UNIVERSITY OF SOUTHAMPTON

Towards Unconstrained Ear Recognition

by

John D. Bustard

A thesis submitted for the
degree of Doctor of Philosophy

in the
Faculty of Physical and Applied Sciences
Department of Electronics and Computer Science

April 2011

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES
DEPARTMENT OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by John D. Bustard

Humans can recognise individuals in many different situations. Automated vision-based biometric systems, which identify individuals from an image of a particular physical feature, aspire to a similar level of performance but currently have to impose constraints to achieve satisfactory recognition rates. These include limitations on the background of the image in which a feature is located, the lighting on the feature, its degree of occlusion, its viewed angle, and the properties of the camera that captures it. The computational cost of any recognition system is also an issue.

This thesis examines ways of reducing such constraints. Its particular focus is the recognition of individuals from the unique signature provided by their ears.

Specifically, the work develops techniques to support a hypothesis that:

The constraints on the use of ear-based biometric systems can be relaxed significantly through the introduction of robust recognition techniques.

Two novel techniques designed to improve robustness are described: (i) a fully automated 2D recognition system to reduce sensitivity to noise and occlusion; and (ii) the use of a 3D model to allow for variations in both pose and lighting;

The thesis begins by summarising current progress in the general field of biometrics and in the associated techniques for robust recognition. Each technique is then described in successive chapters, identifying related work, explaining the technique in detail and evaluating its performance. Future work will focus on developing algorithms to enable the 3D model to be accurately fitted to images. A number of developments in this area are outlined in the appendix.

While these techniques have been developed for ear recognition they also contribute to the general research challenge of recognising any object in any environment.

Contents

Acknowledgements **ix**

1 Introduction **1**

1.1 Biometrics 1

1.2 Ear Recognition 3

1.3 Robust Recognition 4

1.4 Contributions 5

1.5 Publications 5

1.6 Thesis Structure 6

2 Robust Recognition **9**

2.1 Biometric Systems 9

2.2 Unconstrained Factors 11

2.2.1 Background 11

2.2.1.1 Position and scale 11

2.2.1.2 Rotation 11

2.2.2 Occlusion 12

2.2.3 Lighting 12

2.2.4 Pose 13

2.2.5 Camera 13

2.2.5.1 Field of view 14

2.2.5.2 Focus and motion blur 14

2.2.5.3 Resolution 14

2.2.5.4 Colour sensitivity 15

2.2.5.5 Noise 15

2.2.6 Within class variation 15

2.2.6.1 Intra-subject variation 15

2.2.6.2 Identity invariant detection 16

2.2.6.3 Identification 16

2.3 Review of Detection and Recognition Techniques 16

2.3.1 Image Processing and Feature Extraction 17

2.3.1.1 Filtering 17

Noise 17

2.3.1.2 Super resolution 17

Resolution 17

2.3.1.3 Normalisation 18

Position and Scale 18

Rotation 18
Lighting 18
Colour Sensitivity 18
2.3.1.4 Robust distance measures 19
Occlusion and Noise 19
2.3.1.5 Edges 19
Lighting 19
2.3.1.6 Sub-regions 19
Occlusion 19
2.3.1.7 Area transform 20
Position and Scale 20
Rotation 21
2.3.1.8 Histograms 22
2.3.1.9 Feature points 22
Position, Scale and Rotation 22
Occlusion 23
Pose 23
2.3.1.10 Shape from X 23
Lighting 24
Pose 24
2.3.2 Model based techniques 25
2.3.2.1 Template matching 25
Position and Scale 25
2.3.2.2 Hough transform 25
Position, Scale and Rotation 25
Occlusion 26
2.3.2.3 Model fitting to interest points 26
2.3.2.4 Active shape models 26
Position, Scale and Rotation 27
2.3.2.5 Subspace 27
Lighting 27
Pose 28
Noise 29
2.3.2.6 Active appearance models 29
Position, Scale and Rotation 29
Resolution 30
2.3.2.7 Morphable models 31
Lighting 31
Pose 31
Intra Subject Variation 32
2.3.3 Summary 32
2.4 Detection and Recognition using combined techniques 34
2.4.1 Non-contact biometrics 35
2.4.2 Unconstrained recognition 37
2.5 Ear Recognition 38
2.5.1 Ear Recognition Robustness 40
2.5.1.1 Position, rotation and scale 40

2.5.1.2 Occlusion 43
2.5.1.3 Pose 43
2.5.1.4 Lighting 44
2.5.1.5 Within class variation 44
2.6 Conclusions 46

3 SIFT Based Registration and Recognition 47
3.1 Technique 48
3.1.1 Building the Gallery Database 48
3.1.2 Feature Detection 49
3.1.3 Registration Calculation 50
3.1.4 Distance Measure 52
3.2 Evaluation 53
3.2.1 Recognition 53
3.2.2 Robustness 55
3.2.2.1 Gallery 55
3.2.2.2 Results 55
3.3 Conclusions 58

4 Ear Model Construction 61
4.1 Related Work 62
4.2 Defining the Problem 64
4.3 Technique 66
4.3.1 Base Mesh 66
4.3.2 Training Data 67
4.3.3 Preprocessing 67
4.3.4 Automated surface classification 68
4.3.5 Feature Points 69
4.3.6 Initial Registration 70
4.3.7 Precise Registration 71
4.3.8 Optimisation Based Fitting 72
4.3.9 Optimisation Algorithm 73
4.3.10 Sparsity 74
4.3.11 Robustness 74
4.3.12 Data Normalisation 75
4.4 Morphable Model Evaluation 75
4.4.1 Automated classification accuracy 75
4.4.2 Fitting consistency 76
4.4.3 Model metrics 76
4.4.3.1 Generalisation 76
4.4.3.2 Total Variance 77
4.5 Conclusions 78

5 Conclusions and Discussion 81
5.1 SIFT Based Ear Recognition 81
5.2 Morphable Model of the Face and Ear 82
5.3 Further Work 83

