Robust signatures for 3D face registration and recognition
Nair, Prathap M

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

For additional information about this publication click this link.
https://qmro.qmul.ac.uk/jspui/handle/123456789/487

Information about this research object was correct at the time of download; we occasionally make corrections to records, please therefore check the published record when citing. For more information contact scholarlycommunications@qmul.ac.uk
Robust signatures for 3D face registration and recognition

A thesis presented to the University of London
by
Prathap M Nair
for the degree of
Doctor of Philosophy
in
Electronic Engineering

Queen Mary University of London
Mile End Road
E1 4NS, London, UK
October 2009
Robust signatures for 3D face registration and recognition

Abstract

Biometric authentication through face recognition has been an active area of research for the last few decades, motivated by its application-driven demand. The popularity of face recognition, compared to other biometric methods, is largely due to its minimum requirement of subject co-operation, relative ease of data capture and similarity to the natural way humans distinguish each other.

3D face recognition has recently received particular interest since three-dimensional face scans eliminate or reduce important limitations of 2D face images, such as illumination changes and pose variations. In fact, three-dimensional face scans are usually captured by scanners through the use of a constant structured-light source, making them invariant to environmental changes in illumination. Moreover, a single 3D scan also captures the entire face structure and allows for accurate pose normalisation.

However, one of the biggest challenges that still remain in three-dimensional face scans is the sensitivity to large local deformations due to, for example, facial expressions. Due to the nature of the data, deformations bring about large changes in the 3D geometry of the scan. In addition to this, 3D scans are also characterised by noise and artefacts such as spikes and holes, which are uncommon with 2D images and requires a pre-processing stage that is specific to the scanner used to capture the data.

The aim of this thesis is to devise a face signature that is compact in size and overcomes the above mentioned limitations. We investigate the use of facial regions and landmarks towards a robust and compact face signature, and we study, implement and validate a region-based and a landmark-based face signature. Combinations of regions and landmarks are evaluated for their robustness to pose and expressions, while the matching scheme is evaluated for its robustness to noise and data artefacts.
# Table of Contents

Abstract .......................................................... ii
Previously Published Work ........................................... v
Acknowledgments ....................................................... vi
Dedication ............................................................. vii
List of abbreviations ............................................... viii
Glossary ............................................................... ix

1 Introduction ......................................................... 1
   1.1 Imaging faces .................................................. 1
   1.2 Main contributions ............................................. 3
   1.3 Outline of the thesis .......................................... 4

2 State of the art ...................................................... 5
   2.1 Introduction .................................................... 5
   2.2 3D acquisition ................................................ 6
      2.2.1 3D scanners .............................................. 8
      2.2.2 Softwares ................................................ 9
   2.3 3D representations .......................................... 10
      2.3.1 Point clouds .............................................. 11
      2.3.2 Wireframe models ....................................... 12
      2.3.3 Polygonal models ..................................... 13
   2.4 Features ....................................................... 14
      2.4.1 Surface curvature ....................................... 15
      2.4.2 Extended Gaussian Images ............................. 16
      2.4.3 Facial profiles .......................................... 18
      2.4.4 Statistical shape descriptors ......................... 19
      2.4.5 Local indexing methods ............................... 22
   2.5 Face classification .......................................... 25
      2.5.1 Holistic methods ....................................... 25
      2.5.2 Feature-based methods ................................ 28
      2.5.3 Multimodal methods ................................... 30
   2.6 Summary ....................................................... 33
3 Face signatures 35
  3.1 Introduction ................................................. 35
  3.2 Datasets ..................................................... 35
    3.2.1 GavabDB ................................................. 36
    3.2.2 BU-3DFE ................................................. 38
  3.3 Feature detection ............................................ 39
    3.3.1 Model creation .......................................... 39
    3.3.2 Model fitting ........................................... 41
    3.3.3 Evaluation ............................................. 48
  3.4 Face detection ............................................... 53
    3.4.1 Evaluation .............................................. 55
  3.5 Signatures ................................................... 57
    3.5.1 Region-based signatures ................................. 58
    3.5.2 Landmark-based signatures ............................... 60
  3.6 Summary ..................................................... 64

4 Experimental results 65
  4.1 Introduction .................................................. 65
  4.2 Experimental setup ........................................... 65
  4.3 Region-based signatures ..................................... 66
  4.4 Landmark-based signatures ................................... 70
    4.4.1 Signature selection ..................................... 71
    4.4.2 Signature performance .................................. 75
  4.5 Summary ..................................................... 80

5 Conclusions 81
  5.1 Summary of achievements ..................................... 81
  5.2 Future work .................................................. 83

Bibliography 84
Previously Published Work

Journals


Conferences


Miscellaneous


Electronic preprints are available on the Internet at the following URL:

http://www.elec.qmul.ac.uk/staffinfo/andrea/publications.html
Acknowledgments

First, I would like to express my deep gratitude to my supervisor Dr Andrea Cavallaro for his continuous support throughout my time at Queen Mary. His guidance and dedication have greatly contributed to not only my research, but also my development at both a personal and professional level. Research and professional ethics and practices that I have imbibed from him will be carried forward in all of my future career.

Next I would like to thank my second supervisor Prof. Alan Pearman and independent assessor Dr Rob Donnon for their constructive feedback during various stages of my research. Part of the work presented in this thesis has been done in collaboration with other colleagues and students working under my supervision. Working with them has been greatly rewarding and will be among the most memorable experiences of my Ph.D.

During my stay at Queen Mary I witnessed the growth of my research lab, the greater research group and the department and university as a whole. I had the pleasure of coming in contact with many fellow researchers from all around the world, some for a brief period and others from much longer. Most of them became, and will remain, good friends of mine. Following the merging of the Electronic Engineering and Computer Science departments towards the end of my Ph.D., I came in contact with many more interesting people who greatly contributed to both my research and social life. I would like to thank all of these unnamed people for being not only colleagues but also friends.

Last, but most certainly not least, I express my immense gratefulness towards my parents and family to whom I dedicate all my achievements for their continuous support and motivation. My living away from home, while being hard on my parents, never affected their undying support whenever I was in need. My sister, while not being the most vociferous person, has been a great support in her own unique way. My extended family here in London were always there for me and I never felt lacking of family warmth even while being away from my home town. A special thanks goes out to my aunt Madhuri, to whom I will always remain indebted for her support during the time of my Ph.D.
To my loving parents and family
**List of abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DMM</td>
<td>3D Morphable Model</td>
</tr>
<tr>
<td>AAM</td>
<td>Active Appearance Model</td>
</tr>
<tr>
<td>ADP</td>
<td>Average Dynamic Precision</td>
</tr>
<tr>
<td>ASM</td>
<td>Active Shape Model</td>
</tr>
<tr>
<td>CMC</td>
<td>Cumulative Match Curve</td>
</tr>
<tr>
<td>EER</td>
<td>Equal Error Rate</td>
</tr>
<tr>
<td>EGI</td>
<td>Extended Gaussian Image</td>
</tr>
<tr>
<td>GPA</td>
<td>Generalised Procrustes Analysis</td>
</tr>
<tr>
<td>ICP</td>
<td>Iterative Closest Point algorithm</td>
</tr>
<tr>
<td>ILD</td>
<td>Inter-Landmark Distance</td>
</tr>
<tr>
<td>LBP</td>
<td>Local Binary Pattern</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>LFA</td>
<td>Local Feature Analysis</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PDM</td>
<td>Point Distribution Model</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>S-ICP</td>
<td>Selective Iterative Closest Point Algorithm</td>
</tr>
<tr>
<td>SLDA</td>
<td>Subspace Linear Discriminant Analysis</td>
</tr>
<tr>
<td>SPS</td>
<td>Surface Point Signature</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TPS</td>
<td>Thin Plate Spline</td>
</tr>
</tbody>
</table>
Glossary

$H$ Mean curvature.

$K$ Gaussian curvature.

$L$ Number of training shape vectors.

$N$ Number of 3D points in a shape vector.

$\Omega$ Point distribution model.

$\Psi$ 3D face mesh.

$\Phi$ 3D face dataset.

$\omega$ Shape vector.

$\phi$ Eigen vectors of the Point Distribution Model.

$b$ Shape parameters of the Point Distribution Model.

$\gamma(v_i)$ Curvedness index at vertex $v_i$.

$\partial_s(\Psi, \Psi')$ Symmetric Hausdorff distance between 3D face meshes $\Psi$ and $\Psi'$.

$\rho(v_i)$ Shape index at vertex $v_i$. 
Chapter 1

Introduction

A man finds room in the few square inches of his face for the traits of all his ancestors; for the expression of all his history, and his wants.

*Ralph Waldo Emerson, Conduct of Life*

1.1 Imaging faces

In a world of seven billion humans, each human face is unique. It is a small part of us (roughly 1/10th the vertical height of the body) that reveals our personality, genetic and cultural identity. We have evolved to express, and read in others, almost all our emotions and needs using our face. There are seven recognised prototypic facial expression, i.e., anger, fear, happiness, sadness, disgust, surprise and contempt, and it is estimated that we can generate about 7,000 discrete expressions [1]. While recognising other humans and their expressions comes naturally for the average human, this task is far from trivial for a machine. Computer vision is in a constant struggle to match (and perhaps someday outperform) the vision and cognitive power of humans.

Since the production of the first permanent photograph in the early 19th century [2] began the human fascination for capturing images of themselves. Through the years photographic images have been used mainly for art and capturing personal portraits, and most of the development has been in improving the technology behind capturing images and making it cheaper and more accessible to everyone. The advent of digital imaging has resulted in an exponential growth in the use and our dependence on images for our everyday life.

Digital images are currently not only used at a personal level, but also for most scientific studies, medical diagnoses and research, and surveillance and security, to name
just a few applications. Even though images are but a two-dimensional projection of the three-dimensional world we live in, in most cases the human brain quite easily constructs depth information from them giving us a clear understanding of the image. However, the loss of depth information has always been a major hurdle in understanding the true 3D world using computer vision.

Recent advances in imaging technology has helped in overcoming this problem by capturing 3D data. 3D facial surface data in particular are increasingly used in animation, medical and security applications. In medical and dental applications, this technology is widely employed for the assistance of the clinical diagnoses and the treatment planning through facial asymmetry analysis in orthogenetic operation; for the assessment of a treatment regime when a new operation process is employed; and for the monitoring and study of facial growth when a specified orthodontic treatment is being applied [3–7]. In security applications, 3D face scans are being used for person recognition [8–10] as they overcome the limitations presented by pose and illumination variation, and also the use of cosmetics.

In this thesis we focus on the analysis of face scans towards 3D face recognition. The term recognition is generally used to refer to both identification and verification. However in literature, recognition mostly refers to identification and authentication is used instead of verification. In identification the task is to find the identity of a person from a database. In this case the biometric information of the person is matched against all the information in the database. The closest match found in the database, in terms of similarity, is used to assign an identity to the person. Whereas in authentication the person’s identity is known and the task is to confirm if the person is who he claims to be [11,12]. So in this case the task is to accept or reject the claimed identity on the basis of the similarity estimated between the information at hand and the information of that person in the database.

Three-dimensional face scans are usually captured by scanners through the use of a constant structured-light source, making them invariant to environmental changes in illumination. A single 3D scan captures the entire face structure and allows for accurate pose normalisation. In addition, due to the nature of the 3D data, they are better suited for describing surface-based events. However, one of the biggest challenges posed by the use of 3D geometrical information is its sensitivity to large local deformations (due to facial expressions or medical conditions) and outliers (due to the acquisition process). Deformations due to expressions and medical conditions bring about large changes in the geometry of the scan. Also, the type of noise and artifacts present in 3D data, such as spikes and holes, are uncommon with 2D images and requires a pre-processing stage
specific to the scanner.

1.2 Main contributions

In this work we address the 3D face analysis problem by answering the following questions:

- Can we define a robust approach for detecting salient feature points in a mesh representing a 3D face?
- Can we devise a 3D facial scan registration approach that is robust to pose and expressions?
- Can we uniquely represent a 3D facial scan of an individual with a compact signature?

The thesis focuses on the registration of 3D faces to make the process invariant to changes due to expressions or medical abnormalities, and a recognition approach that allows for a compact face signature. The main contributions are as follows:

1.1. Feature detection: We present and extensively evaluate an approach for the detection of landmarks and the segmentation of regions on 3D facial scans \cite{J1,C3} that, unlike existing approaches \cite{13,14}, is independent of texture, pose or orientation information. The approach is based on a 3D Point Distribution Model (PDM) that is fitted to the region of interest using candidate vertices extracted from curvature-based low-level feature maps.

1.2. Face registration: We propose an expression-invariant face registration technique for the accurate matching of 3D faces with missing data and deformations through the use of prior anthropometric knowledge and localized landmarks \cite{C2}. Unlike the traditional Iterative Closest Point (ICP) algorithm \cite{15}, our approach is independent of the initial orientation of the faces and robust to large facial deformations.

1.3. Face recognition: We present a novel approach to 3D face recognition using compact face signatures based on automatically selected 3D landmarks \cite{C1} and represent the face geometry with inter-landmark distances within selected regions of interest. This approach allows us to achieve robustness to expression variations and because of its compactness the face signature can be stored on RFIDs and 2D barcodes.
1.3 Outline of the thesis

The thesis is organised as follows. Chapter 2 discusses the 3D acquisition process, the different ways 3D data can be represented, the features that can be used to represent facial characteristics and the related work in 3D face recognition. In Chapter 3 we motivate and describe our approach for finding features on 3D faces (landmarks and regions), the model fitting strategy and the face signatures we propose for characterising and identifying faces. Chapter 4 presents the evaluation of the proposed approach and in particular the person identification results using the face signatures and classification approach presented in the thesis. Finally, in Chapter 5 we summarise the achievements of this thesis and we discuss possible extensions of this work.
Chapter 2

State of the art

2.1 Introduction

The 3D shape of a human face can be represented as a non-rigid free-form surface. Currently, most of the static 3D sensors acquire data from only the visible part of the human facial surface from the viewpoint of the camera lenses. It is also possible to acquire full 3D model of a human head from multi-view stereo systems or using rotating tables during the scanning process. However, multi-view or rotating sensors are not practical for identification scenarios. Therefore, static 3D sensors such as laser scanners or structured light based stereo systems produce, what is popularly known as, 2.5D surface data. 2.5D surface is usually defined as having at most one z-depth measurement from a given (x,y) coordinate. It is possible to generate full 3D face models by combining several 2.5D images. In the 3D face recognition literature, the term 3D is commonly used to denote 2.5D data. A discussion on the 3D acquisition process and a brief description on some of the commercially available 3D scanners is provided in Sec. 2.2. The 3D data can be further represented in different ways, depending on the acquisition process and intended use. Three commonly used representations are the point cloud, wireframe and polygonal mesh methods. We highlight the main characteristics of these three representations in Sec. 2.3.

The extraction of facial features is often an important stage in face recognition systems. These features are usually representative of the face geometry and in many cases are used to perform the matching of the faces. In cases where they are not directly used to perform the matching, they are still instrumental in the pre-processing stages (such as face detection and normalisation). We discuss commonly used features extracted from 3D faces in Sec. 2.4 of this chapter.
Existing approaches for 3D face recognition can be broadly classified into two main categories, namely: (i) holistic, and (ii) feature-based. Holistic approaches match faces through direct surface comparison or by projecting the face data into appropriate feature spaces. Feature-based approaches on the other hand compare extracted features that represent the geometry of the face. A description of these two categories of recognition approaches and related literature is provided in Sec. 2.5.1 and Sec. 2.5.2. A third category in existing literature is that of multi-modal methods. These are recent methods that aim at the fusion of 2D texture and 3D shape information. While in this thesis we do not work with multimodal data, we provide a discussion on existing work in this area in Sec. 2.5.3 for completeness.

2.2 3D acquisition

Dense surface acquisition is one of the most challenging tasks in computer vision. A 3D imaging system is required to capture three-dimensional coordinates from the visible object surface in a scene. Research over the last two decades has led to a number of high speed and high precision 3D sensors.

The triangulation based sensors observe the object from at least two different angles. In order to obtain three-dimensional measurements, point correspondences have to be established, allowing a 3-D shape to be reconstructed in a way that is analogous to the way the human eye works. Triangulation sensors can be further classified into active and passive. Active triangulation systems illuminate the scene rather than relying on natural or uncontrolled lighting. A stereo camera is the prime example of passive optical triangulation, where two or more cameras are used to view a scene. Determining the correspondence between left and right views for a binocular system by means of image matching is however a difficult process. For optimal 3-D reconstruction of objects, passive stereo vision techniques depend on texture information on surfaces [16].

