Palpebrale superius (Left and Right)
19, 20 - Orbitale superius (Left and Right)
21, 22 - Superciliare (Left and Right)
23 - Nasion
27, 28 - Alare (Left and Right)
40, 41 - Labiale superius prime (Left and Right)
52, 53 - Otobasion superius (Left and Right)
60, 61 - Preaurale (Left and Right)

This reduced the landmark list from sixty-one to thirty points, Table 5.3. A detailed landmark placement manual (Appendix B) was produced for these thirty landmarks to ensure all available observers followed the same procedures.

Landmark	Name
1	Glabella
2	Sublabiale
3	Pogonion
4	Endocanthion Left
5	Endocanthion Right
6	Exocanthion Left
7	Exocanthion Right
8	Centre point of pupil Left
9	Centre point of pupil Right
10	Palpebrale inferius Left
11	Palpebrale inferius Right
12	Subnasion
13	Pronasale
14	Alare crest Left
15	Alare crest Right
16	Highest point of columella prime Left
17	Highest point of columella prime Right
18	Labiale superius
19	Labiale inferius
20	Stomion
21	Cheilion Left
22	Cheilion Right
23	Superaurale Left
24	Superaurale Right
25	Subaurale Left
26	Subaurale Right
27	Postaurale Left
28	Postaurale Right
29	Otobasion inferius Left
30	Otobasion inferius Right

Table 5.3 - The reduced list of landmark points, which were collected on the whole Geometrix® database

The thirty landmark points were collected from the main facial image database, further details given in §2.4.3. Prior to this data collection the technique was validated for use with multiple observers by carrying out a similar study looking at intra-observer and inter-face variability in a small sample of ten faces (§2.6) to ensure that the thirty landmark points collected from different observers were comparable, section 5.3.

5.3 Repeatability of Technique for Data Collection

5.3.1 The Data and Procrustes Registration

There were six different observers available to carry out the placement of 3D landmark points on the images in the Geometrix® facial database. Prior to data collection a small subset of faces (§2.6) were used as a test dataset to check the repeatability of the landmark collection technique. Data collected by all six observers were checked to ensure that inter-observer consistency of landmark placement was of a suitable standard. Any single observer may be highly accurate at recording similar representations of a configuration on multiple occasions (this was also investigated); however representations must also agree with other observers who use the same technique. This also applies to other areas of forensic evidence evaluation, e.g. in fingerprint analysis several experts have to be in agreement that prints match, and results do not only depend on the opinion of one person.

The data (described fully in §2.6) were Procrustes aligned to remove the rotation and location information, §3.4; leaving behind the underlying shape of the configurations (as in §5.2.1 scale was also retained).

5.3.2 PCA and Exploration of Variability

PCA was carried out on the Procrustes registered tangent coordinates, for further details see §3.7. Figure 5.12 shows the first few PC scores; different symbols represent the six different observers. The percentage of the overall variation in the Procrustes aligned coordinates explained by each set of PCs is displayed above each plot. The first PC plot picked out some differences between observers; seen by a separation in the data symbols. Observers 1 and 2, represented by rings and triangles respectively, were showing some systematic differences to observers 3, 4, 5 and 6. Observers 1 and 2 were the two more experienced observers who compiled the landmark placement manual and agreed the procedures between themselves. The remaining observers only acquired their knowledge by reading the manual, so it was possible that something in there, or the tutorial given to the beginners, was not an accurate description of what was required.

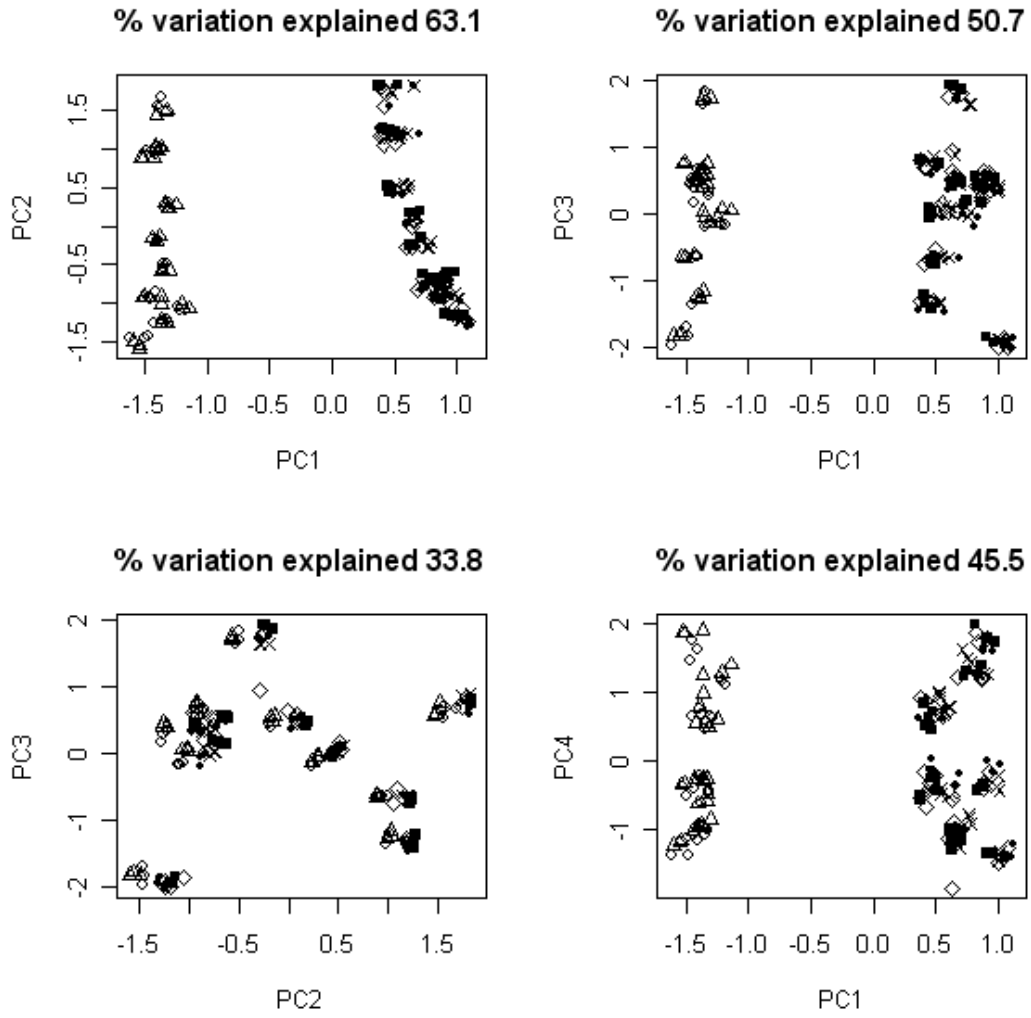


Figure 5.12 – The first few PC scores, observers 1-6 are represented by rings, triangles, crosses, diamonds, solid squares and solid circles respectively. The percentage of variation explained by the two plotted PCs is displayed above each plot.

To check the division seen in observer agreement the data were split into two subsets; observers 1 and 2 and observers 3, 4, 5 and 6. Procrustes alignment was carried out individually on these two subsets and a separate PCA run on each. The PC score plots showed that data from observers 1 and 2 produced comparable results across these first few PCs. No systematic differences were found when only comparing observers 3, 4, 5 and 6 either, therefore it was likely that these four beginner observers had been taught some part of the process differently to the way the experienced observers (1 and 2) actually carried the process out.

To examine where the disagreement lay between the two groups of observers, in terms of facial landmark points, the raw landmark data were plotted, as previously in §5.2.2.

The mean face shapes for each pair of the observers were compared using vector plots as in Figure 5.13.

In Figure 5.13 grey dashed lines join the mean landmarks for observer 1 and black vectors drawn from these indicate the corresponding position of the mean landmark for observer 3. The longer vectors show larger differences in landmark position between observers. Plots were drawn to compare all pairs of observers and all showed similar patterns between the experienced and beginner observers.

Figure 5.13 shows that the differences in location of most landmarks are only small; however three landmarks showed large disagreement between observers, these were the left and right alares and the pronasale, all around the tip of the nose. The same inconsistencies were picked up on plots showing the mean data from the other pairs of beginner observers Vs experienced observers. The middle plot in Figure 5.13 suggested that the problem was an issue with the ordering of landmark labels. Returning to the original data source the facial images with associated landmark points for each observer were viewed in the Forensic Analyzer® program, described in §2.4.3. It was clear that observers 1 and 2 had numbered the alare left, alare right and pronasale 13, 14 and 15 respectively, while observers 3-6 had labelled these points 14, 15 and 13 respectively. The mislabelling of these points was corrected, the PCA plots of the first few scores are displayed in Figure 5.14, and the percentage of variation explained by each plot is also displayed.

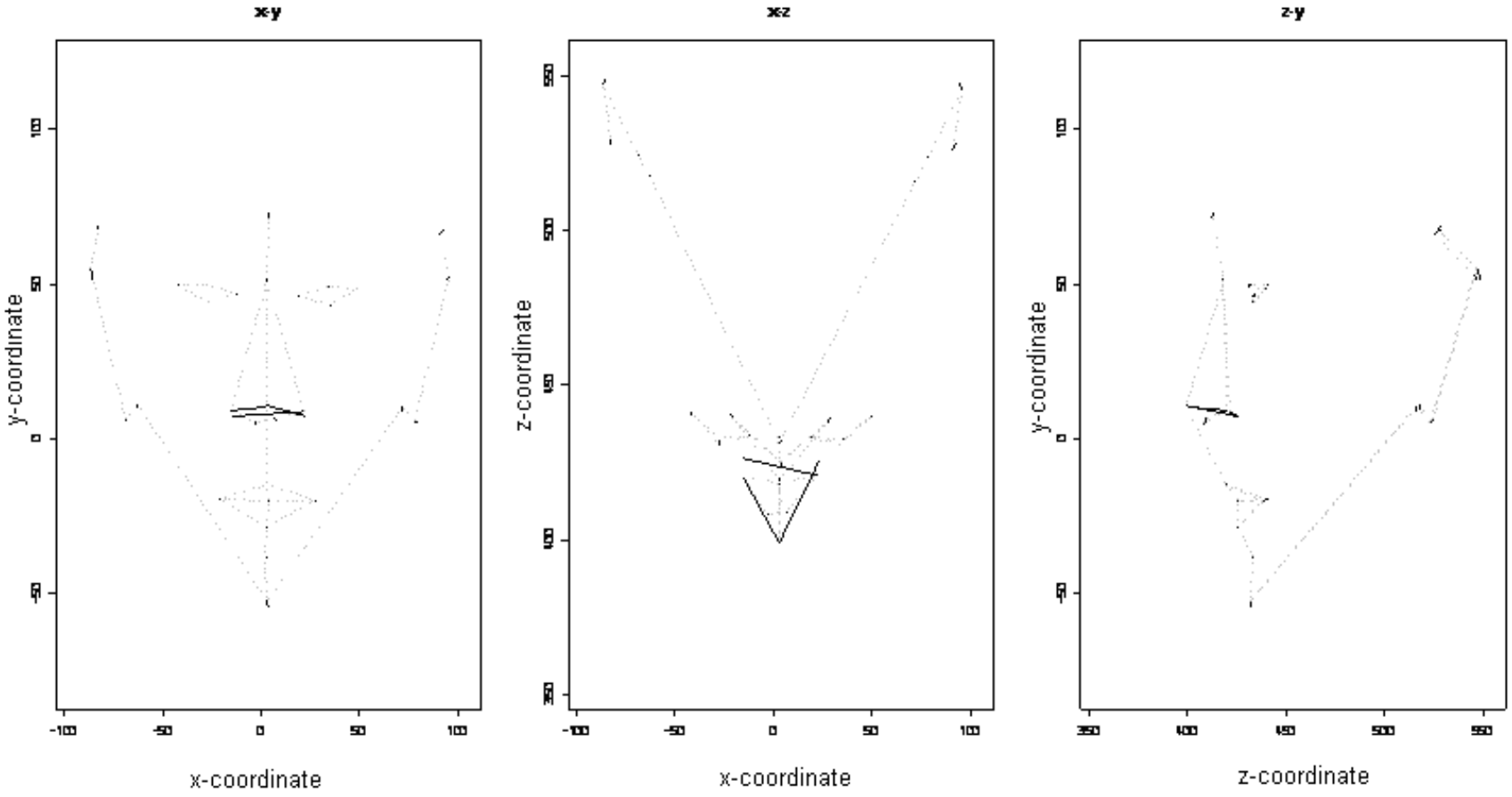


Figure 5.13 - Differences between Observers 1 and 3. Grey dashed lines indicate observer 1 shape, black lines indicate differences of observer 3 from observer 1 (see text)

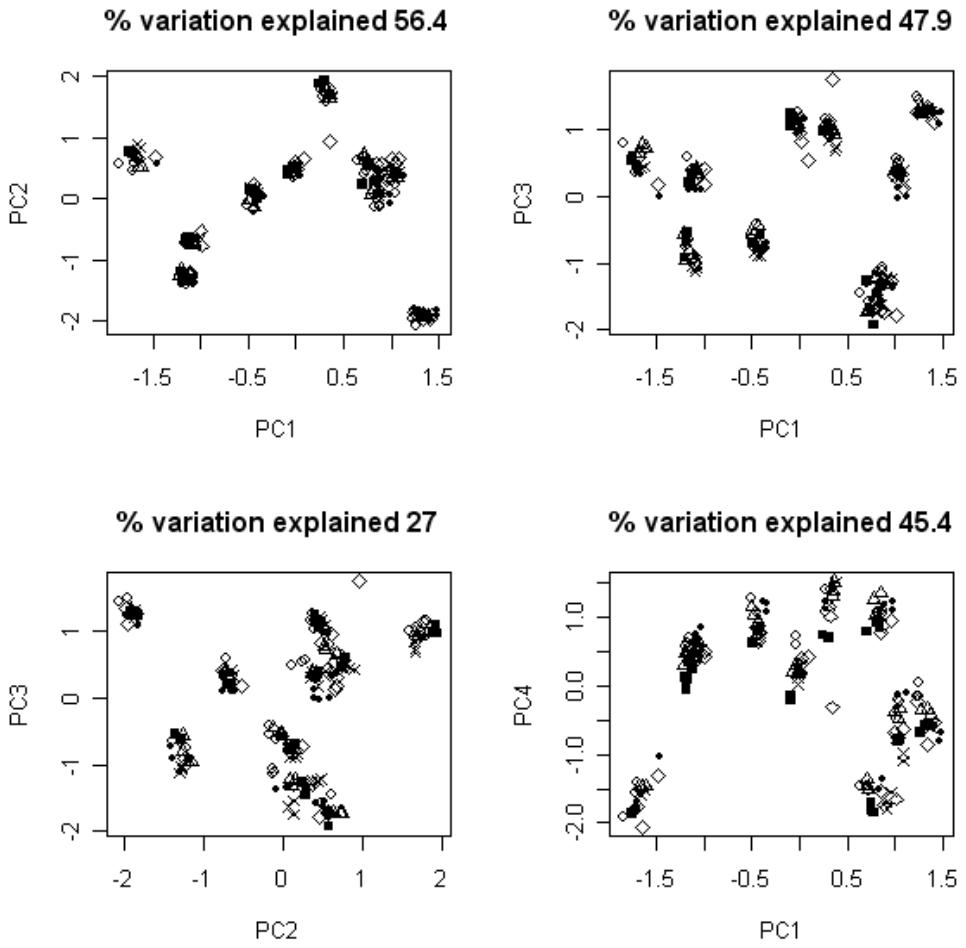


Figure 5.14 - The first few PC scores after correcting for mislabelled points; observers 1-6 are represented by rings, triangles, crosses, diamonds, solid squares and solid circles respectively. The percentage of variation explained by the two plotted PCs is displayed above each plot.

Figure 5.14 shows that there were no longer any obvious differences between the data from different observers, indicated by a random distribution of symbols. There were clearly small clusters of data in the plots; to investigate these points were relabelled numbering them 1-10 to represent the ten faces investigated, Figure 5.15.

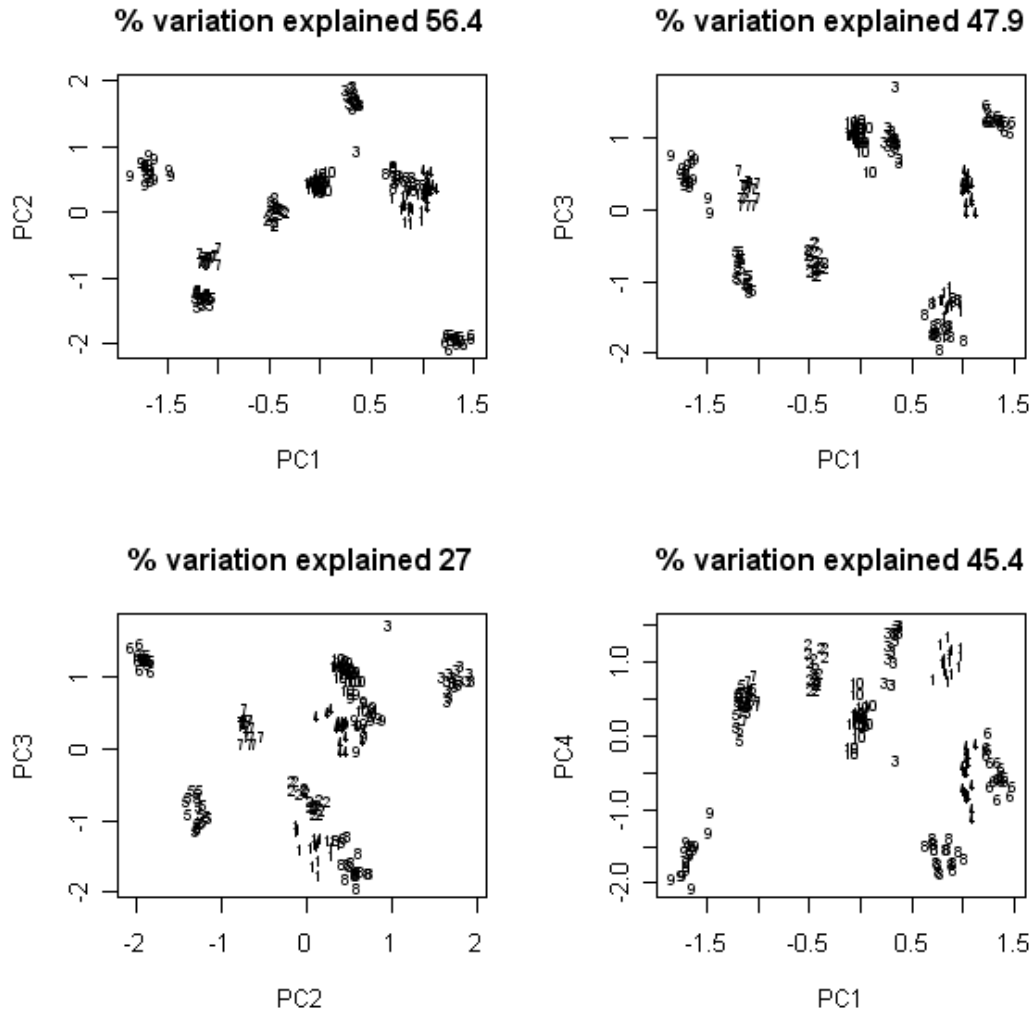


Figure 5.15 – The first few PC scores after correcting for mislabelled points, numbers 1-10 represent the 10 different faces under investigation. The percentage of variation explained by the two plotted PCs is displayed above each plot.

Figure 5.15 shows that the different sets of measurements taken from each of the ten different faces (numbered 1-10) were clustered into distinct groups. There was some overlapping between the clusters in the first two plots with those for faces 1, 8 and 4 positioned close to one another. The third plot, showing the second and third PCs, defined the face clusters the best, showing no overlapping. This was only a very small sample of data and inevitably as more faces are analysed the chances of getting the data so clearly separated will decline. To investigate these groups further the following subsection uses cluster analysis to search for similar groups in the data.

5.3.3 Cluster Analysis - Groups of Similarity

Figure 5.16 shows a dendrogram of the results of a Ward's cluster analysis carried out as an alternative method to search for groups of similar and dissimilar data in the data. Figure 5.17 shows the first branch of Figure 5.16 enlarged, at the lowest level there were many clusters of size three, each containing three observations from the same observer. This means that the different measurements from one observer were more similar to each other than they were to the measurements taken by other observers from the same face.

Figure 5.18 shows the same dendrogram as in Figure 5.16, only it is labelled differently. Labels 1-10 represent the ten faces investigated in the study. Figure 5.19 is an enlargement of Figure 5.18 and shows that all measurements taken on a face (by all observers) were more similar to each other than they were to the measurements taken on the other faces. In other words even though six different observers were used it was still possible to classify the data clearly into the ten different face shapes.

To investigate the structure of the dendrogram in Figures 5.16 and 5.18 the biographic information was sought for each of the faces analysed. All ten subjects were of white British ethnicity, however three faces were female and seven were male. On further investigation of the dendrogram it was found that the highest level cluster separated the sexes into two clusters, however one male face was on the female cluster. The separation of sexes is most likely to be because the scale information was retained in the Generalized Procrustes analysis and, as is seen later in §6.2.2 and §6.2.3, generally females are smaller than males. The male on the female cluster was perhaps a small male.

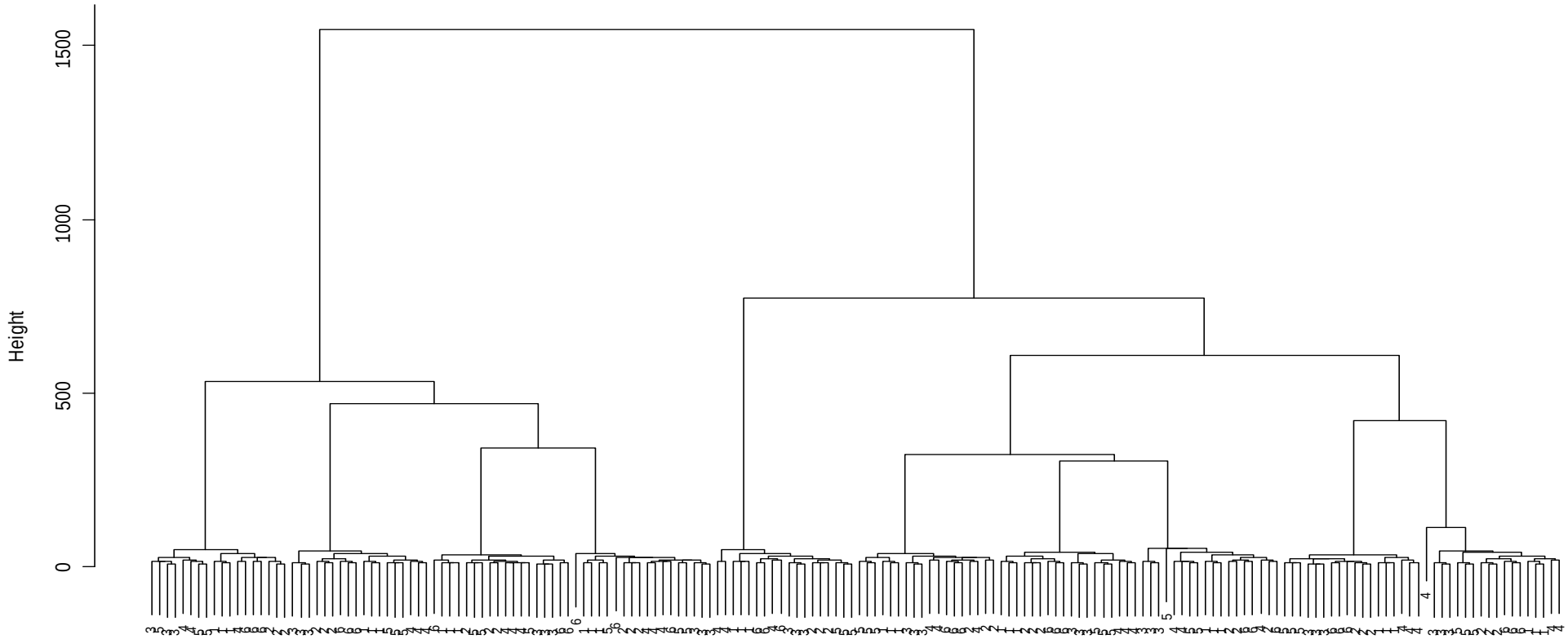


Figure 5.16 – Dendrogram of Wards' cluster analysis, numbers 1-6 represent the different observers. Height refers to the square error of the clusters, which are added to those of their lower clusters

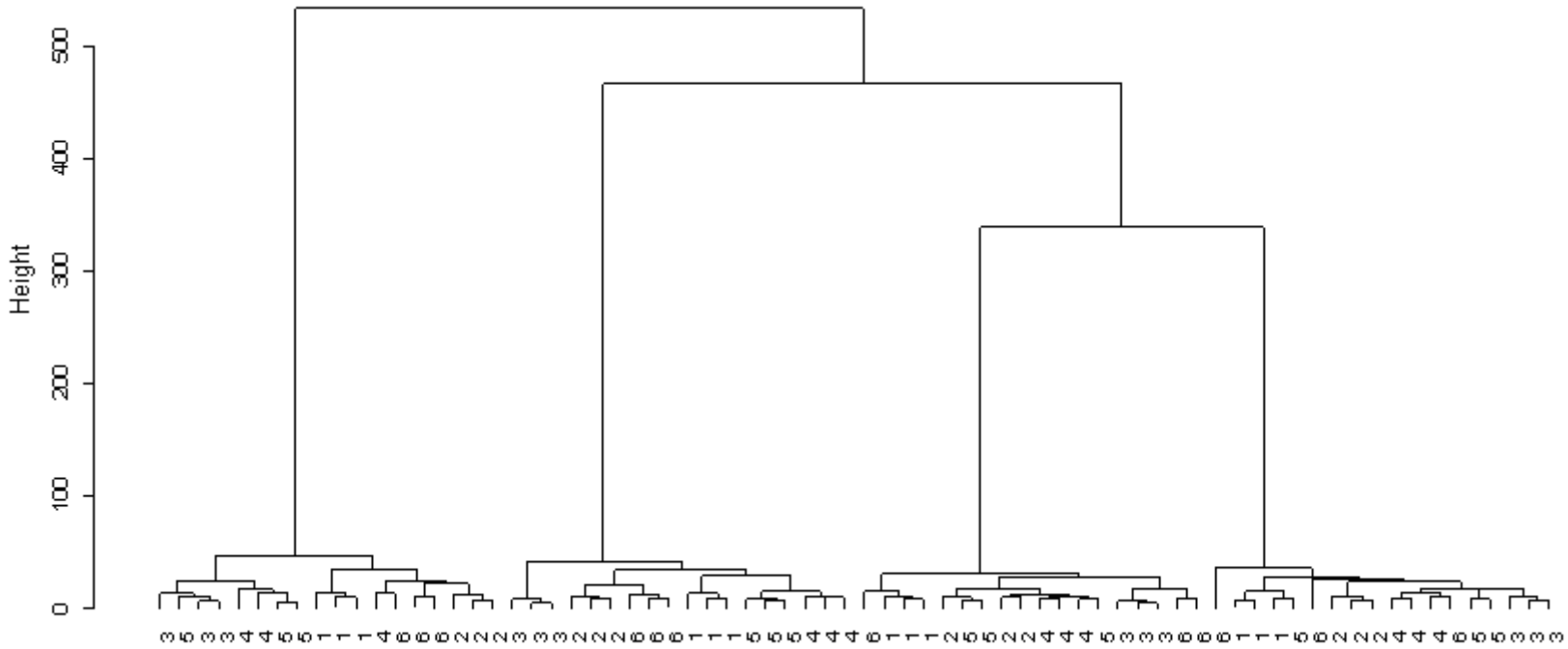


Figure 5.17 - Enlargement of the left hand branch of the dendrogram in Figure 5.16 to examine clusters at the lowest level, numbers represent different observers. Height refers to the square error of the clusters, which are added to those of their lower clusters

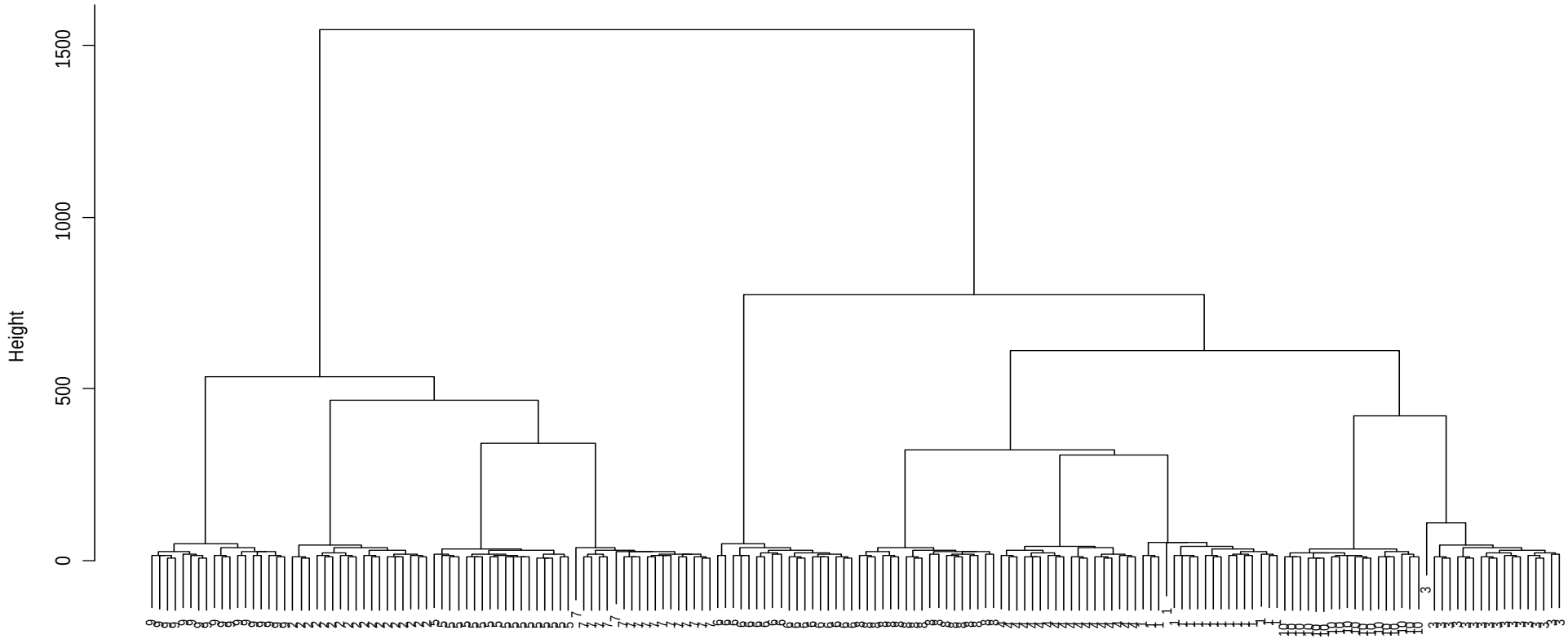


Figure 5.18 - Wards Cluster dendrogram; numbers represent the faces 1-10. Height refers to the square error of the clusters, which are added to those of their lower clusters.

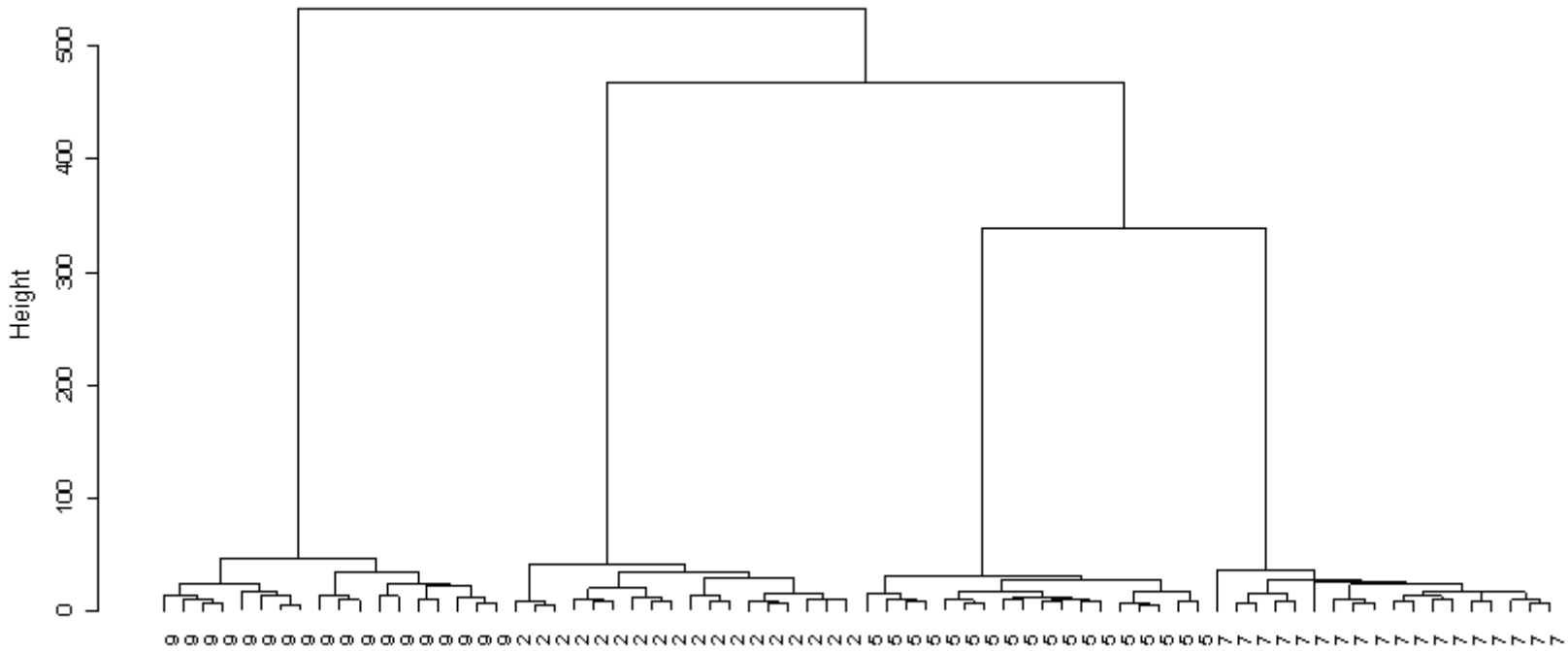


Figure 5.19 - Enlargement of the first branch of the dendrogram in Figure 5.18, numbers represent different faces. Height refers to the square error of the clusters, which are added to those of their lower clusters.

5.4 Summary

This chapter has given full descriptions of preliminary investigations into the facial variation in the available data (§5.2) and demonstrated the repeatability of the data collection technique (§5.3). A set of thirty landmark points has been chosen for collection from the full Geometrix® database (§5.2.6, Table 5.3). These points were thought to be the most appropriate for facial matching. Points were chosen based on the consistency of multiple measures taken by two observers, and also the influence of each point in discriminating between faces (§5.2.5).

When assessing the repeatability of the method some in consistencies in the ordering of the landmark labels for different observers were flagged (§5.3.2). The labels of three landmarks (left and right alare and pronasale) were redefined for all six observers. After this amendment the data from all six observers' was found to be comparable, i.e. observers were producing repeatable configurations. For a small subset of ten faces, a Wards cluster analysis classified the different faces into distinct groups despite multiple observers placing the landmark points (§5.3.3). Further observers employed on the project were asked to collect multiple measurements of the landmark points from the ten faces in this subset. The cluster analysis was then rerun to ensure that new observers were producing configurations which were in line with the other current observers.

Following the results of the work summarized in this chapter the collection of the thirty chosen landmark points for facial matching (§5.2.6) was carried out on the main facial database (§2.4.3). Chapter 6 explores this large facial database looking at differences in size and shape with respect to the age, sex and ethnicity of the subject (§6.2.2). The variation of the individual landmark points are explored through principal components analysis (PCA) of the tangent shape coordinates (§6.2.6, §6.4.1) and a model for the facial shape data is proposed (§6.5). Obtaining the parameters of the data model allows the MVNLR procedure (§3.8.4.1) to be applied to the facial data to quantify likelihoods of facial matches or exclusions, this is carried out in chapters 7 and 8.

6 Facial Variation in the Geometrix® Database

6.1 Introduction

This chapter explores the large sample of faces available to assist in the quantification of facial matches. The data variation in the sample is examined, anomalies in the dataset were uncovered and corrected where possible and a proposed model for the data is given. Such a model provides a measure of the shape variation in the known sample population and enables likelihood estimates of how similar two face shapes are.

The main purpose of the chapter was to investigate the shape variation in the large sample of facial shapes held in the Geometrix® facial database (§2.4). Prior to the development of techniques for facial matching a basic understanding of facial shape and how individual faces varied was necessary. Originally examination of all thirty landmarks points collected (§5.2.6, Table 5.3) in three dimensions was carried out and complete data were examined; i.e. only those faces where all thirty landmark points were collected (§6.2). The data were aligned using generalized Procrustes methods, §3.6, §6.2.1. Two methods of alignment (with and without removing scale) were carried out to explore both size and shape variation (§6.2.4).

Differences in facial variation between different sexes, age groups and ethnicity groups were explored by examining the mean centroid size and standard deviation of the different groups, the Procrustes shape distance and also the root mean squared shape distance between the means of the groups (§6.2.2). Size differences between male and female faces were plotted (§6.2.3). Further investigation of the shape variability was carried out through a principal components analysis (PCA) of the tangent shape coordinates (§3.7, §6.2.4). Plots of the PC scores exposed certain outliers in the dataset, these were explored to reveal some anomalies in the landmark data, the database was cleaned appropriately (§6.2.5).

Further investigation of the principal components (PCs) was carried out, the loadings for each PC were plotted to look at the variability of the individual facial landmark points (§6.2.6). This gave an impression of which landmark points would be good to use for facial matching in terms of the points which varied between different faces.

After reassessment of the requirements for the data and methods proposed (§3.8.4) for facial matching a subset of the database was selected to only include faces for which there were replicated measurements available, (§6.3). Also the landmark points which could not be seen in anterior view were excluded leaving twenty-two potential landmarks for anterior facial matching (§6.4). Analysis of the PC loadings for these twenty-two landmarks showed which points would be good to use for anterior view facial matching (§6.4.1). Chapter 7 uses this information and expands further to select a 'best' set of facial landmarks which optimises the results for matching landmark configurations from anterior images.

Finally the chapter checks that the complete anterior facial data is adequately modelled by a multivariate normal distribution (§6.5) in order to be able to use this as the background data for the MVNLR procedure (§3.8.4) proposed for facial matching. These checks proved to be successful (§6.5) and chapter 7 goes on to apply and extend the MVNLR procedure to carry out anterior facial matching.

6.2 Complete Data

6.2.1 The Data and Procrustes Registration

The general Procrustes methods described in chapter 3 have a requirement that all configurations in the data to be Procrustes registered must be complete, i.e. all landmark coordinates must be present for all configurations. This is because generalized Procrustes matching works by carrying out the superimposition of all n configurations placed in optimal positions by translating, rotating and rescaling each figure so as to minimise the sum of squared Euclidean distances between all k landmark points, Dryden and Mardia (1998). Inspection of the Geometrix® facial landmark database showed that there were 3254 configuration that had 3D locations for all thirty landmark points; these are referred to as the 'complete' data.

Generalized Procrustes analysis (GPA) (§3.6) was carried out to align the complete configurations prior to exploring the structure of facial shape variability. The procedure was carried out both with and without scaling, to see the effect this had on resulting aligned data. It was found that without removing the scale the size of the female face

was clearly smaller than that of the male face, §6.2.3. Removing the scale showed that the underlying facial shapes between the sexes were comparable (§6.2.4).

6.2.2 Differences in Overall Shape and Size

6.2.2.1 Age and Sex Differences

To examine differences in face shape and size between subjects of different sexes and ages the data were grouped into ten year age bands and separate Procrustes alignments (without scaling to compare size information) were carried out for each sex and age group. Table 6.1 lists the number of observations in each age group (n), the mean centroid size of the group (Sbar), standard deviation (sd(S)), the Procrustes shape distance between the means for males and females (Procdist(m,f)) and also the root mean squared shape distance (RMS); subscript _m and _f represent males and females respectively.

Age group	n_m	n_f	Sbar_m	sd(S)_m	Sbar_f	sd(S)_f	Procdist(m,f)	RMS_m	RMS_f
15-24	293	204	402.6	13.3	381.9	10.9	0.015	0.072	0.072
25-34	516	243	413.3	12.1	383.4	12.2	0.022	0.072	0.065
35-44	941	340	415.6	12.1	382.8	11.4	0.021	0.069	0.068
45-55	315	107	415.7	12.3	386.3	12.3	0.023	0.071	0.069
55-65	153	57	418.2	12.5	388.2	8.4	0.021	0.073	0.065
65-74	69	12	416.3	12.1	395.5	11.6	0.031	0.086	0.064
75+	18	0	413.1	14.1	-	-	-	0.073	-

Table 6.1 – Summary of size and shape by sex and age group: number of observations in each age group (n), the mean centroid size of the group (Sbar), standard deviation (sd(S)), the Procrustes shape distance between the means for males and females (Procdist(m,f)) and also the root mean squared shape distance (RMS); subscript _m and _f represent males and females respectively.

Table 6.1 shows that the Procrustes distance between the mean shapes of males and females was the least in the youngest (15-24) age group, suggesting that perhaps the feminine or masculine shape may not have fully developed for this age group. The Procrustes distance was fairly consistent (around 0.02 in standardised units) across age groups 25-65, suggesting age was not a factor attributable to the mean shape difference between males and females between twenty-five and sixty-five. After age sixty-five the Procrustes distance increases, however the number of females in this age group was

only twelve compared to sixty-nine males. This indicates that male and female faces become increasingly different with age after sixty-five, though with the small numbers of females this is only an indication. The root mean squared distances indicated little difference in overall shape variability between the sexes in first two age groups; however this showed a steady increase in differences with age between the sexes.

Figure 6.1 shows the mean centroid size across the different age groups for both sexes. It shows that male face size was always larger than female regardless of what age the subjects are, this phenomenon is known as sexual dimorphism and is well known in anthropology. There was some evidence for a trend toward increasing mean facial size with age, this was slight, more obvious for females and the last two age groups for males appeared to deviate from the trend. Looking at Table 6.1 the standard deviations in size were fairly consistent across all sex and age groups, although females in age groups 15-24 and 55-65 showed a little less size variability. When adding or subtracting twice the standard deviation (2*sd(S)) from the mean centroid size the estimates for male and female size overlap, so it is unlikely that observed differences are significant.

