Geometric Expression Invariant 3D Face Recognition using Statistical Discriminant Models

Jacey-Lynn Minoi

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

of the

Imperial College London

March 2009

Department of Computing

Imperial College London
Abstract

Currently there is no complete face recognition system that is invariant to all facial expressions. Although humans find it easy to identify and recognise faces regardless of changes in illumination, pose and expression, producing a computer system with a similar capability has proved to be particularly difficult. Three dimensional face models are geometric in nature and therefore have the advantage of being invariant to head pose and lighting. However they are still susceptible to facial expressions. This can be seen in the decrease in the recognition results using principal component analysis when expressions are added to a data set.

In order to achieve expression-invariant face recognition systems, we have employed a tensor algebra framework to represent 3D face data with facial expressions in a parsimonious space. Face variation factors are organised in particular subject and facial expression modes. We manipulate this using single value decomposition on sub-tensors representing one variation mode. This framework possesses the ability to deal with the shortcomings of PCA in less constrained environments and still preserves the integrity of the 3D data. The results show improved recognition rates for faces and facial expressions, even recognising high intensity expressions that are not in the training datasets.

We have determined, experimentally, a set of anatomical landmarks that best describe facial expression effectively. We found that the best placement of landmarks to distinguish different facial expressions are in areas around the prominent features, such as the cheeks and eyebrows. Recognition results using landmark-based face recognition could be improved with better placement.

We looked into the possibility of achieving expression-invariant face recognition by reconstructing and manipulating realistic facial expressions. We proposed a tensor-based statistical discriminant analysis method to reconstruct facial expressions and in particular to neutralise facial expressions. The results of the synthesised facial expressions are visually more realistic than facial expressions generated using conventional active shape modelling (ASM). We then used reconstructed neutral faces in the sub-tensor framework for recognition purposes.
The recognition results showed slight improvement. Besides biometric recognition, this novel tensor-based synthesis approach could be used in computer games and real-time animation applications.
Acknowledgements

The first person I would like to thank is my supervisor, Prof Dr Duncan Gillies for his excellent style of supervision. I owe him an enormous debt of gratitude for his guidance, overflowing patience and support at all stages of my research. Not only has he taught me the importance of conducting meaningful research, thinking critically about my research area, giving invaluable comments and feedback throughout, but as well he has been very supportive and encouraging. Thanks also to Prof Dr Daniel Rueckert, Dr Papatheodorou, Dr Maja Pantic and Dr Carlos Thomaz for their useful advice and help, which benefitted my research a lot.

Special thanks to Sylvester Arnab and Hassan Amin for the stimulating research and technical discussions. To friends I have made during my stay in London - Zia Wadud, Eunice Yeo, Edna Gias, Fitri Mohamad, Jimmie Vosmeer, Dana Lim, Jeffray Chung, Aaron Chaiithiwet, Sye Loong Keoh, Surapa Thiemjarus, Kevin Gregory, Eldos Then, Anita Cheah, Michelle Chan, Esther Tjokro, Supakit Tubanthoeng, Marco Visentini Scarzanella, Marco Caproni, Raghavendra Chandrashekara, Prakich Treetasayuth, PK and Kim, Terry Scott, Felix and Sophina, (and many more) - thank you for the friendship and the moral support. Anders Ericsson, Paul-Amaury, Jung-tai Chen and Maria Murgasova, I am very glad for the lunch and coffee breaks, and Capital FM 95.8 for my daily dose of music to keep me sane especially during my writing-up stage!

I would also like to thank the department’s postgraduate administrators and the CSG staff for their administrative and technical support.

The biggest thanks to my supportive and loving family back in Borneo. Without their care, I doubt I would be able to go through the research process with a smile. Their nagging is the only source of motivation for me to quickly finish off this thesis!

Lastly, to Yayasan Sarawak Tunku Abdul Rahman and Universiti Malaysia Sarawak for the
full scholarships and for giving me the opportunity to pursue my research at Imperial College London.
# Contents

1 **Introduction** ................................................................. 1
   1.1 Face recognition in brief ................................................. 4
   1.2 Motivation ........................................................................ 6
   1.3 Challenges in face recognition ........................................... 7
   1.4 Aims .............................................................................. 8
   1.5 Contributions .................................................................. 9
   1.6 The organisation of the thesis ............................................ 10

2 **Facial Expressions** .............................................................. 12
   2.1 The Measurement and Coding of Facial Expressions ............... 14
       2.1.1 Facial Anatomical Analysis ........................................ 16
       2.1.2 Facial action coding system (FACS) ............................ 20
   2.2 Facial Expression Analysis .................................................. 25
       2.2.1 Facial expression recognition ...................................... 25
       2.2.2 Facial expression modelling and animation .................... 27
       2.2.3 Issues with facial expression analysis .......................... 31
   2.3 Conclusions .................................................................... 32

3 **Reviews of Existing Face Recognition Systems** .................. 33
   3.1 Face recognition by the human brain ................................... 35
   3.2 2D face recognition .......................................................... 36
   3.3 3D face recognition .......................................................... 42
       3.3.1 3D acquisition systems .............................................. 42
CONTENTS

3.3.2 Surface-based techniques ........................................... 45
3.3.3 Statistical techniques ............................................. 48
3.3.4 Model-based techniques .......................................... 50
3.4 Recognition performance measurements .......................... 54
3.5 Discussion and Conclusions ......................................... 55

4 Pre-processing ............................................................. 56
4.1 Problems in 3D face datasets ........................................ 56
4.2 Registration .............................................................. 57
  4.2.1 Registration transformation types ................................ 58
    4.2.1.1 Rigid transformation ..................................... 58
    4.2.1.2 Affine transformation .................................. 59
    4.2.1.3 Non-rigid transformation ................................ 59
  4.2.2 VisionRT 3D face scans .......................................... 60
  4.2.3 Binghamton 3D face scans ....................................... 60
  4.2.4 Notre Dame 3D face scans ....................................... 63
4.3 The pre-processing steps ............................................. 64
  4.3.1 Rigid registration ................................................. 65
  4.3.2 Non-rigid landmark registration ................................. 67
  4.3.3 Establishing correspondences ................................... 67
4.4 Experiments and Results .............................................. 70
4.5 Conclusions ............................................................. 72

5 Statistical 3D facial analysis .......................................... 74
5.1 Introduction to landmarking ......................................... 74
5.2 Craniofacial anthropometry ........................................... 76
5.3 Background on landmark-based face applications ................ 78
  5.3.1 Face registration ................................................. 78
  5.3.2 Face recognition applications .................................. 79
  5.3.3 Facial expression applications .................................. 83
5.4 Issues in Landmarking .......................................................... 89
5.5 Our proposed landmarking approach ..................................... 89
  5.5.1 Experiments and results ................................................ 90
5.6 Discussion ......................................................................... 96
5.7 Conclusions ....................................................................... 99

6 3D Face Recognition Using Multilinear Decomposition 101
  6.1 Review of Principal Component Analysis ............................... 102
  6.2 Review of Singular Value Decomposition ............................... 104
  6.3 The Fundamental Concepts of Multilinear Algebra .................. 107
    6.3.1 Tensor Flattening ....................................................... 107
    6.3.2 Mode-n Product ....................................................... 111
    6.3.3 Tensor Decompositions ............................................. 112
    6.3.4 An example of HOSVD computation .............................. 115
  6.4 Related research on tensor decompositions ............................. 118
  6.5 Experimental Method .......................................................... 121
  6.6 Results and Discussion ........................................................ 124
  6.7 Conclusions ....................................................................... 141

7 Face recognition using statistical modelling approaches 143
  7.1 Statistical Shape Models ..................................................... 144
  7.2 Statistical Discriminant Models .......................................... 147
    7.2.1 Fisherfaces Method .................................................... 148
    7.2.2 The Maximum uncertainty LDA-based Approach .......... 148
    7.2.3 Statistical Discriminant Analysis ................................ 149
    7.2.4 Tensor-based Statistical Discriminant Analysis ............. 151
  7.3 Experimental Work .............................................................. 152
  7.4 Results and Discussion ......................................................... 153
    7.4.1 Reconstruction of Faces with Expression ...................... 153
    7.4.2 Reconstruction of Neutral Faces .................................. 163
7.4.3 Quantitative Comparison of the Models in Face Recognition . . . . . . . 165
7.5 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 167

8 Conclusions and Further Works 168
8.1 Summary of contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . 169
8.2 Limitations and further work . . . . . . . . . . . . . . . . . . . . . . . . . . 170
  8.2.1 Extending the 3D face database . . . . . . . . . . . . . . . . . . . . . . . 170
  8.2.2 Improving the pre-processing method . . . . . . . . . . . . . . . . . . . 170
  8.2.3 Improving the reconstruction method in the tensor-based SDA approach 171
  8.2.4 Improving facial expression recognition . . . . . . . . . . . . . . . . . . 171
  8.2.5 Mapping face texture . . . . . . . . . . . . . . . . . . . . . . . . . . . . 172
List of Figures

1.1  The facial muscles .......................................................... 2
1.2  Variation in appearance over different subjects [117] .................. 7
1.3  Extreme variation in appearance of the same subject under different lighting conditions and poses [117] ............................................. 8
1.4  Face change over a 2-year interval [229] .................................. 8
1.5  Face change in 2-year intervals from 1978 to 2004 [229] ............... 9

2.1  Guillaume-Benjamin Duchenne and his assistant electrically stimulate the face of a live subject to display an expression [42] ............................... 13
2.2  The study of facial muscles (image taken from Duchenne [42]) .......... 13
2.3  The skull anatomy (taken from Gray’s Anatomy [126]) .................... 17
2.4  The anatomy of facial musculature. The * indicates the mastication muscles while others are facial expression muscles (images taken from Dreamstime [81]) 17
2.5  The facial sphincter muscles .................................................. 18
2.6  The temporalis muscle with the zygomatic arch and masseter muscle removed (image taken from Gray’s Anatomy [126]) ......................... 19
2.7  The pterygoideus externus and pterygoideus internus with the zygomatic arch. A portion of the ramus of the mandible has been removed (image taken from Gray’s Anatomy [126]) ......................... 19
2.8  The eleven main muscles that influence facial expression formation (taken from Erfan [91]) ............................................................. 20
2.9  The Action Units to describe facial expressions as described in FACS .... 22
2.10 Action Unit 2 and a combination of Action Unit 1+2 ....................... 24
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.11</td>
<td>Different synthesized facial expressions (taken from the Barrett Fox Character Animator [108])</td>
</tr>
<tr>
<td>3.1</td>
<td>A face recognition pipeline</td>
</tr>
<tr>
<td>3.2</td>
<td>(a) the VisionRT stereo camera system (b) captured images.</td>
</tr>
<tr>
<td>3.3</td>
<td>(a) Lee and Milios 3D shape acquisition system (b) Left: captured images. Right: Reconstructed 3D shape and texture mapping</td>
</tr>
<tr>
<td>3.4</td>
<td>An example of a 3D face model acquired using the Minolta Vivid scanner (a) surface model (b) mesh model</td>
</tr>
<tr>
<td>3.5</td>
<td>Point signature examples (a) peak, (b) ridge, (c) saddle, (d) pit, (e) valley, (f) roof edge (image taken from [64])</td>
</tr>
<tr>
<td>3.6</td>
<td>The non-rigid area of the face are removed for recognition purposes [63]</td>
</tr>
<tr>
<td>3.7</td>
<td>Texture mapping of the facial surface, A, and its canonical form, B, which results the flattened texture, C, and the canonical image, D (image taken from [46])</td>
</tr>
<tr>
<td>3.8</td>
<td>The overall steps in image manipulation to generate 3D face models used for face recognition (image taken from Blanz [36])</td>
</tr>
<tr>
<td>3.9</td>
<td>(a) Component-based face recognition using the 3D morphable model (b) Ten of the 2D components (image taken from Heisele et al. [142])</td>
</tr>
<tr>
<td>4.1</td>
<td>The transformation T(a) transforms point a in face A into its corresponding location in face B</td>
</tr>
<tr>
<td>4.2</td>
<td>An example of a non-linear transformation</td>
</tr>
<tr>
<td>4.3</td>
<td>An example of raw reconstructed faces. The images on the left are without texture, those on the right have texture</td>
</tr>
<tr>
<td>4.4</td>
<td>Different poses captured with the VisionRT 3D camera</td>
</tr>
<tr>
<td>4.5</td>
<td>The 3D face imaging system setup</td>
</tr>
<tr>
<td>4.6</td>
<td>Raw face models</td>
</tr>
<tr>
<td>4.7</td>
<td>First and third rows are the happy expression with four levels of intensity. The second and fourth rows are the surprise expression with four levels of intensity</td>
</tr>
</tbody>
</table>
4.8 An example of a raw reconstructed faces taken from Notre Dame database, which contains only neutral faces .......................... 64
4.9 The 13 landmark placements on a face surface ........................... 66
4.10 Rigid registration using landmarks information. The top row shows the two faces aligned to the mean landmarks. The bottom row shows a frontal 2D polygon projection of the outer landmarks of the same polygon before and after registration (image taken from [229]) .......................... 66
4.11 The top image shows point-pairing before non-rigid registration and the bottom image shows the point-pairing after non-rigid registration (image taken from [229]) .................................................. 67
4.12 A free-form deformation and the corresponding mesh of control points (image taken from [229]) .................................................. 68
4.13 Non-rigid registration using landmarks information. The top row shows the two faces aligned to the mean landmarks. The bottom row shows a frontal 2D polygon projection of the outer landmarks of the same polygon before and after non-rigid registration (image taken from [229]) .......................... 68
4.14 The distance colour map after non-rigid registration (image taken from [229]) .......................... 69
4.15 The difference between landmark-based non-rigid registration and non-rigid surface registration (image taken from [229]) .................................................. 70
4.16 The overall pre-processing steps (image taken from [229]) .................................................. 70
4.17 The variation described by the modes of the landmark-based model .......................... 71
4.18 The first 4 most discriminant components between $\pm 3 \sqrt{\lambda}$, where $\lambda$ is the standard deviation .................................................. 71
4.19 The shape distribution of the landmark-based model projected on the first two most discriminant modes .................................................. 72
5.1 Craniofacial landmarks superimposed on a frontal face image (image taken from [174]) .................................................. 77
5.2 Left: Main craniofacial landmark around the head. Right: Profile landmarks (image taken from [174]) .................................................. 77
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Reference</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>Craniofacial landmarks from bottom-up view</td>
<td>[174]</td>
<td>78</td>
</tr>
<tr>
<td>5.4</td>
<td>Face detection and landmark placements</td>
<td>Tao and Veldhuis [274]</td>
<td>79</td>
</tr>
<tr>
<td>5.5</td>
<td>Face scan with nine landmarks</td>
<td>Hutton [151]</td>
<td>79</td>
</tr>
<tr>
<td>5.6</td>
<td>Thirteen manually selected landmarks</td>
<td>Papatheodorou [229]</td>
<td>80</td>
</tr>
<tr>
<td>5.7</td>
<td>Three-dimensional polyhedral face shape (a)</td>
<td></td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Polyhedral face shape based on a set of landmarks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.8</td>
<td>The location of the twenty-nine landmarks used by Shi et al. [264]</td>
<td></td>
<td>81</td>
</tr>
<tr>
<td>5.9</td>
<td>Face image showing the facial landmark locations</td>
<td></td>
<td>82</td>
</tr>
<tr>
<td>5.10</td>
<td>Feature points on a contour line</td>
<td>[62]</td>
<td>83</td>
</tr>
<tr>
<td>5.11</td>
<td>Feature point marking by Cohn et al. [68]</td>
<td></td>
<td>84</td>
</tr>
<tr>
<td>5.12</td>
<td>The contour generated from the eleven fiducial points for muscle contraction estimation</td>
<td>Terzopoulos and Waters [278]</td>
<td>84</td>
</tr>
<tr>
<td>5.13</td>
<td>Thirteen landmark points</td>
<td>Pighin [240]</td>
<td>85</td>
</tr>
<tr>
<td>5.14</td>
<td>The extracted canonical points</td>
<td>[96]</td>
<td>85</td>
</tr>
<tr>
<td>5.15</td>
<td>Template-based approach of 5 regions on a face</td>
<td>[35]</td>
<td>86</td>
</tr>
<tr>
<td>5.16</td>
<td>Feature points for front and profile views</td>
<td>[228]</td>
<td>86</td>
</tr>
<tr>
<td>5.17</td>
<td>Feature points</td>
<td>[318]</td>
<td>87</td>
</tr>
<tr>
<td>5.18</td>
<td>(a) Feature points on a 3D face model and (b) 2D face image</td>
<td>[268]</td>
<td>87</td>
</tr>
<tr>
<td>5.19</td>
<td>(a) Manually placed features on a subject; (b) 3D polyhedral surface; (c) 3D mask fitted on the face markers (d) 3D face surface</td>
<td>[163]</td>
<td>88</td>
</tr>
<tr>
<td>5.20</td>
<td>Markers put on the face and the result of the alignment on specific features</td>
<td>[247]</td>
<td>88</td>
</tr>
<tr>
<td>5.21</td>
<td>The landmarks overlaid on to the face and the face muscles</td>
<td>[246]</td>
<td>91</td>
</tr>
<tr>
<td>5.22</td>
<td>The spread of the thirty-three landmarks in a 3-dimensional space</td>
<td></td>
<td>93</td>
</tr>
<tr>
<td>5.23</td>
<td>The spread of the thirty-three landmarks in a 2-dimensional space</td>
<td></td>
<td>94</td>
</tr>
</tbody>
</table>
5.24 The neutral mean and expression mean of the thirty-three landmarks viewed in
a 2-dimensional space .......................................................... 94
5.25 Landmarks variation to the mean data ........................................ 95
5.26 Facial expression variances ..................................................... 95
5.27 The displacement of expressions from neutral expressions ............... 96
5.28 The displacement of landmarks sorted in increasing order ............... 96
5.29 The spread of landmarks on all the neutral faces (2 per subject) in the x-y plane 97
5.30 The spread of mean landmarks (averaged for each subject) in the neutral faces
in the x-y plane .................................................................. 97
5.31 The spread of mean landmarks of smiling and frowning faces against the overall
expression mean in the x-y plane ................................................. 98
5.32 The spread of landmarks in smiling and frowning against the expression mean
and the neutral mean in the x-y plane .......................................... 98
5.33 Landmarks with the least spread, number 3, 5 and 19. The landmarks represent
the inside corner of the eyes and the tip of the nose (landmark 19) .......... 99
6.1 Visualisation of the order of tensor. A vector is a $1^{st}$ order tensor, a 2D matrix a
$2^{nd}$ order tensor and a 3D matrix is a $3^{rd}$ order tensor etc. ..................... 108
6.2 (a) illustrates a 3-mode tensor $\mathcal{A}$ with 3 different elements (axis) which give 3
mode spaces as shown in (b), (c) and (d) (adapted from Lathauwer et al. [184]) 108
6.3 Flattening of a $(I_1 \times I_2 \times I_3)$-tensor in 3 ways to obtain matrices comprising
$(I_1 I_2 I_3)$ matrix mode-1 vectors $A_1$, $(I_2 I_3 I_1)$ matrix mode-2 vectors $A_2$ and $(I_3 I_1 I_2)$
matrix mode-3 vectors $A_3$ ......................................................... 110
6.4 Visualisation of a three-mode SVD array ........................................ 114
6.5 3D visualisation of tensor matrix $\mathcal{A}$ ............................................ 116
6.6 An example of a data tensor model (image taken from [296]) ............... 120
6.7 Visualisation of the decomposition of tensor $\mathcal{A}$ into three subspace matrices and
a core tensor $\mathcal{D}$ ..................................................................... 122
6.8 Samples of pre-processed face surfaces ........................................ 125
6.9 Each emotional expression is managed into slices. The concatenation of the
slices forms the sub-tensor model with a dimension of $6 \times (54 \times 5090 \times 3)$,
where 3 is the x,y,z- coordinates ........................................... 127
6.10 The eigenfaces taken from three rows of -3.$\lambda$ to +3.$\lambda$ eigenmodes .......... 127
6.11 A partial visualisation of sub-tensor for a disgust expression (from -3.$\lambda$ to +3.$\lambda$
eigenmodes), which makes the eigenvectors person-specific. ................. 128
6.12 The rank 1 recognition rate of recognising neutral faces in a pool of all expres-
sioned faces ................................................................. 128
6.13 The recognition results gathered from using the sub-tensor SVD and the PCA
under varying facial expressions .............................................. 130
6.14 Energy captured by retaining the number of eigenvectors with the largest eigen-
values of 54 subjects .......................................................... 131
6.15 Energy captured by retaining the number of eigenvectors with the largest eigen-
values of 54 subjects with an expression. Interestingly, the first eigenvector
captures more than 90% of the shape variations. .............................. 131
6.16 Energy captured by retaining the number of eigenvectors with the largest eigen-
values of 54 subjects with 2 facial expressions. .............................. 132
6.17 The recognition results gathered from using the sub-tensor SVD and the PCA
under varying facial expressions without intensity level 3 expressions ........ 132
6.18 ROC graph for (a) angry expressions, (b) disgust expressions, (c) fear expres-
sions, (d) happy expressions, (e) sad expressions, (f) surprise expressions. The
numbers shown on the graphs are the normalised thresholds. The probability
of correctly recognising a subject and falsely accepting a subject is measured
based on several leave-one-out experiments. The correct operating threshold
depends on the application. .................................................. 133
6.19 Results of recognition using PCA ........................................ 134
6.20 Results of recognition using PCA ........................................ 135
6.21 Comparative recognition results between PCA and sub-tensor SVD .......... 135
6.22 Variance distribution of recognising faces .................................. 137
6.23 Variance distribution of recognising facial expressions ............... 137
6.24 Some samples of 3D face database ........................................ 138
6.25 Landmarking on 3D face surface ........................................... 139

7.1 The geometric overview of the SDA method between two expression classes. 150
7.2 The geometric idea of the tensor model for multiple facial expression classes 152
7.3 The reconstruction of the largest PCA modes on the VisionRT face datasets . 154
7.4 Reconstruction using the principal modes on NotreDame face datasets .... 155
7.5 Reconstruction using the principal modes of Binghamton University face dataset 157
7.6 Reconstruction using the SDA most discriminant direction for facial expressions. Left to right (top) $-3sd_1$ to $+3sd_1$ and (bottom) $-3sd_2$ to $+3sd_2$, where $sd_1$ is the smiling axis and $sd_2$ is the frowning axis ................. 158
7.7 Reconstructions from the Binghamton University faces using the most discriminant direction between neutral and angry expressions. Left to right $-3sd_1$ to $+3sd_1$ of the angry expression. ................................. 159
7.8 Reconstruction along the most discriminant direction between surprise and angry expressions. .................................................. 159
7.9 Reconstruction using the SDA most discriminant direction when using the Notre Dame data along the smile(left)/frown(right) axes. ................. 160
7.10 The reconstruction of angry faces over the four different level of intensities. 161
7.11 Illustration of SDA expression synthesis on 3D subjects reconstructed from the FERET database ................................................. 162
7.12 Illustration of (a) original faces with expressions (b) exaggerated and neutralised examples of those faces. ..................................... 163
7.13 The near neutral face reconstruction from a smiling expression of a subject. 165
7.14 The overall mean of the neutral and the mean of smiling expressions captured from the Imperial College face database. ............................ 165
7.15 The face recognition rate over the different numbers of eigenmodes for five expressions. ....................................................... 166
8.1 The samples of the captured RAW faces of Binghamton 3D face database . . . 172
List of Tables

2.1 Description of the eleven facial muscles that contribute to facial expressions . . 21
2.2 Basic muscle groups involved in facial animation [181] . . . . . . . . . . . . . 23
2.3 The description of Actions units . . . . . . . . . . . . . . . . . . . . . . . . . 24

3.1 Overview of 2D face recognition techniques . . . . . . . . . . . . . . . . . . . 40
3.2 Overview of 3D face recognition techniques . . . . . . . . . . . . . . . . . . . 52

4.1 The 13 anatomical landmark points . . . . . . . . . . . . . . . . . . . . . . . . 65

5.1 The measured input error on the 33 landmarks using standard deviation (s.d.) . 92

6.1 Recognition rate results of PCA and sub-tensor HOSVD . . . . . . . . . . . . 136
6.2 Recognition rate results of Group A . . . . . . . . . . . . . . . . . . . . . . . 140
6.3 Recognition rate results of Group B . . . . . . . . . . . . . . . . . . . . . . . 140
6.4 Recognition rate results by differing number of subjects . . . . . . . . . . . . 140

7.1 The RMS along the most discriminant component for angry expressions of Level 1. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 161
7.2 The RMS of Type 1 and Type 2 (measured in mm) . . . . . . . . . . . . . . . 164
Chapter 1

Introduction

Recognising faces is an innate ability in all humans. A baby learns quickly to recognise his or her mother, and in a high school reunion dinner, despite the changes due to ageing over the years, the use of beauty products and a variety of facial expressions, we can still recognise friends and familiar faces. Our brain has the capability to capture and recall a person we have seen often. The task of recognising faces is done quickly and effortlessly by the brain. However, the brain mechanisms that are responsible for the recognition process are so complex that despite the work of neurobiologists, mathematicians and computer scientists, it is still not possible to replicate them accurately.

Over the last three decades, the face research community has matured with the development of numerous tools and algorithms for face recognition applications. Despite the availability of such applications, there is not yet a face recognition tool that can be used for all situations (or in a general environment). While each of the existing tools can achieve a high recognition rate, they only work in constrained environments with limited or specific light intensity distribution, controlled location and head pose, and specific facial expressions [157]. Face image variations are a significant problem in real-life face recognition applications [119], and there is a need to take them into account.

Face recognition researchers have developed increasingly complex and more accurate 2D face recognition systems, however, 2D systems are intrinsically susceptible to errors caused by face variations. These errors result from the lack of data defining the shape of a face. With the advancements of 3D sensor and camera technologies, 3D face recognition systems were
introduced with the aim of using the additional geometric data to eliminate some of the intrinsic problems associated with 2D face recognition [13, 322, 43]. It is evident that the 3D geometry of a face is invariant to changes in lighting conditions. As 3D models can be registered to a base pose, the system can be made head pose-invariant. Capturing texture information remains possible with the 3D data, and this can also be useful for recognition purposes.

Despite the advantages of using 3D models, facial expression variation is still a problem. Facial expression is a dynamic variation on a face controlled by facial muscles (as in Figure 1.1), from which it is possible to identify the cognitive activity, emotional states, personality and intention of a person. Expression plays a communicative role in interpersonal relations. Besides facial expressions, other gestures of non-verbal communication are cues in face-to-face interaction. The gestures are bodily movements intended to express meaning such as waving goodbye, pointing, drooping the head and others. These non-verbal cues are useful in helping listeners to elicit the intended meaning of spoken words. Research conducted by Mehrabian [208] revealed that although humans have verbal language, messages shown on the face provide extra information supplementing verbal communication. The author stated that 55% of effective face-to-face human communication depends on facial expressions, while only 45% relies on languages and non-verbal body gestures.

Figure 1.1: The facial muscles
Interestingly, humans can recognise the different facial expressions of an unfamiliar person and recognise a familiar person regardless of the person’s facial expressions. In the interaction between human and machines, a duality exists making it possible to automatically recognise faces and facial expressions in natural human-machine interfaces. These interfaces may be useful in behavioural science, robotic and medical applications. For example, in a case that involves negotiation, it would be useful to have a system that is able to recognise the emotional state of a person. Likewise, robotics for clinical practice could also benefit from the ability to recognise expressions.

In the literature on automatic face recognition, many researchers often use similar techniques for both face recognition and facial expression recognition. For example, in facial expression recognition, expression information is the centre element in recognition even though faces carry other information such as the identity of a person. On the other hand, in face recognition the variability that arises from facial expression is unwanted and the uniqueness of a face is the central recognition criterion. However, removing facial expression variability means working in a constrained environment and recognising faces invariant of facial expressions will always be a problem. Besides facial expressions, there are other factors that are not yet resolved, for example the differences arising from age, illness, growth or shaving of beards or facial hair, make-up and facial deformation due to speech.

This thesis investigates the possibility of accounting for different forms of variability in recognising faces. This can be done by creating a subspace for each face variation and establishing a relationship among the subspaces. Here, we are looking at using a more advanced algebraic and statistical framework to analyse and represent 3D face models in a parsimonious fashion. We employ multilinear algebra, using higher order tensors. This tensor framework possesses a remarkable ability to deal with the shortcomings of principal component analysis in less constrained situations. A tensor is defined as a multilinear representation of a set data. We use them to perform $N$-mode orthogonal decomposition. The $N$-modes represent different variations of collections of face models with multiple formation factors. The factors may include changes in pose, illumination and facial expression. They can be represented across the various modes without destroying the detail of each other. By employing tensor decomposition, results
show better recognition rates than by employing the popular eigenfaces technique.

### 1.1 Face recognition in brief

The extensive literature on face recognition and facial expression recognition is discussed in Zhao et al. [322], and Fasel and Luettin [102]. Most face recognition deals with recognising a person under minimal expression change and only works with well framed images with neutral expression. There are many successful methods that have been developed but none of these work simultaneously for both face and facial expression recognition.

Chang et al. [57] applied existing face recognition techniques to recognise a person in a database containing a set of different faces and facial expressions, which were taken over a time lapse. The result showed that the recognition rate without any expression was around 90%. However, when facial expressions were included the recognition rate dropped in between 25% to 50%. Recent studies done by Givens et al. [119] and Chang et al. [57] reported that facial expression change is one of the most important factors affecting recognition.

Facial expressions have been widely studied and analysed. There are six common expressions related to the emotional states of happiness, sadness, anger, disgust, fear and surprise [157, 84]. However, the variation of expressions covers more than the emotion category. Verbal and non-verbal communication and the physiological activities (e.g. pain and tiredness) also create facial expressions. The variation of face features can be classified based on Action Units [87]. An Action Unit (AU) is an element of visible facial movement and an expression is the result of a combination of several AUs. AUs are described in a framework known as the Facial Action Coding System (FACS) [87].

The FACS approach was designed for human observers to classify the different facial expressions generated by facial muscle motions. FACS has been deployed directly or indirectly in some facial expression research, however, no face recognition techniques have yet employed FACS. This could be due to the complex computations when utilising AUs. In order to study the change of expressions based on AUs, landmarks can be used. These landmarks can be placed on facial features.

On the human face, there are permanent and salient features. These features are formed by
1.1 Face recognition in brief

the shape and placement of the bones of the skull, the cartilage and the soft tissue. The per-
manent features of a face are referred to as *physiognomy*. Facial features that might uniquely 
identify a person include the forehead, eyebrows, nose, cheeks and mouth. The science of mea-
suring the salient facial features on the face is known as the craniofacial anthropometry [174].

Principal Component Analysis (PCA) can be used to model variations [288]. PCA is a 
popular linear algebra technique which is commonly used and addresses the total variation of a 
data set efficiently. In the context of face recognition, there are often two or more independent 
dimensions of change, for example subject and expression, which influence the recognition of a 
face. PCA will not separate these different sources of variation. However, multilinear analysis 
methods have the potential to characterise different forms of variations.

Bilinear models were used by Tenenbaum and Freeman [276]. The concept of multilinear 
algebra was introduced by Tucker [287]. Multilinear models can represent data that is organised 
according to different types of variation and furthermore allows the manipulation of the data 
independently on those different dimensions.

The proposed research will look into representing faces using a multilinear algebra tech-
nique based on higher-order tensors. A tensor is a structure generalising vector spaces. Higher 
Order Singular Value Decomposition (HOSVD) will be used to manipulate the tensor model. 
HOSVD is a general $N$-factor analysis method. Vasilescu and Terzopoulos [294] used HOSVD 
in 2D face recognition. Further work by Macedo et al. [200] followed Vasilescu and Terzopou-
los but used different data. Wang and Ahuja [297] worked on synthesizing facial expressions 
based on the results gathered from HOSVD. The results of all these authors have shown that the 
multilinear approach is promising in handling multiple independent factors simultaneously. In 
this thesis, we extend the method to three dimensional face data.