One of the most common forms of active range sensing is optical triangulation. The fundamental principle is illustrated in Fig. 2.1 [17]. A focused beam of light illuminates a tiny spot on the surface of an object. Typical surfaces scatter this light in many directions and a camera records an image of the spot. We can compute the center pixel of the spot and trace a line of sight through that pixel until it intersects the illumination beam at a point on the surface of the object. The camera center of the lens lies at \((0, 0, 0)\). The point \((x, y, z)\) is projected onto the image plane at pixel \((u, v)\), such that \(\frac{u}{x} = \frac{f}{z}\) and \(\frac{v}{y} = \frac{f}{z}\), where \(f\) is the focal length of the camera. Let \(\theta\) be the projection angle. The \((x, y, z)\)
coordinates of the surface point can be computed as:

\[
x = \frac{b}{(f \cot \theta - u)} \cdot u, \tag{2.1}
\]

\[
x = \frac{b}{(f \cot \theta - u)} \cdot v, \tag{2.2}
\]

\[
x = \frac{b}{(f \cot \theta - u)} \cdot f, \tag{2.3}
\]

To scan the entire surface instead of one point, the beam can be spread into a plane of laser light. The light will cast a stripe onto the surface of the object, which is then imaged by a conventional video camera. Each camera scanline can be treated separately, the center of the imaged light can be found, the line of sight can be intersected with the laser plane. Thus, each image gives us a range profile (one point per scanline), and by sweeping the light over the surface of the object, we can capture its shape.

To produce well-behaved surfaces and to speed up the evaluation steps, active triangulation systems project specific light patterns onto the object. The light patterns are distorted by the object surface and these distorted patterns are observed by at least one camera and then used to reconstruct the object’s surface. Particularly useful is a set of techniques, known as coded light techniques, that project a set of well-defined binary patterns [18]. Within this sequence time-encoded correspondence information is included.
2.2.1 3D scanners

Initial 3D data acquisition methods suffered from drawbacks such as high cost and lack of portability [9]. However, recently many commercial providers have made affordable and easy to use scanners available for a variety of applications. This section provides a brief description of some of the 3D scanners available in the market.

Konica Minolta

Konica Minolta [19] has a range of 3D scanners that are designed for different applications. Their range 5 and 7 scanners are used mainly for reverse engineering and inspection in industrial applications. They can be used in all stages of manufacture from development to prototyping. The Konica Minolta VI-9i range is designed to be used for quality inspection and reverse engineering of industrial products such as automotive parts, cast and moulded products, mock-ups, etc. Finally, their VI-910 range has been designed particularly for capture and measurements of the human body towards research and development, design, game software and the apparel industry, to name a few applications.

Polhemus

Polhemus [20] provides a very versatile hand-held 3D scanner called FastSCAN\textsuperscript{TM}. FastSCAN is lightweight and portable and can instantly acquire three dimensional surface images when the handheld laser scanning wand is swept over an object. The scanner can be used to digitise a human shape for animation, multimedia, custom apparel design, biomedical research and forensics. It can scan a complicated surface with convolutions that would normally be obscured from a conventional turntable-based laser scanner.

3Dshape

3Dshape [21] is a German company providing a number of scanners for both industrial and medical applications. Two of their products in particular, FaceSCAN\textsuperscript{3D} and BodySCAN\textsuperscript{3D}, are designed specifically for capture and measurements of the human body. The FaceSCAN\textsuperscript{3D} sensor was specially developed for the measurement of faces, in the field of maxillofacial surgery for example. It maps the entire face quickly and precisely from ear to ear. BodySCAN\textsuperscript{3D} delivers 3D images of body parts such as the breast, back or thigh for 3D visualisation and for comparison of pre- and postoperative conditions in clinical trials, as well as for individual consultation.
Cyberware

Cyberware [22] provides two scanners, namely *Model PX* and *Model WBX*. The PX scanner is designed to offer maximum coverage scanning of the human head and face. The system is designed for applications that require increased scan coverage of the top of the head and under cut areas of the chin. Projects which require high quality data from the eye and mouth area can also benefit from the fine sample pitch of the new Head and Face Color 3D Scanner. The system incorporates a rugged, self-contained optical range-finding system, whose dynamic range accommodates varying lighting conditions and surface properties. The scanning process captures an array of digitised points, with each point represented by X, Y, and Z coordinates for shape and 24-bit RGB values for colour.

The WBX scanner is a whole body scanner that is fast, accurate and easy to use. It takes hundreds of thousands of measurements of the human body in less than 20 seconds. Four scanheads collect high-speed 3D measurements every 2 mm from head to toe to create an accurate 3D data set. In this case as well, the scanning process captures an array of digitised points, with each point represented by X, Y, and Z coordinates for shape and 24-bit RGB values for colour. The WBX motion system is designed for heavy use applications and can reliably scan at rates up to one subject every 30 seconds.

Vitronic

Vitronic [23] offers systems for measuring the human body for scientific as well as commercial purposes. In case of all 3D complete body scan systems, irrespective of the application area and its complexity, 360 photos are undertaken in a short measuring time. Along with the body measurements, the human body surface is also registered. The acquired data is then available in registered form and can be directly further processed or archived. The Vitronic head scanner is designed keeping in mind security applications. The scanner captures head and busts of people and can be used in a mobile form or in a permanent location.

2.2.2 Softwares

Almost all of the commercially available 3D scanners are provided along with coupled software. The software assists the 3D capture process through an interface, and in addition, provides tools for the analysis of the scans. These tools allow conversion between data representations and file formats, and include functionalities such as smoothing,
decimation, clipping, registration and geometric measurements, to name a few. Among these, smoothing and decimation are two important 3D pre-processing techniques that are used extensively in 3D face recognition systems. Smoothing of the 3D data [24, 25] is necessary to reduce the noise and spikes that are present in the output of almost all available scanners. The data also usually contain a number of holes, due to occlusions during capture, and hole-filling [26, 27] is required. Noisy spikes and holes are especially common in scans that are captured in a non-cooperative practical environment. Figure 2.2 highlights the problem of spikes and holes prevalent in captured 3D data. Surface decimation [28] is another pre-processing step that is usually applied to reduce the resolution of dense surface scans, to make them easier to process.

In addition to the mentioned pre-processing techniques, many available software provide extended 3D object analysis software that allow registration and geometric measurements of the scans. Registration is the process where two or more scans are fitted in exact alignment with one another [15]. These functionalities in commercial software provide solutions for industrial applications for matching scans with reference data and production-line inspection. Table 2.1 summarises the capabilities of some of these available softwares.

### 2.3 3D representations

An important factor when dealing with 3D data is taking into account how the data is represented. A point cloud, wire-frame mesh and shaded polygonal model are
Table 2.1: Comparison of software capabilities provided with commercial 3D scanners

common representations for rendering 3D face data. The 3D shape of the face is often also sensed in combination with a 2D intensity image. In this case, the 2D image can be thought of as a texture map overlaid on the 3D shape. This section describes these 3 representations, their characteristics and applications.

2.3.1 Point clouds

Many 3D acquisition systems provide 3D point clouds as raw data, possibly coupled with 2D texture information. Thus, for many 3D face recognition systems, point cloud or point set data is the default input data representation [29–31]. The popularity of the point cloud representation scheme is due to i) its generality: almost every 3D acquisition device produces (x,y,z) coordinates without any higher-level information such as connectivity, and ii) its simplicity: point coordinates, if sampled with good accuracy, are simple and sufficient to represent a complex surface. On the other hand, point cloud-based representation has several drawbacks, most notably: i) since there is no connectivity information, search for nearest points may be cumbersome, and efficient search algorithms usually require advanced data structures such as kd-trees, and ii) storage requirements are high.

Point clouds themselves are generally not directly used in most 3D applications, and therefore are usually converted to polygonal mesh models. Applications where point clouds are directly used generally rely on first registering a mesh model with the point cloud and subsequently using the mesh for processing [10]. In industrial metrology or inspection, the point cloud of a manufactured part can be aligned to a CAD model (or even another point cloud), and compared to check for differences. In medical imaging, point clouds can be used to represent volumetric data [32]. Figure 2.3 shows a sample face point cloud.
2.3.2 Wireframe models

In this scheme models are represented as a collection of vertices and edges (edges can be line segments or curves) (see Fig. 2.4). Due to its simple structure they are relatively easy to process, render and work with especially in real-time systems. A wire frame model allows visualisation of the underlying design structure of a 3D model. However the use of wireframe models can result in ambiguous situations thereby limiting their use. Fig. 2.5 shows an illustration of its ambiguous nature. The solid shown consists of 16 vertices and 32 edges. It is evident that the cube in the middle represents a hollow opening but the direction of the opening is not evident from the representation data. The three different possibilities depicted are mere rotations of each other but knowing the correct orientation is vital especially if the object is part of a bigger scene of objects [33].

Wireframe models are comparatively simple and fast to calculate and so they are often used when a high screen frame rate is needed. They are the popular choice for
rendering a particularly complex 3D model or in real-time systems [34]. If greater detail is required, surface textures can be overlaid on the wireframe model.

2.3.3 Polygonal models

Polygonal mesh representation is probably the most common scheme in use today since it combines the versatility of the wireframe model while making up for its shortcomings [35]. This scheme is just an extension of the wireframe model where the object is represented with a collection of vertices and polygons. Since polygons are planar, for a generic curved surface an approximation of the shape has to be made. To increase the resolution of the approximation, more polygons will have to be used. Objects created with polygon meshes must store different types of elements. These include vertices, edges, faces, polygons and surfaces. In many applications, only vertices, edges and either faces or polygons are stored.

In most cases surface shading is used in the display of polygonal mesh models [36]. Shading is the moment in the rendering process when visible surfaces are assigned a shading value. The value is calculated based on the relationship between the surface normals and the light sources. There are number of shading techniques, such as faceted, smooth and specular shading, to name a few. Faceted shading assigns a single and constant shading value to each polygon on the surface according to the angle of its normal in relation to the light source. Smooth surface shading assigns a continous shading value that blends throughout the visible polygons on the surface. Specular shading creates surfaces with highlights that are found in reflective surfaces.

The most commonly used polygon is the triangle and many algorithms have been
Figure 2.6: A sample polygonal mesh representation of a scanned face. The shown face has been rendered with smooth shading.

designed for the efficient division of any given surface into a set of finite triangles, a process called triangulation. Triangulation techniques include the commonly used Delaunay triangulation [37] and more recent techniques such as Marching triangles [38], Marching cubes [39], and the Ball-Pivoting algorithm [40]. An example of a polygonal model with smooth shading is shown in Fig. 2.6.

2.4 Features

According to the representation schemes used, several feature extraction methods can be applied to construct feature spaces. For example, statistical dimensionality reduction techniques such as Principal Component Analysis (PCA) [41] or Linear Discriminant Analysis (LDA) [42] can be applied to depth images, or curvature-based features can be extracted from surface scans. Surface curvatures are heavily used in segmentation and facial feature localisation, but their usage for the representation and identification of human faces is very rare [43]. Statistical models such as Active Shape Models (ASM) [44] are extensively used as shape descriptors for 2D face analysis. The use of these shape models has recently been adapted for 3D data [45, 46] and reconstruction of 3D meshes [47, 48]. Local indexing methods [49–51] form another category of shape descriptors that provide a very concise representation of 3D object geometry, and are particularly useful when dealing with large databases of high resolution data. This section describes these different features that can be extracted from 3D faces and that can be used to represent the geometry of the face.
2.4.1 Surface curvature

Curvature can be defined to be the degree by which a geometric object deviates from being completely flat [52]. The word flat might have very different meanings depending on the objects considered (for a curve it is a straight line and for a surface it is a Euclidean plane). The main curvatures that are used in the study of surfaces in 3D geometry are the Principal Curvatures, Mean Curvature and Gaussian Curvature.

If \( \hat{n} \) is the unit tangent vector of a regular surface (see Fig. 2.7), then the normal curvature is

\[
K(\hat{n}) = S(\hat{n}) \cdot \hat{n},
\]

(2.4)

where \( S \) is the shape operator. The shape operator is the negative derivative of the unit normal vector field of the surface.

The maximum and minimum of the normal curvature gives the Principal curvatures of the surface, \( \kappa_1 \) and \( \kappa_2 \). The Gaussian curvature \( (K) \) and Mean curvature \( (H) \) are related to \( \kappa_1 \) and \( \kappa_2 \) by,

\[
K = \kappa_1 \cdot \kappa_2,
\]

(2.5)

\[
H = \frac{1}{2}(\kappa_1 + \kappa_2).
\]

(2.6)

Surfaces that are initially flat will have a minimum and maximum curvature of zero and consequently a Gaussian curvature of zero. It should be noted that neither the
Guassian or Mean curvature is scale invariant. Also, due to the discontinuity in digital data, curvatures can only be estimated. Scale invariance is a much desired feature property and has been previously introduced in other geometric invariants. For example, shape characterisation based on moments was presented in [53]. In [54] invariance concepts contained in affine differential geometry were generalised and extended. Multi-scale properties were presented in [55], where feature points were combined using a geometric hashing algorithm in a way that is scaling invariant. For 3D curvature features, a scale invariant measure was proposed in [56], based on a curvature ratio \( \kappa_3 \) defined as 
\[
\kappa_3 = \min(|\kappa_1|, |\kappa_2|) / \max(|\kappa_1|, |\kappa_2|).
\]
The proposed curvature measure was proved to be scaling invariant for any smooth 3D set.

Curvature-related descriptors are attractive since they are invariant to rotations and, therefore, they are frequently used in segmenting 3D surfaces [57]. For face analysis, regions with the highest curvature values (on a smoothed surface) correspond to prominent features like the nose and the inside corners of the eyes and hence aid in their detection. Existing papers [43, 58] that have evaluated the discrimination power of facial features based on surface curvatures, suggest that concave and convex regions with high curvature values provide compact descriptions of face surfaces and can be used as robust and stable facial features for recognition. Figure 2.8 illustrates an example 3D face showing calculated principal, mean and Gaussian curvature values.

### 2.4.2 Extended Guassian Images

By translating the surface normals of an object to a common point, a representation of the distribution of surface orientation is formed, called the Extended Gaussian Image (EGI). A Gaussian Image is the mapping of normal vector information onto a unit sphere, such that tails of the vectors lie at the centre of the sphere and the heads on the surface of the sphere appropriate to the orientation. The process is extended by assigning a weight to each point on the Gaussian sphere equal to the area of the surface having the given normal. Weights are represented by vectors parallel to the surface normals, with length equal to the weight [59].

The discrete representation of the EGI is a 3D orientation histogram. In order to construct and view the 3D orientation histogram used for symmetry detection, it is necessary to tessellate the surface of a unit sphere. Each point on the Gaussian sphere lies in a particular bin of the histogram. The bins used are the facets of the tessellated sphere. An example of a tessellated EGI corresponding to a model of a human head is shown in Fig. 2.9.
Figure 2.8: The main types of curvature used in 3D face analysis: (a) Minimum Principal curvature; (b) Maximum Principal curvature; (c) Mean curvature; and (d) Gaussian curvature.

Figure 2.9: Example of a Tessellated EGI: (a) wire-frame model of a human head; (b) the corresponding EGI [59].
As mentioned earlier, the advantages 3D data has to offer such as viewpoint invariance and lighting and pose invariance is due to the fact that curvature and shape possess these characteristics. Since curvature is a second order derivative its determination from a high resolution 3D scan is usually a computationally expensive procedure. The use of EGIs provides an efficient mapping of object curvature information where the model can be stored as surface normal vector histogram. This mapping allows easy retrieval of the surface normals for later processing and easy identification of both the identity and the orientation of objects. The search space for retrieval or symmetry evaluation can be reduced considerably depending on the tessellation resolution of the unit sphere. However EGIs posses the disadvantage of being able to uniquely describe only convex objects, with infinite number of non-convex objects being capable of having the same EGI. Since in human faces, convex regions are believed to change shape less than other regions in response to changes in facial expression, the use of EGIs is justified to some extent in providing an expression invariant representation.

### 2.4.3 Facial profiles

Facial profiles are defined as 2D curves extracted from the facial surface [60–62]. Automated profile-based face analysis techniques can be grouped into three main approaches [63]: tangency-based, where lines tangent to the profile are used; curvature based, where the curvature along the profile is used; and template-based where a template of mapped onto a profile for extraction of further features.