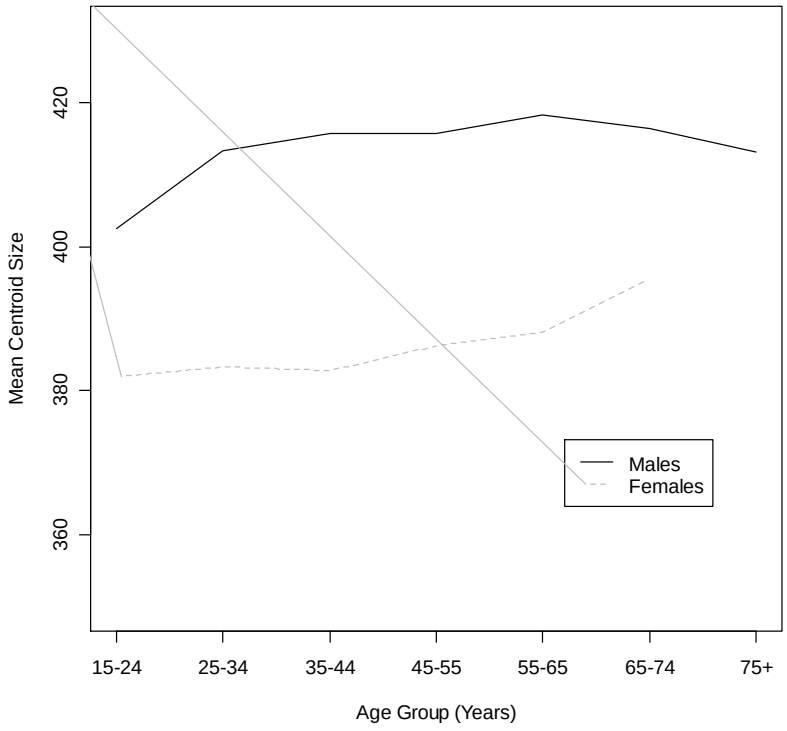


Figure 6.1 - Mean centroid size of faces in different age groups, males and female

6.2.2.2 Ethnicity and Sex Differences

Table 6.2 compares face shape and size for the fifteen different ethnic groups (defined previously in Table 2.1), again males and females were examined separately. It should be noted that the number of observations in ethnic group 1 (white British) far exceeded the numbers in all other groups, so observations made are unreliable to assume for the whole population. Also, the subject of putting oneself into one of fifteen ethnic groups is a very individual thing, what a person feels is their actual ancestry may not fit into any of the groups hence they choose a 'best' estimate, so the ethnicity data is not very reliable.

Table 6.2 shows that ethnic group 4 (white and black African) had the smallest mean centroid size for males and the largest mean centroid size for females. The Procrustes shape distance between the means for males and females was the smallest for white British and the largest for ethnic group 13 (other black background). There were no females in the dataset of ethnic group 11 (Caribbean).

Ethnic group	n_m	n_f	Sbar_m	sd(S)_m	Sbar_f	sd(S)_f	Procdist(m,f)	RMS_m	RMS_f
1	2086	866	413.7	12.9	383.6	11.7	0.019	0.072	0.069
2	112	46	414.5	14.0	384.1	10.1	0.028	0.074	0.064
3	2	4	412.9	1.4	396.0	18.1	0.061	0.020	0.055
4	4	2	399.6	2.0	401.8	0.2	0.056	0.058	0.028
5	8	2	413.6	18.0	382.1	1.1	0.067	0.060	0.019
6	2	4	409.3	0.6	396.0	1.7	0.068	0.025	0.060
7	20	9	404.0	10.7	378.7	8.4	0.034	0.068	0.070
8	12	5	405.3	10.1	373.0	6.5	0.038	0.064	0.054
10	15	3	403.6	15.0	381.3	8.3	0.054	0.066	0.047
11	11	0	423.1	13.9	-	-	-	0.091	-
12	9	4	421.5	14.1	390.3	5.2	0.045	0.065	0.062
13	2	3	409.3	1.0	381.1	14.8	0.118	0.023	0.044
14	19	11	412.2	15.1	380.9	13.6	0.041	0.088	0.059
15	3	19	421.1	9.7	412.2	15.1	0.072	0.032	0.088

Table 6.2 - Differences in size and shape between sex and ethnic groups: number of observations in each ethnic group (n), the mean centroid size of the group (Sbar), standard deviation (sd(S)), the Procrustes shape distance between the means for males and females (Procdist(m,f)) and also the root mean squared shape distance (RMS); subscript _m and _f represent males and females respectively.

6.2.3 Size Differences between Males and Females

§6.2.2 established that there were differences in size between male and female faces. To see which specific aspects of the face were different in size plots comparing the mean face shapes for each sex were drawn for the three orthogonal views of the Procrustes aligned data preserving the scale, Figures 6.2 and 6.3.

Figures 6.2 and 6.3 showed that landmark points located in the centre of the face and around the eye and nose areas were the least variable between the sexes. The largest differences were in the ears, lips and chin areas. Females had smaller ears than males, which were also located closer to the centre of the face, thus indicating an overall smaller facial width than the males. Also, the female landmarks around the chin and lips were closer to the centre of the face than the males, indicating an overall shorter facial length than the males.

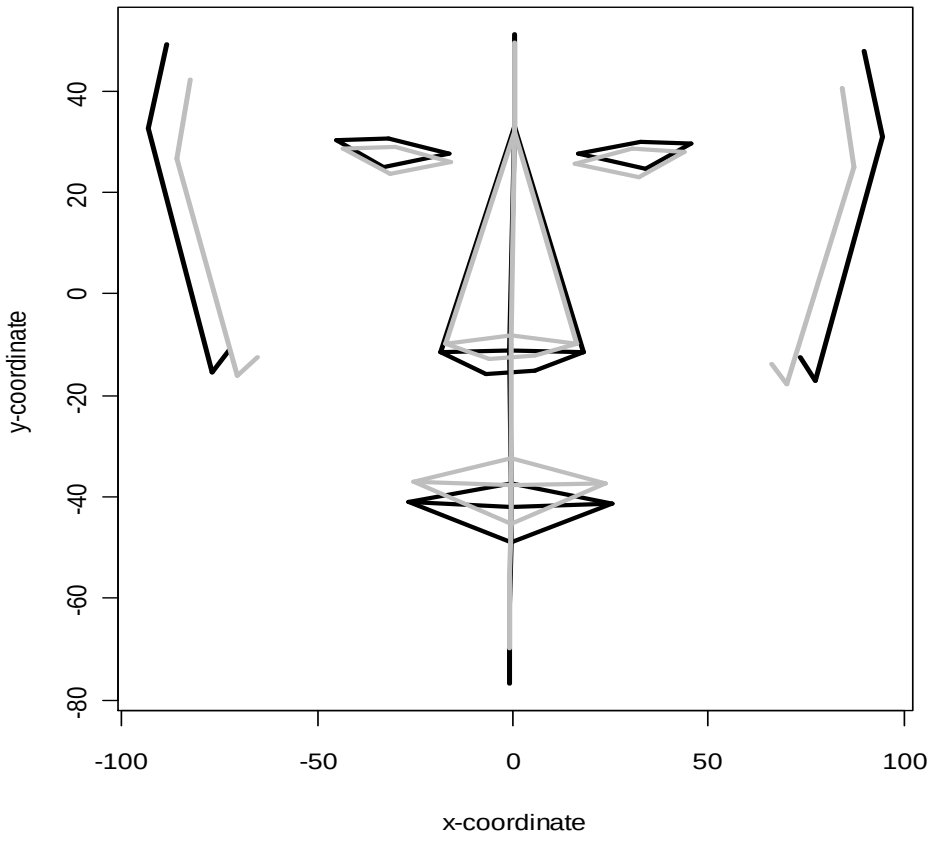
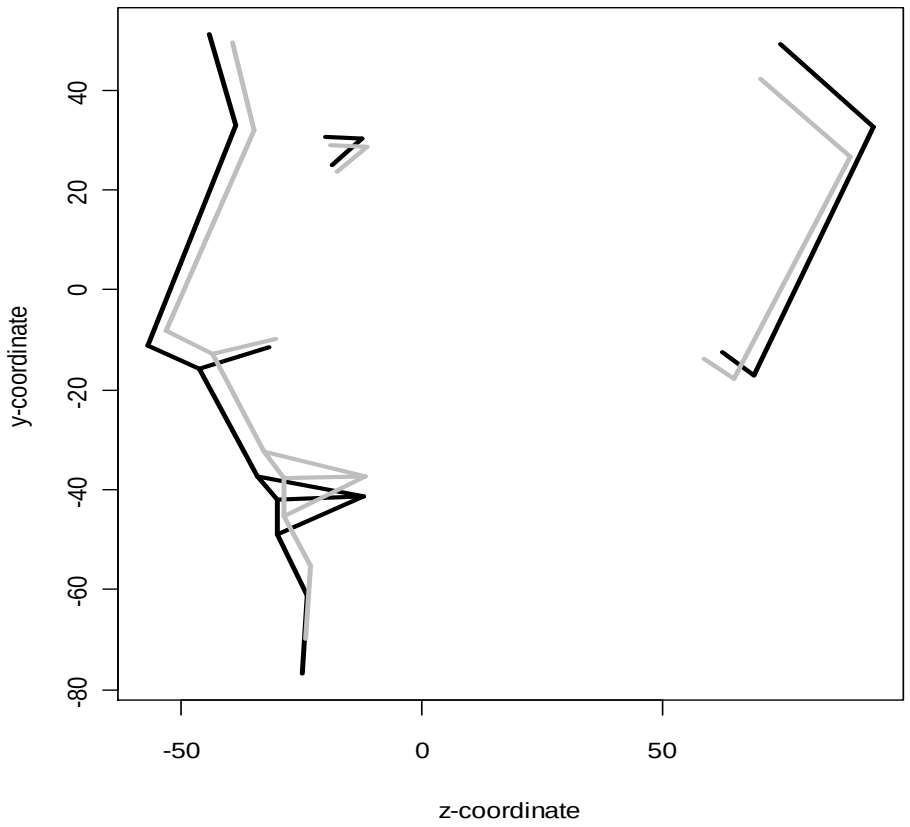


Figure 6.2 - Mean face shape for males (black) and females (grey) for the Procrustes aligned data preserving scale.


y-coordinate
40
20
0
-20
-40
-60
-80
-50
0
50
z-coordinate

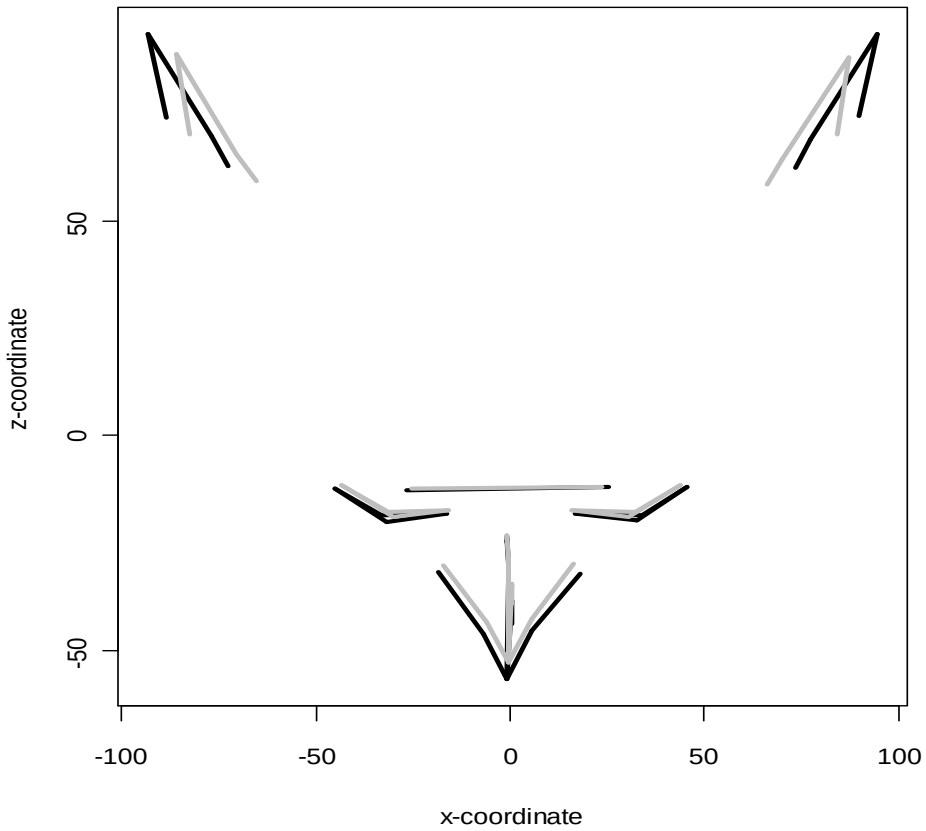


Figure 6.3 – Mean face shape for males (black) and females (grey) for the Procrustes aligned data preserving scale

6.2.4 PCA - Facial Shape Variability

After a Procrustes alignment preserving the scale information a PCA of the sample covariance matrix in Procrustes tangent space was carried out to analyse the main modes of shape variation, §3.7. Figure 6.4 is a scree plot of the cumulative sum of the percentage of data variation that was explained by the PCs. Figure 6.4 depicts a smooth curve and does not show a definite kink or cut-off point to indicate an efficient number of PCs that would explain the majority of the variation in the data. About 90% of the data variation was explained by the first twenty-five PCs, thus the dimensionality of the problem could be reduced greatly whilst keeping the majority of important information from the data.

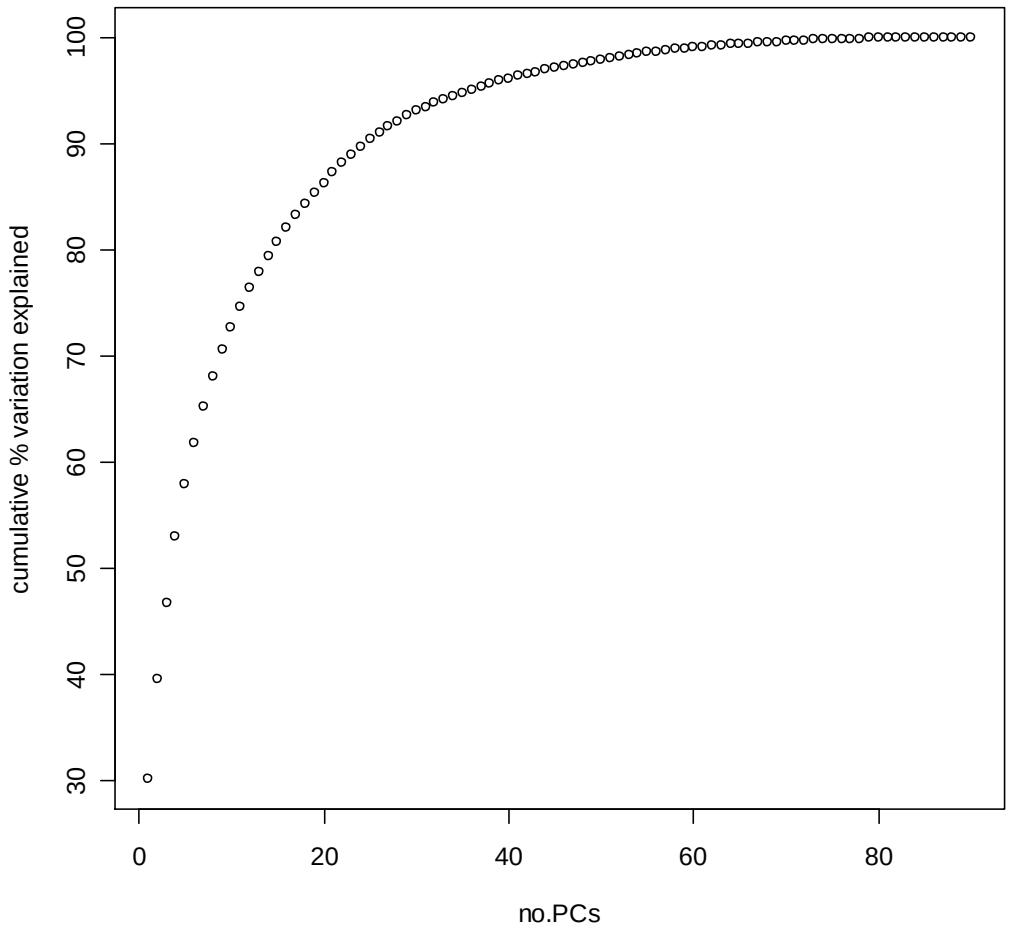


Figure 6.4 – Scree plot to show cumulative amount of variation explain by PCs of Procrustes aligned data (preserving size); 30 landmark points in 3D (i.e. 90 variables)

Plotting the scores for the PCs can show valuable information about the structure of variability in the data, such as groups of similarity. Figure 6.5 shows plots for the first few PCs scores for the Procrustes tangent coordinates from the GPA without the removal of scale.

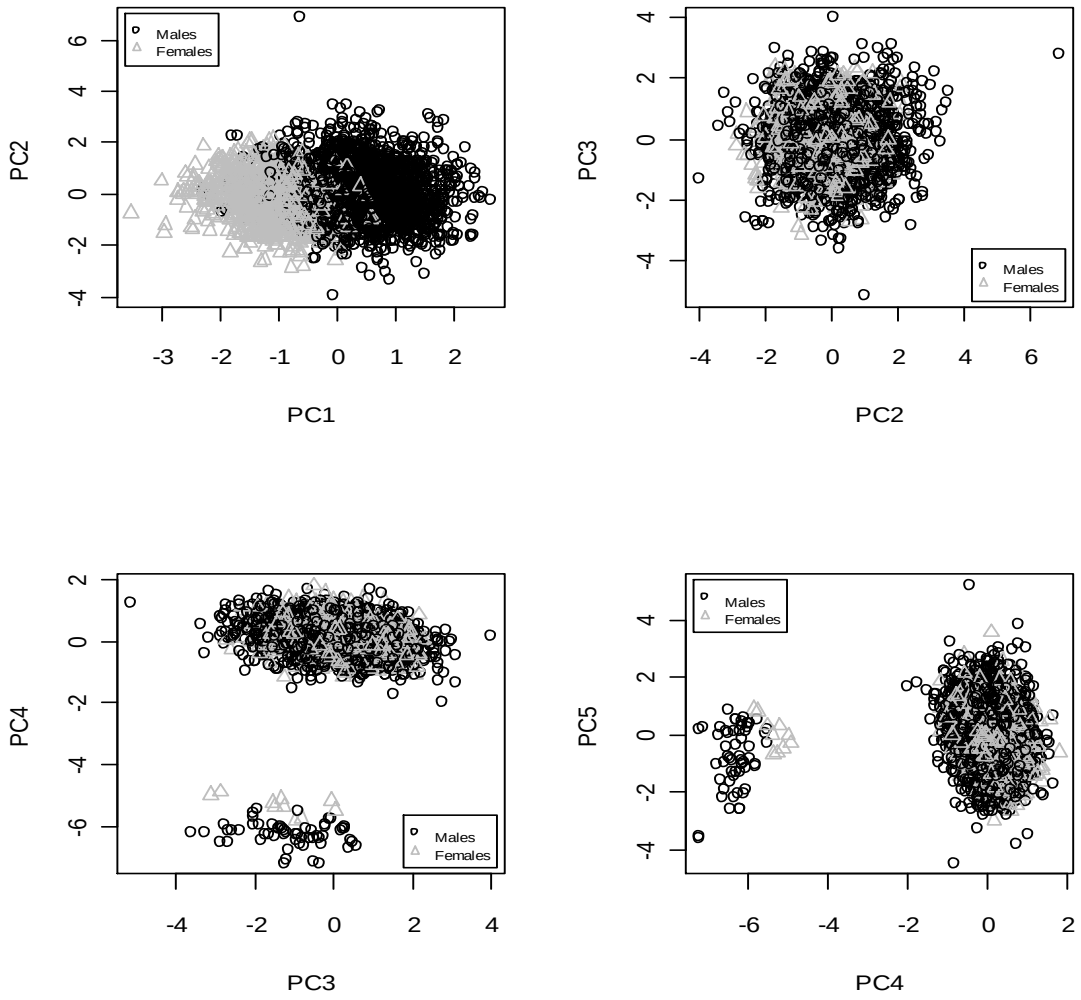


Figure 6.5 – First few PC score plots for Procrustes aligned (preserving scale) complete configurations, males and females are represented by circles and triangles respectively.

Figure 6.5 shows that on PC 1 differences between the sexes were evident when the size component was preserved during the Procrustes alignment. Although the two groups had some overlap a clear distinction between the sexes was seen. Also seen in the last two plots in Figure 6.5 is a secondary cluster of points containing both males and females, these subjects are evidently outliers to the majority of data and require further investigation.

A different Procrustes alignment was carried out on the complete data, this time removing the scale information. A PCA was performed on the resulting tangent coordinates and plots of the first few PC scores are displayed in Figure 6.6.

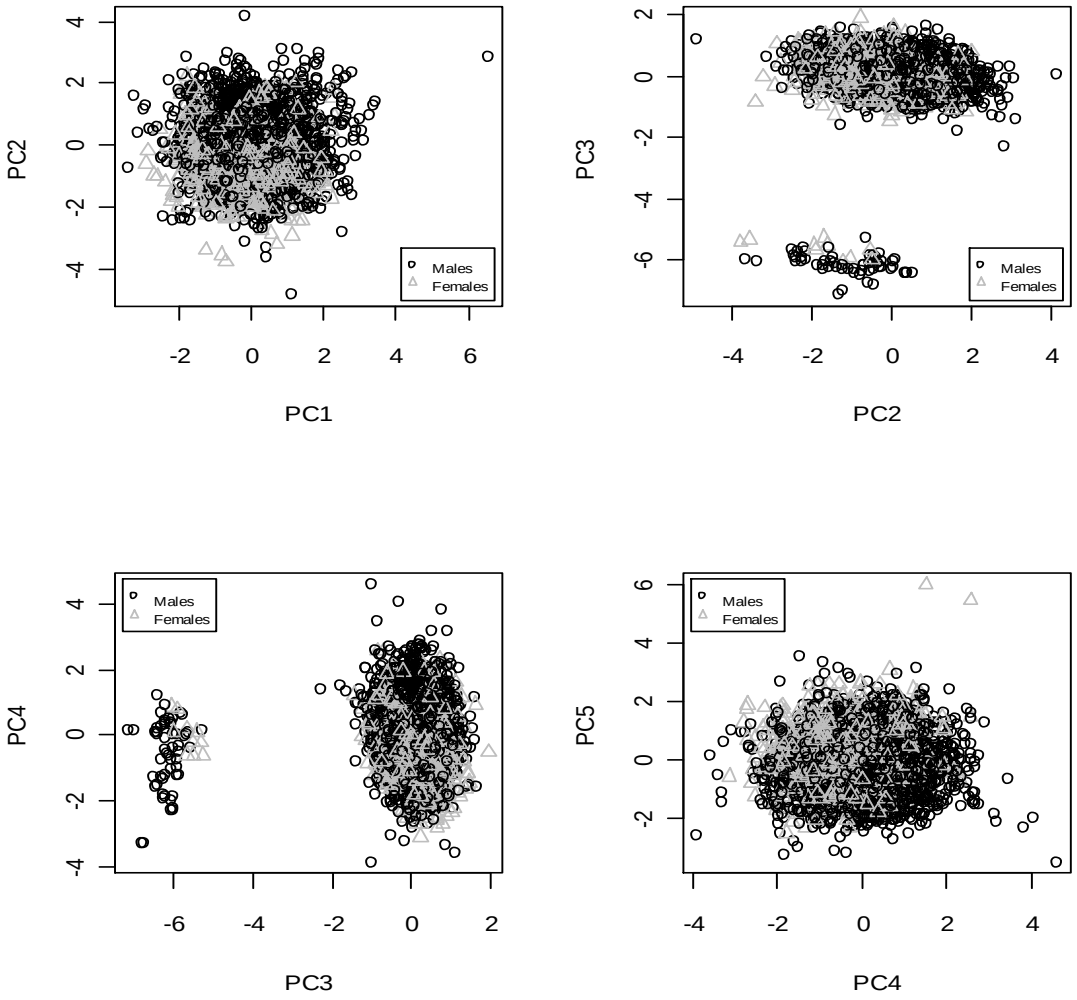


Figure 6.6 - PC score plots for Procrustes aligned complete data including the removal of scale; circles represent males, triangles females.

Figure 6.6 shows that the distinction of the two groups for males and females that was apparent on PCs 1 and 2 in the first PC score plots (Figure 6.5) was no longer present when the scale was removed in the Procrustes alignment. This suggests that when the scale was preserved during the GPA the first two PCs represent the size or scale of the face. Figure 6.6 indicates that when scale is removed the underlying shapes of male and female faces were comparable showing no discrete groupings on the PCs. Figure 6.6 shows that there was a still a secondary cluster of points containing data of both sexes,

this cluster was evidently separate to the majority of data and these observations must differ in facial shape (and not just size) from the majority. It was thought that perhaps the secondary cluster represented data that were of a different ethnic origin to white British, this is investigated further in §6.2.5.

These analyses suggest that for facial comparisons a Procrustes alignment removing scale should be carried out on the data prior to any comparisons, to remove differences attributed to size between males and females. When thinking about facial comparison data the images for analysis will probably be of unknown scale, e.g. images taken from CCTV footage where the distance of the subject from the camera is unknown, and so it will be necessary in that respect to remove the scale information.

6.2.5 Examination of Outliers and Data Cleaning

The PC score plots in §6.2.4 picked up on a group of points that seemed to lie far away from the majority of data. Investigations into the collected demographic data (§2.4) for each of these faces showed that the observations in question were of various sex, age and ethnic groups and no particular group dominated the outliers. It was thought that the subjects could have some unusual facial features or characteristics, which could be the reason they are outliers. Such features would be beneficial to the application of facial comparison, as we would expect unusual faces to only match with faces that are equally unusual, so false positive results would be low. To investigate the facial landmarks the original source data (facial images with landmark points displayed) for the outlying group were examined for data inconsistencies.

The facial images were opened within the Forensic Analyzer® program and the landmark points were checked. It was found that for several images a couple of labels had been switched, i.e. the landmarks were in a correct landmark position with the incorrect label; Figure 6.7 shows an example of this where landmarks 3 and 13 have been swapped. Other faces showed that one or several landmark points were in an incorrect landmark position and so had been misplaced; see Figure 6.8 for an example where landmark 26 has been incorrectly positioned.

The mislabelled points were corrected; however configurations where points were misplaced were excluded from the database, as the original observer who located the points was not available to correct the position. This method of data checking and cleaning was also applied when additional data for facial comparisons (§2.7.2) was added to the main facial database (§2.4) to ensure that the model was still a reasonable fit (§7.2.2.2).

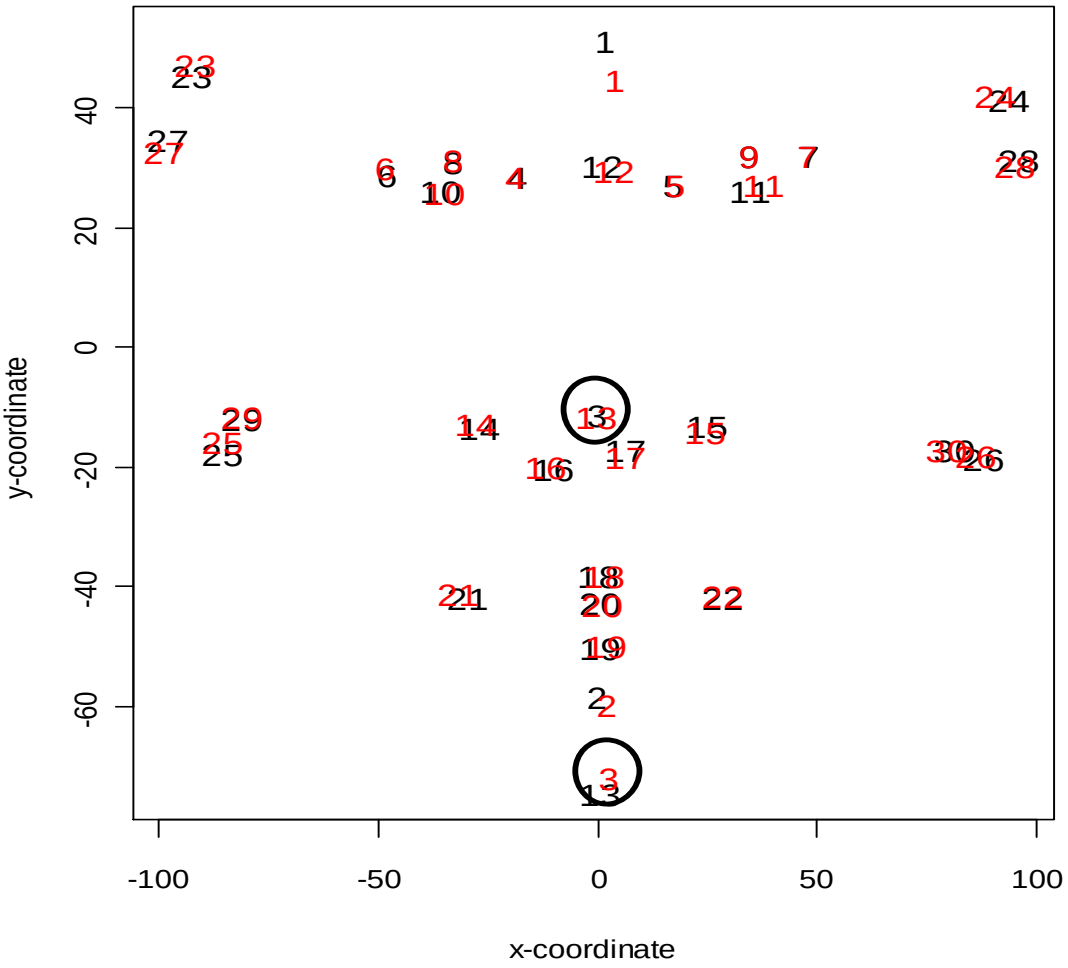


Figure 6.7 – Procrustes rotated x and y coordinates of duplicated landmarks (observation number 29 in black and 30 in grey) from one face. Observation 29 has mislabelled landmarks 3 and 13; this can be corrected by swapping over the landmark labels for this configuration

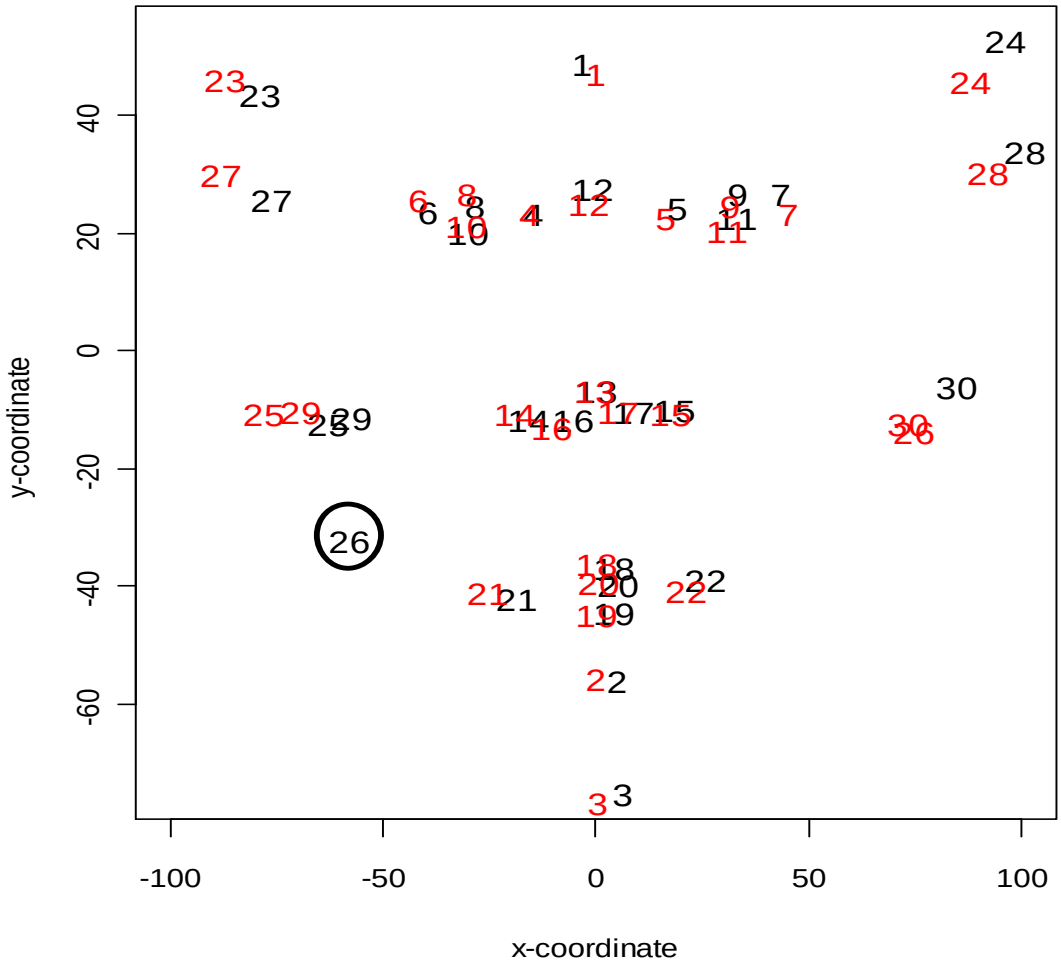


Figure 6.8 - Procrustes rotated x and y coordinates of duplicated landmarks (observation number 2052 in black and 2053 in grey) from one face. Landmark 26 has been misplaced for observation 2052, this cannot be corrected for and so the configuration must be excluded from further analyses

6.2.6 Variability of Individual Facial Landmarks

Looking at the Procrustes aligned data removing scale, comparing only the underlying shape of faces (not shape and size); a further exploration of variation was carried out on the loadings of the PCs of the facial landmarks. As with multivariate data inspection of the loadings for each PC can be carried out for shape analysis. There is an added advantage with the shape data; in that the geometrical properties of the configurations are preserved (up to similarity transforms) therefore visualizing a typical face shape (e.g. the mean face) with superimposed PC loadings can help to directly interpret the effect of each PC on the parts of the face.

Plots to show the loadings for the first four PCs are displayed in Figure 6.9 (anterior XY orthogonal view) and Figure 6.10 (profile ZY and overhead XZ orthogonal views). Landmarks of the mean face from the sample are plotted points joined by grey lines to show the orientation of the face; black arrows are drawn from plotted points to indicate the direction and magnitude of the PC loadings. Loadings were scaled to show the differences better, those greater than five were indicated by a black solid arrow, smaller than five by a grey dashed arrow. Shown below each plot is the percentage of variation represented by the PC, the total amount of variation explained by the first four PCs was 39%.

Figures 6.9 and 6.10 showed that the most variable points on PCs 1 and 2 were the landmarks around the ears; it could be that the ears are more unique to individuals or that the scanner or software for collecting the landmark points was not as accurate in the profile views. PC3 showed that landmarks around the nose, the pronasale and the alares (left and right) were the most variable points. PC4 had points in the chin area showing the most variation.

In terms of facial comparisons the facial areas used to quantify matches will need to show enough variation to be able to distinguish between faces. So the areas found to vary in the PC loadings plots should be ones which would be useful in terms of facial matching. Unfortunately the ear landmarks are going to be difficult to locate, as data for facial comparisons will be in the form of 2D views of the face, e.g. from CCTV, and so it is unlikely that the ears will be visible.

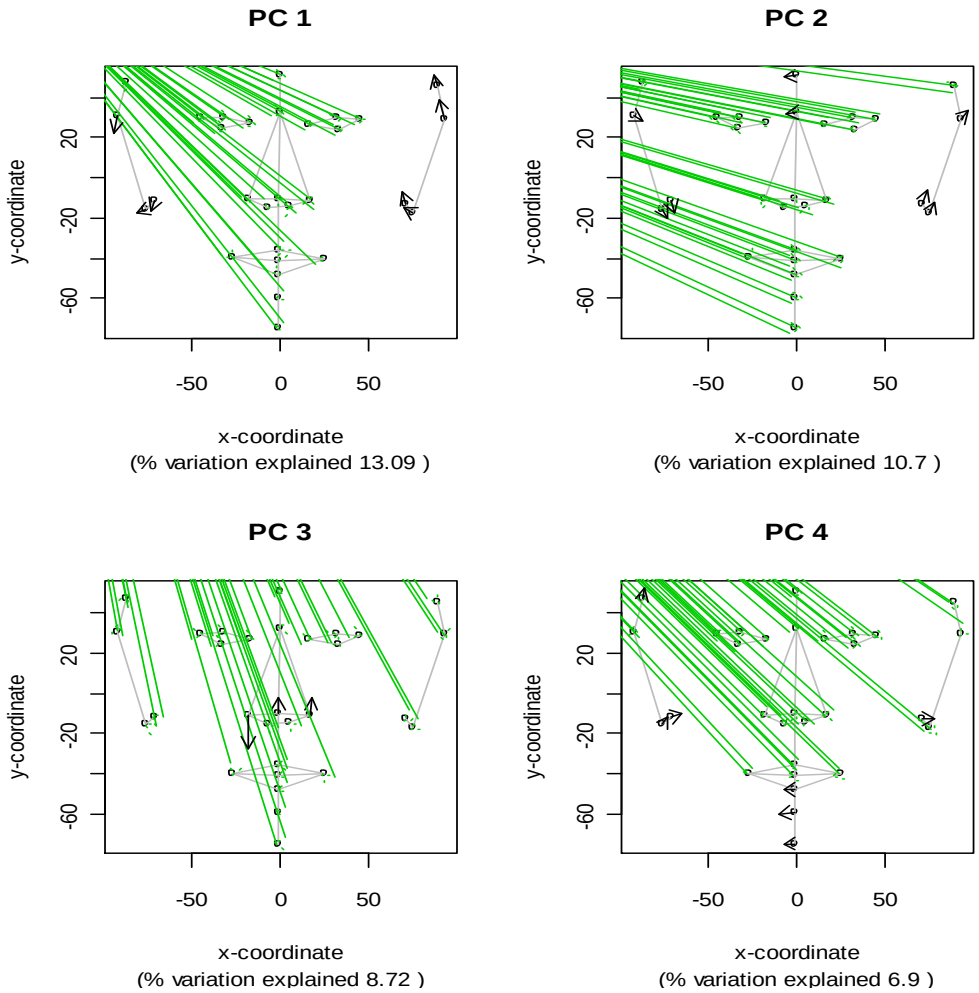
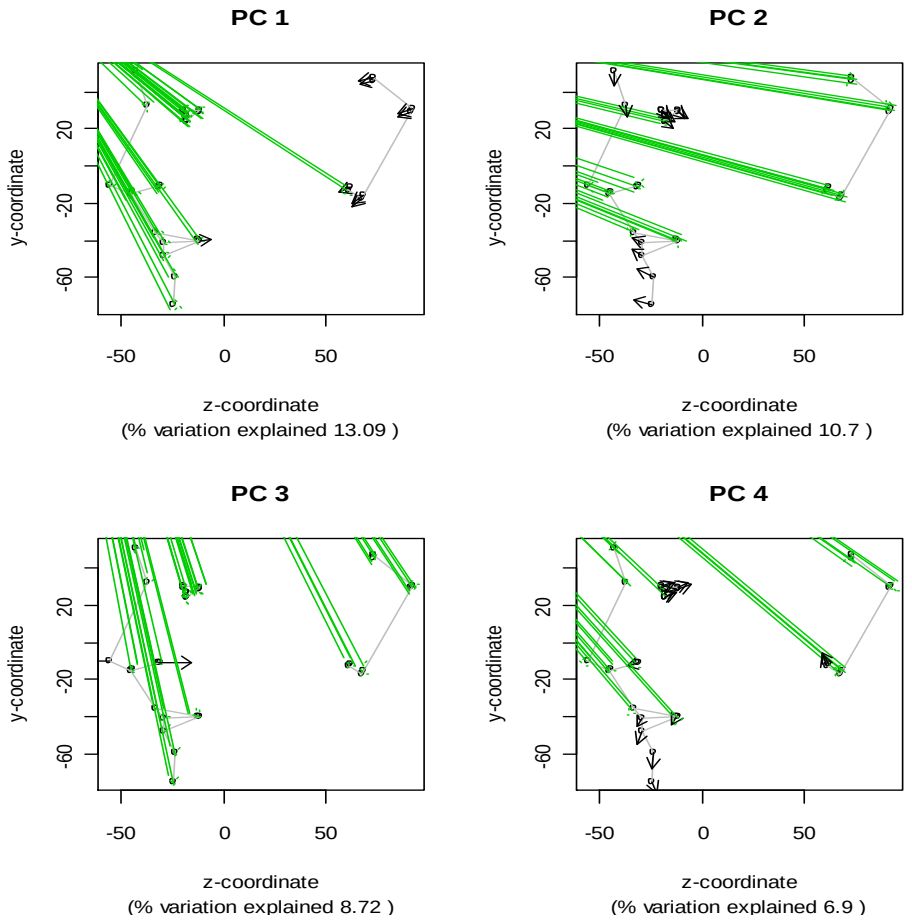


Figure 6.9 –Loadings for first four PCs for the XY anterior facial view, points represent the mean face shape in the aligned data with solid vectors indicating the direction and magnitude of the loadings.


PC 1
y-coordinate
20
-20
-60
-50
0
50
z-coordinate
(% variation explained 13.09)
PC 2
y-coordinate
20
-20
-60
-50
0
50
z-coordinate
(% variation explained 10.7)
PC 3
y-coordinate
20
-20
-60
-50
0
50
z-coordinate
(% variation explained 8.72)
PC 4
y-coordinate
20
-20
-60
-50
0
50
z-coordinate
(% variation explained 6.9)

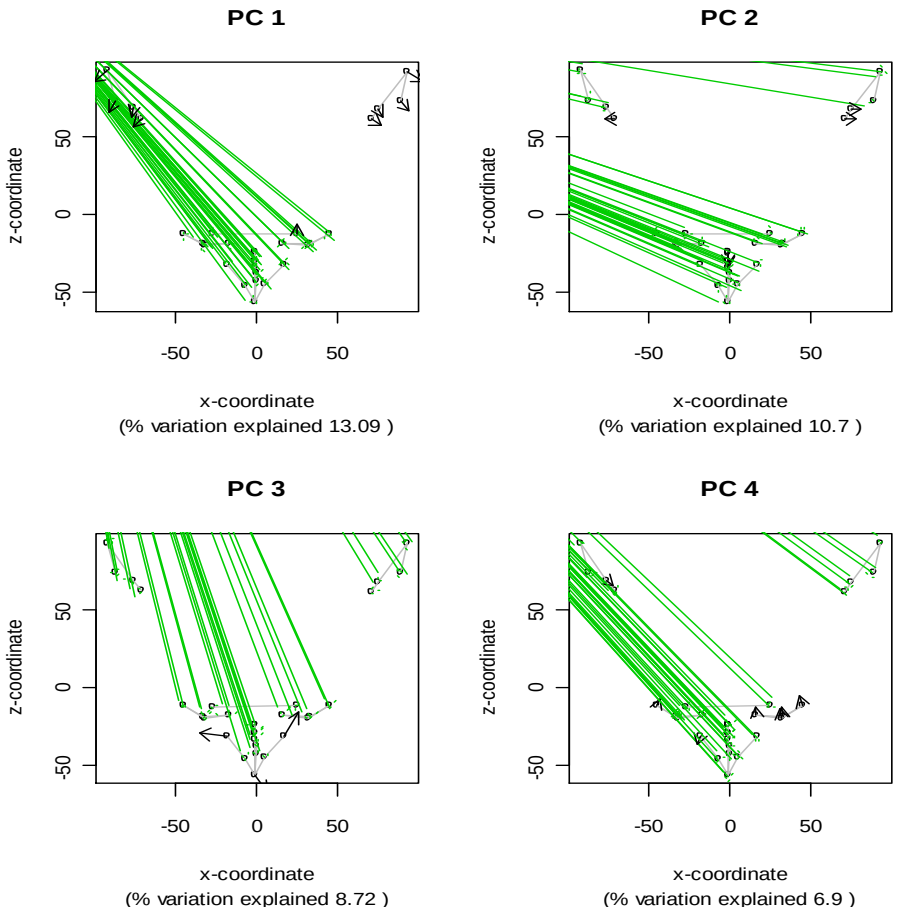


Figure 6.10 - Loadings for first four PCs for the ZY and XZ facial views, points represent the mean face shape in the aligned data with solid vectors indicating the direction and magnitude of the loadings.