5.3.1 Completion of the Morphable Model 83
5.3.2 Morphable Model Fitting 83
5.3.3 Comprehensive Evaluation 85
5.3.4 Optimised Configuration Parameters 85
5.3.5 Further Modelling Techniques 86
5.4 Concluding Remarks 86

A Model Initialisation 87
A.1 Existing Research 87
A.1.1 Occlusion Robust Detection 87
A.1.2 Precise Localisation 88
A.1.3 Model Fitting 90
A.2 Technique 92
A.2.1 Inverse Compositional Alignment with Model Likelihood 92
A.2.2 Training Data 94
A.2.3 Robust Learning 94
A.2.4 Model Registration 95
A.2.5 Robust Non-Linear Optimisation Fitting 95
A.3 Evaluation 96

References 99

Acknowledgements

Firstly I would like to thank my supervisor, Professor Mark Nixon who has always been encouraging and helpful and who, along with the other members of the Information: Signals, Images and Systems research group, has made the development of this PhD the most enjoyable experience of my working life.

Thanks also to my parents for providing emotional support throughout my PhD. In particular, thanks go to my father, who has provided invaluable advice. Thanks to his reassurance and guidance, the development of this work has been much smoother and less stressful than it would have been otherwise.

And lastly, and most importantly, I would like to thank my wife, Alison, who throughout our time together has been endlessly loving and supportive. Without her I would have achieved very little.

Chapter 1

Introduction

Ideally, automated vision-based biometric systems should match human recognition performance over the same diverse range of observation conditions. Currently, however, constraints need to be imposed on lighting and other factors to make accurate recognition possible. This thesis considers techniques for relaxing such constraints. The work is based on recognising individuals from their ears. The sections that follow provide a brief history of biometrics, especially the use of the ear, and highlights the factors that affect accurate recognition. The remainder of the chapter summarises the contributions of the research and the overall structure of the thesis. The work can be seen as a contribution to both robust recognition and a further validation of the ear as a biometric.

1.1 Biometrics

The field of biometrics can be traced back to the 1800's when a Parisian anthropologist, Alphonse Bertillon, developed a means for identifying subjects from body measurements [15]. Unfortunately, the complexity of his approach limited its use and its accuracy. Also, the technique was effectively discredited in 1903 when its use in Fort Leavenworth, Kansas failed to distinguish between two prisoners with the same name (William West) and body measurements. As a result, attention switched to fingerprinting for identification.

The use of fingerprints for biometrics was developed by Henry Faulds in 1880. His work led to the widespread adoption of fingerprint analysis in criminal investigations. Fingerprint recognition was one of the first biometrics to be automated, with functioning systems appearing as early as the 1960's. This technology has become an invaluable tool for law enforcement and is used routinely as part of criminal investigations. In particular, the FBI's Integrated Automated Fingerprint Identification System (IAFIS) is currently the largest biometric system in the world, containing the fingerprints and criminal history of more than 47 million subjects.

Another early automated biometric was the face. Work in this area was initially motivated by the key role that face recognition has in human interaction. Face recognition continues to be an active research area and has led to the creation of many practical, commercial face recognition systems such as FaceIt[1] and Animetrics[2].

One of the weaknesses of using either the face or fingerprints for recognition is the ease with which these physical features can be altered or damaged. A more secure alternative is the retinal scan, first developed by the EyeDentify corporation. This uses the pattern of blood vessels at the back of the eye for identification. The first practical retinal system was developed in the US in 1981[3] and has been used for high security applications ever since. Examples include verification to prevent welfare fraud, and access control to secure environments such as prisons or government agencies, including the FBI, CIA and NASA[4].

One drawback of the retinal approach, however, is that the scanning procedure is invasive, requiring close contact between the eye and the sensor. A more acceptable alternative is iris recognition, a concept first patented by Leonard Flom and Aran Sarir in 1987 [48]. To exploit their patent they worked with John Daugman to develop an automated system, the first working version of which was produced in 1995 [35]. This achieved high accuracy for very large datasets[5]. Again, iris recognition is used for high security applications and is an alternative to passports in some airports. However, one limitation of the approach is that it requires close cooperative interaction to be effective.

In 1998, following terrorist attacks on US institutions, DARPA began researching human identification at a distance[6]. This work sought to identify individuals remotely to provide an early warning for military and homeland defence. It included research in the new field of gait recognition [97]. The initial study, which ran from 2000 to 2004, demonstrated the viability of the technique for identification. Since then, research in gait recognition has continued and it remains one of the few biometrics that can be used over large distances. However, further work is needed to bring these experimental systems into commercial use.

The emergence of new biometric technologies has led to new applications, with examples ranging from the protection of personal information on encrypted data storage[7] to the prevention of vote rigging in elections[8].

A significant limitation of existing commercial biometric systems, however, is their dependency on controlled environments and cooperative interaction to produce accurate

[1] FaceIt http://www.l1id.com/pages/71-facial-screening

[2] Animetrics http://www.animetrics.com

[3] Patent No. 4877322, Robert V. Hill, 1989

[4] Retinal Scan http://en.wikipedia.org/wiki/Retinal_scan

[5] Independent Testing of Iris Recognition Technology http://www.biometricgroup.com/reports/public/ITIRT.html

[6] Human ID at a distance http://w2.eff.org/Privacy/TIA/hid.php

[7] BioDisk http://www.card-media.co.uk/biodisk.htm

[8] Mexico deploys multi-biometric voting system, Biometric Technology Today, 2006

results. Recent research has therefore considered ways of increasing system robustness to reduce such constraints. One innovation has been the use of the ear as an alternative biometric feature. It offers a number of advantages. In particular, it has a wide variation in appearance between individuals and, like the face, its structure is visible at a distance. However, unlike the face, its appearance does not alter with expressions. Also, it is rarely disguised by makeup and is believed to remain similar in appearance with age, although ears do continue to grow in size.

This thesis examines ear recognition in detail and contributes to the field through the development of techniques that improve robust recognition accuracy, making progress towards fully unconstrained recognition.