For profile-based schemes the accurate extraction of the profile is the main hurdle
to be overcome. One approach is to first find the symmetry plane that cuts the face into two similar parts [64]. Then the nose tip can be used for the extraction of profiles. Figure 2.10 shows example vertical profiles of a sample face taken from the inner and outer eye corners and the nose-tip. Planes can be used to intersect the face surface, following pre-alignment, for the extraction of profiles [65]. A number of heuristics are usually used to aid the process. Once the profiles are detected there are number of ways to match them for face recognition. Usually a matching coefficient is used between corresponding profile points that can be based on curvature [64], area [60] or distances [66].

### 2.4.4 Statistical shape descriptors

Statistical models such as Active Shape Models (ASM) [44], Active Appearance Models (AAM) [67] and 3D Morphable Models (3DMM) are extensively used for 2D face analysis [68]. The shape model used in these approaches, called Point Distribution Model (PDM), aims to perform image interpretation using prior statistical knowledge of the shape to be found. In ASMs the shape of a face is represented by a vector consisting of the positions of the landmarks, \( S = (x_1, y_1, ..., x_n, y_n) \), where \((x_j, y_j)\) denotes the 2D image coordinate of the \( j^{th} \) landmark point. All shape vectors of faces are normalised into a common coordinate system. Principal Component Analysis (PCA) is then applied to this set of shape vectors to construct the face shape model, denoted as: \( S = \overline{S} + P_S B_S \), where \( S \) is a shape vector, \( \overline{S} \) is the mean shape, \( P_S \) is a set of orthogonal modes of shape variation, and \( B_S \) is a set of shape parameters.

In AAMs, in order to construct the appearance model, the example image is warped to make the control points match the mean shape. Then the warped image region covered by the mean shape is sampled to extract the gray level intensity (texture) information. Similar to the shape model construction, a vector representation is generated, \( G = (I_1, ..., I_m)^T \), where \( I_j \) denotes the intensity of the sampled pixel in the warped image. PCA is also applied to construct a linear model \( G = \overline{G} + P_G B_G \), where \( \overline{G} \) is the mean appearance vector, \( P_G \) is a set of orthogonal modes of gray-level variation, and \( B_G \) is a set of gray-level model parameters.

Thus, the shape and texture of any example face can be summarised by the vectors \( B_S \) and \( B_G \). The combined model is the concatenated version of \( B_S \) and \( B_G \), denoted as follows:

\[
B = \frac{W_S B_S}{B_G} = \frac{W_S P_S^T (S - \overline{S})}{P_G^T (G - \overline{G})},
\]  
(2.7)
where $W_S$ is a diagonal matrix of weights for each shape parameter, as a normalisation factor, allowing for the difference in units between the shape and gray scale models. PCA is applied to vector $B$ also, $B = QC$, where $C$ is a vector of parameters for the combined model.

Given a test image and the face model, the metric used to measure the match quality between the model and image is $\Delta = |\delta I|^2$, where $\delta I$ is the vector of intensity differences between the given image and the synthesised image generated by the model tuned by the model parameters, called the residual. The AAM fitting seeks the optimal set of model parameters that best describes the given image. Figure 2.11 shows the AAM and ASM fitting process on a sample face.

3DMMs are closely related to AAMs where a 3D model is used to estimate the 3D parameters in a 2D image and to recognise and segment faces. To handle the extreme image variations induced by shape and texture parameters, a common approach used in 3DMMs, is to use generative image models. The general strategy in existing techniques is to fit the generative model to a test image, thereby parameterising it in terms of the model. The morphable face model is based on a vector space representation of faces [70]. The coordinate and texture values of all the $n$ vertices of a reference face are concatenated to form shape and texture vectors

$$S_0 = (x_1, y_1, z_1, ..., x_n, y_n, z_n)^T, \quad (2.8)$$

$$T_0 = (R_1, G_1, B_1, ..., R_n, G_n, B_n)^T. \quad (2.9)$$
Chapter 2: State of the art

Figure 2.12: In 3DMMs, the fitting process finds shape and texture coefficients $\alpha$ and $\beta$ such that rendering $R_p$ produces an image $I_{\text{model}}$ that is as similar as possible to $I_{\text{input}}$ [71].

Vectors $S_i$ and $T_i$ of the subjects $i = 1...N$ in the database are formed in a common coordinate system. Convex combinations of the examples produce novel shape and texture vectors $S$ and $T$. The shape and texture information can then be combined independently:

$$S = \sum_{i=1}^{N} a_i S_i, \quad T = \sum_{i=1}^{N} b_i T_i. \quad (2.10)$$

During the process of fitting the model to a test image, not only the shape and texture coefficients $\alpha_i$ and $\beta_i$ are optimised, but also the following rendering parameters, which are concatenated into a vector $\rho$: the head orientation angles $\phi$, $\theta$ and $\gamma$, the head position $(P_x, P_y)$ in the image plane, size $s$, colour and intensity of the light sources $L$, as well as colour constant, and gain and offset of colours, shown in Fig. 2.12.

The primary goal in analysing a face is to minimise the sum of square differences over all colour channels and all pixels in the input image and the symmetric reconstruction,

$$E_I = \sum_{x,y} \|I_{\text{input}}(x,y) - I_{\text{model}}(x,y)\|^2. \quad (2.11)$$

It should be noted that the model can only deform in ways observed in the training set. If the object in an image exhibits a particular type of deformation not present in the training set, the model will not fit to it. This is true of fine deformations as well as coarse ones. However, using enough training examples can usually overcome this problem. Another drawback of the approach is the amount of labelled training examples required to build a good model, which is very time consuming to generate. As mentioned earlier, PDM-based statistical shape models have recently been adapted and used with 3D data. In [45] and [46] the model is fit to 3D data with texture, while [47] and [48] deal with construction of the 3D models.
2.4.5 Local indexing methods

Almost all indexing schemes in use for 3D data have a very similar stratagem. Invariant features for a given model are computed and conglomerated into a look-up table with references to the model. Orientation parameters are also usually encoded along with each table entry. Indexing methods are found to be very useful, when working with large model databases, than other matching schemes where models are considered separately.

Chua and Jarvis [49] present a new form of point representation for describing 3D free-form surfaces called the point signature. Neighbourhood surface structure of interest points are estimated by plotting the distance profile of a circle of points to a plane defined by that circle of points. Interest points chosen can be at points of discontinuities, and need not be restricted to smooth regions as is the case for the computation of principal curvatures using differential geometry. This method serves to describe the structural neighbourhood of a point in a more complete manner than just using the 3D coordinates of the point being invariant to rotation and translation. Recognition is achieved by matching the signatures of data points representing the sensed surface to the signatures of data points representing the model surface.

Yamany and Farag [51] present a representation scheme called Surface Point Signature (SPS) which they claim outperform the point signature scheme by providing more discriminating power. For a given patch the distance of points in the patch to the centre of gravity, and the angles made by line segments from the centre of gravity and the normal at the centre of gravity, is encoded. This information is encoded as a 2-D image corresponding to point on the surface used as centre of gravity for the patch. They state that this image is unique for this point and is independent of the object translation or orientation in space. However their paper does not show quantitative analysis of their testing nor comparative analysis to show this method outperforms the point signature scheme.

Spin-images [50] provide a 2D representation for a 3D surface. An oriented point is a point with an associated direction. An oriented point is defined on a mesh surface taking a point $p$ and the surface normal $\hat{n}$ at this point. The tangent plane is approximated as the fitting of a plane to the point $p$ and its closest neighbours, i.e. the points directly connected to $p$ and the normal is calculated of this plane. Calculus does not provide any way to determine the inward or outward direction of the normal (also due to the lack of rigour in the definition of “inward” and “outward”) but for the purpose of spin-image computation this needs to be chosen to have consistent features that allow for a meaningful
Figure 2.13: Examples of point signatures: (a) peak, (b) ridge, (c) saddle, (d) pit, (e) valley, (f) roof edge [49].

comparison; usually outward normals are used. An easy and empirical way to determine the direction of the normal is: after computing the normal of $p$ the centroid $c$ of the mesh is calculated as the point in space whose coordinates are the mean of the coordinates of all the points in the mesh; then the normal vectors to the coordinates of $p$ are summed to get the point $q$. If the distance of $q$ from the centroid $d(q, c)$ is bigger than $d(p, c)$ then the normal computed is outward, otherwise each of the components of the vector is multiplied by $-1$ to get the normal pointing outward. Thus using a plane and a normal, a partial coordinate system can be defined. Consequently, each other point $x$ in the mesh can be represented by means of two coordinates $\alpha$ and $\beta$: one is the radial coordinate i.e. the unsigned distance from the normal vector, the other is the elevation coordinate i.e. the
Chapter 2: State of the art

Figure 2.14: Spin image: $\alpha$ is the distance of $x$ from the normal in $p$, beta is the distance of $x$ from the plane tangent in $p$ [50].

Signed distance from the tangent plane.

$$\alpha_p(x) = \sqrt{\|x - p\|^2 - (\hat{n} \cdot (x - p))^2} \quad (2.12)$$

$$\beta_p(x) = \hat{n} \cdot (x - p) \quad (2.13)$$

where $\hat{n}$ is the normal in $p$. These values are then binned in a two-dimensional histogram. The parameters used in the generation of spin-images are the bin size, the image width and the support angle.

A bigger bin size wipes out small differences in point positions and so it affects the descriptiveness of the spin-image. The bin size is usually chosen as a multiple of the mesh resolution to avoid dependance on the mesh resolution. Image width is the number of rows and columns of a spin-image: a bigger width will produce a more global description but will be more likely to be affected by cluttering. Given a bin size $b$ and a width $W$ each point $x$ is binned in the bin $(i_x, j_x)$ if

$$i_x = \left\lfloor \frac{\alpha}{b} \right\rfloor \quad j_x = \left\lfloor \frac{\beta - \frac{W}{2}}{b} \right\rfloor \quad (2.14)$$

Support angle $A_s$ is used to reduce the coeffect of self-occlusion and given two points $p$ and $q$, $q$ will be in the spin-image of $p$ if

$$\arccos(n_p \cdot n_q) < A_s \quad (2.15)$$

Spin-images were first suggested in [50], where they were used for surface matching, and were subsequently used for object recongnition in cluttered scenes.
2.5 Face classification

Techniques used for classification (recognition) of 3D faces can be broadly divided into two groups: Holistic and feature-based. Holistic approaches use the entire face data and not local features directly (the features may be used for pre-processing stages such as face normalisation). These approaches can be further classified into two groups: (i) methods based on the direct comparison of selected regions or of the whole surface [9,31]; and (ii) methods based on projecting the faces onto appropriate spaces [8,72–74]. Feature-based approaches on the other hand are based on the comparison of features such as landmarks and contours which represent the geometry of the face. In addition some recent work focuses on the use of multimodal data, shape and texture, which can be classified into a third category. A description of these face recognition strategies and existing work is provided in the remainder of this section.

2.5.1 Holistic methods

Methods that fall under this category either analyse data in 3D Euclidean space directly via matching facial surfaces or profiles or project the faces into a classification space.

Surface matching

The most common approach to compare surfaces is through a rigid registration via the Iterated Closest Point (ICP) algorithm [31,75]. The ICP algorithm, proposed by Besl and McKay [15], uses the Euclidean distance between the closest points on two surfaces as a cost function to be minimised. Let S be a 3D point set \( \{\vec{s}_1, \vec{s}_2, \vec{s}_3, \ldots, \vec{s}_N\} \) and M be another 3D point set. Let \( CP(\vec{s},M) \) be the closest point in M to \( \vec{s} \), and \( T^{[0]} \) be an initial estimate of the rigid transformation. This iterative algorithm has three basic steps:

- Compute the set of correspondence, \( C = \bigcup_{i=1}^{N} \{(\vec{s}_i, CP(T^{[k-1]}(\vec{s}_i), M))\} \)
- Compute the new Euclidean transformation \( T^{[k]} \) that minimises the mean square error between point pairs
- Repeat for \( k = 1, \ldots, k_{\text{max}} \) or until convergence \[15\]

X. Lu, D. Colbry, and A.K. Jain [76,77] have extensively used the ICP algorithm in their work for 3D face recognition. In their work, the test scan is registered onto the images in the database followed by evaluation of the match. Initial coarse matching is
achieved by finding corresponding landmarks in both scans such as the eyes, nose, etc, followed by fine registration using ICP. The same authors have further improved their work to allow for a distinction between changes that could take place between two scans of the same person and two scans of different people. In their most recent paper [78] a face surface matching framework has been proposed that accommodates both rigid and non-rigid variations in the scans. The rigid registration is achieved by a modified ICP algorithm and Thin Plate Spline (TPS) model is applied to estimate the non-rigid deformation. A Support Vector Machine (SVM) is used for the classification of intra-subject deformation and inter-subject deformation.

Medioni and Waupotitsch [79] perform 3D face recognition using ICP matching of face surfaces. Their work uses a stereo-based system for the reconstruction of the 3D model from two 2D images with experiments carried out both indoors and outdoors. Experiments with seven images each from a set of 100 subjects are reported. They compare their results with the commercial 2D face recognition system Identix FaceIt and an equal error rate of better than 2% is reported.

Achermann and Bunke [80] report on a method of 3D face recognition that uses an extension of the Hausdorff distance matching. They perform registration by assuming that a face can be roughly approximated by a plane that is parallel to the focal plane of the camera. Hence a plane is fitted to the given set of data and pre-registration is achieved by rotating this plane around the z- and y-axis such that it becomes parallel to the focal plane of the camera. They report on experiments using 240 range images, 10 images of each of 24 persons, and achieve 100% recognition for some instances of the algorithm.

The ICP algorithm has the benefit of not requiring any priory semantic knowledge of the data. In general spatial matching methods do not require pre-calculations of second order measures such as curvature and hence are usually less computationally expensive. However in the case of the ICP it is vital that the two point sets are roughly pre-aligned or the algorithm can easily converge on a local minimum.

Ayyagari et.al [81] presents a method to achieve automatic registration of 3D face point sets through a criterion based on Gaussian fields. The method defines a simple energy function, which is always differentiable and convex in a large neighbourhood of the alignment parameters. This overcomes the limitation of the ICP cost function being undifferentiable allowing for the use of powerful standard optimisation techniques. The new method thus overcomes the necessity of close initialisation, which is required by Iterative Closest Point algorithm. In Table 2.5.1 we summarise and compare existing methods for 3D registration.
Table 2.2: Comparison of landmark detection and registration algorithms (Key: CM: Curvature maps; ORG: Object registration; FR: Face recognition; FRG: Face registration; FD: Face detection)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Application</th>
<th>Landmark detection</th>
<th>Registration</th>
<th>Robustness to prior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Approach</td>
<td>Thresholds</td>
<td>approach</td>
</tr>
<tr>
<td>[81]</td>
<td>FRG</td>
<td>-</td>
<td>-</td>
<td>Gaussian field criteria</td>
</tr>
<tr>
<td>[82]</td>
<td>FR</td>
<td>CM</td>
<td>Yes</td>
<td>ICP</td>
</tr>
<tr>
<td>[57]</td>
<td>FD</td>
<td>CM</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>[83]</td>
<td>ORG</td>
<td>-</td>
<td>-</td>
<td>Levenberg Marquardt algorithm</td>
</tr>
<tr>
<td>[84]</td>
<td>ORG</td>
<td>Integral volume descriptors</td>
<td>No</td>
<td>ICP</td>
</tr>
<tr>
<td>[85]</td>
<td>FR</td>
<td>CM</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>[86]</td>
<td>FR</td>
<td>CM</td>
<td>Yes</td>
<td>ICP</td>
</tr>
<tr>
<td>[87]</td>
<td>ORG</td>
<td>-</td>
<td>-</td>
<td>EGI (coarse) + ICP (fine)</td>
</tr>
<tr>
<td>[58]</td>
<td>FR</td>
<td>CM</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>[88]</td>
<td>ORG</td>
<td>Surface signatures</td>
<td>No</td>
<td>ICP</td>
</tr>
</tbody>
</table>

**Projection into classification spaces**

One of the most popular classification spaces is the Principal Component Analysis (PCA) [41] subspace. PCA is a dimensionality reduction technique which allows a high dimensional dataset to be represented in terms of a few directions of variability. It can be regarded as a multivariate procedure that aims to rotate the data such that maximum variability in the data can be projected onto the axes. Essentially, it can be regarded as a transformation of a set of correlated variables into a set of uncorrelated variables which are then ordered by approximating the data in the subspace reducing variability, and the last of these variables can be removed with minimum loss of significant data.