There are plenty of 2D face data sets available [127] and a commonly used one is the FERET 
database [103]. Some 2D face databases cover a large variety of variations which includes 
different poses, different facial expressions, and time lapse variations. By contrast, 3D face 
databases are few in number and do not cover a large variety of variation factors. One publicly 
available 3D face database is the FRGC [109], from the University of Notre Dame. However, 
this database contains only 3D faces with neutral expression. Recently, many face researchers
1.2 Motivation

As real world applications for face recognition continue to increase, the need for an accurate, easily trained recognition system becomes more important, for example in the development of automatic visual surveillance systems. Current systems for face detection and identification are primarily based on classifying faces in a set scenario often involving neutral expression and frontal face pose. Such scenarios make the process of identifying an individual a lot simpler. However, in a real surveillance environment, subjects do not limit or restrict their facial expressions, and current recognition systems will be confused by data variance in subspaces that make detecting and identifying an individual very difficult. In order to organise and account for the different forms of variation, a more advanced algebraic and statistical framework needs to be used.

Face recognition also can be applied for medical science and other disciplines. Some of the research tries to identify subtle changes in the face associated with growth and ageing. Andresen et al.’s [10] work compares the changes of mandibles of children over time. Hutton [151] used a statistical model to analyse a natural way of expressing growth changes with age. This technique is useful in identifying abnormal growth in faces. Simulating age changes can also be useful in the search for missing children [224].

Research in the area of face recognition has been complemented by interest in facial expres-
sion research. Synthesizing facial expressions to portray emotions, analysis of facial actions and recognising emotions are common topics in facial expression research nowadays. The field of face recognition research is growing quickly and is involving other disciplinary fields from psychology to computational science. Face recognition is not only useful in security and surveillance, but also in applied clinical and facial expression animation.

To develop robust and efficient applications, a face recognition system should be able to deal with a large variety of face variations. As facial expression variation is known as the most difficult problem in face recognition, a novel approach is needed to address this problem.

### 1.3 Challenges in face recognition

Figure 1.2 shows variation over different subjects. Lu [196] claimed that inter-subject variations can be small, while Adini et al. [5] demonstrated that the variations between images of the same face due to illumination and pose changes could be larger than images of different faces. This is illustrated by Figure 1.3.

![Variation in appearance over different subjects](image)

**Figure 1.2: Variation in appearance over different subjects [117]**

In face recognition, regardless of inter- or intra-subjects, facial expression contributes significantly to the variation. This is true for either 2D or 3D face information. Many face recognition systems sidestep facial expression problems [24, 57, 119, 157] by training with only neutral or relaxed faces. If facial expressions are included in the training set for these systems, the recognition performance declines.

Human faces come in different shapes and sizes depending on the age, sex and lifestyle of the subjects. Generally, children have a rounder face in comparison to adults. As children
Figure 1.3: Extreme variation in appearance of the same subject under different lighting conditions and poses [117]

Figure 1.4: Face change over a 2-year interval [229]

grow into young adulthood, the skeletal structures grow along with an increase of muscle tissue and change in volume of fatty tissues. In adulthood, particularly in men, it can be seen that the cartilage and nose continue to grow slowly. In later life, wrinkles become more visible. Figure 1.5 shows how three faces changed over a period of twenty-six years. From the figure, we will also notice that the different ‘choice of lifestyle’ can contribute to the variability of faces. The lifestyle choices include changes in facial hair, receding hair-lines, changeable hair styles, eye glasses and plastic surgery. All these variables increase the difficulty in recognising of human faces.

The above problems in face recognition are well known, and yet computer recognition of human faces is still far from providing solutions to all the global variability.

1.4 Aims

In this thesis, we aim to show that it is possible to develop an expression-invariant face recognition system by making use of facial expression information. In using the word ‘possible’ we
mean that a practical implementation that runs in a reasonable length of time can be achieved. Facial expression information is to be described by the geometric surface mesh information of the face described by 3D coordinates. The term ‘making use’ implies that we will take advantage of the facial expression information available instead of trying to avoid or eliminate it.

1.5 Contributions

In this thesis, we propose and investigate several novel ideas that can be used in 3D face recognition applications. The first is to identify and use a set of landmark points that contain high expression information. Many face and facial expression researchers have made their own arbitrary choice of landmarks. In order to select landmarks, we began with prominent facial features and identified a simple way to manually locate landmarks on less palpable areas. The selected landmarks used points from the FACS framework and craniofacial anthropometry landmarks. The complete set of landmark points was then statistically analysed using facial expressions and those landmark placements that are highly mobile in different facial expressions were identified as the best set to use.

The second method we developed is to organise the face data using a tensor model, according to the variation found in the 3D face databases. We proposed a novel method of expression-

Figure 1.5: Face change in 2-year intervals from 1978 to 2004 [229]
invariant face recognition by introducing a sub-tensor SVD algorithm that subsumes the conventional eigenface approach. In comparison to the conventional approach, the sub-tensor SVD technique can effectively manipulate data independently across the different dimensions. This means that the sub-tensor can find the principal axes of variation across the various modes (subject, poses, facial expressions) and represent how the various factors interact with one another. In contrast, the PCA based eigenface approach represents the principal axes of variation across all the face data. With the sub-tensor SVD method, the intra-class and inter-class variations become more manageable. This algorithm has shown promising results especially in face recognition and recognising similar facial expressions of the subjects.

Due to the limited number of facial expressions in 3D face databases, we developed a third method based on 3D statistical face modelling which can be used for synthesizing expressions. This is based on a linear statistical approach to model the facial expression changes by moving along the most discriminant vector directions. The innovation here is that we employed different dimensions of the tensor model with the aim of finding the changes that characterise the facial expression transformations. We found that our synthesised expressions are visually more realistic than those generated using the Active Shape Modeling (ASM) approach[71]. The main strength of our approach is that it is able to extract individual facial expression information efficiently and create controlled transformations of expression within the 3D face models.

Many papers on expression synthesis report that deforming faces from one expression to another expression is simpler than neutralizing expressions. We investigated the neutralisation of faces using the proposed tensor-based statistical approach. We also compared the results of the neutralised faces with the ASM approach. We showed that the proposed method achieves the synthesis of neutral faces better than using ASM. We also explored face recognition performance on the neutralised expressions. The recognition results yield slightly higher accuracy than the PCA approach.

1.6 The organisation of the thesis

The organisation of this thesis is as follows:
Chapter 2 reviews the concept of how human facial expressions are generated and analysed in animation and recognition applications.

Chapter 3 discusses existing work on face recognition applications including the different algorithms used particularly in the area of 3D face recognition.

Chapter 4 describes the methods used for pre-processing raw 3D face data to remove surface defects, normalise the face surfaces and establish correspondence between surface points. These pre-processing methods were developed by a colleague and were simply applied without change in our work.

Chapter 5 evaluates the best placement of landmarks for distinguishing facial expressions by applying quantitative analysis on facial expression datasets.

Chapter 6 presents the novel sub-tensor SVD approach for expression-invariant face recognition.

Chapter 7 investigates the manipulation of facial expressions for face synthesis and recognition. The expression reconstruction approaches considered are the ASM, tensor-based SDA and statistical deformable models using registration.

Chapter 8 describes the contribution of this work and proposes possible extensions and improvements that could be attempted in the future.
Chapter 2

Facial Expressions

Facial expressions provide important information in social environments. Expression is a primary channel of nonverbal communication [7]. The human face contains not only information about the identity, gender and age of a person but also their emotions and intentions. The ability to discriminate accurately between expressed emotions is an important part of interaction and communication with others. Even though humans have acquired powerful capabilities of language, the role of facial expressions in communication remains substantial.

In Darwin’s book *The Expression of the Emotion in Man and Animals* [78], the similarities and differences of the displayed emotions among human and animals are discussed. He concluded that they both produce similar facial movements for expression. He claimed that certain fundamental facial expressions perform behavioural functions and are innate and therefore the same for all people.

Darwin’s study of facial expressions was motivated by Guillaume Duchenne [42] who was a pioneering neurophysiologist working on perception and communication of human facial affect. He used electrical stimuli on subjects (see Figure 2.1) to determine which facial muscles are responsible for different facial expressions. Figure 2.2 shows photographs of expression generation taken from his book *Mecanisme De La Physionomie Humaine Ou Analyse Electrophysiologique Des Passions*. His work confirmed the idea that a distinct muscular pattern has evolved for each expression.

In the 20th century, many studies relating to emotion-based facial expression and interhuman communication have been carried out especially among psychologists and cognitive sci-
Figure 2.1: Guillaume-Benjamin Duchenne and his assistant electrically stimulate the face of a live subject to display an expression [42]

Figure 2.2: The study of facial muscles (image taken from Duchenne [42])
entists. Ekman [84] reinvestigated Darwin’s work and initially suggested that cultural environments, as well as the biological, can influence the meaning and generation of facial expressions. This contradicted the evidence that facial expressions are innate [89, 155, 6, 179, 34, 110]. However, further evidence to support the hypothesis that facial expressions are innate emerged from studies on neonates. Results have shown that spontaneous facial actions are organised without being taught and this can be seen in the uniformity for expression patterns created by blind infants and children [59].

Overall, the notion that certain fundamental expressions are innate has received robust support from cross-cultural research. Following the studies of a large number of literate and pre-literate societies, Ekman and Friesen [86] reported that people of all races, including isolated tribes, are able to recognize facial expressions based on emotions, in particular anger, disgust, fear, happiness, sadness, and surprise. These emotions have become widely accepted as the six basic emotions, which Ekman et al. [88] claims to be universal across cultures and human ethnicities [84, 190].

Parke [232] applied the analysis of facial expressions in computer science by generating simple looking 3-dimensional face models with artificial expressions. His work was inspired by the findings of the cognitive scientists. Since then, automatic facial expression analysis has been widely researched following the advancements in face detection and recognition applications with the availability of cheap computational power. Suwa et al. [272] investigated automatic facial expression analysis from a sequence of images using a set of selected feature points whose movements were compared and analysed against the original facial expressions.

### 2.1 The Measurement and Coding of Facial Expressions

Tian et al. [283] listed two approaches that can be used to extract facial features for facial expression analysis: geometric feature-based methods and appearance-based methods. The geometric features are based on the shape and locations of facial components (such as the eyes, mouth, nose) while the appearance-based methods extract information by applying filters (such as Gabor wavelets) to images of either the whole face or sections of it. The extracted information from these approaches is represented by feature vectors. The feature vectors contain facial
action information that can be used to analyse emotion-based facial expression and facial muscle actions [227, 281].

Cohn [66] divides facial expression measurement approaches into judgment-based and sign-based methods. Judgment-based methods involve the interpretation of facial expressions in terms of predefined emotions and mental categories. Sign-based methods concern just facial motion and deformation and are independent of any interpretation. Results are simply coded into visual classes. The judgement-based methods infer what underlies a displayed facial expression, for example personality, while sign-based methods describe just the physical facial appearance change.

The most commonly used expression descriptors in judgment based approaches [66] are emotion-based. The different types of descriptors can be found in Ekman and Friesen [89], Plutchik [243] and Russell [254]. Ekman and Friesen [89] proposed six basic emotion-based facial expressions. These emotions however may not always be displayed and there are cases where the displayed facial expression does not reflect a person’s emotion [88]. Ekman’s emotion-based facial expressions are limited to only six expressions and they do not include other possible facial expressions (such as agreement, boredom and sleepiness). Thus, Ekman’s basic emotions may not be sufficient for an automated facial expression analysis system.

Plutchik [243] extended the universal expressions by introducing an emotion wheel. He suggested eight primary emotions: acceptance, anger, participation, disgust, fear, joy, sadness and surprise, and by mixing these primary emotions suggested twenty-four secondary emotions. For example, the combination of ‘joy’ and ‘acceptance’ produces ‘love’. Unfortunately, this theory is based on the inner state of the person, and, like Ekman’s six basic emotions, may not be sufficient to fully describe facial expressions [291].

Many emotion models have been developed by cognitive scientists. Russell [254] proposed a systematic method of classifying all emotions within a two-dimensional space. Emotions ranging from sadness to happiness, and boredom to frantic excitement lie around two bipolar factors called valence and arousal. The emotions are incorporated with intensity and they can be correlated. Similar to prior emotion models, this theory is not guaranteed to describe all facial expressions.
The aforementioned, judgment-based approaches are based on feelings. In contrast, the FACS framework is independent of any feeling or interpretation. The framework associates facial expression changes with the actions of the facial muscles. Since FACS is based on facial muscles, it is suitable to describe all facial expressions. We will use the FACS framework to analyse facial expressions. The next subsection explains the analysis of facial muscles followed by how the FACS and the facial muscles are correlated.

### 2.1.1 Facial Anatomical Analysis

The generation of facial expressions is complex, involving the use of specific sets of facial muscles and interaction between the muscles and the skull bone structure. These components all contribute to the different expressions and other changes of facial appearance caused by mastication, speech, yawning and so on.

Figure 2.3 shows the human skull anatomy. There are two classes of bones in the head: the cranial and facial. The cranial bone is typically located around the brain, on the top and back parts of head. The facial bones form the upper and lower jaw, the cheek bones, part of the nose, orbits, and the top part of the mouth. There are eight cranial bones and fourteen facial bones. Some parts of the skull can be easily identified by palpation (examining or exploring by touching). The mandible or jaw bone is the largest and strongest facial bone. This bone forms the chin and the overall appearance of the mouth area either opened or closed.

In the face, there are twenty-six main muscles. Muscles are made up of fibrous tissue which produces movement by contraction or relaxation, either singly or in combination. Figure 2.4 illustrates the frontal and profile views of the locations of the facial muscles.

The facial muscles can be divided into two main groups based on their actions: subcutaneous and masticators muscles. These muscles are either linear or sphincter muscles. Linear muscles pull in a linear fashion, and sphincter muscles are ringlike structures which contract to close or open a passage. Most of the facial muscles are linear muscles except for the orbicularis oris (an ellipse around the mouth) and orbicularis oculi (circles around the eyes), which are sphincter muscles (as shown in Figure 2.5).

The subcutaneous muscles are located just beneath the skin and they move the elements
Figure 2.3: The skull anatomy (taken from Gray’s Anatomy [126])

Figure 2.4: The anatomy of facial musculature. The * indicates the mastication muscles while others are facial expression muscles (images taken from Dreamstime [81])
2.1 The Measurement and Coding of Facial Expressions

Figure 2.5: The facial sphincter muscles

of the skin. They are distributed throughout the face. They play a major part in creating expressions often causing one or multiple folds in the skin. The muscles that contribute to the production of facial expressions include frontalis, orbicularis oris, orbicularis oculi, buccinator and zygomaticus.

There are two regions of masticator muscles placed on both sides of the face. The first region is connected to the coronoid process which is a thin triangular eminence that varies in shape and size on top of the temporalis of the mandible surrounding the jaw area. They serve for chewing and mastication. This region contains the masseter muscle (see Figure 2.4). The second region is at the temporal fossa (it is located on the right top of the mandible (see Figure 2.3 profile view of the skull), which contains the temporalis muscles.

Figure 2.6 shows the temporalis muscle. The muscles of mastication are the strongest muscles in the body. Their function is to close and open the jaw in chewing and crushing food.

Two mastication muscles, hidden behind the masseter muscle and between the mandible and zygomatic arch, are the pterygoideus externus and pterygoideus internus. Figure 2.7 shows the pterygoidei muscles. The pterygoideus externus assists in the opening of the mouth, but its main action is to draw forward the condyle and articular disk so that the mandible protrudes. The pterygoideus internus also assists the pterygoideus externus action. The pterygoidei also
2.1 The Measurement and Coding of Facial Expressions

Figure 2.6: The temporalis muscle with the zygomatic arch and masseter muscle removed (image taken from Gray’s Anatomy [126])

acts to move the mandible from side-to-side. This action is particularly important during the trituration of food due to chewing, crushing and grinding with the teeth.

Figure 2.7: The pterygoideus externus and pterygoideus internus with the zygomatic arch. A portion of the ramus of the mandible has been removed (image taken from Gray’s Anatomy [126])

During mastication, the temporalis muscle depresses the zone where it is located and elevates slightly the skin on the Zygomatic Arch. The mastication muscles are known to be wide, thick and strong. The contracting temporalis muscle creates a range of expressions, threat, wrath, etc. It is also interesting to note that the muscles in the mouth area are not solely for speaking and generating facial expressions, but are also used for masticating and positioning the jaw.
Only a limited range of movements make up facial expressions. There are about nineteen muscles in the face and not all people possess all of them [77]. Faigin [99] identified eleven muscles that are responsible for facial expressions. These muscles are shown in Figure 2.8.

![Figure 2.8: The eleven main muscles that influence facial expression formation (taken from Erfan [91])](image)

The function of each of the muscle is listed in Table 2.1. Faigin’s summary of the facial musculature does not include any insight into which facial muscles work together to create a facial expression.

### 2.1.2 Facial action coding system (FACS)

The FACS was first proposed by Ekman and Friesen [87] to explain and measure the behaviour of facial muscles in creating facial expressions. The contraction of one or more muscles of the face is known as an Action Units (AU). The AUs were identified by face anatomists and facial behavior psychologists. In total, there are twenty-eight AUs to describe facial expressions (see Figure 2.9).

Action units describe specific small facial movements. Facial expressions are unique and they are described by an AU or a combination of AUs. Each AU contains additional information which scores the intensity of a facial expression on a five-point scale [85]. The scored information does not convey any mental activity information or facial behaviour. This means that the
Table 2.1: Description of the eleven facial muscles that contribute to facial expressions

<table>
<thead>
<tr>
<th>Facial muscles</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Orbicularis oculi</td>
<td>This muscle is attached to the inner orbit and the skin of the cheek. It is used for squinting by contracting the eye.</td>
</tr>
<tr>
<td>2. Levator Palpebrae</td>
<td>This muscle is attached to the upper eyelid. It is used to raise the eyelid for example in surprise.</td>
</tr>
<tr>
<td>3. Levator Labii Superioris</td>
<td>The muscle is divided into three branches. The inner branch originates at the base of the nose, the middle branch starts on the bottom edge of the orbicularis oris, the outer branch is connected to the zygomatic arch. It is used as the sneering muscle.</td>
</tr>
<tr>
<td>4. Zygomatic major</td>
<td>This muscle is attached to the top of zygomatic arch. It is responsible for smiling.</td>
</tr>
<tr>
<td>5. Risorius/platysma</td>
<td>This muscle originates over the rear of the jaw. It is used in conjunction with the risorius to stretch the mouth for example in crying.</td>
</tr>
<tr>
<td>6. Frontalis</td>
<td>This muscle originates near the top of the skull and terminates under the eyebrows. It is responsible for the expression of surprise. It raises the eyebrows and tightens back the scalp.</td>
</tr>
<tr>
<td>7. Orbicularis oris</td>
<td>This muscle lies around the mouth. It is used for tightening or curling of the lips.</td>
</tr>
<tr>
<td>8. Corrugator</td>
<td>This muscle is found on the nasal bridge, attached to the skin between the eyebrows. It is used to lower the inner end of the eyebrows for frowning and to wrinkle the forehead.</td>
</tr>
<tr>
<td>9. Triangularis</td>
<td>This muscle is located along the lower margin of the jaw and is inserted into the corner of the mouth. It generates expressions of sadness.</td>
</tr>
<tr>
<td>10. Depressor labii inferioris</td>
<td>This muscle originates along the bottom of the chin and is inserted into the lower lip. It is used when speaking to pull the bottom of the lip down.</td>
</tr>
<tr>
<td>11. Mentalis</td>
<td>This muscle stems from just below the teeth on the lower jaw and concludes at the ball of the chin. It is used to close or protrude the lips and for pouting.</td>
</tr>
</tbody>
</table>
2.1 The Measurement and Coding of Facial Expressions

Figure 2.9: The Action Units to describe facial expressions as described in FACS

interpretation of facial expressions still depends on human expert [85].

From the FACS rules, the facial areas that contribute to facial expressions are around the eyebrows, eyes, the upper part of the nose and the mouth. As illustrated in Figure 2.9, AU1 to AU7 [in Column 1], AU9, AU41 to AU44 [in Column 2] encode major changes around the top part of the face including the eyes, eyebrows, forehead and the top part of the nose. The eyebrows, for example, can either be pulled downwards or upwards (and these activities are encoded in AU1 - AU4 [in Column 1], AU9 [in Column 2]) which also cause the skin on the forehead to wrinkle horizontally and/or the skin on the nose to wrinkle (encoded in AU9). The eye alone has many activities (such as closing and opening, slitting, squinting, blinking or winking) and AU42 to AU44 (in Column 2) describe these activities and the muscle responsible is the musculus orbicularis oculi.

The mouth and nose areas are described by the remaining AUs. The mouth activities include tightening, pulling, pressing, parting, stretching of lips and mouth, and pushing the jaw forward. These are encoded in AU10 to AU28 [in Columns 3 to 6]. Such activities around the mouth also affect the surrounding areas of the nose and cheek, such as the change of appearance of the nostrils (AU9 in Column 2, AU10 and AU11 in Column 3), cheeks (from AU12 - AU28 in Columns 3 to 6) and chin (for example in AU15 [Column 4] and AU25 to AU27 [Column 6]).
From action units analysis, it can be seen that the face areas that change in facial expressions are the eyes and forehead areas, and the mouth and cheek areas.

As AUs are based on muscle activations, not all combinations produce facial expressions [181]. According to Lander [181], specific AUs and combinations of AUs generate facial expressions which can easily be understood especially by those who do not have any knowledge of muscle movements. Table 2.2 summarizes a facial muscle chart involving facial expressions which are commonly used within the computer graphics community in face modeling and animation. Table 2.3 shows the AU rules that represent emotions based facial expressions.

### Table 2.2: Basic muscle groups involved in facial animation [181]

<table>
<thead>
<tr>
<th>Action</th>
<th>Muscle Name</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raise Inside Brow L/R</td>
<td>Frontalis Medial portion</td>
<td>Many expressions</td>
</tr>
<tr>
<td>Raise Outside Brow L/R</td>
<td>Frontalis Lateral portion</td>
<td>Many expressions</td>
</tr>
<tr>
<td>Tighten Inside Brow Frown</td>
<td>Corrugator Supercilii + Procerus</td>
<td>Anger, Pain, Disgust</td>
</tr>
<tr>
<td>Eyes Wide L/R</td>
<td>Levator Palpebrae Superioris</td>
<td>Surprise, Fear, Shock</td>
</tr>
<tr>
<td>Eyes Squint L/R</td>
<td>Orbicularis Oculi orbital portion</td>
<td>Anger, Thought, Concentration</td>
</tr>
<tr>
<td>Eyelid Close L/R</td>
<td>Orbicularis Oculi palpebral portion</td>
<td>Blink, Wink</td>
</tr>
<tr>
<td>Nostril Flare L/R</td>
<td>Dilator Naris + Levator Labii Superioris Alaeque Nasi</td>
<td>Disgust</td>
</tr>
<tr>
<td>Purse Lips</td>
<td>Incisivus Labii</td>
<td>Kiss, Anger, &quot;oo&quot;, Whistle</td>
</tr>
<tr>
<td>Smile Corner L/R</td>
<td>Zygomaticus Major</td>
<td>Smile</td>
</tr>
<tr>
<td>Corner mouth down into sadness L/R</td>
<td>Depressor Anguli Oris + Zygomaticus Minor + Depressor Anguli Oris + Mentalis</td>
<td>Sadness</td>
</tr>
<tr>
<td>Top Lip Up L/R</td>
<td>Levator Labii Superioris</td>
<td>Disgust, Part lips for sounds</td>
</tr>
<tr>
<td>Lower Lip Down L/R</td>
<td>Depressor Labii Inferioris</td>
<td>Part lips for sounds</td>
</tr>
<tr>
<td>Tighten Lips U/L</td>
<td>Orbicularis Oris</td>
<td>&quot;p&quot;, &quot;b&quot;, &quot;m&quot;, anger</td>
</tr>
<tr>
<td>Jaw Open</td>
<td>Digastricus</td>
<td>Speaking, surprise</td>
</tr>
</tbody>
</table>

The variety of facial expressions results from the magnitudes of muscle actions. Relying solely on individual AU, it is not possible to describe and interpret all the facial expressions accurately [118]. For example, a smile comes with a variety of intensities. When smiling, the facial muscles involved are the zygomaticus major and minor, the orbicularis oculi, the levator labii superioris, the levator anguli oris and the risorius. Depending on the intensity
2.1 The Measurement and Coding of Facial Expressions

Table 2.3: The description of Actions units

<table>
<thead>
<tr>
<th>Emotions</th>
<th>Action Units (AUs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>6, 12, 25, 26</td>
</tr>
<tr>
<td>Sadness</td>
<td>1, 4, 6, 7, 11, 15, 17, 25, 26</td>
</tr>
<tr>
<td>Surprise</td>
<td>1, 2, 5, 26, 27</td>
</tr>
<tr>
<td>Anger</td>
<td>2, 4, 5, 7, 10, 17, 22, 23, 24, 25, 26, 27</td>
</tr>
<tr>
<td>Disgust</td>
<td>9, 10, 16, 17, 25, 26</td>
</tr>
<tr>
<td>Fear</td>
<td>1, 2, 4, 5, 20, 25, 26, 27</td>
</tr>
</tbody>
</table>

scores, the coded AU combinations can be (AU6 + AU12), (AU6 + AU12 + AU25), (AU12 + AU25/AU26).

The FACS framework is the most popular facial action coding system for analysing facial activities. However, there are a few issues that make it difficult to apply, such as having to combine the different AUs. This can be computationally complex and many of the resulting AU combinations are redundant. According to Scherer and Ekman [260], there are more than 7000 different AUs combinations that have been derived. In addition, the effects of combining some of the AUs may produces similar appearances. For example as shown in Figure 2.10 the combination of AU1 and AU2 produces a similar appearance to AU2 alone.

![Figure 2.10: Action Unit 2 and a combination of Action Unit 1+2](image)

Multiple occurrences of AUs may also interfere with each other [118], which leads to a different facial appearance. Besides that, not all appearance changes may be apparent on the face or in any given images (especially in low resolution images). Hence, effective study of facial expressions requires the use of higher dimension face models, particularly those that have
geometric detail such depth information. Given such information, AUs can be used to advantage in describing all visible facial movements.

2.2 Facial Expression Analysis

Facial expression analysis can be based on the positions and movements of salient facial feature points or on muscle movement information [95, 312]. Optical flow is commonly used to capture the muscle movement. Mase and Pentland [204] analysed expressions based on the estimated 2D muscle movement. Using facial expression analysis, automatic facial expression recognition can be developed. Essa and Pentland [95] used muscle motion models to analyse the facial muscle movements for the purpose of recognising expressions.

There are various uses of the term ‘facial expression analysis’. We use it to mean automatic detection of facial feature changes using visual information. Facial expression analysis is sometimes used to mean human emotion analysis. Emotion analysis requires a higher level of knowledge which incorporates context, cultural factors, personality, gender and so on [53, 255, 256]. Our work simply concerns the geometric changes that occur in expressions, and does not involve any of this higher knowledge.

Conventional methods for characterising facial movement and deformation have relied on hand-crafted rules and emotional state representations [204, 313, 35]. They examine the geometric relationships among facial features [278, 272]. Many automatic facial expression analysis systems directly interpret the basic emotions but only a small number interpret observed facial expressions [225, 147, 199, 96, 323, 169, 319, 149, 299]. Black and Yacoob [35], and Pantic and Rothkrantz [227] used facial expression dictionaries of facial actions to identify emotions. Hoch [146] resorted to FACS to describe facial expressions. More recent approaches have taken advantage of the FACS framework and facial muscles information [96].

2.2.1 Facial expression recognition

There are a number of literature reviews that can be found on automatic recognition of facial expression [258, 102, 227, 272]. Theoretically, recognising facial expressions based on ex-
tracted features is the last step in facial expression analysis [283]. Classifiers such as neural networks [281, 282], support vector machines (SVM) [20, 107], statistical pattern recognition [68] or Bayesian networks [65], are commonly applied in recognising facial expressions. A number of expression recognition systems use rule-based classification from facial actions [227, 67].

In Rowley et al. [251], sixteen AUs from the FACS framework were recognised by using a three-layered neural network obtaining a 95.5% of recognition rate. Similar work was undertaken by Cohn et al. [68], Donato et al. [80] and Lien et al. [193], who also recognised AU combinations. Ford [107] employed SVMs to classify the six primary emotions and neutral expressions. Recognition has been done by using MLR (multinomial logistic ridge regression [39]) which obtained 91.5% accuracy in recognition. The use of neural network classifiers to recognise emotion-specific expressions was also investigated by Tian and Bolle [280], Kobayashi et al. [172] and Rosenblum et al. [249].

Cohen et al. [65] employed Bayesian networks and Hidden Markov Models (HMM) to a sequence of images to recognise the six basic emotion-based expressions. The overall recognition rates were 73.2% on the Cohn-Kanade database [67] and 66.5% on the UIUC-Chen database [61]. The use of HMMs to recognise AUs with SVM and PCA classifiers was also done by Bartlett et al. [20]. They achieved 98.1% recognition rate for blinking and non-blinking but the recognition rate for brow-up, brow-down and no brow motion was 70.1%.

Using a rule-based classification system, Cohn et al. [67] were able to recognise three AUs of the eyes with an accuracy of 98%. However, the accuracy across the categories in the brow regions (brow-up, brow-down and no brow motion) was only 57%. From their analysis, the brow action motion is not a reliable point of estimation because the motion is too small to detect. This statement was also confirmed by the human FACS coders, who found that the detection of brow movements has difficulties especially when the face is occluded with eyeglasses or the head is tilted [283].

It is important to note that the figures reported in much of the above work is obtained from different data sets and different experimental conditions and are not in general comparable. Since the recognition is limited to facial action units and emotion-specific expressions, it is not adequate for real-life settings. In a direct task of recognising AUs, facial expression recognition
systems such as [281, 313] can recognise a number of AUs individually and in combinations. Presently, the work on spontaneous facial expression analysis is emerging and potentially will have significant impact across a range of theoretical and applied facial expressions recognition topics.

2.2.2 Facial expression modelling and animation

Facial animation is complex and difficult to do realistically. The facial features that contribute the most to facial expressions are the eyelids, eyebrows and the mouth. Wrinkles and bulges also contribute to the change of facial appearance. Movements or the flow of features can be measured and then used to animate facial expressions. This approach is known as feature-based deformation. Platt and Badler [242] used landmark information to deform face models. Following on from this was a work on facial animation using pseudo muscles by Waters [302]. Work by Guenter at al. [131] and Pighin et al. [239] used facial movement information. According to Fasel and Luettin [102], the deformation approach does not necessarily require extensive facial movement which makes the animation process faster and simpler. However, this approach is unreliable in creating more exaggerated facial expressions.

A large number of face modelling and facial animation works have employed the muscle-based approach [277, 188]. Synthetic facial movements are generated by mimicking the contraction of facial muscles. This can be done by firstly defining the functionality and locations of the facial muscles on the face model and then applying a combination of muscle contractions [302]. The combinations of these muscles are defined by AUs. Using AUs reduce the amount of work in characterising facial expression data.

Many face animators, for example in Fox [108], imitate facial muscle movements to generate facial expressions. An illustration of the different facial appearances generated by individual facial muscles can be seen in Figure 2.11. Similar to the prior muscle-based animation problems, this approach creates only a limited set of facial expressions.

Multi-layer models supplement the use of facial muscles. A multi-layer model is built from the anatomical structure of the face, facial muscles, soft tissue, skin etc. [188, 277, 265, 304]. Terzopoulos and Waters [277] proposed facial animation by contracting synthetic facial mus-
Figure 2.11: Different synthesized facial expressions (taken from the Barrett Fox Character Animator [108])
2.2 Facial Expression Analysis

cles embedded in a face skin model. This approach improves the realism of the synthetic facial expressions, however the use of the sophisticated biomechanical models requires accurate simulation methods and high computational costs [188]. Furthermore, animating expressions in complex multi-layer structures requires nonlinear methods to simulate dynamic deformation of the skin. Failing to create a detailed skin deformation (such as wrinkles) may result in less realistic expressions [92].