6.3 Complete Data with Replicates

To gain good estimates of probabilities for facial matches all possible sources of variation should be included in a model from which to deduce these probabilities. For the Geometrix® facial database there were $r = 2$ duplicate measures taken on each face, by either the same or different observers. This data permits the assessment of intra-facial variability, which can be carried out by examining the differences between the duplicate measurements taken of the same face. There were 1286 faces for which replicated measurements were available for the full set of thirty landmarks, i.e. 2572 configurations. This number is considerably less than that of all complete configurations, which was 3254; this suggests that the observer carrying out the first capture of measurements did not always agree with the observer obtaining the second set of measurements. This subjectivity in itself is important, as before the data can even be analysed differences in opinion between observers are affecting results (i.e. how many faces can be used in the comparison analyses). The data were Procrustes aligned, so it also had to be complete; therefore this subset of the landmark database will be referred to as the 'complete replicated' data.

6.4 The Anterior Facial View for 2D Facial Matching

The real life application of facial matching would involve 2D images, these would probably be best in the anterior (subject forward facing) facial view. This view is when the majority of landmark points are visible and easily located, with the exception of the ear landmarks all the other thirty points chosen in chapter 5 (Table 5.3) for data collection are visible in the anterior view. When carrying out a Procrustes alignment of 2D anterior data it is necessary to remove the factor of scale. This is because when an image is obtained from an external source (other than the Geometrix® database, which was acquired under controlled conditions) the distance of subject to camera is unknown, i.e. the scale is unknown.

§6.2.6 showed that the landmarks with the greatest variability were the points around the ears, which as previously discussed are not easily placed in an anterior view. Therefore the ear landmarks were excluded to look at the variability of the remaining points for the anterior view. A Procrustes alignment (removing scale) was carried out on

the remaining k = 22 anterior landmarks for the n = 2572 complete replicated configurations, a PCA was carried out on the resulting tangent coordinates.

6.4.1 Variability of Anterior Facial Landmarks

The loadings for the first twelve PCs are displayed in Figures 6.11, 6.12 and 6.13; the plots show the anterior (x, y) landmarks of the mean face from the complete replicated data joined by grey lines to show the orientation of the face, black arrows are drawn from the plotted points to indicate the direction and magnitude of the PC loading scores. Loadings were scaled to show differences better, landmarks with loadings greater than 0.05 are indicated by a black solid vector, smaller than 0.05 are a grey dashed vector. Also shown below each plot is the percentage of variation explained by the PC. The total amount of variation explained by these first twelve PCs is 85%.

Table 6.3 summarizes in words the parts of the face which varied the most for each of the PCs in the anterior analysis. PC 1 accounted for 26% of the data variation and the three points around the tip of the nose were the landmarks showing the most variation. The landmarks that occur in this area are difficult to place when the images are in the anterior view. For instance the tip of the nose is extremely subjective to place when looking from in front of the subject. Even when looking from the profile view there is nothing to indicate whether the point will lie in the centre of the nose if the subject turned to face the camera. A combination of the two (anterior and profile) views best locates these landmarks; this was not carried out here and could explain why these landmarks showed high variation. These landmarks will also be affected by the situation of the subject head, i.e. tilting or turning of the head away from the central forward facing position. PC 2 showed that 16% of the facial variation occurred mainly around the outer eyes. Additional areas of the face that showed variability were regions which are affected by facial expression; the mouth, chin and outer eye regions all have more mobility than other facial areas if a subject smiles.

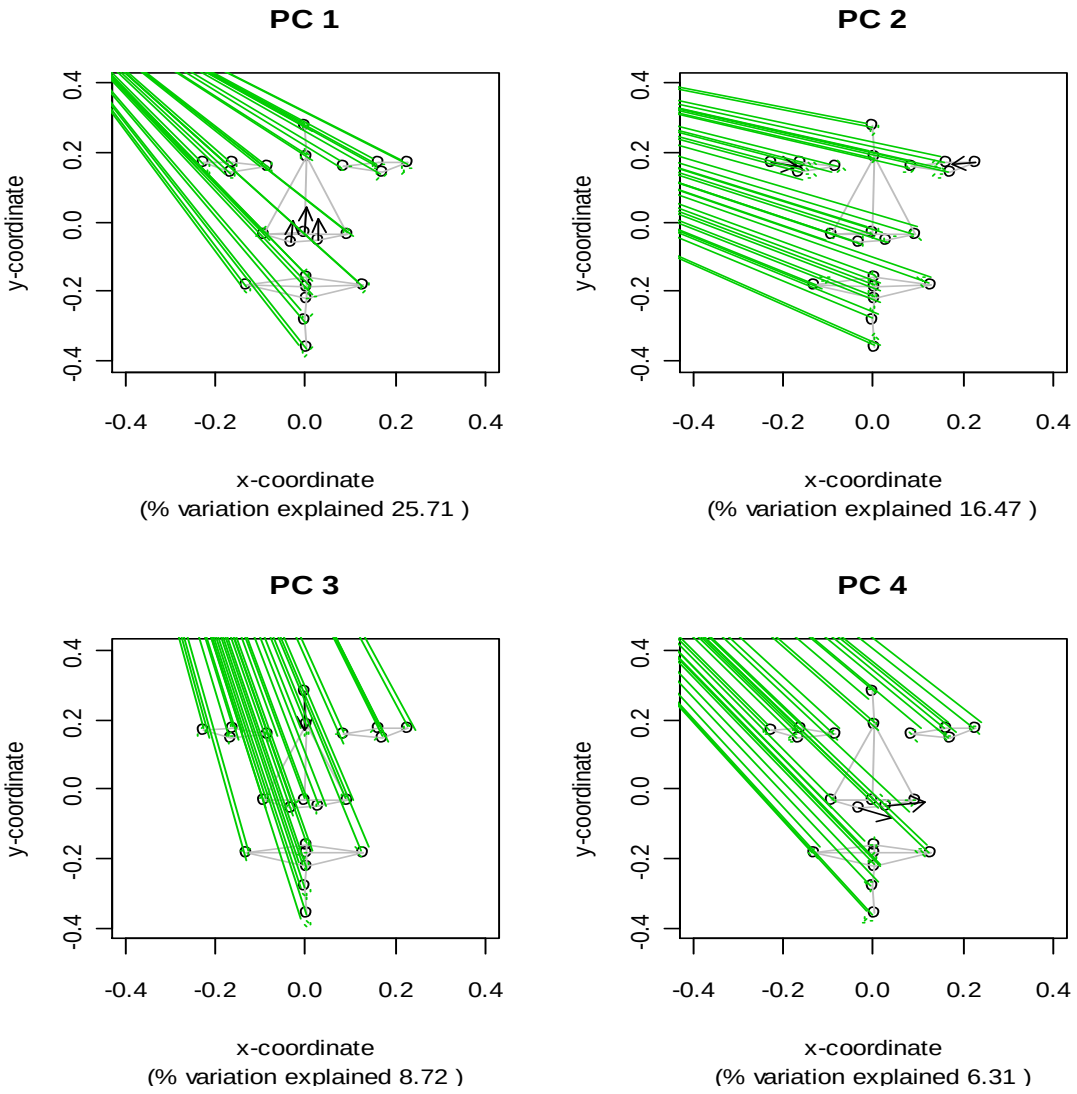


Figure 6.11- PC plots (PCs 1-4) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows.

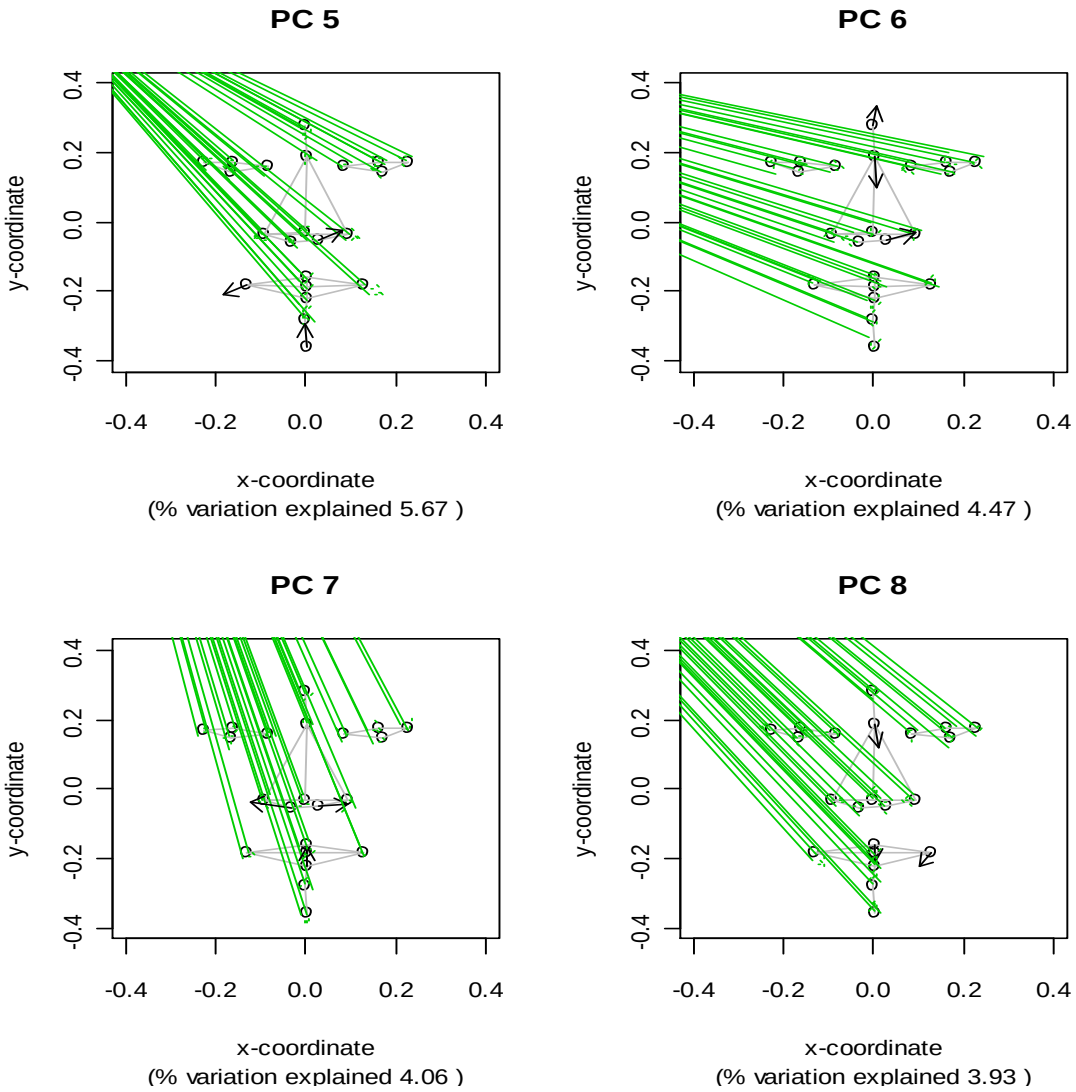


Figure 6.12 –PC plots (PCs 5-8) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows.

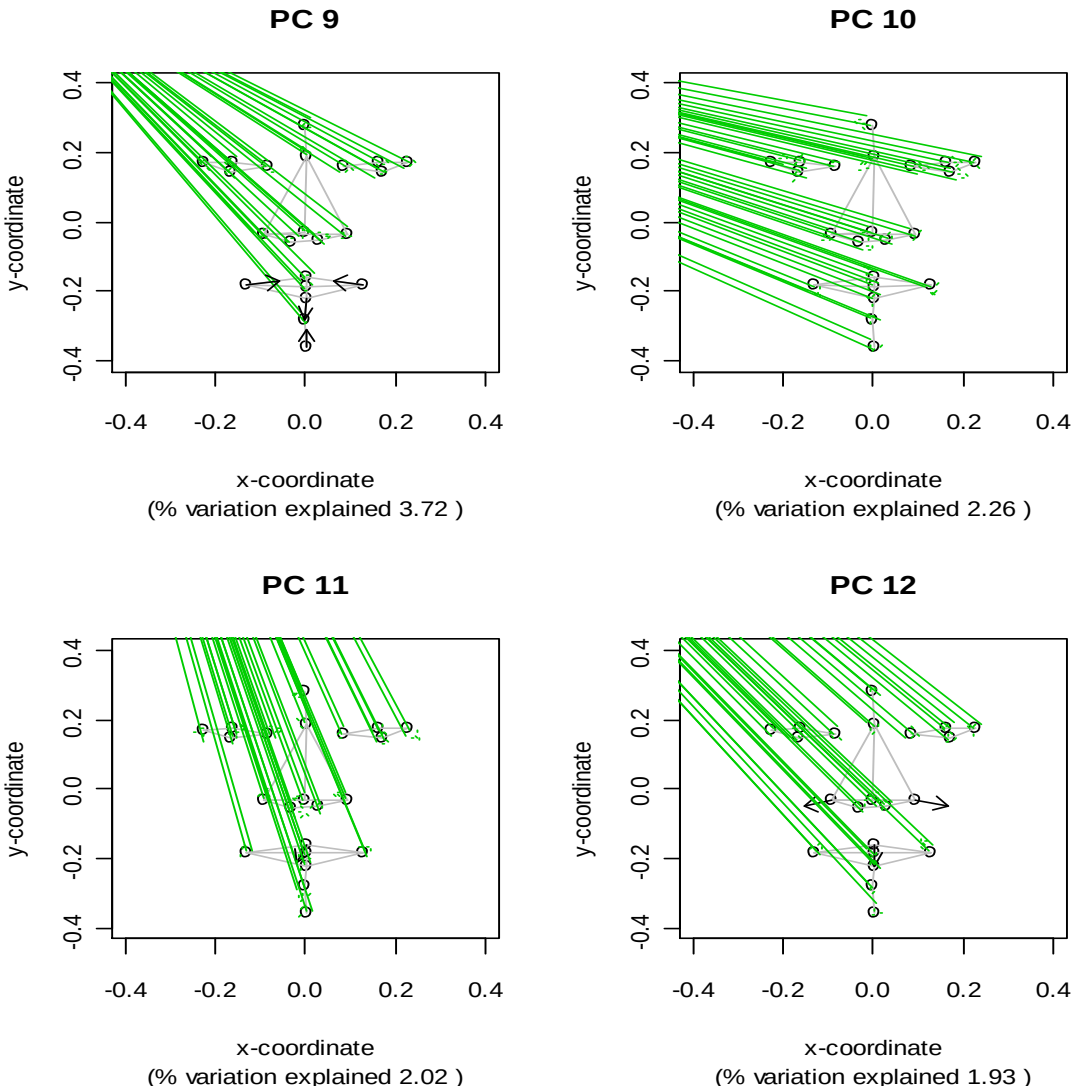


Figure 6.13 - PC plots (PCs 9-12) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows.

Table 6.3 could be used to assess which facial landmarks will be useful for facial identification. Requirements of a 'good' identification landmark are that it has to be accurate to place and yet sufficiently variable between faces to be able to distinguish individuals. Therefore the variation around the landmarks needs to be attributed to genuine facial variation and not to observer error. The landmark points which do not vary greatly in the twelve PCs in Figures 6.11 to 6.13 are unlikely to contribute very much in terms of distinguishing faces from one another, as they don't vary sufficiently.

PC	% Variation Explained	Landmarks Affected	Facial Area Affected
1	25.7	Crista philtra (left and right), pronasale	Nostrils and tip of the nose
2	16.5	Exocanthion (left and right)	Outer eyes
3	8.7	Glabella	Centre of forehead
4	6.3	Crista philtra (left and right)	Nostrils
5	5.7	Crista philtra (right), chelion (left), pogonion	Nostril, outer lips, chin
6	4.5	Glabella, subnasion, crista philtra (right)	Forehead, nostril
7	4.1	Crista philtra (left and right), labiale inferius	Nostrils, lower lip
8	3.9	Subnasion, chelion (right), labiale superius	Forehead, lips
9	3.7	Chelion (left and right), sublabiale, pogonion	Outer lips, chin
10	2.3	None above threshold of 0.05	
11	2.0	Stomion	Centre of lips
12	1.9	Alare (left and right)	Nose width

Table 6.3 – Summary of which facial landmarks vary the most on each PC

6.5 A Multivariate Normal Model for Facial Shape

The complete replicated anterior facial data (n = 2572 landmark configurations of k = 22 landmarks) was modelled using a multivariate normal distribution. For a multivariate normal distribution the squared Mahalanobis distances are Chi-squared distributed with the degrees of freedom being equal to the dimension of the space of the distribution (here this is 2*k, k landmarks in 2 dimensions). To check the fit of the multivariate normal model for the facial data the sorted squared Mahalanobis distance, of each configuration to the mean, was plotted against the appropriate Chi-squared distribution. A QQ plot of this is displayed in Figure 6.14; the linearity of the plot indicates the model is a reasonable enough fit to continue with this model in light of the fact there is no straightforward alternative.

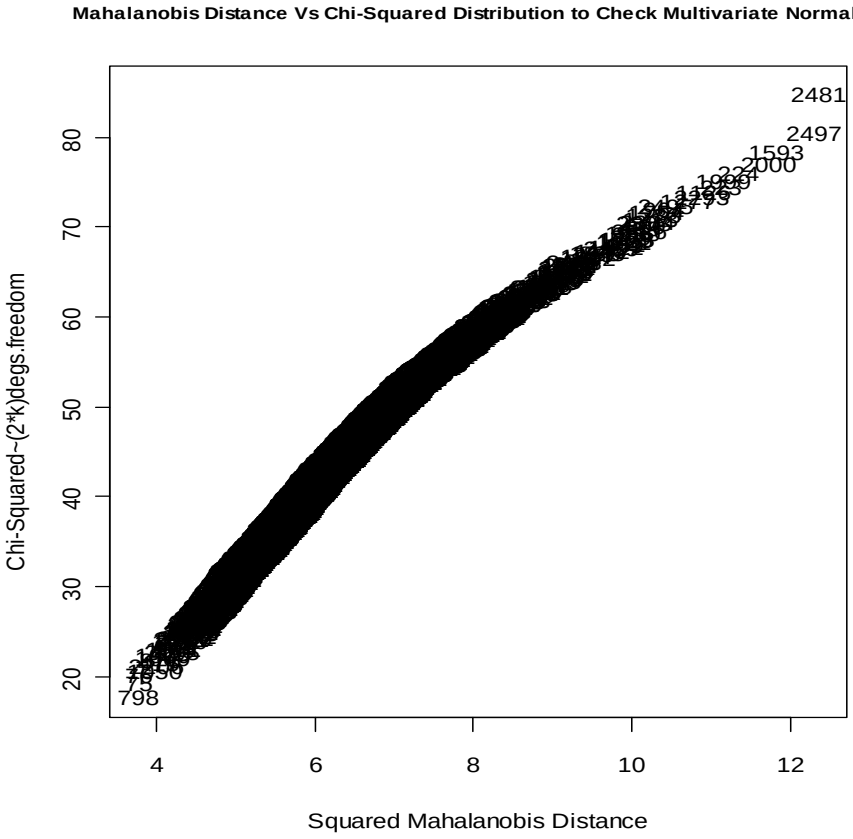


Figure 6.14 – QQ plot to check multivariate normality of complete replicated data configurations (*n* = 2572*)* of 22 landmark points

To develop a method for quantifying facial matches, estimates of the multivariate normal distribution parameters (including the covariance matrices for within-face and between-face variation) for the Geometrix® facial database could be used to derive likelihood estimates for how similar two faces are.

6.6 Summary

This chapter has examined the variability of the facial landmark data in the main Geometrix® database (§2.4) available to obtain estimates for the likelihood of facial matches based on the known facial variation in the large sample. Procrustes methods (§3.4) were used to align the data and differences in size and shape have been examined (§6.2.2). Male faces were found to be generally larger than female faces though not by a significant amount (§6.2.3). There were no overall shape differences between males and females once the scale factor was removed from the data (§6.2.4).

Several observations in the database were found to contain errors (§6.2.5). The data were cleaned where possible (correcting for mislabelled landmark points), or excluded where mistakes could not be corrected (incorrectly positioned landmark points).

It was revealed that transforming the Procrustes registered facial landmark data onto principal components (PCs) explained different aspects of shape variation in the main facial database (§6.2.6, §6.4.1). Groups of the original landmark variables showed large variation on particular PCs, with variation in specific facial areas being seen in each PC (Table 6.3).

Looking at the variability of the landmarks, via the PC loadings, indicated which points varied the most in the background data (§6.2.6, §6.4.1). The most appropriate of these to use for facial matching will be where the high variability is attributable to actual facial differences and not observer differences, therefore the consistency of placement of the points should be good.

The facial landmarks found to vary the most in 3D were the points around the ears (§6.2.6). When thinking about matching facial images obtained from other sources outside of the Geometrix® database the anterior (forward facing) facial view is likely to be the best to use, as this view encompasses the most landmarks (§6.4). The ear landmarks should be excluded from anterior facial matching, as they are not easily placed in this view. In addition if using the anterior facial views the data must be Procrustes aligned including the removal of scale information, as the distance of the subject to camera will be unknown and unlikely to be the same in two images obtained from different sources.

Sufficient knowledge about the available facial data and variation of landmark points has been acquired to enable the data to be used to quantify facial matches. A multivariate normal model seems to fit the facial landmark data well (§6.5). Obtaining parameter estimates for this model will enable measures for the estimation of how similar two faces are. The following chapter goes on to use the model for the data to estimate likelihood ratios for anterior facial matches using the MVNLR procedure (§3.8.4).

7 Likelihood Ratios for Quantifying Facial Matches

7.1 Introduction

This chapter applies and extends the method of using likelihood ratios to evaluate the multivariate evidence for facial matches. The MVNLR method (§3.8.4) was applied to two test datasets (§7.2.1, §7.2.2). The first consisted of ten faces from the main facial database (§2.4), measurements collected by two different observers were compared for matches. The second dataset was from the FBI and consisted of sixty facial images of which some were known matches, some were known exclusions and some were undecided (§2.7.1). Assessment of the LR results for these datasets revealed that some extensions of the MVNLR method were required to accommodate the large amount of information available (§7.3). The rotation of the shape data onto principal components was necessary to overcome the high correlation in the data, thus making the principal components (PCs) the matching variables. Checks to ensure the facial comparison data was in line with multivariate normal model for the background data were performed. Data anomalies were corrected or excluded from the facial comparisons (§7.2.2.2).

There was evidence to suggest that a subset of PCs would improve the MVNLR procedure for matching faces (§7.2.3). It was seen that the inclusion of certain PCs resulted in a decrease in the LR, and hence the evidence for a match, across a number of known matches. These troublesome PCs were perhaps associated with facial expression (the lower lip) and the subject's position away from the 2D anterior view (nose width) (§6.3, §7.2.3). There was also indication that a threshold for the LR was required, where a match is only confirmed if the LR is greater than the threshold (§7.2.1.1).

A novel method to select which variables to include in a subset has been proposed. A match/exclusion ratio (MER) was used to measure the subset performance (§7.3.2). The MER measures the magnitude of the average known match LR against that of the average known exclusion LR for a set of facial comparisons. Subsets with MER values close to one were sought, indicating they were equally fair for both hypotheses of facial matching. A further selection criterion was the magnitude of LRs produced by a subset to indicate the amount of evidence to support known results. Preferable subsets produced strong evidence, i.e. large values of LRs for known matches and small values of LRs for known exclusions.

An investigation into LR thresholds to confirm matches and exclusions was carried out (§7.3.3). The ratio of true to false results and the average strength of the match results were seen to vary for different thresholds. The most suitable threshold to apply to the facial data was found to be LR>300 for a match and hence LR< 1/300 for an exclusion. Values between these two thresholds were not reliably associated with positive (LR>1) or negative (LR<1) results, these were therefore deemed to have 'insufficient evidence to support either hypothesis'.

Thirteen different subsets of facial landmarks were investigated (§7.3.3, §7.4.1), taking into account what was learnt in §6.4.1 in terms of the facial variation explained by each of the landmarks. Landmarks were Procrustes registered and transformed onto PCs. All subsets of the first ten PCs were examined for use as matching variables to optimize the facial matching results for a set of known matches and exclusions (§7.3.1, Table 7.7). Subsets were taken of PCs rather than the original landmarks because certain landmarks were important on some PCs and not on others (§6.4.1). To exclude such landmarks totally from the analysis could lose valuable information on facial variation and so was thought inappropriate. All subsets were examined rather than the first k PCs (k =1 – 10) since 'better' results were obtained by excluding some intermediate PCs (§7.2.3).

Subsets that fit both the MER and LR magnitude criteria (§7.4.2.2, Results Appendix D) were deemed good and evaluated on matching performance using the FBI anterior facial data (§2.7.2, §7.2.2) which contained known matches and exclusions. The numbers of true and false results were examined (§7.4.1, §7.4.2.3, Results Appendix D) and a 'best' subset was selected as the one with the lowest false (both positive and negative) rates (§7.4.1, §7.4.2.3).

The chapter concludes with an examination of the robustness of the methods. This was carried out by dropping a number of faces from the background data and recalculating the LRs for facial matches in the FBI anterior data. The performance of the 'best' subset (§7.4.2.3) was reassessed and the number of false results examined (§7.4.3). It was found that the more faces that were dropped from the background data the higher the false positive and negative results, indicating that the method was sensitive to changes in the background data and relied on the previous analyses (§6) being carried out on all available data.

7.2 LRs for Quantifying Facial Matches

7.2.1 Test Data 1 – Matching Data from Two Observers

The data (outlined in §2.7.1) consisted of a sample of ten different facial images taken from the main facial database (§2.4). Two observers each measured the landmark positions on these images twice. The measurements from observer one were compared with those from observer two to look for facial 'matches'.

The anterior (x, y) facial view was taken for analysis, so landmark configurations were in 2D. The total number of available matching variables (§3.8.4) was sets of (x, y) coordinates for each of the twenty-two landmark points (so, $p = 44$). The twenty sets of landmark measurements (two sets for each of the ten faces) for observer one (n_1) and the twenty sets of landmark measurements for observer two (n_2) were added to the (x, y) coordinates of the 2572 configurations from the Geometrix® facial database, giving a total of $N = 2612$ face shape configurations in the background population (§3.8.4). The data were aligned using generalized Procrustes analysis (GPA, §3.6). Scale was preserved here, as it was assumed that all the subjects were positioned at the same distance from the 3D scanner when the facial images were captured. As mentioned in §6.4 scale would usually have to be removed, as other anterior images for comparison are unlikely to be obtained under controlled conditions.

Calculations of the MVNLR procedure (outlined in §3.8.4) were attempted on the $p = 44$ matching variables. A principal components analysis (PCA) on the Procrustes registered data produced a set of uncorrelated variables. More than 80% of the variation was contained in the first five principal component (PC) scores. These were taken for the LR analysis.

The LR analysis was then carried out by taking each of the twenty test faces in turn as the control face, which was assessed against each of the other nineteen test faces (the recovered face) to evaluate with a LR the strength of evidence for a facial match. A total of 190 (1/2 x 20 x 19) LRs were calculated. It was known that the control and recovered data for ten LRs came from the same source (i.e. were a known match), therefore there were 180 sets of control and recovered measurements that were known exclusions.

7.2.1.1 Test Data 1 - Results

In the 190 runs of the MVNLR procedure there were ten LRs greater than one and therefore favouring the H_p over the H_d, indicating the possibility of a facial match. All ten of these results were the known matching observers measurements matches using data from the two different observers (§2.7.1). There were no false positive or false negative results. The LR results for the ten found matches are displayed in Table 7.1, sorted by the strength of support for H_p. The LR values had a large range, from a match with LR>100,000, which is obviously much more supported than a match with an LR>16. This suggests that perhaps a result should only be confirmed a match if the LR is above a certain value, probably greater than one. This is explored further in §7.3.3.

Face	N	Likelihood Ratio
2	2612	163222.7
3	2612	135449.7
5	2612	27977.4
4	2612	16905.5
10	2612	11906.1
7	2612	5879.8
6	2612	3494.3
1	2612	632.5
9	2612	537.5
8	2612	16.1

Table 7.1 - Likelihood ratio results for 10 matches (LR>1) found from pair-wise comparisons of test data 1 (10 faces measured by two observers), using the first 5 PC scores

The LRs were recalculated using the first twenty PC scores instead of the first five, this reduced the original number of $p = 44$ characteristics by twenty-four variables. The first twenty PCs represented about 96% of the variation in the data. The LR results are displayed in Table 7.2, again only the ten simulated known matches were shown to have a LR greater than one, and there were no false positive or false negative results. Table 7.2 shows that the ranking of the strength of support for H_p from the LR tests (in terms of how much greater than one the LR statistics were) has changed.

Comparing Tables 7.1 and 7.2 it can be seen that even with $p = 5$ characteristics all the LR results for the manufactured matches were larger than one (ranging from 16 to 163000) and so were more in favour of H_p over H_d. When p was extended to twenty characteristics the results were all much greater than one and probably unnecessarily large. All these results were taken from extremely controlled data and an investigation

into more real-life data was necessary, which follows in §7.2.2. Still this small simulated dataset has proved that the MVNLR procedure (§3.8.4) works for evaluating matches in facial shape data.

Face	N	Likelihood Ratio
6	2612	2.02E+26
4	2612	2.23E+23
10	2612	3.91E+21
3	2612	7.71E+19
5	2612	1.04E+19
2	2612	4.52E+18
7	2612	1.45E+18
1	2612	1.62E+15
8	2612	1.15E+14
9	2612	1.19E+09

Table 7.2 - Likelihood ratio results for 10 matches (LR>1) found from pair-wise comparisons of test data 1 (10 faces measured by two observers), using the first 20 PC scores

The wide range of LR values (Tables 7.1 and 7.2) indicates that perhaps a threshold for confirming a match is required, this is explored in §7.3.3. The ranking of the facial matches in terms of the strength of evidence (i.e. size of LR) for the matches differs between Table 7.1 and Table 7.2. This indicates that perhaps the inclusion of certain PCs does not have the effect of improving the LR, this is investigated further in §7.2.3.

7.2.2 Test Data 2 - FBI Suspects Data

Duplicate measurements taken by one observer from sixty facial images provided by the FBI (displayed in Figures 13.1 – 13.2 and described further in §2.7.2) were added to the existing Geometrix® data to form the background database (N = 2692 configurations) from which to calculate LRs for facial matches. The sixty images for comparison contained known, possible and supposed matches and exclusions of pairs (or more) of faces that visually looked similar. Procrustes alignment of this background data was carried out, including the removal of scale from the data. A principal components analysis (PCA) was carried out on the Procrustes registered data, and the first five PC scores were taken as the p = 5 data characteristics in the LR analysis. These PCs contained around 62% of the variation in the data, around 20% less than was represented by the first five PCs of test data 1 (§7.2.1) taken from the facial image

database. This is as expected as all images in that dataset were taken under controlled conditions.

7.2.2.1 Test Data 2 - Results

The LR analysis was carried out by taking each of the sixty faces in turn as the control face and assessing this against each of the other fifty-nine faces in the dataset (the recovered face) by calculating a LR for the evidence to support a facial match. A total of 1770 (1/2 x 60 x 59) tests were carried out. According to the LR results there were 137 facial matches found in the 1770 tests (i.e. 137 LR results were greater than one). These matches were all examined by going back to the original data sources (i.e. the images) and by referring to a list of notes the FBI had sent with the images to explain the facial comparison cases (§2.7.2).

Of the 137 LR results that were greater than one it was found that 54 (39%) were 'yes' definite known matches, 67 (49%) were 'no' false positive results, 8 (6%) were 'possible' matches and 8 (6%) were 'supposed' matches. These results are displayed in Table 7.3 and an explanation of these comments is given in §2.7.2. Also displayed are the exclusions where LR<1 (i.e. results which favour the defence hypothesis H_d that the faces do not match). The percentage of false negative results was around 3%, much lower than for the false positives. However, the number of known exclusions was much greater than the number of known matches.

	Yes	**No**	**Possible**	**Supposed**	**Unverifiable**
LR>1	54	67	8	8	0
% LR>1	39.4	48.9	5.8	5.8	0.0
LR<1	56	1546	18	10	3
% LR<1	3.4	94.7	1.1	0.6	0.2

Table 7.3 – Numbers and percentages of 'matches' (LR>1) and 'exclusions' (LR<1) from the FBI anterior test data (n = 60 faces) using p = 5 PCs as the number of matching variables. Columns indicate true matches (yes), true exclusions (no), possible, supposed and unverifiable matches – explained meanings in §2.7.2.

The LRs were recalculated increasing the number p of variables used in the calculations to be $p = 20$, the first twenty PCs from the PCA carried out on the Procrustes registered data. These twenty PCs represented 94.4% of the variation in the data. These variables

produced 63 resulting LR 'matches' (where the LR was greater than one). Of these 56 (89%) were true positive matches, 3 (5%) were false positives, 4 (6%) were 'possible' matches and none were 'supposed' matches. Table 7.4 summarizes all the results.

	Yes	No	Possible	Supposed	Unverifiable
LR>1	56	3	4	0	0
% LR>1	88.9	4.8	6.3	0.0	0.0
LR<1	54	1613	23	14	3
% LR<1	3.2	94.5	1.3	0.8	0.2

Table 7.4 – Numbers and percentages of 'matches' (LR>1) and 'exclusions' (LR<1) from the FBI anterior test data (n = 60 faces) using p = 20 PCs as the number of matching variables. Columns indicate true matches (yes), true exclusions (no), possible, supposed and unverifiable matches.

Comparing Tables 7.3 and 7.4 it is seen that almost the same number of true positive matches were obtained through both runs of the LR calculations, p = 5 resulted in 54 true positive matches whereas for p = 20 there was 56. The main difference in increasing the number of variables to use in the LR calculations was that it greatly reduced the number of false positive results, when p = 20 only 5% of results were false positive. This improvement may also be seen for $5 < p < 20$, therefore the number of variables to use for facial matching using LRs will be investigated further in §7.2.3. LR results for known matches are examined increasing the number of p matching variables incrementally from 2 up to 20.

7.2.2.2 Checking the Model Fit and Data Cleaning

Test data 1 was taken from the main facial database (§2.4) and the multivariate normal model had been shown to be a good fit for this data, §6.5. Test data 2 did not come from the same source and so a check was performed to ensure that the assumptions of normality in the LR calculations were appropriate. The FBI anterior test data was added to the Geometrix® data to make the background population and (as previously in §6.5) a QQ plot of the sorted squared Mahalanobis distance of each configuration to the mean was plotted against the appropriate Chi-squared distribution, Figure 7.1.

The multivariate normal model seemed to fit the data well with the exception of three observations which lay far away from the model, Figure 7.1. To see if there were any discrepancies in the position of landmark points (in a similar way to how outliers were

dealt with in §6.2.5) the two duplicate measures for each of the three faces corresponding to the outlying observations were plotted on the same plot. It was found that two of the outliers had one landmark misplaced, these two mistakes could not be corrected for and so these faces were excluded from further analyses. The remaining outlier had simply mislabelled four landmarks, swapping the label for landmark 8 with landmark 10, and that for landmark 9 with landmark 11. This data anomaly could be corrected as the points were in the correct location, just labelled incorrectly.

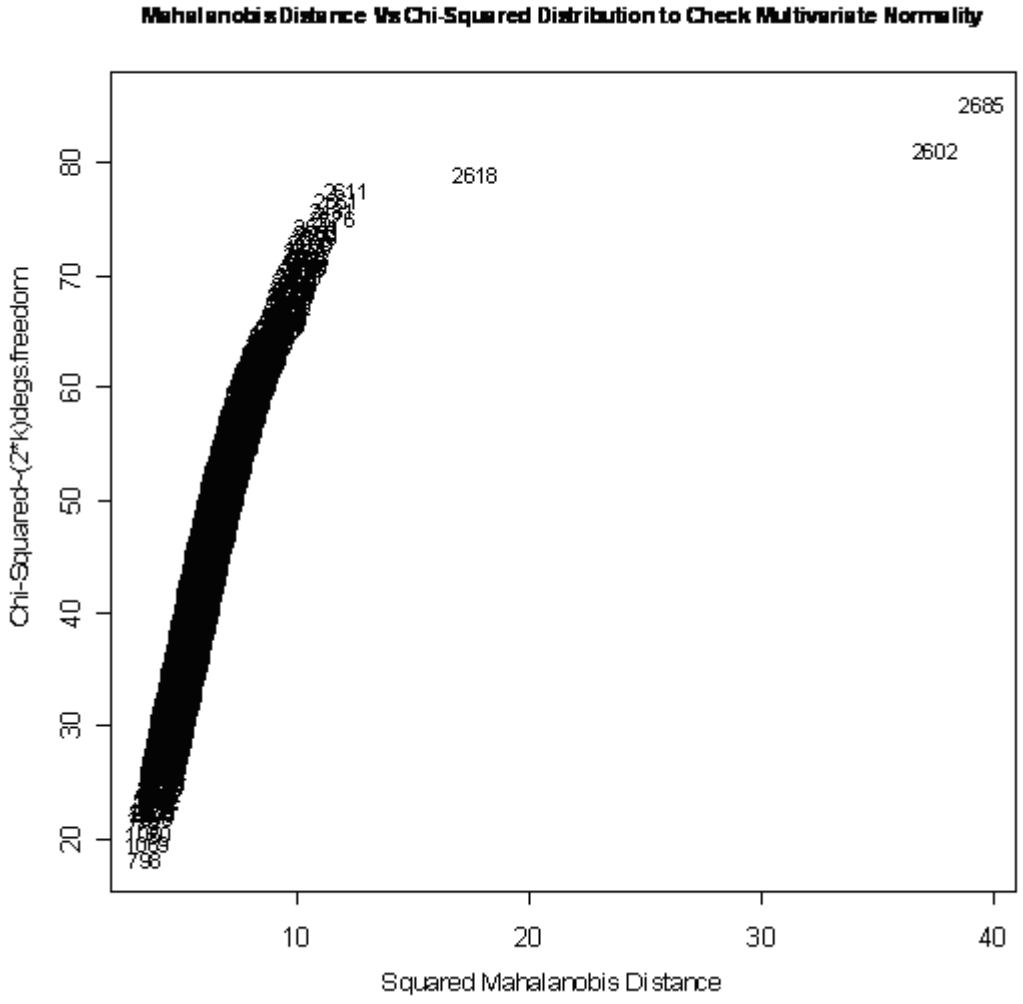


Figure 7.1 – QQ plot to check multivariate normality of the background data plus the facial comparison data (control and recovered), three clear outliers are apparent.

7.2.2.3 Results after Data Cleaning

The LR analysis was repeated taking each of the fifty-eight remaining FBI anterior faces in turn as the control face and testing it one by one against each of the other fifty-seven test faces in the dataset (the recovered face). A total of 1653 (1/2 x 58 x 57) tests were carried out on the cleaned data. The LRs using the first five PCs and also the first twenty PCs as matching variables were calculated and resulting numbers of matches and exclusions are displayed in Table 7.5. Results showed that removing the three outliers (§7.2.2.2) had an effect on the facial matches obtained. Using the first twenty PCs as the matching variables before the data were cleaned there were fifty-two true matches and sixty-three false matches (Table 7.5). After the data were cleaned there were fifty-one true matches and only three false matches (Table 7.5). Similarly for exclusions prior to data cleaning there were 1436 true exclusions and fifty-five false exclusions (Table 7.4), whereas with the cleaned data there were 1496 true exclusions and fifty-six false exclusions. So, even though just three points were incorrectly positioned in the background database of twenty-two points for $n = 2692$ configurations, the effect this had on the multivariate model and estimates of covariance produced quite different match results. In other words the LR method was found to be sensitive to inconsistencies within the data.

	No. PCs	Yes	No	Possible	Supposed	Unverifiable
LR>1	5	54	47	7	9	1
% LR>1	5	45.8	39.8	5.9	7.6	0.8
LR<1	5	53	1452	19	9	2
% LR<1	5	3.5	94.6	1.2	0.6	0.1
LR>1	20	51	3	3	6	0
% LR>1	20	81.0	4.8	4.8	9.5	0.0
LR<1	20	56	1496	23	12	3
% LR<1	20	3.5	94.1	1.4	0.8	0.2

Table 7.5 – Numbers and percentages of 'matches' (LR>1) and 'exclusions' (LR<1) from the cleaned FBI anterior test data ($n = 58$ faces) using $p = 5$ and $p = 20$ PCs as the number of matching variables. Columns indicate true matches (yes), true exclusions (no), possible, supposed and unverifiable matches.

Table 7.6 displays the actual known number of matches and exclusions for the fifty-eight FBI anterior test faces; these were obtained through reading the FBI case notes accompanying the images. Although using the first twenty PCs as matching variables produced 81% of match results (LR>1) which were genuine (Table 7.5), only fifty-one

out of one hundred and seven matches had been identified. An examination into the factors that could be attributed to these known matches not being identified by the LR method is carried out in chapter 8.

Match?	Number	%
Yes	107	6.5
No	1499	90.7
Possible	26	1.6
Supposed	18	1.1
Unverifiable	3	0.2

Table 7.6 – The actual number and percentage of known matches and exclusions in the fifty-eight FBI anterior faces

7.2.3 Evidence for Potential Improvements to the Method

A selection of ten images (§2.7.3, Confidential Appendix, Figure 13.3) that were known facial matches were selected from the sample of FBI anterior images to illustrate the effect of increasing the number of PCs used as the p matching variables on the resulting LRs. It was thought that $p = 5$ were too few variables and produced too many false positive results (§7.2.1.2, §7.2.2.3). Very few false positive results were obtained when using $p = 20$ variables (§7.2.1.2, §7.2.2.3), however it could be more efficient to use less than twenty variables if equally low numbers of false positives results could be obtained.