PCA aims to reduce the dimensionality while retaining as much information as
is desireable or possible. It computes a compact and optimal description of the data set. The first principal component is the combination of variables that explains the greatest amount of variation. The second principal component defines the next largest amount of variation and is independent to the first principal component. There can be as many possible principal components as there are variables.

Like PCA, discriminant analysis is a statistical technique to classify objects into mutually exclusive and exhaustive groups based on a set of measurable object’s features. Linear Discriminant Analysis (Fisher Analysis) [42] is most suited in the scenario where the within class frequencies are unequal and their performances have been examined on randomly generated test data. This method tends to maximise the between-class variance ratio whilst decreasing the within-class variance in any particular data set thereby ensuring that the items belonging to one class are stacked close and the classes are spread apart as far as possible. The prime difference between LDA and PCA is that PCA does more of feature classification and LDA does data classification. In PCA, the shape and location of the original datasets changes when transformed to a different space whereas LDA does not change the location but only tries to provide more class separability and draw a decision region between the given classes. LDA by increasing the between-class variance tries to define the distribution of features with in the data [42].

Bronstein et al. [8] approach 3D face analysis with the analysis of geometric invariants to allow for natural deformations due to facial expressions. They propose effective multi-modal 2D+3D recognition using eigen decomposition of flattened textures and canonical images. Variation due to facial expressions is one of the main reasons for the failure of most face recognition systems and herein a novel approach is provided to overcome this limitation. They perform testing on a 3D face database consisting of 64 children and 93 adults (115 males and 42 females) with results showing examples of correct and incorrect recognition by different algorithms, but do not report any overall quantitative performance results for any algorithm.

2.5.2 Feature-based methods

Gordon [85] points out that through curvature estimation the shape of facial features can be more accurately described as opposed to 2D images. In his work facial scans are segmented using Gaussian, Mean and Principal Curvatures into concave, convex and saddle regions. Ridge and valley lines are also determined using the Principal curvatures. Following the segmentation scalar features such as Left eye width, Right eye width, Eye separation, Total width of eyes, Nose height, Nose width, Nose depth, and Head width
are estimated and stored in a 24 x 24 (symmetric) matrix. The Euclidean distance metric is used for comparison. Experiments are reported with a test set of three views of each of eight faces, and recognition rates as high as 100% are reported. The author admits failure of the system when images with large facial expression variations are introduced.

Moreno et al. [58] perform face recognition using an approach similar to the one adopted by Gordon by segmenting facial scans using Gaussian and Mean curvatures and then creating a feature vector based on the segmented regions. A feature vector with 86 features is formed, each ranked according to their discriminating power. They report results on a dataset of 420 face meshes representing 60 different persons, with some sampling of different expressions and poses for each person. They report 78% recognition on the subset of frontal views, and 93% overall recognition.

Zhang et al. [62] approach 3D face recognition by first finding a plane of bilateral symmetry through the face and calling the intersection of this plane with the surface as Symmetry Profile. Three points are estimated on the Symmetry Profile using Mean curvature estimation and are used to transform the image into a new co-ordinate system called the Face Intrinsic Coordinate System. All database scans and test scans are aligned using this system. The symmetry profile and two transverse profiles and then used for the authentication stage. They validate their method with testing 213 face surfaces, which come from 164 individuals which they state covered a wide ethnic and age variety and variable facial expressions achieving 90% recognition rate.

Lee and Milios [89] once again make use of Mean and Gaussian curvatures to segment convex regions in facial scans following which an EGI is created for each convex region. A match between a region in a test image and an image from the database is done by correlating EGIs. A graph matching algorithm is used that utilises the correlation matrix between convex regions of the two faces and also imposes relational constraints to account for the relative spatial locations of these regions. Difficulty with EGI interpolation at regions with high curvature due to inadequate sampling rate at highly curved is acknowledged. They also admit that their matching algorithm suited only pair-wise matching but was unsuitable for finding the best match from a database.

Sun and Sherrah [59] use EGIs to devise a 3D symmetry detection algorithm where the symmetry detection problem is converted to the correlation of the Gaussian image. They observe that an object’s orientation histogram exhibits the same symmetry as the object and this fact can be utilised in detecting the planes of reflective and rotational symmetry. The largest degree of symmetry in the object scan is retrieved from the corresponding bin direction in the EGI histogram by searching near the principal axis di-
rections using correlation operations. The authors point out that the symmetry property of the orientation histogram is a necessary-but-not-sufficient condition for 3D symmetry detection. Hence there could exist cases where the orientation histogram exhibits symmetry while the object is non-symmetric and additional checks are required. However for convex objects, histogram symmetry provides a sufficient condition for object symmetry.

Tanaka et al. [43] also perform curvature-based segmentation and represent the face using EGIs. Two unit spheres are used; one for ridge lines and one for valley lines, constructed by mapping the principal curvatures and their direction at every point. Recognition is then performed using Fischer’s spherical correlation of the EGIs. Experiments are reported with a set of 37 images from a National Research Council of Canada range image dataset [37], and 100% recognition is reported.

Wang et al. [14] use Gabor filter responses in 2D and point signatures in 3D to perform multimodal face recognition. The feature points chosen in both 2D and 3D are selected from areas that are most stable to changes due to facial expressions, for example, 2D feature points are avoided from near the contours of the face and 3D feature points are avoided from near the mouth. Ten feature points in 2D and 4 feature points in 3D are used in generating the feature vector. The feature vector is comprised of the Gabor filter response of the 2D feature points and point-signatures of the 3D feature points in a weighted combination, the weights correspond to the discriminating power of each feature. For a given test facial image, the best match in the model library is identified according to similarity function or Support Vector Machine (SVM). Experimental results involving 50 persons with different facial expressions and extracted from different viewpoints were demonstrated with recognition rates exceeding 90%.

Table 2.3 summarises and compares the mentioned 3D face recognition algorithms.

### 2.5.3 Multimodal methods

Up until now, we have discussed a number of 3D face recognition approaches according to their shape representations. However it is possible to combine different matchers with the aim of increased classification rate. For this purpose, a number of systems propose to fuse different shape- or texture-based individual matchers. For instance, Tsalakanidou et al. [72] propose a classic approach where shape and texture images are coded using PCA and their scores are fused at the decision level. Their experimental findings confirms that using both of the modalities is better than using shape or texture only. In later work, Malassiotis et al. [97] present a pose correction and illumination correction scheme for
Table 2.3: Comparison of 3D face recognition algorithms (Key: PCA: Principal Component Analysis; ICA: Independent Component Analysis; ICP: Iterative Closest Point algorithm)

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Technique</th>
<th>Database</th>
<th>Classes</th>
<th>Number of training</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>[31]</td>
<td>ICP</td>
<td>MSU</td>
<td>100</td>
<td>200</td>
<td>598</td>
</tr>
<tr>
<td>[90]</td>
<td>PCA + ICA (depth-maps)</td>
<td>FSU_3D</td>
<td>37</td>
<td>185</td>
<td>37</td>
</tr>
<tr>
<td>[72]</td>
<td>PCA (depth-maps)</td>
<td>XM2VTS</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>[73]</td>
<td>PCA (depth-maps)</td>
<td>In-house</td>
<td>198</td>
<td>198</td>
<td>198</td>
</tr>
<tr>
<td>[91]</td>
<td>PCA (depth-maps + local variation)</td>
<td>3D_RMA</td>
<td>30</td>
<td>150</td>
<td>30</td>
</tr>
<tr>
<td>[92]</td>
<td>Distances between landmarks (geodesic)</td>
<td>In-house</td>
<td>105</td>
<td>1128</td>
<td>663</td>
</tr>
<tr>
<td>[85]</td>
<td>Distances between landmarks (Euclidean) + curvature values</td>
<td>In-house</td>
<td>8</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>[43]</td>
<td>Surface curvatures</td>
<td>NRCC_3D</td>
<td>37</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>[10]</td>
<td>Annotated deformable model fitting</td>
<td>FRGCv2</td>
<td>466</td>
<td>3541</td>
<td>4666</td>
</tr>
<tr>
<td>[8]</td>
<td>Bending-invariant canonical forms</td>
<td>In-house</td>
<td>30</td>
<td>65</td>
<td>30</td>
</tr>
<tr>
<td>[93]</td>
<td>Matching Ridge lines</td>
<td>GavabDB</td>
<td>61</td>
<td>-</td>
<td>427</td>
</tr>
<tr>
<td>[94]</td>
<td>PCA</td>
<td>GavabDB</td>
<td>61</td>
<td>244</td>
<td>61</td>
</tr>
<tr>
<td>[95]</td>
<td>Multi Library Wavelet Neural Networks</td>
<td>GavabDB</td>
<td>61</td>
<td>244</td>
<td>61</td>
</tr>
<tr>
<td>[96]</td>
<td>Multiview keypoint matching</td>
<td>GavabDB</td>
<td>61</td>
<td>61</td>
<td>366</td>
</tr>
</tbody>
</table>

3D face recognition. The proposed algorithm starts with locating the facial region by a statistical modeling of the head and torso points using a mixture of Gaussians assumption [98]. After detecting the head region, an automatic algorithm is used to locate the nose tip and the nose ridge line. Using the coordinates of these features, pose correction is carried out. After pose correction, illumination compensation is done by rendering a novel image illuminated from a frontal direction. Given normalised shape and texture images, an embedded hidden Markov model-based classifier produces similarity scores and these scores are fused by a weighted sum rule. Experiments carried out on two databases (each has 20 subjects) demonstrate that correction of pose and illumination increases the correct identification rates.

A similar approach which uses PCA is given in [73], where Chang et al. use
PCA-based matchers for shape (depth image) and texture modalities. The outputs of these matchers are fused by a weighted sum rule. The experimental results obtained on a database containing 198 subjects reveal that fusing modalities achieves 97% identification rate whereas individual 2D and 3D modalities have 96% and 91% identification rates, respectively.

Subspace-based representations are frequently used for the fusion. BenAbdelka-der and Griffin [99] use Local Feature Analysis (LFA) technique instead of the classical PCA to extract features from both shape and texture modalities. This classifier combines texture and shape information with the sum rule. Another interesting variant in this work is the data-level fusion. The depth image pixels are concatenated to the texture image pixels to form a single vector. LDA is then applied to the concatenated feature vectors to extract features. Authors report 100% and 98.58% accuracies for the LFA-based and LDA-based fusion methods, respectively, for a face database of 185 persons. These accuracies improve the best single modality (texture) rates by 0.24 and 1.36 percent for the LFA and LDA methods, respectively.

A prominent example of shape and texture feature fusion is presented in [100]. Wang and Chua [100] select 2D Gabor wavelet features as local descriptors for the texture modality, and use point signatures as local 3D shape descriptors. These feature-based representations are matched separately using structural Hausdorff distance, and then their similarity scores are fused at the score-level by using a weighted sum rule. The authors had previously used 3D Gabor features instead of point signatures as local shape descriptors in [101] in the same setting.

Li et al. [102] present a system which learns discriminative 2D and 3D features using the AdaBoost algorithm. Local Binary Pattern (LBP) features are first extracted from 2D texture images and 3D depth images. LBP features are local descriptors and their contribution to the identification task is learned automatically by the AdaBoost algorithm. Using AdaBoost, a number of weak learners are produced from 2D and 3D modalities. Each weak learner is responsible for a local region in the images. In the first part of their experimental results, the authors demonstrate that LBP-based features are superior to a PCA-based baseline algorithm. In the second part of the proposed algorithm, the authors use the AdaBoost algorithm to fuse combined 2D and 3D features at the feature level. Experimental results on a 3D face database which contains 2,305 images shows that AdaBoost-based learned fusion scheme obtains better identification rate than sum rule-based fusion of PCA matchers.

Kakadiaris et al. [10] present a multimodal identification system which fuses
shape, texture and infrared imagery. The 3D shape-based identification algorithm fits an annotated deformable model to a face, and computes the deformation image. The deformation image is then coded using Haar wavelets. For the thermal image modality, first a segmentation is carried out to locate skin pixels. After segmentation, a binary image indicating the presence of vessels is computed around the forehead region. This binary image thus represents the facial vasculature. The matching scores produced from shape, texture, and thermal modalities are first normalised, and then fused using product rule. 3D-shape based classifier obtains 99.3 percent rank-1 identification rate on the FRGC v1.0 database. On a second database (University of Houston face database, 88 subjects, 1-5 scans per subject, totally 356 scans, with expression variations, gallery set: 62, probe set: 223), the fused system obtains 98.22 percent rank-1 identification rate.

2.6 Summary

This chapter has reviewed the principles behind capturing three-dimensional object scans, the different ways 3D data are represented, features used to represent facial characteristics and literature of 3D face recognition approaches. A brief summary of commercially available 3D scanners was provided along with the 3D acquisition theory in Sec. 2.2. A discussion about the three commonly used 3D data representations, namely point clouds, wireframe models and polygonal mesh models, was provided in Sec. 2.3. Features used to characterise 3D faces were described in Sec. 2.4. Finally, the algorithms for face recognition from literature were categorised into holistic-based approaches and feature-based approaches and discussed in Sec. 2.5. A discussion on recent multi-modal approaches that use both shape and texture information was also presented in Sec. 2.5.

From the review of the available methods for analysis of 3D facial scans it is seen that while the use of 3D data overcomes some of the limitations of 2D data, most existing algorithms are not capable of handling variations of the human face due to facial expressions and other deformations. Approaches such as the Extended Gaussian Image cannot be used for the mapping of concave regions on the face that are most susceptible to changes due to expressions. Similarly the spatial matching methods that assume the face to be a rigid body also fail due to the same reason. There is still need for a system that can model facial characteristics while being invariant to deformations due to expressions. Section 3.5 will discuss our proposed methodology for expression-invariant 3D face recognition.

Efficient algorithms are also required for the quick detection of semantic feature
points on the face that are robust to variations in pose and orientation of the 3D scan. Most existing systems are tested on databases of scans that are pre-processed with roughly known orientation, thereby allowing for many assumptions to be made in the feature detection algorithms. While the Point Distribution Model (Sec. 2.4.4) has been adapted for 3D data, its fitting is dependent on the availability of texture information, which is not always available. The fitting of a PDM without the use of texture is still an open research problem. Hutton et al. [103] use a hybrid of the ICP and ASM fitting to achieve non-rigid registration of a dense surface model on 3D faces. While this method does not require texture, it imposes constraints on the initial orientation of the face and is not scale invariant. In Sec. 3.3.1 and Sec. 3.3.2 we describe our approach for achieving pose, orientation and scale invariant feature detection based on the fitting of a PDM without the use of texture information.
Chapter 3

Face signatures

3.1 Introduction

In this chapter we introduce and discuss our algorithms for face recognition. The localisation of specific anthropometric locations (landmarks) and regions on faces plays an important part in our algorithm, as with most existing approaches. In existing literature, landmarks are used to aid the ICP algorithm in achieving rough alignment of meshes, and by themselves provide valuable semantic information. The detection and segmentation of faces is also often an important step prior to landmark localisation and registration. Face segmentation is required when the data contains more than one person or includes other body parts.

The sections that follow describe the datasets used in our work, our Point Distribution Model (PDM) based feature detection algorithm, face detection and segmentation algorithm and our proposed face recognition approaches.

3.2 Datasets

Most of the existing 3D databases contain few 3D meshes per subject and there are not many variations among the different samples of a person. Some of the databases offer variations related to certain aspects but not with respect to others, and in general, the variation rank is reduced. Among the presented databases, very few capture images with some kind of 3D facial expression. In the work presented in this thesis, we use 2,933 polygonal face meshes from the BU-3DFE [104] database, GavabDB [105] database and an in-house dataset obtained by optical surface scanning [3]. The face meshes have a wide range of expressions with varying intensities of expressions, and varying degrees of noise.
3.2.1 GavabDB

The GavabDB database (Fig. 3.2) consists of 427 face meshes, corresponding to 7 scans of 61 subjects with different poses and facial expressions. The dataset provides a challenging collection of faces with the samples of each subject having varying degrees of noise, holes and presence of other body parts and clothing. Faces captured in GavabDB typically contain between 8K and 15K vertices and extraneous parts such as the shoulders and the neck. It is the most expression-rich and noise-prone 3D face database currently available.

1http://shapes.aim-at-shape.net/index.php
Figure 3.2: Sample scans from the GavabDB database showing varying amounts of noise, holes (missing data), presence of non-facial regions and exaggerated expressions.