1.2 Ear Recognition

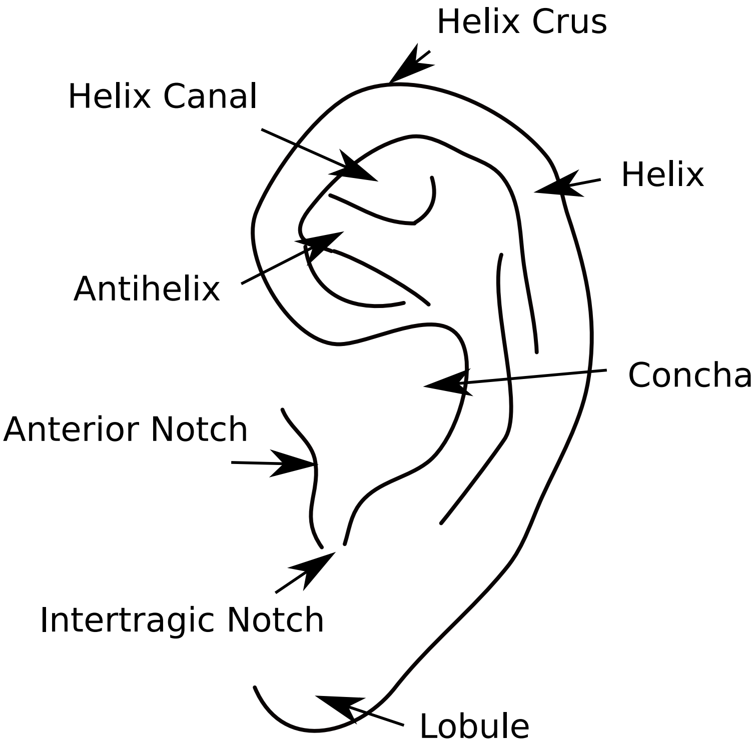


Figure 1.1: The main anatomical features of the ear

Ear height was included in the first biometric system developed by Alphonse Bertillon [15], but it was not until 1955 that a criminologist, Alfred Iannarelli, developed a practical recognition process based solely on the ear [57]. In developing this process, he gathered and analysed over 10,000 ear photographs to demonstrate that they could be used for accurate recognition. Like fingerprints, ear prints have been used in the police service as a forensic tool, and in 1967 their analysis provided key evidence in a criminal prosecution [81]. Ear prints have continued to be used in cases as recently as 2008; however, at least one conviction has been overturned on appeal due to insufficient ear print quality[9].

In 1998, Burge and Burger proposed the first computerised ear recognition system [22]. Their technique used an adjacency graph built from Voroni regions of ear curve segments. Although their paper did not include recognition results, it prompted further studies into the effectiveness of ears as a biometric. Force fields [56], neural networks [76], genetic algorithms [115], geometric features [31], active shape models [69] and shape from shading [26], have all been used to achieve accurate recognition on small collections of ear images taken under controlled conditions. However, the recognition performance of these techniques is greatly reduced when subjects are recorded under less constrained conditions, as, for example, with different poses or with different lighting [29]. Current approaches tend to be highly sensitive to these variations.

Addressing such sensitivity is the focus of this thesis. The overall hypothesis is that:

The constraints on the use of ear-based biometric systems can be relaxed significantly through the introduction of robust recognition algorithms.

1.3 Robust Recognition

There are five main factors that affect accurate ear recognition:

- *Background:* the difficulty of finding the ear in a specific context that may be cluttered by other objects.
- *Occlusion:* the difficulty of finding the ear when partly obscured, for example by hair, a hat or earrings.
- *Lighting:* the amount of light on an ear, and the direction and colour of that light.
- *Pose:* the angle at which the ear is viewed (out of plane rotations).
- *Camera:* the particular attributes of the camera, including its field of view, its sensing resolution, colour sensitivity and any noise in the image produced.

To reduce the impact of these factors, this thesis examines two novel approaches to robust ear recognition. The first uses SIFT feature points [68] to detect and align known samples of subjects' ears with an image to be identified. SIFT points are a highly robust means of matching distinctive points between images. They define both a location, which includes a position, scale and orientation, and an associated signature calculated from the image region around the point. SIFT points have been shown to retain similar signatures under a wide range of variations, including pose, lighting, field of view and resolution [75].

[9] State v. David Wayne Kunze, Court of Appeals of Washington Division 2, 1999

This technique represents a significant step towards confirming the research hypothesis. However, the approach has some drawbacks. In particular, it remains sensitive to light direction and large pose variation. In addition, the approach is too computationally expensive for large datasets.

One way to address these issues is to use a Morphable Model [17] of the 3D surface of the ear. Optimisation techniques can be used to adjust the model so that it corresponds with an ear within an image. The resulting shape parameters provide a unique signature that can be used for recognition. This signature remains constant regardless of the ear's pose or lighting conditions. Also, as the signature is a short vector it can be compared very rapidly with a large database of existing subjects, resulting in a significant efficiency gain over the SIFT based technique.

1.4 Contributions

The work described makes two significant research contributions:

- *The creation of a novel ear image registration and recognition algorithm, which uses SIFT point correspondences to calculate homographies and robust image distances to rank ears.* Analysis on a range of datasets demonstrates the technique to be robust to background clutter, viewing angles up to 13 degrees and with up to 18% occlusion. In addition, recognition remains accurate with ear images as small as 20x35 pixels.
- *The development of a process for the construction of a general model of ear appearance from partial, noisy and occluded range images.* This approach includes the identification of a key set of ear feature points to enable accurate registration. It also provides a partially automated technique for pre-processing training images to remove occlusions and noise. In addition, the approach includes a novel technique for evaluating the quality of the registration of training scans, and metrics to assess the resulting model quality. These metrics are used to quantify the improvement in the generalisation capabilities of the model as the training set size is increased. The evaluation shows that the described technique extracts a consistent shape associated with the identity of the individual and that within the error margins of the registration process, 160 training samples are close to achieving convergence of the model.