The LR method was run nineteen times on the ten known matches (§2.7.3, Confidential Appendix Figure 13.3) from the FBI anterior data. Each run was for a different number of matching variables (PCs) from $p = 2$ to $p = 20$. The results were collated and for one particular face (labelled 1 in Confidential Appendix Figure 13.3), which was a known match with the nine other faces (labelled 2-10 in Confidential Appendix Figure 13.3), 2*log(LR) was plotted against the number of PCs used as matching variables in the LR calculation, Figure 7.2.

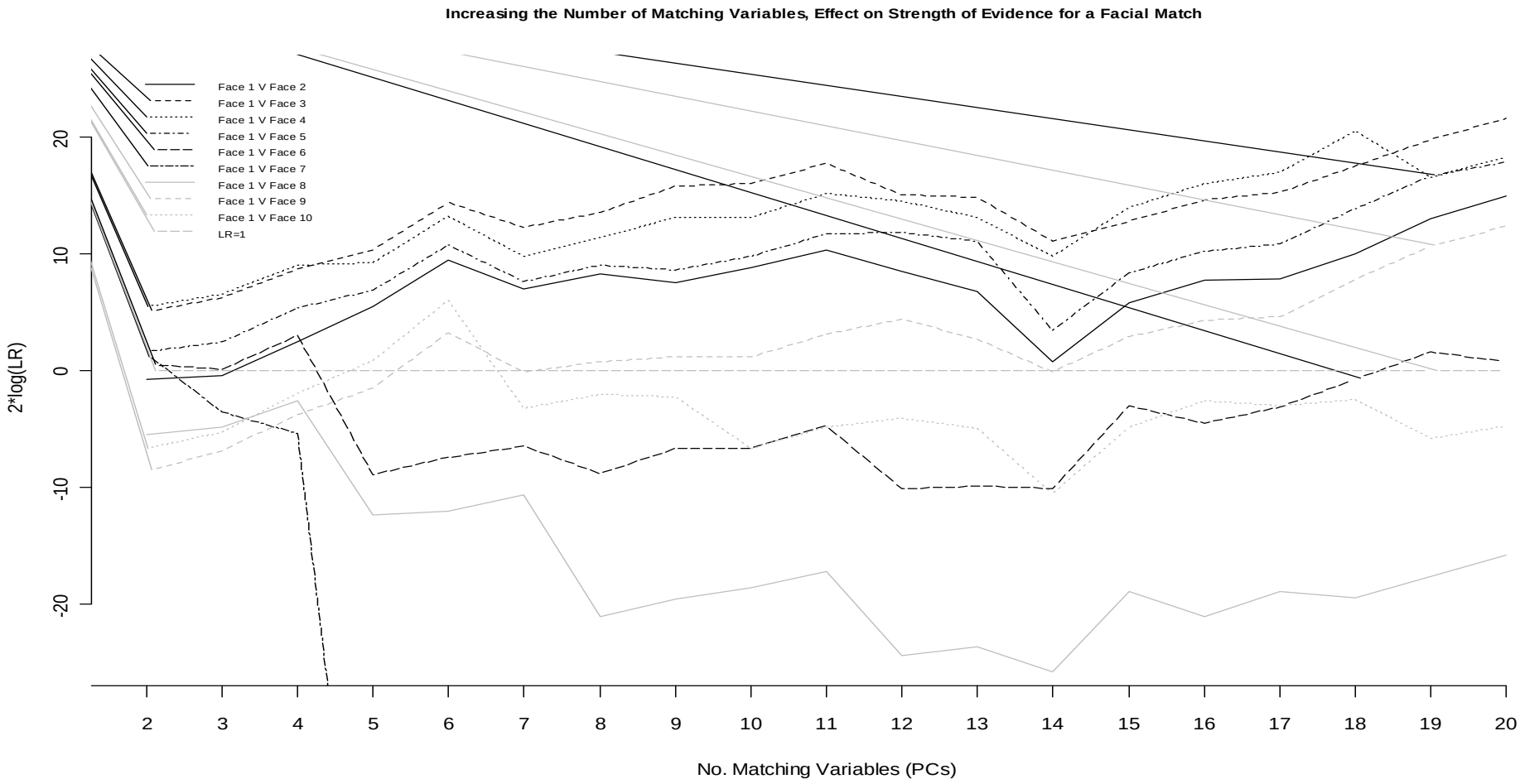


Figure 7.2 – Graph to show how LR results varied as the number of PCs to use as matching variables in the LR procedure were increased.

Figure 7.2 shows a plot of performance of the LR method against dimensionality, generally such a plot shows improving performance as the dimensions of the problem increase. Figure 7.2 shows that the increase in evidence for a match is not monotonic across the increase in PCs used for the LR calculations. For PCs = 7, 12, 13 and 14 there are breaks in the general increasing trend for several of the facial comparisons. Also, for four of the facial comparisons the LR results were false negative (LR<1), these were when face 1 was compared with faces 6, 7, 8 and 10. Looking at the images in Confidential Appendix Figure 13.3 it can be seen that faces 6 and 7 do not depict a neutral facial expression; the subject was smiling which could have an effect on the position of certain facial landmarks around the lips and eye areas. The LR procedure used a model for face shape where the subject had a neutral facial expression; this could explain why for images 6 and 7 a match was not determined. It is less easy to see how images 8 and 10 were different to image 1, the subject's shoulder position to camera was not strictly forward facing, although for image 9 this was also the case and this was seen to match with image 1 (LR>1) from PC7 onwards. It was also interesting to note that image 1 matched (LR>1) with both images 2 and 4 even though the subject was wearing glasses in these two images. If glasses do not obstruct the landmark points the face is not actually altered in terms of shape by wearing them, unlike a change in facial expression which causes landmarks to move position.

The plot in Figure 7.2 was repeated for different sets of known facial matches, it was seen that the PCs showing peaks and troughs in the strength of evidence for a match were not consistent across all facial comparisons. However, some PCs showed troughs that were common to many known match pairs, for example PCs 7 and 12 in Figure 7.2. Referring back to the variation of the facial landmark points (§6.4.1, Table 6.3) these PCs show variation in the lower lip (labiale inferius) and the nose width (left and right alare). These facial features are likely to be affected if the subject in the image has their mouth open (lower lip) or is positioned away from the anterior position in the 2D view (nose width). The removal of such PCs may produce better overall match results if the trough can be removed from the affected facial comparisons without removing other important evidence from the known match pairs that did not show the trough. The merits of removing such features also apply to the real life application, as photos from scenes of crime might show alternative facial expressions to 'neutral' and are likely to be of different facial angles than the anterior view.

For other sets of known matches the LR results were always shown to take a value less than one, hence giving false negative results. Looking at the source images (Confidential Appendix, Figures 13.1 and 13.2) for these data it was seen that the angle of presentation of the subject's head to the camera could be affecting these results; certain images were obviously taken at different angles. A further investigation into this is carried out in §8.4.

7.3 Suggested Improvements to the Method

7.3.1 Selection of a Subset to Optimise Match Results

Usually when selecting a subset of variables to use in data analysis we are looking to preserve as much information as possible in the data, yet reduce the dimensionality of the problem. In other words discard variables that don't add enough value to the results of an analysis. Typically the measure of information increases monotonically as the dimension increases, e.g. percentage of variation explained by successive numbers of PCs. By contrast, here where the criteria are that we want to obtain LRs for facial matching of the 'best' quality possible this is not the case. By 'best' we mean that we want a subset that 'provides the best quality of evidence to support a match or exclusion'. §7.2.3 suggested that taking the first few PC scores was inappropriate, as inclusion of certain PCs showed a decrease in the quality of evidence for a facial match (Figure 7.2). Therefore a subset of PC scores that produces the 'best' quality LRs and therefore evidence will be more appropriate for facial matching. It will be seen that 'best' means achieving a balance between conflicting requirements of providing evidence for matches and for exclusions (§7.3.2, §7.4).

7.3.2 A New Method for Subset Evaluation – Match: Exclusion Ratio

"A measure of evidential strength is required to tell us about the strength of evidence in support of a guilty or innocent proposition, without it actually telling us how likely or unlikely the proposition of guilt or innocence itself is", Lucy (2005). In other words some kind of impartial measure of the strength of evidence is required, to present to a court of law. Therefore the subsets of facial matching variables should produce a quality of evidence that not only gives a high LR result for known facial matches, but also gives a low LR result for the known facial exclusions. To achieve such a balanced result a

match/exclusion ratio was calculated for each subset, this was used as a measure to assess the suitability of the subset for facial matching.

Usually with a subset selection procedure there is some kind of penalty applied for number of variables in the subset, e.g. Akaike's information criteria (AIC). The facial data had previously shown that LR results did not increase monotonically with the number of (the first few) PCs used in LR calculations (§7.2.3). Certain PCs showed a decrease in the strength of evidence for a facial match, so it was not necessarily appropriate to penalize for the number of PCs in the subset, as more did not necessarily equate to better quality results. A different penalty measure was required. It was thought that given the severity of false results here (i.e. the miscarriages of justice that may be carried out based on a false facial identification), any subsets that produced averagely false results, either false positive or false negative, should be immediately discarded.

Hand (1997) says the key to effective feature selection is in finding a 'good' measure to separate one subset from another. A method to measure which subsets produce good quality evidence was required. 'Good' in the sense of being easy to calculate and optimize, as well as being related to the criterion of interest; here this was getting accurate facial matching results from the LRs. With this in mind it was thought that statistical evidence for a courtroom should not be biased for or against a trial outcome; it should provide independent evidence evaluation for both the prosecution and defendant. Therefore what was needed from a 'good' subset of PCs in this case was one that was equally good at picking out or excluding a facial match correctly. A simple method to measure this would be to examine the ratio of the magnitude of the average match LR in relation to the average exclusion LR for each subset, i.e. a match/exclusion ratio (MER).

$$ME\ ratio = \frac{Average\ LR(known\_match)}{\left(Average\ LR(known\_exclusion)\right)^{-1}}$$

'Good' subsets were those with a non-biased MER of or close to one, i.e. the subset was equally as good at quantifying a match as it was at excluding a non-match. In addition to this the average LR values also needed to provide strong support for the match or exclusion hypothesis, i.e. large LR values for matches, small LR values for exclusions.

The methods essentially first carried out a feature extraction; a PCA to get important derived PCs from the original raw measured landmarks. This was then followed by a subset selection on the extracted PCs, using the MER as the performance measure and selection criterion.

The 'best' subsets of PCs were determined for subsets of twenty-two, fifteen, eleven and ten landmarks; the results and reasoning behind the choice of landmarks for these are in Appendix D and §7.4.2. There were thirteen different subsets that were found worthy, in terms of the MER criterion, of further investigations into their matching performance, §7.3.3, §7.4.1.§

7.3.3 LR Thresholds for Claiming Matches and Exclusions

An investigation into how LR threshold affected subset performance was carried out. This sort of trade off between sensitivity and specificity is often shown with a receiver operator characteristic (ROC) curve, however usually one is only interested in the positive results (both true and false). Here were are equally interested in the negative results (i.e. the facial exclusions), so instead of an ROC curve the ratio of true results (both positive and negative) to false results was plotted against the LR threshold used to quantify a match for each of the thirteen 'best' subsets evaluated for performance, Figure 7.3. In general as the LR threshold was increased the true to false result ratio also increased, Figure 7.3. A threshold of 1 for a match was clearly too low and the ratio of true to false results was less than 5:1 for all subsets, increasing to 100 improved this for all subsets investigated and almost doubled the true to false ratio. Further increasing also improved the ratio of true to false results though not to as great an extent. The 'best' subset was the one with the highest true to false ratio. With the exception of using a threshold of 1, all other thresholds investigated showed that the 'best' subset was subset 6 (Figure 7.3), which consisted of PCs 1, 3, 4, 7, 9 and 10 from eleven landmarks (§7.4.2). As the threshold for a match was increased the number of results for which there were insufficient evidence to confirm either hypothesis also increased. If a threshold is taken to be too large there will be too many cases where there is insufficient evidence. Looking at Figure 7.3 increasing the LR threshold to greater than 300 did not improve much further the ratio of true to false results, so a threshold of 300 seems to optimise the facial matching results. In a court room it would be up to the prosecution or

defence to decide what they deem to be a suitable threshold for a ‘match’ LR, a good thing to do would also be to quote the percentage of true and false results obtained when using such a threshold.

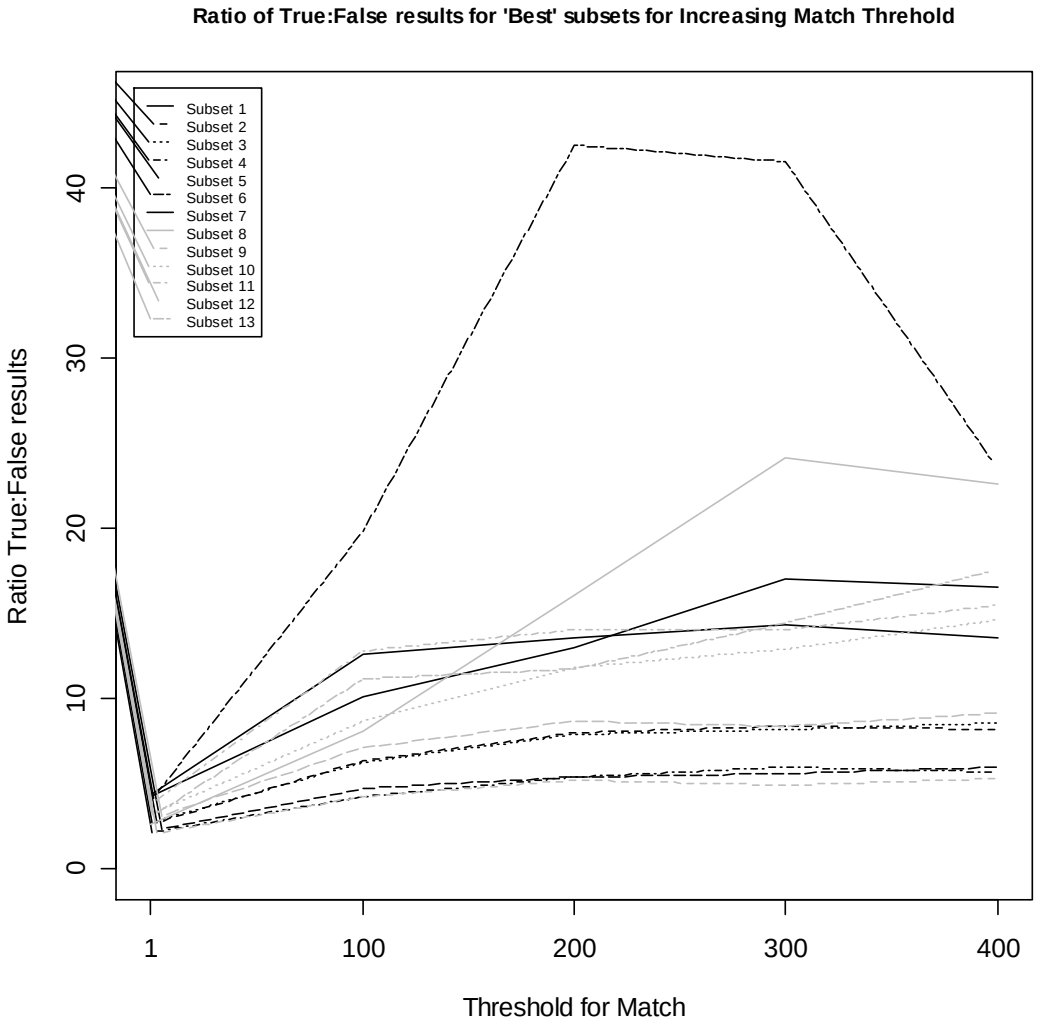


Figure 7.3 – Ratio of true to false results obtained for each of the thirteen subsets investigated for various LR thresholds used to quantify a ‘match’. The corresponding exclusion thresholds were (1/threshold for match).

Figure 7.4 shows the strength of the average LR (§7.3.2) for known matches for each of the thirteen subsets of matching variables which were evaluated. Obviously the subsets with the larger LRs were deemed better performers. Subsets 5 and 6 clearly had the ‘best’ results in providing on average the strongest evidence for known matches for all of the thresholds investigated. Subset 6 was PCs 1, 3, 4, 7, 9 and 10 from eleven landmarks (§7.4.2). As the MER (§7.3.2) was used to pick out the ‘best’ subsets, which

are equally as good at confirming either hypothesis, the average strength of evidence for facial exclusions will also show the same two subsets as the 'best' performers. Again when the LR threshold was increased to be greater than 300 this showed little improvement in the average match LR (Figure 7.4), so this strengthened the support for LR>300 being the optimum threshold for quantifying facial matches.

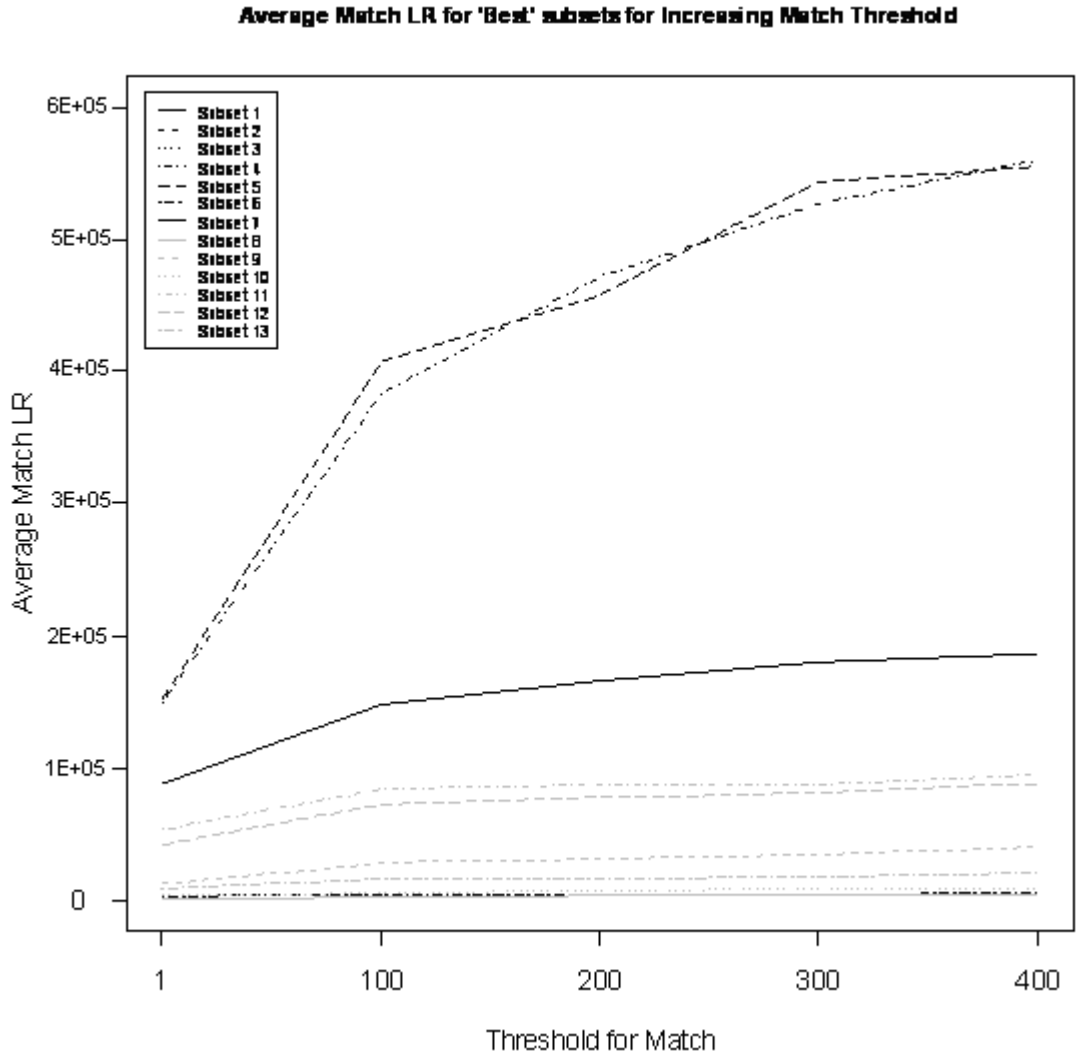


Figure 7.4 – Average strength of evidence for a match for each of the top thirteen subsets.

7.4 Applying the New Methods to Facial Data

To explore possible 'best' subsets of PCs a small subset of the FBI anterior facial data were arbitrarily selected. Using the fifty-eight FBI anterior faces (§2.7.2, §7.2.2.2) where the images were known matches or exclusions would take many hours. Instead the twenty comparisons in Table 7.7 were chosen, ten of which were known matches

and ten of which were known exclusions (§2.7.4, Confidential Appendix C Figures 13.4 – 13.23). Table 7.7 shows these facial comparison pairs, along with the LR result that were obtained by using the first twenty PCs for the twenty-two anterior facial landmarks, to give an idea of the strength of evidence that was available for each match or exclusion. It was thought that the number of matching variables should be able to be reduced without sacrificing the power of the results obtained for these comparison pairs.

Face i	Face j	Known Result	LR using 22 landmarks, PCs 1-20
A14.tif	A15.tif	match	3.04E+11
A06.tif	A09.tif	match	2.25E+11
A13.tif	A14.tif	match	1.03E+11
A07.tif	A08.tif	match	5.27E+09
A15.tif	A16.tif	match	3.21E+10
A13.tif	A17.tif	match	1.05E+07
A08.tif	A16.tif	match	3.18E+06
A06.tif	A10.tif	match	3.41E+03
A05.tif	A08.tif	match	9.66E+03
A22.tif	A23.tif	match	5.46E+05
A02.tif	A51.tif	exclusion	2.77E-158
A02.tif	A50.tif	exclusion	0.00E+00
A14.tif	A51.tif	exclusion	8.55E-147
A13.tif	A51.tif	exclusion	1.35E-143
A15.tif	A51.tif	exclusion	9.06E-140
A05.tif	A19.tif	exclusion	9.13E-47
A05.tif	A24.tif	exclusion	8.15E-41
A05.tif	A21.tif	exclusion	2.72E-23
A07.tif	A42.tif	exclusion	4.41E-48
A05.tif	A25.tif	exclusion	1.28E-22

Table 7.7 - 20 facial comparisons (Confidential Appendix, Figures 13.4 – 13.23) used to test LR matching results for different subsets. The LR when using PCs 1-20 from 22 anterior landmarks is given.

There are many methods available for variable selection. The common techniques are stepwise and backwards methods, however both of these do not examine many of the potential subsets. Here it was decided to examine all possible subsets of the first ten PCs; these represent around 80% of the variation of the data. Also the PCs showing peaks or troughs in the strength of evidence (§7.2.3) were not consistent for different pairs of known matches, so looking at all possible subsets was more preferable than a backwards or stepwise approach. The number of possible subsets containing two or greater variables (at least two are needed to be able to 'match' configurations) of the first ten PCs was 1013.

The landmark configurations for the facial comparison pairs from Table 7.7 were taken as the control and recovered data for LR calculations. This gave twenty facial

comparisons for each of the 1013 subsets, making 20260 facial comparisons. This took two hours to run on a minimum specification computer. There were 3306 pair wise comparisons for the fifty-eight FBI anterior faces, so over three million facial comparisons for these pairs for all possible subsets of the first ten PCs. It is estimated that this would take approximately 300 hours to run.

A program was written to calculate LRs for the facial comparison pairs for every possible subset (of size two or greater) of the $p = 10$ variables. This produced results for the 1013 subsets for each comparison from which to choose a 'best' subset. The LRs for pairs of known matches were averaged for each subset. Similarly an average LR result was obtained for pairs of known exclusions. This meant one average LR result for known matches and one average LR test result for known exclusions for each of the 1013 subsets. In principle the individual LRs for each pair could have been used as a basis for the evaluation of subsets but averaging simplified the task.

MERs near to one were examined to find the 'best' subset in terms of being as good at finding a true match as it was at excluding a true non-match, therefore producing an equally fair strength of evidence estimate in respect of the two hypotheses for the prosecution and defence. In addition to the MER, subsets that produced LRs with large magnitude were considered better i.e. in terms of providing stronger evidence for a either a match or exclusion.

7.4.1 Summary of Facial Matching Performance for all Subsets Investigated

The 'best' subsets found from each of four landmark subsets investigated (described further in §7.4.2 and Results Appendix D) were used to run the LR facial comparisons on all pair wise combinations of the fifty-eight FBI anterior faces (§2.7.2, §7.2.2.2); the numbers of true and false results were examined to assess how well the subset performed at matching and excluding (§7.4.2.3 and Results Appendix D). Thirteen different subsets of matching variables were selected as the 'best', Table 7.8 summarises the subsets performance. §7.4.2 gives further details of all results for one particular subset of eleven landmarks and Appendix D lists the results for all other subsets investigated. Different thresholds to determine a 'match' were examined, these were 1, 100, 200, 300 and 400.

PCs in Subset	Landmarks	Exclusion Threshold	Yes	No	Possible	Supposed	Unverifiable	Match Threshold	Yes	No	Possible	Supposed	Unverifiable	False positives	False negatives
1,3,4,7,9,10	11	LR<0.00333	2.37	95.81	1.05	0.56	0.21	300	87.76	2.04	2.04	8.16	0.00	2.04	2.37
1,3,4,7,9,10	11	LR<0.01	2.55	95.59	1.03	0.62	0.21	100	86.67	5.00	1.67	6.67	0.00	5.00	2.55
2,3,6,7,8,9	22	LR<0.00333	1.55	97.67	0.39	0.16	0.23	300	78.85	5.77	15.38	0.00	0.00	5.77	1.55
1,2,3,4,5,6,7,8	11	LR<0.00333	2.61	95.24	1.21	0.74	0.20	300	82.35	7.84	9.80	0.00	0.00	7.84	2.61
2,3,4,6,7,10	10	LR<0.00333	2.58	95.39	1.29	0.54	0.20	300	80.33	9.84	9.84	0.00	0.00	9.84	2.58
3,4,5,6,7,8,9	15	LR<0.00333	2.22	95.69	1.18	0.69	0.21	300	83.33	10.61	6.06	0.00	0.00	10.61	2.22
3,6,7,8,9,10	11	LR<0.00333	1.95	95.75	1.30	0.79	0.22	300	87.69	10.77	1.54	0.00	0.00	10.77	1.95
2,5,6,7,8,9,10	22	LR<0.00333	2.25	96.01	0.94	0.58	0.22	300	73.91	10.87	2.17	13.04	0.00	10.87	2.25
2,3,4,6,7,10	10	LR<0.01	2.75	95.17	1.28	0.60	0.20	100	79.17	11.11	1.39	8.33	0.00	11.11	2.75
3,4,5,6,7,8,9	15	LR<0.01	2.52	95.23	1.36	0.68	0.20	100	81.69	11.27	1.41	5.63	0.00	11.27	2.52
3,6,7,8,9,10	11	LR<0.01	2.19	95.34	1.41	0.85	0.21	100	83.33	13.89	2.78	0.00	0.00	13.89	2.19
1,2,3,4,5,6,7,8	11	LR<0.01	2.59	95.29	1.19	0.73	0.20	100	77.42	14.52	8.06	0.00	0.00	14.52	2.59
2,5,6,7,8,9,10	22	LR<0.01	2.54	95.70	0.99	0.56	0.21	100	68.85	16.39	3.28	11.48	0.00	16.39	2.54
3,4,6,8,9,10	10	LR<0.00333	2.14	96.23	0.93	0.50	0.21	300	72.22	18.06	9.72	0.00	0.00	18.06	2.14
2,3,6,7,8,9	22	LR<0.01	1.80	97.37	0.38	0.23	0.23	100	66.15	18.46	3.08	12.31	0.00	18.46	1.80
3,4,6,8,9,10	10	LR<0.01	2.38	95.80	1.05	0.56	0.21	100	70.37	20.99	8.64	0.00	0.00	20.99	2.38
1,2,3,9	15	LR<0.00333	2.57	95.30	1.18	0.73	0.22	300	58.33	25.00	16.67	0.00	0.00	25.00	2.57
1,2,3,9	15	LR<0.01	2.76	95.14	1.16	0.73	0.22	100	57.81	29.69	12.50	0.00	0.00	29.69	2.76
2,3,7,8,9	10	LR<0.00333	2.24	95.74	1.12	0.67	0.22	300	62.86	30.00	1.43	5.71	0.00	30.00	2.24
2,3,4,6,7,10	10	LR<1	2.97	94.64	1.49	0.71	0.19	1	58.10	32.38	2.86	6.67	0.00	32.38	2.97
2,3,7,8,9	10	LR<0.01	2.27	95.75	1.10	0.66	0.22	100	59.09	34.09	2.27	4.55	0.00	34.09	2.27
1,3,4,7,9,10	11	LR<1	2.86	94.87	1.43	0.65	0.19	1	55.26	34.21	3.51	7.02	0.00	34.21	2.86
1,2,3,4,5,6,7,8	11	LR<1	2.91	94.51	1.55	0.84	0.19	1	59.05	34.29	1.90	4.76	0.00	34.29	2.91
3,4,5,6,7,8,9	15	LR<1	2.92	94.55	1.49	0.84	0.19	1	55.36	37.50	2.68	4.46	0.00	37.50	2.92
2,5,6,7,8,9,10	22	LR<1	3.08	94.88	1.18	0.66	0.20	1	46.51	41.09	6.20	6.20	0.00	41.09	3.08
3,4,6,8,9,10	10	LR<1	3.04	94.84	1.26	0.66	0.20	1	43.57	45.71	5.00	5.71	0.00	45.71	3.04
3,6,7,8,9,10	11	LR<1	2.76	94.54	1.58	0.92	0.20	1	49.24	46.21	1.52	3.03	0.00	46.21	2.76
2,3,6,7,8,9	22	LR<1	2.65	95.79	0.95	0.41	0.20	1	37.57	49.17	6.63	6.63	0.00	49.17	2.65
1,2,3,9	15	LR<1	3.10	94.74	1.28	0.67	0.20	1	35.67	55.56	4.09	4.68	0.00	55.56	3.10
2,3,7,8,9	10	LR<1	3.01	94.66	1.37	0.75	0.21	1	32.64	60.62	3.11	3.63	0.00	60.62	3.01

Table 7.8 - Summary of 'best' subsets tested for performance by matching the 58 anterior FBI faces. Subsets of PCs were obtained by first varying the number of original landmarks taken for analysis. Various LR thresholds for matches and exclusions were explored; percentages of results obtained for each subset and threshold are given (yes, no, possible, supposed and unverifiable matches) and the percentage of false positive and negative results.

Table 7.8 provides a summary of the performance of thirteen different subsets which were investigated as potential facial matching variables. Displayed are the number of landmarks included in the subset, which of the first ten PCs were included in the subset, the LR threshold to confirm matches and exclusions (§7.3.3), and the percentages of true and false results obtained when the subset was used to find matches in the FBI test data (§7.2.2). Table 7.8 is sorted in order of the 'best' subsets in terms of those which produced the least false positive results. It can be seen that the majority of the top performing subsets were when a threshold of 300 was used to claim facial matches, the false positive rates ranged from just 2% to 11% for the top eight performers. The bottom performing subsets were when a threshold of 1, for the nine worst performers the number of false positive results ranged from 32% to 60%, this is further evidence that a threshold greater than one was required.

7.4.2 Eleven Landmarks

This subsection details how the 'best' subsets of PCs were determined for a subset of eleven landmarks. Other subsets of landmarks were investigated; the results for these are in Appendix D.

7.4.2.1 The Data

To investigate improving the facial matching results obtained from the subset of fifteen landmarks (Appendix D, §14.2.1) the subset of points was reduced to eleven landmarks. The PC loadings for fifteen landmarks (Figures 14.4 and 14.5) showed that certain landmark points varied very little between faces, and therefore were unlikely to contribute much to aid in distinguishing one face from another. Landmarks found to show low variation (a threshold of <0.05 for the PC loadings was used) in the first few PCs were excluded from the subset to reduce the number of landmarks from fifteen to ten; these were (from Table 14.1, Figure 14.1) the centre point of the pupil (left and right), the endocanthion (left and right), the labiale superius. Also, the sublabiale (Table 14.1, Figure 14.1) was added back into the set of landmarks, as it was felt that with the ten landmarks there was no measure of the chin area. This produced a new subset of eleven landmarks which were Procrustes registered (§3.6), a PCA was carried out on the aligned data (§3.7) and all subsets of PCs were investigated in the same way as with

previous subsets of landmarks (Appendix D). Results for subsets of PCs from eleven landmarks are in the following subsections. It was this subset of eleven landmarks which was found to contain the subset of PCs which performed 'best' at facial matching with the available test data.

7.4.2.2 Selecting the 'best' subsets

The methods described in §7.3 were used to search for the best subsets of facial matching variables from the first ten PCs in the eleven facial landmarks described in §7.4.2.1. All subsets of at least size two were used to quantify the LRs for the twenty known facial matches and exclusions, Table 7.7. An average LR for the known matches, average LR for the known exclusions and the MER (§7.3.2) were calculated for each subset. The MERs for all subsets investigated ranged between 6×10^{-7} and 95032. The average match LR ranged between 4.3 and 2586807 and the average exclusion LR ranged between 1.1×10^{-10} and 25.4. Sixty-three subsets out of the 1013 investigated produced false results on average; these were immediately excluded from the list of potential subsets for facial matching.

The remaining subsets were examined and those with a MER close to one were selected as potentially good for facial matching. The 'best' nine subsets (the nine with MER closest to one) are displayed in Table 7.9, i.e. the 'best' in terms of being equally as good at matching known matches as excluding known exclusions. Three of the nine subsets in Table 7.9 were selected as the best three for eleven landmarks; these were ones which provided strong evidence to support both hypotheses (i.e. a large LR for matches and a small LR for exclusions). These best three were subsets 6 (PCs 1, 3, 4, 7, 9 and 10), 7 (PCs 3, 6, 7, 8, 9 and 10) and 8 (PCs 1, 2, 3, 4, 5, 6, 7 and 8), these were evaluated for performance using the PCs they contain to quantify matches in the fifty-eight FBI anterior test faces (§7.2.2).

Subset number	PCs	Average Exclusion LR	Average Match LR	MER
1	2,4	0.1807	5.3	0.9516
2	1,5,10	0.0188	51.7	0.9695
3	2,3,4,5,7	0.0003	3443.3	0.9699
4	2,3,4,6,7,8,9	0.0000	34879.4	0.9831
5	5,6	0.1848	5.4	0.9961
6	1,3,4,7,9,10	0.0002	5070.7	1.0051
7	3,6,7,8,9,10	0.0000	40721.6	1.0310
8	1,2,3,4,5,6,7,8	0.0000	175067.7	1.0332
9	1,2,3,4,6,9	0.0009	1097.2	1.0364

Table 7.9 - 'Best' subsets in terms of MER and average LRs for known matches and exclusions, 11 landmarks.

7.4.2.3 Subset Performance

The best three subsets of the eleven landmarks in terms of the selection criteria set in §7.3 were subsets 6, 7 and 8, see Table 7.9 (§7.4.2.1). To explore how effective these subsets were at matching faces the PCs in the subsets were used in LR calculations to carry out facial comparisons on all pairs of faces in the FBI anterior dataset (§2.7.2). This meant the comparison of 1653 pairs of the fifty-eight faces, where some pairs were known matches and some known exclusions. The percentages of correctly classified matches and exclusions were examined for several levels of LR. Following the large range of LR results seen for the known matches in §7.2.1.1 and §7.2.2.1 a number of different thresholds were investigated to find the optimum LR threshold for quantifying matches and exclusions (§7.3.3). It was thought that although any LR greater than one was more in favour of a match than it was an exclusion this threshold was not sufficiently large enough to encompass false matches or exclusions that have LRs close to one.

Tables 7.10-7.12 show the percentages of true matches (LR>1 and 'Yes') and exclusions (LRs<1 and 'No') and also false positive (LR>1 and 'No') and negative (LR<1 and 'Yes') results quantified from the LRs produced by subsets 6, 7 and 8 respectively. There were also some 'supposed', 'possible' and 'unverifiable' matches, where the FBI notes supplied insufficient information to confirm either hypothesis, full details given in §2.7.2. We have not used the distinction between 'supposed' and 'possible', as these are only a word of the data provider, these results could be further investigated however this was not pursued here.

Tables 7.10-7.12 show that when considering all results for matches a threshold of LR>1 produced very high false positive results (34.2 – 46.2%). Increasing the threshold to LR>100 between 5 – 14.5% results were false positives and increasing again to LR>300 produced only 2 – 10.7% false positive results. Similarly for the exclusion results when considering all LR<1 the number of false negatives was low compared to the false positives (2.8 – 2.9%). Decreasing the exclusion threshold to LR<0.01 reduced the false negatives to between 2.2 – 2.6% and decreasing again to LR<0.00333 barely improved the results to produce between 2 – 2.6% false negatives. The performance in terms of confirming exclusions was much better than for confirming matches. The threshold for quantifying matches was increased further to LR>400, however the optimum threshold for confirming matches was found to be LR>300 (further details in §7.3.3)

Using a threshold of LR>300 for quantifying matches and LR<0.00333 for quantifying exclusions it was found that subset 6 (Table 7.9) performed the best with a very low false positive rate of 2% (Table 7.10). Even though this was the 'best' found subset it only succeeded in correctly identifying 43 out of the 107 known matches and 1368 out of the 1499 known exclusions in the FBI anterior data (§2.7.2, Table 7.6), chapter 8 investigates factors which may influence this.

Threshold	Yes	No	Possible	Supposed	Unverifiable
LR>1	55.26	34.21	3.51	7.02	0.00
LR>100	86.67	5.00	1.67	6.67	0.00
LR>300	87.76	2.04	2.04	8.16	0.00
LR<1	2.86	94.87	1.43	0.65	0.19
LR<0.01	2.55	95.59	1.03	0.62	0.21
LR<0.00333	2.37	95.81	1.05	0.56	0.21

Table 7.10 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 6 (PCs 1, 3, 4, 7, 9 and 10 from 11 landmarks)

Threshold	Yes	No	Possible	Supposed	Unverifiable
LR>1	49.24	46.21	1.52	3.03	0.00
LR>100	83.33	13.89	2.78	0.00	0.00
LR>300	87.69	10.77	1.54	0.00	0.00
LR<1	2.76	94.54	1.58	0.92	0.20
LR<0.01	2.19	95.34	1.41	0.85	0.21
LR<0.00333	1.95	95.75	1.30	0.79	0.22

Table 7.11 – Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 7 (PCs 3, 6, 7, 8, 9 and 10 from 11 landmarks)

Threshold	Yes	No	Possible	Supposed	Unverifiable
LR>1	59.05	34.29	1.90	4.76	0.00
LR>100	77.42	14.52	8.06	0.00	0.00
LR>300	82.35	7.84	9.80	0.00	0.00
LR<1	2.91	94.51	1.55	0.84	0.19
LR<0.01	2.59	95.29	1.19	0.73	0.20
LR<0.00333	2.61	95.24	1.21	0.74	0.20

Table 7.12 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 8 (PCs 1, 2, 3, 4, 5, 6, 7 and 8 from 11 landmarks)

7.4.2.4 Relating the Results back to the Matching Variables

To visualize what was happening in terms of the variation being represented in the variables (PCs) in the 'best' found subset for facial matching (§7.4.2.3) plots of the directional PC loadings were examined (Figures 7.5 and 7.6). Plots show the average location of the eleven landmarks in the subset (points) with arrows drawn in the direction of variation. Landmarks with large variation (>0.05) are indicated by black solid arrows. The best performing subset (6) contained PCs 1, 3, 4, 7, 9 and 10 from the eleven facial landmarks (§7.4.2.1). Table 7.13 lists the areas of the face that show variation for each PC included in the 'best' found subset for facial matching.

PC	Area of face showing variation
1	Eyes and chin – far extremities of the face
3	Width of the lips
4	Width of the lips and thickness of lower lip
7	Eyes
9	Width of nose, width of lips
10	Width of the lips and thickness of upper lip

Table 7.13 - PCs included in the 'best' found subset for facial matching and the facial variation represented by each of these PCs

Of the PCs which were not included in the 'best' subset, PCs 2, 6 and 8 all represented variation in parts of the face which had already been explained in the PCs which were included (Table 7.13). PC 5 represented variation in the distance between the lower lip and the chin, the fact that this variable was not included and is not explained in the 'best' subset variables suggests that this measurement is not a useful one to use in facial identification.

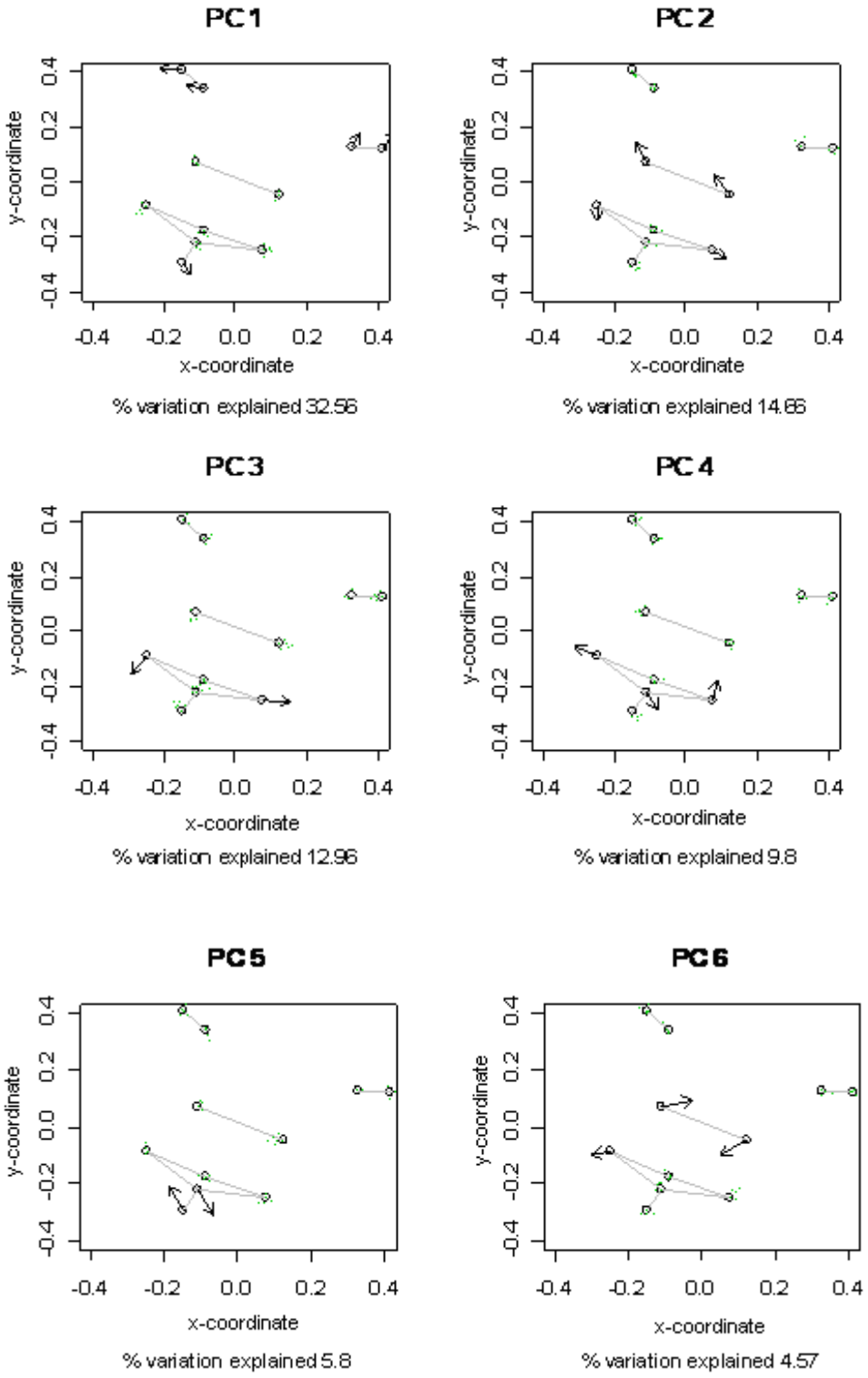


Figure 7.5 - Plots (PCs 1-6 from eleven landmarks) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows.