The geometry warping approach is another method for synthesizing facial expressions. The facial expression information is measured from two images - one with a neutral expression and another with a particular expression. The calculated facial movement difference vectors are transferred to a target image of a neutral face [211, 239, 263, 233, 305]. The facial movement differences can be controlled by using linear interpolation. The disadvantage of this approach is that only facial expressions in-between existing examples can be created. Also, linear interpolation is not a perfect solution for generating in-between expressions [92]. Recent work has been undertaken to overcome this weaknesses by using nonlinear interpolation or by combining linear and nonlinear interpolation.

Parke and Waters [233] used simple geometric interpolation to synthesize expressions on 3D face models. The feature points are manually digitised on each face model. This was followed by automatic expression synthesis where the data of real actors were captured and analysed [305, 26, 93]. The captured face surfaces are represented using a structured mesh, along with texture information. Work on synthesizing facial expressions has been investigated using a number of different 3D imaging modalities including range scanning [188, 278, 162], CT scanning [173], laser scanning [188, 302], and 3D stereo camera scanning [211].

Apart from using 3D scanners, 3D data can be obtained from video data with multiple viewpoints, where the shape and features of the face can be calculated from a set of images. The 3D video-based method has also been applied to study and analyse expressions. The expression information is then used to generate a set of expression data for a person [239, 241].

Segmenting face models into smaller regions is also employed with the aim of synthesizing only the relevant parts of the face contributing to a facial expression. Joshi et al. [161] applied this approach on 3D face models. Blanz et al. [36] employed a morphable technique to animate
facial expressions on existing images and videos. The advantage of the morphable modelling approach is that it can work on any face without acquiring examples of expression data of the person.

Real-time facial expression tracking captures motion data that can be used to generate facial animation [93, 56, 131, 177]. Accurate tracking of feature points or edges is important to maintain a consistent and accurate expression generation. Methods such as snakes, or deformable minimum-energy curves, are used to track the underlying muscles mechanisms in video sequences [201, 277, 278]. Markers are placed or painted in the face. These markers are extensively used to aid tracking facial expressions from video sequences [194, 219].

The optical flow method can also be used to perform feature tracking in video sequences. The advantage of this approach is that it obviates the need for markings on the face which can be intrusive and impractical. Essa et al. [94] utilised optical flow on a set of measured scores against a set of previously trained templates. Eisert and Girod [83] animated expressions using optical flow data over the whole face.

Vlasic at al. [296] mapped facial movements from a recorded video to a target face using an optical flow-based tracker to estimate 3D shape movements. In addition, they used a multilinear model to manage the face attributes separately. Theoretically, by using multilinear models on a larger collection of faces with different expressions, faces with any expressions can be generated. However, the collection used in Vlasic’s work is limited in size. Nevertheless, the advantage of this technique is that visible facial markers or special face-scanning equipment is not required.

The facial animation approaches described above have strengths and weaknesses. In order to simulate a realistic facial expression, a large collection of facial expression examples is required. When using muscle information, accurate muscle descriptions or templates are needed for producing visually correct facial movements. Using fewer facial expression resources means that expressions appear artificial. Additionally, acquiring a large collection of face data requires a massive amount of time, proper planning and using the right equipment to capture face models of high quality. Methods to optimise the animation computation may also be needed to allow real-time facial animations.
2.2.3 Issues with facial expression analysis

Despite the emerging advances and successes in automatic facial expression analysis, there are still issues that remain open for further research. The use of automatic facial expression recognition systems is very restricted due to the limited robustness and hard constraints imposed on the recording conditions [226]. These include centering the face and minimizing head pose changes. Systems are not robust to analyzing facial expressions irrespective of their sources, motion artefacts and facial feature deformation with varying expression intensities.

Recently, multiple cameras have been used to increase the resolution and improve the accuracy [133]. Furthermore, it is a known fact that facial expressions are produced as a result of 3D geometric shape variations occurring due to specific movements of the non-rigid parts of the face. These shape variations are often such that they can not be captured accurately using standard still or video cameras, and hence cannot be accurately modelled using 2D information because of their inherent 3D nature. The use of 3D models is therefore appealing because of their ability to represent the problem and their independence from viewpoints and pose [198]. Also, with the advancement of data capturing hardware, 3D face models can be created via 3D cameras or scanners to produce quality face models in which fine details are visible such as wrinkles and bulges [92].

A few approaches in the literature have tried to deal with some of the technical issues by using active appearance models [69], local parametric models [211, 302], 3D motion models and feature point tracking approaches based on facial actions or facial deformations [240, 301, 305, 56]. More work needs to be carried out to enhance the quality and accuracy of facial expression analysis for the purpose of recognition and animation.

Another issue concerns human perception and understanding how humans recognise facial expressions [59, 78, 291]. Generated facial expressions may differ depending on the context. Understanding how humans recognise expression can improve the techniques used for automatic facial expression analysis.
2.3 Conclusions

This chapter presented an overview of facial expression analysis and its use in face and facial expression recognition and animation. The existing reviews [92, 102, 118] in this area cover two perspectives: facial modelling and facial animation. We are more interested in the geometry of facial expressions and their synthesis.

Facial animation applications are used extensively in various areas including the movie industry, computer games, medicine and telecommunication. However, the current state of the art systems in animation have not resolved fully the research issues mainly caused by technological limitations.
Chapter 3

Reviews of Existing Face Recognition Systems

Humans can easily detect and recognise a single face image of a familiar individual. However, when it comes to recognising a face from a pool of face databases, humans find it difficult because of the overwhelming data volume. This is especially true when recognising a face that is not particularly familiar. This may be because humans have limited capacity to remember many faces. Humans tend to recognise faces by their particular characteristics such as attractiveness, gender or distinct visual details.

Over the decades, automated face recognition has advanced to the point where face recognition systems are often used in the real world setting. The rapid development of face recognition is due to many factors and each or combinations of these factors contribute to where face recognition systems stand nowadays. Such factors are the availability of large face databases, the development and improvement of classification and evaluation algorithms, and the availability of high speed and high capacity hardware. The development is paralleled by the emergence of face recognition conferences, published papers and successful applications of face recognition within the last 20 years. Protocols for face acquisition, verification and evaluation were also introduced with the aim of developing reliable and robust face recognition systems.

In this chapter, the different techniques used in automatic face recognition for 2D and 3D face models are discussed. The comparison of the different data sets used, the set-up of the
experimental conditions and the experimental results are presented to ensure readers can understand and reproduce the results.

Generic face recognition can be split into two problems: face detection and normalisation, and face identification. This is illustrated in Figure 3.1. Faces are first detected and segmented from cluttered images. The face is then normalised. The extracted facial features are analysed for identification purposes. The identification process is done by comparing and matching the acquired facial feature vector with the feature vectors stored in the database. A matched pattern identifies the subject.

![Figure 3.1: A face recognition pipeline](image)

Over the years, the development of face recognition systems have tended to use rather small sets of face images. Nevertheless, their findings have contributed to many improvements in the recognition algorithms. Developing face recognition systems can involve the use of static images, dynamic videos, range data and three-dimensional face data. To date, a number of face databases are available ranging from still images, dynamic images (video) to 3D face models. However, many of them are not publicly available. Those databases that are available do not necessarily suit all the requirements of face recognition research or applications. For example, there are far more 2D image databases than video images and 3D face geometry databases.
3.1 Face recognition by the human brain

The human brain has a high capability to recognise faces in images with widely varying resolution, illumination, pose and expression. Neurobiologists have found face-specific areas in the brain [303]. The fusiform gyrus is sometime also called the fusiform face area because it is active during face recognition. Its activity can be seen by functional magnetic resonance imaging (fMRI) when a person is trying to remember a face [289]. It is also active when distinguishing other animals or objects [114].

According to Biederman [32], the brain recognises a subject by face matching. The retinal image of a subject travels through the optic nerve to the visual cortex. The face-specific area contains hundreds of millions of neurons, each of which is tuned to detect contrast between light and dark. Distinct features trigger the neurons to fire and this creates a firing pattern that represents the faces or objects. The pattern is then processed by networks of neurons to classify and to recognise the face even if it appears in a different orientation or with different illumination. The face information may be extracted by a number of different mechanisms. There are various hypothesized mechanisms explaining recognition [114] including magnetoencephalographic (M170), the existance of neural regions in the brain sustaining face recognition and interpretations of actions in specific brain areas [141, 115, 90, 311].

Harmon and Julesz [135] studied the limits of human face recognition and have shown that people are able to identify familiar faces even from poor quality images, as low as $16 \times 16$ pixel resolution. An interesting experiment was conducted by Burton and Wilson [48] in which they obscured faces giving only body posture to a group of participants. The results showed dramatically low recognition rates. These experiments demonstrated that faces carry highly significant information for recognition purposes compared to other parts of the body.

Perceiving faces is associated with the ability to extract spatial relationships between different features. With human cognitive abilities, general information such as race or gender may well contribute to better recognition of a person. In summary, the identification process is still an open research area with many hypotheses to be investigated.
3.2 2D face recognition

The 2D face recognition systems can be divided into feature-based, holistic and hybrid recognition methods. The feature-based method is similar to the hypothesized brain mechanism in identifying and recognising distinct features on the face and matching them to stored data. The holistic approach does not extract features but maps the overall face, and hybrid-based recognition methods combine both approaches.

A recent survey on 2D face recognition was written by Zhao et al. [322], where they gave an up-to-date review of existing literature and insights into the studies of machine recognition over 30 years. Early research on face recognition mainly used feature-based methods. The feature information in a feature-based method usually involved either manually or automatically allocated anatomical landmarks placed on facial elements (e.g. the eyes, nose, corners of the mouth, etc.). Kelly [167] extracted the distances between these feature points and used this information for recognition. Kanade [164] developed an automatic facial feature detection system using computed feature parameters against the parameters of known faces. Other work that used feature information either for automatic detection or face recognition was done by Cox et al. [74, 180] and Cootes et al. [70, 120, 191].

Nefian and Hayes [221] and Samaria [259] used strips of pixels across the facial features instead of locating specific facial feature points. Wiskott et al. [306] applied elastic bunch graph matching on the computed feature information. Instead of comparing the distances between points, structure matching methods applied to Gabor wavelets can be used. The face representation is in blocks and the extracted local features are represented by wavelet coefficients, which are used for comparing face sets.

Some feature-based techniques require careful landmarking in order to accurately measure the differences between faces [222]. In addition, the number of landmarks and their locations are important in extracting the geometric distance information. However, research has shown that the use of landmark features and their relationship is insufficient for face recognition [288]. An alternative, holistic technique uses the overall representation of face images to produce a vector space structure. As faces have regular structures, some features may appear redundant. By exploiting this information, a face can be described with fewer parameters.
An image is considered high in dimension and processing faces using all the pixels or points is computationally expensive. By taking advantage of the redundancy, Kirby and Sirovich [266, 170] developed a holistic technique to reduce the dimension of images, which was extended by Turk and Pentland [288]. They used principal component analysis (PCA) to analyse the distribution of the face image in a compact space and compare the similarity of feature vectors in this space. In the original publication of Turk and Pentland, sixteen subjects were digitised under three variations in head orientations, head sizes and lighting conditions. Their face recognition system managed to achieve 85% accuracy with head orientation variation, 64% with head size variation and 96% with lighting variation. Moghaddam and Pentland [217] applied PCA on two variations and a Bayesian approach and their results have shown better recognition rates than using PCA alone [238]. The PCA technique has been applied in many other face recognition systems using different face variations [288, 192, 244]. Most systems use a nearest neighbourhood approach for classification.

Independent component analysis (ICA) has been used in other holistic approaches. It is similar in concept to PCA, however the distribution of the faces need not be Gaussian and higher order statistics can be used. The lack of restriction to a Gaussian distribution promotes statistical independence [152]. Bartlett et al. [20] applied ICA for face recognition, and their results show similar performance to PCA based methods.

Another popular projection method is LDA (Linear Discriminant Analysis). Its goal is to find the optimal projection to represent the face space which maximises the ability to discriminate between face classes. LDA has been used extensively in face recognition applications [24, 273, 322]. To make the LDA classification more effective, PCA is applied to first reduce the dimension of the images, then LDA is employed to obtain the optimal face space. A comparative study has shown that this technique produces lower error rates than using only PCA [24]. The combination of PCA and LDA is known as Fisherfaces.

The above three linear projection methods have advantages and disadvantages. The PCA technique is easy to implement and it provides a general overall representation. However, it is not the best space for class separation. The problem with ICA is that there is no general closed-form solution for the independent components. However, its strength lies in utilising higher-
order statistics. The LDA approach is good for class separation, but requires an assumption that the classes have the same covariance structure. Moreover, LDA may not be computable in small sample size problems, where the dimensionality of the sample class space is bigger than the number of training samples.

The development of linear projection methods for face recognition did not stop with the above three techniques. Vasilescu and Terzopoulos [294] proposed multilinear tensor decomposition of image ensembles with an aim of reducing the confusion of multiple factors. Yang et al. [314] proposed PCA on images that have been generalised in a 2D matrix. Motivated by this, Kong et al. [176] applied it with LDA in face recognition. Support vector machines (SVM) were also used to decompose a multi-dimensional space with multi-hyperplanes placed to achieve the maximum separation distance between classes [236]. The reported recognition rate for SVM was 78% while PCA’s rate on the same data was 54%. The SVM-based algorithm used by Phillips et al. [236] makes use of the radial basis kernel while PCA identifies faces with a $L_2$ nearest neighbour classifier.

The surface of the human face is a complex manifold, and linear separation methods may not be able to deal with its nonlinear nature. For this reason, kernel PCA [261] and kernel LDA [210] methods were introduced to map the nonlinear class boundaries to a linearly separable space. The suitability of a kernel and its optimal parameters are empirically determined [196]. Kernel methods perform better using the data on which they are trained but may worsen in cases where unseen faces are tested. This is due to their complex class boundaries which are prone to over-fitting [229]. To limit the kernel flexibility, non-linear manifolds can be learnt. The learning can be done by applying machine learning techniques such as ISOMAP [275], LLE [250], LDA+geodesic distance [315] etc. However, effective learning of these nonlinear manifolds requires a high number of training face images in all their variations. Unfortunately, existing face data sets are still limited in size and diversity. Therefore, researchers are looking into the possibility of synthesizing faces based on the available samples using a combination of classifier methods. In [295, 320], the authors claimed that this approach may be helpful in enhancing the performance of face recognition systems. However, researchers such as [223, 18] claimed the opposite. They suggested that the nose, eyebrows and eyes are
more important when it comes to recognition. For example, removing eyebrows can disrupt the accuracy of recognition.

The performance of face recognition using 2D images drops drastically when encountering variations in pose, illumination and facial expression [238]. A number of methods have been proposed to handle such variations. One of the earliest work using a hybrid approach was by Baymer [30], in which they coupled 2D face recognition interactively with the estimated poses. Besides pose variation, an illumination-based synthesis method was proposed by Georghiades et al. [116] to handle both pose and illumination problems in 2D face recognition.

The active shape model (ASM) was first proposed by Cootes et al. [71] to estimate and deform face shapes based on shape parameters controlling the eigenmodes gathered from PCA. Using the mean shape and the modes, a set of global shape-free models can be constructed. This method was extended to include texture variation. This extended model is known as an active appearance model (AAM) [70]. Lanitis et al. [182] used both the shape and texture information for automatic face recognition. Faggian et al. [98] used AAM on video images for recognition purposes.

Optical flow has also been used to synthesize poses using correspondence information between an image and a reference face. This approach requires high computational power [31, 257]. Jonnes and Poggio [159] employed a stochastic gradient algorithm to match the synthesized image with the original image at every iteration. This application was later extended to a 3D-based model. Other methods for synthesizing variants for face recognition can be found in Ezzat and Poggio [97] and Maurer and Malsburg [206]. In Table 3.1, a summary of 2D face recognition methods is given.
### Table 3.1: Overview of 2D face recognition techniques

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dataset size</th>
<th>Methods</th>
<th>Reported performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Feature-based approaches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kanade [164]</td>
<td>&gt; 800 faces</td>
<td>Geometry distances</td>
<td>75% correct identification</td>
</tr>
<tr>
<td>Kelly [167]</td>
<td>N/A</td>
<td>Geometry distances and angles</td>
<td>N/A</td>
</tr>
<tr>
<td>Cox et al. [74]</td>
<td>188 images of 47 subjects</td>
<td>Template matching and geometric mixture distances</td>
<td>100% for template matching, 90% for geometric matching</td>
</tr>
<tr>
<td>Nefian &amp; Hayes [221]</td>
<td>400 images of 40 subjects</td>
<td>Hidden Markov Model</td>
<td>98%</td>
</tr>
<tr>
<td>Samaria [259]</td>
<td>N/A</td>
<td>Hidden Markov Model</td>
<td>87% and 95%</td>
</tr>
<tr>
<td>Wiskott et al. [306]</td>
<td>115 faces with 15 degree rotation, 110 faces with 30 degree rotations</td>
<td>Elastic bunch graph &amp; gabor wavelet</td>
<td>86.5% and 66.4% for matching of 15 degree and 30 degree rotations faces</td>
</tr>
<tr>
<td><strong>Holistic-based approaches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kirby &amp; Sirovich [170]</td>
<td>100 images</td>
<td>Eigenfaces</td>
<td>82.95%</td>
</tr>
<tr>
<td>Turk &amp; Pentland [288]</td>
<td>2500 images of 16 subjects</td>
<td>Eigenfaces</td>
<td>96%, 85%, 64% correct classifications averaged over lighting, orientation and size variations</td>
</tr>
<tr>
<td>Moghaddam &amp; Pentland [217]</td>
<td>&gt; 2,000 images</td>
<td>Probabilistic eigenfaces</td>
<td>97%</td>
</tr>
<tr>
<td>Li &amp; Lu [192]</td>
<td>1,079 images of 137 subjects</td>
<td>Point-to-line distance based</td>
<td>55.6% versus 65.4% error rates of proposed method and of the standard method</td>
</tr>
<tr>
<td>Bartlett et al. [21]</td>
<td>425 images</td>
<td>ICA-based feature analysis</td>
<td>89%</td>
</tr>
</tbody>
</table>
## 3.2 2D face recognition

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dataset size</th>
<th>Methods</th>
<th>Reported performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belhumeur et al. [24]</td>
<td>330 images of 5 subjects</td>
<td>Fisherfaces/subspace LDA</td>
<td>99.6%</td>
</tr>
<tr>
<td>Swet &amp; Weng [273]</td>
<td>1.316 images</td>
<td>Fisherfaces/subspace LDA</td>
<td>95%</td>
</tr>
<tr>
<td>Zhao et al. [322]</td>
<td>1,195 images</td>
<td>Fisherfaces/subspace LDA</td>
<td>100%</td>
</tr>
<tr>
<td>Vasilescu &amp; Terzopoulos [293]</td>
<td>161 images of 23 subjects</td>
<td>Higher Order Singular Value Decomposition &amp; SVM</td>
<td>78% for SVM vs 54% for PCA</td>
</tr>
<tr>
<td>Yang et al. [314]</td>
<td>400 images of 40 subjects</td>
<td>2DPCA</td>
<td>96%</td>
</tr>
<tr>
<td>Kong et al. [176]</td>
<td>1,040 images of 50 subjects</td>
<td>LDA</td>
<td>93-98%</td>
</tr>
<tr>
<td>Scholkopf et al. [261]</td>
<td>N/A</td>
<td>Kernel PCA</td>
<td>N/A</td>
</tr>
<tr>
<td>Tenenbaum et al. [275]</td>
<td>698 images</td>
<td>ISOMAP</td>
<td>&lt; 0.2 residual variance</td>
</tr>
<tr>
<td>Yang [315]</td>
<td>400 images of 40 subjects</td>
<td>LDA &amp; geodesic distance</td>
<td>9.7 EER</td>
</tr>
<tr>
<td>Vetter &amp; Poggio [295]</td>
<td>100 images of 50 subjects</td>
<td>Linear transformation</td>
<td>N/A</td>
</tr>
<tr>
<td>Zhao &amp; Chellappa [320]</td>
<td>216 images of 108 subjects</td>
<td>Subspace LDA</td>
<td>93%</td>
</tr>
</tbody>
</table>

### Hybrid-based approaches

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dataset size</th>
<th>Methods</th>
<th>Reported performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beymer [30]</td>
<td>930 images of 62 subjects</td>
<td>Pose-based synthesis</td>
<td>98.7%</td>
</tr>
<tr>
<td>Georghiades et al. [116]</td>
<td>4,050 images of 45 subjects</td>
<td>Illumination-based synthesis (Cones Gen.)</td>
<td>97%</td>
</tr>
<tr>
<td>Cootes et al. [70]</td>
<td>100 images</td>
<td>AAM</td>
<td>6% EER</td>
</tr>
<tr>
<td>Lanitis et al. [182]</td>
<td>300 images of 30 subjects</td>
<td>Flexible appearance (shape and texture) models</td>
<td>92-97%</td>
</tr>
<tr>
<td>Faggian et al. [98]</td>
<td>17 subjects</td>
<td>AAM</td>
<td>92-95%</td>
</tr>
<tr>
<td>Beymer &amp; Poggio [31]</td>
<td>10 views per subject</td>
<td>Example-based technique</td>
<td>85-98%</td>
</tr>
</tbody>
</table>
### 3.3 3D face recognition

Three-dimensional models offer extra information which 2D images do not have (such as depth). Bowyer et al. [43] compared the use of both 2D images and 3D models in face recognition systems. Two dimensional face recognition is severely limited by facial variations [196, 5] but 3D models are more robust [238, 196]. This section examines recognition using 3D face data [160]. Face recognition using 3D face information was first initiated in 1989 by Cartoux et al. [54]. In the 1990s, only a few research results appeared using 3D face models. However, 3D face recognition has become increasingly popular. This is due to the availability of cheap hardware to acquire 3D faces. In the next sections, we present the different 3D sensors and the relevant techniques used in 3D-based face recognition systems. The techniques can be divided into three categories: surface-based, statistical and model-based approaches.

#### 3.3.1 3D acquisition systems

A number of high speed and high precision 3D sensors have been developed recently. In obtaining 3D measurements, at least two cameras are used, corresponding points in the images are established and the 3D coordinates are reconstructed.

A stereo-based system consists of at least three cameras with a light source that projects structured light in a speckle pattern on the human face. A stereo camera pair finds the depth information based on the camera calibration and using the projected speckle to assist point matching. A third camera captures the object texture. A commercial stereo-based system of this

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dataset size</th>
<th>Methods</th>
<th>Reported performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sali &amp; Ullman [257]</td>
<td>N/A</td>
<td>Example-based technique</td>
<td>N/A</td>
</tr>
<tr>
<td>Jones &amp; Poggio [160]</td>
<td>100 faces</td>
<td>Stochastic gradient algorithm</td>
<td>N/A</td>
</tr>
<tr>
<td>Ezzat &amp; Poggio [97]</td>
<td>N/A</td>
<td>Image-based model</td>
<td>N/A</td>
</tr>
<tr>
<td>Maurer &amp; Malsburg [206]</td>
<td>50-80 pairs of images</td>
<td>Single-view based on rotation in depth</td>
<td>53%</td>
</tr>
</tbody>
</table>
kind is the VisionRT (VRT3D) which has three video cameras and a projector (see Figure 3.2). The system is fast in acquisition and reconstruction. In addition, it is cheap and portable. One drawback of using 3D camera systems is the limited depth of field. Faces need to be placed at a specific distance from the camera for optimal face surface reconstruction. The camera needs to be calibrated correctly and missing data due to occlusion or improperly reflected regions may cause errors. Papatheodorou and Rueckert [229] utilised the VRT3D to capture 3D face surfaces in the Imperial College database.

![Figure 3.2: (a) the VisionRT stereo camera system (b) captured images.](image)

Stereo-based cameras can also be built by hand. Lee and Milios [186] set-up a camera system (see Figure 3.3), which uses calibrated cameras and a robust stereo reconstruction method to create a dense and accurate 3D representation. The system works both indoors and outdoors in a variety of conditions. The system takes two images with an angular baseline of a few degrees between $8^\circ$ to $15^\circ$. The correspondence between similar features of two face surfaces can be established. However, this self-built system had problems on areas with high curvatures (e.g. the tip of the nose). Archermann et al. [13] used a coded light approach to acquire grey level range images. This 3D acquisition system is low in cost and relatively high in speed.

Another common form of 3D acquisition system is the laser scanner. It is based on optical triangulation illuminating a tiny spot on the surface of the object. The camera tracks the image of the spot, finding the centre pixel. It then traces a line of sight through that pixel until it intersects with the illumination beam at a point on the surface. The beam is scanned in a series of planes crossing the object surface, allowing the complete object surface to be captured.

Examples of commercial laser systems are the Minolta Vivid series [213] and Cyberwave 3D
3.3 3D face recognition

![3D face recognition system](image)

Figure 3.3: (a) Lee and Milios 3D shape acquisition system (b) Left: captured images. Right: Reconstructed 3D shape and texture mapping

Scanner [76]. The Cyberware scanner used by Archermann et al. [13], Beumier and Acheroy [29] and Gordon [123] claimed to capture 3D face information accurately and provides a panoramic view of the object. The disadvantages of this system are the high cost and slow acquisition time. Lu and Jain [198], Srivastava et al. [269] and Lapreste et al. [183] used the Minolta Vivid 700 scanner. The reconstructed 3D geometric information is in the form of polygonal meshes. An example is illustrated in Figure 3.4. The number of mesh points depends on the resolution of the scanner. Other available 3D scanners are the 3DMD [1], Geometric [73], A4Vision [153] and Genex [154].

Laser scanners provide depth information of the anatomical structures instead of its visual appearance. Other range imaging technologies are available, but are less commonly used. Range images have some advantages over 2D intensity images. They are robust to illumination and colour changes because each pixel corresponds to a depth value rather than a colour. The work on using range images on face recognition was first proposed by Laprest et al. [183].

The quality of the reconstructed 3D model is highly dependent on the 3D sensors being used. Laser-based scanners provide a high quality of reconstruction. However, the scanner is bulky and costly. On the other hand, stereo-based camera systems are cheap and fast, but more
3.3 3D face recognition

Figure 3.4: An example of a 3D face model acquired using the Minolta Vivid scanner (a) surface model (b) mesh model

error prone.

3.3.2 Surface-based techniques

Surface-based approaches rely on curvature, profile lines and distance metrics to describe faces. Local surface-based methods use local features of the face. The facial features (e.g. the eyes, nose and mouth) are extracted from each face and their characteristics are compared. Lee and Milios [186] used facial features for establishing correspondence between human faces and for comparing their similarity. The convex regions that correspond to the features are computed using the extended Gaussian images (EGI) method. The extracted convex regions are used to match the computed convex regions from other faces. Their method explores the feasibility of region-based matching of 3D range images in face recognition. Their proposal was unable to deal with areas of relatively high curvature e.g. the tip of a nose, when applying EGI interpolation. This was because of the scanning resolution.

Gordon [123] demonstrated 3D face recognition using surface curvature and depth values to segment faces into its features. The feature descriptors are extracted from a face and they are mapped onto a vector. Each face is represented by a vector placed in a feature space which is used for the recognition stage. Recognition rates of 80% to 100% were reported using three views of eight faces with the condition that the features are correctly detected and no facial expression variation exists. According to Gordon, 3D face recognition holds more advantages over intensity image based face recognition in terms of the rich information gathered on the
region of the features. Even with limited sample numbers, the overall computation process was lengthy and costly. In addition, his approach works with fixed facial expressions.

Another work using 3-dimensional feature descriptors for recognising faces was by Moreno et al. [218]. Based on the mean and curvature of the Gaussian, facial features can be extracted and segmented. Each image represents a feature vector that contains eighty-six features. When using the Fisher discriminant, thirty-five most discriminant eigenvectors were computed and provided 78% accuracy.

Chua and Jarvis [64] proposed using additional features which can be extracted locally. The technique uses point signatures. Examples of point signatures are peak, saddle, pit, valley, roof edge and ridge. These point signatures are used to compare the signatures of data points on other face surfaces [63]. To improve the recognition performance, the point signatures are rigidly registered to each other and only the rigid parts of the face that do not change much with expression (such as the eyes, nose and forehead) are included (as shown in Figure 3.5 and Figure 3.6).

![Point signature examples](image taken from [64])
Wang et al. [300] also based their recognition system on point signatures. Chua and colleagues [63] only manipulate geometric information, while Wang and his fellow researchers used both 3D shape and 2D texture features. They used Gabor filters to get the feature points and PCA to place the facial feature vectors into a subspace. The classification was done using SVM and DDAG (decision directed acyclic graphs). Using fifty subjects with a set of facial expressions in different viewpoints, the recognition performance was more than 90%. This approach demonstrated that using a combined feature vector can significantly improve the recognition rate.

An interesting piece of work by Xu et al. [309] combined the 3D point cloud taken from a regular mesh fitting and a 3D mesh represented in the form of Gaussian-Hermite moments, using a PCA space. Using the nearest-neighbourhood distance measurement, the recognition rate for thirty faces reached 92%. However, as the number of faces used increased, the recognition rate decreased. This effect could be due to the limited local information.

Global methods use the whole face instead of only selected facial features. Cartoux et al. [54] proposed the use of the whole surface on top of the extracted local curvature values to calculate similarity distances and their recognition rate reaches 100%, but only when using five subjects. Pan et al. [112] based their recognition on a priori knowledge and used the Hausdorff distance for similarity measurement. On a face dataset of thirty subjects, 3.2% equal error rate (EER) was reported while 5% of EER was measured when using PCA. The equal error rate (EER) is the point where the false acceptance rate is equal to the false rejection rate. The Hausdorff distance is defined as the maximum distance between two point sets on surface A and surface B. The details of this measurement can be found in Rucklidge’s paper [252]. Other work that applied Hausdorff distance includes Lee and Shim [187], Russ et al. [253] and Achermann.
and Bunke [12].

The popular ICP (iterative closest point) algorithm was also used to generate a distance subspace and to recognise 3D faces. Medioni et al. [207] obtained a recognition performance of 90% with 100 subjects. Papatheodorou and Rueckert [229] also applied ICP in recognition and achieved a rate of between 95-97%. Maurer et al. [205] also applied similar techniques. Lu et al. [197] extended the method by performing hybrid ICP. In their experiment, a recognition rate of 92% was reached using eighteen faces of different semi-profile faces and facial expressions.

To overcome facial expression changes, non-rigid transformation further explained in Section 4.2.1.3 was employed by Lu and Jain [198] to register face surfaces. They also applied deformation using a small training set with the aim of distinguishing the identity of the subject from the facial expressions. A SVM was used to perform the classification. The recognition performance was measured on a hundred subjects and a rate of 89% was achieved.

Beumier and Archeroy’s [29] recognition method is similar to Gordon’s [123] approach. They employed a vertical profile of the 3D model for recognition purposes. The coefficients of the feature vectors were estimated using Fisher’s method [176]. When using twenty-seven subjects, the EER rate dropped to 1.4%. Profile-based 3D face recognition was also employed by Wu et al. [308]. The calculated EER results were similar to Beumier and Archeroy [29].

### 3.3.3 Statistical techniques

When applying PCA on 3D models, there are a few issues that require attention. Different modalities such as depth require correct alignment. Failing to do this results in low recognition performance.

Hesher et al. [144] explored using different numbers of eigenvectors, computed from PCA, on range images. The face dataset they used had six different facial expressions for each of the thirty-seven subjects. They concluded that the number of training and query faces can influence the recognition rates. Using a higher number of training samples and a lower number of query images resulted in higher recognition rates. Their recognition results reached 90% accuracy when using 185 training images and 37 query images.

Some early work using PCA combined both 3D face models and 2D face images. Tsalakanidou
et al. [286] combined texture and depth modalities for face recognition. They used PCA on forty subjects and a recognition rate of 99% was achieved. Archermann et al. [13] also applied PCA and used HMM for face recognition. They claimed that the result of their experiment, based on their strategy of using only 24 data items, was very promising. However, no recognition rate was presented in the report.