The database was captured using a Minolta Vi-700 laser range scanner. The subjects, of which 45 are male and 16 are female, are all Caucasian. Each subject was scanned nine times for different poses and expressions, namely six neutral expression scans and three scans with an expression. The neutral scans include two different frontal scans, one scan while looking up \( (\approx +35^\circ) \), one scan while looking down \( (\approx -35^\circ) \), one scan from the right side \( (\approx +90^\circ) \), and one from the left side \( (\approx -90^\circ) \). The expression scans include one with a smile, one with a pronounced laugh, and an arbitrary expression freely chosen by the subject.

The expressions in this dataset vary considerably, including sticking out the tongue and strong facial distortions. Additionally it has strong artifacts due to facial hair, motion and the bad scanner quality. This dataset is typical for a non-cooperative environment. The problem with most other 3D face datasets is that they contain only limited variability. For example, some datasets contain variations in head orientation, but the variations are quite limited; others contain scans of faces with expressions, but the expressions are mild. The GavabDB dataset, in contrast, was deliberately designed with the intent of introducing considerable variability in head position, orientation, and facial
Chapter 3: Face signatures

3.2.2 BU-3DFE

The BU-3DFE database contains face meshes from 100 individuals (56 females and 44 males) with 25 face meshes per person. This is a face expressional database where the 25 meshes per person corresponds to five expressions (anger, disgust, fear, happiness, sadness and surprise) with four degrees (intensities) of expressions and one neutral expression. The database consists of a variety of ethnic/racial ancestries, including White, Black, East-Asian, Middle-east Asian, Hispanic Latino, and others. Each subject performed seven expressions. With the exception of the neutral expression, each of the six prototypic expressions (happiness, disgust, fear, angry, surprise and sadness) includes four levels of intensity. Therefore, there are 25 instant 3D expression models for each subject, resulting in a total of 2,500 3D facial expression models in the database. The 3D range data is scanned by a 3DMD static digitizer [109], which uses a random light pattern projection in the speckle projection flash environment. Associated with each expression shape model, is a corresponding facial texture image captured at two views (±45°). As a result, the database consists of 2,500 two-views texture images and 2,500 geometric shape models. The model resolution is in the range of 20,000 to 35,000 polygons, depending on the size of the face being scanned.

This database is provided with ground-truth of annotated landmarks for each face, which include anthropometric landmarks with points interpolated between them, which we used for the training and evaluation of our facial model. Figure 3.3 shows a
sample subject with the range of expressions and intensities, together with sample landmarks.

3.3 Feature detection

For the detection of facial features we use a PDM to represent the shape of the region of interest that includes the required landmarks, along with statistical information of the shape variation across the training set. This section describes the generation of the PDM and the methodology used to fit the model.

3.3.1 Model creation

In order to build the PDM $\Omega$, a training set of $L$ face scans were manually landmarked with $N$ points representing the region of interest. Each training shape is stored as an $3 \times N$ element vector $\omega$, where

$$
\omega = (x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_N, y_N, z_N). \tag{3.1}
$$

Training shapes are then aligned to the same co-ordinate frame (registered) so that global transformations are eliminated and statistical analysis is carried out only on shape variations. Generalised Procrustes Analysis (GPA) [110] is carried out to align the training shapes to their mutual mean in a least-squares sense, via similarity transformations. This minimises the sum of distance of each shape $\omega^k$ to the mean $\overline{\omega}$, i.e $D = \sum_{i=1}^{3 \times N} |\omega^k(i) - \overline{\omega}(i)|^2$, using the following steps [44]:

1. Translate each shape $\omega^k$ so that its center of gravity $c_{\omega^k}$ is at the origin:

$$
\forall \omega^k : d(c_{\omega^k}, \overline{\omega}) = 0. \tag{3.2}
$$

2. Choose one example, $\overline{\omega}_0$, as an initial estimate of the mean shape $\overline{\omega}$ and scale so that $|\overline{\omega}| = 1$ to define the default reference frame.

3. Align all the shapes with the current estimate of the mean shape, via similarity transformation.

$$
\min_{\theta, t, s} [D = \sum_{i=1}^{3 \times N} |\omega^k(i) - \overline{\omega}(i)|^2]. \tag{3.3}
$$
4. Re-estimate the mean from aligned shapes:

$$\overline{\omega} = \frac{1}{L} \sum_{k=1}^{L} \omega_k.$$  \hspace{1cm} (3.4)

5. Apply constraints on the current estimate of the mean $\overline{\omega}$ by aligning it with $\overline{\omega}_0$ and scaling so that $|\overline{\omega}| = 1$. If $|\overline{\omega}^{(e-1)} - \overline{\omega}^{(e)}| > \epsilon$, where $e$ is the iteration number and $\epsilon$ a specified tolerance, then return to 4.

Statistical shape models adopt PCA for dimensionality reduction. Using PCA the variations of the shape cloud can be estimated along the principal axes of the cloud formed by the training shapes in the $(3 \times N)$-D space, where $N$ is the number of landmarks in each sample. This results in a parameterised model, $\omega = \Upsilon(b)$, where $b$ is a vector of parameters. First the mean of the data is computed using Eq.(3.4). Next the covariance of the data is computed as

$$Z = \frac{1}{L-1} \sum_{k=1}^{L} (\omega_k - \overline{\omega})(\omega_k - \overline{\omega})^T.$$  \hspace{1cm} (3.5)

Finally, the eigenvectors, $\phi_i$, and the corresponding eigenvalues, $\lambda_i$, of $Z$ are computed and sorted so that $\lambda_i > \lambda_{i+1}$.

If $\phi$ contains the $t$ eigenvectors corresponding to the largest eigenvalues, any shape similar to the ones in the training set $s$ can be approximated, using

$$\omega \approx \overline{\omega} + \phi b$$  \hspace{1cm} (3.6)

where $\phi = (\phi_1|\phi_2|\ldots|\phi_t)$ and $b$ is a $t$ dimensional vector given by $b = \phi^T(\omega - \overline{\omega})$. The vector $b$ defines a set of parameters of a deformable model. By varying the elements of $b$ we can vary the shape of the model. The variance of the $i^{th}$ parameter, $b_i$, across the training set is given by $\lambda_i$. The mean shape is the shape that results from setting all parameters to zero.

We use the BU-3DFE database and the corresponding ground-truth for the training of our facial model. We use 48 ground-truth landmarks from the eyebrows, eyes and nose regions (Fig. 3.4) provided with the dataset and include an additional landmark at the nose-tip. Our training set used is composed of landmarks from 150 faces, corresponding to all the expressions of 6 individuals, out of the total of 2500 faces of 100 individuals. We retain 98% of the training variance, which corresponds to 45 eigen-modes ($t = 45$), and ignore 2% variation as noise [44]. By varying the first 3 parameters ($b_1$, $b_2$ and $b_3$)
Figure 3.4: Sample face from the BU-3DFE database showing the 49 landmarks used to train the facial model.

Figure 3.5: Effects of varying the first 3 parameters of the PDM: (top) $b_1$; (middle) $b_2$; (bottom) $b_3$.

separately, we can generate shape examples as shown in Fig. 3.5. It can be seen that large variations in the shape and scale are regulated by varying the first parameter alone (Fig. 3.5 (top)). The second parameter mainly affects the shape of the nose (Fig. 3.5 (middle)), while the third parameter mainly affects the orientation of the eyebrows (Fig. 3.5 (bottom)).

### 3.3.2 Model fitting

To fit the model $\Omega$ we isolate candidate vertices on a face mesh using curvature based feature maps, avoiding the need of a texture map. The inner eye and nose tip areas on a face are normally unique based on local curvature and can be robustly isolated from other vertices. The block diagram of the model fitting algorithm is shown in Fig. 3.6.
Isolation of candidate vertices

In order to characterise the curvature property of each vertex on the face mesh, two features maps are computed, namely the \textit{shape index} and the \textit{curvedness index} \cite{111}. These features maps are derived based on the principal curvature values, $\kappa_1(.)$ and $\kappa_2(.)$, at all the vertices of the mesh. The shape index, $\rho$, at a vertex $v_i$, is defined as

$$\rho(v_i) = \frac{1}{2} - \frac{1}{\pi} \tan^{-1} \left( \frac{\kappa_1(v_i) + \kappa_2(v_i)}{\kappa_1(v_i) - \kappa_2(v_i)} \right),$$

(3.7)

where $\kappa_1(v_i) \geq \kappa_2(v_i)$; $\rho(.) \in [0, 1]$. The feature map generated by $\rho(.)$ can describe subtle shape variations from concave to convex thus providing a continuous scale between salient shapes. However, $\rho(.)$ does not give an indication of the scale of curvature present in the shapes. For this reason, an additional feature is introduced, the curvedness of a surface. The curvedness of a surface, $\gamma(.)$, at a vertex $v_i$, is defined as

$$\gamma(v_i) = \frac{\sqrt{\kappa_1^2(v_i) + \kappa_2^2(v_i)}}{2}.$$  

(3.8)

The low-level feature maps are computed after Laplacian smoothing that reduce outliers arising from the scanning process. A comparison between feature maps generated with a smoothed and non-smoothed surface scan is shown in Fig. 3.7 (left-middle).

To reduce the computational overhead through the reduction of outlier candidate vertices, the original mesh is first decimated \cite{28}. Then the feature maps are averaged across vertex neighbors according to

$$\hat{\rho}(v_i) = \frac{1}{P} \sum_{p \in P(v_i)} \rho(v_p),$$

(3.9)
Figure 3.7: Comparison between feature maps generated on: (left) original mesh (83K vertices); (middle) smoothed mesh; (right) decimated mesh (12K vertices).

\[ \tilde{\gamma}(v_i) = \frac{1}{|P(v_i)|} \sum_{p \in P(v_i)} \gamma(v_p), \] (3.10)

where \( P(v_i) \) is the set of \( P \) neighbouring vertices of \( v_i \).

If \( \tilde{\gamma}(.) > \gamma_s \), then \( v_i \) is in a salient high-curvature region. The condition \( \tilde{\rho}(.) < \rho_c \) identifies concave regions; while \( \tilde{\rho}(.) > \rho_n \) identifies convex regions. We can therefore relax thresholds to segregate candidate inner eye vertices from the nose tip ones. The thresholds \( \gamma_s = 0.08, \rho_c = 0.25 \) and \( \rho_n = 0.85 \) were found to be adequate for the entire database.

Second order neighbourhoods for feature averaging and a decimation of 80% was also used. Note that decimation needs to be done after the extraction of the feature maps, otherwise the resulting features would not characterise the original surface (Fig. 3.7 (right)). Likewise, the neighbourhood averaging of the feature maps is done post decimation. If it is done before decimation, the consistency of features in a neighbourhood
Figure 3.8: Effect of averaging and decimation on candidate vertex (red - eye; green - nose tip) detection: (a) without averaging and decimation, (b) with averaging and without decimation, (c) with decimation and without averaging, (d) with averaging and then decimation, (e) with decimation and then averaging.

would remain and outlier candidate vertices would not be eliminated. Note that the smoothed and decimated mesh is only used for the isolation of the candidate vertices, whereas the original mesh is used for the detection of the landmarks. Examples of scans with candidate vertices isolated are shown in Fig 3.8; regions in green are candidate nose tip vertices and regions in red are candidate eye vertices.

Global fitting

The PDM $\Omega$ is fitted onto a new mesh $\Psi$ by performing similarity transformations of the model, estimated using three control points of the mean shape $\overline{\Omega}$, which are the inner eye points ($\omega_r$ and $\omega_l$) and the nose tip point ($\omega_n$), with $\{\omega_r, \omega_l, \omega_n\} \in \overline{\Omega}$. A further reduction in outlier candidate combinations is performed at this stage by checking the triangle formed by each combination of 2 inner eye ($\alpha_r, \alpha_l$) and 1 nose tip ($\alpha_n$). A
plausible inner eye-nose triangle should be acute angled with,

\[
\begin{cases}
 d_{rl}^2 + d_{rn}^2 > d_{ln}^2 \\
 d_{rl}^2 + d_{ln}^2 > d_{rn}^2 \\
 d_{rn}^2 + d_{ln}^2 > d_{rl}^2 
\end{cases}
\]

where \( d_{rl}, d_{rn} \) and \( d_{ln} \) are the lengths of the sides of the triangle. Figure 3.9 shows illustrations of plausible acute angled candidate combinations in white and outlier combinations in yellow.

Plausible combinations of the candidate inner eye vertices and candidate nose tip vertices on \( \Psi \) are used as target points to transform the model. Next the remaining points of \( \Omega \) are moved to the closest vertices on \( \Psi \). \( \Omega \) is then projected back into the model space and the parameters of the model, \( b \), are updated. This selective search is performed until a face-fit is found. A face-fit is defined as one that results in all parameters \( b_i \) satisfying the condition

\[
|b_i| \leq +d_m \sqrt{\lambda_i}, \quad \forall \ i = 1, ..., t. \tag{3.11}
\]

where \( d_m \) is a suitable limit on the allowed standard deviation. An alternate approach to selecting a face-fit would be to choose the model fit such that the Mahalanobis distance \( (D_m) \) of the model parameters is less than a suitable value, \( D_{max} \):

\[
D_m^2 = \sum_{i=1}^{t} \left( \frac{b_i^2}{\lambda_i} \right) \leq D_{max}^2. \tag{3.12}
\]

In the final stage of the fitting algorithm, first the 2nd-order neighbourhood of the control points \( \omega_r, \omega_l \) and \( \omega_n \) are isolated as shown in Fig. 3.10. These vertices now
Figure 3.10: Landmark localisation: candidate vertices for refining the face-fit are obtained from 2nd order neighbourhoods (green) of the three control points (red).

Algorithm 1 Global Fitting

$E$: Set of candidate eye vertices; $F$: Set of candidate nose vertices
$x$: number of candidate eye vertices; $y$: number of candidate nose vertices
$\hat{C}_\Psi(x)$: Closest point to $x$ on $\Psi$

1: for $i \leftarrow 1, x$ do
2: $\alpha_r = E(i)$
3: for $j \leftarrow 1, x$ do
4: $\alpha_l = E(j)$
5: for $k \leftarrow 1, y$ do
6: $\alpha_n = F(k)$
7: Estimate $T_{\theta,t,s} \rightarrow D = |\alpha_r - \omega_r|^2 + |\alpha_l - \omega_l|^2 + |\alpha_n - \omega_n|^2$
8: $\hat{\Omega} = T_{\theta,t,s}(\Omega)$
9: for $p \leftarrow 1, N$ do
10: $\omega(p) = \hat{C}_\Psi(\hat{\omega}(p))$
11: end for
12: $\hat{\Omega} = T_{\theta,t,s}(\hat{\Omega})$
13: $b = \phi^T(\hat{\Omega} - \Omega)$
14: end for
15: end for
16: end for

Transformation with minimum $\nu$, where $\nu = \sum_i b_i$ is chosen as best fit

form the new candidate vertices that are used for a selective model fitting search. The final model fit is the transformation with the minimum deviation from the mean shape, while respecting the constraint of Eq. (3.11). This algorithm is summarised in Algorithm 1. Sample snapshots of the evolution of the model with different combinations of candidate vertices are shown in Fig. 3.11. Sample face meshes with the fit model are shown in Fig. 3.12.
Figure 3.11: Example of evolution of the shape model $\Omega$ during fitting (a)-(d).

Figure 3.12: Examples of scale and pose invariant model fitting on faces with different expressions (top), and faces with different pose and scale (bottom).
3.3.3 Evaluation

To evaluate the accuracy of the landmark localisation, we compare the landmarks localised on 2350 faces of the BU-3DFE database with the provided ground-truth. These faces are independent of the 150 faces used in training the model. We measure the model fitting accuracy based on the localisation errors of 5 key landmarks, i.e., outer eye points ($\alpha_{or}, \alpha_{ol}$), inner eye points ($\alpha_r, \alpha_l$) and nose tip point ($\alpha_n$). To account for different head sizes, the error is normalised by the distance between $\alpha_r$ and $\alpha_l$ in each scan. The mean normalised error for each landmark, across the test set, is estimated along with the standard deviation and used in the final comparison. We introduce a detection failure criterion, wherein if the distance between a landmark and the ground-truth is larger than a certain threshold ($\tau_P = 0.5$), it is deemed to be a failure.