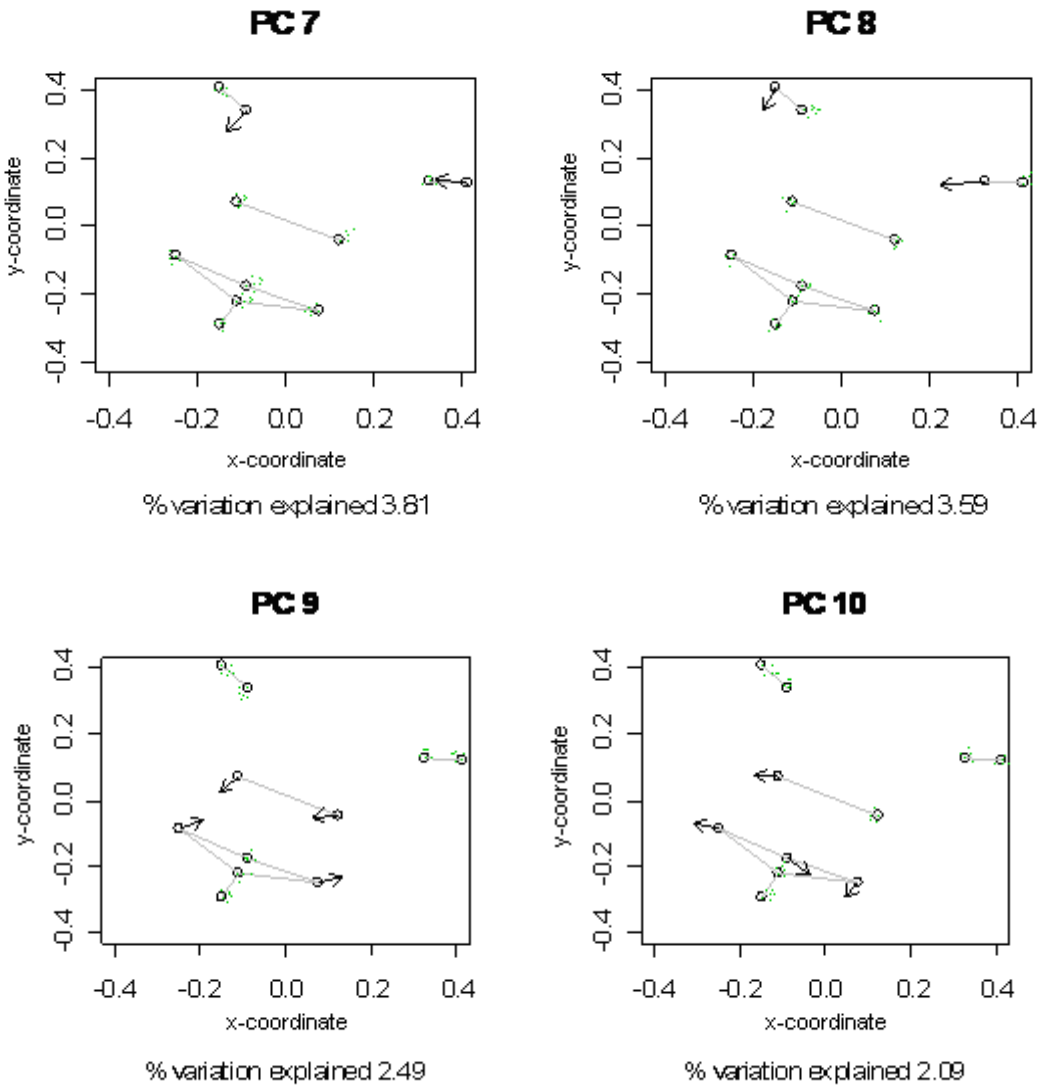


Figure 7.6 - plots (PCs 7-10 from eleven landmarks) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows.

7.4.3 Robustness of the 'Best' Subset

One way of evaluating the robustness of the 'best' found subset (§7.4.2.3) is to perturb the covariance matrix of the background database. It is unknown how much the covariance matrix should be perturbed to simulate real life facial variation outside the sample collected, jack-knifing, bootstrapping or cross-validation could be carried out to overcome this. An easier method to adopt here was simply to drop some faces from the background database to see how this affected the matching results obtained from the 'best' found subset.

Duplicate measurements from 1286 faces were originally included in the background database, as described in §2.4. The 'best' subset of PCs 1, 3, 4, 7, 9 and 10 for eleven landmarks (§7.4.2.3) was used to quantify the number of facial matches and exclusions in the fifty-eight FBI anterior test faces, using a threshold of LR>300 to quantify a match and LR<0.00333 to quantify an exclusion (§7.3.3). This quantification was repeated when randomly excluding a certain number of faces from the background data, Table 7.14. The number of faces to be excluded was increased from ten up to seventy-five; for each number excluded the process was repeated three times and matching results were obtained. The average percentages of false results (both false positive and false negative) over the three triplicate measurements were examined. Table 7.14 displays the average percentages of matches and exclusions for each LR threshold and number of faces excluded from the background data.

Figure 7.7 shows the percentage of false results obtained by the 'best' subset as the number of faces excluded from the background data was increased. Included are the original results obtained for the 'best' subset i.e. excluding zero faces from the background data. For exclusion results it was seen that randomly dropping ten faces from the background data increased the false exclusion results by almost ten percent, subsequent increases in the number of faces dropped showed a more gradual increase in the false exclusion rate up to 15.5% for dropping seventy-five faces from the background data. For facial matches randomly dropping up to twenty-five faces from the background data appeared to have little effect on the false match results which ranged between 1% and 2%. Dropping more than thirty-five faces increased the false match rate to 6% and dropping more than sixty faces showed a dramatic increase in the false match rate, which rose to 25% for dropping seventy-five faces from the background data. To put this into context, dropping seventy-five out of 1286 faces from the background data is excluding just six percent of the data. If this has such an effect on the false result rates then it casts doubts on the performance of the 'best' subset. A more in depth study would be beneficial, perhaps taking a completely different set of data and performing the investigation into the most appropriate subset of facial variables to use to maximise facial matching results.

No. Faces Excluded from background data	Threshold	Average % Match	Average % Exclusion
0	<0.00333	2.37	95.81
10	<0.00333	11.09	88.45
15	<0.00333	9.54	90.07
20	<0.00333	10.42	89.15
25	<0.00333	9.53	90.08
35	<0.00333	12.43	87.09
50	<0.00333	11.07	88.47
60	<0.00333	14.85	84.60
75	<0.00333	15.47	84.14
0	>300	87.76	2.04
10	>300	97.30	2.28
15	>300	97.51	1.24
20	>300	97.98	0.81
25	>300	97.72	1.14
35	>300	93.60	6.40
50	>300	93.99	5.10
60	>300	85.55	14.45
75	>300	73.48	25.30

Table 7.14 – Sensitivity of the 'best' subset; % of matches and exclusions for fifty-eight FBI test faces when a number of faces were randomly excluded from the background database.

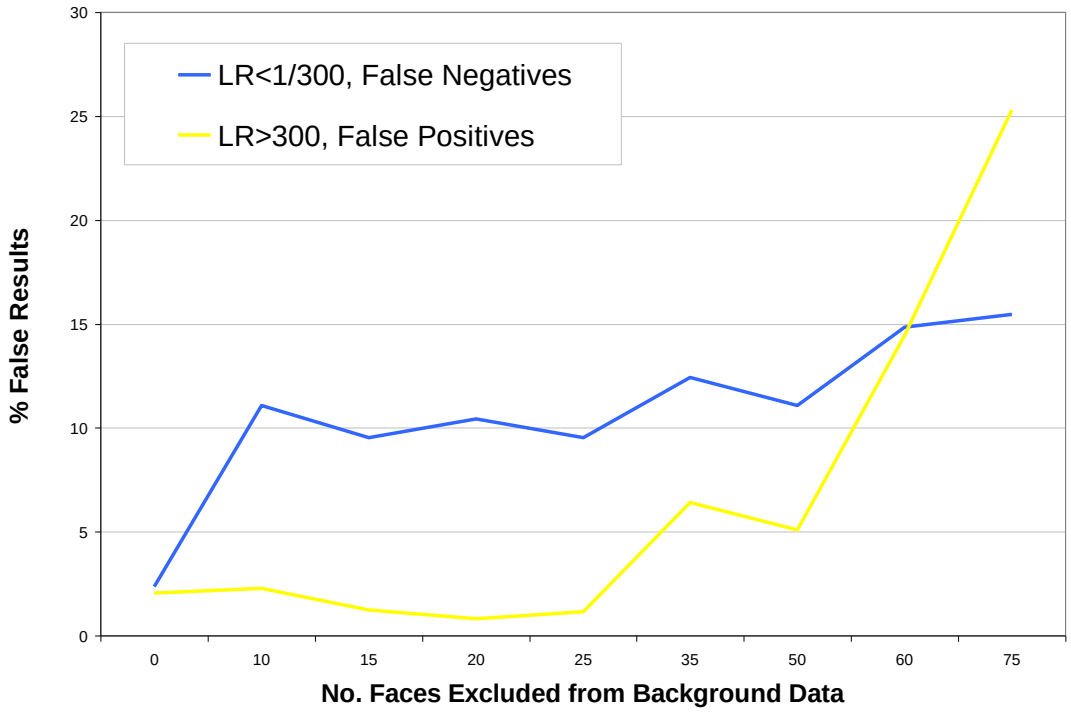


Figure 7.7 – Percentage of false positive and negative results for the fifty-eight FBI test faces. Using the 'best' subset of matching variables and excluding faces from the background database to check sensitivity of results.

7.5 Summary

The work throughout this chapter has extended the Aitken and Lucy (2004) multivariate normal likelihood ratio method for evaluating evidence (§3.8.4). The method was found to work well with the facial data (§7.2), however certain extensions had to be made to accommodate the large number of variables in the dataset (§7.3, §7.4). Firstly, the Procrustes tangent coordinates were transformed onto principal components and the PC scores were used as the matching variables rather than the original variables (§7.2.1). A further extension was that a subset of the variables (PC scores) was found to perform better than taking all or the first few. This was because increasing the number of PCs cumulatively did not necessarily mean the LR monotonically got larger, as the inclusion of certain PCs weakened the results (§7.2.3). Two test datasets were examined, the first simply to check whether the method worked and the second to apply the method to some real life case data from the FBI (§7.2.2), a number of which were known matches.

The LR method with just five matching variables proved to produce a substantial number of false results when tested with the anterior FBI data (§7.2.2.1). Increasing the number of matching variables to twenty produced much fewer false results, although the method only identified around half of the known matches from the whole test data (§7.2.2.1). Possible factors that may determine which known matches are recognized and which are not include facial expression and the rotation of the face to camera, these issues are investigated in chapter 8.

It was found that checking the facial comparison data fitted the multivariate normal model for the background population was an important aspect of the LR procedure (§7.2.2.2). This was important because the checks highlighted any erroneous landmark configurations which could be investigated. It was seen that when a configuration had just one landmark point misplaced it lay far from the multivariate normal model (§7.2.2.2). Model checking identified problems with landmark positions, which were either corrected or excluded from analyses; this ultimately improved the match results obtained (§7.2.2.3). It is recommended that before any facial comparisons are carried out the comparison data should be added to the background data and the data checked for errors using a Chi-squared quantile plot (Figure 7.1). Such a plot shows how well the data follow a multivariate normal distribution and data points deviating far from the model can be examined, as these could influence matching results considerably.

The choice of how many matching variables to use has been addressed (§7.2.1.1, §7.2.2.1 §7.2.3), and also which variables were the 'best' in terms of producing 'good' evidence for matches (§7.3). With the test data increasing the number of matching variables generally increased the percentage of correct results obtained, however not monotonically so, some variables appeared to reduce the quality of known facial matches (§7.2.3). It was therefore decided that a subset of PCs would be the most appropriate variables to use for facial matching (§7.3.1).

Several criteria were chosen to determine the 'best' subsets, §7.3.2, §7.3.3. Firstly the results obtained from matching on a particular subset of variables were required to be relatively unbiased for either hypothesis. In other words the subset should be equally as good at picking out matches as it was at excluding exclusions. A basic measure to address this issue was the match/exclusion ratio (MER) which compares the average match LR to the average exclusion LR, good subsets were those with an MER close to one (§7.3.2, §7.4.2.2). In addition to this the actual values for the average match LR and average exclusion LR were examined, obviously the better subsets were those which had greater support for either hypothesis after already measuring the bias of the subset (§7.3.2, §7.4.2.2). All potentially good subsets which fulfilled these selection criteria were examined for performance by using the variables to search for matches in the fifty-eight FBI test faces (§7.4.2.3 and Appendix D). The number of true and false results indicated how well each subset performed.

A 'best' subset of matching variables was found, the match results for this subset were 98% true matches (§7.4.2.3). The subset was impartial for either hypothesis H_p or H_d (as the MER of the best subset was very close to one), however only around half of all known matches in the sample were picked up (§7.4.2.3). Reasons for these missed matches are explored in chapter 8 to explore why certain facial pairs are less likely to match than others.

When subsets were evaluated for performance using a threshold of LR>1 to confirm facial matches the number of false positive results ranged from 32% to 60%, §7.4.1. Increasing the match threshold reduced the numbers of false positives to improve results, §7.3.3. A threshold of LR>300 found the optimum true results, increasing the threshold further than this did not improve results to a great extent (§7.3.3). In a courtroom it would be up to a prosecution or defence team to decide what they deem to

be a suitable threshold for a ‘match’ LR, a good thing to do would also be to quote the percentage of true and false results obtained when using such a threshold (e.g. here a threshold of 300 to confirm matches and 1/300 to confirm exclusions gives false results 4% of the time).

Obviously using only twenty facial comparison cases to quantify the ‘best’ subset of matching variables is a limitation, ideally we would want to use all possible comparisons from the FBI anterior data where the known matches and exclusions exist, however computationally this would be a large task. The program to compare all possible subsets for the twenty comparisons took around two hours to run.

The ‘best’ subset of eleven landmarks (§7.4.2.1-§7.4.2.3) was derived from knowledge and experience of the landmark data. Prior knowledge and judgement of landmark location, ease of landmark placement and determination in the anterior facial view were all taken into account when choosing the subset of facial landmarks. The strength of results (average LRs for matches and exclusions) along with the MER (§7.3.2) were then used to evaluate a subset of PCs from the subset of facial landmarks.

The sensitivity of the ‘best’ subset was checked by randomly dropping a number of faces from the background data, in order to change the covariance structure, and repeating the matching on the fifty-eight FBI test faces (§7.4.3). The false results were examined and it was found that dropping just 6% of the background data increased the false exclusion rate from 2% to 15% and the false match rate from 2% to 25%. When such increases were seen when dropping this small amount of data it was thought that taking a completely different set of data was likely to require a whole new investigation into the most appropriate subset of facial variables to use to maximise facial matching results.

8 Performance Evaluation on Selected Subsets and Further Data

8.1 Introduction

In this chapter we investigate a variety of factors which influence the success of the core procedure. Modifications to the core procedure have been suggested, these may be required to reduce the unwanted influence. Influential factors were differences in observers measuring landmarks, different images and the angle of presentation in the images both around a vertical and a horizontal axis perpendicular to the line of view (and a combination of both of these). Possible extensions to the method that were considered to improve matching results were to include observer error in the model for calculating LRs (§8.3.1.2) and to use averages of sets of facial landmarks (§8.3.1.1, §8.6).

Data which were obtained from a different source to that analysed so far were examined for matches. This enabled an assessment of how well the developed methods handle external data obtained under different conditions, such as that which would be acquired in a real life situation. The performance of the method was inspected more closely by relating results back to the source images (given in Confidential Appendix C) to explore possible factors which could influence the outcome of results.

Three different datasets containing known matches were used to assess the quality of results obtained from the core method. These were data obtained from twins and matched controls (§2.7.5), multiple images obtained from a different data source (§2.7.6) and multiple images of an FBI agent (§2.7.2).

The performance of the method at matching twins (§8.2) using three sets of twins and controls from the Geometrix® database (§2.7.5) was assessed. The second dataset examined contained multiple images of faces obtained using an ordinary 2D digital camera as part of a separate study carried out by the IDENT project (Evison and Vorder Bruegge, 2008) (§2.7.6). These data were collected by three different observers from those who collected the background data. The LR results were analysed to see whether the method worked for matching different images of the same face when the data were

collected by one observer. The performance of matching was also investigated for data collected by two different observers from one facial image (§8.3.1).

The method was also used to match a third dataset consisting of multiple images of one face (agent Vorder Bruegge) from the FBI anterior dataset (§2.7.3, §8.4). The multiple images were known to be taken at varying times and consisted of the subject posed in different facial positions and expressions. The investigation of this dataset was primarily concerned with factors which influenced whether the method picked up a known match or not, as it was found in §7.4.2.3 that the 'best' subset only found 43 out of 107 of the known matches in the anterior FBI data (§2.7.2). It was revealed that one particular factor influencing match results could be the angle of presentation of the subject face to camera during the image capture procedure. Section 8.5 uses the available 3D data (§2.4) to simulate the rotation of ten faces and compare these with the original ten faces to investigate the effect of rotation on the matching results. Faces were rotated around a vertical or horizontal axis perpendicular to the line of view (or a combination of both of these).

Section 8.6 suggests an extension to the method which looks at taking the average of several different sets of facial landmarks and using this average to match faces.

8.2 Twins from the Geometrix® database

Visually twins look very similar; it is sometimes difficult for the human eye to tell them apart. The modified LR facial matching procedure (§3.8.4, §7.3, §7.4) using the 'best' subset of matching variables (§7.4.2.3) was used to try and distinguish between twins. The background data collected with the Geometrix® scanner was known to contain three sets of twins (§2.7.5). Additionally three pairs of non-related controls were taken to match the sex and age of the twins. All images in this subset are displayed in the Appendix C, Figures 13.24 – 13.29. The subset of twin and control data which was complete for the eleven landmarks in the 'best' subset (§7.4.2.1) was taken from the facial database. The interest lay in how strong the evidence for a facial match was for the identical twins in comparison with the non-identical and matched controls using the 'best' subset of matching variables.

The background data consisted of duplicated measurements for 2640 faces from the Geometrix® database (§2.4) and fifty-eight faces from the FBI anterior data (§2.7.2), eleven facial landmarks (§7.4.2.1) from the faces were Procrustes aligned and a PCA was carried out to obtain PC scores. The PC scores for PCs 1, 3, 4, 7, 9 and 10 (§7.4.2.3) were taken as the six variables to carry out the matching algorithm for calculation of the LR for strength of evidence for a match for the six pairs of faces in the twins and controls test data. The resulting LRs are displayed in Table 8.1.

Twins?	Face i	Face j	Likelihood Ratio
Non-identical	1870	1871	4.73E-16
Identical	1860	1861	6.92E-10
Identical	1459	1460	1.38E-07
Control	2553	2540	6.55E-08
Control	2221	927	2.94E-08
Control	472	90	0.000529

Table 8.1 - Facial matching results comparing three sets of twins and age and sex-matched controls

Table 8.1 shows the LR results for facial comparisons for the three sets of twins and the three sets of controls. There is more evidence to exclude all six comparisons than there is to suggest that any of them are a match. On examination of the images for the twins (Confidential Appendix, Figures 13.24 – 13.29) the non-identical set do not even visually look alike, so it is not surprising that they were excluded. The identical twins had evident facial similarities; however it was still easy to tell them apart through visual examination. For both sets of identical twins one of the pair appeared to be smiling more than the other and also the head position to camera appeared different. Both of these factors may influence the accuracy of matching results and will be investigated further in the following sections (§8.3 - §8.5). Also, for the data collection multiple observers placed the two measures of landmark points on the twins' images. The subset of variables used for facial matching (§7.4.2.3) was deemed the 'best' on the basis of its performance on the FBI anterior test data (§2.7.2), this data was collected by just one observer. This is an obvious limitation when using the subset to match data collected by multiple observers.

8.3 Other Data from Multiple Images of Like Faces

Some other data from a different source were obtained to test further the matching method (§3.8.4, §7.3, §7.4) and 'best' subset of matching variables (§7.4.2.3). The data are described fully in §2.7.6 and the images can be seen in Appendix C, Figure 13.31. The performance of the LR matching procedure was assessed in terms of how well different images of the same face matched and also how well configurations from different observers placing landmarks on the same photograph matched. The data here were collected using a different piece of software and by three different observers than were used to collect the data for the main database (§2.4), an important check to test the method did not depend on the software used to collect data or on observers opinion.

The 135 configurations from the multiple image data were added to 4544 configurations from the Geometrix® facial database (duplicate landmark configurations for 2272 faces) and 116 configurations from the FBI anterior test data (duplicate landmark configurations for fifty-eight faces) to form the background data used to calculate LRs (§3.8.4, §7.2). The 'best' subset of variables (§7.4.2.3) was used to carry out the facial comparisons; PCs 1, 3, 4, 7, 9 and 10 of the Procrustes tangent coordinates from the eleven facial landmarks (§7.4.2.1).

The three triplicate measures taken by each observer on each photograph were counted as three different sets of measurements to compare. Therefore there were forty-five sets of different measurements to compare with one another, when actually there were only five different faces. Treating the data in this way allowed the assessment of how well different observers' data matched, as well as how well different photographs of the same faces matched. 990 (pair wise comparisons for forty-five faces) facial comparisons were carried out using the LR method (§3.8.4, §7.3, §7.4). Applying an ANOVA model here would allow the consideration of the various sources of variation; this could be useful in assessing relative likelihoods of different sources (both images and measures), however again this would force the formality of a statistical test on the court. Additionally, more data transformations would have to be carried out to make valid distributional assumptions.

8.3.1 Results

The results were analysed twice, once for data collected from two different images by the same observer and once for data collected from one image by two different observers. The results for the matches and exclusion conclusions obtained for the known facial matches are summarized in Tables 8.2 and 8.3. A threshold of LR>300 was used to determine a facial match and LR<0.00333 to determine a facial exclusion. Those with LRs in between these thresholds are regarded as 'insufficient evidence' to confirm either a match or exclusion. Table 8.2 displays results for comparing landmark data from two different images of alike faces using the same observer to collect the data from both images. Table 8.3 displays results for comparing two different landmark configurations taken from the same image by two different observers. All results from both Tables 8.2 and 8.3 should conclude that each face matches with the nine other faces it was compared with.

Face	Matches (LR>300)	Insufficient Evidence	Exclusions (LR<0.00333)	LRs in range	
1	2	7	0	9.86E-01	6.09E+02
2	2	6	1	6.25E-04	4.88E+03
3	2	4	3	2.69E-56	2.31E+04
4	5	4	0	4.10E-02	7.23E+04
5	5	4	0	1.31E-02	2.73E+04

Table 8.2 – Number of facial matches evaluated using the 'best' subset of matching variables (§7.4.2.3) with the LR procedure (§3.8.4, §7.3, §7.4). Results are for known facial matches comparing two different photos of the same face, data were collected by one observer.

Table 8.2 shows that the number of facial matches found by the 'best' subset of variables (§7.4.2.3) differed considerably for the five different faces analysed. Out of the nine comparisons for each face either two or five matches were ascertained. One and three exclusions were made respectively for faces 2 and 3. All other results in Table 8.2 showed insufficient evidence to support either hypothesis.

Face	Matches (LR>300)	Insufficient Evidence	Exclusions (LR<0.00333)	LRs in range	
1	2	2	5	3.50E-09	1.10E+03
2	3	6	0	2.41E-01	6.57E+03
3	1	3	5	4.14E-71	5.92E+02
4	3	0	6	9.11E-08	8.92E+03
5	0	4	5	8.88E-16	4.52E+01

Table 8.3 - Number of facial matches evaluated using the 'best' subset of matching variables (§7.4.2.3) with the LR procedure (§3.8.4, §7.3, §7.4). Results are for known facial matches comparing the data from one photograph collected by two different observers.

Table 8.3 shows that for facial comparisons of data collected by two different observers from one photo many of the results gave sufficient evidence to exclude the two faces as a match. A lot of results also supported neither hypothesis. Comparing these results with Table 8.2 implies that the LR procedure for facial matching does not perform as well on data collected by a number of different observers as it does on data collected from multiple images by the same observer, in other words the error associated with image capture is less than that associated with landmark capture. The Geometrix® facial landmark database (§2.4) was collected by six different observers, this is the background data used for the LR calculations. Two possible extensions to the method are explored in the following subsections. The first simply takes averages of the measurements obtained by different observers, §8.3.1.1. The second adapts the multivariate normal model to include variation attributed to observer error would improve the LR facial matching procedure, this is explored in §8.3.1.2.

8.3.1.1 Averaging Comparison Data Collected by Different Observers

To improve the poor matching results seen with data collected by different observers (§8.3.1) a simple option would be to take an average of the facial landmark coordinates across observers. To investigate this the average of each different landmark measurements was taken for each pair of observers, i.e. the average of the first measure of observers 1 and 2, 1 and 3 and 2 and 3, then the same for the second and third measurements. The facial comparisons were run using these averaged data; the results are in Tables 8.4.

Comparing Tables 8.3 and 8.4, the numbers classified as matches, exclusions and insufficient evidence for either hypothesis were identical for the original data and the pair wise observers' averages. The values for the LR statistics changed slightly, though the strength of evidence was still about the same, e.g. a match of 10^3 or an exclusion of 10^{-8} for face 4, Table 8.4. So, taking average landmark configurations over pair wise observers did not alter any results in terms of the number of matches and exclusions obtained.

Face	Matches LR>300	Insufficient Evidence	Exclusions LR<0.00333	LRs in range	
1	2	2	5	3.17E-09	1.10E+03
2	3	6	0	8.61E-02	3.98E+03
3	1	3	5	3.67E-71	5.84E+02
4	3	0	6	8.05E-08	8.93E+03
5	0	4	5	7.83E-16	4.53E+01

Table 8.4 - Number of facial matches evaluated using the 'best' subset of matching variables (§7.6.1.2) with the LR matching procedure (§3.8.4, §7.2). Results are for known facial matches comparing the data from one photograph; data were collected by two different observers and then averaged.

8.3.1.2 Extending the Model to Include Observer Error in the Background Data

The LR method used for all previous facial comparisons (chapter 7, Appendix D, §8.2) assessed the within face and between face covariance structures in the background data, as explained in chapter 3 (§3.8.4). It could be argued that a more appropriate model for the data would be to nest the factor observer into the within and between face covariance matrices, as not always the same observer placed both duplicate measures on the faces on the background data (§2.4). This was not pursued in detail, instead a simplified approach of taking averages was used (§2.7.6, §8.4).

Table 8.5 shows the results for using the simplified extended LR model to compare the multiple images data. When compared with the previous results (Tables 8.3 and 8.4) Table 8.5 shows that extending the model did reduce the number of false exclusions that were ascertained, however it also decreased the number of correct match results. The greatest difference with including observer error in the model was that there were many more results that had insufficient evidence to support either hypothesis. This was to be

expected, using multiple observers would inevitably produce more diverse information than if the same observer collected all the data.

Face	Matches LR>300	Insufficient Evidence	Exclusions LR<0.00333	LRs in range	
1	0	9	0	1.08E-02	1.50E+02
2	3	6	0	1.78E+01	9.56E+02
3	1	4	4	5.55E-49	5.78E+02
4	3	6	0	5.63E-01	1.34E+04
5	0	6	3	4.42E-04	7.26E+01

Table 8.5 - Number of facial matches evaluated using the 'best' subset of matching variables (§7.4.2.3) with the LR matching procedure (§3.8.4, §7.3, §7.4) extended to include observer error in the data model. Results are for known facial matches comparing the data from one photograph collected by two different observers.

8.4 Multiple Images of Agent Vorder Bruegge

Part of the FBI anterior test data (§2.7.2) consisted of multiple images of an FBI agent (§2.7.3) who was involved with the IDENT project (Evison and Vorder Bruegge, 2008) and the collection of the Geometrix® data (§2.4). The eleven anterior landmarks in the 'best' subset (§7.4.2.1) were placed on fourteen facial images of agent Vorder Bruegge, (§2.7.3, Appendix C Figure 13.30); these along with all the other FBI anterior data were added to the Geometrix® data to use as the background database for the facial matching calculations. GPA was carried out to align all faces and a PCA was carried out on the tangent coordinates of the aligned background database to get the matching variables found to perform the 'best'; PCs 1, 3, 4, 7, 9 and 10 (§7.4.2.3). These matching variables were used to quantify matches between all pair wise comparisons of the fourteen test images. Obviously the real results were that all faces matched with each other, interest lay in whether the method would produce results that confirmed all the matches. The optimum thresholds of LR>300 and LR<0.00333 were used to confirm matches and exclusions respectively.

Table 8.6 shows the LR results for the number of matches and exclusions found for each of the fourteen images of agent Vorder Bruegges face using the 'best' subset of matching variables. Although all images were known to match with each other all had at least two exclusion results. There were also a number of cases which had insufficient

evidence to support either a match or exclusion. The images with the most exclusion results were numbers 8 and 9 (Appendix C, Figure 13.30) where the face was smiling; twelve and eleven of the thirteen comparisons were excluded respectively. Interestingly when the two smiling images were compared to each other there was insufficient evidence to support either a match or exclusion. On inspection of the two images it appeared as though one was more rotated away from the anterior position than the other.

Images 3 and 5 show the subject wearing glasses (Appendix C, Figure 13.30); these images did not appear to produce any worse match results than images without glasses. On inspection it was clear that the glasses did not apparently mask the position of any of the landmarks placed, this may not be the case for all glasses. It is possible that the position of the landmarks seen through the glass could be subject to a greater error but this only affects four landmarks.

From these results there was evidence that facial expression and the rotation of the facial position to camera had an effect on the matching results. This seems highly likely as the model for face shape was based on background data where faces held a neutral expression; therefore there is no information on facial variation for varying facial expressions.

The issue of how the position of the subject's head to camera affects the matching results is investigated in the following section. The rotation of a set of faces is simulated with the available 3D data (§2.4) and the original faces are compared to the rotated faces.

Face	Matches LR>300	Insufficient Evidence	Exclusions LR<0.00333	LRs in range		Comments
1	1	9	3	2.39E-38	391.6584	old photo
2	7	4	2	1.73E-30	1223.304	
3	8	3	2	4.86E-29	13305.51	glasses
4	8	3	2	9.57E-30	2545.135	old photo
5	8	3	2	4.58E-27	2545.135	glasses
6	9	2	2	8.34E-28	13305.51	
7	1	6	6	7.27E-07	53053.85	subject rotated
8	0	1	12	1.95E-39	0.00499	smiling
9	1	1	11	9.64E-10	53053.85	smiling
10	9	1	3	1.62E-33	12293.09	
11	6	4	3	1.66E-33	52134.3	
12	9	1	3	1.95E-39	9279.646	
13	8	2	3	2.06E-38	7631.597	
14	7	4	2	1.33E-29	52134.3	

Table 8.6 – Number of facial matches and exclusions obtained through evaluating LR s using the 'best' subset of matching variables (§7.4.2.3) to compare fourteen images of agent Vorder Bruegge (Confidential Appendix Figure 13.30).

8.5 The Affects of Head Orientation on Matching Results

Figure 8.1 - Two anterior views of the face, images from the Geometrix® scanner central camera pair

An observation made during examination of the anterior images from the Geometrix® data (§2.4) was that in the two anterior images available, which were taken from a centre camera pair, the same face looked quite different in the two camera views, Figure 8.1. This was apparent despite the fact that the angle between cameras was small

(approximately nine degrees, Figure 8.2). For 3D data (§2.4) these differences in head orientation can be corrected for through the Procrustes alignment of the 3D coordinates. When we deal with 2D images, as we have established will be the case for the majority of facial comparisons, any rotation of the head which is not in the *xy* plane can not be corrected for in Procrustes alignment without the third dimension.

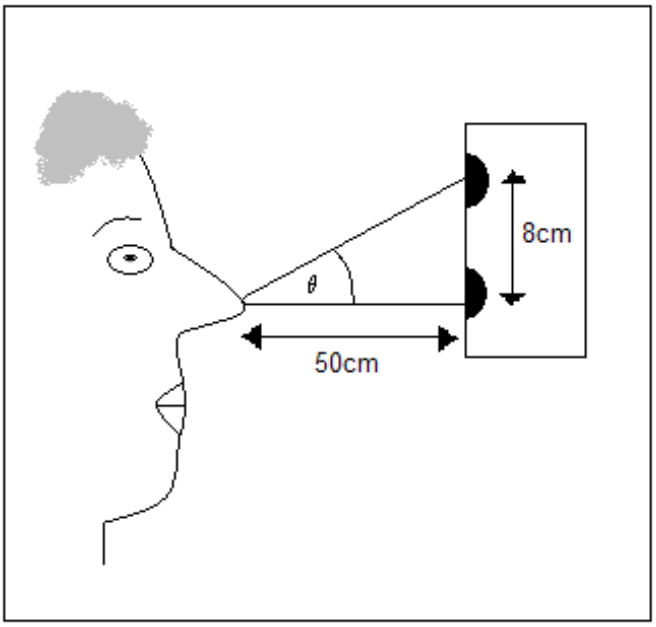


Figure 8.2 Estimation of angle between two central pair cameras

This raises the question of what tolerance in facial angle to the camera could be allowed for two different 2D images of the same face to still be declared a match. The following section explores whether 2D images, as would be obtained from real life situations, would be sufficient for accurate facial matching when the angles of the subject to camera are unknown. We look at how critical the subject's facial angle to camera is for obtaining accurate match results using the LR method (§3.8.4, §7.3, §7.4) with the 'best' subset of matching variables (§7.4.2.3).

8.5.1 The Data

A small subset of ten faces was arbitrarily selected from the Geometrix 3D database (§2.4). The set of eleven most reliably placed and most variable facial landmarks (§7.4.2.1) were selected for comparing the images, these landmarks were rotated by multiplying data by a rotation matrix. A number of different rotations were simulated, the tilting of the head either upwards or downwards about the *x*-axis and the rotation of the head left and right about the *y*-axis. The tolerance for each of these angles was investigated by testing whether the original face matched with the rotated face for a variety of different angles. The matching was carried out using the LR method (§3.8.4, §7.3, §7.4) with the 'best' found subset of matching variables (§7.4.2.3) and a threshold of LR>300 to quantify a match (§7.3.3).

The background data for the LR calculations were duplicated landmark coordinates from 1306 faces from the Geometrix database, plus duplicated measurements for the subset of ten faces and the ten faces that were generated from rotating these. A match threshold of 300 was taken, i.e. if a likelihood ratio (LR) of greater than 300 was obtained then the faces were assumed a good match.

8.5.2 Translation to X, Y, Z Axes for Rotation

The coordinate system for the scanner (§2.4.2), which was used to obtain coordinates for the facial landmarks, was based on the tip of the nose being point (0, 0, 0). So, to rotate about the *x*-axis the pivot of the rotation would be a horizontal axis going through the tip of the nose. Observing a subject tilting the head upwards and downwards it is clear that this axis is not the most appropriate in terms of the position of pivot by which the head tilts. Instead the axis of rotation to rotate the head upwards or downwards was taken as the line through the two landmark points left subaurale and right subaurale (Table 2.1). These points are at the base of the ears, and were chosen as the most appropriate of the points available in terms of being closest to the position of pivot by which the head tilts. The left and right rotations about the *y*-axis were carried out through an axis of rotation through the points of the glabella and the pogonion (Table 2.1), points down the facial midline. It is thought that actually this left/right pivot is more likely to be through the centre of the top of the head rather than in line with the

face, though there were no available landmark points for which to find this axis, so a best estimate was used.

The axes of rotation were not exactly on the *x*, *y*, and *z* axes but parallel to them, therefore before landmark configurations were rotated each landmark point first had to be transformed so that the axis of rotation was in the position of the *x*, *y* and *z* axes. If the axis of rotation goes through the two points $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$ then the translation matrix, *T*, for translating a point *P* to new point *P*' is given by:

$$T = \begin{bmatrix} 1 & 0 & 0 & x_2 - x_1 \\ 0 & 1 & 0 & y_2 - y_1 \\ 0 & 0 & 1 & z_2 - z_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \ldots(8.1)$$

$$\text{where } T \bullet P = P' \qquad \ldots(8.2)$$

8.5.3 Rotations

The rotation about the *x*-axis was carried out using the rotation matrix in expression 8.3, and rotation about the *y*-axis using the matrix in expression 8.4.

x-axis rotation matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & \sin\theta & 0 \\ 0 & -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \ldots(8.3)$$

y-axis rotation matrix

$$\begin{bmatrix} \cos\theta & 0 & -\sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ \sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \ldots(8.4)$$

8.5.4 Inverse Transformation

Finally, after the rotation, the inverse transformation of the rotated landmark points (multiplication by T^{-1}) is required to get the landmarks back to the original coordinate system, which is used to carry out the facial comparisons.

8.5.5 Results

8.5.5.1 x-axis Rotations

Figure 8.3 shows the original coordinates for a face (circular points) along with the rotated coordinates of the same face tilted upwards (squares) or downwards (triangles). Only the (*x*, *y*) anterior coordinates are shown and the different rotations look like translations of the original face, however examination of distance matrices of relative distances between the twenty-two landmark points show that actually the relative position of points varies for each rotation, as the axis of rotation is not central to all points.

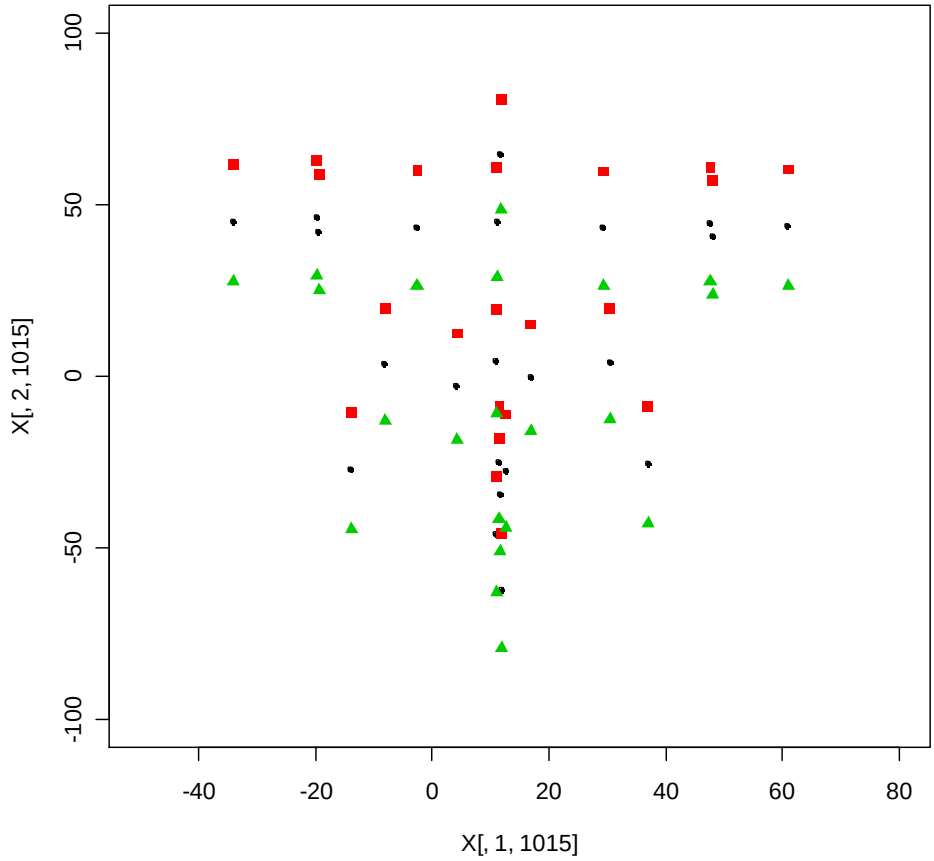


Figure 8.3 - Example of the x-y anterior landmarks of one of the test faces in the original orientation (circles) and generated angles of downwards (triangles) and upwards (squares) tilts by 2 degrees about the x-axis.

Table 8.7 shows the results for ten facial comparisons of each original face compared with the subsequent rotated face simulating a downward head tilt. The number of degrees by which the faces were rotated was varied from two to ten and the number of matches and exclusions, when using a 300 LR threshold, were quantified.

Rotated Degrees	Matches, LR>300	Insufficient Evidence	Exclusions, LR<0.00333	Min LR	Max LR
2	4	6	0	1.24E+01	2.14E+03
3	0	10	0	1.09E-01	5.44E+01
4	0	8	2	1.32E-04	2.82E-01
5	0	0	10	2.05E-08	3.18E-04

Table 8.7 - Number of matches and exclusions when comparing each original face with its downward rotation about the *x*-axis

The results in Table 8.7 show that when the faces were rotated by a downwards angle of just two degrees about the *x*-axis, only four out of ten faces matched the original pre-rotated faces. The remaining six comparisons showed insufficient evidence to confirm either a match or exclusion result. When the angle of rotation was increased to three degrees all ten comparisons had insufficient evidence to confirm either result. A further increase to four degrees confirmed two exclusions and rotations of five degrees or greater excluded all ten facial comparisons.

The results were very similar for the simulated upward rotation of the face, Table 8.8. Again for ten facial comparisons, each face was compared with its' subsequent rotation face simulating an upward head tilt. The number of degrees by which the faces were rotated was varied from two to ten and the number of matches and exclusions, when using a 300 LR threshold, were quantified.