Chang et al. [57] applied PCA on a combination of 3D face models and 2D face images. The depth information gathered from the range images was normalised using a set of selected landmarks. Experiments were conducted to evaluate the accuracy of the sensor in capturing the depth information of the face. The results provide a valuable insight into the quality of depth representation that is required for face recognition. Their experiments indicated that 3D-based techniques are more sensitive to low quality data than 2D-based techniques. Using similar low resolution, 89% accuracy was achieved with 2D images, whereas 3D range images only achieved 61%.

Bronstein et al. [45, 46] applied a geodesic distance representation so that faces change very little due to facial expressions. The results from the approach are canonical representations. Figure 3.7 shows the flattened faces and their corresponding flattened texture with their canonical images. PCA was applied on the canonical representation, and with thirty subjects they achieved a recognition rate of 100%. They also claimed that their approach is able to distinguish identical twins.

Figure 3.7: Texture mapping of the facial surface, A, and its canonical form, B, which results the flattened texture, C, and the canonical image, D (image taken from [46])

An improvement on Bronstein’s approach was made by Pan et al. [112] in which they claimed to provide a more flexible and adaptive solution. A region of interest (ROI) was
normalised to a reference plane, triangulated and parameterised into an isomorphic 2D planar circle. PCA was performed on the planar circle representation of the depth values. The results showed a 95% recognition rate compared to using PCA techniques which achieved a 90% recognition rate.

Gökberk et al. [121] utilised LDA and PCA as data reduction on depth and profile information. The reported recognition rate was 96.3%. The authors also classified the information using ICP on mesh and point cloud representations. By combining those representations, 99.1% recognition rate was reported.

### 3.3.4 Model-based techniques

Model-based approaches aim to construct models of human faces that are able to capture facial variations. Prior knowledge is required for the reconstruction of 3D data from 2D images. The early work on reconstruction based on 3D human face surfaces was by Blanz and Vetter [38]. They used a 3D morphable face model created from a set of 3D faces with a range of appearances [36]. Their approach used linear 3D models and assumed that the linearity extends to the 2D projection of 3D [295]. The principle is similar to the ASM by Cootes et al. [70], except that instead of using all the pixel information, they extracted features of the objects. The approach used by Blanz and Vetter is sometimes called *interpretation through synthesis*. The process of morphing objects can be generalised into two main parts: 1) generating a statistical shape model 2) using the parameters to morph the model. Figure 3.8 illustrates the overall approach of the *interpretation through synthesis*. The representation of faces in terms of model coefficients \( \alpha_i, \beta_i \) for 3D shape and texture is independent of viewpoint. For recognition, all probe and gallery images are processed by the model fitting algorithm. The morphable model generates a transformed front view of faces which is used in recognising a set of frontal gallery views.

Blanz et al. [37] used a gallery of 68 subjects illuminated from the same lighting direction which was queried with 4,420 images of the same subjects in 3 poses under 22 different illumination changes. Using images for which the model fitted well earns a recognition rate of 92.5%. This recognition rate is considered high in view of the pose and illumination variations. If the model fitted poorly, a recognition rate of 82.6% was achieved.
The approach used by Blanz and colleagues requires a high cost of computation. Due to this limitation, Huang et al. [150] proposed a component-based technique which combines both Blanz and Vetter [36] 3D morphable models and the component-based detection method by Heisele et al. [142]. Heisele’s component-based method decomposes a face into a set of ten anatomical facial features, as shown in Figure 3.9. Three dimensional morphable models are used to synthesize arbitrary images under varying pose and illumination. The components of the face are trained using a SVM classifier and the recognition rate was 88%.

Lu et al. [198] made use of 3D face models in face recognition. The models are taken from different viewpoints. The technique integrated shape matching with a constrained appearance-based method. Landmark areas were used for registering and aligning the surfaces from different viewpoints. A hybrid ICP method was used for surface matching and point-to-point distance was used as a similarity metric. A 98% recognition rate was obtained when combining ICP and LDA. However, the recognition rate dropped to 91% when emotional expressions were included in the training set. The authors then explored the use of Active Appearance Models introduced...
by [70] to synthesize expressions, aging and illumination variations.

Passalis et al. [234] used statistical data constructed from an average 3D mesh and local deformations of landmarks placed on the surface vertices. They used the wavelet transform to compress the encoded shape information. A recognition rate of 90% was achieved when the extracted signature of the encoded 3D shape was compared to the gallery or training of 3D shapes.

All the 3D based methods introduced so far are summarised in Table 3.2 displaying the methods used, the dataset sizes and the recognition performance.

Table 3.2: Overview of 3D face recognition techniques

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dataset size</th>
<th>Methods</th>
<th>Reported performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee and Milios [186]</td>
<td>6 faces of 6 subjects</td>
<td>Extended Gaussian Images (EGI) interpolation</td>
<td>N/A</td>
</tr>
<tr>
<td>Gordon [123]</td>
<td>26 faces of 26 subjects</td>
<td>Surface curvature and depth values</td>
<td>80-100%</td>
</tr>
<tr>
<td>Moreno et al. [218]</td>
<td>60 subjects</td>
<td>Mean and curvature of Gaussian</td>
<td>78%</td>
</tr>
<tr>
<td>Chua &amp; Jarvis [64]</td>
<td>15 models</td>
<td>Point signatures</td>
<td>66-79%</td>
</tr>
<tr>
<td>Wang et al. [298]</td>
<td>300 faces of 50 subjects</td>
<td>SVM &amp; DDAG</td>
<td>&gt; 90%</td>
</tr>
<tr>
<td>Reference</td>
<td>Dataset size</td>
<td>Methods</td>
<td>Reported performance</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------------------------------</td>
<td>----------------------------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Chua et al. [63]</td>
<td>36 faces of 6 subjects</td>
<td>Point signatures</td>
<td>83 - 94%</td>
</tr>
<tr>
<td>Xu et al. [310]</td>
<td>120 subject (max)</td>
<td>Point cloud in the form of Gaussian-Hermite moments</td>
<td>92%</td>
</tr>
<tr>
<td>Cartoux et al. [54]</td>
<td>42 subjects</td>
<td>Extracted local curvature</td>
<td>98 - 100%</td>
</tr>
<tr>
<td>Pan et al. [112]</td>
<td>943 faces of 276 subjects</td>
<td>PCA</td>
<td>3.2% EER</td>
</tr>
<tr>
<td>Lee &amp; Shim [187]</td>
<td>84 faces of 42 subjects</td>
<td>Curvature-based method</td>
<td>98%</td>
</tr>
<tr>
<td>Russ et al. [253]</td>
<td>200 subjects</td>
<td>Hausdorff approach</td>
<td>93.5-98%</td>
</tr>
<tr>
<td>Archermann &amp; Bunke [12]</td>
<td>240 faces of 24 subjects</td>
<td>Hausdorff approach</td>
<td>100%</td>
</tr>
<tr>
<td>Medioni &amp; Waupotitsch [207]</td>
<td>700 faces of 100 subjects</td>
<td>ICP</td>
<td>91%</td>
</tr>
<tr>
<td>Papatheodorou &amp; Rueckert [230]</td>
<td>216 faces of 54 subjects</td>
<td>ICP</td>
<td>95 - 97%</td>
</tr>
<tr>
<td>Maurer et al. [205]</td>
<td>4,009 faces of 466 subjects</td>
<td>Hybrid ICP</td>
<td>87% verification</td>
</tr>
<tr>
<td>Lu et al. [197]</td>
<td>63 faces of 10 subjects</td>
<td>Hybrid ICP</td>
<td>96%</td>
</tr>
<tr>
<td>Lu &amp; Jain [198]</td>
<td>700 faces of 100 subjects</td>
<td>ICP</td>
<td>91%</td>
</tr>
<tr>
<td>Beumier &amp; Acheroy [29]</td>
<td>81 faces of 27 subjects</td>
<td>ICP</td>
<td>1.4% EER</td>
</tr>
<tr>
<td><strong>Statistics-based approaches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hesher et al. [144]</td>
<td>222 faces of 37 subjects</td>
<td>PCA</td>
<td>90%</td>
</tr>
<tr>
<td>Tsalakanidou et al. [286, 285]</td>
<td>80 faces of 40 subjects</td>
<td>PCA</td>
<td>99% 2D+3D, 93% 3D only</td>
</tr>
<tr>
<td>Chang et al. [58]</td>
<td>951 faces of 200 subjects</td>
<td>PCA</td>
<td>99% 2D+3D, 93% 3D only</td>
</tr>
<tr>
<td>Bronstein et al. [46]</td>
<td>220 faces of 30 subjects</td>
<td>Canonical geodesic distance</td>
<td>100%</td>
</tr>
</tbody>
</table>
### 3.4 Recognition performance measurements

Recognition performance of a biometric system can be assessed in terms of either identification or verification. A gallery $G$ containing $n$ labeled faces makes up the database and is used to determine the eigenspace of a given transformation vector for recognition. Queries can be made with unknown probes for identification or labeled probes for verification.

Identification can be either open-set or closed-set [130]. Closed-set identification is used when it is known that the probes always belong to the gallery data set. In this case the closest gallery sample to the query is matched with it. In an open-set identification, the probes may or may not be represented in the gallery. Thus, a test is required to determine whether the closest match is sufficiently accurate to be considered significant.

Verification is the process in which a person’s identity is known a priori. The verified query pattern is compared with the known person’s data. A distance threshold is applied to verify the probe or to classify it as an impostor. The threshold can be set to determine the required

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dataset size</th>
<th>Methods</th>
<th>Reported performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pan et al. [112]</td>
<td>943 faces of 276 subjects</td>
<td>PCA+Canonical geodesic distance</td>
<td>95%</td>
</tr>
<tr>
<td>Gokberk et al. [121]</td>
<td>579 faces of 106 subjects</td>
<td>PCA+LDA</td>
<td>96%</td>
</tr>
<tr>
<td><strong>Model-based approaches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blanz et al. [37]</td>
<td>4,420 faces of 68 subjects</td>
<td>3D Morphable model</td>
<td>93%</td>
</tr>
<tr>
<td>Huang et al. [149]</td>
<td>200 faces of 10 subjects</td>
<td>Component based 3D Morphable model</td>
<td>88%</td>
</tr>
<tr>
<td>Heisele et al. [142]</td>
<td>N/A</td>
<td>Component based + SVM</td>
<td>88%</td>
</tr>
<tr>
<td>Lu et al. [197]</td>
<td>598 faces of 100 subjects</td>
<td>ICP+LDA</td>
<td>98% neutral faces, 91% expressioned faces</td>
</tr>
<tr>
<td>Passalis et al. [234]</td>
<td>4,007 faces of 446 subjects</td>
<td>Deformable model + Wavelet transform</td>
<td>90%</td>
</tr>
</tbody>
</table>
operating conditions for the system. If the selected threshold is small, the number of impostors is low, but correct probes may be rejected. However, if we increase the threshold, the number of false negatives reduces at the expense of accepting more impostors.

3.5 Discussion and Conclusions

Comparing different techniques based on their published performance can be misleading. Experimental results are presented using different data and experimental conditions. According to Xu et al. [309] the size of the dataset will influence the performance of the algorithms significantly. Using an algorithm on a small dataset will produce a higher recognition rate than on a large dataset. Xu and colleagues found a recognition rate of 96.1% when using 30-person datasets but when using the same algorithm on a dataset with 120 subjects, the recognition rate dropped to 72.4%. In addition, the number of samples per subject can influence the evaluation scores. If the training set contains only one face sample per subject, the recognition rate will be lower than the case where there are many faces of the same subject in the training set [58]. If the training and query datasets contain only neutral faces, the recognition rates will be higher than if the query face set contains several facial expressions. The quality of the captured face data may also influence the recognition accuracy. For example, a set of face data taken from VisionRT 3D stereo camera produce less accurate results than a similar set taken with a laser scanner.

Many 3D face recognition algorithms are restricted to neutral face expressions [64, 123, 229, 207]. In recent work, researchers made use of facial expression information in the face recognition algorithms [63, 196, 212, 198, 45, 144]. In some cases, facial expressions are avoided by excluding the non-rigid area of the face from the recognition algorithms [63, 45]. Other work in recognising faces with expression makes use of deformation [38, 198, 234]. The literature reveals the importance of accounting for facial expressions in face recognition systems, both for recognition and expression identification.
Chapter 4

Pre-processing

This chapter discusses the pre-processing techniques that we used on raw 3D data scans that were used during the project. The 3D data were acquired using a VisionRT 3D system and a 3DMD range scanner. The VisionRT 3D face data belongs to Imperial College London [230] and the 3DMD faces were obtained from Binghamton University, New York [317]. We applied pre-processing on those datasets. Each of the 3D face surfaces was stored using Cartesian coordinates.

A pre-processing step is required to regularise the surfaces and to correct flaws in the raw geometric data caused by hair, clothing boundaries and reconstruction errors. The pre-processing step includes cropping the face regions. Reconstruction error correction will include the removal of holes and spikes, and the triangulated meshes will be standardised to a set of corresponding vertices.

The pre-processing methods we used were researched and developed by our colleague Dr Papatheodorou, in his PhD work [229]. They use rigid and non-rigid registration to align, and then correct and standardise the 3D face models. Although we did not contribute to this work, it is included here as it is an important step that may affect the experimental results presented in this thesis.

4.1 Problems in 3D face datasets

A common problem with raw 3D surfaces are reconstruction errors causing holes and spikes in the geometry. Facial hair and eyes cause difficulties in reconstruction. Each acquired surface is tessellated in a different way, using a different space coordinate system. The pre-processing
step is required to remove errors from the face data, and to regularise the data to have the same surface tesselation in a common coordinate system.

Patching holes can be done using interpolation techniques [40]. Gaussian smoothing and linear interpolation techniques are used in range data [12, 21, 25, 29, 31, 40, 43] to cover the holes. Closest distance measurement is used to regularise the meshes [48, 33].

Most aligning algorithms employ either the centres of mass, the nose tip or the eyes as reference points, or fit a plane to the face. Different subjects are aligned by registration. Registration creates an accurate correspondence between faces with different geometric shapes or facial expressions. Rigid registration finds the closest alignment without changing the shape. One of the commonly used rigid registration algorithms is the Iterated Closest Point (ICP) algorithm [33, 38, 40, 45]. Non-rigid registration using warping can align the surfaces better. A popular non-rigid registration method is the thin-plate splines (TPS) technique [38].

Registration can be based on all the surface points or just a few landmarks. Work that uses landmark information can be found in [14, 25, 37, 38]. Most landmarks are entered manually but a few systems use automatic landmarking [27, 33, 49].

### 4.2 Registration

The purpose of registration is to find the transformation $T$ that will map a surface point, $a$ in one data set to its corresponding point, $b$ in another data set. It can be written as:

$$b = T(a)$$

where $a = (a_x, a_y, a_z)$ and $b = (b_x, b_y, b_z)$ are corresponding points on two surfaces. Figure 4.1 illustrates the transformation in 2-dimensional space.

In the case of faces, rigid registration can correct pose variations [229] without achieving exact alignment. Non-rigid registration is used to align corresponding areas exactly with each other, for example, in the same face with different expressions. The next subsection explains the different registration transformation types.
4.2 Registration

4.2.1 Registration transformation types

The transformation model imposes mathematical constraints on the type of geometric distortions that can be used during the registration procedure. The number of parameters needed to describe a transformation is called the degree of freedom. For example, if a transformation utilises only translation on a 3D face surface, then it will have three degrees of freedom. The domain of the transformation is global if the transformation applies to the entire data and local if the transformation is applied to only a part of the data.

Transformations that preserve parallel lines are called affine. Rigid transformations do not change shapes, but only translate and rotate them. Non-rigid transformations freely deform objects.

4.2.1.1 Rigid transformation

A rigid transformation preserves all object distances and internal angles. It is expressed as a combination of rotation $R$ and translation $t$ using Eq 4.1:

$$ b = Ra + t $$

Figure 4.1: The transformation $T(a)$ transforms point $a$ in face A into its corresponding location in face B
The expanded matrix vector form is written as:

\[
\mathbf{T}_{\text{rigid}}(a_x, a_y, a_z) = \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix}
\] (4.3)

where the $3 \times 3$ matrix is an orthogonal rotation matrix and $t = [t_x, t_y, t_z]^T$ is the displacement vector describing the translation component.

### 4.2.1.2 Affine transformation

Affine transformation is another geometric transformation that preserves parallel lines but not lengths and angles. It can be expressed by:

\[
b = \mathbf{A}a + t
\] (4.4)

where matrix $\mathbf{A}$ is a matrix represented as:

\[
\mathbf{T}_{\text{affine}}(a_x, a_y, a_z) = \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix} \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix}
\] (4.5)

where $\mathbf{A}$ is a $3 \times 3$ matrix describing the scale, shear and rotation components of the transformation.

### 4.2.1.3 Non-rigid transformation

A non-rigid transformation is a free-form geometric transformation in which a set of control points or landmarks are used to interpolate points within an area. A non-rigid transformation is adopted when one surface must be freely deformed to fit another. For example, in face models, the affine transformation cannot map a neutral face on a smiling face. Instead, non-rigid transformation is required as illustrated by Figure 4.2.

Spline-based transformation is a simple form of non-rigid transformation which maintains topology and smoothness. A detailed explanation of spline functions used in computer graphics and modelling can be found in [19]. A regular grid of control points is used to define the object.
space. The position of each object point is defined in terms of the local position of the control grid, and movement of the grid points causes regular displacement of each object point in the space.

### 4.2.2 VisionRT 3D face scans

The VisionRT 3D camera, described in the previous chapter, has a speed acquisition of up to 30 frames per second and the speed of data reconstruction is less than 5 seconds on a 1GHz machine. This allows real-time acquisition in realistic scenarios. The high speed data acquisition can prevent motion artifacts from being introduced. The RMS error is just under 1 mm for a typical 3D face acquisition. Figure 4.3 displays an example of a captured and reconstructed face surface.

The subjects in the Imperial College database were mostly students within an age range of 18-35 years. Each subject was acquired in several different head positions and three facial expression poses. The facial expressions were smiling, frowning and neutral. The different head positions were frontal, 45° left and right turn and a 45° upward tilt. An example of data captured with the VisionRT 3D is shown in Figure 4.4.

Each raw reconstructed face model has between 8,000 and 12,000 polygon points. In total, the face dataset has 57 subjects with 4 females and 53 males. The subjects were also classified in terms of their ethnicity as 8 South Asians, 6 East Asians, 1 Afro-Caribbean and 42 Caucasians.

### 4.2.3 Binghamton 3D face scans

The Binghamton’s 3D facial range data was captured with a 3D face imaging system, the 3DMD digitiser as shown in Figure 4.5, using a random light pattern speckle projection. The 3D
Figure 4.3: An example of raw reconstructed faces. The images on the left are without texture, those on the right have texture.

Figure 4.4: Different poses captured with the VisionRT 3D camera.
imaging system has six digital cameras and two light pattern projectors positioned on each side. The system automatically merges all six synchronised camera images to produce a 3D surface mesh. Each surface has between 20,000 and 35,000 polygons depending on the size of the subject’s face. The mesh resolution is higher than the 3D models captured using the VisionRT 3D system. The texture size is around 1300 by 900 pixels. Figure 4.6 shows some raw face models.

![Figure 4.5: The 3D face imaging system setup](image)

![Figure 4.6: Raw face models](image)

Each subject was asked to present seven facial expressions which were neutral, happiness, surprise, fear, sadness, disgust and anger. Each of these expressions (excluding neutral) had four different intensity ranges: low, middle, high and highest (see Figure 4.7). The intensity of the expressions varies at the subject’s discretion. As the 3D acquisition system does not capture facial expression dynamically, the subjects were asked to freeze and pause on each expression
for a few seconds. There were 100 subjects including undergraduates and graduates from the departments of Psychology, Arts and Engineering. The database consists of about 60% female and 40% male subjects with a variety of ethnic ancestries including Caucasian, Afro-Caribbean, East Asian, Middle Eastern, Hispanic and others. In total, the 3D face database contains 2,500 3D face models.

Figure 4.7: First and third rows are the happy expression with four levels of intensity. The second and fourth rows are the surprise expression with four levels of intensity

4.2.4 Notre Dame 3D face scans

The Notre Dame University face database was acquired using the Minolta VIVID 910 camera [214], which uses a structured light sensor to scan surfaces. The light reflected from the surface is captured by a CCD camera, which is then used to reconstruct the surface by inferring the 3D shape from the distortion of the light pattern. The camera is also able to capture colour texture information.

The reconstructed surfaces have higher resolution in comparison to the VisionRT surfaces. A typical face has about 20,000 points before any processing. The facial features are better defined with fewer holes or artifacts created. The drawbacks of the face database are that it
4.3 The pre-processing steps

Papatheodorou’s pre-processing technique uses registration to establish correspondence. First, a 3D rigid registration is carried out, then non-rigid registration aligns the face exactly with a reference face. The tessellation of the reference face can be mapped to the new subject face, establishing correspondence between each of the control points and the texture. Accurate registration can only be achieved through the use of known corresponding landmarks between the reference surface and a new subject. A set of landmarks must be manually placed on each face surface at anatomically distinctive points. The process of meticulous manual landmarking is very slow. Many authors have tried to automate the landmarking process, however the automatic methods are high in error [71, 151].

Landmarks are placed on feature points of the 3D surface. In the pre-processing step, the
eyebrow regions are not selected as part of the feature points since the borders of the eyebrows are not as constant as the eyes or mouth. In addition, they are difficult to identify manually especially with low resolution images and surface meshes. The face surface can be rotated within the face space to ease the selection of points. The face texture is also used to make the landmark placement easier. Previous work on 3D face modelling has shown that there is little improvement in using more than 11 landmarks [151]. Papatheodorou used 13 landmarks. An average of three minutes is taken to manually landmark a single face surface accurately. Table 4.1 shows the landmarks that were used and Figure 4.9 shows an example of a manually landmarked face.

Table 4.1: The 13 anatomical landmark points

<table>
<thead>
<tr>
<th>Points</th>
<th>Landmark No.</th>
<th>Landmark Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glabella</td>
<td>5</td>
<td>Area in the centre of the forehead between the eyebrows, above the nose which is slightly protuding.</td>
</tr>
<tr>
<td>Eyes</td>
<td>1-4</td>
<td>Both the inner and outer corners of the eyelids are landmarked.</td>
</tr>
<tr>
<td>Nasion</td>
<td>6</td>
<td>The intersection of the frontal and two nasal bones of the skull where there is a clearly depressed area directly between the eyes above the bridge of the nose.</td>
</tr>
<tr>
<td>Nose tip</td>
<td>7</td>
<td>The most protuding part of the nose.</td>
</tr>
<tr>
<td>Lower philtrum</td>
<td>8</td>
<td>The middle point at the base of the philtrum.</td>
</tr>
<tr>
<td>Lips</td>
<td>9-12</td>
<td>Both left and right corners of the lips as well as the top point of the upper lip and the lowest point of the lower lip.</td>
</tr>
<tr>
<td>Gnathion</td>
<td>13</td>
<td>The lowest and most protuding point on the chin.</td>
</tr>
</tbody>
</table>

Assuming that the face is a rigid body and the sensor captures the faces perfectly, then the mean square error between the points on a source surface to the points on a target surface can be used to determine the identity of the subject.

4.3.1 Rigid registration

Once the landmarks have been placed on every single face surface, the mean landmark is calculated. Based on the computed mean, all the face surfaces are registered to one another using rigid transformation on each landmark. As a result, all face surfaces are brought into close alignment using only the landmarks to calculate the rigid transformation. Figure 4.10 shows a
4.3 The pre-processing steps

Figure 4.9: The 13 landmark placements on a face surface

rigid transformation based on landmarks.

Figure 4.10: Rigid registration using landmarks information. The top row shows the two faces aligned to the mean landmarks. The bottom row shows a frontal 2D polygon projection of the outer landmarks of the same polygon before and after registration (image taken from [229])
4.3 The pre-processing steps

4.3.2 Non-rigid landmark registration

After the anatomical landmarks have been rigidly aligned, the next step is to use non-rigid registration to maximise the correspondence between the faces. This step is important to pair-up the corresponding points on all the surfaces because face surfaces are not of the same size and shape and rigid registration alone is unable to pair-up all the points correctly with other face surfaces. Figure 4.11 shows the correspondence points before and after non-rigid registration. The triangle points are the landmark points and the rest of the circle points are points around the landmarks.

Free-form deformation (FFD) using a multi resolution B-spline approach was used to model the local non-rigid deformation. The deformation grid control points are gradually adjusted to reduce the distance between the corresponding landmarks as shown in Figure 4.12. Figure 4.13 illustrates faces before and after non-rigid landmark registrations.

4.3.3 Establishing correspondences

Once the landmarks have been rigidly and non-rigidly registered, a correspondence can be determined for every point on the surfaces. This is done by using a face template. The template is a
4.3 The pre-processing steps

Figure 4.12: A free-form deformation and the corresponding mesh of control points (image taken from [229])

Figure 4.13: Non-rigid registration using landmarks information. The top row shows the two faces aligned to the mean landmarks. The bottom row shows a frontal 2D polygon projection of the outer landmarks of the same polygon before and after non-rigid registration (image taken from [229]).
typical face, not part of the population, that is free from artifacts. It was manually processed to make the point distribution on its surface regular, and then warped to the mean landmarks using the B-spline transformation.

Given a face $A$, the resulting face after rigid and non-rigid landmark-based registrations to the template face is denoted $A'$. Both $A'$ and the face template are in close alignment, particularly near the landmarks as illustrated in Figure 4.14.

![Figure 4.14: The distance colour map after non-rigid registration (image taken from [229])](image)

Once registered, the template mesh is used to resample surface $A'$. This step automatically standardises the number of points and can also be used to cover holes and remove spikes from $A'$. The covering of holes can be done by linking up the points around the edge of a hole. A more accurate point based registration can then be carried out to bring the surfaces into more accurate alignment.

Finally, the template can be subjected to the inverse of the registration transformation to create a new clean surface $A''$ with the geometry of $A$ but the topology of the template mesh.

Figure 4.15 displays a face where each point is colour coded to show the distance between a face surface $A$ and a reference surface $Y$. The red colour signifies the smallest distance while the blue signifies the larger distance. It can be seen that the mean distance between the face and the base mesh has reduced globally after the non-rigid surface registration. The red stripes on the face in (b) shows that non-rigid surface registration brings two surfaces closer globally while in face (a), the closest points between two surfaces are around the landmarks.
4.4 Experiments and Results

Once the pairing has been established, the point coordinates of the surface are copied over to the template mesh. The complete overall pipeline of the aforementioned pre-processing steps is as shown in Figure 4.16.

![Diagram](image)

Figure 4.16: The overall pre-processing steps (image taken from [229])

4.4 Experiments and Results

After the correspondence of each point has been established and each surface has the same number of points, PCA can be applied to the set of 3D face surfaces. Figure 4.17 describes the variance over the eigenvalue components of 150 subjects using the Notre Dame 3D face datasets. Figure 4.18 shows the shape variations between -3, 0, and +3 standard deviations using a landmark-based registration model.
4.4 Experiments and Results

Figure 4.17: The variation described by the modes of the landmark-based model

<table>
<thead>
<tr>
<th>Landmark Registration-Based Principal Shape Modes</th>
<th>$-3\sqrt{\lambda}$</th>
<th>mean</th>
<th>$+3\sqrt{\lambda}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mode 1</td>
<td><img src="image1" alt="mode 1" /></td>
<td><img src="image2" alt="mode 1 mean" /></td>
<td><img src="image3" alt="mode 1 +3sqrtlambda" /></td>
</tr>
<tr>
<td>mode 2</td>
<td><img src="image4" alt="mode 2" /></td>
<td><img src="image5" alt="mode 2 mean" /></td>
<td><img src="image6" alt="mode 2 +3sqrtlambda" /></td>
</tr>
<tr>
<td>mode 3</td>
<td><img src="image7" alt="mode 3" /></td>
<td><img src="image8" alt="mode 3 mean" /></td>
<td><img src="image9" alt="mode 3 +3sqrtlambda" /></td>
</tr>
<tr>
<td>mode 4</td>
<td><img src="image10" alt="mode 4" /></td>
<td><img src="image11" alt="mode 4 mean" /></td>
<td><img src="image12" alt="mode 4 +3sqrtlambda" /></td>
</tr>
</tbody>
</table>

Figure 4.18: The first 4 most discriminant components between $\pm 3\sqrt{\lambda}$, where $\lambda$ is the standard deviation
The population distribution on the first two principal modes of variation is shown in Figure 4.19. It can be seen that the population forms a peak, which means that there are no sub-clusters in the population distribution and all the faces are spread uniformly in a two dimensional space.

![Shape distribution using landmark-based registration](image)

Figure 4.19: The shape distribution of the landmark-based model projected on the first two most discriminant modes

### 4.5 Conclusions

In this chapter, the pre-processing method done of Papatheodorou [229] was explained. Pre-processing is a necessary step before further analysis can be made. Since we have three different 3D face datasets, captured using three different 3D acquisition systems, pre-processing can be used to clean, regularise and establish correspondences between the face surfaces.

The pre-processing technique was evaluated using PCA to investigate the distribution in the face space. The pre-processing methods produced encouraging results and the cleaned faces were visually acceptable. The overall pre-processing process takes about 30 minutes on 50 subjects. The only tedious and painstaking task is the manual landmarking process, which
takes an average of 3 minutes to place landmarks on each face surface. Once all the 3D face datasets have been pre-processed, further analysis can be made. The next chapter will discuss the analysis of facial movements in different facial expressions.
Chapter 5

Statistical 3D facial analysis

In this chapter, we discuss quantitative analysis of facial expressions using a collection of 3D face surface datasets. Each face surface is recorded in an array of surface points in three dimensional space. The movement of landmark points is analysed according to the motion of facial features.

In this experiment, we used the Imperial College 3D faces. This dataset was chosen due to the availability of the facial expression information and higher quality of expression texture. A choice of landmarks was made based on craniofacial anthropometry and Facial Action Coding System (FACS) frameworks. Each 3D surface was annotated with the same set of chosen landmarks. The spread and the variance of those points with different subjects and facial expressions were studied and statistically analysed. The aim of this experiment is to identify those face areas containing high facial expression information which may be useful for facial expression analysis, face and facial expression recognition and facial expression animation. This chapter will also discuss current face applications that are based on landmark information.

5.1 Introduction to landmarking

Landmarks are defined as characteristic points that can be used to establish correspondence between different objects within populations of the same class [82]. There are various synonyms for landmarks, such as points, nodes, vertices, anchor points, fiducial markers, model points, markers, key points, etc. There are three basic groups of landmarks:

1. Anatomical (traditional) landmarks

   Anatomical landmarks are points that have a unique identifiable place on the object, for
example, the corner of the eyes and the tip of the nose. Those points can be located easily regardless of the orientation of the object. The landmarks may be independent of any coordinate system or dependent on a particular orientation or a coordinate system. For example, the points at the corners of the eyes are independent of the position and the orientation of the face. Points on the mandible, such as the tip of the chin may rotate relative to the rest of the face as the jaw moves.

2. Fuzzy landmarks (Mathematical landmarks)

Fuzzy landmarks are points corresponding to an average estimation of the location within the identifiable area on the object [290, 301]. They can be points that are mathematically calculated and located on the object. As fuzzy landmarks are often located on featureless structures, there is a possibility of high placement error [290]. Fuzzy landmarks are used on smooth surfaces where there are insufficient anatomical landmarks [189]. High curvature points or extremum points are the examples of fuzzy landmarks. Areas around the nose and eyebrows on a face have high curvature points and hence can be used as fuzzy landmarks. The placement of fuzzy landmarks can be improved by taking the average of multiple landmarkings.