Figure 3.13 (top) highlights the influence of the size, $L$, of the training set on the overall performance. It can be noticed that the accuracy improves, with a decrease in both the mean and standard deviation of the error, with the use of a larger training set $L$. This is because more shape variability is captured in the model without incurring in over-training. A sample face demonstrating the improvement in model fitting accuracy with increase in training is shown in Fig. 3.14. It can be seen that with a training set as small as 5 samples the model fitting accuracy is very poor with inaccurate landmark localisation (Fig. 3.14(a)). The model fitting accuracy progressively improves on increasing the amount of model training (Fig. 3.14(b-d)).

Figure 3.13 (middle) shows the influence of varying the number of model points $N$. Once again an improvement is seen on increasing the number of model points $N$, as a better description of the shape of interest is captured. The corresponding percentage of failed detections is shown in Table 3.1. The error graph and table show the 3 control points $\alpha_r, \alpha_l$ and $\alpha_n$ to be most stable, while the outer eye points ($\alpha_{or}, \alpha_{ol}$) have a higher mean error with a larger standard deviation and also more failed detections. This is because the 3 control points are localised using the initially isolated candidate locations and the locations are effectively refined in the model fitting process, thereby making them more robust than the outer eye points. The localisation accuracy for the outer eye points can be improved if candidate positions can be isolated robustly for these points and their use is incorporated into the model fitting process.
Figure 3.13: Normalised distance (error) between automatically detected landmarks and ground-truth landmarks, on 2350 faces of the BU-3DFE database, on varying number of training samples 'L', model points 'N' and variance 'σ' of additive white noise: (left) mean ; (right) standard deviation; (top) comparison with varying L (N=49); (middle) comparison with varying N (L=150); (bottom) comparison with additive white noise with variance σ (L=150, N=49).
Figure 3.14: A sample face showing model fitting on varying number of training samples $L$: (a) $L = 10$; (b) $L = 15$; (c) $L = 20$; (d) $L = 25$.

The evaluation of the robustness of the proposed landmark detection method can be seen in Fig. 3.13 (bottom). The figure shows the influence of additive white noise with variance $\sigma$. It can be seen that the algorithm achieves stable detections even up to $\sigma = 1$, with only a marginal increase in the mean and standard deviation of the error. Figure 3.15 shows examples of model fitting on two sample faces with $\sigma = 0.5$ and $\sigma = 1$, where the model fitting can be seen to be almost identical. This behaviour is also seen when applying spatially non-uniform noise (Fig. 3.16). The non-uniform noise was applied by varying the noise variance at each mesh vertex in proportion to the z-coordinate of that vertex. The resulting noisy mesh has higher noise at the regions of interest for the model fitting (assuming a frontal face which is the case in our testset). Even in the case of non-uniform noise, the mean and standard deviation of the error (Fig. 3.17) is seen to be stable across the variation of $\sigma$. The stability of the model fitting is due to the mesh Laplacian smoothing that is part of the model fitting pre-processing stage (Sec. 3.3.2) which effectively removes the applied noise. This results in consistent model localisation and the error seen is only due to model points that are localised on noisy peaks of the non-smoothed mesh.

We also analysed the effect of expressions on the accuracy of landmark local-
Table 3.1: Landmark detection accuracy error as a function of the number of model points ($N$), with $L = 150$ and a failure criterion $\tau_P = 0.5$ (Key: $\alpha_r$, right inner eye; $\alpha_l$, left inner eye; $\alpha_{or}$, right outer eye; $\alpha_{ol}$, left outer eye; $\alpha_n$, nose tip)

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\alpha_r$</th>
<th>$\alpha_l$</th>
<th>$\alpha_{or}$</th>
<th>$\alpha_{ol}$</th>
<th>$\alpha_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>12.0%</td>
<td>10.2%</td>
<td>13.6%</td>
<td>14.4%</td>
<td>2.4%</td>
</tr>
<tr>
<td>27</td>
<td>11.6%</td>
<td>9.1%</td>
<td>12.5%</td>
<td>12.5%</td>
<td>1.4%</td>
</tr>
<tr>
<td>37</td>
<td>11.3%</td>
<td>7.9%</td>
<td>12.5%</td>
<td>12.5%</td>
<td>1.4%</td>
</tr>
<tr>
<td>49</td>
<td>9.3%</td>
<td>7.6%</td>
<td>12.5%</td>
<td>12.5%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Figure 3.15: A sample face showing model fitting on varying variance $\sigma$ of additive white noise: (a) $\sigma = 0.5$ ; (b) $\sigma = 1.0$.

Figure 3.16: A sample face showing model fitting on varying variance $\sigma$ of additive spatially non-uniform white noise, where $\sigma$ is: (a) 50% of the z-coordinate ; (b) 80% of the z-coordinate

Figure 3.18 shows the mean normalised error and standard deviation with the 7 expressions (Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprised). The results show that least error was obtained with neutral and surprise expressions, while the highest error was obtained with anger, distress and happy expressions.

Table 3.2 shows the absolute mean error (in mm) in landmark localisation ob-
Figure 3.17: Normalised distance (error) between automatically detected landmarks and ground-truth landmarks on addition of spatially non-uniform white noise with variance $\sigma$, where $\sigma$ is the percentage of the z-coordinate ($L=150$, $N=49$).

Figure 3.18: Effect of expressions on landmark localisation accuracy. Normalised distance (error) between automatically detected landmarks and ground-truth landmarks: (left) mean; (right) standard deviation. (Key - AN: Anger; DI: Disgust; FE: Fear; HA: Happiness; NE: Neutral; SA: Sadness and SU: Surprise).

obtained on using the best model ($L=150$, $N=49$) and compares our approach with a state-of-the-art method replicating [13]. The main reasons for the improvement in detection accuracy is due to the relaxation of feature thresholds and invariance of the model fitting to pose variations.
Table 3.2: Comparison of detection accuracy error with an approach replicating [13], showing absolute mean distance (in mm).

<table>
<thead>
<tr>
<th></th>
<th>( \alpha_r )</th>
<th>( \alpha_I )</th>
<th>( \alpha_{or} )</th>
<th>( \alpha_{ol} )</th>
<th>( \alpha_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-statistical</td>
<td>25.01</td>
<td>26.68</td>
<td>31.84</td>
<td>34.39</td>
<td>14.59</td>
</tr>
<tr>
<td>PDM (N=49, L=150)</td>
<td>12.11</td>
<td>11.89</td>
<td>20.46</td>
<td>19.38</td>
<td>8.83</td>
</tr>
</tbody>
</table>

Figure 3.19: Normalised histogram for 2350 face (blue) and 2300 non-face (red) fits in terms of: (left) standard deviations of the spread of model parameters; (right) Mahalanobis distance.

### 3.4 Face detection

Face detection is performed during the model global fitting stage where the transformations between model points and candidate vertices are classified based on the upper-bound of the deviation of the parameters from the mean model. Figure 3.19 (left) shows an analysis of the deviation of the model parameters with face-fits and non-face-fits. For face-fits, ground-truth from 2350 faces of the BU-3DFE database (that were not used in training the model) was projected into the model space and the parameters \( \mathbf{b} \) calculated. Non-face fits are based on the model parameters from the model fitting on outlier face candidate vertices. The Equal Error Rate (EER) is at 11.80 standard deviations. Figure 3.19 (right) shows the distribution of the Mahalanobis distance for each face and non-face fit, where it can be seen that there exists a wide separation between the distances obtained from each class. The threshold \( D_{\text{max}} \) on the Mahalanobis distance in the range of 600-700 would separate face-fits from non-face-fits.

The Mahalanobis distance can be used if model localisation accuracy is not required and the goal is only to classify faces from non-faces. The localisation is however usually important, as also in the case of the proposed algorithm, to accurately segment the
face and for further refinement for landmark localisation. To make our algorithm robust to non-faces while being accurate for faces in terms of localisation, we use Eq. 3.11 instead of using the Mahalanobis distance. We restrict the deviation of each model parameter to the range of 3 standard deviations ($d_m = 3$). While the ground-truth parameters are seen to deviate above 3 standard deviations, model fitting is still obtained on these faces with a only a marginal decrease in accuracy. For landmark localisation, the fitting accuracy will be refined in the final stage of the fitting algorithm as described in Sec. 3.3.2. Figure 3.20 shows two example faces with ground-truth in red and face-fit in green where the fitting was restricted with $d_m = 3$. In these two examples, 60% (27 out of 45) ground-truth parameters were found to be greater than 3 standard deviations.

For detecting multiple faces, all candidate vertices within the bounds of a fit model are removed and model fitting is repeated with the remaining candidate vertices. For face segmentation, we place a sphere of radius $r$ at the midpoint of the line joining the nasal bridge and nose tip, and the intersection boundary of the face and the sphere is used to segment the face [112]. To account for different face sizes, $r$ is set in proportion to the size of the fit model. A histogram analysis of the ratios of top and bottom face heights to nose length was performed using 2927 faces from the BU-3DFE and GavabDB databases. Figure 3.21 shows the histograms of the ground-truth ratios of top face height $l_t$ to nose length $l_n$ (blue), and bottom face height $l_b$ to nose length $l_n$ (red). The length of the nose $l_n$ is defined as the Euclidean distance between the nose tip and nasal bridge landmarks. A larger variation in the bottom face height ratio is noticed due to expressions causing jaw movements. To make our algorithm robust to all face sizes, we set the value of $r = 2.6 l_n$. Figure 3.22 illustrates the segmentation procedure with the cropping sphere centered on the nose tip. The coloured portion of the face shows the segmented region.
3.4.1 Evaluation

The accuracy of the proposed face detector is evaluated on a dataset of 827 meshes of faces and non-faces. The dataset was composed of 427 faces from the GavabDB database and 400 object scans from the NTU-dataset. We varied the allowed parameter limit \( d_m \) from 0 to 5 with different combinations of training sets, \( L \), and configurations.
of model points, $N$. Figure 3.23 shows the different model configurations used. It was seen that better accuracy was obtained on using a larger training set (Fig. 3.24 (left)). Lower number of training samples results in a decrease in true acceptance rates, due to the limited shape variability captured by the model. On the other hand, model configurations with lower number of points also results in lower true acceptance rates along with higher false acceptance rates (Fig. 3.24 (right)). This is because models with lower number of points have a limited face-shape description leading to more false positives. We restrict our model to 150 training samples and 49 points to limit the complexity in training and model fitting. Figure 3.25 shows the precision, recall and accuracy on varying $d_m$ with the best model configuration ($L=150$, $N=49$). The optimum results were obtained with $d_m = 3$ and $d_m = 3.5$ where all the faces were classified correctly, with only 2 non-faces being detected as faces. Figure 3.26 shows examples of face detections from the GavabDB database, with the corresponding segmented faces. Note that we are able to detect faces even in the presence of large holes near the eye regions, thus overcoming the limitation presented in [57]. The average runtime for the PDM fitting over the GavabDB database was 121 seconds on a 3.20 GHz Intel Pentium 4 CPU.

Figure 3.27 (a) shows visual face detection and segmentation results on a scene containing the head and neck of three people, with varied orientation and hair coverage. The automatically segmented faces are displayed below. A composite scan with three faces at different scales and the corresponding segmented faces is shown in Fig. 3.27 (b). Notice the scale invariance of the facial model fitting and subsequent face segmentation. Finally, Fig. 3.27 (c) shows the face detection results on non-face objects. It is possible to notice that despite changes in scale and facial expressions, the faces are detected and correctly segmented from the rest of the body.
3.5 Signatures

As mentioned earlier, our aim is to devise a face signature that is compact in size and overcomes the limitations posed by facial expressions. This section describes our region-based and landmark-based face signatures. Combinations of regions and landmarks are evaluated for their robustness to pose and expressions, while the matching scheme is evaluated for its robustness to noise and data artifacts.

We represent a 3D face dataset as $\Phi = \{ (\Phi^1_{N_1}, C^1), (\Phi^2_{N_2}, C^2), \ldots, (\Phi^M_{N_M}, C^M) \}$, where $\Phi^i_{N_i} = \{ \Psi^i_1, \Psi^i_2, \ldots, \Psi^i_{N_i} \}$ are the $N_i$ faces of a person $C^i$ with $M$ people in the dataset. Then given any face $\Psi^i_j \in \Phi$ we aim to retrieve the identity and set of faces
3.5.1 Region-based signatures

The proposed region-based identification approach is based on rigid registration with prior anthropometric knowledge, which utilises an adaptation of the ICP algorithm. The algorithm uses the detected landmarks from the PDM fitting to first perform a coarse registration. Coarse registration is based on the best fit mapping in a least squares sense. The landmarks are then used to segment specific \textit{stable} regions on the face, which are...
robust to expressions and facial deformations. These regions are finally used to achieve fine registration. The registration approach is summarised in Fig. 3.28.

Stable regions \((R_s)\) include the region around the inner eye points, the nasal bridge between these points and around the eyebrow region (Fig. 4.1 (a)). The nose region (Fig. 4.1 (b)) is also relatively stable to most natural expressions. Vertices from these regions are localised using the fitted model and selected for ICP registration. We will refer to this approach as Selective-ICP (S-ICP).

Once the reference scan and the test scan are registered, the next step is to evaluate the distance between the two meshes as a measure of similarity. To this end, the symmetric Hausdorff distance \([113]\) is used. Let \(\Psi\) and \(\Psi'\) be the two facial meshes and \(\partial(v_i, \Psi')\) be the distance between a vertex \(v_i \in \Psi\) and \(\Psi'\). If we define

\[
\partial(v_i, \Psi') = \min_{v'_i \in \Psi'} (\|v_i - v'_i\|),
\]

then the Hausdorff distance, \(\partial(\Psi, \Psi')\), is given by

\[
\partial(\Psi, \Psi') = \sum_{v_i \in \Psi} \max[\partial(v_i, \Psi')],
\]

and the symmetric Hausdorff distance, \(\partial_s\), is then given by

\[
\partial_s(\Psi, \Psi') = \max[\partial(\Psi, \Psi'), \partial(\Psi', \Psi)].
\]

Figure 3.29 shows a comparison the proposed registration approach with a registration method based on the traditional ICP. The proposed registration approach is referred to as S-ICP, while the ICP registration is referred to as R-ICP. The R-ICP algorithm samples randomly 200 vertices from across the scan, while S-ICP uses around 100 vertices from the specific local regions for the ICP step. In both methods the ICP step is executed for 50 iterations. The first row of Fig. 3.29 shows the registration results with two identical meshes, one of which is slightly rotated in the YZ direction (where Y is the vertical axis and Z is the depth axis). In this case R-ICP outperforms S-ICP as there are no local deformations and the global minimum of the distance function corresponds with the best match.

Rows 2 and 3 of Fig. 3.29 show the matching of two meshes of the same person, taken at different time instances. In the second row, S-ICP clearly distinguish the regions of change around the cheeks and the eyes. This is consistent with the morphological
knowledge of the person under study as the cheek region consists of soft tissue which has changed over time, while the upper regions of the face consist of hard tissue which is known to be invariant. A change is seen in the eye region as the eyes are closed in one scan, and open in the other. R-ICP fails due to the presence of these local deformations and outliers. The third row shows the accuracy of using anthropometric information in the registration. Here the person changes pose with an intentional deformation of the left cheek so that the entire mouth region is displaced. The best registration is achieved by S-ICP, accurately highlighting the regions where deformations occurred.

Finally, figure 3.30 shows quantitative results where a scan was modified with synthetic deformations and then registered to its original self using both methods. The figure shows a scan in which cylindrical deformations were made, to partially overlap the semantic regions, with a total MSE = 1.00. S-ICP provides the best match with a measured MSE = 1.10, while R-ICP fails with measured error of 1.31. The evaluation of the registration approach for the recognition of faces is presented in Sec. 4.3.

### 3.5.2 Landmark-based signatures

The proposed landmark-based identification approach aims at finding a compact face representation, comprising of facial landmarks, that is robust to expression changes.
Given a set $S = \{\omega_1, \omega_2, ..., \omega_N\}$ of $N$ 3D landmarks on a face mesh $\Psi$, where $\omega_i = (x_i, y_i, z_i)$ represents the $i^{th}$ landmark, we extract geometrical information describing the face morphology. To this end, we compute the inter-landmark distances (ILDs), $d_{ij}$, between pairs of landmarks (see Fig. 3.31) and generate a feature vector, $\Delta$, of dimension $N(N-1)/2$, represented as
Inter-landmark distances (ILDs) between pairs of landmarks are used to represent the geometry of the face.

\[ \Delta = (d_{1,2}, d_{1,3}, \ldots, d_{1,N}, d_{2,3}, \ldots, d_{2,N}, \ldots, d_{(N-1),N}), \]  

where \( d_{i,j} = ||\omega_i - \omega_j|| \). When working with a sparse subset of facial points, inter-landmark distances are used instead of the point positions directly as this allows us to effectively capture the face geometry. This also allows us to be invariant to the initial orientation of the face and registration of the faces in the database is not required. Moreover, the use of distances allows scale invariance through a simple normalisation scheme. We choose the Euclidean distance for its simplicity of computation and robust representation of face geometry than, for example, geodesic distances. In fact, geodesic distances are highly sensitive to expressions, noise and the resolution of the face meshes. Moreover, the use of the Euclidean distance allows us to obtain a more concise signature as only \( N \) landmarks need to be stored, whereas the \( \frac{N(N-1)}{2} \) ILDs can be calculated at the recognition stage.