Rotated Degrees	Matches, LR>300	Insufficient Evidence	Exclusions, LR<0.00333	Min LR	Max LR
2	5	5	0	1.60E+01	1.49E+03
3	0	10	0	1.99E-01	3.75E+01
4	0	8	2	4.42E-04	3.03E-01
5	0	0	10	1.78E-07	1.70E-03

Table 8.8 - Number of matches and exclusions when comparing each original face with its upward rotation about the *x*-axis

8.5.5.2 Y-axis Rotations

Figure 8.4 shows the original coordinates for a face (circular points) along with the rotated coordinates of the same face turned to the right (squares) or left (triangles). Only the (*x*, *y*) anterior coordinates are shown and the different rotations look like translations of the original face, however examination of distance matrices of relative distances between the landmark points show that actually the relative position of points varies for each rotation, as the axis of rotation is not central to all points.

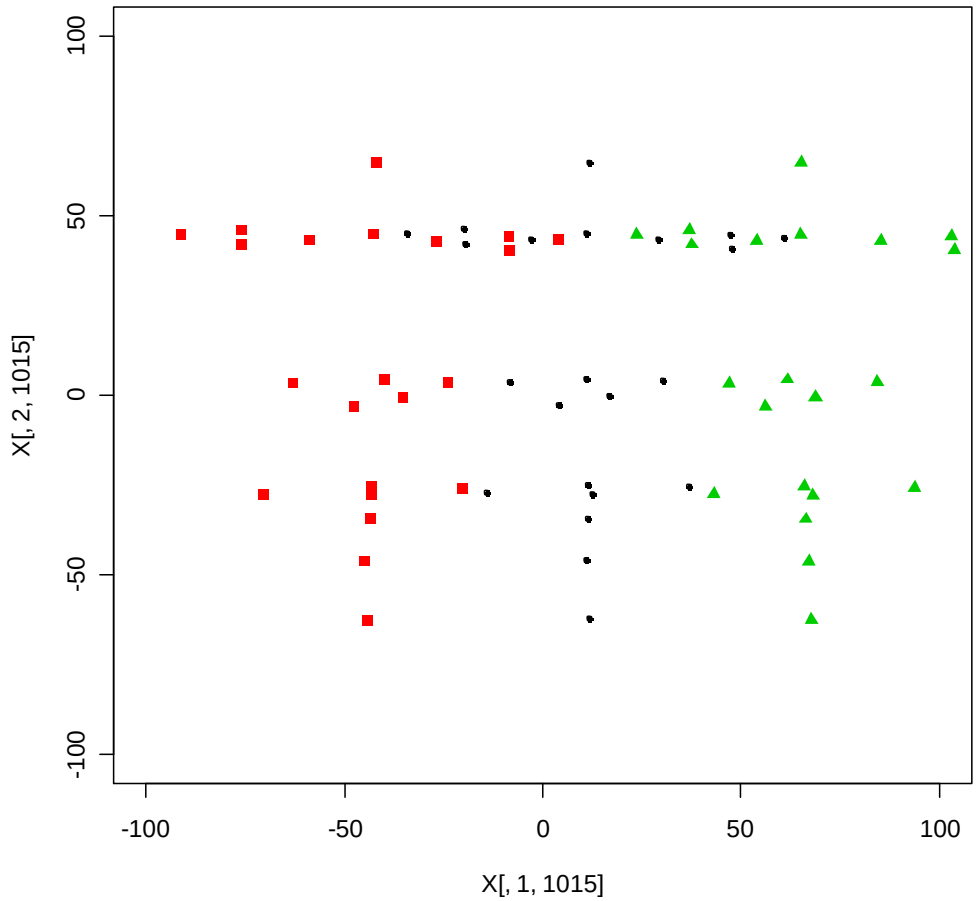


Figure 8.4 - Example of the x-y anterior landmarks of one of the test faces in the original orientation (circles) and generated angles of left (triangles) and right (squares) tilts by 7 degrees about the y-axis.

Table 8.9 shows the results for the ten facial comparisons, one for each original face compared with its' subsequent left turn rotation. The number of degrees by which the faces were rotated was varied from two to eight and the number of matches and exclusions, when using a 300 LR threshold, were quantified. Table 8.10 shows the comparative results for the ten facial comparisons with the simulated right turn rotation.

Rotated Degrees	Matches, LR>300	Insufficient Evidence	Exclusions, LR<0.00333	Min LR	Max LR
2	8	2	0	1.90E+02	1.44E+04
3	6	4	0	3.97E+01	4.18E+03
4	3	7	0	4.04E+00	6.67E+02
5	0	10	0	1.84E-01	5.46E+01
6	0	10	0	3.37E-03	3.52E+00
7	0	7	3	2.21E-05	1.23E-01
8	0	0	10	4.64E-08	1.91E-03

Table 8.9 - Number of matches and exclusions when comparing each original face with its rotation to the left about the *y*-axis

Rotated Degrees	Matches, LR>300	Insufficient Evidence	Exclusions, LR<0.00333	Min LR	Max LR
2	8	2	0	1.43E+02	9.35E+03
3	5	5	0	3.08E+01	2.09E+03
4	0	10	0	2.77E+00	2.42E+02
5	0	10	0	1.02E-01	1.36E+01
6	0	9	1	1.39E-03	6.25E-01
7	0	3	7	6.14E-06	1.26E-02
8	0	0	10	7.58E-09	9.92E-05

Table 8.10 - Number of matches and exclusions when comparing each original face with its rotation to the right about the *y*-axis

As with the *x*-axis rotations (Tables 8.7 and 8.8) the rotations about the *y*-axis, in either the left or right direction, reduced the number of LR confirmed matches (Tables 8.9 and 8.10). There was found to be insufficient evidence to support either hypothesis with simulated head rotations of just two degrees. The evidence for all matches disappeared for simulated rotations of four degrees or greater. All matches were incorrectly classified as exclusions for simulated rotations of eight degrees or greater.

8.5.5.3 Rotation in both x and y directions

Another extension of interest to this work is to investigate what happens if the subject both turns and tilts the head slightly. To estimate the effect of this the original faces were multiplied by both the *x*-rotation matrix and the *y*-rotation matrix (expressions 8.3 and 8.4). It should be noted that the results here are very approximate, as the pivot for facial rotation in these directions is unknown.

Tables 8.11-8.14 show results for comparisons of the ten faces with the rotated faces simulating the four directions of rotations through both *x* and *y* directions. All results showed that if the head was rotated by just two degrees in two directions very few matches were confirmed using the LR method with the 'best' subset of matching variables.

Rotated Degrees	Matches, LR>300	Insufficient Evidence	Exclusions, LR<0.00333	Min LR	Max LR
2	5	5	0	9.42E+00	1.16E+03
3	0	10	0	4.81E-02	1.76E+01
4	0	8	2	2.72E-05	6.88E-02
5	0	0	10	1.65E-09	8.55E-05

Table 8.11 – Facial matches and exclusions obtained when comparing each original face with its rotation when the face had been rotated in both the x and y directions upwards and left.

Rotated Degrees	Matches, LR>300	Insufficient Evidence	Exclusions, LR<0.00333	Min LR	Max LR
2	3	7	0	4.98E+00	1.18E+03
3	0	10	0	1.71E-02	1.70E+01
4	0	6	4	5.90E-06	4.14E-02
5	0	0	10	1.91E-10	2.38E-05

Table 8.12 - Facial matches and exclusions obtained when comparing each original face with its rotation when the face had been rotated in both the x and y directions downwards and right.

Rotated Degrees	Matches, LR>300	Insufficient Evidence	Exclusions, LR<0.00333	Min LR	Max LR
2	1	9	0	2.15E+00	4.74E+02
3	0	9	1	1.24E-03	1.91E+00
4	0	0	10	2.27E-08	1.47E-03

Table 8.13 – Facial matches and exclusions obtained when comparing each original face with its rotation when the face had been rotated in both the x and y directions downwards and left.

Rotated Degrees	Matches, LR>300	Insufficient Evidence	Exclusions, LR<0.00333	Min LR	Max LR
2	0	10	0	1.94E+00	2.48E+02
3	0	9	1	1.47E-03	1.21E+00
4	0	0	10	5.42E-08	6.72E-04

Table 8.14 - Facial matches and exclusions obtained when comparing each original face with its rotation when the face had been rotated in both the x and y directions upwards and right.

These findings are extremely important because acquiring images for real-life facial comparisons it is very unlikely that the face is going to be positioned in exactly the anterior view for both images to be compared. The angle of face to camera is likely to be unknown for comparison images and it is not easy to see when the facial angle to camera differs by an angle of nine degrees (Figure 8.1), which is obviously more than two degrees that have been found to affect the matching results.

8.6 Averaging faces

Jenkins and Burton (2008) have outlined an approach for facial recognition which involves taking the average face of a number of different images. This approach could be applied here by taking the average landmarks from a number of different images of the same face. The real life application of facial comparison would inevitably utilize CCTV technology and images from CCTV would be perfect to take an average of landmarks from the face in a series of stills taken from a CCTV video, where the face is in differing angles towards the camera.

To investigate the plausibility of this approach the set of fourteen images of agent Vorder Bruegge (§2.7.3, §8.4, Appendix C, Figure 13.30) were taken and the LR method for facial matching was reapplied taking averages of the facial landmarks. Initially the investigations we did in §8.4 showed that when comparing single images to one another eighty-two matches, fifty-six exclusions and forty-four results of insufficient evidence were obtained from the LR method. Obviously here all comparisons should be matches, however it was noted that the facial expressions and facial angles to camera were not consistent for all fourteen photos in the dataset (Appendix C, Figure 13.30).

Instead of doing single image to single image comparisons (§8.4, Table 8.15 rows 1 and 2), an average was taken of the landmarks from seven of the photos in the dataset, this was the control face. This control face was then compared with the individual sets of landmarks from each of the remaining seven (recovered) faces in the dataset. All possible combinations of seven faces were averaged and compared against all other faces; this gave a total of 24024 facial comparisons, Table 8.15. Taking an average of the control face in this way gave a marginal improvement in the number of true matches found (increasing results from 45% to 47%), however it dramatically reduced the

number of false exclusions (31% was reduced to 11%) thus increasing the number of cases where there was found to be insufficient evidence to support either hypothesis (from 24% to 43%).

After examining the dataset of images it was decided to exclude the two images numbered 8 and 9 (Appendix C, Figure 13.30), which depicted the subject with a smiling expression. It was thought that averaging a set of images could overcome the issue of differing facial angles to camera, as the same shape appears in the image just rotated at a different angle. However different facial expressions essentially cause different face shapes and taking an average of six neutral expressions and one smiling expression will inevitably produce a different shaped average from simply taking an average of seven neutral faces. After excluding the two smiling images, the average of six faces from the dataset was used as the control and compared to all the remaining images in the same way as before, Table 8.15. This time there were a total of 5544 comparisons. Excluding the two smiling images dramatically improved the LR matching results, using the average of six faces for the control face 84.7% of results obtained were true matches, 5.6% were false exclusions and 9.7% had insufficient evidence. This strongly strengthens the case that facial expression has a very important effect on facial matching and all facial images being compared should depict the same facial expression to ensure good quality results.

Next, as well as taking an average for the control face in the comparisons, an average of the landmarks for the remaining six photos was also taken for the recovered face. This was carried out for every combination of six faces from the dataset, giving a total of 924 comparisons. Here all the results (100%) were true positives, there were no false results. The number of faces used in the averages was investigated to find the optimum. It was found that using averages of just three facial images for both the control and recovered faces gave LR matching results where 92.1% were true positives, only 7.9% were insufficient evidence and there were no false exclusion results.

These results are very encouraging; when recovered facial images are obtained for comparison, taking landmark measurements from several images (e.g. obtained easily from CCTV of a crime being committed) dramatically improves the accuracy of the matching method even if only one control image is available.

Data	Control Images Averaged	Recovered Images Averaged		Match (LR>300)	Insufficient Evidence	Exclusion (LR<0.00333)
All dataset	1	1	no. cases	82	44	56
All dataset	1	1	%	45.1%	24.2%	30.8%
All dataset	7	1	no. cases	11250	10226	2548
All dataset	7	1	%	46.8%	42.6%	10.6%
Excluding smiles	6	1	no. cases	4696	539	309
Excluding smiles	6	1	%	84.7%	9.7%	5.6%
Excluding smiles	6	6	no. cases	924	0	0
Excluding smiles	6	6	%	100.0%	0.0%	0.0%
Excluding smiles	2	2	no. cases	2234	736	0
Excluding smiles	2	2	%	75.2%	24.8%	0.0%
Excluding smiles	3	3	no. cases	17024	1456	0
Excluding smiles	3	3	%	92.1%	7.9%	0.0%

Table 8.15 – Match and exclusion results for the fourteen images of Agent Vorder Bruegge (Appendix C, Fig. 13.30), various different averages were taken of the landmark configurations to see the effect on the number of matches obtained.

8.7 Summary

This chapter has examined further the performance of the LR method for facial matching (§3.8.4, §7.3, §7.4) using the 'best' found subset of variables (§7.4.2.3) to quantify matches in some different datasets containing multiple images of like faces. This was a good test to observe how the methods would cope with data obtained externally and under different conditions to that analysed so far, such as that acquired in a real life situation.

It was discovered that multiple images of the same face (§2.7.6, §8.3) matched relatively well when the landmark points on both images were placed by the same observer (§8.3.1). Taking two sets of landmark measurements from one image, with each set positioned by a different observers, produced comparatively worse match results (§8.3.1), suggesting that the LR matching procedure needed to account for observer error in the model used to calculate the LRs (§3.8.4.1). Applying this extension did not appear to improve the number of true matches found, however the number of results that were found to have insufficient evidence to support either hypothesis was increased (§8.3.1.2). In terms of evaluating other types of forensic evidence there is a need to account for observer subjectivity, in particular fingerprint analysis has been in the news as being perhaps less accurate than everyone believed.

The match results for three sets of twins (§2.7.5) were poor, no matches were found using the LR procedure with 'best' found subset of matching variables. Visual examination of the source images exposed visible differences in facial position and expression (§8.2, Confidential Appendix C Figures 13.24 – 13.29). Also, on examination of the data for the twins it was found that multiple observers placed the two measures of landmark points on the images. This is a limitation since the subset of variables used for facial matching (§7.4.2.3) was deemed the 'best' on the basis of its performance at matching the FBI anterior test data (§2.7.2), which was collected by only one observer.

One key discovery was that an important part of the data checking should be to ensure that the facial expressions of the subjects in all the images in the dataset of both control

and recovered data are consistent. Differences (particularly where the subject was smiling) were seen to effect the facial matching results (§8.4, §8.6).

Another important finding was that the facial position to camera (rotation) was found to have a large effect on whether two images matched (§8.5). 3D geometry was used to rotate facial landmarks from images (§8.5) and the MVNLR procedure was applied to evaluate matches between the original and rotated landmarks. It was found that rotating landmarks by merely three degrees caused false negatives (exclusions) to occur between the original and rotated known matches.

It was also revealed that the LR matching method could be improved by taking the average of landmarks from three different images of the control and recovered faces (e.g. from stills of CCTV video footage), §8.6. This was determined on a small dataset and needs to be investigated further by collecting multiple images of many more faces to match.

9 Discussion

9.1 Introduction

This chapter brings together the main points and findings from the research study and critically evaluates what has been achieved. A summary of each chapter is given (§9.2) recalling the key rationale of each main section and any issues which need to be addressed. The key steps in the procedure developed for anterior 2D facial comparisons using the Geometrix® database (§2.4) are given (§9.3). A set of conclusions have been drawn (§9.4) reviewing the aims that were set out in the introduction (§1.1), including limitations of the methods (§9.4.1) and suggestions for improvements (§9.4.2). Ideas for the further development of this work are considered (§9.5), along with potential applications of the work in other areas (§9.6).

9.2 Summary by Chapter

Chapter 1 reviews current methods in facial identification and demonstrates that these are rudimentary, unscientific and untested. Other forms of forensic evidence such as DNA and trace evidence are presented as probabilistic measures. The need to develop a reliable quantitative technique for comparing and identifying faces is explained. The means for carrying out a statistical analysis of face shape based on well-defined anthropometrical landmark points are summarised, along with some recognized techniques in presenting forensic evidence to courts. It was established that methods could be applied to model face shape in a population and calculate the likelihood of a face shape occurrence. Two face shapes could then be compared by means of a likelihood ratio (LR). In order to achieve this some knowledge of the population variation in face shape in a large sample of people is required.

Some criteria for a facial identification technique to be admissible in court were outlined. These include the production of precise protocols and guidelines for the placement of landmark points and the analysis of the resulting coordinate data, in order for any developed technique to be accepted and adopted. The landmark points for investigation should be easily placed by an observer with no previous expertise, in order to make the technique practical. Any statistical methods used to affirm or reject a facial

match to aid in facial identification in court should be easily explainable to the judge and jury, who will most probably have no expertise in the field of statistics and probability. More importantly the methods must be robustly tested and subjected to peer review in order for them to become accepted within the forensic and statistical scientific communities.

Some potential problems to address when developing a facial comparison method were thought to be properties that will affect the outcome, for example, facial expression and lighting conditions during image capture. However it was thought that to develop a method that will be invariant under such properties was impractical and it would be better to first carry out a study examining face shape controlling for a 'natural', Farkas (1994), facial expression and constant camera and lighting conditions. Once it is known how the face varies in shape between individuals under these controlled conditions the extra factors could then be brought in.

Chapter 2 fully describes the wide range of facial image and landmark data available for use throughout the research project. Descriptions of anthropological facial landmarks which may be suitable for the comparison of face shapes were given, an initial set of sixty-one facial landmarks to explore for facial matching were listed (§2.2).

The data for a pilot study (§2.3, §4) to confirm that the methods proposed for facial matching (§3) were suitable for use with landmark data were summarized. A detailed account of the large Geometrix® facial database was presented. These data, available for the main component of the research, facilitate the calculation of population estimates of facial variation (§6), which can be used to quantify the likelihood that two faces 'match' (§7, §8). Descriptions of how the images were collected (§2.4.2) and how the landmark coordinates were measured on these images (§2.4.3) were provided. Other image and facial landmark data were also described (§2.5, §2.6, §2.7). These data were used to check the reliability of the landmark data and chose a set of landmark points which would be the most appropriate for facial matching (§2.5, §5.2); to validate that the data collection procedure was repeatable when multiple observers took the landmark measurements from the facial images (§2.6, §5.3); and to test the performance and accuracy of developed facial matching methods (§2.7, §7, §8). The data for testing the methods were from other sources external to the main Geometrix® data and consist of known facial matches or exclusions. The landmark measurements for these data were

collected using different software to the main Geometrix® data, therefore could verify whether proposed methods for the analysis and comparison of shapes could handle data acquired from different sources, which is an important requirement of the techniques.

Observed flaws in the design of experiment for the collection of the main facial database were highlighted. There was some bias in the sample in terms of the ethnic and age distributions of the faces in the sample (§2.4.1). If the results for the available data are reasonable then further exploration of different ethnic and age groups could be carried out at a later stage. There were ten different photographers collecting image data and this was not controlled for. The way that the landmark data was measured (§2.4.3) could also have been improved by using a more orthogonal design for a better assessment of inter-measurement variability. If each observer were given a set of faces and they were responsible for taking both sets of the two landmark measurements from the set then inter-measurement variability for each observer could have been measured.

Chapter 3 reviews the statistical theory needed to extract facial shape information from landmark coordinate data using methods in statistical shape analysis (§3.4 - §3.6). Methods for examining the structure of shape variability in a data set by using multivariate methods such as principal components analysis (PCA) to examine the tangent space coordinates were detailed (§3.7).

Methods for modelling the extracted shape data as a multivariate normal distribution and using the associated model parameters to estimate the likelihood of two facial shapes being quantified a ‘match’ or ‘exclusion’ were given (§3.8.4). The LR tests whether landmark configurations from two faces are more similar to each other than they are to all other faces in the known population sample (§3.8.4). Here the known population sample is the main Geometrix® database (§2.4). It was found that several modifications to the method were required for use with the facial data, these are suggested in §7.3 and applied in §7.4 and chapter 8.

Chapter 4 details a pilot study which was carried out to check whether the data available and the proposed methods were feasible for use with facial comparison. In summary, the results showed that statistical shape analysis and likelihood ratios were effective methods to use in quantifying facial matches. Precise empirical measurements of coordinates of attributes of the face are used, which gives these methods a clear

advantage over the current techniques used for facial identification. It also means previous studies of facial variation could be used to permit the probability of a credible match to be empirically established; a database of measurements could be expanded as more facial data is collected to improve the model used in the matching.

Multivariate analysis of variance (MANOVA) results established there was strong evidence that aligned and transformed facial landmark data showed sufficient variation between faces to overcome any variation attributed to taking multiple landmark measures and scans of the same image. The results were promising even though only a proportion of facial features (ten facial landmark points) were examined due to censoring in the images of the dataset. This censoring also implied that partial facial matching could be done successfully; which is useful for real life crimes where the perpetrators may mask their facial features in someway.

Cluster analysis and the proposed LR methods were applied to find possible facial matches in the data. Although a cluster dendrogram is a good visualisation tool there were several pairs found in the data that were false matches, indicating that clustering may not be a statistically appropriate technique for matching the facial shape data. Also when looking at displaying much larger datasets dendrograms are inappropriate. LRs for evaluating the strength of a facial match when modelling the data with a multivariate normal distribution were calculated for pairs of faces. The top three matches found using this method (i.e. the highest three LRs) were confirmed visually as true matches. There were also false positive results, where the LR result was greater than one however the two images were not matches. By applying a threshold to the LR results the false positive rates could be reduced, here LR>200 would have been an appropriate level to only select the true matches.

Chapter 5 explored the variation in sixty-one landmark points proposed for facial comparison (§2.2, §5.2). A subset of thirty of these points were chosen for collection from the main database (§2.4.2), these were thought the most appropriate for facial matching based on the consistency of multiple measures taken by two observers, and also the influence of each point in discriminating between faces (§5.2.5). Based on the outcome of this work a manual (Appendix B) to assist observers in locating the points effectively was written.

The repeatability of the data collection technique was demonstrated, as data collected from different observers on the thirty chosen points was comparable (§5.3). For a small subset of ten faces, a Wards cluster analysis classified the different faces into distinct groups despite multiple observers placing the landmark points (§5.3.3). Following the results of this work further observers employed on the project were asked to collect multiple measurements of the landmark points from the ten faces in this subset. The cluster analysis was then rerun to ensure that new observers were producing configurations which were in line with the other current observers. As one important criterion of the developed facial comparison methods were they had to be independent of observer judgement, and so an improvement on the subjective 'expert witness' evidence.

Chapter 6 explored the large Geometrix® facial database looking at differences in size and shape with respect to the age, sex and ethnicity of the subject (§6.2.2) after Procrustes methods (§3.4) were used to align the data. The variation of the individual landmark points were explored through PCA of the tangent shape coordinates (§6.2.6, §6.4.1) and a multivariate normal model for the facial shape data was found to fit the data well (§6.5). Thus a multivariate normal model was suitable for modelling Procrustes corrected facial landmark data. Obtaining the parameters of this model allows the multivariate normal likelihood ratio (MVNLR) procedure (§3.8.4.1) to be applied to the facial data to quantify likelihoods of facial matches or exclusions.

Several observations in the database were found to contain errors (§6.2.5), these were seen as outliers in the principal component (PC) score plots. The data were cleaned where possible (correcting for mislabelled landmark points), or excluded where mistakes could not be corrected (incorrectly positioned landmark points). An important issue when bringing in new facial data is therefore to be aware of the potential data problems and examine the PC scores for such outliers.

Transforming the Procrustes registered facial landmark data onto PCs and examining the loadings explained different aspects of shape variation in the main facial database (§6.2.6, §6.4.1). Groups of the original thirty landmark variables showed large variation on particular PCs, with variation in specific facial areas being seen in each PC (Table 6.3). The facial landmarks found to vary the most in 3D were the points around the ears

(§6.2.6). It was determined here that facial images obtained in real-life for comparison, from other sources outside of the Geometrix® database, would be in only 2D. The anterior (forward facing) facial view is likely to be the best 2D view to use for facial comparison, as this view encompasses the most landmarks (§6.4). Therefore the ear landmarks were excluded from anterior facial matching, as they are not easily placed in this view. In addition it was also determined that when using the anterior facial views the data must be Procrustes aligned including the removal of scale information, as the distance of the subject to camera will be unknown and unlikely to be the same in two images obtained from different sources.

Chapter 7 applied the likelihood ratio method (§3.8.4) to compare faces in various datasets containing known facial matches (§2.7). The method modelled the tangent coordinates of the main facial database (§2.4) with a multivariate normal distribution and then used model parameters to obtain likelihoods of facial matches or exclusions (§7.2). Certain extensions to the method had to be carried out to handle the large complex dataset (§7.3, §7.4). The data were transformed onto principal components (PCs) to overcome the high correlation in the data. A further extension was that a subset of the variables (PC scores) was found to perform better than taking all or the first few. This was due to the fluctuation of LR results for known matches across the PCs; the LR did not monotonically increase with the number of matching variables used (§7.2.3). Subsets of different landmark points (§7.4.2 and Appendix D) and then subsets of the PCs were investigated to find a set of variables that optimised the results of facial matching for some known matches (§7.4.2.3). A novel method for subset selection has been proposed (§7.3.2) and a LR threshold of 300 (above which to confirm matches) has been suggested (§7.3.3).

The LR method with just five matching variables proved to produce a substantial number of false results when tested with the anterior FBI data (§7.2.2.1). Increasing the number of matching variables to twenty produced much fewer false results, although the method only identified around half of the known matches from the whole test data (§7.2.2.1). Possible factors that may determine which known matches are recognized and which are not were identified to include facial expression and the angle of the face to camera during image capture.

It was found that checking the facial comparison data fitted the multivariate normal model for the background population was an important aspect of the LR procedure (§7.2.2.2). It was seen that when a configuration had just one landmark point misplaced it lay far from the multivariate normal model (§7.2.2.2). Model checking identifies problems with landmark positions, which can then either corrected or excluded from analyses; ultimately improving the match results obtained (§7.2.2.3). It is recommended that before any facial comparisons are carried out the comparison data should be added to the background data and the model checked for errors which could influence the results considerably.

The choice of which variables were the 'best' in terms of producing 'good' evidence for matches (§7.2.3) was addressed. Several criteria were chosen to determine the 'best' subsets, §7.3. Firstly the results obtained from matching on a particular subset of variables were required to be relatively impartial for either hypothesis. In other words the subset should be equally as good at picking out matches as it is at excluding exclusions. A measure to address this issue was devised as the match/exclusion ratio (MER), §7.3.2. This compares the average match LR to the average exclusion LR; 'good' subsets were chosen as those with an MER close to one. In addition to this the actual values for the average match LR and average exclusion LR were examined, obviously the better subsets were those which had greater support for either hypothesis after already measuring the bias of the subset. All potentially 'good' subsets which fulfilled these selection criteria were examined for performance by using the variables to search for matches in the fifty-eight FBI test faces (§7.2.2, §7.4.2 and Appendix D). The number of true and false results indicated how well each subset performed.

A 'best' subset of matching variables was found, the match results for this subset were 98% true matches (§7.4.2.3). The subset was impartial for either hypothesis H_p or H_d (as the MER of the best subset was very close to one), however only around half of all known matches in the sample were picked up. When subsets were evaluated for performance using a threshold of LR>1 to confirm facial matches the number of false positive results ranged from 32% to 60%, §7.4.1. Increasing the match threshold reduced the numbers of false positives to improve results. A threshold of LR>300 found the optimum true results, increasing the threshold further than this did not improve results to a great extent (§7.3.3, §7.4.1). In a courtroom it would be up to a prosecution or defence team to decide what they deem to be a suitable threshold for a 'match' LR, a

good method of practice would also be to quote the percentage of true and false results obtained when using such a threshold.

The ‘best’ subset of eleven landmarks (§7.4.2.1) was derived from knowledge and experience of the landmark data. Prior knowledge and judgement of landmark location, ease of landmark placement and determination in the anterior facial view were all taken into account when choosing the subset of facial landmarks. The strength of results (average LRs for matches and exclusions) along with the MER (§7.3.2) were then used to evaluate a subset of PCs from the subset of facial landmarks. A limitation here was that only twenty facial comparison cases were used to quantify the ‘best’ subset of matching variables (§7.4). Ideally all possible comparisons available where known matches and exclusions exist should be used to determine a ‘best’ subset, however computationally this would be a large task and beyond the boundaries of the current project.

The sensitivity of the ‘best’ subset was checked by randomly dropping a number of faces from the background data, in order to simulate change in the covariance structure, and repeating the matching on the fifty-eight FBI test faces (§7.4.3). The false results were examined and it was found that dropping just 6% of the background data increased the false exclusion rate from 2% to 15% and the false match rate from 2% to 25%. When such increases were seen when dropping this small amount of data it was thought that taking a completely different set of data was likely to require a whole new investigation into the most appropriate subset of facial variables to use to maximise facial matching results.

Chapter 8 further examined the performance of the LR method for facial matching (§3.8.4, §7.3, §7.4) using the ‘best’ found subset of variables (§7.4.2.3) to quantify matches in some different datasets containing multiple images of like faces (§2.7). This was a good test to observe how the methods would cope with data obtained externally and under different conditions to that analysed so far, such as that acquired in a real life situation. Factors which affected the results were explored and suggestions for improving the method were made.

It was discovered that multiple images of the same face (§2.7.6, §8.3) matched relatively well when the landmark points on both images were placed by the same

observer (§8.3.1). When one image was taken and landmark points measured on this by two different observers the match results were comparatively not as good (§8.3.1). This suggested that the LR matching procedure needed to account for observer error in the model used to calculate the LRs (§3.8.4.1, §7.3). Applying this extension did not appear to improve the number of true matches found, however the number of results that were found to have insufficient evidence to support either hypothesis was increased (§8.3.1.2). In terms of evaluating other types of forensic evidence there is a need to account for observer subjectivity, in particular fingerprint analysis has been in the news as being perhaps less accurate than everyone believed.

The match results for three sets of twins (§2.7.5) were poor, no matches were found using the LR procedure with 'best' found subset of matching variables. Visual examination of the source images exposed visible differences in facial position and expression (§8.2, Confidential Appendix C Figures 13.24 – 13.29). Also, on examination of the data for the twins it was found that multiple observers placed the two measures of landmark points on the images. This is a limitation since the subset of variables used for facial matching (§7.4.2.3) was deemed the 'best' on the basis of its performance at matching the FBI anterior test data (§2.7.2), which was collected by only one observer.

Other key discoveries were that it is important to ensure that the facial expressions for all the images in the dataset of both control and recovered data are consistent, as differences were found to effect the facial matching results (§8.4, §8.6). The facial position to camera (rotation) was found to have a significant effect on whether two images matched (§8.5). Matching started to fail when images were rotated just three degrees. It was also uncovered that the LR matching method could be improved greatly by taking the average of landmarks from three images of the control and recovered faces (e.g. from stills of CCTV video footage), §8.6.

9.3 The Anterior 2D Facial Comparison Method with the Geometrix® Database

9.3.1 Procedure

1. Acquire data for comparison: two or more facial images to compare to one another and quantify whether they could be a 'match'.
2. Acquire the population sample of facial measurements from the Geometrix database (§2.4), which contains multiple measurements on eleven anterior landmarks (§7.4.2.1) from images.
3. Take multiple measurements (at least two sets) of the eleven anterior landmarks (§7.4.2.1) on the data for comparison where possible, i.e. only if landmarks are clearly visible.
4. Add 3 to 2 and align with generalized Procrustes analysis (GPA) removing scale, location and rotation. NB If not all eleven points are measured in 3 then only use the landmarks which were measured on all images in 1, as GPA can not deal with missing values.
5. Transform 4 onto tangent shape coordinates and carry out a PCA on these to get the PC scores.
6. Select PC scores 1, 3, 4, 7, 9 and 10 as the six matching variables and apply the MVNLR procedure to compare these variables for pairs of images from 1 to get a LR for each comparison.
7. LRs>300 define that a comparison is a match, LRs<0.00333 define that a comparison is an exclusion, any results 0.00333>LR>300 have insufficient evidence to confirm either hypothesis.

9.3.2 Key Points and Suggestions for Improvement

The procedure (§9.3.1) is sensitive to facial expression and rotation. It should be ensured that all comparison images are in an anterior position with a neutral facial expression in order to get the best possible results from the Geometrix® database (§2.4).

The Geometrix® database is available to researchers (Evison and Vorder Bruegge, 2008), however if a different population sample is used the procedure outlined above such be extended to select the most appropriate set of facial matching variables for the new sample. The analyses carried out in chapters 6 – 8 should be replicated to examine the facial variation in the new sample and select a subset of matching variables which optimizes the matching results for the different set of data. These recommendations are made on the basis of the findings of the robustness analyses in §7.4.3.

If possible it is recommended to take landmark measurements from multiple images of the faces to be compared. It was seen in §8.6 that taking an average of the landmark measurements from more than three images of the same faces noticeably improved the match rate.

9.4 Conclusions

It was mentioned in the introduction to this research that in order for any developed technique to be accepted and adopted by facial imaging analysts, precise protocols and guidelines for the placement of landmark points and the analysis of the resulting coordinate data should be provided. The landmark placement manual (Appendix B) was written to cover the data collection protocol requirement and the procedure in §9.3.1 gives guidelines for the analyses. It was noted in §9.3.2 that the procedure is sensitive and if a different database of face shapes is used then a more thorough investigation is required, whereby a different subset of facial matching variables may be found appropriate.

When selecting the set of landmark points for investigation a key requirement was that the landmarks should be easily placed by an observer with no previous expertise. In the field of facial identification, landmark observers are likely to be police officers or

lawyers, and not forensic anthropologists. It was seen in chapter 5 that the landmark collection procedure, following the instructions in the landmark placement manual (Appendix B), was validated and the data on thirty different points was deemed comparable for six different observers, four of whom had no previous expertise.

A substantial study looking at population variation in face shape in a large sample of people has to be carried out in order to gain better knowledge of the face shape system. Obviously the more data that are used in the sample to model for the LRs the more the matching results will be appropriate for applying to the general population. After carrying out some robustness analyses (§7.4.3) it is clear that the sample collected and used throughout the study, although large, is not extensive enough to be able to represent the general population. There were also limitations in the distribution of ethnicity and age.

The large database used for examining face shape controlled for a 'natural', Farkas (1994), facial expression and constant camera and lighting conditions. However, additional images brought in to test the methods did not always follow these conditions.

It was suggested in the aims that statistical methods used to affirm or reject a facial match to aid in facial identification in court should be easily explainable to the judge and jury, who will most probably have no expertise in the field of statistics and probability. Likelihood ratios are used in courtrooms already and so must be accepted and easy enough to explain. What may be more complicated to explain with this research is the series of data transformations that occur before the likelihood ratio calculations are carried out. These include the generalized Procrustes analysis, then a transformation onto tangent space, then a PCA from which a subset of PCs are taken as the matching variables.

It was also a requirement that the methods developed must be robustly tested and subjected to peer review in order for them to become accepted within the forensic and statistical scientific communities. As many tests as possible have been carried out here with the data and time restrictions available. What would be necessary is for an independent individual to review the methods and feedback any issues or concerns.

9.5 Limitations

A flaw in the design of the whole study is that the facial matching has only been carried out in the 2D anterior view (i.e. with the subject looking towards the camera). On examination of images from different views of the face (§2.4.2, Figure 2.4) it is much harder to visually confirm alike faces when two images have been taken at angles that are not anterior. In real life criminal situations, for example where we wish to identify someone from some CCTV footage, it is unlikely that a criminal will look directly at the camera in this way.

The procedure only correctly identified around half of the known matches in the FBI anterior test data. It was ascertained that reasons for this could have been down to the position of the face to camera during image capture and also to facial expression. More test data should be used on the method, where these factors have been controlled for.

9.6 Future Work

9.6.1 Ideas for Further Development

Many things could be investigated with this large and complex dataset. The biographical information and information on blood related individuals has not been thoroughly explored.

One way of possibly improving the matching results would be to add in additional information into a facial comparison, for example if the two faces thought to be a match were known to be males, aged thirty and of white British ethnicity then a better population sample to use for the LR calculations in such a case would be to only select the white British male individuals who were around aged thirty. The Geometrix® sample is limited in terms of other ethnic data; however an intelligent search for white British or a specific age and sex group is possible.

Now it is known how the face varies in shape between individuals under controlled conditions (‘natural’ facial expression and constant camera and lighting conditions)

extra factors could now be investigated. Some other factors that could affect the outcome of facial matching are facial expression and lighting conditions during image capture. Darwin (1872) said emotions can be classified into six types - anger, fear, sadness, disgust, surprise and enjoyment. Psychologists have carried out work to suggest that any facial expression can be categorised into one of these emotions (Eckman, 1993). A set of individuals could be asked to be photographed holding these facial expressions and an investigation into the movement and variation of the landmark points could be carried out. Similarly a set of subjects could be photographed at various times of the day in various degrees of light to see at what stage landmark points start to become difficult to place.

Further investigations could be carried out to see if the techniques applied in this study could also be applied to non-anterior facial images. This has not been investigated here, although §8.4 uses three dimensional facial data to simulate what happens to matching results when the face is rotated a few degrees away from the anterior view.

10 References

Aitken, C.G.G. and Lucy, D. (2004) *'Evaluation of trace evidence in the form of multivariate data'* Appl. Statist, **53**, 109–122.

Aitken, C.G.G. and Taroni, F. (2004) *'Statistics and the Evaluation of Evidence for Forensic Scientists'*, Wiley.

Bergen County Technical School (2004) *'The Forensic Science Project'* (last updated 06/03/04, accessed 20/04/05)

URL http://www.bergen.org/EST/Year5/fingerprint.htm

Berry, D. A. (1991) *'Inferences in forensic identification and paternity cases using hypervariable DNA sequences'* (with discussion). Statist. Sci., **6**, 175–205.

Berry, D. A., Evett, I.W. and Pinchin, R. (1992) *'Statistical inference in crime investigations using deoxyribonucleic acid profiling'* (with discussion). Appl. Statist., **41**, 499–531.

Bock, M. and Bowman, A.W. (2005). *'On the measurement and analysis of asymmetry with applications to facial modelling'* Appl. Statist, **55**, 77–91.

Bookstein, F, L., (1978) *'The Measurement of Biological Shape and Shape Change'* Lecture notes on Biomathematics, Vol. 24. Springer-Verlag, New York.

Bookstein, F. L., (1986) *'Size and Shape Spaces for Landmark Data in Two Dimensions'* (with discussion), Statist. Sci. **1**, 181–242.

Bookstein, F. L., (1991) *'Morphometric Tools for Landmark Data: Geometry and Biology'* Cambridge University Press, Cambridge.

Cootes, T.F. Edwards, G.J. and Taylor, C.J. (2001) '*Active Appearance Models*', IEEE PAMI, Vol.**23**, No.6, pp.681-685

Darwin. C. R. (1872) *'The Expression of the Emotions in Man and Animals'*

Dryden, I. L. and Mardia, K. V. (1998) *'Statistical Shape Analysis'*, Wiley.

P. Ekman. (1993) Facial expression of emotion. *American Psychologist*, 48, 384-392

Evison M. P. and Vorder Bruegge R. W, (2008) '*The Magna Database: A Database of Three-Dimensional Facial Images for Research in Human Identification and Recognition'* Forensic Science Communications, volume **10** number 2.

Evett, I.W., Cage, P. E. and Aitken, C. G. G. (1987) *'Evaluation of the likelihood ratio for fibre transfer evidence in criminal cases'* Appl. Statist., **36**, 174–180.

Farkas, L.G. (1994) (ed.) *'Anthropometry of the head and face'*, New York: Raven Press.

Fraser, N., Yoshino, M., Imaizumi, K., Blackwell, S.A., Thomas, D.L., Clement, J. G. (2003) *'A Japanese computer-assisted facial identification system successfully identifies non-Japanese faces'* Forensic Sci. Int. **135**, 122–128.

Galton F. (1892). *Fingerprints*. MacMillan and Co, London

Goodall, C. R. (1991) *'Procrustes methods in the statistical analysis of shape'*, Journal of the Royal Statistical Society, Series B, **53**, 285–339.

Hallinan, P.L., Gordon, G.G., Yuille, A.L., Giblin, P. and Mumford, D, (1999) *'Two- and three-dimensional patterns of the face'*, Natick, MA: AK Peters.

Hancock, P.J.B. (2000) '*Evolving faces from principal components*'. Behavior Research Methods, Instruments and Computers, **32**-2, 327-333.

Hand D.J. (1997) *Construction and assessment of classification rules*, Wiley.

Harley, I. (2004) *'Forensic Photogrammetry: a personal perspective'* Geomatics World, Vol.12, issue 2 (accessed 01/04/05)

URL **http://www.pvpubs.com/read_article.asp?I27&article_id=148**

Jenkins. R and Burton, A. M. (2008) '*100% Accuracy in Automatic Face Recognition*' ***Science*** **319** (5862), 435.

Kendall, D.G. (1977) *'The Diffusion of Shape'* Adv.Appl.Prob **9**, 428–430.

Kendall, D.G. (1984) *'Shape Manifolds, Procrustean Metrics and Complex Projective Spaces'* Bull. London Math. Soc., **16**, 81–121.

Kendall, D.G., (1989) *'A Survey of the Statistical Theory of Shape'* Statist. Sci. **4**, 87–120.

Lucy, D. (2005) *'Introduction to Statistics for Forensic Scientists'*, Wiley.

Mardia, K.V. and Dryden, I.L. (1989) *'Shape Distributions for Landmark Data'* Adv.in Appl.Prob. **21**, 742–755.

Mckie, I. (2005) *'shirleymckie.com'* (accessed 20/04/05)
URL http://www.shirleymckie.com/

Morecroft, L. C. (2002) *'Identification of Faces',* MSc Statistics dissertation, University of Sheffield.

Nordberg, P. (2005) *'The Daubert Worldview'* (work in progress, last revised 04/03/05, accessed 19/04/05)
URL http://www.daubertontheweb.com/

Porter, G. and Doran G. (2000) *'An anatomical and photographic technique for forensic facial identification'* Forensic Science International, 114, 97-105.

Schofield, D. and Goodwin, L. (2006), *'Facing the Future: Errors Involved in Biometric Measurement of the Human Face'*, Forensic Science: Classroom to Courtroom, Proceedings of The 18th International Symposium of the Forensic Sciences, Fremantle, Western Australia, 2nd - 7th April 2006.

Solomon, C.J, Gibson, S.J and Pallares-Bejarano, A. (2005) '*EigenFit – the generation of photographic quality facial composites*' The Journal of Forensic Science, in press

U.S Department of Justice (FBI) (1988) *Facial Identification Catalogue* Graphic design Unit, Special projects Section, Laboratory Division.