3. Constructed landmarks (Pseudo landmarks)

Constructed landmarks are points on the object at positions defined using a combination of anatomical landmarks and geometric information. Constructed landmarks can be placed manually or computed automatically. For example, landmarks around the cheek areas are constructed using the location of two or more traditional landmarks. Midpoints between the two traditional landmarks are identified and their location are recorded as a constructed landmark. The error in locating constructed landmarks depends on the error in the location of each of the traditional landmarks used in the construction [189].

The studies that use landmark data ensure that the selected landmarks are homologous. Homologous landmarks are those that share a common cause or descent. This means that each landmark placed on an object will have a corresponding landmark on another object. The choice of landmark points depends greatly on the study of interest and on the basis of scientific
5.2 Craniofacial anthropometry

The study of measuring living creatures is known as anthropometry. In the case of human faces, the geometric measurement process is called craniofacial anthropometry. Landmarking on human face was first carried out at Charles University, Prague in the 1960s. In craniofacial anthropometry, the landmark points are selected by identifying visible or palpable features on the human face. These landmarks are measured using specific procedures and measuring instruments.

There are fifteen craniofacial measurements - the head measurements, head inclinations, face horizontal measurements, face vertical measurements, face lateral measurements, surface measurements, angles and inclinations, orbit measurements, orbit inclinations, nose measurements, nose inclinations, nasal angles, orolabial measurements, orolabial inclinations and ear measurements. Each measurement consists of additional sub-measurements (refer to [174] for the measurements), and specific instruments are required to take each measurement. Handling the different instruments makes the process of measurement strenuously time-consuming. The detailed process to undertake each measurement can be found in Kolar and Salter [174].

A total of seventy-one craniofacial landmarks have been defined around the head, face, orbits, nose and mouth region and ear. Figure 5.1, 5.2 and 5.3 illustrate landmarks displayed on the face of a child. The labels signify the measurement names. Craniofacial landmarks are taken on soft tissue reflecting the underlying bony face structures. In order to get an accurate measurement, the landmarks are first marked on the skin with a dermographic pen while the person is kept in a rest position with neutral expression.

The application of craniofacial anthropometry started with studies of children with cleft lip and palate in the 1960s. It was extended in the 1970s and 1980s to record surgical results through pre-operative and post-operative measurements. Farkas [100] together with Hajnis [101] developed a set of measurements to record the growth and development of children with cleft lip and nose. They carried out craniofacial anthropometry in the role of cleft evaluation that involved soft tissue anthropometry. Similarly, Kolar and Salter [174] worked on cleft
Figure 5.1: Craniofacial landmarks superimposed on a frontal face image (image taken from [174])

Figure 5.2: Left: Main craniofacial landmark around the head. Right: Profile landmarks (image taken from [174])
5.3 Background on landmark-based face applications

We now consider landmarking for our three main applications: face registration, face recognition and facial expression analysis.

5.3.1 Face registration

In face registration, we need to ensure that anatomical features correspond correctly across subjects. Landmark points are used to coarsely align all faces together for further recognition and synthesis processes. The landmarks selected should be placed on prominent anatomical facial features and at extreme points. The three often cited craniofacial landmarks are on the both corners of the eyes and the tip of the nose. Other features such as the mouth corners, the bridge of the nose and the tip of the chin are also suitable for registration.

Tao and Veldhuis [274] considered five prominent features (the tip of the nose, the corners
of the eyes and mouth, as shown in Figure 5.4 as landmark points. Hutton [151] used nine landmark points (see Figure 5.5) placed at both corners of the eyes and mouth, the glabella, the tip of the nose and chin. Papatheodorou [229] chose thirteen craniofacial landmarks as shown in Figure 5.6.

![Figure 5.4: Face detection and landmark placements (image taken from Tao and Veldhuis [274])](image)

![Figure 5.5: Face scan with nine landmarks (image taken from Hutton [151])](image)

### 5.3.2 Face recognition applications

In face recognition applications, in order to achieve a better recognition performance, high resolution face data is required. Thus, the denser the set of landmarks, the better. The term resolution denotes either the number of pixels acquired from a digital camera or the number of surface points obtained from a 3D scanner. It is recommended that a face image requires more than 50 pixels between the eyes [41]. Anything less makes detecting and recognising
5.3 Background on landmark-based face applications

Figure 5.6: Thirteen manually selected landmarks - chosen because of their anatomical distinctiveness (image taken from Papatheodorou [229])

faces difficult. Baker and Kanade [16, 15] investigated and developed methods to increase and enhance the resolution of image data using the hallucination algorithm which firstly recognises features then applies smoothing.

If the resolution is high, for example with point separations of 1 mm, computational demand can increase. Studies by [41, 321, 178, 298] have shown that reducing the image resolution can still achieve high recognition performance.

Besides using all the surface points or pixel data, a selection of landmarks can also be used in face recognition applications [122, 47, 75]. Landmark-based recognition is performed in the same way as geometric recognition. For example, three-dimensional landmark positions can produce a simple polyhedral representation of a face, as illustrated in Figure 5.7(b) from a dense surface point map of a face as shown in Figure 5.7(a).

Landmark-based face recognition has been of interest to many face recognition researchers as the number of face datasets and the resolution of face images and 3D models increases. Fiducial landmarks aim to reduce high dimensionality problems. By using them, recognition computation may become faster and more efficient [264, 238]. The results of the face recognition vendor test (FRVT 2002), indicated that the performance of face recognition degrades as the size of a face search space increases [238]. Landmarking was proposed to allow a more effective search within a face space. Precise landmarks are essential for effective face recognition performance [248, 28]. This suggest that the placement of landmarks and their geometry should
5.3 Background on landmark-based face applications

Shi et al. [264] studied the effectiveness of landmark-based face recognition. They identified twenty-nine craniofacial landmark positions over a distribution of 944 face images (see Figure 5.8). Brunelli and Poggio [47] extracted thirty-five geometric features based on the mouth and nose, eyebrows and face outline. Craw et al. [75] investigated thirty-four manually selected landmarks in different automatic face recognition codings (as illustrated in Figure 5.9). Ferrario et al. [104] defined twenty-two landmarks to study sexual dimorphism on human faces.

Figure 5.7: (a) Three-dimensional polyhedral face shape (b) Polyhedral face shape based on a set of landmarks

Figure 5.8: The location of the twenty-nine landmarks used by Shi et al. [264]
A study done by Goldstein et al. [122] showed how well human faces can be identified by both human and machine using descriptive feature information extracted from photographed faces. Twenty-one feature descriptions were extracted from the computer training. Even though the descriptive features were static values, those selected features are similar to the craniofacial landmarks and they were locations around the head, eyebrows, eyes, nose, mouth, chin, ears, cheeks and forehead.

Kaya and Kobayashi [166] also conducted a similar experiment that explores the parameter characteristics of faces. Nine landmark parameters were manually extracted from 2D photographed images. These parameters were the internal bi-ocular breadth, external bi-ocular breadth, nose breadth, mouth breadth, cheek bone breadth, jaw bone breadth, distance between the lower lips and chin, distance between the upper lip and nose, and the height of the lips. Similar to Goldstein and colleagues’ work, the parameters were based on the eyes, nose, mouth, cheek, jaw and chin.

Some psychologists assessed what facial cue saliency is used by humans to recognise a face. Davies et al. [79] asked a number of people to describe faces and their features. They gathered that the thirteen most frequently pointed out features were (in no particular order of importance) the hair, eyes, nose, eyebrows, face shape, chin, lips, mouth, ears, face lines, complexion, forehead and cheeks. Rhodes [128] further studied these features and concluded that the main appearance that characterised faces are the eyes, eyebrows and mouth. These are known as the first-order features. The rest of the features were based on the spatial relationships of the first-order features.
Studies have also been done on recognition from profile images of faces [165, 136, 185, 307, 50]. The side view of faces reveals depth and curvature information that is not seen in the frontal view. Previous work on facial profiles extracted fiducial points by heuristic rules. Harmon [136] manually drew the outline of profile photos and nine landmark points were selected: the forehead, bridge, nose top, nose bottom, upper lips, mouth, lower lip, chin and throat. A set of eleven features were derived from these fiducial points. Those features are made up from the distance between the feature points and the measurement of areas, angles and curvatures. Later, Harmon et al. [136, 185] analysed recognition performance using seventeen geometric profile features. Wu and Huang [307] extracted twenty-four profile features for face recognition. Chen et al. [62] extracted eleven feature points from a contour line between two feature points. The feature points are as shown in Figure 5.10.

![Feature points on a contour line](image_taken_from_62)

Most profile-based face recognition methods depend on the location of the fiducial points. Automatic detection of landmark points on the profiles is known to be more difficult and can be unreliable in comparison to manual landmarking especially in subjects with a well-defined nose, protruding lips or flat chin.

### 5.3.3 Facial expression applications

In facial expression analysis, the anatomy and physics of facial muscles is important [92, 301, 239]. Facial muscle movements correspond to the generation of different facial expressions. Facial expression analysis and expression recognition therefore needs landmarks that carry information about the movements. There are similar landmarking requirements for methods of
5.3 Background on landmark-based face applications

synthesizing facial expressions. Landmark points are chosen to characterise actions of the different facial muscles. Cohn [68] identified forty-six landmarks located around the palpable features to analyse expressions from a sequence of image. Figure 5.11 displays the manually landmarked fiducial points, which can be edited easily.

Terzopoulos and Waters [278] used eleven dynamic fiducial points placed on the left and right eyebrows, the nasolabial furrows and the chin (see Figure 5.12). Pighin [240] identified thirteen feature points shown in Figure 5.13. They used these points to estimate the face position, face orientation and facial expression given a sequence of video images. Emotional expressions are animated from the acquired different expressions measured from individual photographs.

Essa and Pentland [96] developed an automatic landmark placement method using 2D spatio temporal motion energy information. Eight landmarks were selected and they are located on
distinct features on face (see Figure 5.14). The landmarks contain facial motion information which is useful for expression extraction and classification.

Black and Yacoob’s [35] approach used areas around the craniofacial points, as shown in Figure 5.15. The models are used to measure deformations such as affine facial motions (e.g. head movement and rotation) and a curvature model of nonrigid motions of facial features around the eyes and mouth.

Another work that captures visual facial information using landmark points is by Pantic and Rothkrantz [228]. They used a dual-view face model (a frontal-view and a profile-view) to extract facial features in order to reduce the ambiguities of face geometry. The number of the
acquired landmarks for the frontal-view is ten and nineteen landmarks for the profile-view (see Figure 5.16). These landmarks are placed on palpable features.

Zhang et al. [318] proposed a technique to synthesize detailed facial expressions by analysing the motion of feature points that are manually labelled on a face. They used one hundred and thirty-four selected feature points, as shown in Figure 5.17.

Similar to Zhang’s approach, Song [268] proposed a generic method to transfer expressions based on local geometry vertex coordinates. These coordinates are the fiducial points labelled on 3D surfaces. The feature point sets are displayed in Figure 5.18 and include the contour of the eyebrows, eyes, the tip and bridge of the nose, mouth, jaw, cheek and forehead.

Blanz et al. [36] worked on animating photo-realistic expressions by transferring geometric expression descriptions captured from manually landmark points on the face of a subject. The landmarks are placed at the bottom half of the face around the cheeks and mouth.
Figure 5.17: Feature points (image taken from [318])

Figure 5.18: (a) Feature points on a 3D face model and (b) 2D face image (images taken from [268])
Facial analysis also includes the analysis of changes of face shapes and face features due to speech. Kalberer and Gool [163] worked on animating 3D faces shapes during speech. The detailed analysis of the 3D face shape is based on fiducial points placed on the face and lips. Thirty-eight points are placed on the lips and a hundred and twenty-four points are placed on other parts of the face that are influenced by speech (see Figure 5.19). The geometric descriptions are learned so that speech can be animated.

![Figure 5.19: (a) Manually placed features on a subject; (b) 3D polyhedral surface; (c) 3D mask fitted on the face markers (d) 3D face surface (image taken from [163])](image)

Interesting work by Reveret and Essa [247] animated the motion around the lips to create a realistic animation. The lip model is represented as a parametric surface guided by thirty markers on one side of the face and then mirrored to the other side of the face. Figure 5.20 shows the result of alignment against the manually landmarked points.

![Figure 5.20: Markers put on the face and the result of the alignment on specific features (image taken from [247])](image)

Basu [22] addressed the problem of lip movements in speech production. In their work,
a 3D model of lip motion is tracked and reconstructed from video for each subject. The lip motions are extracted by tracking ink markers painted on the lip surface. The observed features are then trained and learned using a statistical approach involving modeling the subspace of lip motions to describe and fit the observations in the video stream.

5.4 Issues in Landmarking

We have seen that the choice of landmarks is specific to the aim of the research. For example, in face registration applications, only a few landmarks are required and usually placed on well defined craniofacial features. Whereas, in facial expression analysis, more landmarks are required and they are constructed on the basis of craniofacial landmarks. They are placed on mobile areas around the eyebrows, cheeks and mouth.

Unfortunately, most expression analysis applications have not thoroughly studied the process of placing landmarks or verifying that the placements are effective. Many have selected those landmarks located on facial muscles or that are FACS-based. However, they have not verified that those landmarks correctly characterise facial muscle activities. The issues in landmarking led us to the following research questions:

- Which landmark placements have maximum face and facial muscle motion information?
- Is it adequate to only use craniofacial landmarks?
- How are the facial, muscle-based, landmarks selected?
- Is the error between corresponding landmarks minimal when repeating similar landmarking processes on similar faces?

5.5 Our proposed landmarking approach

In our work, we select a number of landmarks placed on both distinct features as well as areas around distinct features. These landmarks are now analysed in terms of facial muscle motion based on FACS. In practice, landmark points must be easy to identify and also mobile in the different facial expressions. Locating prominent facial features, such as at the corners of the
eyes and mouth, and the tip of the nose, is easier, in comparison to identifying areas where no palpable features are visible and available. The areas around the forehead, cheek and mouth make an important contribution to facial expression analysis.

A set of craniofacial landmarks is used as the basis to construct landmarks on facial muscles. The most commonly used craniofacial landmarks are the corners of the eyes, the bridge of the nose, the tip of the nose, the corners of the mouth and the chin. Since we are dealing with 3D face surfaces, we have included features which are not distinct in 2D images but are seen on 3D face surfaces. They are the glabella, subnasal, the top and bottom of the eyes and the mouth. In total, we chose thirteen craniofacial landmarks. The selected landmarks are similar to those used in the pre-processing step (see Figure 5.6).

From the selected craniofacial landmarks, we can construct pseudo landmarks. The constructed landmark positions were selected by analysing the rules of FACS. Mathematical formulae using a distance measure were used to locate the constructed landmarks. The pseudo landmarks are placed on AU areas around the eyebrows and eyes, the cheeks and the mouth. These areas have no palpable features, but linear-based pseudo landmarks can be computed on the basis of the selected craniofacial landmarks. For example, one pseudo landmark on the cheek is located by drawing a vertical line through the middle of the eyes to meet the horizontal line at the subnasal level (see Figure 5.21, landmark 30 and landmark 31). These landmark points are constructed in the neutral expression examples only. They will move relative to the craniofacial basis with expression changes.

Figure 5.21 shows the craniofacial and pseudo landmarks. In the initial stage of our landmark-based expression analysis, we selected twenty pseudo landmarks that incorporate the AU rules. In total, we have thirty-three landmarks.

5.5.1 Experiments and results

Craniofacial and pseudo landmarks were manually placed on each of the 3D face surfaces tested. The landmark placements were analysed in two main experiments. The aim of the experiment is to analyse the facial expression variance invariant of gender. In the first experiment, ten volunteers manually placed landmarks on the face surfaces. The aim of the first experiment is to
analyse the consistency in landmarking by different individuals. Table 5.1 shows the discrepancies found in each of the thirty-three landmarks, measured by the standard deviation.

The results show that the precision (measured in standard deviations from the mean) of manual landmarking was about 0.97 mm for a resolution of 5090 mesh points. All the volunteers did the landmarking without having to repeat the landmarking process due to error in placing the landmark points. When asked about the difficulty of identifying the landmarks, all the test volunteers said that the landmarks are all simple to locate.

The second experiment is to test the variation of those landmark points across the different facial expressions. The aim is to see how much information about expressions is conveyed by each landmark. The landmarks are gathered from the first experiment and additional landmarks are captured from all the faces.

Figure 5.22 illustrates the movement of the 3D landmarks across the different facial expressions. The numbers represent the landmark’s numbering in sequence and the different coloured symbols signify the different type expressions. All of the thirty-three landmarks are weighted
Table 5.1: The measured input error on the 33 landmarks using standard deviation (s.d.)

<table>
<thead>
<tr>
<th>Landmark No</th>
<th>s.d.</th>
<th>Landmark No</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.67</td>
<td>18</td>
<td>0.69</td>
</tr>
<tr>
<td>2</td>
<td>1.7</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.0</td>
<td>20</td>
<td>0.77</td>
</tr>
<tr>
<td>4</td>
<td>1.97</td>
<td>21</td>
<td>0.66</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>22</td>
<td>1.47</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>23</td>
<td>0.3e-05</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>24</td>
<td>0.3e-05</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>25</td>
<td>1.69</td>
</tr>
<tr>
<td>9</td>
<td>2.28</td>
<td>26</td>
<td>1.67</td>
</tr>
<tr>
<td>10</td>
<td>0.8</td>
<td>27</td>
<td>1.68</td>
</tr>
<tr>
<td>11</td>
<td>2.48</td>
<td>28</td>
<td>1.49</td>
</tr>
<tr>
<td>12</td>
<td>1.78</td>
<td>29</td>
<td>1.79</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>30</td>
<td>1.38</td>
</tr>
<tr>
<td>14</td>
<td>1.59</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>1.77</td>
<td>32</td>
<td>0.89</td>
</tr>
<tr>
<td>16</td>
<td>0.55</td>
<td>33</td>
<td>0.43</td>
</tr>
<tr>
<td>17</td>
<td>0.87</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 5.22: The spread of the thirty-three landmarks in a 3-dimensional space

The spread of landmarks across expressions can also be viewed in a 2-dimensional space as shown in Figure 5.23.

There are three facial expressions in the analysis and they are neutral, smiling and frowning. The mean expression is the average of all the expressions:

$$\bar{x} = \frac{\sum x}{N}$$  \hspace{1cm} (5.1)

The placements of the craniofacial and pseudo landmarks are analysed using positional variance. The variation of landmarks is compared to the neutral face. We can use the variance to identify those points or areas that do not contribute to expressions. Figure 5.24 plots the 3D coordinates of the expression mean and the neutral mean landmarks.

The expression mean is the mean of all the facial expressions. The neutral mean is the mean of the neutral expressions in the 3D face database (see Figure 5.25).

The dispersion of the thirty-three landmarks is measured statistically across all expressions of the fifty subjects. The variance for each landmark is computed. The landmark positions are against the overall mean.
5.5 Our proposed landmarking approach

Figure 5.23: The spread of the thirty-three landmarks in a 2-dimensional space

Figure 5.24: The neutral mean and expression mean of the thirty-three landmarks viewed in a 2-dimensional space
then averaged and the mean is used as the origin to measure the variation across expressions. Figure 5.26 shows the variance of smiling and frowning expressions. The high variances on landmark 21 and 23 are due to the two corners of the mouth.

Figure 5.26 and Figure 5.27 shows the landmark displacements across the expressions. High variance change can be seen between landmark 21 and landmark 29 (area around the mouth). The variance can also be sorted in an increasing order as shown in Figure 5.28. From this graph, we can see that the landmarks around the mouth and eyes change significantly with the different generated expressions.
5.6 Discussion

There are two types of variation in the experiments: geometric variation and facial expression variation. The geometric variation includes positional variation of facial features. Figure 5.29 illustrates the overall variation in the neutral faces. Figure 5.30 shows the mean landmark positions of each subject. The inclusion of different genders and races increases geometric variation. For example in the available face datasets, Chinese faces have a smaller nose and shorter nose bridge in comparison to Caucasian faces.

Facial expression variation is illustrated in Figure 5.23. It shows the dispersion of the landmarks across the expression data particularly around the mouth. There is an obvious variation between smiling and frowning (see Figure 5.31). In Figure 5.32, the distinct separation between smiling and frowning can be seen on areas around the mouth, and they can roughly be separated by a horizontal y-axis (see landmarks 21 to 26, 28 and 29).

Unfortunately, not all the craniofacial and pseudo landmarks contain high expression information. The inside corner of the left and right eyes (landmarks 3 and 5) and the tip of the nose...
5.6 Discussion

Figure 5.29: The spread of landmarks on all the neutral faces (2 per subject) in the x-y plane

Figure 5.30: The spread of mean landmarks (averaged for each subject) in the neutral faces in the x-y plane
Figure 5.31: The spread of mean landmarks of smiling and frowning faces against the overall expression mean in the x-y plane.

Figure 5.32: The spread of landmarks in smiling and frowning against the expression mean and the neutral mean in the x-y plane.
(landmark 19) have the least positional change to expression variance (as shown in Figure 5.33) and the variance is as shown in Figure 5.28.

Figure 5.33: Landmarks with the least spread, number 3, 5 and 19. The landmarks represent the inside corner of the eyes and the tip of the nose (landmark 19)

5.7 Conclusions

The conclusions that can be reached from the analysis of the selected craniofacial and pseudo landmarks are as follow:

- **There are significant variations in some landmark positions due to facial expressions.**
  
The corners of the mouths and some of the pseudo landmarks are highly mobile and could have significant discriminant power in recognising expressions.

- **The tip of the nose, the inside corners of the eyes and the bottom eyes do not contribute much to facial expressions.**
This could be due to the minimal facial muscle activity around the area. As for the tip of the nose, there is no facial muscle there. The variation is caused by the neighbouring facial muscles.

- **Facial expressions are similar in all individuals.**
  This is shown by the fact that the expression mean is quite close to the neutral mean.

- **Smiling and frowning can be easily distinguished by landmark movements.**
  Around the mouth area, especially at the corners, the landmarks all move up when smiling and down when frowning.

These experiments were designed to determine the best placement of landmarks for distinguishing different facial expressions. They show that there is a noticeable difference in the displacement of the landmarks under different expressions. Identifying landmarks that are mobile in facial muscle movements is a good strategy for characterising expressions. Mobility of the different landmark points will be an important factor in their ability to discriminate different expressions. However, it will not be the only one. Further work on the discrimination power of landmarks requires them to be tested in the context of a recognition experiment.
Chapter 6

3D Face Recognition Using Multilinear Decomposition

Principal component analysis (PCA) is a commonly used technique to represent the total variation of a dataset efficiently. However, PCA’s good performance is limited to problems with a few linearly separated modes of variation. In the context of face recognition, there are often several independent dimensions of change (for example subject, gender, illumination, texture, pose and facial expression) that can influence the accuracy of face recognition. In order to separate these different sources of variation well, the tensor model has been introduced. Tensor models organise data according to the different types of variation which later allow easy manipulation of data independently across different dimensions. Each dimension of a multi-dimensional array represents one type of variation. The tensor model is manipulated using multilinear algebra which is fully described by Tucker [287, 184]. By employing a tensor model to represent 3D face data, large intra-subject and inter-subject variations may become manageable. Using a tensor model to characterise facial expression variation should result in better recognition performance.

The tensor framework is used in conjunction with multilinear singular value decomposition (SVD) methods with the aim of improving the recognition rate of 3D face surfaces invariant of different facial expressions. The advantage of using SVD is that the uniqueness and the orthogonality of the matrices are assured.

In this chapter, we make use of a sub-tensor SVD within the tensor model where the identity of subjects and facial expression factors can be separated for better recognition performance. The two main goals of this work are: firstly, to use the tensor framework for recognising faces invariant of different facial expressions, and secondly, to analyse a set of facial dynamic vectors
classified in a tensor model for recognition purposes.

6.1 Review of Principal Component Analysis

Processing data within a high dimensional space may lead to computational complexity of \( O(Nd^2) \). With that, a method is required to reduce the dimensionality of data. The dimension reduction removes redundant or unimportant data while preserving effective classifications. Principal component analysis achieves this reduction by creating an eigenspace. The concept of PCA was originally introduced by Karl Pearson [235] in 1901 although the method is also attributed to Hotelling [148] and Karhunen-Loeve who worked independently on similar problems. The aim of PCA is to reduce the dimensionality of the original space with a minimum loss of information by finding the projection directions that maximise the total scatters across all classes.

The first standard eigenspace-based face recognition method was developed by Sirovich and Kirby [266] and is known as eigenfaces. The eigenface approach used principal component analysis to find a smaller representation of the original space. However, the projection directions that maximise the total scatter also retain unwanted variations such as lighting and facial expressions.

Following from Sirovich and Kirby’s work was an extended PCA approach proposed by Turk and Pentland [288] for face recognition research. Since then, PCA has been the basis of numerous research projects in face detection and recognition. Because of its popularity, PCA serves as a benchmark for evaluation of new algorithms [24, 237, 273, 322].

Given that there are \( M \) 3D face surfaces and each surface, \( \Gamma_m \) consists of \( n \) points \( \{p_j = (p_{x_j}, p_{y_j}, p_{z_j})\} \) where \( j = 1 \ldots n \), the average 3D surface is calculated using:

\[
\bar{\Gamma} = \frac{1}{M} \sum_{j=1}^{M} \Gamma_j
\]  

(6.1)

The mean centered face data can be calculated using vector differences:

\[
\Omega_j = \Gamma_j - \bar{\Gamma}
\]  

(6.2)
The covariance matrix of each point is then computed by:

$$
\Sigma = \frac{1}{M} \sum_{j=1}^{M} \Omega_j \Omega_j^T
$$

(6.3)

The eigenvectors $\Phi_j$ and the eigenvalues $\Lambda_j$ can be obtained from the covariance matrix. The non-zero eigenvalues $\Lambda_j$ and associated eigenvectors $\Phi_j$ are ordered from the largest values to the smallest values. The eigenvector with the largest eigenvalue is the vector oriented in the direction of the most variance in the data. All the eigenvectors with the corresponding eigenvalues form the basis subspace known as eigenspace.

All surfaces are then projected into a space by

$$
\varphi_k = \Phi_k^T (\Gamma - \bar{\Gamma})
$$

(6.4)

where $k = 1, \ldots, m$. The value of $m$ is the number of non-zero (or non-trivial) eigenvalues and depends on the application. Every face in the new space is described as a vector of weights $\varphi^T = [\pi_1, \pi_2, \ldots, \pi_m]$. The weights dictate the magnitude that the principal eigenfaces contribute in describing the input surface.

According to Hutton [151], the number $m$ of eigenfaces required for good recognition should ensure that 98% of the population variation can be described:

$$
\sum_{k=1}^{m} \frac{\lambda_k}{\sum_{j=1}^{M} \lambda_j} \geq 0.98
$$

(6.5)

The similarity between two face surfaces, $\varphi_A$ and $\varphi_B$, can be measured using a distance metric on the weights. There are three main distance measurements: Euclidean distance, Mahalanobis distance and MahCosine distance. The Euclidean distance is defined as:

$$
Euclidean \ distance \ (\varphi_A, \varphi_B) = \|\varphi_A - \varphi_B\| = \sqrt{\sum_{i=1}^{m} (\varphi_{Ai} - \varphi_{Bi})^2}
$$

(6.6)

Turk and Pentland [288] used the Euclidean distance to classify a new face projected into an eigenspace. They identified four possibilities:

- the face is near the feature-space mean and near a known face class
- the face is near the feature-space mean but not near a known face class
• the face is not near the feature-space mean or a face class

• the face is not near the feature-space mean but is near a face class

In the Mahalanobis metric, the variance along each dimension is normalised to 1. The two face shape parameters are compared by dividing the differences with the corresponding standard deviation, $\sigma$. The Mahalanobis distance for classes with zero co-variance is calculated using:

$$ Mahalanobis \ distance \ (\varphi_A, \varphi_B) = \sqrt{\sum_i^m (\varphi_{Ai} - \varphi_{Bi})^2 / \sigma_i^2} $$

(6.7)

The MahCosine distance metric measures the cosine of the angle between surfaces after they have been transferred to the normalised eigenspace:

$$ MahCosine \ distance \ (\Gamma_A, \Gamma_B) = \cos(\theta_{\Gamma_A, \Gamma_B}) = \frac{\phi_A \cdot \phi_B}{||\phi_A|| \cdot ||\phi_B||} $$

(6.8)

Some face researchers have reported an improvement when using MahCosine distance than other distance metric measurements (Euclidean and Mahalanobis distances) \[60\] depending on the applications and the variations. However, the most popularly used distance measurement is the Euclidean distance because the processing is faster than other distance measures.

### 6.2 Review of Singular Value Decomposition

Another statistical method that is similar to the PCA is Singular Value Decomposition (SVD). SVD is a more widely applied technique for diagonalising matrices, and can be applied to a rectangular $m \times n$ matrix $A$ using the following equation:

$$ A = UDV^T $$

(6.9)

where matrix $A \in \mathbb{R}^{m \times n}$ is the product of two orthonormal matrices, $U \in \mathbb{R}^{m \times m}$, $V^T \in \mathbb{R}^{n \times n}$ and a pseudo-diagonal matrix $D = diag(\sigma_1, ..., \sigma_\rho) \in \mathbb{R}^{m \times n}$, with $\rho = \min(m, n)$. SVD is illustrated in the tableau below:
The columns in $U$ are called the left singular vectors and the rows of $V^T$ are the right singular vectors. Both the $U$ and $V^T$ matrices are orthogonal. In relation to PCA in the case where principal components are calculated from the covariance matrix, the eigenvectors of $AA^T$ yields the columns of $U$ and the eigenvectors of $A^TA$ make up the rows of $V^T$. The singular values in the $D$ matrix are square roots of the eigenvalues from $AA^T$ or $A^TA$. Similar to PCA, the diagonal elements can be arranged in descending order.

Equation 6.9 can be rewritten as the weighted sum of the corresponding outer products by denoting the column vectors of $U$ and $V^T$ by $u_i$ and $v_i^T$, and $\sigma_i$ the magnitude of the diagonal elements of singular values in $D$ sorted in a decreasing order, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\rho} \geq 0$. The $\rho_0$ is used as the index of the last non-zero singular value, that is $\sigma_i = 0$ for $i > \rho_0$ and $|\sigma_i| > 0$ for $i \leq \rho_0$.

$$A = \sum_{i=1}^{\rho_0} \sigma_i u_i v_i^T$$  (6.11)

If the number of rows $m$ of matrix $A$ is greater than the number of columns $n$ (see Equation 6.10), SVD can still be performed whereby the last $(m-n)$ rows of $D$ will be zero and can therefore be omitted, along with the last $(m-n)$ columns of $U$.

SVD can be used in applications that identify and order the data, according to its variance, along different dimensions. This means that a high dimensional and high variable set of data points can be reduced to a lower dimensional space that exposes the substructure of the original data, in order to find the best compact approximation. This process is known as data reduction, which can also be carried out in PCA. Any variation below a particular threshold, $\epsilon_e$, can be ignored without impairing the data representation [215].

Depending on the application, the SVD method may be used when the dimensionality of
the original data is very large, and therefore the covariance matrix is large [215]. The effective dimensionality of the data can also be obtained from the rank $r$ of matrix $A$. The exact rank of $A$ is the non-zero singular values of the diagonal matrix $D$. If $A$ is nonsingular, then the rank would be $n$. And if $A$ is singular, then the rank would be less than $n$. According to Michael et al. [209], this important organisational ability of SVD can detect weak signals in data.

SVD can be used to provide a measure of the linear independence between the column vectors of the matrix and this measure is called a condition number denoted as:

$$\text{cond}(A) = \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}$$

(6.12)

where $\sigma_{\text{min}}$ and $\sigma_{\text{max}}$ are the smallest and the largest singular values of $A$.

If the $\text{cond}(A)$ is too large, matrix $A$ is said to be ill-conditioned. If there are one or more zero singular values, the condition number becomes infinite. Hence, a threshold $\epsilon_e$ is usually defined to identify the index of the smallest singular value from which the condition of $(\sigma_{\text{max}}/\sigma_{\text{min}}) > \epsilon_e$ is computed.