1.5 Publications

There have been four publications from the work so far:

- The SIFT based technique was presented at the Biometric Theory, Applications and Systems (BTAS) conference in 2008 [23].
- A more detailed analysis was then published in a special issue of IEEE Transactions on Systems, Man, and Cybernetics [25].
- The SIFT based technique was also used to investigate the benefits of fusion approaches within a multi-biometric system as described in the chapter of the book Multibiometrics for Human Identification" [95].
- The Morphable Model technique was presented at the 2010 conference for Computer Vision and Pattern Recognition [24].

1.6 Thesis Structure

Chapter 2 reviews the literature on robust recognition within images, particularly with respect to the ear. After introducing the process common to all vision-based biometric systems, the factors constraining effective detection and recognition are discussed. The techniques available for easing these constraints are then reviewed in detail. This review is summarised in tables that assess the potential contribution of each technique. As no technique emerges which is robust to all constraining factors, it is suggested that they be used in combination. As a baseline for improvement, the performance of current non-contact biometrics is summarised, followed by a review of existing research on ear biometrics. The conclusion is that ear recognition, in general, is currently sensitive to the identified constraining factors but that the use of two proposed techniques, based on SIFT point analysis and Morphable Models have the potential to achieve significant improvements.

Chapter 3 describes the first ear recognition technique. It covers the pre-processing of enrolled ear images as well as an overview of the SIFT feature detector. It then explains how SIFT correspondences are clustered to produce ear hypotheses, followed by a consideration of how these correspondences can be used to calculate homographies that align the enrolled ears with the test image. The chapter concludes by showing how the aligned images can be compared robustly to determine the most likely identity associated with the ear. The chapter includes a complete set of recognition results and an assessment of the recognition sensitivity to pose, lighting, occlusion, background clutter, noise and resolution.

Chapter 4 describes the 3D model construction process. By using a Morphable Model approach it is possible to achieve robust and efficient recognition using large datasets. The chapter outlines the existing approaches to Morphable Model construction, noting that existing Morphable Models either explicitly or implicitly avoid accurate ear modelling. The description of the new technique begins with an explanation of the preprocessing algorithm that is used to remove noise and occluders. This is followed by details of the optimisation-based registration algorithm, including the adaptations that have been made to make computation efficient, as well as describing the outlier detection process, which addresses any occluders misclassified by the preprocessing stage. The chapter concludes by describing the evaluation metrics and scope for further improvements.

Chapter 5 concludes the thesis, outlining the future work necessary to construct a fully automated Morphable Model recognition algorithm and providing a summary of the contributions that have resulted from the development of the techniques presented. The chapter also outlines possible directions for further research such as the development of a technique to fit Morphable Models to images. Initial work in the development of this technique is presented in the appendix.

Chapter 2

Robust Recognition

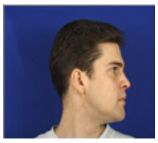
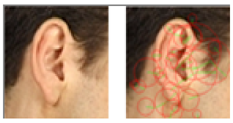
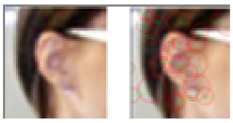
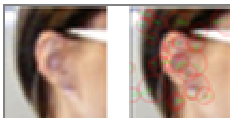
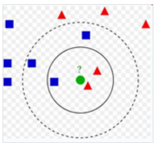
This chapter reviews the literature on robust recognition within images, particularly with respect to the ear. After introducing the process common to all vision-based biometric systems, the factors constraining effective detection and recognition are discussed. The techniques available for easing these constraints are then reviewed in detail. This review is summarised in tables that assess the potential contribution of each technique. As no technique emerges which is robust to all constraining factors, it is suggested that they be used in combination. As a baseline for improvement, the performance of current non-contact biometrics is summarised, followed by a review of existing research on ear biometrics. The conclusion is that ear recognition, in general, is currently sensitive to the identified constraining factors but that the use of two proposed techniques, based on SIFT point analysis and Morphable Models have the potential to achieve significant improvements.

2.1 Biometric Systems

Biometric systems are used to identify individuals, either to *verify* who they claim to be or to *recognise* them from a stored database of identities. Both approaches require *enrolment*, which is the process of constructing a database of known subjects. These stages are summarised in Table 2.1. The first column of the table illustrates the typical image inputs to an ear recognition system. The second column shows a possible output image from ear detection and feature extraction. The last column then suggests how the extracted features are used to enrol, validate or recognise an input ear.

Depending on how the system has been developed, enrolment may involve substantial processing of the recorded biometric feature and measurement of the feature under multiple conditions, such as varying pose and lighting. In the majority of existing systems, this stage is conducted under controlled conditions to maximise the quality of the enrolment data.

Table 2.1: The key stages of a biometric system

	Record Subject	Detect Feature & Extract Signature	Process Signature	Result
Enrolment				Add to database
Verification				Compare to enrolled data
Recognition				Classify

Both verification and recognition can be implemented with similar systems, the main difference being the increased computational cost associated with recognition as it requires a person to be compared against an entire enrolment database rather than a single entry. In both cases, the biometric system must record a feature of the subject, for example an image of their ear, and then compare it with the enrolled information to determine the subject's identity.

If feature appearances were sufficiently controlled, the comparison could simply be between the pixel values of a test image and those in the enrolled image. If equal, the verification succeeds. Similarly, for recognition, as soon as an image is matched, the search of the enrolment database would be complete. In practice, however, even with controlled recognition systems, this level of accuracy is never achieved. In particular, when using unconstrained systems, ears will not necessarily be in the same position. As a result, they must be detected within an image prior to being compared against the gallery.

Detection and validation are closely related problems. In both cases, information extracted from an image must be classified to determine if the image is within a set of appearances. The main difference is that when the system is detecting a class of object, such as an ear, any possible ear appearance should return a positive match. In contrast, when performing validation, only the appearances of one particular person's ear should produce a positive response.

2.2 Unconstrained Factors

Detection and recognition are difficult problems due to the very large number of appearances that a class or an object of a class can produce. For example, the appearance of an ear varies significantly due to lighting and pose. Also, the ear may be located at any position, rotation or scale within an image. This section considers these difficulties and compares the principal approaches that have been developed to address them.