Venables, W.N. and Ripley, B.D. (2001) *Modern Applied Statistics with S-PLUS, 3rd Ed* (New York: Springer)

Yoshino, M., Matsuda, H., Kubota, S., Imaizumi, K., Miyasaka, S. and Seta, S. (1997) *'Computer-assisted skull identification system using video superimposition'* Forensic Sci. Int. **90,** 231–244.

Yoshino, M., Matsuda, H., Kubota, S., Imaizumi, K. and Miyasaka, S. (2000) *'Computer-assisted facial image identification system using a 3D physiognomic rangefinder'* Forensic Sci. Int. **109,** 225–237.

Yoshino, M., Noguchi, K., Atsuchi, M., Kubota, S., Imaizumi, K., Thomas, C.D., and Clement, J.G. (2002) *'Individual identification of disguised faces by morphometrical matching'* Forensic Sci. Int. **127,** 97–103.

http://www.cs.cf.ac.uk/Dave/AI2/node174.html

http://www.innocencenetwork.org.uk/

http://www.innocenceproject.org/

http://life.bio.sunysb.edu/morph/

11 Appendix A – Data Collection Questionnaires

You are being asked to participate in a research study. Before you decide, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

What is the purpose of this study?

The faces of perpetrators of often-serious crimes are regularly caught on Closed Circuit Television (CCTV) or security cameras. You may have seen such images on television on *BBC Crimewatch UK* or *Crime stoppers*. Little scientific work has been done that can be used to help the police or a jury decide whether the face in the CCTV image is the same as that of someone suspected of committing a crime. We think there is some danger of both unwarranted convictions and acquittals in cases involving facial comparison.

This research project will develop a means of comparing faces. This will be accomplished by measuring the distances between certain points on the face. We will be using the information from your 3D photograph to take these measurements. This will help us understand how people's faces vary, and help us to create better methods of recognizing and distinguishing between faces scientifically.

Who is organizing and funding the research?

The University of Sheffield is organizing the research project with other scientists at Nottingham and Kent Universities. An international consortium of agencies whose goal is the prevention and detection of crime, and/or the administration of justice sponsors the project—including the US Federal Bureau of Investigation. The project will follow the guidelines of the UK Police Information Technology Organisation. Authority for this research project rests with the University of Sheffield.

Why have I been chosen?

We are gathering information from volunteers who are willing to participate in our project.

Do I have to take part?

Participation in the project is purely voluntary. It is up to you whether or not you take part. If you do decide to take part you will be given this information sheet to keep and be asked to sign a consent form.

What will happen to me if I take part?

We will take your 3D photograph and record the following biographical information: your age, sex, ancestral affiliation (ethnicity), and whether any of your relatives are also volunteering. We need to record this information as these factors can effect face shape. The 3D photograph and biographic information will be kept in a secure database. The sponsors will routinely be provided access to and keep this 3D photograph and biographical information database after the project is over so that it can continue to be used by researchers interested in crime prevention and detection. It will not be used for any purpose other than scientific and technical research. If you initially decide to take part you are still free to withdraw at any time without giving a reason and your database record will be destroyed.

In addition to the biographical information identified above, we will also record your name. We need to record your name in case you ask us to remove your data later on. If you want further information about the project we will also record your email address. Your name and email address (if you provide it) will be stored in a secure database that will be separated from the database containing your 3D photograph and biographical information. The University of Sheffield organizers will maintain control and access to this separate database. A unique key, allocated by the University of Sheffield researchers, will reside in both databases, providing us with the ability to destroy your database record, should you request it.

Except as described above, your 3D photograph will not be made public or distributed outside of the scientific, technical or research community and we will not publish any other personal information that will allow you to be specifically identified with your 3D photograph. Your name and email address will not be

made public or distributed beyond the Sheffield University researchers engaged in this project.

What will happen to the results of the research study?

The results will be part of a research project due to be completed in Autumn 2005. The scientific results will be published and it is intended that new tools for comparing faces which result from this research will be made available to police, defence lawyers and courts.

Who has reviewed this study?

The Research Ethics Committee at the University of Sheffield has reviewed this study.

Contact for further information:

Dr Martin Evison, Department of Forensic Pathology, The University of Sheffield, The Medico-Legal Centre, Watery Street, Sheffield, S3 7ES, United Kingdom. Tel. +44 114 2738721, Fax. +44 114 2798942, Email. m.p.evison@sheffield.ac.uk.

THE UNIVERSITY OF SHEFFIELD

Computer-Assisted Facial Comparison Project

Biographic Details

CONFIDENTIAL

Please provide the following information:-

Age: ☐☐ Sex (m / f): ☐

How would you describe your ancestry / ethnicity?

Please tick ✓ in the box that applies to you.

		Code	
White			
British	☐	01	*Please describe*
Any other White background (Please describe)	☐	02	
Mixed			
White and Black Caribbean	☐	03	
White and Black African	☐	04	
White and Asian	☐	05	*Please describe*
Any other Mixed background (Please describe)	☐	06	
Asian or Asian British			
Indian	☐	07	
Pakistani	☐	08	
Bangladeshi	☐	09	*Please describe*
Any other Asian background (Please describe)	☐	10	
Black or Black British			
Caribbean	☐	11	
African	☐	12	*Please describe*
Any other Black background (Please describe)	☐	13	
Chinese or other ethnic group			
Chinese	☐	14	*Please describe*
Any other (Please describe)	☐	15	

Are any of your blood relatives participating in this study? Please give details:

Name	Date of Birth

THE UNIVERSITY OF SHEFFIELD

Computer-Assisted Facial Comparison Project

Consent Form

CONFIDENTIAL

Family Name:

First Name:

Date of Birth: d d m m y y

Please read the statement and tick ✓ in the boxes that apply to you:-

I have read and understood the above information. I consent to the collection, long term storage and retention and use of my image and biographic information for scientific and technical research in the UK and elsewhere. I certify that the information disclosed by me is true and accurate to the best of my knowledge. ☐

I would like to be kept informed about this study. I agree for my email address to be kept on computer by the University of Sheffield in order to facilitate their disclosure of periodic updates, time and resources permitting. I understand that this e-mail information will be stored separately and will be controlled and accessed only by the organizers of this project, and will not be distributed to any other party. ☐

Email:

Signature:_____

Signature of Parent or Guardian if under 18:_____

Date: _____

Thank you for participating in this study!

For project use only

Scanner: ☐ 1 Geometrix 2 Cyberware 3 Both **Location:** ☐ 1 Magna

Key: 0

Date: d d m m y y

Operator:

12 Appendix B – Landmark Placement Manual

LANDMARKING PROTOCOL

This report is based on the landmarking procedure developed for the *IDENT* project during the initial investigative studies. When determining the landmarking method, both intra- and inter-observer errors were taken into consideration. This protocol is a guide for all landmarkers using the *IDENT* system, and should be followed closely to ensure consistency of the results.

Each of the cameras was given an abbreviated name, see Table 1. The majority of the anthropological information given is taken from Farkas (1994), with author additions and amendments where necessary, See Tables 2-19. When "left" and "right" are discussed, it must be noted that this is standard anatomical siding, and therefore the volunteers left and right, not as the pictures are viewed. In order to facilitate the correct placement of landmarks, the camera views selected favour the correct aspect of the face.

The aim of the Landmarking Protocol is to furnish the technician with all of the information required to assist the correct, accurate, and replicable placement of each landmark. Any questions regarding the method of landmarking should be forwarded to the authors of this protocol, Xanthé Mallett or Lucy Morecroft.

Table 1. Camera Name Abbreviations.

CAMERA VIEW	ABBREVIATION
LEFT PROFILE	LP
RIGHT PROFILE	RP
LEFT TOP	LT
RIGHT TOP	RT
LEFT BOTTOM	LB
RIGHT BOTTOM	RB
CENTRE TOP	CT
CENTRE BOTTOM	CB

FIGURES

Below are the Tables and Figures which give details of all of the landmarks included in the *IDENT* technique. The variables are listed in order of anatomical area from which the landmark is taken. Each variable is represented individually in a Table, followed by a visual representation of the correct position for each landmark. Included in the Tables are the name of the region, the number of the variable, the abbreviated term, information regarding which variables are bilateral, details from Farkas (1994), any additional notes for the placement of each landmark, the camera selection, and Figure number for each landmark.

Table 2. Landmark 1, Glabella.

REGION	NUMBER	ABBRV.	LEFT (L) & RIGHT (R)	FARKAS DESCRIPTION (1994)	NOTES	CAMERA SELECTION		FIGURE NUMBER
						1	2	
Head	1	g		The most prominent midline point between the eyebrows and is identical to the bony glabella on the frontal bone.		RP	CB	1

Figure 12.1. The Glabella (g).

Table 3. Landmark 2, Sublabiale.

Region	Number	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Face	2	sl		Determines the lower border of the lower lip or the upper border of the chin.		RP	CB	2

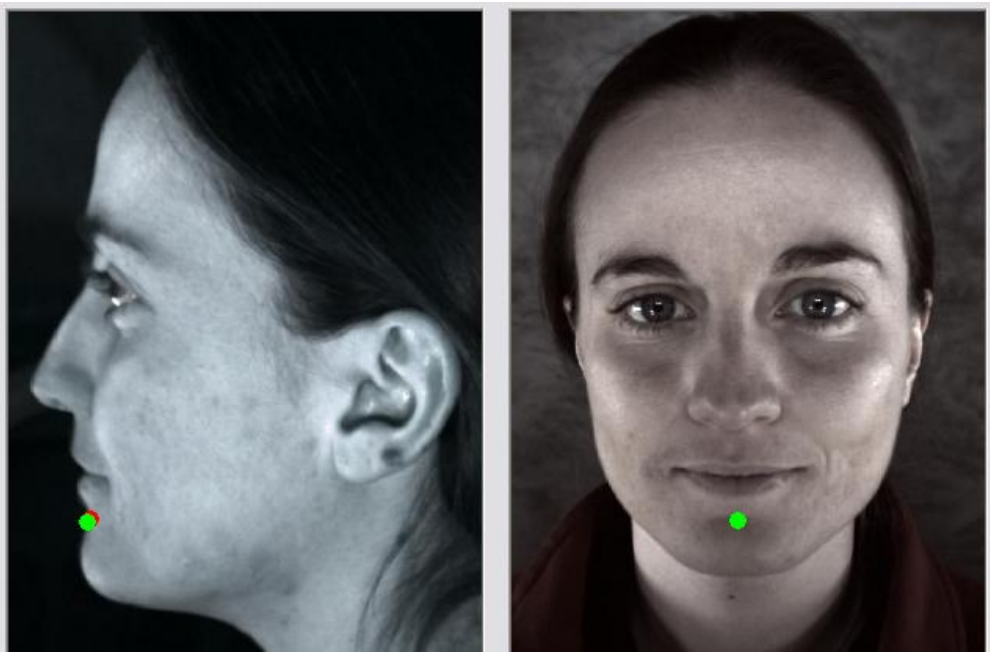


Figure 12.2. The Sublabiale (sl).

Table 4. Landmark 3, Pogonion.

Region	Number	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Face	3	pg		The most anterior midpoint of the chin, located on the surface in front of the identical bony landmark on the mandible.		LP	CB	3

Figure 12.3. The Pogonion (pg).

Table 5. Landmarks 4-5 Endocanthion, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Orbit	4-5	en	✓	The point at the inner commissure of the eye fissure.	Not placed on the eye itself, but the most medial corner of the actual fissure.	CB	RT	4
						CB	LT	5

Figure 12.4. Endocanthion (en), Left.

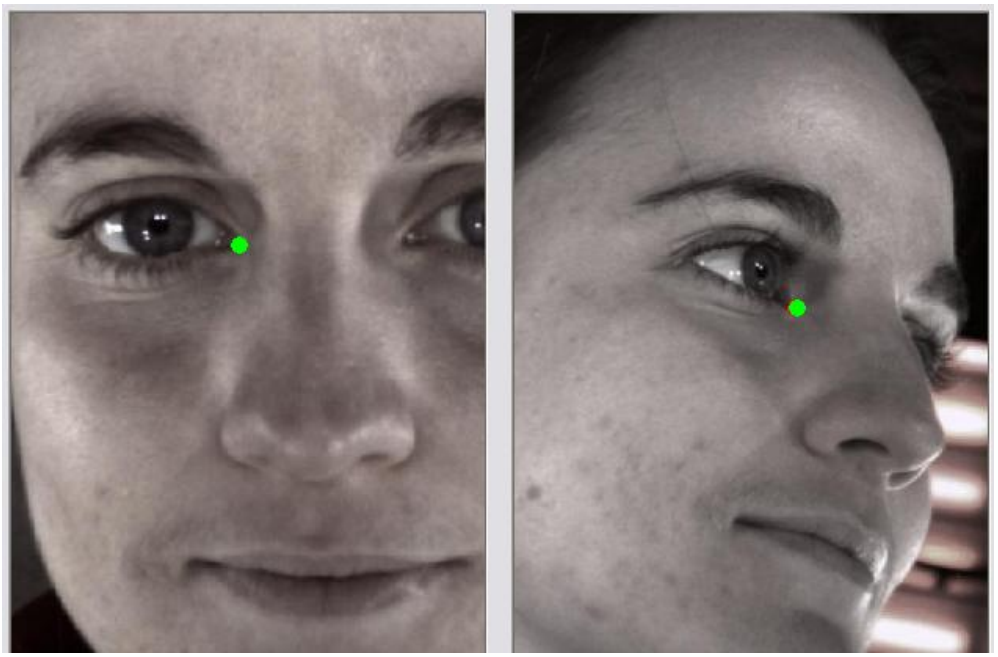


Figure 12.5. Endocanthion (en), Right.

Table 6. Landmarks 6-7, Endocanthion, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Orbits	6-7	ex	✓	The point at the outer commissure of the eye fissure. The soft exocanthion is slightly medial to the bony exocanthion.	Not placed on the eye itself, but the most lateral corner of the actual fissure.	CB	RT	6
						CB	LT	7

Figure 12.6. Exocanthion (ex), Left.

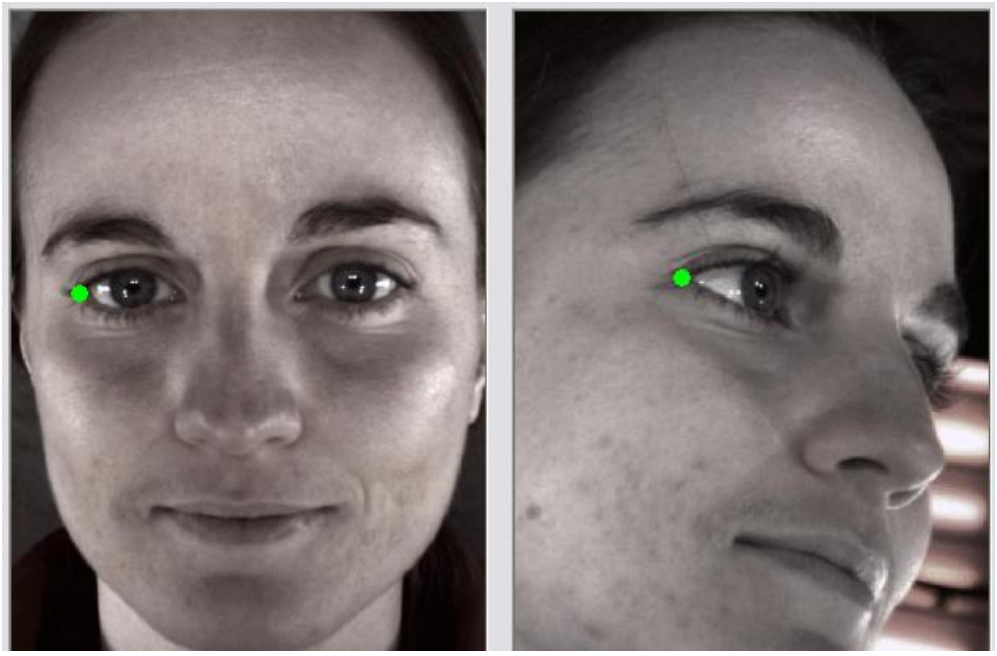


Figure 12.7. Exocanthion (ex), Right.

Table 7. Landmarks 8-9, Pupil, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Orbits	8-9	p	✓	Determined when the head is in the rest position and the eye is looking straight forward.		CB	RT	8
						CB	LT	9

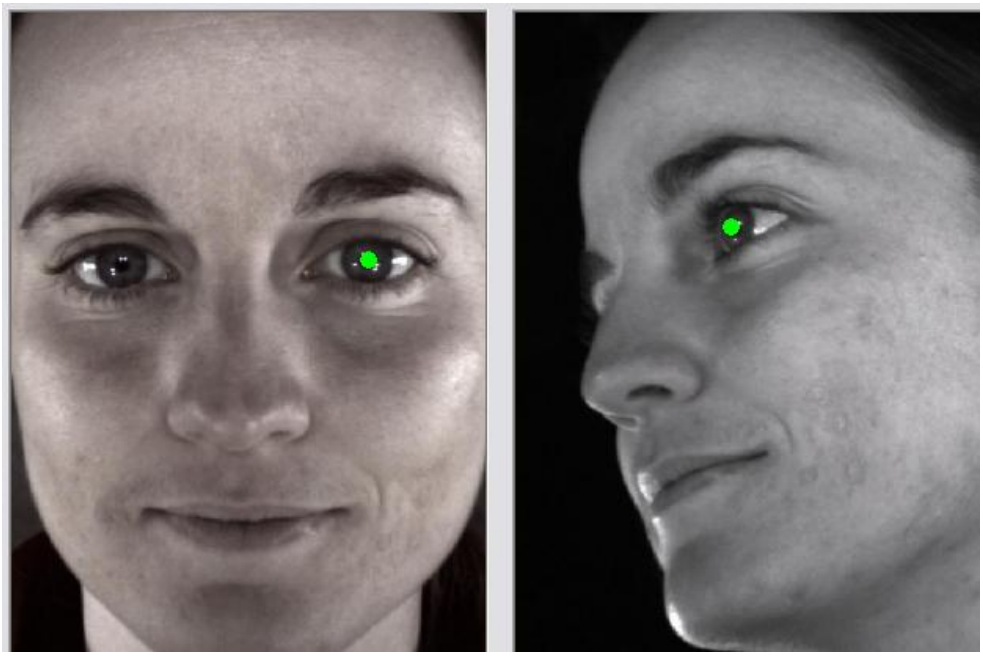


Figure 12.8. Pupil (p), Left.

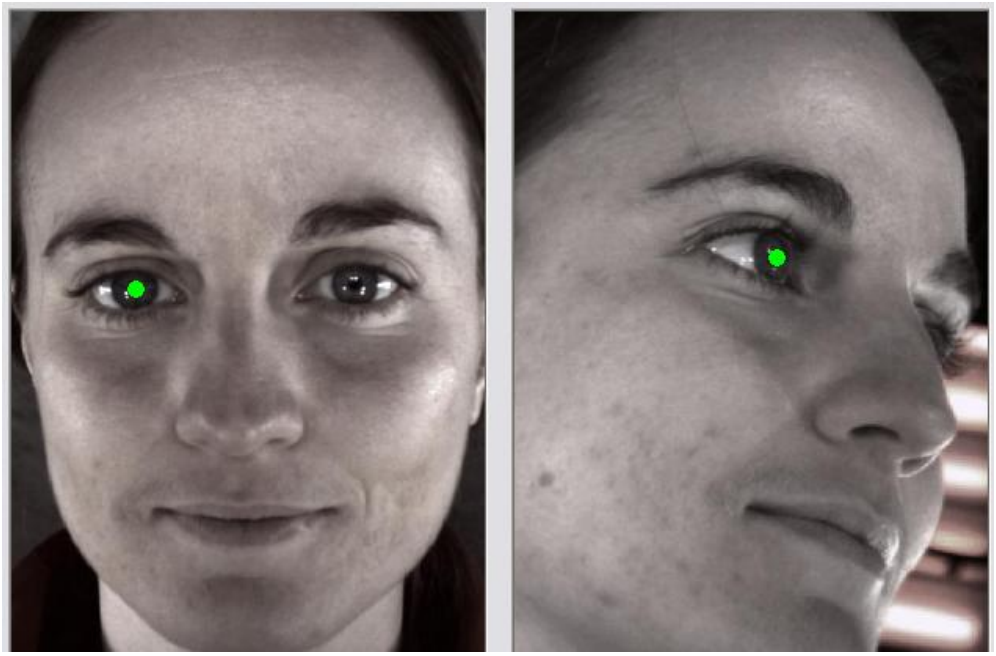


Figure 12.9. Pupil (p), Right.

Table 8. Landmarks 10-11, Palpebrale Inferius, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Orbits	10-11	pi	✓	The lowest point in the midportion of the free margin of each lower eyelid.		CB	RT	10
						CB	LT	11

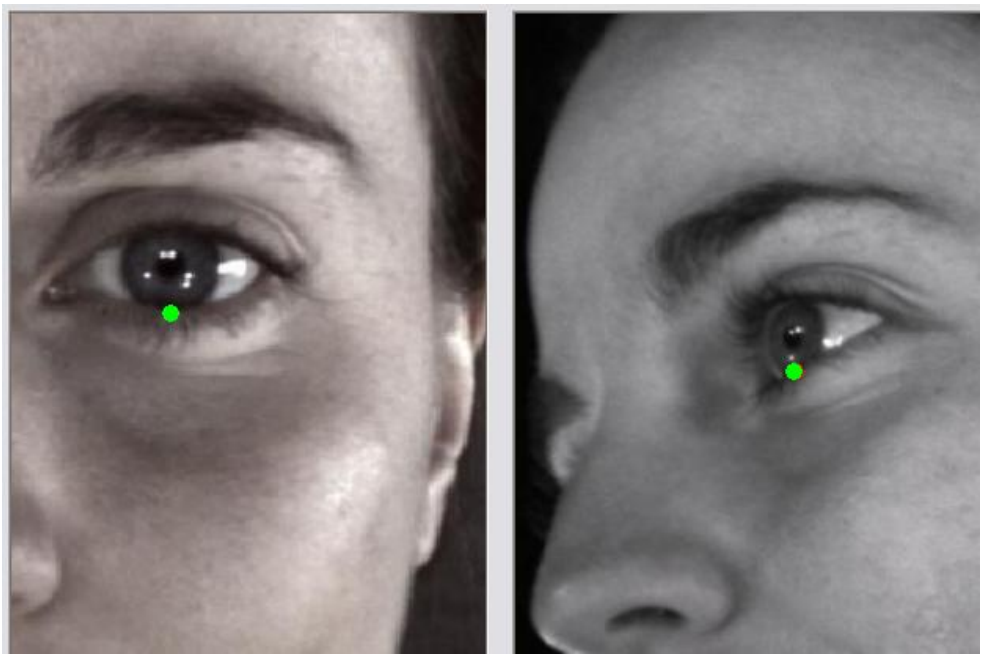


Figure 12.10. Left Palpebrale Inferius (pi), Left.

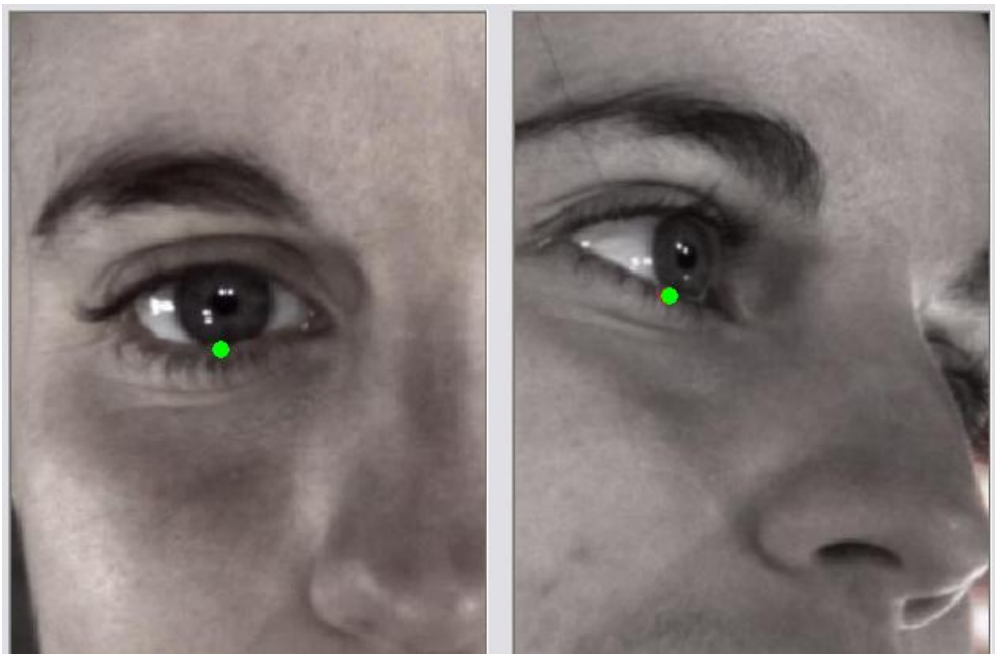


Figure 12.11. Palpebrale Inferius (pi), Right.

Table 9. Landmark 12, Sellion.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Nose	12	se		The deepest landmark located on the bottom of the nasofrontal angle. Commonly also marked as "m" (median). The point usually occurs somewhere between the levels of the supratarsal fold of the eyelash.		RP	CB	12

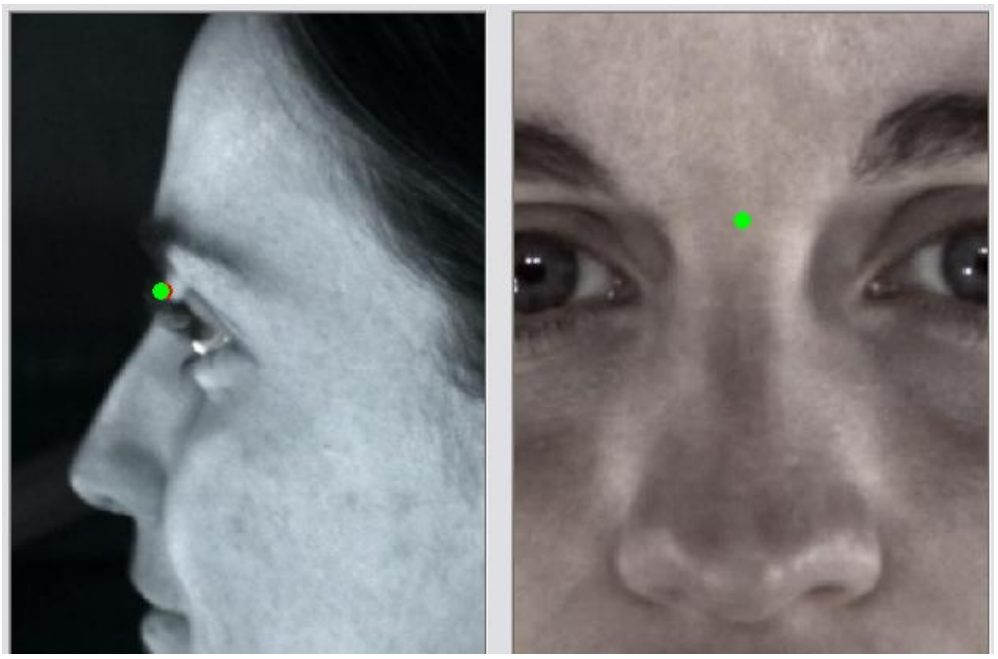


Figure 12.12. Sellion (se).

Table 10. Landmark 13, Pronasale.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Nose	13	prn		The most protruded point of the apex nasi. This point is difficult to determine if the nasal tip is flat. In the case of the bifid nose, the more protruding tip is chosen for prn		LP	CB	13

Figure 12.13 - Pronasale (prn).

Table 11. Landmarks 14-15, Alar, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Nose	14-15	al	✓	The most lateral point on each alar contour.		CT	CB	14
						CT	CB	15

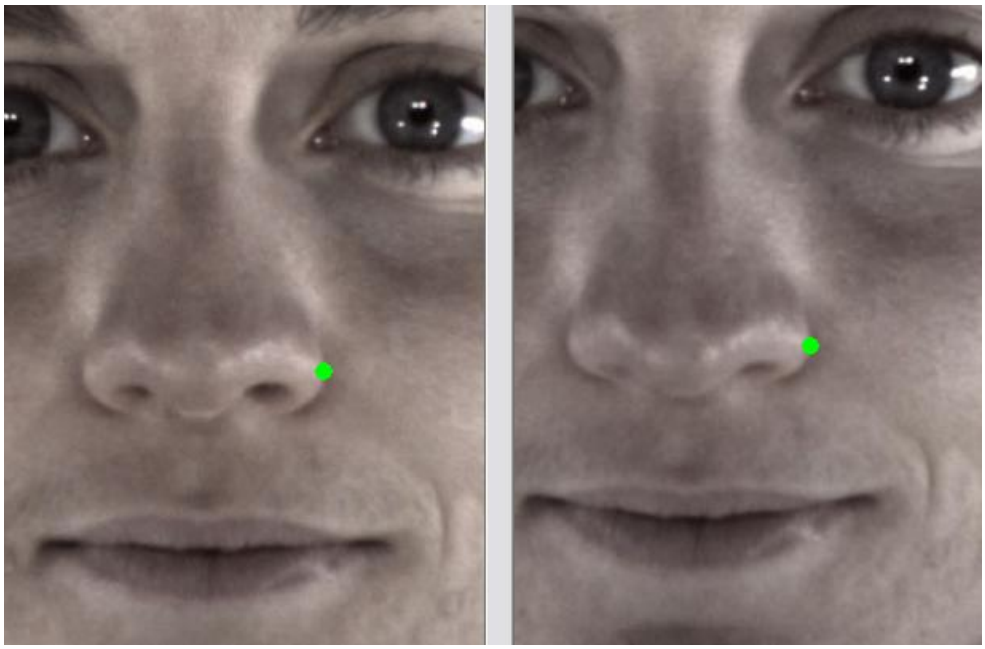


Figure 12.14. Alar (al), Left.

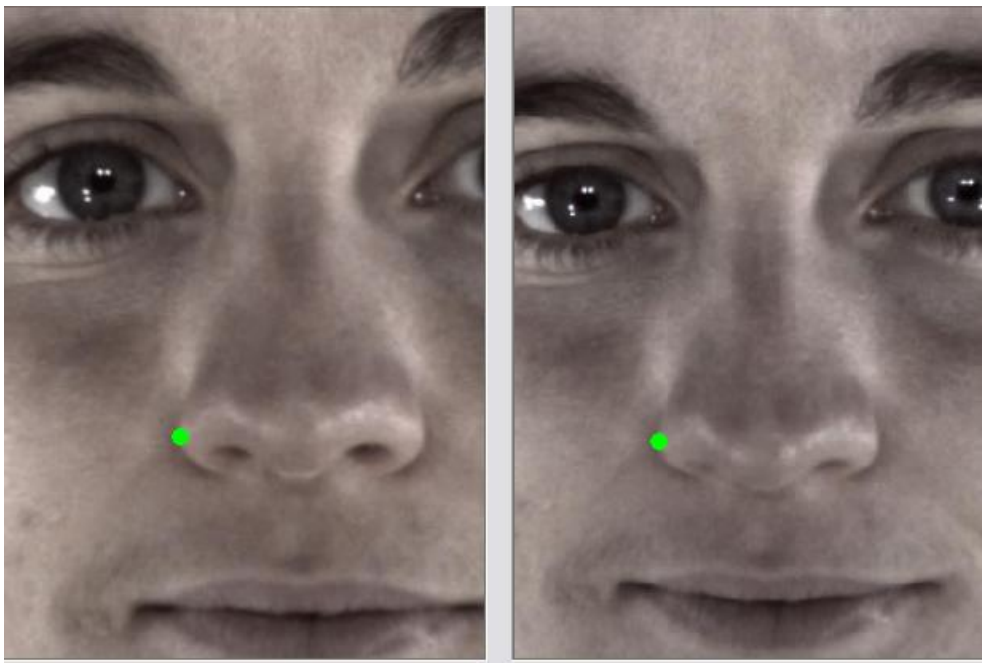


Figure 15. Alar (al), Right.

Table 12. Landmarks 16-17, Highest Point of the Columella, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Nose	16-17	c’	✓	The point on each columella crest, level with the top of the corresponding nostril.		RT	RB	16
						LT	LB	17

Figure 16. Highest point of left columella (c').

Figure 17. Highest point of the right columella (c').

Table 14. Landmark 20, Labiale Superius.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Mouth	20	ls		The midpoint of the upper vermillion line.	Instead of' 'cph'	CT	RB	20

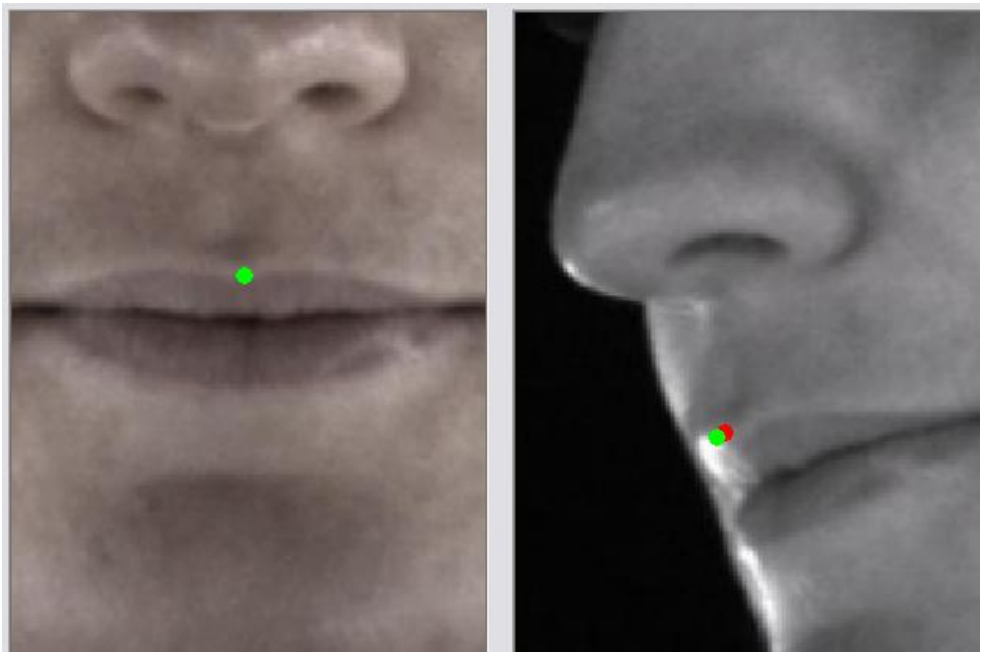


Figure 20. Labiale Superius (ls).

Table 15. Landmark 21, Labiale Inferius.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Mouth	21	li		The midpoint of the lower vermillion line.		CT	RB	21

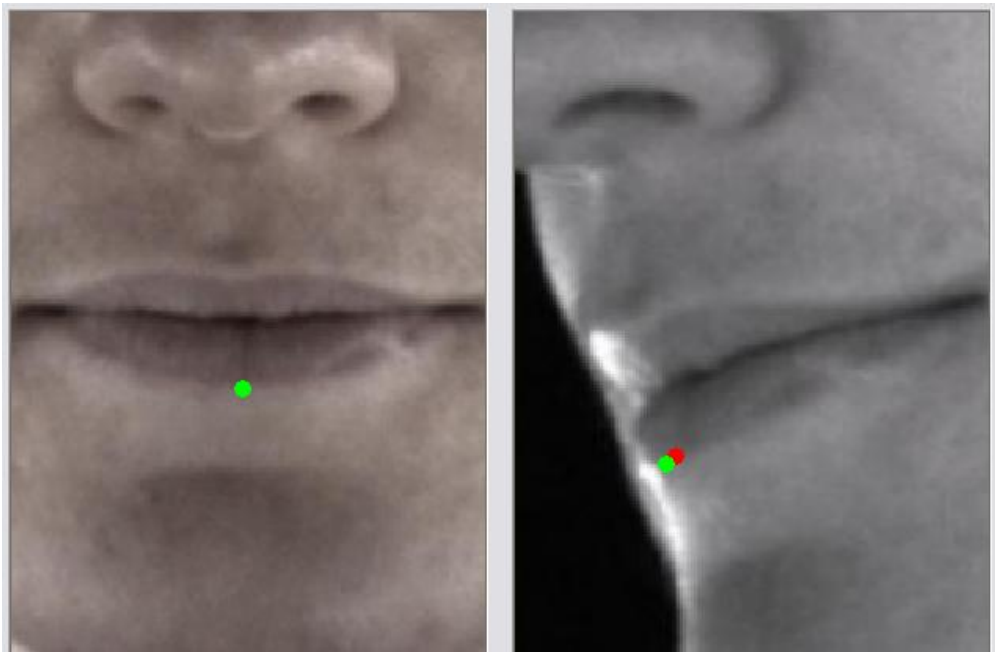


Figure 21. Labiale Inferius (li).

Table 16. Landmark 22, Stomion.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Mouth	22	sto		The imaginary point at the crossing of the vertical facial midline and the horizontal labial fissure between gently closed lips, with teeth shut in the natural position.		CT	LP	22

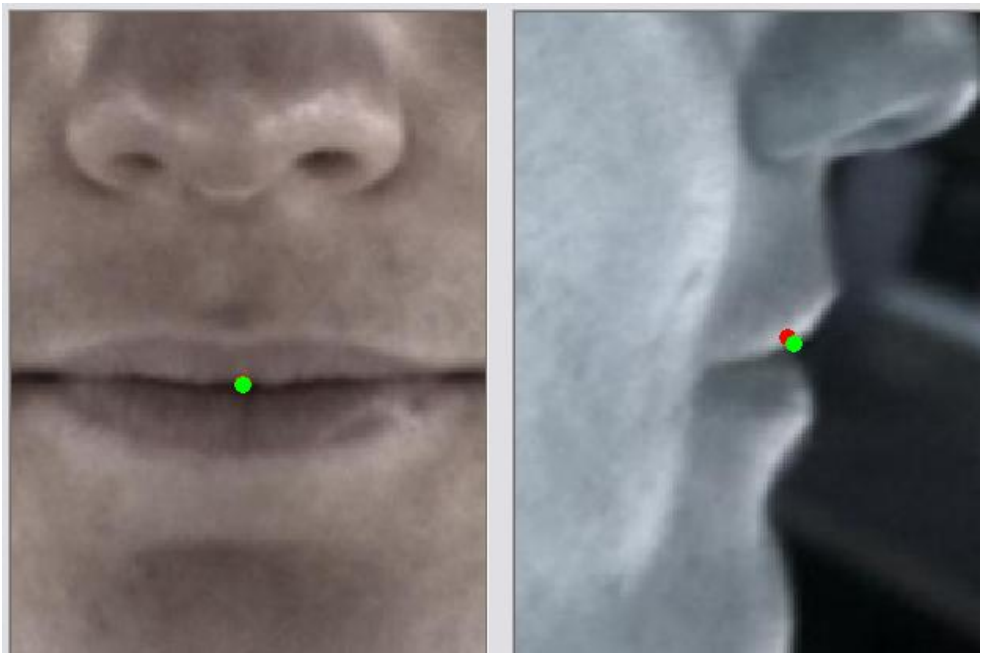


Figure 22. Stomion (sto).

Table 17. Landmarks 23-24, Cheilion, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Mouth	23-24	ch	✓	The point located at each labial commissure.	The edge of the mouth, not the lips. Includes the shadowed area at the very corners of the mouth; place landmark at most lateral point.	CT	RT	23
						CT	LT	24

Figure 23. Cheilion (ch), Left.

Figure 24. Cheilion (ch), Right.

Table 17. Landmarks 25-26, Superaurale, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Mouth	25-26	sa	✓	The highest point on the free margin of the auricle.		RP	CT	25
						LP	CT	26

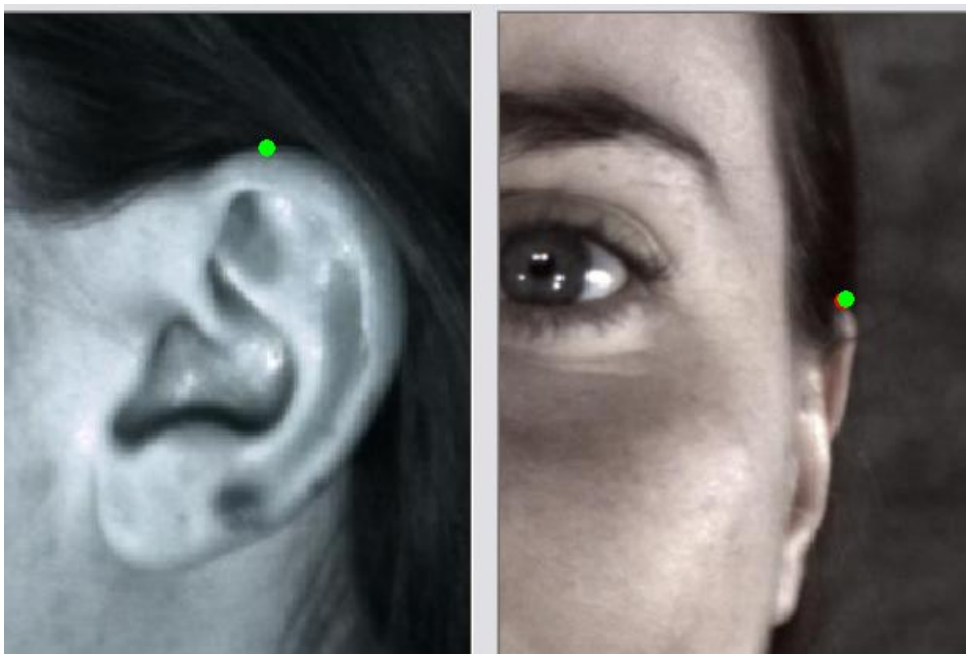


Figure 25. Superaurale (sa), Left.

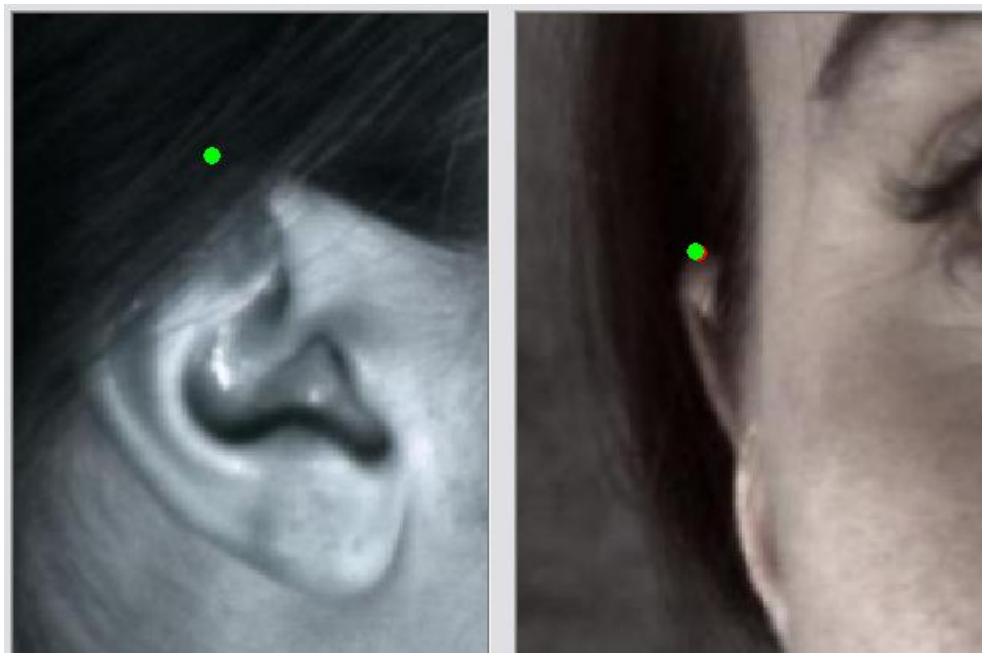


Figure 26. Superaurale (sa), Right.