The two sets of eigenvectors gathered from SVD and PCA are identical. This shows that either SVD or PCA can be used to extract the eigenvectors. Any projection of a 3D face surface into the new feature space requires the generation of an $(m \times n)$ matrix using the equation below:

$$\psi_k = \Xi((\Gamma - \bar{\Gamma}))$$

(6.13)

where $k = 1, \ldots, m$ and $\Xi$ are eigenvectors taken from either $U$ or $V^T$ depending on the rank $r$ of the matrix $A$. If $m > n$, the matrix $V^T$ would be used. When $r < n$, the matrix $U$ is employed.

Since SVD is similar to the PCA method, face recognition researchers could apply either approach. According to Belhumeur et al. [24], neither approach is optimal from the discrimination standpoint. The mathematical background for both methods are similar in that the corresponding eigenvalues of SVD are the square roots of the eigenvalues of the covariance matrix of PCA. The next section describes the use of multilinear algebra to reconstruct a tensor model.
6.3 The Fundamental Concepts of Multilinear Algebra

Multilinear algebra is the algebra of higher-order tensors that defines multilinear operators over a high dimensional space. It offers a unifying mathematical framework suitable for addressing a variety of computer vision problems [294]. Multilinear analysis manipulates a basic object known as the tensor. A tensor is a multidimensional matrix or mode-\(n\) matrix, and is useful for the description of higher order quantities.

The notation of multilinear algebra distinguishes scalars, vectors, matrices and higher-order tensors, as shown below:

- scalars are represented by \(a, b, \ldots\) (lower case letters)
- vectors are represented by \(\mathbf{a}, \mathbf{b}, \ldots\) (bold lower-case letters)
- matrices are represented by \(A, B, \ldots\) (italic upper-case letter)
- higher order tensors are represented by \(\mathcal{A}, \mathcal{B}, \ldots\) (calligraphic upper-case letters)

In multilinear algebra, the \(N^{th}\) order tensor is written as \(\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}\), and its elements are indexed as \(\mathcal{A}_{i_1, i_2, \ldots, i_N}\) or \(a_{i_1, i_2, \ldots, i_N}\). Multilinear algebra is a higher order generalisation of vectors (1\(^{st}\) order tensors), matrices (2\(^{nd}\) order tensors) and cubes (3\(^{rd}\) order tensors), which are illustrated in Figure 6.1. The \(N^{th}\) order tensor has mode-\(n\) spaces. For example, if \(N = 2\), two mode spaces exist, which is a matrix defined in terms of a set of mode-1 vectors as the column vectors and a set of mode-2 vectors as the row vectors.

In a matrix, the tensor element can be expressed as \(B = [b_{ij} \ldots b_{NN}]\), where the element of the matrix \(B_{ij}\) has \(i\) row index and \(j\) column index. In the case of a 3\(^{rd}\) order tensor \(\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}\), there are three mode spaces. Mode-1 corresponds to the column space, mode-2 corresponds to the row space and mode-3 corresponds to the depth space. This is illustrated in Figure 6.2.

6.3.1 Tensor Flattening

A tensor \(\mathcal{A}\) can be flattened into a matrix form, \(A_n\) along any dimension \(n\) where \(n = I_1, I_2, \ldots, I_N\). Flattening (also known as unfolding or matricizing) of an \(N^{th}\)-order tensor is a process to reduce the tensor representation along row and column spaces. In general, flattening has a cyclic nature and it is defined as follows [184]:

\[
A_n = \begin{bmatrix}
A_{n1} & A_{n2} & \cdots & A_{nN}
\end{bmatrix}
\]
6.3 The Fundamental Concepts of Multilinear Algebra

Figure 6.1: Visualisation of the order of tensor. A vector is a $1^{st}$ order tensor, a 2D matrix a $2^{nd}$ order tensor and a 3D matrix is a $3^{rd}$ order tensor etc.

Figure 6.2: (a) illustrates a 3-mode tensor $\mathbf{A}$ with 3 different elements (axis) which give 3 mode spaces as shown in (b), (c) and (d) (adapted from Lathauwer et al. [184])
6.3 The Fundamental Concepts of Multilinear Algebra

Assuming an $N^{th}$ order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, the unfolded matrix $A_n \in \mathbb{R}^{I_n \times (I_{n+1} + I_{n+2} + \ldots + I_N I_1 I_2 \ldots I_{n-1})}$ contains the element $a_{i_1 i_2 \ldots i_N}$ at the position with row number $i_n$ and column number equal to

$$j = (i_{n+1} - 1)I_{n+2}I_{n+3} \ldots I_N I_1 I_2 \ldots I_{n-1} + (i_{n+2} - 1)I_{n+3}I_{n+4} \ldots I_N I_1 I_2 \ldots I_{n-1} + \ldots + (i_N - 1)I_1 I_2 \ldots I_{n-1} + (i_1 - 1)I_2 I_3 \ldots I_{n-1} + (i_2 - 1)I_3 I_4 \ldots I_{n-1} + \ldots + i_{n-1}$$

Figure 6.3 illustrates flattening of a $(I_1 \times I_2 \times I_3)$ tensor along the three mode spaces. For example, a tensor $\mathcal{A}$ matrix of $(3 \times 2 \times 3)$ is $a_{111} = -a_{113} = a_{122} = -a_{212} = -a_{312} = -a_{313} = -a_{323} = 1, a_{213} = -a_{222} = a_{311} = 2, a_{121} = a_{321} = 3, a_{223} = 4$, and $a_{112} = a_{123} = a_{211} = a_{221} = a_{322} = 0$. The unfolded matrix $A_2$ is:

$$\begin{pmatrix}
1 & 0 & 2 & 3 & 0 & 3 \\
0 & -1 & -1 & 1 & -2 & 0 \\
-1 & 2 & -1 & 0 & 4 & -1
\end{pmatrix}$$

An $N^{th}$ order tensor $\mathcal{A}$ has rank-1 when $\mathcal{A} = u_1 \circ u_2 \circ \ldots \circ u_N$, where $\circ$ denotes the tensor product and $u_i$'s are vectors of $\mathbb{R}^{I_i}$. The rank of an $N^{th}$ order tensor $\mathcal{A}$, denoted

$$R_n = rank_n(\mathcal{A}) = rank(A_n), \quad (6.14)$$

is the minimal number of rank-1 tensors ($\mathcal{A}$) that satisfy a linear combination:

$$\mathcal{A} = \sum_{r=1}^{R} \sigma_r u^{(r)1} \circ u^{(r)2} \circ \ldots \circ u^{(r)n} \quad (6.15)$$

where $\sigma$ denotes the singular values.
Figure 6.3: Flattening of a \((I_1 \times I_2 \times I_3)\)-tensor in 3 ways to obtain matrices comprising \((I_1 \times I_2 I_3)\) matrix mode-1 vectors \(A_1\), \((I_2 \times I_3 I_1)\) matrix mode-2 vectors \(A_2\) and \((I_3 \times I_1 I_2)\) matrix mode-3 vectors \(A_3\)
6.3.2 Mode-n Product

In multilinear algebra, the analysis of the matrices can be done by manipulating the mode spaces via linear transformation, which is referred to as the mode-\(n\) product. The product of a tensor \(\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}\) and a matrix \(M \in \mathbb{R}^{J \times I_n}\) is written as a multiplication with subscript: \(\mathcal{A} \times_n M\). Let tensor \(\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N}\) whose entries are computed by multiplying a tensor and a matrix be:

\[
(\mathcal{A} \times_n M)_{i_1 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} = \sum_{j} a_{i_1 \ldots i_{n-1} j i_{n+1} \ldots i_N} m_{j_n i_n}
\]  

(6.16)

The notation shows a linear transformation of the vectors in the tensor mode-\(n\) space by the matrix \(M\). The mode-\(n\) product is expressed in tensor notation as:

\[
\mathcal{B} = \mathcal{A} \times_n M
\]  

(6.17)

or equivalently it can be written using flattened matrices:

\[
B_n = MA_n
\]  

(6.18)

The mode-\(n\) product has the following properties:

1. given a tensor \(\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}\) and two matrices, \(U \in \mathbb{R}^{J_m \times I_m}\) and \(V \in \mathbb{R}^{J_n \times I_n}\) \((m \neq n)\)

\[
\mathcal{A} \times_m U \times_n V = (\mathcal{A} \times_m U) \times_n V = (\mathcal{A} \times_n V) \times_m U = \mathcal{A} \times_n V \times_m U
\]  

(6.19)

2. given a tensor \(\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}\) and two matrices, \(U \in \mathbb{R}^{J_m \times I_m}\) and \(V \in \mathbb{R}^{K_n \times J_n}\)

\[
\mathcal{A} \times_n U \times_n V = \mathcal{A} \times_n (VU)
\]  

(6.20)

The products of matrices are computed using either the Kronecker, Khatri-Rao or Hadamard products. Given matrix \(A \in \mathbb{R}^{I \times J}\) and matrix \(B \in \mathbb{R}^{K \times L}\), the Kronecker product is denoted as \(A \otimes B\), where the result would be a matrix of size \((IK) \times (JL)\) and is as shown as follows:
A \otimes B = \begin{bmatrix} a_{11} B & a_{12} B & \ldots & a_{1J} B \\ a_{21} B & a_{22} B & \ldots & a_{2J} B \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1} B & a_{I2} B & \ldots & a_{IJ} B \end{bmatrix} = \begin{bmatrix} a_1 \otimes b_1 & a_1 \otimes b_2 & a_1 \otimes b_3 & \ldots & a_J \otimes b_{L-1} & a_J \otimes b_L \end{bmatrix} \quad (6.21)

While given matrix $A \in \mathbb{R}^{I \times K}$ and matrix $B \in \mathbb{R}^{J \times K}$, the Khatri-Rao product is a columnwise matrix product and is denoted by $A \odot B$. The resulting matrix of size $(IJ) \times K$ is:

$$A \odot B = [a_1 \otimes b_1 \ a_2 \otimes b_2 \ a_3 \otimes b_3 \ldots a_K \otimes b_K] \quad (6.22)$$

The Hadamard product is the element wise matrix product. Given two matrices, $A$ and $B$ of the same size $(I \times J)$, the Hadamard product is:

$$A * B = \begin{bmatrix} a_{11} b_{11} & a_{12} b_{12} & \ldots & a_{1J} b_{1J} \\ a_{21} b_{21} & a_{22} b_{12} & \ldots & a_{2J} b_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ a_{IJ} b_{11} & a_{IJ} b_{12} & \ldots & a_{IJ} b_{IJ} \end{bmatrix} \quad (6.23)$$

More detailed properties of these matrix products can be found in [267, 245, 175]. For computing the product of a tensor and matrix, the Kronecker product is used. Equation 6.17 can be expanded for all $n \in \{1, \ldots, N\}$ to:

$$B = \mathcal{A} \times_1 M^1 \times_2 M^2 \ldots \times_N M^N \quad (6.24)$$

$$B_n = A_n M_n (A^N \otimes \ldots \otimes A^{n+1} \otimes A^{n-1} \ldots \times A^1)^T \quad (6.25)$$

6.3.3 Tensor Decompositions

Although the work on tensor decompositions and models has been available for over four decades, very little application of it has been published. Tensor decompositions started back
in 1927 by Hitchcock [145]. In the 1960s, the tensor concept was revived by Tucker [287], followed in the 1970s by Carroll and Change [51] who introduced the Canonical decomposition (CANDECOMP) model and Harshman [137] who, proposed the Parallel factors (PARAFAC) tensor decomposition. A popular tensor decomposition used is the TUCKER tensor model proposed by Tucker [287].

The TUCKER decomposition is a form of higher-order principal component analysis. It is modeled as follows, given an \( I \times J \times K \) tensor \( \mathcal{A} \):

\[
\mathcal{A} = D \times_1 U \times_2 V \times_3 W \tag{6.26}
\]

\[
\mathcal{A} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \sigma_{ijk} (u_i \circ v_j \circ w_k) \tag{6.27}
\]

where \( 1 \leq i \leq I, 1 \leq j \leq J, 1 \leq k \leq K \). Each of the \( U, V, W \) factor matrices can be thought of as the principal components in each mode. The tensor, \( D = (\sigma_{ijk}) \), contains the eigenvalues and it is called the core tensor. The elementwise of the TUCKER decomposition is written as:

\[
a_{ijk} = \sum_{i=1}^{R_1} \sum_{j=1}^{R_2} \sum_{k=1}^{R_3} \sigma_{ijk} u_i v_j w_k \tag{6.28}
\]

for \( i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K \).

In a mode-3 tensor, the flattening matrices for each mode (based on Equation 6.27) are:

\[
A_1 = UD_1(V \otimes W)^T \tag{6.29}
\]

\[
A_2 = VD_2(U \otimes W)^T \tag{6.30}
\]

\[
A_3 = WD_3(U \otimes V)^T \tag{6.31}
\]

The original Tucker method was presented only for a three-mode case [287]. However, it can easily be extended to an \( N \) dimension case. The generalised \( N \)-mode tensor is decomposed as:
the element wise equation of the decomposition is:

$$a_{i_1i_2...i_N} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \ldots \sum_{r_N=1}^{R_N} \sigma_{r_1r_2...r_N} a_{i_1r_1}^1 a_{i_2r_2}^2 \ldots a_{i_Nr_N}^N$$  \hspace{1cm} (6.33)$$

where $i_n = 1, \ldots, I_n, n = 1, \ldots, N$. Hence, the flattened version of Equation 6.32 is:

$$A_n = U_n D_n (U_n^T \otimes \ldots \otimes U_n^{n+1} \otimes U_n^{n-1} \otimes \ldots \otimes U_1^T)^T$$  \hspace{1cm} (6.34)$$

In an $N$-mode tensor model, the core tensor is not necessarily of the same dimension as the tensor $\mathcal{A}$, and it is not unique. The $N$-mode TUCKER tensor decomposition is also known as the higher-order SVD (HOSVD) from the work of De Lathauwer, De Moor and Vandewalle [184].

In Equation 6.34, $U_n$ is the left singular values of the SVD.

Figure 6.4 illustrates an $N$-mode SVD for the case of $N = 3$. Unlike the two dimensional SVD, the core tensor $\mathcal{D}$ does not have a simple diagonal structure but it contains a set of the basis matrices for all the mode matrices.

Figure 6.4: Visualisation of a three-mode SVD array

The core tensor can be expressed in the form of:

$$\mathcal{D} = \mathcal{A} \times_1 U_1^T \times_2 U_2^T \times_3 \ldots \times_N U_N^T$$  \hspace{1cm} (6.35)$$
By applying the flattening matrices (as shown in Figure 6.3), the generalised $N$-mode SVD can be rewritten as follows:

$$A_n \xrightarrow{\text{data}} U_n \xrightarrow{\text{basis vectors}} D_n (U^{m-1} \otimes ... \otimes U^1 \otimes U^N \otimes ... \otimes U^{m+2} U^{m+1})^T \xrightarrow{\text{coefficients}}$$ (6.36)

Recall that $\otimes$ denotes the Kronecker product defined in Equation 6.21. Equation 6.36 can be interpreted as a standard linear decomposition of the data ensemble. The multi-factored decomposition shown in Equation 6.36 is a generative Higher-order Singular Value Decomposition (HOSVD) model where the mode subspaces with their corresponding core tensor can be obtained.

HOSVD is a generalisation of SVD [175]. The computation of the left singular vectors of $A_n$ (that is the computation of the SVD on each flattened matrix) is efficient, making the HOSVD a preferred used tensor decomposition. Given that the left singular matrices are orthogonal, the HOSVD core tensor is orthogonal in each dimension. The HOSVD algorithm is as below:

**Listing 1** HOSVD algorithm  
1: Apply matrix flattening on the tensor matrix $A_{(n)}$  
2: for $n = 1, \ldots, N$ do  
3: Compute the left singular value of matrix $U_n$  
4: end for  
5: Calculate the core tensor $D$ based on $D = A \times_1 U_1^T \times_2 U_2^T \times_3 \ldots \times_N U_N^T$

There are a number of other tensor decomposition techniques. Most of them are related to the CP and TUCKER models. They include the Individual Differences in Scaling method (INDSCAL) [51], Parallel Factors for Cross Products (PARAFAC2) [138], CANDECOMP with Linear Constraints (CANDELINC) [52], Decomposition into Directional Components (DEDICOM) [139], PARAFAC and Tucker2 (PARATUCK2) [140], PARALIND [44], [4] etc. A brief review of each of these tensor decomposition can be found in [175]. Due to the popularity of the tensor framework, the next section will further explain the proposed novel N-mode SVD tensor decomposition, following the concept of the tensor model, which is used in this research.

### 6.3.4 An example of HOSVD computation

Define a $(2 \times 3 \times 3)$ tensor shown in figure 6.5 with the following values $A$:
6.3 The Fundamental Concepts of Multilinear Algebra

The tensor can be flattened in three ways. Matrix unfolding $A_1$ is equal to

$$
\begin{pmatrix}
1 & 1 & 2 & 1 & -1 & 0 & 1 & -2 & -1 \\
1 & 4 & 2 & 2 & -2 & 2 & 1 & -5 & 2
\end{pmatrix}_{2\times9matrix}
$$

Matrix unfolding $A_2$ is

$$
\begin{pmatrix}
1 & 4 & 1 & 4 & 2 & 2 \\
1 & 2 & -1 & -2 & 0 & 2 \\
1 & 1 & -2 & -5 & -1 & 2
\end{pmatrix}_{3\times6matrix}
$$

And matrix unfolding $A_3$ equals to

$$
\begin{pmatrix}
1 & 1 & 1 & 4 & 2 & 1 \\
1 & -1 & -2 & 4 & -2 & -5 \\
2 & 0 & -1 & 2 & 2 & 2
\end{pmatrix}_{3\times6matrix}
$$

Computing $U_1$, the left singular matrix of $A_1$, we get:

$$
\begin{pmatrix}
0.3256 & 0.9455 \\
0.9455 & -0.3256
\end{pmatrix}
$$
In the same way, we calculate $U_2$ and $U_3$ as:

$$U_2 = \begin{pmatrix}
-0.6564 & 0.7055 & -0.2673 \\
0.2546 & 0.5406 & 0.8018 \\
0.7102 & 0.4582 & -0.5345
\end{pmatrix}$$

$$U_3 = \begin{pmatrix}
0.1893 & -0.7572 & 0.6252 \\
0.9816 & 0.1622 & -0.1008 \\
0.0251 & -0.6327 & -0.7740
\end{pmatrix}$$

The flattened core tensor can be computed using:

$$D_1 = U_1^T \times A_1 \times (U_2 \otimes U_3)$$

$$D_2 = U_2^T \times A_2 \times (U_3 \otimes U_1)$$

$$D_3 = U_3^T \times A_3 \times (U_1 \otimes U_2)$$

Hence,

$$D_1 = \begin{pmatrix}
0.3256 & 0.9455 \\
0.9455 & -0.3256
\end{pmatrix} \times \begin{pmatrix}
1 & 1 & 2 & 1 & -1 & 0 & 1 & -2 & -1 \\
4 & 2 & 2 & -2 & 2 & 1 & -5 & 2
\end{pmatrix} \times$$

$$\begin{pmatrix}
-0.124 & 0.497 & -0.410 & 0.134 & -0.534 & 0.441 & -0.051 & 0.202 & -0.167 \\
-0.644 & -0.107 & 0.066 & 0.693 & 0.114 & -0.071 & -0.262 & -0.043 & 0.027 \\
-0.017 & 0.415 & 0.508 & 0.018 & -0.446 & -0.546 & -0.007 & 0.169 & 0.207 \\
0.048 & -0.193 & 0.159 & 0.102 & -0.409 & 0.338 & 0.152 & -0.607 & 0.501 \\
0.250 & 0.041 & -0.026 & 0.531 & 0.088 & -0.055 & 0.787 & 0.130 & -0.0801 \\
0.006 & -0.161 & -0.197 & 0.014 & -0.342 & -0.418 & 0.020 & -0.507 & -0.621 \\
0.134 & -0.538 & 0.444 & 0.087 & -0.347 & 0.287 & -0.101 & 0.405 & -0.334 \\
0.697 & 0.115 & -0.071 & 0.450 & 0.074 & -0.046 & -0.525 & -0.087 & 0.054 \\
0.018 & -0.449 & -0.550 & 0.012 & -0.290 & -0.355 & -0.013 & 0.332 & 0.414
\end{pmatrix}$$
6.4 Related research on tensor decompositions

Applications of tensor decomposition have ranged from chemical sciences [11] to computer vision and visualisation [294, 296, 297]. The majority of the applications used CANDECOMP/PARAFAC (CP) or TUCKER decompositions which are higher order generalisation of singular value decomposition and principal component analysis.

The CP decomposition was first introduced in the field of psychometrics in 1970 by combining the work of Carroll and Chang [51] and Harshman [137]. Carroll and Chang [51] proposed CANDECOMP for analysing multiple similarity and dissimilarity from a set of auditory tones from Bell Labs to other auditory tones from other countries. Harshman [137] introduced PARAFAC with an aim of eliminating the ambiguity associated with two-dimensional PCA. He applied it to vowel-sound data where the data is organised into three modes - mode-1 is the different individuals, mode-2 is the different vowels and mode-3 is the measured format.

Andersen and Bro [9] surveyed the use of CP in chemometrics and they have shown that the CP model is particularly useful in the work of modelling fluorescence excitation-emission data. Appellof and Davidson [11] pioneered the use of CP in chemometrics. In neuroscience, Möcks [216] discovered a unique and effective way to choose the number of factors when using CP decomposition in brain imaging applications. Field and Graupe [105] compared the work of Möcks with other work on CP in brain imaging, discussing its practicality and utility for event-related potentials.

Other applications that employed CP decomposition for image compression and classification are Shahsua and Levin [262] for recognising objects in images, Furukawa et al. [111] in building a compressed image texture database, Bauckhage [23] for recognising coloured objects.
and Acar et al. [2, 3] applied it in communication data in online chatrooms.

The second most cited tensor decomposition is the TUCKER decomposition [175]. Henrion [143] used TUCKER decomposition for N-way PCA in chemical analysis. Kiers and Mechelen [168] have employed a three-way TUCKER model in psychometrics by comparing the results with two-way techniques. The three-way TUCKER decomposition is sometimes called a trilinear model and the two-way decomposition is known as bilinear decomposition. Tenenbaum and Freeman [276] have applied bilinear models to analyse a two factors task. According to the authors, a bilinear model can provide a richer description of interactions by allowing factors to modulate each other’s contributions multiplicatively in comparison to using PCA alone. Another work that employed bilinear decomposition is by Magnus and Neudecker [202] for analysing statistical matrices. Marimont and Wandell [203] applied a bilinear model to characterise two independent factors. In 2000, Tenenbaum and Freeman [276] continued the work of Magnus and Neudecker [202] by applying a bilinear model to face image recognition. The performance of bilinear models has also shown higher recognition results than those using linear principal components model alone [276]. Kiers and Van Mechelen explained the usefulness of the three-way technique compared to employing a bilinear decomposition method.

Many applications have more than three factors, hence an $N$-way array TUCKER decomposition is required. De Lathauwer and Vandewalle [184] applied TUCKER decomposition to signal processing. Nagy and Kilmer [220] proposed the use of the TUCKER model to construct Kronecker product approximations for preconditioners in image processing. Tucker’s work was also used in text analysis by Liu et al. [195] and Sun et al. [270, 271].

In face recognition applications, the use of tensor decomposition is rare, and most face recognition systems employ PCA. Considering that faces have facial variations, the conventional eigenfaces technique is not ideal. Vasilescu and Terzopoulos [294] pioneered the use of an $N$-dimension TUCKER decomposition which they called tensorfaces. Their concept of the tensorfaces model is similar to the multilinear model as explained in Section 6.3. Vasilescu and Terzopoulos proposed an $N$-mode SVD face image representation model and decomposed the $N$-dimensional tensor using Equation 6.32. They structured the data using the selected factors in one $N$-mode tensor. Their recognition performance using tensor data is significantly better.
6.4 Related research on tensor decompositions

than the standard eigenfaces technique [294].

Vasilescu also applied tensorfaces for compression and removing of irrelevant effects [294], and analysing human motion [292]. Gralewski et al. [125] applied TUCKER decomposition to facial emotion signature analysis from a sequence of video image footage. Wang and Ahuja [297] applied the TUCKER method to generate facial expressions while Vlasic et al. [296] used it to transfer faces with expressions and visemes (speech-related mouth articulations) from a video. They managed the face variation factors in bilinear and trilinear TUCKER models. Figure 6.6 shows that the face scans are arranged appropriately according to facial expressions, identity and pixels.

![Figure 6.6: An example of a data tensor model (image taken from [296])](image)

Based on the work of Vlasic et al. [296], the TUCKER model can be used to estimate and manipulate the face attributes that are not available in the model. From the listed attribute parameters (such as facial expressions and visemes), the system can then change the appearance of the face. The advantage of this overall model is that even with missing face data, the TUCKER model can still be built based on the imputation\(^1\) process.

A continuation of Vlasic et al. [296] was carried out by Macedo et al. [200] who adopted a similar idea utilising photograph data. They extracted and analysed the expressions from a photograph of a given subject and transferred the estimated expression parameters onto another person’s face. Their work combined both multilinear analysis and active appearance models (AAM) as describes by Cootes [70].

\(^1\)Imputation or estimate is a method to find the missing data based on the other attributes or parameters
Along similar lines, Wang and Ahuja [297] used TUCKER decomposition to analyse 2D facial expressions and synthesize the facial expression of a person from the resulting subspaces. Firstly, they applied PCA onto the mode then HOSVD to only deal with the reduced dimension 3rd-order tensor as follows:

\[ \mathcal{A} = D \times_1 U_{\text{subject}} \times_2 U_{\text{expression}} \times_3 U_{\text{feature}} \]

where the \( U_{\text{subject}} \) is denoted as the subject’s subspace, \( U_{\text{expression}} \) as the facial expression subspace and \( U_{\text{feature}} \) is the facial feature subspace. The face and facial expression recognition accuracy obtained when employing HOSVD was high. They also use HOSVD for learning the attributes of facial expressions to enable them to synthesize facial expressions.

From the above, we see that the tensor framework is effective especially when the application involves higher order tensors or multiple factors. A face, for example, contains multiple factors such as facial features, facial expressions, textures, visemes, identity, age and many more. Currently, very few face recognition researchers deploy tensor decomposition models. The ones that were discussed either used HOSVD alone or a combination of other tensor models, and they focussed mainly on synthesising facial expressions. The limitation of the existing tensor models is that they require a full data tensor to work efficiently. A full data tensor has no missing data in any subspace. Hence, we require the introduction of the sub-tensor method.

### 6.5 Experimental Method

We adopted tensor framework explicitly accounting for the subject factor and the facial expression factor. Firstly, we define a 3\textsuperscript{rd} order tensor \( \mathcal{A} \in \mathbb{R}^{\text{sub} \times \text{exp} \times \text{pts}} \) which includes an entire training set of three dimensional face surface data. The training faces are arranged by hand into a tensor \( \mathcal{A} \) which is a \( I_{\text{sub}} \times I_{\text{exp}} \times I_{\text{pts}} \) tensor where the \( I_{\text{sub}} \) is the number of persons or subjects, \( I_{\text{exp}} \) is the number of expressions and \( I_{\text{pts}} \) is the number of landmark points. The multilinear decomposition of \( \mathcal{A} \) results in:

\[ \mathcal{A} = D \times_1 U_{\text{sub}} \times_2 U_{\text{exp}} \times_3 U_{\text{pts}} \] (6.37)
The core tensor $D$ manages the interaction between the indices of the 3-mode matrices. The $U_{sub}$ is the mode-1 matrix spanning the subject vector space. The $U_{exp}$ is the mode-2 matrix spanning the expression vector space and the $U_{pts}$ is the mode-3 matrix spanning the landmark points space. The core tensor $D$ has the same dimension as tensor $A$. The visualisation of the 3rd order tensor $A$ is as illustrated in Figure 6.7.

Figure 6.7: Visualisation of the decomposition of tensor $A$ into three subspace matrices and a core tensor $D$

The flattened 3-mode matrices are $A_{sub}$, $A_{exp}$ and $A_{pts}$. Mode-1 vector $A_{sub}$ comprises of a $(I_{sub} \times I_{exp} \times I_{pts})$, mode-2 vector $A_{exp}$ comprises of $(I_{exp} \times I_{sub} \times I_{pts})$ and mode-3 vector $A_{pts}$ as $(I_{pts} \times I_{sub} \times I_{exp})$. The left singular value matrices are obtained by computing the SVD on each flattening of the $A$ matrix to obtain $U_{sub}$, $U_{exp}$ and $U_{pts}$. The matrix $U_{sub}$ spans the $(I_{sub} \times I_{sub})$-mode of the expression space, matrix $U_{exp}$ spans the $(I_{exp} \times I_{exp})$-mode and matrix $U_{pts}$ spans $(I_{pts} \times (I_{sub} \times I_{exp}))$-mode.

When SVD is performed on mode-3 of the flattened data tensor, $D$, each column of the obtained matrix $U_{pts}$ is identical to the conventional eigenfaces [288, 266]. The difference between multilinear analysis and PCA is that the tensor represents how the indices interact with one another as a whole, while the PCA basis vectors represent only the principal axes of variation across images. Suppose a face database consists of $I_{sub}$, $I_{exp}$ and $I_{pts}$. With PCA, the length of each coefficient vector is $I_{sub} \times I_{exp}$ where each subject has a set of $I_{exp}$ coefficient vectors. In the multilinear analysis, each person is represented with a single vector coefficient.
of dimension $I_{sub}$ relative to the bases tensor of $I_{sub} \times I_{exp} \times I_{pts}$.

Since the left singular value matrices are of different sizes, we use the Kronecker product to compute the product of the matrices. The Kronecker computation is given in Equation 6.21. The product for a mode-1 matrix is $(U_{exp}, U_{lmk})$, mode-2 matrix is $(U_{sub}, U_{lmk})$ and mode-3 matrix is $(U_{sub}, U_{exp})$. The core tensor, $D$, is the SVD of matrices $A_{sub}, A_{exp}$ and $A_{pts}$ following Equation 6.38.

$$D = A \times_1 U_{sub}^T \times_2 U_{exp}^T \times_3 U_{pts}^T$$

The element wise form is:

$$x_i = \sum_{r_1=1}^{I_{sub}} \sum_{r_2=1}^{I_{exp}} \sum_{r_3=1}^{I_{pts}} \sigma_{r_1 r_2 r_3} a_{i r_1}^{sub} a_{i r_2}^{exp} a_{i r_3}^{pts}$$

where $i = 1, \ldots, 3$ and $\sigma$ is the corresponding singular value lying on the diagonals of $D$. The mode matrices are:

$$D_{sub} = U_{sub}^T \cdot A_{sub} \cdot (U_{exp} \otimes U_{pts})$$

$$D_{exp} = U_{exp}^T \cdot A_{exp} \cdot (U_{pts} \otimes U_{sub})$$

$$D_{pts} = U_{pts}^T \cdot A_{pts} \cdot (U_{sub} \otimes U_{exp})$$

When implementing multilinear SVD in real applications, the computed core tensor is large and can get larger particularly when more variability factors are added [212]. This means that significant amounts of memory and high computation power are needed to generate the core tensor. In order to resolve these problems, the multilinear sub-tensor within the tensor model is proposed. The sub-tensor SVD is constructed based on the variation subspace approach according to:

$$B = A \times_2 U_{exp}^T \times_3 U_{pts}^T$$

Here, the mode-2 vector of tensor $B$ contains expression data and the mode-3 vector is the landmark points. The mode-1 vector of $A$ has mode-2 index $i$ and mode-3 index $j$: $x_{ij}$. 
According to the formation of $\mathcal{A}$, $x_{ij}$ is also the sample vector in $i$-th expression from $j$-th subject. The element wise form is as follows:

$$x_{ij} = \sum_{r_1=1}^{I_{exp}} \sum_{r_2=1}^{I_{pts}} \sigma_{r_1 r_2} a_{i r_1}^\text{exp} a_{j r_2}^\text{pts}$$

(6.44)

Firstly, the tensor $\mathcal{A}$ is flattened along the expression mode to extract $I_{exp}$ number of sub-space matrices. The resulting sub-tensor $\mathcal{B}$ has a dimension of $I_{exp} \times (I_{sub} \times I_{pts})$. Instead of having a global tensor model, we aim to construct bilinear sub-tensors as they are easier to manipulate and require less computation and memory requirements.