2.2.1 Background

Background refers to the difficulty of finding the ear in a specific context that may be cluttered by other objects. To remove clutter, an object first needs to be located within an image. Once identified, the object can be extracted and compared using other techniques.

2.2.1.1 Position and scale

The simplest approach to determining an object's location is to examine each possible position at a fixed number of scales. This reduces the problem to detecting an object within a localised region, reducing the large amount of variation due to background clutter. However, depending on the detection technique, this approach may be too computationally expensive and the fixed set of positions and scales may not correspond exactly with the object's location. To address such imprecision, detectors are often trained using multiple randomly offset samples. Once detected, there are two main approaches to improving an object's position. The first is to detect sub-parts of an object by, for example, using a feature point detector, and then using these component locations to estimate the object position more accurately. Alternatively, a model fitting approach can be used to improve an estimate of the object's position and scale.

2.2.1.2 Rotation

Some detection and recognition systems assume that an object has a particular orientation in an image. For example, the Viola Jones face detector [104] assumes faces are upright. For some objects, however, even small rotations can significantly alter pixel values, reducing the accuracy of any subsequent recognition process applied to the detected object. The techniques described for determining position and scale can also be applied to compensate for an object's rotation. In addition, objects can have their rotation normalised using the properties of the examined region. Normalisation can be achieved, for example, by using the mean orientation of edges within the region. Alternatively, rotation invariant signatures can be used. These have the same value regardless of the orientation of an image region.

2.2.2 Occlusion

Occlusion, the obscuring of an object, is particularly problematic for ear recognition due to hair and earrings. There are two main approaches to addressing occlusion. The first is to split the object into parts and then try to detect these parts separately. Alternately, an existing technique can be adapted to ignore, or significantly reduce the influence of, occluding pixels by using robust statistics.

2.2.3 Lighting

Here lighting refers to the amount, direction and colour of the illumination of the ear. Lighting variation is a significant factor in reducing object recognition performance. In particular, self shadowing and specular highlights can produce extreme variations in pixel values, significantly affecting accuracy. This was noted, for example, on one of the first large scale studies of face recognition performance: the face recognition verification test (FRVT) [89]. A more recent version of this evaluation has shown substantial improvement, but recognition rates are still at only a tenth of the accuracy of constrained lighting performance [90].

Broadly, lighting variation can be addressed by one of four approaches:

1. A feature of an image that is relatively unaffected by changes in lighting can be extracted, such as edges. If these features are used for comparison, lighting effects will have much less of an impact on performance.

2. The tested region can be normalised to reduce the effects of brightness and contrast.

3. The space of possible appearances due to lighting can be modelled. The distance of a test image from the model then determines the likelihood that the image represents the object. If modelled appropriately, this distance will remain accurate regardless of lighting conditions.

4. The 3D shape of the ear can be inferred using assumptions about the material properties of the surface and the plausible lighting types. This surface can also be estimated using assumptions about other image effects caused by a change in depth, such as reduced focus. Collectively, these are known as *Shape from X* techniques.

2.2.4 Pose

Here, pose refers to the angle at which the ear is viewed and is also known as its *out of plane rotation*. Pose is similar to lighting in that variations in appearance are strongly dependent on the 3-dimensional nature of the object. A common approach to dealing with pose variation is to create multiple detection/recognition algorithms for use over a range of pose angles. However, this requires that objects are enrolled at multiple poses, which may not be possible if enrolment images are from existing data sources. In that case, there are four main alternatives:

1. Calculate the 3D surface shape using 3D sensors or Shape from X techniques. Once obtained, the surface can be aligned in a similar manner to that described in Section 2.2.1.1 for dealing with position.
2. Similar to the process used for lighting variation (Section), the subspace of appearance due to pose can be estimated.
3. A signature can be produced that is insensitive to minor pose variations through, for example, edge orientation binning [68] or an affine interest point detector [74].
4. Fit the image using a 3D model, such as a Morphable Model [17] or a set of pose specific Active Appearance Models [33].

3D Sensors can produce extremely accurate shape calculations, however, they generally place constraints on lighting conditions. Likewise, many shape from X techniques rely on relatively simple lighting and material properties. Addressing more realistic conditions, such as specular lighting effects or skin subsurface scattering, results in an ill conditioned problem which can fail to produce accurate results. In contrast, constructing and evaluating a subspace can be performed more easily. One difficulty, however, is that the accuracy of the subspace is dependent on the size of the training set. In practical situations this will limit the robustness that can be obtained. Alternatively, the robust signature technique can be used. The technique is only valid for small pose variations but it requires no training data, making it suitable for general recognition tasks where there is only one sample of an object to be recognised. Finally, a Morphable Model technique can be used. This approach can achieve high accuracy with relatively little training data. Unfortunately, however, constructing and fitting a Morphable Model is a complex process.

2.2.5 Camera

Different cameras can produce very different images of the same object (Figure 2.1). The following sub-sections describe these variations and some of the approaches used to address them.

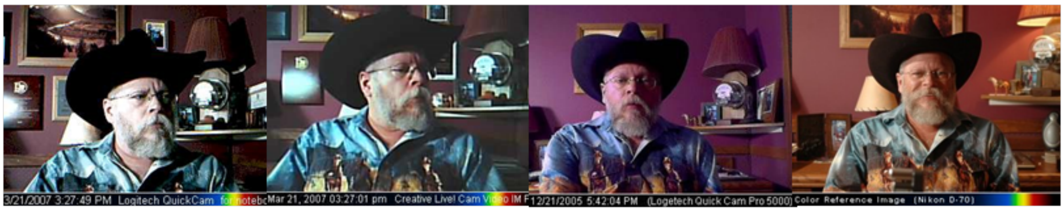

3/21/2007 3:27:49 PM Logitech QuickCam for notebc
Mar 21, 2007 03:27:01 pm Creative Live! Cam Video IM F
12/21/2005 5:42:04 PM (Logetech Quick Cam Pro 5000)
Color Reference Image (Nikon D-70)

Figure 2.1: Similar scenes recorded with different web cameras and a high quality digital camera[1]

2.2.5.1 Field of view

Field of view variations cause similar distortions to those generated by small pose variations. Morphable models can address this variation by including parameters that adjust the camera properties as part of the fitting process. In addition, techniques such as edge histograms and affine signatures are robust to small edge deformations and so remain accurate with small changes in field of view.