Table 18. Landmarks 27-28, Subaurale, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Mouth	27-28	sba	✓	The lowest point on the free margin of the ear lobe.		RP	CT	27
						LP	CT	28

Figure 27. Subaurale (sba), Left.

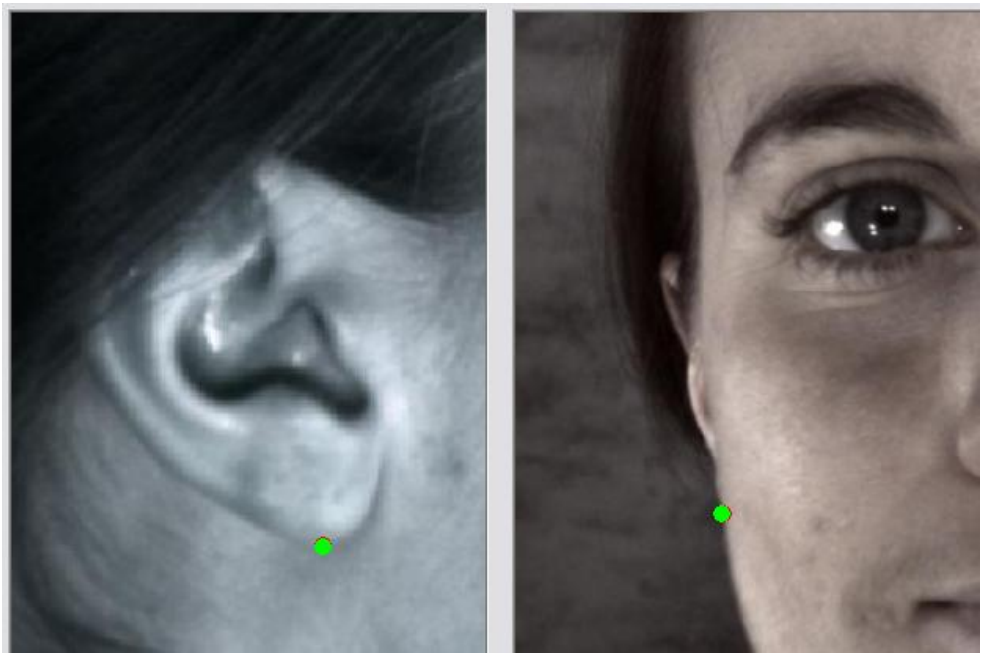


Figure 28. Subaurale (sba), Right

Table 19. Landmarks 29-30, Postaurale, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Mouth	29-30	pa	✓	The most posterior point on the free margin of the ear.		RP	RB	29
						LP	LB	30

Figure 29. Postaurale (pa), Left.

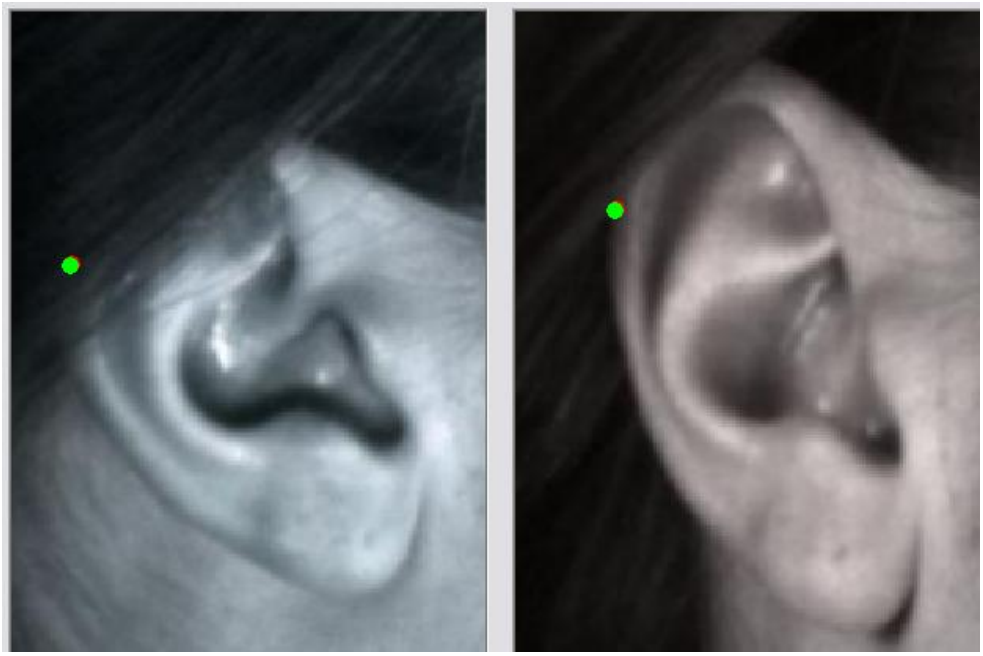


Figure 30. Postaurale (pa), Right.

Table 19. Landmarks 31-32, Otobasion Inferius, Left & Right.

Region	#	Abbrv.	Left (L) & Right (R)	Farkas Description (1994)	Notes	Camera Selection		Figure
						1	2	
Mouth	31-32	obi	✓	The point of attachment of the ear lobe to the cheek. It determines the lower border of the ear insertion.		RP	RB	

Figure 31 Otobasion Inferius (obi), Left.

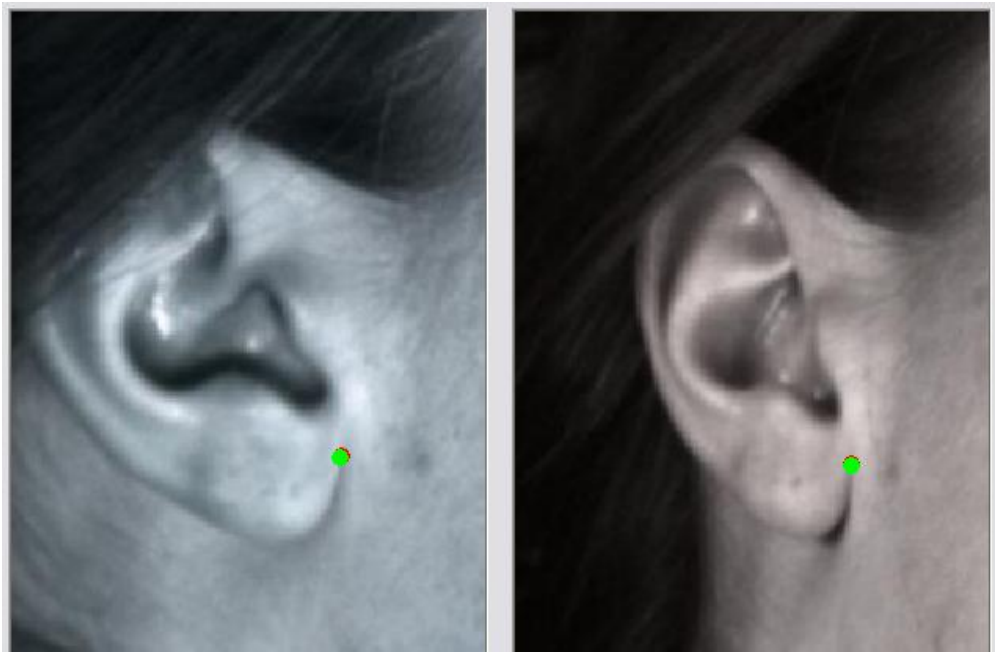


Figure 32. Otobasion Inferius (obi), Right.

13 Appendix C - Confidential Image Data

(See library copy of thesis)

14 Appendix D - Results

14.1 Twenty-two Anterior Facial Landmarks

14.1.1 The Data

Initially the subset of twenty-two anterior facial landmarks (Table 14.1, Figure 14.1) was considered, all subsets of the first ten PCs were investigated as potential facial matching variables. The background data used for the LR calculations were the landmark coordinates for 2572 observations (duplicated measurements from 1286 faces) from the Geometrix® facial database (§2.4) plus 116 observations (replicated measurements from fifty-eight faces) from the FBI anterior test data (§2.7.2). The data were Procrustes aligned to remove the arbitrary differences in scale, rotation and location (§3.6). A PCA was carried out on the tangent coordinates (§3.7) to transform the aligned data into a set of uncorrelated variables (PCs). The first ten PCs were taken as the *p* variables from which subsets to perform the LR calculations were selected, these first ten PCs represented 81.3% of the variation in the data.

Landmark	Name
1	Glabella
2	Sublabiale
3	Pogonion
4	Endocanthion Left
5	Endocanthion Right
6	Exocanthion Left
7	Exocanthion Right
8	Centre point of pupil Left
9	Centre point of pupil Right
10	Palpebrale inferius Left
11	Palpebrale inferius Right
12	Subnasion
13	Pronasale
14	Alare crest Left
15	Alare crest Right
16	Highest point of columella prime Left
17	Highest point of columella prime Right
18	Labiale superius
19	Labiale inferius
20	Stomion
21	Cheilion Left
22	Cheilion Right

Table 14.1 - Twenty-two anterior facial landmarks

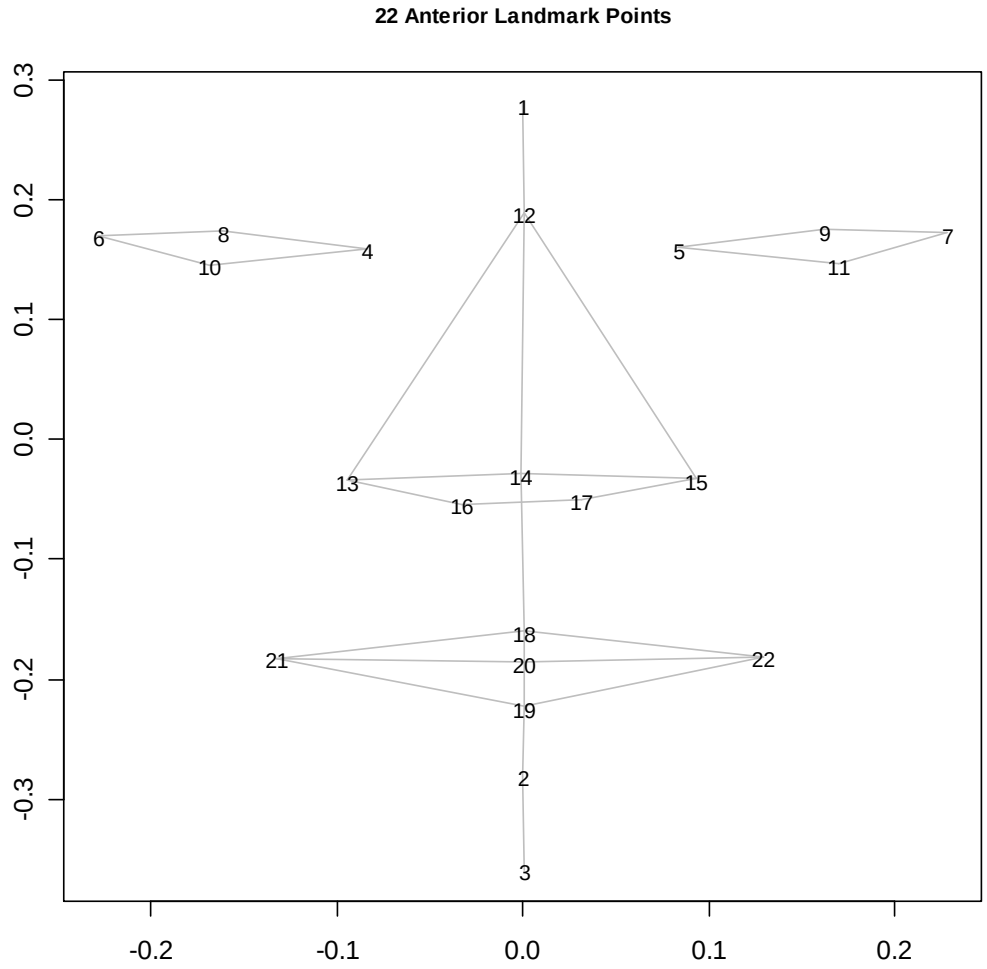


Figure 14.1 – Twenty-two anterior facial landmarks

An examination of the loadings for the first ten PCs of the twenty-two landmarks was carried out to see which PCs provided valuable information to differentiate faces from one another, Figures 14.2 and 14.3. PCs which showed little variation were of less use as facial matching variables.

The landmark configurations for the twenty facial comparison pairs (Table 7.7) were taken with the background data for the twenty-two anterior landmarks. Using the MVNLR method (§3.8.4, §7.2.1.2, §7.2.2.1) LRs were calculated for each facial comparison pair for each of the 1013 possible subsets of PCs. All possible subsets (of size >2) of the first ten PCs were examined using the same twenty facial comparison pairs (§7.4, Table 7.7). An average LR for known matches, an average LR for known exclusions and the MER were calculated (§7.4). The 'best' few subsets in terms of those

with large average LRs and MERs close to one were examined to see how well they performed at matching the fifty-eight FBI anterior faces (§2.7.2).

Results for the 'best' subsets found from twenty-two landmarks can be found in the following subsection.

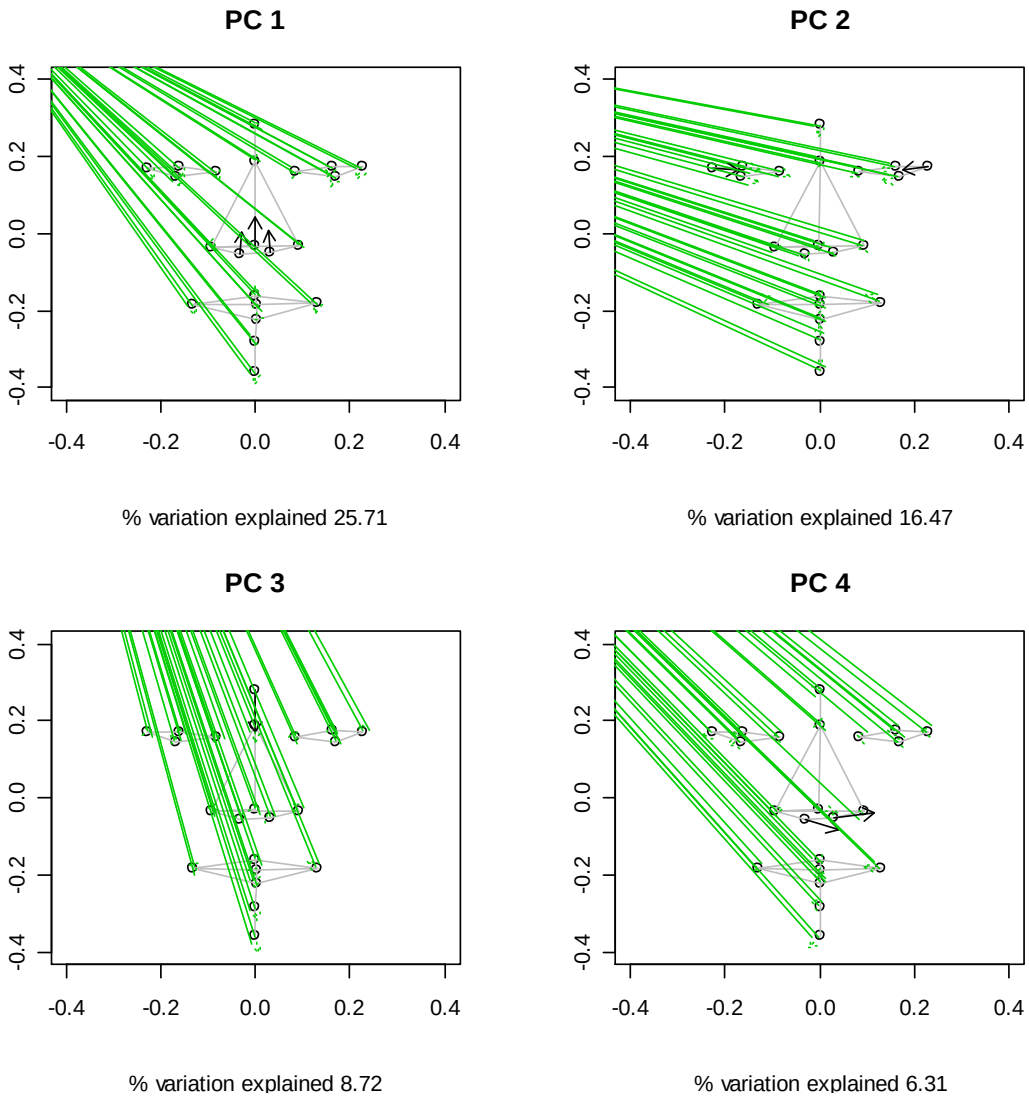


Figure 14.2- PC plots (PCs 1-4) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows.

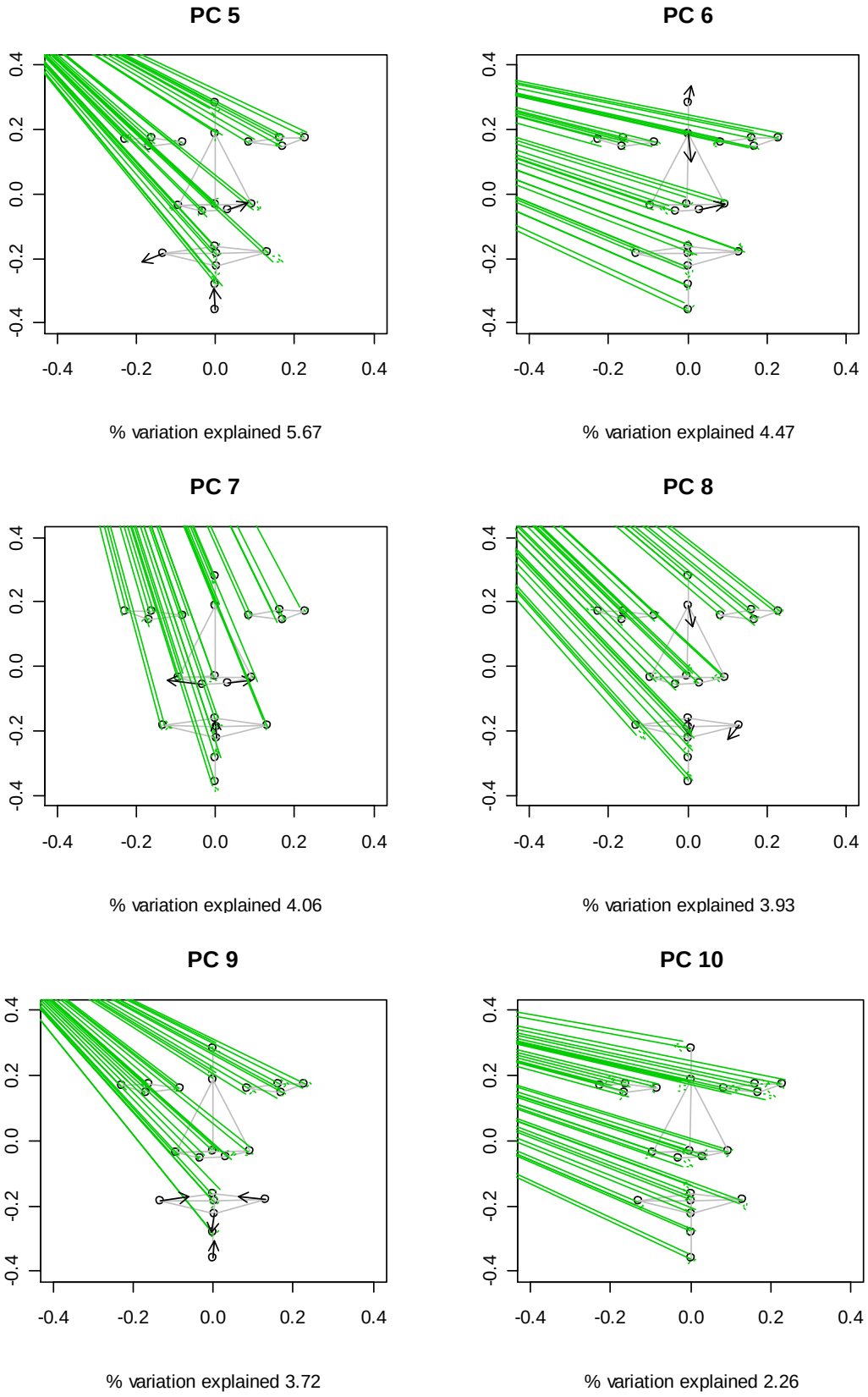


Figure 14.3 –PC plots (PCs 5-10) to show the directional effect of loadings for each facial landmark from the mean face, large loadings (>0.05) are indicated by the solid black arrows.

14.1.2 Results

The methods described in §7.4 were used to search for the best subsets in the twenty-two anterior facial landmarks (§7.5.1). Subsets with MER close to one were examined, Table 14.2. The average LR values for the known matches ranged from 3.4 to 278265, the average LR values for known exclusions ranged from 8.5e-19 to 19.6. Therefore no average results were false for known matches, however some were for known exclusions (i.e. had LRs>1).

168 subsets produced false results for the average LR for known exclusions, these subsets were discarded. The MER for the remaining 845 subsets ranged between 1.3e-14 and 22345. Table 14.2 shows the subsets which had the 'best' MERs, i.e. nearest to one. The two 'best' in terms of those with good magnitude of LR results were subsets 9 (PCs 2, 3, 6, 7, 8 and 9) and 12 (PCs: 2, 5, 6, 7, 8, 9 and 10). These two subsets were examined for their performance at facial matching with the FBI anterior database (§14.1.1).

Subset number	PCs	Average Exclusion LR	Average Match LR	MER
1	57	0.1767	4.6	0.8102
2	158910	0.0046	177.8	0.8136
3	1459	0.0036	225.8	0.8189
4	156	0.0057	149.5	0.8530
5	78	0.2579	3.4	0.8891
6	1810	0.0730	12.3	0.8980
7	135910	0.0044	207.0	0.9049
8	15610	0.0041	222.6	0.9235
9	236789	0.0005	1775.3	0.9648
10	1359	0.0079	137.9	1.0954
11	14589	0.0020	553.9	1.1138
12	25678910	0.0003	3554.7	1.1150
13	3579	0.0168	66.2	1.1156
14	567	0.0172	66.3	1.1425
15	237910	0.0125	94.7	1.1794

Table 14.2 - The top subsets in terms of MER close to one for twenty-two anterior facial landmarks (§7.5.1). Also included are the average LRs obtained for known matches and known exclusions.

14.1.3 Subset Performance

As in §7.6.1.2 to explore how well subsets 9 and 12 performed at matching faces the PCs in the subsets were used in LR calculations to carry out facial comparisons on all pairs of faces in the FBI anterior dataset (§2.7.2). The numbers of matches and exclusions were examined for several thresholds of LR. Tables 14.3 and 14.4 show the percentages of true matches and exclusions and also false positive and negative results obtained from using subsets 9 and 12 respectively to quantify facial matches.

Threshold	Yes	No	Possible	Supposed	Unverifiable
LR>1	37.6	49.2	6.6	6.6	0.0
LR>100	66.2	18.5	3.1	12.3	0.0
LR>300	78.8	5.8	15.4	0.0	0.0
LR<1	2.6	95.8	1.0	0.4	0.2
LR<0.01	1.8	97.4	0.4	0.2	0.2
LR<0.00333	1.6	97.7	0.4	0.2	0.2

Table 14.3 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 9 (PCs 2,3,6,7,8 and 9).

Threshold	Yes	No	Possible	Supposed	Unverifiable
LR>1	46.5	41.1	6.2	6.2	0.0
LR>100	68.9	16.4	3.3	11.5	0.0
LR>300	73.9	10.9	2.2	13.0	0.0
LR<1	3.1	94.9	1.2	0.7	0.2
LR<0.01	2.5	95.7	1.0	0.6	0.2
LR<0.00333	2.2	96.0	0.9	0.6	0.2

Table 14.4 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 12 (PCs 2, 5, 6,7,8,9 and 10).

Using the optimum threshold of LR>300 to quantify matches and LR<0.00333 to quantify exclusions (further details §7.6.2) subset 9 produced 5.8% false positive and 1.6% false negative results and subset 12 produced 10.9% false positive and 2.2% false negative results. Neither of these subsets performed as well as subset 6 from eleven landmarks (§7.6.1.2, §7.6.2).

14.1.4 Relating the Results back to the Matching Variables

To visualize what was happening in terms of the facial variation being included in the matching variables (PCs) in each of the subsets examined for matching performance the plots of the PC loadings were re-examined (§7.5.1, Figures 14.2 and 14.3). PC1 and PC4 didn't appear in either of the 'best' subsets found from twenty-two landmarks and examination of these show that in both the landmarks with the greater variation are around the nose, PC1 showed the alares (left and right) and the highest point of the columella (left) varied the most and PC4 showed the pronasale, alare (right) and submission varied the most (§7.5.1, Figures 14.2 and 14.3).

14.2 Fifteen landmarks

14.2.1 The Data

Next we considered taking first a subset of the original twenty-two anterior landmarks, including only the points which were found to vary a lot between faces and excluding some of the points that were known to be more difficult and subjective to place in an anterior view (§6.4.1). A PCA was carried out on this subset producing different PCs from which to selecting a 'best' subset for facial comparison variables.

Reducing the subset of landmarks was also aided by prior knowledge (§5.2.2, §5.2.4, §5.3) and by re-examining the PC loadings plots for twenty-two landmarks, Figures 14.2 and 14.3. It was seen that several landmarks that showed great variation in the PC loadings were known to be difficult to place in the anterior view; these were the glabella, subnasion, pronasale, pogonion, highest point of the columella prime (left and right) and the sublabiale (Table 14.1, Figure 14.1). The highest points of the columella primes are located in the nostrils, so can not be determined easily from an anterior view. The other points are all located down the facial midline and were excluded because in the anterior view a best guess has to be made for the position of these points. For example the pronasale is the maxima of the curve at the tip of the nose, from an anterior view this is impossible to accurately detect. Even locating the pronasale in views of both the anterior and profile triangulation between a left profile view and the anterior view may produce a different landmark location to that obtained from a right profile and anterior view. It is particularly dependant on the subject's angle to the camera and a few

degrees may alter the landmark location entirely. This was not an issue with the 3D Geometrix® data, as both the left and right profile views were captured simultaneously. However, the location of the facial midline points was determined from the anterior view and only one of the profile views, a more accurate method would be to determine these points using three or more different images to triangulate into the 3D position.

Excluding the aforementioned seven landmarks a subset of fifteen of the original landmarks was chosen and thought to provide the most valuable information for distinguishing between faces. The fifteen landmarks were Procrustes aligned (§3.6). A PCA was carried out to transform the aligned data into a set of uncorrelated data (§3.7). The first ten PCs were taken from which all subsets (of size greater than two) were investigated as the *p* variables for LR calculations. These first ten PCs contained 89.2% of the variation in the data, which is around 8% more than those for all twenty-two landmarks. Average match LRs, average exclusion LRs and MERs (§7.4) were calculated for all subsets and the 'best' few were examined for matching performance. Results for the 'best' subsets of PCs from fifteen landmarks are given in Appendix D.

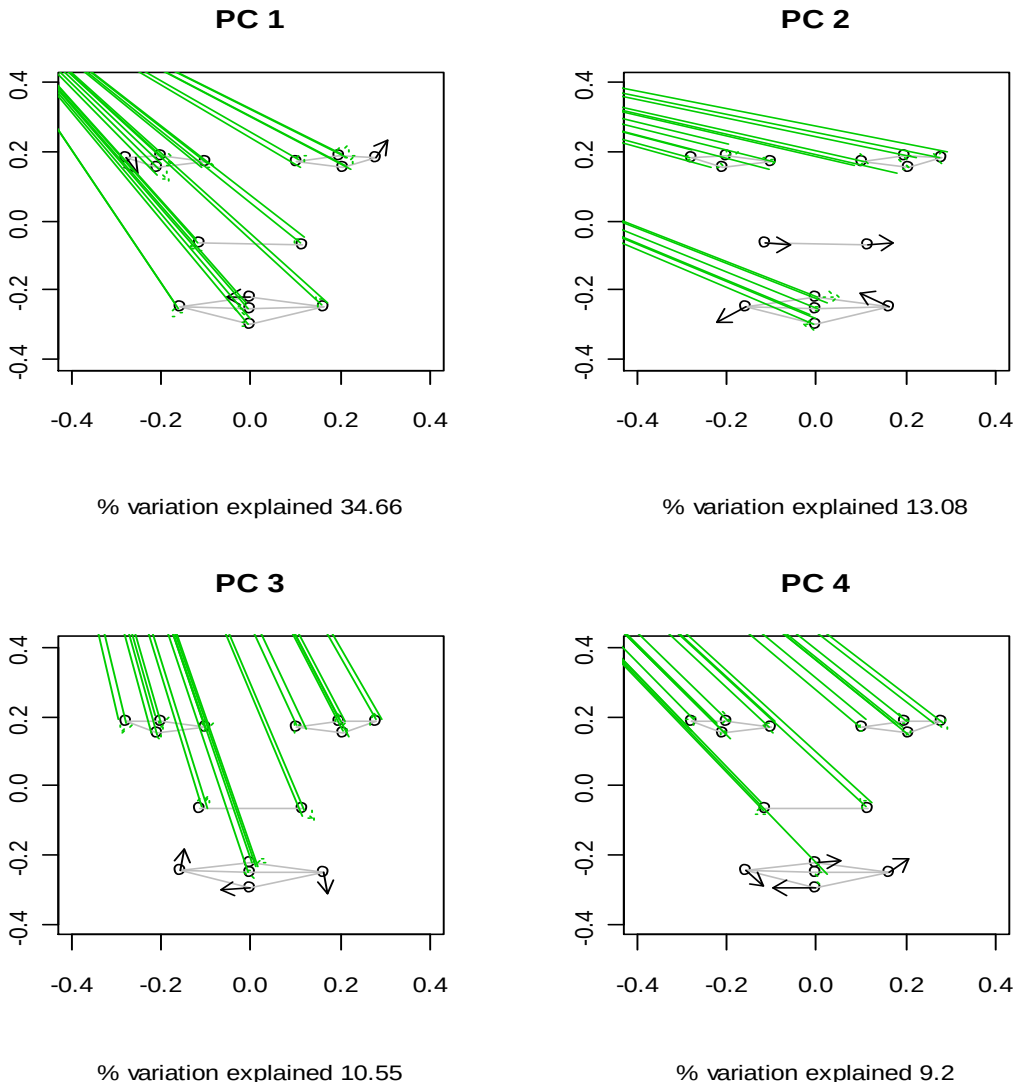


Figure 14.4 - PC loadings for the PCs 1-4 from 15 landmarks

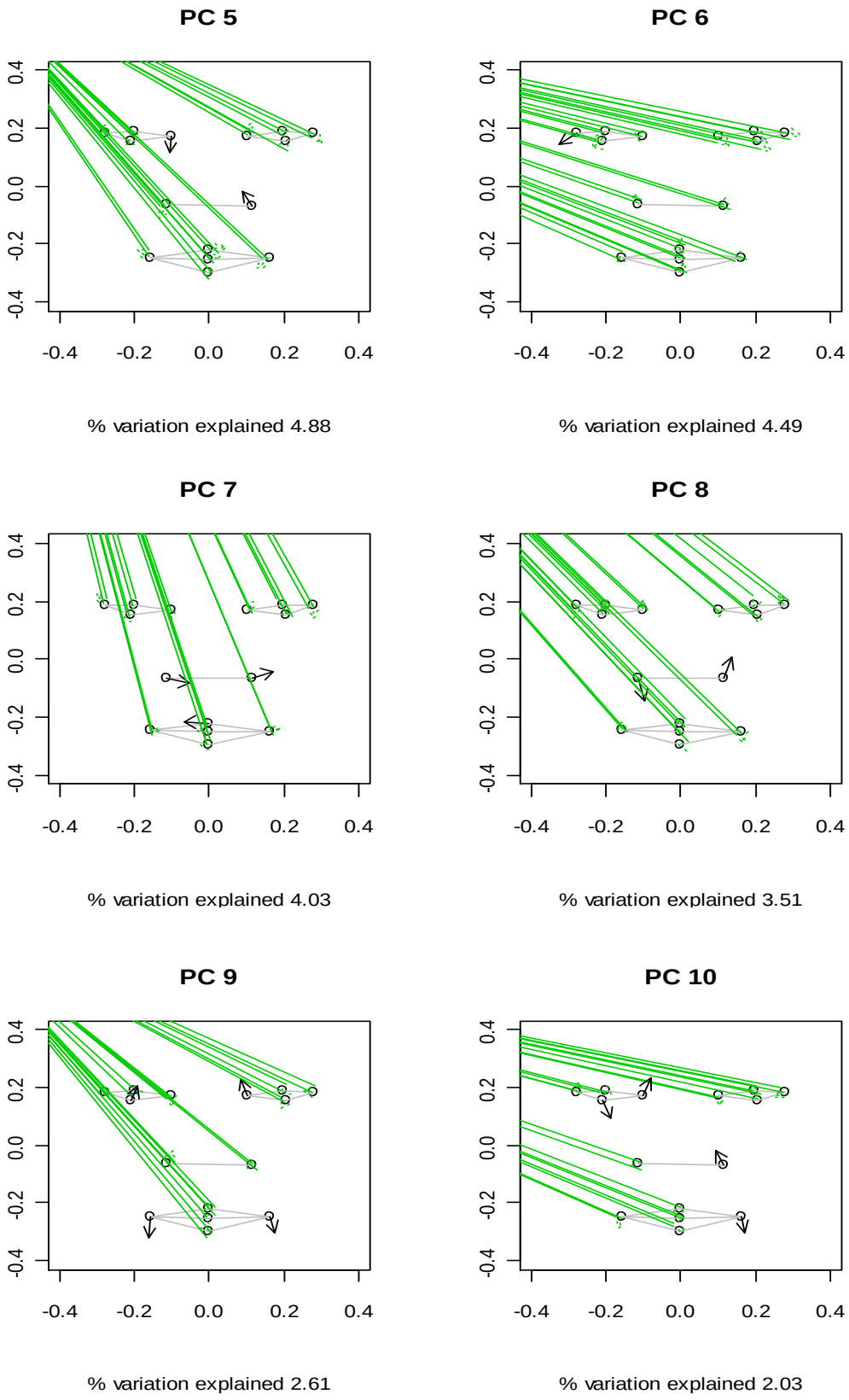


Figure 14.5 - PC loadings for the PCs 5-10 from 15 landmarks

14.2.2 Results

The methods described in §7.4 were used to search for the best subsets in fifteen anterior facial landmarks (§7.5.2). Subsets with MER close to one were examined, Table 14.5. The average LR for known matches ranged between 3.2 and 622462, for known exclusions it ranged between 7.5e-12 and 14; therefore there were average false results for known exclusions. 113 subsets produced average LRs that were false (LR>1 for known exclusions). The MERs ranged from 1.3e-06 to 30838. The top five subsets in terms of MER are displayed in Table 14.5.

Subset number	PCs	Average Exclusion LR	Average Match LR	MER
1	136789	0.00004	23255.5	0.8747
2	1239	0.00018	5447.3	0.9660
3	3456789	0.00004	30574.2	1.0972
4	3469	0.00074	1612.4	1.1960
5	210	0.09506	12.6	1.1969

Table 14.5 - The top subsets of PCs in terms of MER close to one for 15 landmarks (§7.5.2). Average LRs obtained for known matches and known exclusions are also given.

Two subsets (2 and 3) were selected as 'good' ones to explore, 'good' in terms of having MERs closest to one, and also average match and exclusion LRs of a magnitude suggesting very strong evidence (average match LR>5000 and >30000 and average exclusion LR<0.0002 and 0.00004 for subsets 2 and 3 respectively). The performance for subsets 2 and 3 was assessed as before (§7.6.1.2, §14.1.3) in terms of percentages of false positive and negative results obtained when searching for matches in the FBI anterior database (§2.7.2).

14.2.3 Subset Performance

As in §7.6.1.2 and §14.1.3 the performance of matching results obtained from subsets 2 and 3 were investigated, Tables 14.6 and 14.7 respectively. Subset 3 performed on the whole better than subset 2, with 10% of false positives at the LR>300 threshold in comparison to 25% for subset 2. False negative results were also marginally better for subset 3; all subsets investigated had a very low false negative rate in comparison with the false positive. Neither subset performed better than subset 6 from the analysis of eleven landmarks (§7.6.1.1, §7.6.1.2).

Threshold	Yes	No	Possible	Supposed	Unverifiable
LR>1	35.67	55.56	4.09	4.68	0.00
LR>100	57.81	29.69	12.50	0.00	0.00
LR>300	58.33	25.00	16.67	0.00	0.00
LR<1	3.10	94.74	1.28	0.67	0.20
LR<0.01	2.76	95.14	1.16	0.73	0.22
LR<0.00333	2.57	95.30	1.18	0.73	0.22

Table 14.6 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 2 (PCs 1, 2, 3 and 9 from 15 landmarks)

Threshold	Yes	No	Possible	Supposed	Unverifiable
LR>1	55.36	37.50	2.68	4.46	0.00
LR>100	81.69	11.27	1.41	5.63	0.00
LR>300	83.33	10.61	6.06	0.00	0.00
LR<1	2.92	94.55	1.49	0.84	0.19
LR<0.01	2.52	95.23	1.36	0.68	0.20
LR<0.00333	2.22	95.69	1.18	0.69	0.21

Table 14.7 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 3 (PCs 3, 4, 5, 6, 7, 8 and 9 from 15 landmarks)

14.2.4 Relating the Results back to the Matching Variables

The loadings for the PCs for fifteen landmarks are shown in Figures 14.4 and 14.5. The plots indicate which areas of the face have variation in each component. It is interesting to relate these plots to the results in Tables 14.5 and 14.6 to see which PCs (and therefore which areas of the face) were included in the 'best' subsets. Subset 3 containing PCs 3, 4, 5, 6, 7, 8 and 9 therefore excluding 1, 2 and 10 seems to focus on the features within the face, as opposed to the overall width of the face, indicated by the exocanthion (left and right) points in PC1.

14.3 Ten Landmarks

14.3.1 The Data

Although the matching results from subsets of eleven landmarks were found to be good (§7.6.1) one more subset of landmarks was examined to see if the results could be improved any further. Relying on anthropometric knowledge the use of a set of ten

landmarks was investigated. This was obtained from the set of eleven aforementioned by replacing the stomion (landmark 20, Table 14.1, Figure 14.1) with the labiale superius (landmark 18, Table 14.1, Figure 14.1) and excluding the sublabiale landmark (Table 14.1, Figure 14.1). The reasons for these modifications were to provide a measurement of the overall thickness of the lips, as opposed to the thickness of the lower lip provided by the stomion, and to eliminate an intermediate measure of the chin, as the pogonion already encompasses chin length. These ten landmarks were Procrustes registered (§3.6) and then, as with the other subsets (§14.1, §14.2), all subsets of the first ten PCs were investigated for facial matching. Results for the 'best' subsets of PCs from ten landmarks are given in the following section (including average LRs, MERs and assessment of performance at matching the fifty-eight FBI anterior faces (§2.7.2)).

14.3.2 Results

The methods described in §7.4 were used to search for the best subsets in ten anterior facial landmarks (§7.5.4). Here 171 subsets were found to produce averagely false results and so were excluded as potential subsets for matching faces. The average match LR ranged from 2.9 to 1069180 and the average exclusion LR ranged from 2.5e-11 to 20.3. The MERs ranged from 7.3e-08 to 189460. Subsets with MER close to one were examined, Table 14.8.

Subset number	PCs	Average Exclusion LR	Average Match LR	MER
1	34710	0.0027	331.8	0.9011
2	378910	0.0002	4004.4	0.9122
3	12367810	0.0000	153065.1	0.9643
4	23789	0.0003	3073.6	0.9673
5	2346710	0.0002	5490.9	1.0002
6	3468910	0.0001	10675.5	1.0086
7	149	0.1207	8.5	1.0306
8	1249	0.0278	37.2	1.0352
9	35789	0.0006	1810.1	1.0357

Table 14.8 - The top subsets of PCs in terms of MER close to one for 10 landmarks (§7.5.4). Average LRs obtained for known matches and known exclusions are also given.

14.3.3 Subset Performance

As in §7.6.1.2, §14.1.3 and §14.2.3 the performance of matching results obtained from subsets 5 and 6 were investigated, Tables 14.9 and 14.10 respectively. Subset 5 performed on the whole better than subset 6, with 9.8% of false positives at the LR>300 threshold in comparison to 18% for subset 2. Although false negative results were marginally better for subset 6 all subsets investigated had a low false negative rate compared to the false positive rate. Neither subset performed better than subset 6 from the analysis of eleven landmarks (§7.6.1.1, §7.6.1.2).

Threshold	Yes	No	Possible	Supposed	Unverifiable
LR>1	43.57	45.71	5.00	5.71	0.00
LR>100	70.37	20.99	8.64	0.00	0.00
LR>300	72.22	18.06	9.72	0.00	0.00
LR<1	3.04	94.84	1.26	0.66	0.20
LR<0.01	2.38	95.80	1.05	0.56	0.21
LR<0.00333	2.14	96.23	0.93	0.50	0.21

Table 14.9 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 6 (PCs 3, 4, 6, 8, 9 and 10 from 10 landmarks)

Threshold	Yes	No	Possible	Supposed	Unverifiable
LR>1	58.10	32.38	2.86	6.67	0.00
LR>100	79.17	11.11	1.39	8.33	0.00
LR>300	80.33	9.84	9.84	0.00	0.00
LR<1	2.97	94.64	1.49	0.71	0.19
LR<0.01	2.75	95.17	1.28	0.60	0.20
LR<0.00333	2.58	95.39	1.29	0.54	0.20

Table 14.10 - Percentage of true matches (LR>1 and 'Yes'), true exclusions (LR<1 and 'No'), false positive (LR>1 and 'No) and false negative (LR<1 and 'Yes') results obtained from quantifying facial matches using LRs calculated from subset 5 (PCs 2, 3, 4, 6, 7 and 10 from 10 landmarks)