From Equation 6.44, we observe that the $i$-th row vector of $U_{\text{exp}}$ is $v_i^\text{expT}$ and $j$-th row vector of $U_{\text{pts}}$ is $v_j^\text{ptsT}$ and hence it is denoted as:

$$x_{ij} = \mathcal{B} \times_2 v_i^\text{expT} \times_3 v_j^\text{ptsT}$$

(6.45)

Next, the sub-tensor $\mathcal{B}$ is obtained by indexing the basis tensor for a particular expression to a dimension $I_{sub} \times 1 \times I_{pts}$. The sub-tensor $\mathcal{B}$ is flattened along the subject mode to obtain matrix $B$ with a dimension of $I_{sub} \times I_{pts}$. For every expression, there are $I_{exp}$ coefficient vectors. The coefficient vector as in Equation 6.36 is similar to PCA and is denoted as $c_{exp} = B^T x_{ij}$ where $x_{ij}$ is a specific training face of the $j$-th subject with the $i$-th expression.

To recognise a person, we apply a linear projection approach. The projection is computed on each sub-tensor to yield a set of coefficient vectors. Given an unknown face surface, $x$, we project $x$ into the reduced dimensional space to get the coefficient vector, $c_q$, of the unknown subject. We then apply a Euclidean distance scheme for recognition. The best matched coefficient vector yields the smallest value of $||c_{exp} - c_q||$ among all the variant factors.

### 6.6 Results and Discussion

The Binghamton Faces database [316], made up of three-dimensional face surfaces with a larger variation of facial expressions, was used in our experiments. There are 100 subjects with seven different types of facial expressions, namely: neutral, anger, disgust, fear, happiness, sadness and surprise. Six of the emotional expressions contain 4 levels of intensity named low, middle,
high and highest. From the database, we randomly selected a total of 1350 (= 54 subjects \times 25 facial expressions) and implemented pre-processing on the raw face models. Each raw facial mesh has between 18,000 and 23,000 points. After the pre-processing stage, each of the cleaned face surfaces had 5090 points. Figure 6.8 shows samples of faces with the different facial expressions that have been pre-processed. Each of the face models contain both shape and texture information. In this thesis, we only utilise the shape information.

![Figure 6.8: Samples of pre-processed face surfaces](image)

We performed the experiments with several different combinations of the facial expressions. In each experiment, we divide the faces into two groups - the training group and the query group. In the first experiment, we select all the faces with some level of expression as the training group and all the neutral faces as the query group. This first scheme of experiment is called the neutral scheme 1. Here, we attempt to recognise neutral faces that are not in the training set. In the second experiment, called the leave-one-out scheme 2, one of the four faces with expression for each subject is selected as the query face and the remaining faces with expression as the training group. The query face is classified as a particular individual if the distance to any of the other three faces of that individual is the smallest. The third experiment used a similar leave-one-out scheme to the second experiment except that two faces with expression served as training data and one of the remaining two faces for each subject was used as a query face. We call this experiment the leave-one-out scheme 3. The fourth experiment, called the leave-one-
out scheme 4, works by training using only one of the faces with expression and uses one of the remaining faces with expression as the query face. The aim of the last three leave-one-out experiments is to recognise a subject presenting a facial expression that is not available in the training set. In addition, we analysed the facial expression recognised by the system based on the faces available only in the training set. The experiments (except the fourth experiment) used the sub-tensor approach. The results are then compared with the conventional eigenface approach. The fifth experiment was set up to study the use of a set of landmarks as input data for face recognition. This experiment also employs a leave-one-out scheme. The recognition results are compared with results using all surface points, \( I_{pts} \), using PCA. The sixth and final experiment compares the recognition results when varying the number of subjects used. The aim of these two last experiments is to study the influence of face recognition with differing numbers of landmarks and subjects.

The tensor model is formed by arranging the training faces as an \( I_{sub} \times I_{exp} \times I_{pts} \) tensor where the \( I_{sub} \) is the 54 subjects, \( I_{exp} \) is the 25 (= 1 neutral, 6 expressions \( \times 4 \) intensity levels) expressions and \( I_{pts} \) is the 5090 number of (x-, y-, z-) surface points. Based on the training sets used, the tensor model has more than 19 million points, which requires high computation power and memory space. Hence, the multilinear sub-tensor defined in Equation 6.43 is constructed instead. The sub-tensor model has a dimension of \( I_{exp} \times (I_{sub} \times I_{pts}) \) where we set \( I_{exp} = 4 \) (the maximum number of expression levels) and \( (I_{sub} \times I_{pts}) = (54 \times 5090 \times 3) \). Figure 6.9 illustrates the sub-tensor models.

The sub-tensor transforms the eigenfaces in the matrix \( U_{pts} \) into eigenmodes, which represent the principal axes of variation across the subject mode for each of the 6 expressions. By contrast, PCA basis vectors represent only the principal axes of variation across all faces. Figure 6.10 shows some of the PCA eigenfaces. In any column, the first eigenvector depicts the average subject and the remaining eigenvectors capture the variability over subject and facial expressions.

The eigenfaces are the same as those produced by performing PCA on the flattened matrix mode-\( I_{pts} \) [266, 288]. The multilinear analysis creates well separated subject classes by maximizing the ratio of inter-class scatter to intra-class scatter [24, 293]. The sub-tensor eigenfaces...
Figure 6.9: Each emotional expression is managed into slices. The concatenation of the slices forms the sub-tensor model with a dimension of $6 \times (54 \times 5090 \times 3)$, where 3 is the x,y,z-coordinates.

Figure 6.10: The eigenfaces taken from three rows of $-3\lambda$ to $+3\lambda$ eigenmodes.
are shown in Figure 6.11. From the first experiment, Figure 6.12 shows the rank 1 rates of the neutral scheme 1. The experiment was done on 54 subjects, 40 subjects and 30 subjects.

Figure 6.11: A partial visualisation of sub-tensor for a disgust expression (from $-3\lambda$ to $+3\lambda$ eigenmodes), which makes the eigenvectors person-specific.

Figure 6.12: The rank 1 recognition rate of recognising neutral faces in a pool of all expressioned faces

This figure compares the face recognition results of the sub-tensor SVD analysis with the
PCA method, in which the number of eigenvectors are kept between 10 to 50. It can be seen that the overall recognition rates using the sub-tensor model is slightly better than the results obtained using the PCA alone. It is important to note that the total number of sub-tensor eigenvectors is \((I_{\text{sub}} - 1)\) while the PCA has a total of \((I_{\text{sub}} \times I_{\text{exp}})\) eigenvectors. Hence, with 30 subjects, the sub-tensor model feature vector is of length 29. This means that for 30 subjects, we are limited to 29 eigenvectors and for 40 subjects, we are limited to 39 eigenvectors.

Figure 6.13 shows the face recognition results of the leave-one-out scheme 2 (experiment 2). Each graph represents the recognition of a subject within a specific expression space. The displayed recognition rates are rank 1 performance and demonstrate some small gain in accuracy (about 6-7% improvement on average) when using the sub-tensor SVD approach.

Throughout the first two experiments, it can be seen that the recognition results using PCA are constant. This is due to very small shape variance in the third intensity level of the facial expressions. Figure 6.14 shows the eigenvalues due to shape in facial expressions with all of the intensity levels except level 3, using PCA. It shows that the first 10 eigenvectors of individual expressions capture more than 90% of the variance.

As for level 3 intensity of any facial expression, the shape variance for the first mode is as high as 99.2% (see Figure 6.15). Similarly, when training faces with any level of expression intensity together with the level 3 intensity faces, the generated recognition results remain the same for all eigenvectors. This accounts for the results displayed in Figure 6.13.

Figure 6.17 illustrates the recognition results for all expressions excluding level 3 expression intensity. The displayed results are taken from eigenmodes 10 to 50. The receiver operating characteristic (ROC) graphs (probability of verification against false acceptance rate) are shown in Figure 6.18. The ROC curves are plotted using different distance thresholds for identifying subjects from a database. As the distance is increased, we are more likely to correctly identify the subject but we also increase the risk of false identification. It can be seen that the sub-tensor SVD performed better than the PCA on a number of facial expressions.

Figure 6.16 shows the shape variance for any combination of two facial expression intensities except level 3. It takes 20 eigenvectors to capture more than 95% of the shape variance. The recognition results of the leave-one-out scheme 3 experiment are shown in Figure 6.19.
6.6 Results and Discussion

Figure 6.13: The recognition results gathered from using the sub-tensor SVD and the PCA under varying facial expressions
Figure 6.14: Energy captured by retaining the number of eigenvectors with the largest eigenvalues of 54 subjects.

Figure 6.15: Energy captured by retaining the number of eigenvectors with the largest eigenvalues of 54 subjects with an expression. Interestingly, the first eigenvector captures more than 90% of the shape variations.
6.6 Results and Discussion

Figure 6.16: Energy captured by retaining the number of eigenvectors with the largest eigenvalues of 54 subjects with 2 facial expressions.

Figure 6.17: The recognition results gathered from using the sub-tensor SVD and the PCA under varying facial expressions without intensity level 3 expressions.
Figure 6.18: ROC graph for (a) angry expressions, (b) disgust expressions, (c) fear expressions, (d) happy expressions, (e) sad expressions, (f) surprise expressions. The numbers shown on the graphs are the normalised thresholds. The probability of correctly recognising a subject and falsely accepting a subject is measured based on several leave-one-out experiments. The correct operating threshold depends on the application.
this experiment, we used level 2 and level 4 expression intensities as the training set and level 1 as the query set. The presented recognition results are the average rates of six expressions. Both PCA and sub-tensor SVD gave quite similar recognition rates. The resulting performance may be due to the high differences between high intensity expressions and low intensity expressions which then gives a comparatively similar recognition performance.

![Figure 6.19: Results of recognition using PCA](image)

Figure 6.19: Results of recognition using PCA

Figure 6.20 shows the leave-one-out scheme 4 experiment. We used the level 1, 3 and 4 expression intensities in the query datasets. The results clearly show that recognising faces with low intensity expressions (level 1) produces higher rates than higher intensity expressions. The high recognition rate is because the level 1 expressions are more similar to the level 2 expressions used in the training set. The level 4 expressions have more significant differences such as the opening of the mouth.

Table 6.21 illustrates the comparison between recognition results of two methods: PCA with 10 eigenvectors for each subject and sub-tensor SVD with 10 eigenvectors. The results are the average values from all the 6 facial expressions. It can be seen that there is an improvement in the recognition performance using sub-tensor SVD.

Table 6.21 illustrates the comparison between recognition results of two methods: PCA with 10 eigenvectors for each subject and sub-tensor SVD with 10 eigenvectors. The results are the average values from all the 6 facial expressions. It can be seen that there is an improvement in the recognition performance using sub-tensor SVD.

From the comparison between the PCA and sub-tensor SVD methods in face recognition, we conclude that the tensor framework provides better discriminant power and is more effective in recognising a subject with the correct expression. Even though PCA may give high recognition rates for individual subjects, the recognition of a subject with the correct expressions is slightly lower in comparison with the sub-tensor method. This approach is particularly useful
6.6 Results and Discussion

Figure 6.20: Results of recognition using PCA

Figure 6.21: Comparative recognition results between PCA and sub-tensor SVD
for recognising higher intensity expressions that are not available in the training dataset. Table 6.1 compares the results of recognising subjects with level 4 expressions using PCA and sub-tensor SVD methods. The first two columns compare the recognition results using PCA and the last two columns show the recognition results using the sub-tensor method. Columns 2 and 4 show the results of recognising subjects with the same facial expressions. The results of the sub-tensor approach also shows that in recognising the correct expression of subjects, it outperforms the PCA approach.

<table>
<thead>
<tr>
<th>Level 4 expressions</th>
<th>PCA Identity</th>
<th>Expression</th>
<th>Sub-tensor SVD Identity</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>85%</td>
<td>61%</td>
<td>85%</td>
<td>61%</td>
</tr>
<tr>
<td>Disgust</td>
<td>79%</td>
<td>66%</td>
<td>83%</td>
<td>74%</td>
</tr>
<tr>
<td>Fear</td>
<td>90%</td>
<td>72%</td>
<td>90%</td>
<td>78%</td>
</tr>
<tr>
<td>Happy</td>
<td>77%</td>
<td>75%</td>
<td>79%</td>
<td>78%</td>
</tr>
<tr>
<td>Sad</td>
<td>92%</td>
<td>85%</td>
<td>94%</td>
<td>85%</td>
</tr>
<tr>
<td>Surprise</td>
<td>44%</td>
<td>42%</td>
<td>46%</td>
<td>44%</td>
</tr>
</tbody>
</table>

Bootstrapping was applied to estimate the variance of the above results. In order to do so, firstly a number of bootstrap datasets were created by downsampling the original sample to 44 subjects in different ways. We performed the same leave-one-out recognition accuracy test as described above for angry expressions plotted against the number of eigenmodes used. Finally, we compute the mean, the standard deviation and standard error of the accuracy estimated from each bootstrap dataset.

Figure 6.22 shows the result of the variance of the PCA and sub-tensor SVD distributions. The error bars shows the standard deviations. We have also calculated the standard error of the mean (SE). If the sample is small, the SE and the standard deviation tend to be underestimates. In our experiment, since the sample size is > 6, and this would mean that the underestimate would be less than 5% in a normally distributed sample [132]. The calculated sample mean and the SE are printed in Figure 6.22. Based on the sample mean and the SE, approximate confidence intervals for the mean of upper and lower 95% confidence limits can be calculated. Using 20 to 50 eigenmodes with (P ≤ 0.05), the differences are significant while with 10 eigenmode, it is not (P = 0.19).
Figure 6.22: Variance distribution of recognising faces

Figure 6.23 is similar to Figure 6.22 except it displays the results when the correct facial expression is recognised. Using 20 to 50 eigenmodes, the results are statistically significant for confidence limits of 95% with \( P \leq 0.05 \) while with 10 eigenmode, it is not.

Figure 6.23: Variance distribution of recognising facial expressions

There is an improvement in recognising subjects invariant of the expressions when using
sub-tensor SVD. On top of that, recognising both identity and expression also gave higher recognition rates. In both PCA and sub-tensor SVD, the performance of recognising subjects with surprise expressions is below the 50% recognition rates due to the extreme geometric change in the surprise expression which involves opening the mouth, lowering the jaw, widening the eyes and raising the eyebrows.

In experiment 5, we used four scans of twenty subjects from the Imperial College face database. In this experiment, we discarded one expression for testing and the remaining ones were used for training. Examples of the training face surfaces are shown in Figure 6.24. The recognition experiments also employed landmark points. These points are selected based on places that have high variation when changing facial expressions, as discussed in the previous chapter.

Figure 6.24: Some samples of 3D face database
This experiment is divided into two groups - Group A used the whole 5090 points on the surface and Group B used 33 selected landmark points (see Figure 6.25).

![Figure 6.25: Landmarking on 3D face surface](image)

For each group, three sub-experiments were conducted. Each sub-experiment differs in terms of the query face data. In sub-experiment 1, the unseen neutral face data is used as the query set while in the second sub-experiment, a frowning face is used as the query data and finally sub-experiment 3 uses smiling faces. Table 6.2 and Table 6.3 show the results of all the sub-experiments. Each recognition probe was tested with 2 and 19 eigenvectors. It is shown that the placement of landmarks can influence the recognition rate, given that the sub-tensor SVD yields better recognition results than using PCA alone.

The overall recognition results in all three experiments have shown an improvement in rate when using the sub-tensor SVD method. In sub-experiment 2, the recognition results for PCA remain low as we increased the number of eigenvalues, unlike the results from sub-tensor SVD. In PCA, the coefficient vector has a higher dimension. In the sub-tensor, the coefficient vector is smaller in dimension and is dependent on the subjects. Hence, recognising subjects invariant of facial expressions is easier when using sub-tensor. A similar argument can also be applied to sub-experiment 3. Interestingly, recognising subjects with frowning expressions gave higher recognition rates because the frowning expression has low intensity and it is similar to the neutral expression.
Table 6.2: Recognition rate results of Group A

<table>
<thead>
<tr>
<th>Eigenvectors</th>
<th>Sub-exp 1: PCA (%)</th>
<th>Sub-exp 1: Sub-tensor (%)</th>
<th>Sub-exp 2: PCA (%)</th>
<th>Sub-exp 2: Sub-tensor (%)</th>
<th>Sub-exp 3: PCA (%)</th>
<th>Sub-exp 3: Sub-tensor (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>90%</td>
<td>95%</td>
<td>10%</td>
<td>60%</td>
<td>10%</td>
<td>15%</td>
</tr>
<tr>
<td>19</td>
<td>100%</td>
<td>100%</td>
<td>10%</td>
<td>90%</td>
<td>25%</td>
<td>44.4%</td>
</tr>
</tbody>
</table>

Table 6.3: Recognition rate results of Group B

<table>
<thead>
<tr>
<th>Eigenvectors</th>
<th>Sub-exp 1: PCA (%)</th>
<th>Sub-exp 1: Sub-tensor (%)</th>
<th>Sub-exp 2: PCA (%)</th>
<th>Sub-exp 2: Sub-tensor (%)</th>
<th>Sub-exp 3: PCA (%)</th>
<th>Sub-exp 3: Sub-tensor (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>65%</td>
<td>65%</td>
<td>5%</td>
<td>45%</td>
<td>15%</td>
<td>50%</td>
</tr>
<tr>
<td>19</td>
<td>95%</td>
<td>90%</td>
<td>10%</td>
<td>80%</td>
<td>50%</td>
<td>65%</td>
</tr>
</tbody>
</table>

The sixth experiment retrieved the recognition results using the first 10 eigenvectors on a varying number of input query datasets. The first 10 eigenvectors were selected because they encode most of the shape variance (the energy is 90% or more). The average recognition rates of all four levels of expression intensity over 54, 40 and 30 subjects are as shown in Table 6.4. A comparative recognition of results between PCA and sub-tensor SVD methods are also included.

Table 6.4: Recognition rate results by differing number of subjects

<table>
<thead>
<tr>
<th>Expressions</th>
<th>PCA: 54 subjects (%)</th>
<th>Sub-tensor: 54 subjects (%)</th>
<th>PCA: 40 subjects (%)</th>
<th>Sub-tensor: 40 subjects (%)</th>
<th>PCA: 30 subjects (%)</th>
<th>Sub-tensor: 30 subjects (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>87%</td>
<td>88%</td>
<td>90%</td>
<td>87%</td>
<td>90%</td>
<td>90%</td>
</tr>
<tr>
<td>Disgust</td>
<td>83%</td>
<td>87%</td>
<td>85%</td>
<td>90%</td>
<td>86%</td>
<td>90%</td>
</tr>
<tr>
<td>Fear</td>
<td>92%</td>
<td>94%</td>
<td>90%</td>
<td>92%</td>
<td>90%</td>
<td>93%</td>
</tr>
<tr>
<td>Happy</td>
<td>81%</td>
<td>85%</td>
<td>85%</td>
<td>87%</td>
<td>83%</td>
<td>90%</td>
</tr>
<tr>
<td>Sad</td>
<td>94%</td>
<td>94%</td>
<td>95%</td>
<td>92%</td>
<td>93%</td>
<td>96%</td>
</tr>
<tr>
<td>Surprise</td>
<td>68%</td>
<td>72%</td>
<td>75%</td>
<td>77%</td>
<td>73%</td>
<td>76%</td>
</tr>
</tbody>
</table>

The subjects were randomly selected with a balanced number of males and females in the training sets. The results demonstrate higher recognition rates when using the sub-tensor method. As the number of subjects reduces to 30 individuals, the recognition rates for both PCA and sub-tensor SVD increase. The increasing recognition rate could be due to a well separated inter-class and intra-class scatter especially with the lower number of subjects used. The data size with 54 subjects is very high (almost 15 million (\(= 3*5090 \text{ points} \times 54 \text{ subjects} \times 6\)))...
expressions × 4 levels of expression intensity)) while with 30 subjects the data size is around 8 million.

To summarise, the sub-tensor SVD approach can be beneficial in face recognition. The experiments show improved recognition performance over the conventional eigenfaces method. A number of advantages of this approach can be recognised from the analysis. The concept of the sub-tensor SVD is much simpler technically in comparison to the core tensor method. Since we involved only two main factors, that is the subject and facial expressions, the computation time taken is much faster than the time taken to execute PCA. Furthermore, the coefficient vector is relatively smaller in dimension and depends on the number of subjects, $I_{sub}$. While in PCA, the coefficient vector has dimension $(I_{sub} \times I_{exp})$.

In comparison to Vasilescu and Terzopoulos [294]'s work, our sub-tensor method differs in that we compute the global mean before computing PCA (or SVD). This makes sense from the perspective of both the dimensionality of the linear space spanned by the data and from the Gaussian assumption of the distribution of the data. Other tensor related works by Wang and Ahuja [297], and Park and Savvides [231] also recognise the importance of computing the global mean before undertaking the tensor steps. In finding the optimal set of subject vectors for each set of expressions, we calculate the distances between two faces and the closest distance is considered as the best match. This approach can be slow because the search for the overall best match has to be repeated across all the database. Hence, Vasilescu and Terzopoulos proposed an iterative search method while Park and Savvides proposed a restriction to the iterative method by adding weight vectors. This method has the disadvantage that it may leads to underperformance of the HOSVD approach [284, 231].

6.7 Conclusions

In this chapter, we explained the use of multilinear algebra in face recognition. A sub-tensor approach was proposed for representing a collection of 3D face models with multiple formation factors. The sub-tensor defines a multilinear projection onto a tensor space of lower dimensionality that captures the majority of the variation found in the original data representation. In the parsimonious representation multiple factors are separated and decomposed.
The overall sub-tensor SVD results have shown a slight improvement in the recognition accuracy over the original eigenface technique. Next, we look into the possibility of employing the technique in face recognition using deformation information.
Chapter 7

Face recognition using statistical modelling approaches

This chapter explores techniques used for manipulating facial expressions in 3D face data that could be used to improve human face recognition performance. The idea is to adjust the expression of a subject until it provides an intuitive match to a face held in a database. Work on synthesizing realistic facial expressions started more than three decades ago [239, 232]. More recently there has been some work done on synthetic biometrics concerning the generation of artificial biometric data which can be useful for identification and authentication systems, for example generating data that is missing, or has been corrupted.

We investigated methods to generate new facial expressions from a set of real human 3D face surface maps. The methods used include (1) statistical shape modelling and (2) statistical discriminant modelling. In the statistical shape modelling approach, new facial expressions are created by moving the surface points along the appropriate expressive direction in the training set space. In the statistical discriminant model, new facial expressions are synthesized by moving the surface points along the most discriminant direction found from the classes of expressions in the training set. The key advantage of the statistical discriminant approach is that facial expressions of varying degrees can be generated from a small number of examples available in the 3D expression database. The results gathered from the deformation models are analysed and validated by testing their recognition performance.
7.1 Statistical Shape Models

There are a number of approaches to build statistical shape models of the face. The two commonly used approaches are Active Shape Models (ASM) and Active Appearance Models (AAM). Active shape models were introduced by Cootes et al. [70] for modelling anatomical structures from statistical information found in a training set of labelled examples. With ASM, the first step involves creating a training set where landmarks are placed as an outline of the key shape features. Next, these sets of landmarks are iteratively deformed to fit to the corresponding reference landmark sets. The aim of the alignment is to correct scaling, translation and rotation differences. The Procrustes method is used to perform the iteration [124]. The listing below shows the alignment algorithm [70]:

**Listing 2** Alignment algorithm

1: Align (utilising rotation, translation and scaling) each shape with the first shape of the population
2: repeat
3: Calculate the mean shape from the aligned shapes
4: Normalise the orientation, scale and origin of the mean to suitable defaults
5: Realign every shape of the population with the current mean
6: until the process converges

After the shape alignment step, a correspondence is established between each point of each set. The resulting alignment forms a multidimensional space. Let \( \{x_i\} \) denote \( n \) shape vectors where \( i = 0 \ldots N \). Each shape consists of \( m \) 3D landmarks, \( \{p_j = (p_{xj}, p_{yj}, p_{zj})\} \) where \( j = 1 \ldots m \) and \((x, y, z)\) are the coordinates of the landmark. Hence, each vector \( x_i \) consists of landmarks \\

\( (p_{11}, p_{21}, p_{31}, p_{12}, p_{22}, p_{32}, \ldots, p_{xm}, p_{ym}, p_{zm}) \).

Once the points converge into a common coordinate system, each shape \( x_i \) can be represented by a single point in a \( 3m \)-dimensional space.

A Point Distribution Model (PDM) that describes the variation in the coordinates of the aligned landmarks is generated. Assuming that the variation is Gaussian, the center and the principal axes of the ellipsoid are calculated. The Gaussian is infinite, but the magnitudes of the principal axes are limited by a cut off of 3 standard deviations. The centre is the mean shape while the principal axes are directions with some given variance. By using a cut-off on
the spread along the principal axes, a sample may be reduced in its dimensionality. This is the method of PCA that as described in Section 6.1.

After applying PCA to the training set, the approximate distribution of the landmarks can be generated with a linear model in the form:

\[ x = \bar{x} + \Phi_s b_s \]  

(7.1)

where \( \bar{x} \) is the average landmark vector, \( b_s \) is the shape parameter vector of the model, and \( \Phi_s \) is a set of orthogonal modes of variation.

The linear model in Equation 7.1 is called the PDM. A new shape example can be generated by varying the parameters \( b_s \) with suitable limits. Assuming that the distribution is Gaussian, the variance of the \( i^{th} \) parameter of \( b_s \) across the training set is given by \( \lambda_i \). Limits in the variation of \( b_{si} \) are set such that \( b_{si} \leq \pm 3 \sqrt{\lambda_i} \). This ensures that any reconstruction from the PDM remains plausibly within the distribution of the training class.

A similar technique, known as Active Appearance Models (AAM) [70, 72, 120], can be used to model the appearance and the shape of the face. The AAM is a generalisation of the ASM. It is generated by combining a model of shape variation with a model of the appearance variance in a shape-normalised frame. In order to build a statistical model of the appearance, each shape of the population is warped so that its control points match the mean shape.

The AAM approach is applied to face reconstruction using a combination of distinct PCA features [70, 69, 71, 72]. This means that the sources of shape variation have to be calculated separately and then matched together in order to extract and interpret the most expressive differences of the training face data. One of the problems with the AAM approach is that it does not provide an adequate geometric analysis of specific shape variations. Since it is based on PCA, the modes involve all the variables and do not easily separate expression from subject differences. Hence an alternative approach is required to manage the face variabilities.

An alternative approach for modelling shapes makes use of Linear Discriminant Analysis (LDA) [17, 321]. While the objective of PCA is to find modes of shape variation which explain the maximal amount of variance in the population, the primary purpose of LDA is used to classify a shape. Classes can be defined on specific features such as expressions.
In LDA, each 3D-dimensional face shape consists of \( n \) points, \( \{ p_j = (p_{xj}, p_{yj}, p_{zj}) \} \) where \( j = 1 \ldots n \) and there are \( N \) examples used for training. Hence, the original data would be \((N \times n)\) in size. This data is divided into \( g \) groups or classes \( C_1, C_2, \ldots, C_g \).

The group information is used to find the optimal projection vector \( P_{lda} \) for classification of new data. It is computed by maximising the ratio of the determinant of the between-class separability to the determinant of the within-class variability. It is assumed that the true covariance matrices of each class are equal because the same within-class matrix is used for all the classes [158]. \( P_{lda} \) is defined as:

\[
P_{lda} = \arg \max P^T S_b P \frac{1}{P^T S_w P}.
\]

(7.2)

where \( S_b \) is the between-class and \( S_w \) is the within-class covariance matrix. The between-class covariance matrix \( S_b \) is defined as

\[
S_b = \frac{1}{g-1} \sum_{i=1}^{g} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T
\]

(7.3)

and the within-class covariance matrix \( S_w \) defined as

\[
S_w = \frac{1}{N - g} \sum_{i=1}^{g} S_i (N_i - 1) = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T
\]

(7.4)

where \( x_{i,j} \) is the \( m \)-dimensional pattern \( j \) from class \( i \), \( N_i \) is the number of training patterns from each class \( i \) and \( g \) is the total number of classes or groups. The class mean is the \( \bar{x}_i \) and \( S_i \) is the covariance matrix for each class. The estimate of \( S_w \) is often known as the pooled covariance estimate of \( S_p \) since in Fisher’s original work [106], he used the average of the individual class covariance matrices as an estimate of the within class covariance. The grand mean vector \( \bar{x} \) is given by

\[
\bar{x} = \frac{1}{N} \sum_{i=1}^{g} N_i \bar{x}_i = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{N_i} x_{i,j}
\]

(7.5)

where \( N \) is the total number of samples.

Using Equation 7.2, it can be shown that the projection matrix \( P_{lda} \) can be found by solving
the eigenvector problem:

\[(S_w^{-1}S_b)P = PA\]  \hspace{1cm} (7.6)

Thus, the optimal projection \(P\) and \(\Lambda\) is found by evaluating the eigenvectors and eigenvalues of \((S_w^{-1}S_b)\) where \(S_w^{-1}\) is the inverse of \(S_w\). The eigenvectors of \((S_w^{-1}S_b)\) have at most \((g - 1)\) nonzero real corresponding eigenvalues.

One disadvantage of the standard LDA is that the performance degrades if the total number of training observations \(N\) is smaller than the dimension of the feature space \(n\), \((N < n)\). Face recognition often involves small training sets, a large number of features and a large number of classes, and LDA may be limited by singularity of the covariance matrix. This is known as the small sample size problem. The small sample size can also cause poor estimation of matrix \(S_w\) and the inverse \(S_w^{-1}\) may become mathematically unstable due to singularity problem. A rule of thumb for good performance is that \(N\) should be five to ten times \((n + g)\) [156].

### 7.2 Statistical Discriminant Models

The aim of multivariate statistical analysis is to model complex data with multiple dependent variables in a generalised manner [134]. In the case of faces, a well known fact is that the data is highly redundant because every individual has the same features. Having said that, faces too contain information that contributes to differentiating the individuals. Hence, reducing the dimensional space of the face data needs to be done without removing the significant information that is useful for discriminating facial expressions, poses, age, etc.

The statistical discriminant approach uses all the subjects in order to characterise the variations. Statistical discriminant models use a two-stage process, the first stage is to characterise a type of variation and the second stage is to reconstruct faces. In the following subsections, LDA-based methods for face shape modelling are described.
7.2 Statistical Discriminant Models

7.2.1 Fisherfaces Method

The Fisherfaces method is essentially a two-stage dimensionality reduction technique. This approach is most successful for solving small sample size problems in face recognition applications [24, 325]. The Fisherfaces method is also known as the Most Discriminant Features (MDF) method [279].

In the MDF method, the first step involves projecting the original sample space to a lower dimensional space using PCA. The original training space is arranged in \( N \) training samples over the \( x_i \) landmarks points. Next, LDA is applied on the PCA subspace to find the best linear discriminant features. The MDF projection matrix \( P_{mdf} \) is denoted as:

\[
P_{mdf}^T = P_{lda}^T * P_{pca}^T
\]

(7.7)

where \( P_{pca} \) is the projection matrix from the original space to the PCA space and the \( P_{lda} \) is the projection matrix from the PCA subspace to the LDA subspace obtained by maximizing the ratio:

\[
P_{mdf} = \arg \max \frac{|P_{lda}^T P_{pca} S_b P_{pca} P|}{|P_{lda}^T P_{pca} S_w P_{pca} P|}
\]

(7.8)

Fisher’s criterion is maximised when the projection matrix \( P_{lda} \) is composed of eigenvectors of \((P_{pca}^T S_b P_{pca})^{-1}\) with at most \((N - g)\) nonzero corresponding eigenvalues.