2.2.5.2 Focus and motion blur

If an object is not in focus, its appearance will be distorted in a similar way to a focused image blurred with a Gaussian filter [12]. Also, if an object is moving quickly its appearance will be blurred in the direction of motion. These effects will remove details of the object and reduce the accuracy of gallery comparisons or signature generation. In some cases model fitting approaches, such as active appearance or Morphable Models, could be adapted to incorporate a synthesis of these effects in their fitting process. Alternatively, the image can be processed to reduce the blur effects [114]. Other work has explored using these effects to extract additional shape information using *depth from defocus* techniques [88].

2.2.5.3 Resolution

In general, images of objects will have different resolutions according to their distance from the camera and/or the camera's sensor precision. There are three main approaches to addressing this variation. Firstly, an image can be filtered as part of an enlargement process to estimate the high resolution pixel values. Alternatively, if multiple images or video sequences are available, super resolution techniques can be applied [108]. These techniques estimate a higher resolution image from multiple low resolution samples. Finally, if model fitting is used, the region averaging effects of lower resolution sensors can be incorporated into the alignment algorithm to offset their effects [49]. The filtering approach is the simplest of these techniques and can achieve good results provided the

[1] Comparison of web cameras http://cowboyfrank.net/webcams/

object appearance conforms to the smoothness assumptions of the filtering process. However, more accurate results can often be obtained using super resolution or model fitting. The super resolution approach is dependent on the number of images available and their variation. It has the advantage, however, of not requiring a prior model to obtain good results. In contrast, the model fitting technique can achieve highly accurate results using only a single image, but the effectiveness of recognition will depend on the accuracy of the model.

2.2.5.4 Colour sensitivity

When the response of a camera to light frequency and intensity is non-linear, simple image comparison and normalisation techniques can fail to produce accurate results. These non-linear effects can be caused by variations in the sensor, white balance calibration, or automated gain control. In these cases, colour normalisation [46] techniques can be applied. Alternatively, edges can be used. The locations of edges in an image are typically less sensitive to these non-linear variations because of the thresholding techniques used in their detection.

2.2.5.5 Noise

The majority of recognition techniques are designed to either fit a model or measure a difference under the assumption that the main source of error is independent Gaussian noise in each pixel of an image. However, this assumption may be inaccurate when, for example, there are a small number of large error pixels caused by failures in the camera sensor or damage to the lens. In these cases, the noise can be addressed as if it were caused by occlusion. The hidden values of the occluded regions can be estimated using filtering or model fitting approaches.

2.2.6 Within class variation

Within class variation refers to the variation in appearance between different examples of the same object, such as the variation between different people's ears. In the case of detection, within class variation is a form of variation that needs to be reduced, whereas for recognition, the goal is to maximise the difference between different examples so they can be accurately recognised.

2.2.6.1 Intra-subject variation

Implicit in many of the techniques described so far is that the biometric object being detected has a fixed shape and material. However, for most objects there are factors that cause their appearance to vary, such as aging and facial expressions. These effects can be addressed by modelling the space of appearances and separating the variation due to non-identity factors [13]. Alternatively, only the regions that retain a relatively constant shape, such as the ear or nose, can be used to determine identity.

2.2.6.2 Identity invariant detection

In the case of model fitting approaches, the space of possible appearances of an object has already been constructed. In the absence of noise, if a model fails to recreate an image appearance exactly, then the object is not present. In practice, noise in the image and inaccuracy in the model, necessitate placing a threshold on how far an image can deviate from the fitted model appearance before detection is considered to have failed. This threshold value can be estimated by using an additional training data set. For non model-fitting approaches, the set of possible appearances must be estimated and the normalised image, or its signature, compared against this value to achieve detection. By treating each of the elements of the signature, or image, as separate dimensions they can be represented as general vectors. These vectors can then be classified using machine learning algorithms, such as Gaussian mixture models [101] or support vector machines [34].

2.2.6.3 Identification

As with detection, in the absence of noise, only a perfect match between the enrolled and probe vectors will result in a valid recognition. However, all practical systems contain noise and inaccuracies in their models. Therefore it is necessary to find a means of estimating the most likely identity that generated a given vector. Many of the techniques for identity invariant detection can be applied to this problem. When these approaches are used, each subject is treated as a separate object to be detected. In addition, some classification techniques can be extended to distinguish between multiple classes. When techniques cannot be generalised in this way, a multi-class classifier can be created using one binary classifier per enrolled subject. The classifier tests whether the probe represents the subject or the rest of the gallery. By applying each classifier in turn the class can be determined.

2.3 Review of Detection and Recognition Techniques

This section reviews some of the key techniques used in object detection and recognition. The first set of techniques cover general feature extraction and filtering approaches which can reduce variation caused by unconstrained factors. The next set use model fitting approaches, which can be used for both detection and recognition.

2.3.1 Image Processing and Feature Extraction

This subsection describes both image processing techniques and those for general feature extraction. Here, image processing techniques refer to those that alter an input image to reduce variations due to unconstrained factors. These include *filtering*, *super resolution* and *normalisation* algorithms. *Robust distance measures* have also considered here as they are a general approach to removing outlying differences between elements that are being compared.

The reviewed feature extraction techniques process an image to calculate properties which are less affected by unconstrained factors. The features covered are: *edges*, *subregions*, *area transforms*, *histograms*, and *Shape from X techniques*.

2.3.1.1 Filtering

Noise The most common approach to noise is to filter the input. Filters use local information to improve the accuracy of each pixel in an image. Underlying these approaches is a set of assumptions about the objects' likely appearance. For example, Gaussian and Median filtering make the assumption that objects are either smoothly varying or at least locally uniform in appearance. More advanced techniques, such as bilateral filtering [84], make the assumption that images are made up of smooth regions with strong edge separations. To the extent that these assumptions hold, these filters may result in improved performance.