The feature vector \( \Delta \) is normalised with respect to the size of the face to make it scale invariant, thus generating

\[ \hat{\Delta} = \frac{\Delta}{d_S}, \]

where the scaling factor \( d_S \) is the distance between two pre-defined landmarks.

To reduce the dimensionality of the feature space we apply Subspace Linear Discriminant Analysis (SLDA) [114]. SLDA is the projection of the data onto a LDA space after applying PCA. The use of LDA as a feature space is suited for the task of face recognition, especially when sufficient samples per class are available for training. LDA is a supervised learning algorithm that targets data classification more than feature extraction and finds the classification hyperplane that maximises the ratio of the between-class
Figure 3.32: Block diagram of the proposed face recognition algorithm.

variance to the within-class variance, thereby guaranteeing maximal separability. The initial PCA projection allows us to reduce the dimensionality of the data while retaining its discriminative power, which LDA further improves upon by maximising the class separation.

Let $M$ be the number of faces in the training database and $\Delta = (\tilde{\Delta}_1, \tilde{\Delta}_2, \ldots, \tilde{\Delta}_M)$ represent the normalised feature vectors for all the training faces. The initial PCA projection, $\Lambda$, is defined as

$$\Lambda = A^T \tilde{\Delta}, \quad (3.18)$$

where $A$ is the transformation matrix whose columns are the eigenvectors obtained from the covariance matrix, $Z_{\Delta}$, of the data. The LDA projection, $\Gamma$, is defined as

$$\Gamma = B^T \Lambda, \quad (3.19)$$

where the matrix $B$ holds the eigenvectors of $Z_w^{-1}Z_b$. Here $Z_w$ is the within-class covariance matrix and $Z_b$ is the between-class covariance matrix [114].

For classification, we project the probe onto the created LDA-subspace and use the nearest mean classifier. Given a probe face $\Psi^p$ and its landmarks $(\omega^p_1, \omega^p_2, \ldots, \omega^p_N)$, we compute the feature vector $\tilde{\Delta}^p$ of normalised ILDs (Eq. 3.17). $\tilde{\Delta}^p$ is then projected onto the LDA-subspace using Eq. 3.18 and Eq. 3.19. The identity $\Delta^*_p$ is then chosen according
to

\[ \Delta_p^* = \arg \min_j \| \Gamma^p - \bar{\Gamma}^j \|, \]  

(3.20)

where \( \| . \| \) is the Euclidean distance, \( \bar{\Gamma}^j \) is the mean for class \( j \) and \( \Gamma^p \) is the projected probe face. The block diagram of the proposed approach is shown in Fig. 3.32. The evaluation of the approach is presented in Sec. 4.4.

3.6 Summary

In this chapter we discussed our algorithms for 3D face detection, detection of feature points and regions and introduce our proposed face identification algorithms based on regions and landmarks. Face detection is achieved by classifying model fits as face fit and non-face fit based on the model parameters. The face detection is efficient on evaluation with a database of faces and non-faces, and is also demonstrated on scenes with multiple faces and large scale variations. Landmark localisation is performed by finding the model fit that minimises the deviation of the model from the mean shape. The performance of the landmark localisation approach was evaluated with different parameters, models and number of training samples. The algorithm is effective in the fitting of the model and shows significant improvement over a state-of-the-art approach. The region-based face identification algorithm performs accurate registration of 3D faces with incorporation of prior anthropometric knowledge. While the landmark-based signature extracts geometric features of the face in the form of inter-landmark distances (ILDs), to form a concise representation of the face that is robust to facial expressions. In Sec. 4.2 we evaluate the two proposed face recognition approaches.
Chapter 4

Experimental results

4.1 Introduction

In this chapter we demonstrate the performance of the proposed algorithms for region-based (discussed in Sec. 3.5.1) and landmark-based (discussed in Sec. 3.5.2) face recognition. We first describe the experimental setup in Sec. 4.2, followed by the evaluation of the region-based approach in Sec. 4.3 and the landmark-based approach in Sec. 4.4.

4.2 Experimental setup

We demonstrate the performance of the region-based recognition approach on the GavabDB database. As mentioned earlier, this dataset offers the most challenging scenario with missing data, high level of noise and data artefacts, and uncontrolled expressions. The landmark-based recognition approach is evaluated with the BU-3DFE database as this dataset offers a wider range of well classified expressions. The nature of the dataset allows for the analysis of expressions towards finding the most compact expression-invariant signature. For both approaches the feature detection was performed using a PDM (discussed in Sec. 3.3) trained with the ground-truth landmarks of the BU-3DFE database.

The performance of the approaches is evaluated in terms of recognition and retrieval accuracies. Recognition here refers to the accuracy of the retrieved rank-1 identity, while retrieval refers to accuracy of retrieving faces of the same person with most similarity. The retrieval accuracy is measured using the average dynamic precision (ADP) [115]. The ADP is defined as

\[
ADP = \frac{1}{S} \sum_{i=1}^{S} \frac{T_i}{i},
\]  

(4.1)
where \( T_i \) is the number of true positives, with \( T_i \leq i \), and \( S \) is the scope size which refers to the total number of expected true positives. \( S \) is set to 7 in our experiments since we have 7 samples per person. For example, if for a given query the retrieved results correspond to \([1, 1, 0, 1, 1, 0, 1]\) until rank-7 (where 1 is a true positive and 0 is a true negative), the \( ADP = 1 + 1 + 0.67 + 0.75 + 0.8 + 0.67 + 0.71 = 5.56/7 = 0.794 \).

We also use Receiver Operating Characteristic (ROC) curves to demonstrate the accuracy of the algorithms. ROC curve (receiving operating characteristic) analysis has been widely used for the evaluation of face recognition systems. Given the known identity of a person (class) all the faces classified as belonging to that identity are called \textit{positives}. A correct classification is called \textit{true positive} and an incorrect one as \textit{false positive}. By varying a certain classifier-threshold from the minimum to the maximum value of the classifier output, a ROC curve is constructed for this classifier. The curve shows true positive rate (sensitivity) versus false positive rate (1-specificity). The user is then able to decide upon a compromise between sensitivity and specificity achievable simultaneously by the classifier.

### 4.3 Region-based signatures

In this section we evaluate the performance of the proposed region-based face recognition algorithm and show a comparison with the conventional ICP registration approach. First, we find the most robust facial regions to be used in the similarity estimation based on the Hausdorff distance estimation, using different regions of the face. The face was divided into 6 regions (\( R_1 - R_6 \)), separating the forehead, eyes, nose, cheek, mouth and chin, as shown in Fig. 4.1 (d). For a given selection of \( P \) regions \( S_R \in \Psi \) and \( S'_R \in \Psi' \), the similarity \( \Delta \) is defined as,

\[
\Delta = \max \left[ \sum_{v_i \in S_R} \max \left[ \partial(v_i, \Psi') \right], \sum_{v'_i \in S'_R} \max \left[ \partial(\Psi', v'_i) \right] \right]. \tag{4.2}
\]

We first tested the rank-1 recognition accuracy on applying the ICP algorithm over the cropped face and using the 6 facial regions in the similarity estimation. The single region that lead to the worst retrieval is the mouth region (\( R_5 \)), which is the most affected by variations in expressions. The best results were provided by the forehead, eyes and nose regions (\( R_1, R_2 \) and \( R_3 \)) and their combinations. Figure 4.2 shows boot-strapping statistics on the mean rank-1 retrieval accuracy obtained with the use of these regions. The mean rank-1 retrieval score was calculated for each region combination from 100 random subsets.
Chapter 4: Experimental results

Figure 4.1: Regions used for registration and distance estimation: (a) stable regions $R_S$ composed of the region around the inner eye points, the nasal bridge and eyebrow region; (b) nose region; (c) combination of $R_S$ and the nose region; (d) Regions $R_1 - R_6$ used in the distance estimation.

Figure 4.2: Mean boot-strapping statistics of rank-1 retrieval accuracy from 100 runs of random subsets of the GavabDB database, on using $R_1$, $R_2$ and $R_3$ in the similarity estimation, post-ICP registration: (a) $R_1$; (b) $R_2$; (c) $R_3$; (d) $R_1 + R_2$; (e) $R_1 + R_3$; (f) $R_2 + R_3$; (g) $R_1 + R_2 + R_3$. The black line shows the trend of the probability distribution.

of the database, and the shown bootstrap statistics were generated for each combination. Among the individual regions (Fig. 4.2(a-b)), the nose region ($R_3$) was found to be most robust with 86% accuracy achieved at 37.3% probability, and the eyes ($R_2$) performed the worst with 70% accuracy achieved at 28.5% probability. Combinations involving the nose region (Fig 4.2(e-g)) achieved higher accuracies than regions without the nose, with the best retrieval of 86.5% achieved using all three regions at the best probability of 47.8%.

For the finding the best facial regions to be used for the registration we evaluated 3 region configurations (Fig. 4.1 (a-c)). We refer to the configuration with $R_S$ as $S$-
Figure 4.3: Rank-1 recognition (left) and retrieval accuracy (right) obtained using the forehead ($R_1$), eyes ($R_2$) and nose ($R_3$) regions in the similarity estimation.

ICP1, with $R_3$ as $S$-ICP2 and with the combination of both regions as $S$-ICP3. For the detection of landmarks on the face meshes, the PDM $\Omega$ is generated from the BU-3DFE database and the corresponding ground-truth landmarks provided with the database. We use 48 ground-truth landmarks from the eyebrows, eyes and nose regions and include an additional landmark at the nose-tip. Our training set is composed of landmarks from 150 faces, corresponding to 25 scans from 6 individuals each. The 25 scans corresponds to five expressions (anger, disgust, fear, happiness, sadness and surprise) with four degrees (intensities) of expressions and one neutral expression.

Figure 4.3 shows the rank-1 recognition and retrieval accuracy in terms of the ADP of the 4 approaches. The overall best results were obtained on using the nose region for both the registration (S-ICP2) and similarity estimation (R3), with a rank-1 recognition rate of 93.7% and an ADP of 91.1%. For ICP the nose region is the most robust with a recognition rate of 85.2% and an ADP of 83.1%. In the S-ICP1, the region combination of the forehead, eyes and nose ($R_1 + R_2 + R_3$) gives the best result with recognition rate 87.4% and ADP 83.9%. Finally, for S-ICP3 we again achieve most robustness with regions ($R_1 + R_2 + R_3$), with recognition rate 87.8% and ADP 83.9%. We observe a correlation between the regions used in the registration and subsequently in the similarity estimation. In S-ICP2 the nose is used to register the faces and in turn proves to be the most robust in estimating the distances. In S-ICP1 and S-ICP3, parts of the forehead, eyes and nasal bridge is used to register the faces and these regions also proves most robust in the similarity measure. However, recognition and retrieval using $R_3$ outperformed the other regions in terms of robustness.
Figure 4.4: Cumulative match characteristic plot upto rank-10 obtained using the S-ICP2 registration approach, and the forehead (R1), eyes (R2) and nose (R3) regions in the similarity estimation. The results saturate at rank-6 as the database contains 6 samples per subject.

Figure 4.5 shows example results with estimated distances, post-registration, with the 4 approaches. We see here that using ICP on the cropped face (Fig. 4.5 (c)) provides a good but not completely accurate registration. Erroneous regions of change still exist around the nose, cheek and chin areas. The same is noticed with the S-ICP1 approach. S-ICP2 also results in inaccurate distances estimated around the forehead region in this case. The best registration is achieved through the S-ICP3 approach with regions of change only found around the cheek region which is consistent with the expected deformation. Figure 4.6 shows the retrieval results with an example query exhibiting an exaggerated expression. The rank-1 retrieved face corresponds to the query in all cases. ICP performs the worst with only 1 true positive within the top-6 retrieved faces. S-ICP2 outperforms the other approaches with all true positives in the top-6 faces.

Table 4.1 compares the rank-1 recognition accuracy achieved by the proposed approach with other state-of-the-art approaches on the GavabDB database. Moreno et al. [58,116] reported 78% rank-1 recognition (in 2003 and 2005), while Mousavi et al. [94] report 91% in recent work (in 2008). Our approach, S-ICP2, achieves 93.7% rank-1 recognition. The work proposed in [95] achieves higher recognition accuracy than our proposed approach, however requires a large amount of training. The proposed approach on the other hand requires no prior training phase. The main failure mode of the S-ICP2 ap-
Figure 4.5: Registration results using various region configurations and comparison with ICP: (a) ICP; (b) S-ICP1; (c) S-ICP2; (d) S-ICP3.

Table 4.1: Comparison of 3D face recognition algorithms on the GavabDB database

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Technique</th>
<th>Number of training</th>
<th>Number of probe</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>[93]</td>
<td>Matching Ridge lines</td>
<td>-</td>
<td>427</td>
<td>93.5%</td>
</tr>
<tr>
<td>[94]</td>
<td>Principal component analysis</td>
<td>244</td>
<td>61</td>
<td>91%</td>
</tr>
<tr>
<td>[95]</td>
<td>Multi Library Wavelet Neural Networks</td>
<td>244</td>
<td>61</td>
<td>94.8%</td>
</tr>
<tr>
<td>[96]</td>
<td>Multiview keypoint matching</td>
<td>61</td>
<td>366</td>
<td>90%</td>
</tr>
<tr>
<td>Proposed approach</td>
<td>Selective iterative closest point algorithm</td>
<td>-</td>
<td>427</td>
<td>93.7%</td>
</tr>
</tbody>
</table>

...approach occurs in faces with a large amount of noise or holes around the nose region, leading to incorrect PDM fitting and consequently poor registration. Figure 4.7 shows an example query face which resulted in poor retrieval due to the entire nose region being absent.

### 4.4 Landmark-based signatures

In this section, we evaluate the performance of the proposed landmark-based face recognition algorithm first with manual landmarks to find the best configuration of landmark combinations that is expression-invariant. The most robust configuration is then detected automatically. The recognition performance is then compared with the 3D eigenface approach [74]. Moreover, we discuss the influence of the expression intensities in the training, the memory requirements of the automatically detected face signature and its failure modes. The proposed PDM discussed in this section is trained with manually annotated landmarks from 100 face meshes from the BU-3DFE database.
Figure 4.6: Retrieval results of an example query showing robustness of using the nose region $R_3$ in the S-ICP2 approach (faces highlighted in green indicate true positives, while red indicate false positives): (a) ICP; (b) S-ICP1; (c) S-ICP2; (d) S-ICP3.

4.4.1 Signature selection

To determine the landmark combinations leading to the most robust model to expressions, we tested various subsets (regions) of the 83 ground-truth landmarks on all 100 individuals from the BU-3DFE database. The regions included the left and right eyes and eyebrows, the nose, the mouth and the boundary of the face. The scaling distance $d_S$ used for feature normalisation is the distance between the two outer eye points, as shown in Fig. 4.8. An exhaustive combination of landmarks from the five regions results in 31
Figure 4.7: Failure mode: Example scan showing missing nose region which causes failure of the region based approaches due to incorrect PDM fitting and region localisation.

Figure 4.8: Sample face scan showing the landmarks and the scaling distance $d_S$ (dotted line) used in the tests. The number of points in each region of interest are 20 for the mouth, 12 for the nose, 8 each for the left and the right eye, 10 each for the left and the right eyebrow, and 15 for the face boundary.

different models ($2^5 - 1$), ranging from single regions to all the regions. Table 4.2 lists these regions and the number of landmarks used from each region (with $N=83$). Table 4.3 presents the results obtained through our proposed approach, using an exhaustive combination of landmarks from the 7 regions in 31 different models (M1-M31). We found that while the model using all 83 landmarks achieves 100% rank-1 recognition (model M31), the same is also obtained using only the eyes, eyebrows and nose (model M30). Based on this observation we chose the model M30 with the eyes, eyebrows and nose only, as sufficient robust landmarks for our face signature. The choice of this configuration is also motivated by the fact that it is relatively easy to automatically detect these regions compared to detecting the mouth and boundary. The mouth being the least robust region to pose and
Table 4.2: Number of points used in each region of interest

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth</td>
<td>20</td>
</tr>
<tr>
<td>Nose</td>
<td>12</td>
</tr>
<tr>
<td>Left eye</td>
<td>8</td>
</tr>
<tr>
<td>Right eye</td>
<td>8</td>
</tr>
<tr>
<td>Left eyebrow</td>
<td>10</td>
</tr>
<tr>
<td>Right eyebrow</td>
<td>10</td>
</tr>
<tr>
<td>Boundary</td>
<td>15</td>
</tr>
</tbody>
</table>

expression makes it difficult to accurately detect while the boundary is dependent on the output of the scanning device and sensitive to the presence of outliers.