7.2.2 The Maximum uncertainty LDA-based Approach

A method of overcoming the singularity of \( S_w \), based on the idea of enhancing the smaller eigenvalues, was developed by Thomaz et al. [55]. Starting with the pooled matrix, the approach uses the largest dispersion values of the \( S_p \) average eigenvalues to stabilise it. This linear discriminant regularised approach is known as mLDA which stands for Maximum uncertainty LDA method. It relies on the fact that minimising a more difficult but appropriate inflated within-class matrix \( S_w \) also minimises a less reliable shrivelled within-class estimate. Thomaz and colleagues showed that implementing the mLDA improves LDA in face image classification.

The \( \Lambda \) eigenvalues and \( \Phi \) eigenvectors of the original pooled covariance matrix \( S_p \) are com-
computed. The \( \bar{\lambda} \) average eigenvector is calculated. A new set of \( \Lambda^* \) eigenvalues is formed following the largest dispersion eigenvalues, but replacing small eigenvalues with the average eigenvalue. The listing below explains the dispersion algorithm.

**Listing 3** The \( \Lambda^* \) dispersion eigenvalues algorithm

1: Calculate the mean eigenvalues
\[
\bar{\lambda} = \frac{\sum_{i=1}^{N} \lambda_i}{N}
\] (7.9)

2: Sort the eigenvalues from the largest \( \lambda_{max} \) to the smallest \( \lambda_{min} \) eigenvalues

3: repeat
4: if \( \lambda_i < \bar{\lambda} \) then
5: Replace \( \lambda_i \) with \( \bar{\lambda} \)
6: end if
7: until all variables processed
8: Create a new \( \Lambda^* \) matrix in which the new \( \lambda_i \) are arranged diagonally

The modified within class covariance matrix \( S_w^* \) is:
\[
S_w^* = (\Phi \Lambda^* \Phi^T)(N - g)
\] (7.10)

The advantage of the approach is that the method is straightforward and does not require high computation. Furthermore, only those eigenvalues with small and less reliable values are expanded. The dispersion method is believed to provide a parsimonious method of estimating efficient covariance structures for statistical covariance-based classifiers in small sample size problems [279].

### 7.2.3 Statistical Discriminant Analysis

Statistical discriminant analysis (SDA) can be used to find the most significant direction of change between two classes, and to re-construct and visualise intermediate data between the classes. The first stage is the linear classification of training data using a single separating hyperplane, which can be followed by synthesis and visual analysis. Kitani et al. [171] called this approach the statistical discriminant model.

The classes used in the training data describe what is being discriminated, and can be features based on gender, age, facial expression and so on. For example, the classes may be male and female, characterising the variational changes in face shape from male to female. Alternations...
tively we can select the training data to separate smiling expressions from neutral. In each case we construct a discriminant in the most likely direction of change between the two classes.

The initial training set of 3D face data consisting of \( N \) training examples on \( n \) variables is managed by dividing the training data into two groups or classes, \( C_1, C_2 \). The training datasets can be projected from the original vector space of \( N \) by \( n \) to a lower dimensional space using a full rank PCA transformation. The principal component space forms an \( n \times m \) transformation matrix, where \( m = N - 1 \). This step may or may not be necessary to overcome the singularity of the within class covariance matrix. If \( N \geq n \), the PCA transformation is not required. It is possible that after PCA dimensionality reduction, the within-class scatter matrix \( S_w \) may still be less than full rank [326]. If so, the mLDA approach is used to ensure that the scatter matrix \( S_w \) is nonsingular. Ordering the eigenvectors is not necessary for this process. As there are only two classes, \( g = 2 \) and the resulting mLDA is a unidimensional vector of length \( m \). An overview of the first-stage of the SDA algorithm is given in Listing 4.

The next stage of the statistical discriminant analysis method is a reconstruction step. If we project the most discriminant vector found for the two classes into the original data space, we will obtain an \( n \times 1 \) vector. Moving a point in the original data space in this direction will change the point from an example of one class to a maximum likelihood estimate of that point in the other class. For example, changing a face with a smiling expression to a neutral expression. Figure 7.1 shows the overview of the SDA method.

![Figure 7.1: The geometric overview of the SDA method between two expression classes.](image-url)
### Listing 4 The first-stage of statistical discriminant analysis

1: Compute grand mean

\[ \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \]

2: Compute covariance matrix

\[ S = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T \]

3: Compute the between-class scatter matrix \( S_b \) (as in Equation 7.3) and the within-class scatter matrix \( S_w \) (as in Equation 7.4)

4: if \( S_w^{-1} \) is a singularity problem then

5: Compute \( \Lambda^* \) as in Listing 3

\[ \Lambda^* = \text{diag} [\max(\lambda_1, \bar{\lambda}), \max(\lambda_2, \bar{\lambda}), \ldots, \max(\lambda_N, \bar{\lambda})] \]  

(7.11)

6: Implement Equation 7.10

7: Replace the within-class scatter matrix \( S_w \) with the new \( S_w^* \)

\[ W_{lda} = \left| \frac{W^T S_b W}{W^T S_w^* W} \right|. \]  

(7.12)

8: else

9: Compute

\[ W_{lda} = \left| \frac{W^T S_b W}{W^T S_w W} \right|. \]

end if

Assuming that the spread of each class follows a Gaussian distribution, the limits of variation can be set to \( \pm 3sd_i \), where \( sd_i \) is the standard deviation of each class \( i \). By moving along the \( n \times 1 \) dimensional most discriminant features based on the mean of each group and the corresponding standard deviations of each group, the face shapes according to the class variant can be reconstructed in the original face domain. Since there is only one-dimensional data from a two group classification problem, the computation is very fast.

#### 7.2.4 Tensor-based Statistical Discriminant Analysis

In this section we propose extending statistical discriminant analysis to problems formulated with the tensor model. In the previous chapter, the tensor organisation was \( I_{sub} \times I_{exp} \times I_{pts} \), where \( I_{exp} \) is a set of facial expressions. In this experiment, the data tensor can be expanded to include a range of expressions \( I_{sub} \times I_{anger} \times I_{disgust} \times I_{fear} \times I_{happy} \times I_{sad} \times I_{surprise} \times I_{neutral} \times I_{pts} \).
We can index this data tensor to obtain multiple two set facial expression sub-tensors, each with a dimension of $(I_{sub} \times 2) \times I_{pts}$. The conceptual diagram of a sub-tensor model is as illustrated in Figure 7.2.

![Figure 7.2: The geometric idea of the tensor model for multiple facial expression classes](image)

Each sub-tensor can be flattened to obtain a matrix containing specific facial expression transformations. The sub-tensor matrix is $M \times I_{pts}$, where $M$ is the number of 3D face surfaces of the two sets of facial expressions, which makes $M = 108(= I_{sub} \times 2)$ for any two expression classes. The resulting sub-tensor matrix that goes through the first-stage of SDA, as described in Listing 4, has multiple coefficient vectors depending on the choice of facial expression transformations. Then it is followed by the projection.

The advantage of the tensor-based SDA is that we can represent the principal axes of variation across various expression modes to allow us to analyse how the various expression classes interact with one other.

### 7.3 Experimental Work

We carried out deformation experiments using four face databases - the Imperial College, Binghamton, Notre Dame and 2D Feret databases. The first two contain facial expression examples, the other two were used to analyse the realism of the synthesized facial expressions. Only geometric information was used in the deformation experiments. The experimental work is divided into three main parts - (1) synthesizing facial expressions, (2) deforming faces with expression
to a near neutral expression and (3) recognising the deformed faces.

In part (1) of the experiments, we extracted expression information from the Imperial College London and the Binghamton University 3D face databases. In total, there are seven classes of expressions. Using the tensor directions the most discriminant vectors were calculated. By moving along the most discriminant direction, facial expressions can be reconstructed. These synthesised expressions reflect the facial expression changes available in the data bases. These experiments elucidate the subtle changes in facial expressions. We also tested some 3D face surfaces which were re-constructed from 2D FERET face images. The 3D reconstruction was done by a colleague and the details of the technique can be found in [8].

Similar steps were undertaken in the second experiment in which faces with an expression are altered to a near neutral expression. These near neutral expressions were used in the last experiment to recognise subjects in the face database.

7.4 Results and Discussion

The following sections present the results of the expression deformation experiments. The synthesized facial expressions using the proposed tensor-based statistical discriminant analysis are compared to Active Shape Models (ASM).

7.4.1 Reconstruction of Faces with Expression

In the first part of this experiment, we visually examined reconstructed faces using the most variant features captured with ASM. The mean training data is an \( n \)-dimensional point around which we move using \( m \) principal components, reconstructing the respective coordinates in the face space \([70], [69], [71], [72]\). The reconstructed faces are restricted by limiting the change in each principal component to \( \pm 3 \sqrt{\lambda_i} \), where \( \lambda_i \) are the corresponding eigenvalues.

Figure 7.3 illustrates the transformations of the first four most variant components of shape variation of a mixture of all the expressions. The first mode describes the horizontal size variations around the cheek and mouth to create expression changes from a distorted frowning expression to a caricature smile. The second mode modelled vertical variation. The third mode
captures variation around nose and eyes areas. The fourth mode captures horizontal variation.

Figure 7.3: The reconstruction of the largest PCA modes on the VisionRT face datasets

Figure 7.4 illustrates the mode variations on the NotreDame datasets, which contain only neutral faces. The most discriminant directions capture the variations in terms of the size, horizontal, vertical and depth information.
Figure 7.4: Reconstruction using the principal modes on NotreDame face datasets
Figure 7.5 illustrates the most variant shape modes of a mixture of six expressions from the Binghamton University face datasets. The first mode describes the vertical stretch along the centre of the face. The second mode models the variations in the horizontal direction. These variations create opening and closing of the mouth and changes in the size of the nose. The third mode captures variations around the cheek and mouth to create expression changes from a distorted frowning expression to a caricature smile. The fourth mode captures the variations in the horizontal and vertical geometric shape of the face. From the profile view in the figure, we observed that the geometric shape variations caused faces to have different facial expressions.

Examining Figures 7.3 to 7.5, we see that the facial expression groups are not properly organised. In Figure 7.3 the first mode shows changes of facial expression that are not smooth. Moreover, the synthesis of the frowning expression is incorrect due to geometric distortions around the mouth and the chin. Similarly, the generated smiling expressions look abnormal in many instances. The changes of shape are global to the data set and this makes synthesizing individual expressions unrealistic [171]. This can be seen in the lower variance modes in Figure 7.3. Even though subtle changes of expression can be seen visually, the overall geometric face shape changes in other ways. Thus, the ASM approach is not suitable to capture facial expression variations. In contrast, statistical discriminant analysis is proposed for effective capture of facial expression variations, as it is able to find the most characteristic direction of change involved in an expression and the magnitude of that change can be controlled by a single scalar magnitude.

Figure 7.6 presents the SDA reconstructions for smiling and frowning expressions using the information gathered from the two expression groups. We move from one side of the dividing hyper-plane to the other respecting the limits of the standard deviation and the measured mean of each sample group. The figure shows the calculated facial expressions. The top row of the figure displays the synthesized smiles and the row below illustrates the synthesized frowns. By moving from the left to the right of the figure, the expression changes from a mild smile to an exaggerated smile (for the top row) and from an exaggerated frown to a less frown expression (for the bottom row). We can clearly see that the SDA most discriminant direction effectively extracts the 3D facial expression changes. In fact, the SDA most discriminant direction is able
Figure 7.5: Reconstruction using the principal modes of Binghamton University face dataset
to generate a gradual change on facial expressions that is not explicitly present in the training datasets.

![Figure 7.6: Reconstruction using the SDA most discriminant direction for facial expressions. Left to right (top) $-3sd_1$ to $+3sd_1$ and (bottom) $-3sd_2$ to $+3sd_2$, where $sd_1$ is the smiling axis and $sd_2$ is the frowning axis](image)

From the profile view in Figure 7.6, we can also see subtle expression changes in the geometric shape of the face. They show that facial changes are localised around the mouth, cheek, eyes and eyebrows. For example, the smiling changes involve raised cheek and eyebrows, and opened mouth. In frowning expressions, the eyebrows are lowered, the eyes are narrowed slightly, the cheeks are flattened, the mouth is shut and both corners of the mouth are moved downwards. Additionally, there is a visible change in the shape of the face as the chin is pushed downwards while smiling and frowning causes the chin to move upwards and close the mouth.

Figure 7.7 shows the SDA most discriminant features for neutral and angry expression reconstructions. Figure 7.8 illustrates the deformation between an angry and a surprise expression (and vice versa). These expression changes were synthesized using the Binghamton University face database.
7.4 Results and Discussion

Figure 7.7: Reconstructions from the Binghamton University faces using the most discriminant direction between neutral and angry expressions. Left to right $-3sd_1$ to $+3sd_1$ of the angry expression.

In Figure 7.8, artifacts can be seen in the synthesized 3D faces as we move along the discriminant direction towards the extremes $+3sd_1$ and $-3sd_2$. The artifacts become more extreme as we move towards the edge of the face space where the shapes are less well represented.

We have also tested the change of expressions to 3D face surfaces that are not in the training
set. This can be done on the condition that we follow the same spatial normalisation protocols. Once done, the process of transferring the most discriminant feature to a new 3D face surface is similar to the previous experiments. The 3D Notre Dame face datasets was used. Figure 7.9 shows the results when we move the faces across the facial expression hyperplane as calculated previously. The fourth column from the left shows the original faces. As we move to the left from the original face, the expression changes from a neutral face to a smiling face.

![Figure 7.9: Reconstruction using the SDA most discriminant direction when using the Notre Dame data along the smile(left)/frown(right) axes.](image)

In order to ensure that the generated facial expressions are truly representative of a subject, we made use of the four intensity expression levels available in the Binghamton University face database. The Figure 7.10 illustrates the reconstruction of angry expressions using the four levels. Training was done to find the axis between level 1 of anger and level 4 of anger. A subject was chosen from the database and reconstructions were made from each level of expression to the others. The top row shows the four levels of an angry expression of the chosen subject. The remaining rows represent the reconstruction of angry expressions starting from the true subject at each level. The reconstruction of expression is presented over the range of $sd_1$. In this case $sd_2$ is larger and the extreme reconstruction can show distortion. From the figure, it can be seen that the generated expressions are similar to the corresponding real expressions. When
comparing the two face surfaces, the computed RMS is the minimal (average below 0.9mm), as shown in Table 7.1.

Table 7.1 shows the RMS error between the reconstructed expressions and the original expressions for Level 1. It can be seen that the errors are minimal when reconstructing original expression. Similar results are also shown for the remaining levels.

Table 7.1: The RMS along the most discriminant component for angry expressions of Level 1.

<table>
<thead>
<tr>
<th>Intensity levels</th>
<th>$-3sd_1$</th>
<th>$-2sd_1$</th>
<th>$-1sd_1$</th>
<th>$+1sd_1$</th>
<th>$+2sd_1$</th>
<th>$+3sd_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>0.013</td>
<td>0.235</td>
<td>0.517</td>
<td>0.492</td>
<td>0.745</td>
<td>0.0337</td>
</tr>
<tr>
<td>Level 2</td>
<td>1.440</td>
<td>1.397</td>
<td>1.393</td>
<td>1.642</td>
<td>1.793</td>
<td>1.967</td>
</tr>
<tr>
<td>Level 3</td>
<td>1.446</td>
<td>1.402</td>
<td>1.402</td>
<td>1.642</td>
<td>1.783</td>
<td>1.947</td>
</tr>
<tr>
<td>Level 4</td>
<td>2.552</td>
<td>2.462</td>
<td>2.395</td>
<td>2.795</td>
<td>2.942</td>
<td>3.103</td>
</tr>
</tbody>
</table>

The SDA approach can also be used to synthesize realistic facial expressions on single frontal 2-dimensional face images. However, employing SDA on 2D images creates many
unwanted artifacts [171]. Therefore, instead of using 2D face images, we reconstructed 3D face models from the 2D frontal face images and then generate facial expressions on the reconstructed 3D faces. We use the 2D FERET face database. The 3D face reconstruction was done using an analysis-by-synthesis approach based on a statistical shape model [8]. Figure 7.11 shows 3D faces reconstructed from the FERET 2D face dataset with synthesised facial expressions.

![Figure 7.11: Illustration of SDA expression synthesis on 3D subjects reconstructed from the FERET database](image)

The fourth column from the left in the figure displays the original faces. As we move from the original to the left side of the figure, a range of smiling expressions is generated. Similarly when we move to the right side of the figure, frowning expressions are generated. Having texture embedded to the 3D face surfaces makes the expression change smoother and more visible. For example, the raised cheeks and eyebrows, and the opened mouths show an obvious smile. As for frowning, we can see that the eyebrows are lowered, the eyes are narrowed slightly, the cheeks are flatter and the mouths are shut and both corners of the mouth are moved downwards. Since the FERET and NotreDame face databases contain no facial expression information, quantitative analysis of the synthesized expressions could not be done. Assessment of the results is purely visual.
7.4.2 Reconstruction of Neutral Faces

This section describes the second part of our experiments in which we reconstruct near neutral faces from faces with expressions. The steps are similar to those used in the previous section. Figure 7.12 shows examples of the reconstructed near neutral expressions of a subject. Each row in (b) represents an expression. The top row is an expression of disgust, followed by anger, fear, sadness and surprise.

![Illustration of original faces with expressions and near neutral faces](image)

Figure 7.12: Illustration of (a) original faces with expressions (b) exaggerated and neutralised examples of those faces.

As we move towards the left of the figure, the near neutral expression is reconstructed. If we moved towards the right, the relevant expression is exaggerated. By comparing the appearance of the neutral faces with the reconstructions of near neutral faces we see some similarities. For example, the eyebrows and eye areas are shaped similarly to the original neutral face. Interestingly, if the faces with expression have the mouth closed, the synthesised neutral faces are very
close to the original neutral face. However, for this particular subject, a near neutral face cannot be reconstructed from the anger and surprise expressions where the mouth is open.

Table 7.2: The RMS of Type 1 and Type 2 (measured in mm)

<table>
<thead>
<tr>
<th>Test</th>
<th>Type</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>+1</th>
<th>-1</th>
<th>+1</th>
<th>+2</th>
<th>+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digest</td>
<td>Type 1</td>
<td>2.634</td>
<td>2.801</td>
<td>3.017</td>
<td>2.428</td>
<td>2.229</td>
<td>3.423</td>
<td>3.832</td>
<td>4.267</td>
</tr>
<tr>
<td></td>
<td>Type 2</td>
<td>0.204</td>
<td>0.148</td>
<td>0.499</td>
<td>0.906</td>
<td>0.077</td>
<td>1.07</td>
<td>1.586</td>
<td>2.102</td>
</tr>
<tr>
<td>Angry</td>
<td>Type 1</td>
<td>2.163</td>
<td>2.123</td>
<td>2.116</td>
<td>2.18</td>
<td>2.749</td>
<td>2.447</td>
<td>2.59</td>
<td>2.753</td>
</tr>
<tr>
<td></td>
<td>Type 2</td>
<td>1.081</td>
<td>0.837</td>
<td>0.593</td>
<td>0.105</td>
<td>0.0392</td>
<td>0.621</td>
<td>0.893</td>
<td>1.165</td>
</tr>
<tr>
<td>Fear</td>
<td>Type 1</td>
<td>3.349</td>
<td>3.448</td>
<td>3.549</td>
<td>3.666</td>
<td>4.131</td>
<td>4.397</td>
<td>4.681</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type 2</td>
<td>0.631</td>
<td>0.324</td>
<td>0.0779</td>
<td>0.168</td>
<td>0.17</td>
<td>0.971</td>
<td>1.371</td>
<td>1.772</td>
</tr>
<tr>
<td>Sad</td>
<td>Type 1</td>
<td>1.347</td>
<td>1.231</td>
<td>1.205</td>
<td>1.436</td>
<td>1.299</td>
<td>1.203</td>
<td>1.221</td>
<td>1.285</td>
</tr>
<tr>
<td></td>
<td>Type 2</td>
<td>0.7145</td>
<td>0.359</td>
<td>0.00309</td>
<td>0.7075</td>
<td>0.595</td>
<td>0.109</td>
<td>0.135</td>
<td>0.378</td>
</tr>
<tr>
<td>Surprise</td>
<td>Type 1</td>
<td>5.717</td>
<td>5.269</td>
<td>4.871</td>
<td>4.249</td>
<td>4.011</td>
<td>3.883</td>
<td>3.84</td>
<td>3.813</td>
</tr>
<tr>
<td></td>
<td>Type 2</td>
<td>2.6105</td>
<td>1.995</td>
<td>1.38</td>
<td>0.1495</td>
<td>0.509</td>
<td>1.021</td>
<td>1.276</td>
<td>1.531</td>
</tr>
</tbody>
</table>

In Table 7.2, Type 1 is the RMS error of the reconstructed faces shown in Figure 7.12 versus the original neutral face of the particular subject. Type 2 is the RMS error of the original face with expression against the reconstructions of Figure 7.12. The eight magnitude levels represent the levels of expression change. The three columns from the right denote the standard deviation of exaggerated expressions, while column 2-4 represent neutral faces. It can be seen that the RMS values of Type 1 increase as we moved along the magnitude towards the exaggerated expressions. This shows that there is a change of expressions from one class to another class of expressions. While in Type 2, we can see that in column 6 and 7, Type 2 RMS values are closer to zero. This should be the case if the test faces are similar to the original expressions of the specific subject.

The RMS pattern is very much dependent on the intensity of the expressions gathered from the face datasets. In neutralising faces, the rule is to identify the minimum distance between the reconstructed faces with the original neutral faces. However, there may be cases when the minimum RMS error is not the best representation of neutral expression.

Figure 7.13 shows the reconstruction of neutralising faces of a specific sample taken from the Imperial College database. The figure shows a smiling expression being deformed to a neutral expression. Figure 7.14 displays the means of neutral and smiling expressions, which can be used to compare with Figure 7.13.
7.4 Results and Discussion

Figure 7.13: The near neutral face reconstruction from a smiling expression of a subject.

Figure 7.14: The overall mean of the neutral and the mean of smiling expressions captured from the Imperial College face database.

7.4.3 Quantitative Comparison of the Models in Face Recognition

This last experiment aims to recognise individuals using the reconstructed non-expressioned faces. Subjects were taken from the database with the expressions at the fourth level. We chose three appropriate levels of neutralisation based on the previous section. We compared the recognition rates of the proposed facial expression transformation method using the sub-tensor SVD with standard PCA. We used the Euclidean distance to measure the similarities between the faces.

The recognition results of the reconstructed non-expressioned faces were tested on three transformation magnitudes, each being a standard deviation of the distribution of the neutral
faces. Figure 7.15 summarises the face recognition results for all five expressions. As can be seen in this figure: (1) face recognition using the neutralised faces always gives better recognition results than with the unneutralised (2) using 10 eigenmodes, the recognition rates using PCA is about 89% while recognition rates for neutralised faces is more than 90%, (3) as the number of eigenmodes increase, the result of the recognition for sd-1 and sd-2 go as high as 98%. However, increasing the degree of neutralisation does not necessarily increase recognition accuracy. The smaller degree of neutralisation achieves the best result.

![Figure 7.15: The face recognition rate over the different numbers of eigenmodes for five expressions.](image)

We found that by neutralising facial expressions, the recognition rates for subjects with similar facial expressions dropped. This could be because the deformed geometric shape may be closer in distance to other expressions of lower level expression intensity. We have also noticed a decrease in the recognition rate for the surprise expression. As seen in Figure 7.12, the last row represents the neutralisation of a surprise expression and the reconstruction of the neutral face is not accurate. This is because the surprise expression, unlike the other facial expressions in the face database, has high geometric variation caused by the opening of the mouth, dropping of the jaw and raising of the eyebrows.
7.5 Conclusions

The chapter presented the use of statistical modelling approaches to reconstruct and synthesize facial expressions. The tensor-based SDA approach extracts expression discriminant information efficiently, providing the ability to gradually change expressions. The strengths of this work lies in the realism of the expressions generated. We foresee many applications that can employ the method such as improved biometric recognition, computer games and real-time animation.

The proposed method is also novel in generating expressions at varying intensities for a subject without prior examples. We used 3D face data because of the nature of the shape and expression space which makes it possible to generate synthetic faces under a variety of face variation conditions, thus adding further flexibility and realism. This approach differs from many other synthesizing approaches in terms of using the whole face data points instead of selecting feature points or landmarks on the face for shape and texture variations.

From the experiments, we observed that by transferring expressions across subjects using ASM, the overall shape of the face changed, while with the SDA approach, the expression is changed without deforming the overall face shape. Given a large face dataset of expressions of different subjects, predicting the expression and transferring facial expressions to individuals is a straightforward matter. However, this is not the case when deforming from an expressioned face to a non-expressioned face.

According to Blanz [36, 38], neutral face transfer is more difficult than vice versa. When removing expressions, residual differences occur because of extreme expression shape changes such as the opening of the mouth and the closing of the eyes [36]. Examples can be seen in Figure 7.3, Figure 7.4 and Figure 7.5. Unlike the ASM method, the tensor-based SDM approach showed that facial expressions can be efficiently captured in a single linear direction.

We performed face recognition by using the sub-tensor SVD method and compared the recognition results with the PCA. The experimental results show that the face recognition rates of the neutral reconstruction improved slightly especially for neutralised smiling expressions. Recognition rates for the other expressions either remained the same or dropped slightly, depending on the level of deformations.
Chapter 8

Conclusions and Further Works

During this research, we developed various expression-invariant 3D face recognition techniques based on an advanced tensor algebra framework. We showed that the tensor method possesses a remarkable ability to deal with multiple face variations. We explored the sub-tensor approach separating the subject and facial expression factors. We found that the recognition rates obtained using the sub-tensor singular value decomposition method are relatively higher than those obtained by employing PCA alone. Furthermore, we demonstrated that by using the sub-tensor approach, the correct expression of a subject can also be recognised.

We investigated both dense surface point models and the use of a small set of anatomical landmarks that best describe facial expression using the FACS framework. We found that with good placement of landmarks it is possible to distinguish different facial expressions. We demonstrated that landmarks should optimally be placed in particular regions such as the cheeks and eyebrows. We investigated face recognition using landmarks. We found that landmark-based recognition gave slightly higher recognition performance only on faces with very high expression changes, for example smiling subjects. We investigated the possibility of enhancing expression-invariant face recognition by using realistic deformation of facial expressions. The recognition accuracy improved very slightly. Besides biometric recognition, this novel approach could be used in computer games and real-time animation applications.
8.1 Summary of contributions

In Chapter 5, we demonstrated the location and extent of facial expression information that exists in faces. We determined which landmark points are mobile in the different facial expressions. We surveyed craniofacial landmarks, which are placed on prominent and palpable facial features such as the tip of the nose and the corner of the eyes and mouth. In order to maximise expression information, it is important to select landmarks on the facial muscles. We used a set of craniofacial landmarks as a basis to construct a set of pseudo landmarks. The pseudo landmarks were placed in mobile, less palpable areas around the eyebrows, the eyes, the cheek and the mouth, and were consistent with the FACS framework. Since landmarking is difficult around impalpable features, we used a simple method to locate and place those landmarks. We asked ten volunteers to landmark the same data sets and measured the average precision which was about 0.97 mm. The statistical results on the changes in landmark position with expression and subject showed two types of variation: (1) geometric variation that describes the positional changes of facial features and, (2) facial expression variation. It also showed that landmarks placed on prominent face features may contribute to expression variation but not as obviously as the pseudo landmarks. We also demonstrated that the tip of the nose and the inside corners of the eyes have minimal contribution to facial expressions.

In Chapter 6, we presented the use of the tensor model to provide face recognition invariant of facial expression. A tensor model organises face data according to the different types of variation and allows analysis and manipulation of data independently across the different dimensions. We proposed a novel sub-tensor SVD method within the tensor model. The method produced slightly better recognition results than simple PCA and proved to be able to recognise both the subject and the correct expression. In some cases, PCA was found to give higher recognition rates when expression was not considered, but the recognition of a subject with the correct expression was always slightly lower that that achieved by the sub-tensor SVD method. The method is particularly useful when recognising higher intensity expressions that are not available in the training datasets.

We proposed a novel method of expression-invariant face recognition by transferring the input face with an arbitrary expression into its corresponding neutral expression. In Chapter 7, we
introduced a tensor-based SDA method for the reconstruction of expressions. The method can derive a set of matrices describing all facial expression transformations from the tensor. The reconstruction results showed that the tensor-based SDA approach is suitable for capturing facial expression variations, unlike the ASM approach. The expressions that were generated, including neutral expressions, looked more realistic than those generated using ASM. We also tested changing expressions using 3D face surfaces that were not in the training set. The face recognition results using neutralised expressions show slightly higher recognition accuracy when compared with PCA.

8.2 Limitations and further work

8.2.1 Extending the 3D face database

The experimental results reported in this work used a number of 3D face databases. However, only two of the databases contained a variety of facial expressions. These facial expressions are emotion-based expressions and they are only taken once. In order to work on facial expression analysis and recognition, a wider variety of expressions generated using Action Units is desirable.

Greater demographic variety would also be a desirable characteristic for a face database. People from various ethnic backgrounds and of both sexes should be represented in adequate proportions. In addition, the 3D faces should be taken across time lines. It would be useful to investigate another tensor dimension encoding changes due to growth or aging. The face database should also contain more than one expression per subject and include repeat scans.

8.2.2 Improving the pre-processing method

We used the pre-processing method developed by Papatheodorou [229]. The pre-processing steps regularise the face surfaces, clean them by filling holes and removing spikes, standardise the triangulation and create a dense correspondence between all the 3D vertices. The pre-processing was useful, in establishing the dense correspondence between surfaces for modelling and recognition. However, with the Binghamton University 3D face database, the pre-
processing (already applied) was less effective. Some of the pre-processed faces have jagged surfaces. This irregularity will create noise which impairs the statistical modelling approaches. In addition, face surfaces with level 3 expression intensity in this database have high shape coherence creating unnatural recognition results. These problems could be due to an inability to detect defective areas while recovering the shape of the face. The pre-processing method needs to be improved to smooth the jaggedness and holes.

### 8.2.3 Improving the reconstruction method in the tensor-based SDA approach

Modelling facial expressions (and neutralising expressions) is useful for face and facial expression recognition, and animation. The proposed tensor-based SDA approach is able to create realism in synthetic facial expressions (as seen in Chapter 7). However, the recognition results improved only slightly or did not change. The worst result in recognition is in cases with extreme expression changes such as surprise. This is due to problems with the synthesis of the neutral expression from extreme points in the face space. Thus, improving the reconstruction method will be a major factor in improving overall recognition accuracy.

### 8.2.4 Improving facial expression recognition

Since we use the sub-tensor SVD method, this approach has a number of ramifications especially when involving face variations such as facial expressions. We have to search for the overall best match across all the database. Some expressions produce a similar facial appearance to others, for example disgust and anger [53, 66], and this may lead to the incorrect recognition of a person.

In an ideal environment of face and facial expression recognition, where one has access to a large 3D face dataset with a variety of facial variation, one should be able to recognise a face of a subject, with a valid facial expression, as well as recognise the generated facial expressions. Our current facial expression recognition is less effective at distinguishing expressions that appear similar such as disgust and angry expressions, and fear and surprise expressions. Improvement to the facial expression recognition is a desirable component of recognising individual subjects.
8.2 Limitations and further work

8.2.5 Mapping face texture

Another potential improvement to the 3D face databases would be to have texture of higher quality. The current textures used in the existing databases are generally poor quality, being low resolution, grey-scale and unsmoothed (as in Figure 8.1). The experiments performed in this work could be extended to use texture information. A separate texture space could be created in conjunction with the geometric information space. Alternatively, the geometric shape and texture information could be combined in the same tensor model. Using texture information may improve the recognition as well as the deformation of facial expressions.

![Figure 8.1: The samples of the captured RAW faces of Binghamton 3D face database](image)
References


REFERENCES


REFERENCES