Many image transforms, such as wavelets or edge detectors, perform a local Gaussian blur as part of their calculation. However, these approaches may invalidate some of the assumptions used in inverse rendering techniques such as shape from shading or Morphable Model fitting. In these cases, a robust formulation of the model fitting or shading estimate may produce more accurate results.

2.3.1.2 Super resolution

Resolution There has been significant work in the development of super resolution techniques. These approaches combine multiple low resolution images to create a single high resolution output. Wheeler *et al.* applied this technique to face recognition by tracking the face with an active appearance model and then applying a super resolution technique to the pose normalised tracked face [108]. However, this did not improve the recognition performance substantially, raising it from 51% to 56% on average.

2.3.1.3 Normalisation

There are many potential approaches to normalisation. In general, however, they all adjust the values of a region so that its distribution fits a defined ‘normal’ shape. The following subsections describe how various factors can be normalised.

Position and Scale Position and scale of a region can be normalised by finding the mean location and radius of a property within an image. For example, Abate *et al.* used the centre of mass of ear edges to refine their localisation [1]. Under constrained settings this can produce accurate registrations. However, as with most normalisation approaches, occlusion or background clutter will cause this normalisation to fail.

Rotation Regions can have their orientation normalised by extracting a directional property and then using its distribution to determine a canonical rotation. For example, this approach is used to normalise the signature of SIFT descriptors [68]. The SIFT technique calculates the orientation of edges surrounding a detected point. These orientations are then placed in a histogram and smoothed. The peaks of this histogram provide a set of possible orientations for the detected point. Each peak generates a separate signature. The signature for the point is then rotated so that the peak orientation is at zero degrees. By using this normalisation, SIFT points can be matched regardless of their orientation.

Lighting One of the simplest approaches to dealing with lighting variation is to normalise an image by offsetting and scaling the pixel intensities so that the mean and standard deviation of the image are zero and one respectively. This will remove variation between images due to changes in overall brightness and contrast. However, it will not address image variations due to changes in light direction. In addition, if an object region is not completely segmented from the scene, the normalisation process can be corrupted by the background pixel values. This can produce worse results than if no normalisation had been applied. Other, more sophisticated normalisation techniques, such as Retinex [92], attempt to remove variation due to illuminant colour and magnitude. These techniques perform non-linear transformations of an image based on local distributions. However, their results can be very sensitive to noise.

Colour Sensitivity In a study by Finlayson *et al.* colour equalisation was shown to produce the most consistent colour reproduction in a variety of approaches [46]. However, this process is highly non-linear and will impair analysis by synthesis approaches, such as shape from shading or Morphable Model fitting. Colour equalisation was applied to face recognition by King *et al.* [65]. They used the technique with manually registered head photographs that had no background clutter. The images were recorded with a range of lighting conditions and camera sensitivities. When equalisation was applied, the rank 1 recognition rate was improved from 68% to 86%. These results were obtained with a colour Eigen-face recognition technique [102] using a sample of 120 subjects.

2.3.1.4 Robust distance measures

Occlusion and Noise Detection and recognition algorithms can be made robust to occlusion by using robust distance measures. These measures reduce the influence of samples that are far from an image's expected value. For example, two images can be compared robustly by summing the difference between their pixels as measured by an M-estimator [55]. The M-estimator reduces the magnitude of pixels that have larger than expected variation. Other approaches involve excluding a fixed percentage of pixels that contribute most to the error. These techniques can greatly improve robustness to occlusion. However, they can also lead to a reduction in recognition accuracy. This is because when they are applied to unoccluded images they remove the pixels that are most informative in distinguishing between different subjects.

2.3.1.5 Edges

Lighting A common approach to reducing the effects of lighting variation is to use the edges extracted from an image. This approach is based on the assumption that the majority of edges in an image are caused by occluding regions or colour variations across an object surface. The locations of these edges are relatively unaffected by changes in lighting direction. However, as shown in Figure 2.2, strong lighting variation across the face and ear can result in significant changes in edge locations. This is because many of the edges are produced by self-shadowing effects, rather than material borders or occlusion [14]. Many ear recognition approaches have used edges for recognition. For example, the first ear biometric paper by Burge *et al.* [22] extracted ear edges using a Canny edge detector [27] and then created a signature using Voronoi regions calculated from the edges. Ansari *et al.* also used ear edges, but classified them as convex or concave, using the convex edges to register the ear [6].

2.3.1.6 Sub-regions

Occlusion A general approach, suitable for many techniques, is to split a region to be analysed into multiple sub-regions. These regions can then be tested independently. The results of these tests are combined using a voting scheme where only partial agreement is needed for a detection or recognition to be valid. Provided the shapes of the regions correspond well with the occluded areas, the object can be classified correctly.

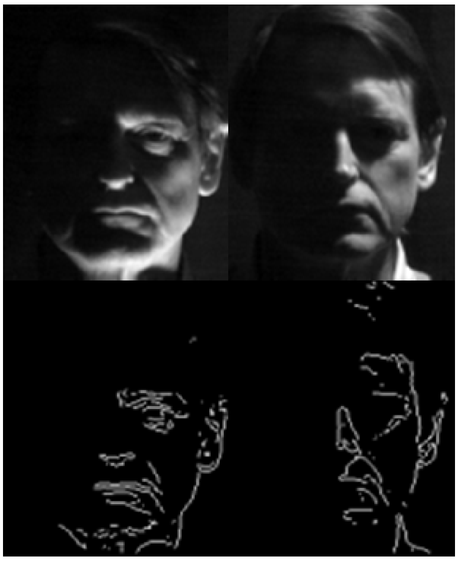


Figure 2.2: The variation in detected edges caused by a change in lighting direction [14]

Some other classifier systems learn these sub-regions as a result of the training process. For example, the Viola Jones detector uses a cascade of weak classifiers to detect a face. The cascade is made up of many localised classifiers that are combined to produce a single accurate detector. Each classifier has an associated weight which is added or subtracted to a confidence value depending on whether it passes or fails the classifier. If the sum of these tests raises the confidence value above a learnt threshold then the object is identified. Each classifier examines a sub-region of the image using an efficiently calculated Haar-like wavelet filter [104]. Because of this structure, a good detection in one region can balance out a failed detection in another. This results in a degree of occlusion robustness, as only a partial match is necessary to detect an object.