**Robustness to noise**

To evaluate the robustness of the different models to noise, we performed experiments with additive Gaussian noise on the manual landmarks upto 15% and observed its effect on the recognition rate. Figure 4.9 shows the 3D plot of the accuracy of the different configuration of landmarks with different levels of noise. The graph presents the average of obtained accuracy on 10 different runs for each level of noise.

We find that the model M20 (eyes and nose) was found to be most robust to noise, while combinations containing the eyebrows and mouth were most sensitive. Also we see that for the chosen configuration M30, the approach is robust with no change in accuracy for upto 2% noise, after which the accuracy drops with higher noise levels. The reason for the eyes and nose to be more robust to noise can be attributed to the spatial distribution of these landmarks, which are more spread compared to the eyebrows and mouth. The landmarks from both the eyebrows and lips of the mouth form two thin lines, and are therefore more sensitive to noise as compared to the eyes and nose. Hence for automatically fitting models that use the eyebrows or mouth, greater accuracy should be ensured for these regions.

Based on our finding, the final choice for our face signature is the combination of the eyes, eyebrows and nose regions (48 landmarks). This is the most compact representation that lead to the best recognition result with the same accuracy as using the full model with 83 landmarks. The single region that lead to the worst recognition results is the mouth region, which is the most affected by variations in expressions. The mouth region was also found to be least robust to noise.
Chapter 4: Experimental results

Table 4.3: Region combinations (using ground-truth landmarks) used to find the model most robust to expressions and noise

<table>
<thead>
<tr>
<th>Model</th>
<th>Eyes</th>
<th>Eyebrows</th>
<th>Nose</th>
<th>Mouth</th>
<th>Boundary</th>
<th>Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td></td>
<td>52.82%</td>
</tr>
<tr>
<td>M2</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td></td>
<td>76.92%</td>
</tr>
<tr>
<td>M3</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>82.82%</td>
</tr>
<tr>
<td>M4</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td></td>
<td>86.41%</td>
</tr>
<tr>
<td>M5</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>91.54%</td>
</tr>
<tr>
<td>M6</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>91.79%</td>
</tr>
<tr>
<td>M7</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>91.79%</td>
</tr>
<tr>
<td>M8</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>93.08%</td>
</tr>
<tr>
<td>M9</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>94.10%</td>
</tr>
<tr>
<td>M10</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>95.64%</td>
</tr>
<tr>
<td>M11</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>96.67%</td>
</tr>
<tr>
<td>M12</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>97.69%</td>
</tr>
<tr>
<td>M13</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>97.95%</td>
</tr>
<tr>
<td>M14</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>97.95%</td>
</tr>
<tr>
<td>M15</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>98.21%</td>
</tr>
<tr>
<td>M16</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td>98.21%</td>
</tr>
<tr>
<td>M17</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>98.46%</td>
</tr>
<tr>
<td>M18</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>98.72%</td>
</tr>
<tr>
<td>M19</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td>98.97%</td>
</tr>
<tr>
<td>M20</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>99.23%</td>
</tr>
<tr>
<td>M21</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>99.23%</td>
</tr>
<tr>
<td>M22</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>99.23%</td>
</tr>
<tr>
<td>M23</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>99.23%</td>
</tr>
<tr>
<td>M24</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td>99.23%</td>
</tr>
<tr>
<td>M25</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>99.49%</td>
</tr>
<tr>
<td>M26</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>99.74%</td>
</tr>
<tr>
<td>M27</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>99.74%</td>
</tr>
<tr>
<td>M28</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>99.74%</td>
</tr>
<tr>
<td>M29</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>100%</td>
</tr>
<tr>
<td>M30</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>100%</td>
</tr>
<tr>
<td>M31</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td>100%</td>
</tr>
</tbody>
</table>
Chapter 4: Experimental results

Figure 4.9: Comparison of results with different models (using ground-truth landmarks) on the addition of Gaussian noise.

Table 4.4: Recognition results from different region combinations using automatically localised landmarks

<table>
<thead>
<tr>
<th>Model</th>
<th>Eyes</th>
<th>Eyebrows</th>
<th>Nose</th>
<th>Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>M3</td>
<td>√</td>
<td></td>
<td></td>
<td>37.98%</td>
</tr>
<tr>
<td>M5</td>
<td></td>
<td>√</td>
<td></td>
<td>56.83%</td>
</tr>
<tr>
<td>M2</td>
<td>√</td>
<td></td>
<td>√</td>
<td>75.41%</td>
</tr>
<tr>
<td>M11</td>
<td>√</td>
<td>√</td>
<td></td>
<td>81.15%</td>
</tr>
<tr>
<td>M20</td>
<td>√</td>
<td></td>
<td>√</td>
<td>87.71%</td>
</tr>
<tr>
<td>M17</td>
<td></td>
<td>√</td>
<td>√</td>
<td>89.89%</td>
</tr>
<tr>
<td>M30</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>96.5%</td>
</tr>
</tbody>
</table>

4.4.2 Signature performance

In order to automatically detect the landmarks of the chosen model, we generate a PDM that includes the desired landmarks along with statistical information of the shape variation over a training set and then fit it to each probe and training mesh.

Region influence

To study the influence landmarks from the 3 different facial regions (eyes, eyebrows and nose) had on the recognition accuracy, we performed an analysis similar to the one used in the selection of the face signature (Sec. 4.4.1). Table 4.4 shows the recognition
accuracies obtained with different combinations of the 3 regions. In contrast to the results obtained with ground-truth landmarks (Table 4.3), it can be seen that the nose is found to be more robust than the eyes and the eyebrows. This is largely due to the fact that the nose is more robustly detected in the model fitting process than the eyes or eyebrows which have a larger amount of error. While the eyes and eyebrows have higher localisation errors than the nose, the combination of the two (M30) performs better than the nose alone. The best result is obtained with use of all three regions (M30) with a recognition accuracy of 96.5%.

**Classifier performance**

To evaluate the feature space and the classifier, we compared the recognition accuracy of using the SLDA projection and those obtained using PCA and LDA individually on the selected 48 landmarks. The feature vectors are projected onto these feature spaces, and classification is performed either by the nearest mean (M) or the nearest neighbour (N) classifier. Figure 4.10 shows that LDA provides a higher accuracy than PCA and that the nearest-mean classifier provides better accuracy than the nearest neighbour classifier. Finally, SLDA outperforms both PCA and LDA.

**Influence of training**

Figure 4.11 shows a comparison of the recognition rates obtained with different training and probe combinations to analyse the influence of the expression intensities
Figure 4.11: Comparison of recognition results using different combinations of probe and training sets on the BU-3DFE database with automatically localised landmarks. The training set was varied to include (red) and exclude (blue): (a) neutral, (b) intensity-1 (lowest intensity), (c) intensity-2, (d) intensity-3, and (e) intensity-4 (highest intensity).

used in the SLDA training. Note that the neutral intensity has the least influence on training (Fig. 4.11(a)), while the inclusion of the remaining intensities have a larger effect (Fig. 4.11(b-e)), because neutral samples are fewer than the other intensities (1 neutral and 6 each of intensity 1-4 per person). A better accuracy is achieved when more samples having a wide expression range are used in the training. More accurate recognition is obtained when the neutral and intensity-1 samples are used as probe, while intensity-4 provides the least accurate results. The best result (96.5% recognition accuracy) is achieved using intensity-1 as probe and the remaining samples in the training. The reduced recognition accuracy for the highly expressive samples is to be attributed to the reduced accuracy in the PDM fitting. Figure 4.12 shows two examples of wrongly classified signatures (red circles), happening for large expression intensities, and compares them with the ground-truth landmarks (green squares). The inaccurate fitting of the PDM here is due to the inaccurate localisation of candidate inner eye vertices. The model fitting in these cases
Influence of feature dimensions

The use of PCA before applying LDA allows us to considerably reduce the dimensions of the feature space while retaining the most relevant information. To analyse the influence of the number of dimensions in the identification accuracy, Figure 4.13 shows and compares the rank-1 recognition results obtained when varying the amount of feature energy retained by the eigenmodes after PCA using the manual and automatic landmarks. The inset in the figure compares the actual number of feature dimensions used to obtain the results in the highlighted region. For manual landmarks, the number of dimensions that lead to the highest accuracy was 115, which corresponds to 10.20% of the original size of the feature vector (1128). In the case of automatic landmarks, the maximum accuracy was obtained with 22.53% (265 modes) of the feature vector. This is due to the fact that automatic landmarks contain a larger amount of noise as compared to the manual landmarks, requiring more information to represent a face. These reduced dimensions correspond to 99.97% and 99.57% of the signal energy for automatic and manual landmarks, respectively.

These results demonstrate that the extensive set of inter-landmark distances, used as the feature set, contains a large amount of redundant and non-discriminatory information that is effectively removed through the application of PCA. While much of these discarded dimensions would contain valuable information related to facial expressions, they do not contribute towards identity recognition and is effectively noise for this purpose.
Figure 4.13: Comparison of rank-1 recognition accuracies, on the BU-3DFE database, of manually and automatically localised signatures on varying the amount of feature energy retained by the PCA eigenmodes. The inset is a comparison of the signature accuracy against the number of eigenmodes for the highlighted region.

**Influence of signature representation**

Figure 4.14 compares the ROC curves obtained with the automatic landmarks of the probe using 32-bit floating point and 16-bit integer representation, where the landmark coordinates were rounded to the nearest integer. The motivation for this experiment was to quantify the decrease in recognition accuracy when reducing the precision of the proposed facial signature. While the storage of the signature with floating point representation requires 588 bytes only, with an integer representation we would achieve a further 50% reduction in the storage requirements. This would allow the 3D face signature to be stored not only on devices such as RFIDs, but also in 2D barcodes. The rank-1 identification rate is 96.5% with the floating point representation and 92.64% with integer representation. With a significant decrease in the signature size (50%) using the integer representation, there was only a 3.86% decrease in rank-1 recognition.

Figure 4.14 also compares the proposed approach with a SVM based classifier and a 3D Eigenface method replicating [74], where depth-maps of entire face meshes were used in the PCA projection. The SVM classifier used a radial basis function as kernel and achieved a rank-1 recognition score of 94.3%. The 3D eigenface method obtains lower accuracy results, with 60.48% rank-1 recognition rate, as it is not capable to properly handle large changes in expressions. The 3D eigen-face approach is also presented in [72] and [73] with multimodal data and, with the use of only the 3D modality, recognition rates of 85% and 88.9% were reported, respectively.
Chapter 4: Experimental results

Figure 4.14: Comparison of results, on the BU-3DFE database, using automatically localised signatures with 32-bit float and 16-bit integer representations, and the baseline 3D eigen-face approach.

4.5 Summary

This chapter presented and discussed the evaluation of our proposed face recognition algorithms. The approaches were evaluated for their robustness to pose, expressions and in the presence of noise and data artefacts. For the region-based approach, facial regions are used in the registration and similarity estimation and we demonstrated that the use of expression invariant regions outperforms use of the entire face. It was seen that the nose is the most robust in all region configurations with which we achieve 93.7% rank-1 recognition on the GavabDB database.

For the landmark-based approach, we generated a concise representation of a face that is robust to noise and facial expressions. Initially using manually localised landmarks we find the model of landmarks most robust to expressions and noise. We also demonstrate the accuracy of using the sub-space LDA transformation as opposed to using PCA or LDA alone. Based on the choice of the robust model we extract them automatically and achieve scale and pose invariant recognition. With the use of the automatically detected landmark-signature we demonstrate 96.5% rank-1 identification rate on the BU-3DFE database.
Chapter 5

Conclusions

5.1 Summary of achievements

We studied and implemented an algorithm for the detection and segmentation of facial landmarks and regions from 3-dimensional face scans. The proposed algorithm eliminates the need for prior knowledge of the orientation and the pose of a scan by using a Point Distribution Model and relaxes the constraints on feature map thresholding. The Point Distribution Model encodes the shape information of the region of interest to be segmented. Feature maps highlighting the curvature properties of the scan are first extracted and then used to isolate candidate points of interest. These points of interest are the inner eye and nose tip vertices of the mesh. The final landmark selection is performed by minimising the deviation between the model and candidate vertices.

The proposed approach is used to develop further algorithms that demonstrated its effectiveness in a face segmentation (detection) application for 3D scenes and in an expression-invariant face recognition application. Face detection is performed using the Point Distribution Model and the detected features by classifying, based on the deviation of the model parameters, model fits into one of two classes, namely face-fit or non-face-fit. Expression-invariant face recognition is obtained with one of two types of face signatures for capturing the 3D geometry of a face and characterising the person identity. The first signature is based on the use of expression-invariant regions of a face. These identified regions and prior anthropometric knowledge are used for face registration. The second signature is based on landmarks. In this case, the face geometry is represented with inter-landmark distances within selected regions of interest. The approaches were evaluated for their robustness to pose, expressions and in the presence of noise and data artefacts. For the region-based approach, we demonstrated that the use of expression invariant regions
outperforms use of the entire face. It was seen that the nose is the most robust in all region configurations with which we achieve 93.7% rank-1 recognition on the GavabDB database. With the landmark-based approach, we generated a concise representation of a face and achieve scale and pose invariant recognition. With the use of the compact landmark-signature we demonstrate 96.5% rank-1 identification rate on the BU-3DFE database. The compactness of the proposed face signature allows its storage on RFIDs and 2D barcodes.

While 3D face recognition is not extensively used in practical applications currently, due to cost of deployment of 3D enrollment and surveillance cameras, there already exists commercially available systems for small scale use. Such systems can be deployed, for example, in companies and universities for access control to systems and buildings. The compactness of our proposed landmark-based face signature and its easy storage on access cards will allow its straightforward use for such systems. In the near future when 3D capture systems are more extensively used, the proposed signature will be a useful additional modality to be stored on national biometric ID cards and passports.

The current trend in face recognition is towards hybrid 2D-3D systems where conventional 2D based methods are combined with the more recent 3D shape based approaches. This allows the combination of different matching schemes with the aim of increased classification rate. 2D based approaches have been extensively studied over the last few decades and much success has been achieved in this area. 3D data now allows use of information that was previously unavailable, thus providing a rich set of facial features. The proposed 3D landmark-based signature approach can overcome current limitations of feature localisation, pose and illumination, and can also be used in conjunction with 2D systems. The use of texture information in our approach, apart from being used as an additional recognition feature, will also aid in refining the model fitting process through an approach similar to 3D Morphable Models. The use of texture will also allow a more robust detection of the mouth region, which has not been addressed in this work. The combination of 2D texture information in the proposed approach as a biometric feature can be achieved at two levels: 1) directly at the feature level where it can be fused with the 3D features; or 2) at the decision level through rank-based or voting scheme approach. A summary of possible future extensions of the proposed work is provided below.
5.2 Future work

- The feature detection algorithm is based on curvature estimation and therefore requires surface scans as input, as opposed to point clouds. A 3D face point cloud requires surface generation or fitting of a dense deformable face model, prior to application of our algorithms. This has not been addressed and can be a further extension of the work.

- One of the limitations of the proposed face signatures is that they are dependent on the availability of eyes and nose regions. If in the cases of missing data arising from occlusions during the scanning process these regions are not present, the feature detection step will fail. An improvement of this work would be to estimate the reliability of regions to be used in the matching.

- We chose not to detect landmarks around the mouth region as it is be most affected by expression with large changes introduced by the opening and closing of the lips. However, if landmarks from the mouth could be detected robustly, this would allow analysis of the deformations introduced in this region. This analysis would aid in designing approaches to extract person-specific face characteristics from the mouth region while being invariant to the large expression-related deformations.

- Texture information was not used in the development of the proposed approach. An extension of this thesis can be to study appropriate fusion methods to integrate textural information in the proposed methods.
Bibliography