2.3.1.7 Area transform

The term *area transform* is used here to describe transforms that convert an image region into a signature vector of values. Each element of the resulting signature is produced by combining many local pixel values, often using a weighted sum. Various area transforms are available, including wavelets and Fourier descriptors. Area transforms are frequently used to summarise large image regions using a relatively small number of signature dimensions.

Position and Scale If the appearance of an object contains high frequency information, misalignments due to variations in scale and position can result in significant differences between gallery and probe images. These differences can be reduced by applying transforms which extract the low frequency components of an image. Such transforms can be applied globally, as in the case of a Gaussian blur, or locally through wavelets. Low frequency components have been used in many detection and recognition techniques. For example, Gabor wavelets are used to generate iris signatures in many state of the art iris recognition systems [36]. They are also used in the elastic bunch graph matching algorithm [109] to perform facial feature detection. They also form an important component of the Viola Jones face detector.

A similar principle underlies the force field transform that has been developed for ear recognition [56]. This approach performs a blur-like operation on an image by treating each pixel as the source of a force field with similar properties to an electrical field. The peaks and troughs of this smoothed image are then used for localisation and comparison. The effect of this processing can be seen in Figure 2.3. When applied to a controlled dataset of 63 subjects, using template matching to compare the peaks and troughs of enrolled images and probes, a 99% recognition rate was achieved.

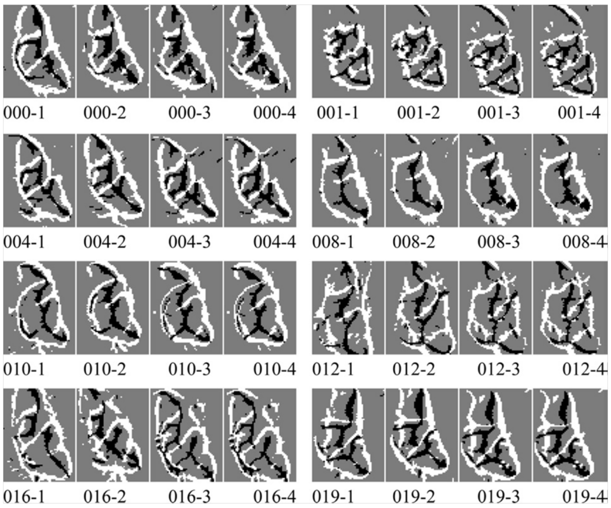


Figure 2.3: Detected peaks and troughs of a set of force field filtered ears [56]

Rotation Area transforms can also be used to calculate a signature that is invariant to rotation. This can be performed by using a property of the Fourier transform. Fourier transforms of 2D signals, such as images, can be separated into magnitude and phase information. The magnitude values remain constant regardless of how the image has been shifted in the x, y plane.

If images are converted into a polar representation (Figure 2.4) the Fourier transform will produce a signature that is invariant to rotation. This approach has been used by Abate *et al.* to recognise ears [1]. On a controlled dataset of 70 subjects they achieved a 96% recognition rate. They also demonstrated that a 15 degree ear rotation was sufficient to halve the recognition rate of a non-rotation invariant approach based on Eigen ears.

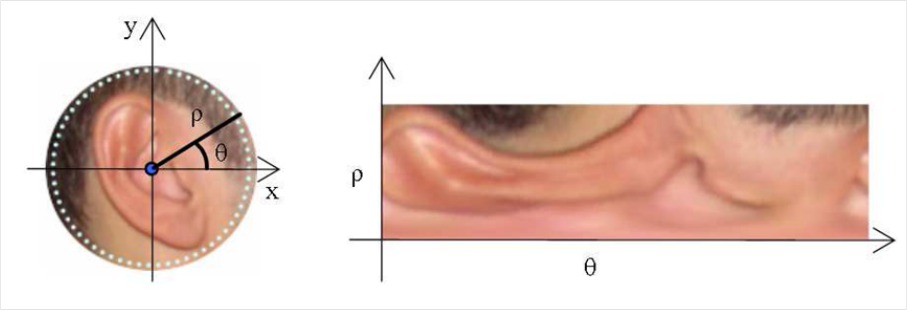


Figure 2.4: An example of preprocessing an ear image into a polar representation before applying a Fourier transform to create a signature [1]

2.3.1.8 Histograms

Many techniques which compare images are sensitive to small local variations in object appearance. One way to reduce this sensitivity is to represent a region using a low resolution histogram. The histogram values are produced by summing all features within a set of discrete locations. As a result, variations that do not cause the properties to move from one region to another will produce the same signature.

This technique is used to describe SIFT feature points. Their signature is calculated by splitting the region around a point into one of 4x4 locations. Each of these locations contains a histogram which sums the contained edge magnitudes into one of 8 different directions. This results in a 128 dimensional signature.

Recent work analysing the accuracy of various feature point techniques has shown that, using this edge histogram technique, SIFT has a matching accuracy of 50% for viewpoint changes up to 50 degrees [75]. A similar approach was used by Jeges *et al.* to detect and initialise ear locations [61]. In an image representation, each pixel was generated by classifying the orientation of the ear image edges into one of eight orientations. By template matching these edge orientation images with probe video sequences, ears were correctly detected and localised at angles of up to 50 degrees.

2.3.1.9 Feature points

Position, Scale and Rotation Feature points are positions on an object which can be reliably detected in images across a range of variations such as position and scale. By matching a set of these points, a transform can be calculated that accurately determines the position, scale and rotation of an object within an image. Feature points can be detected using either custom detectors, such as a Viola Jones eye corner detector [104], or with general feature point techniques, such as SIFT [68](Figure 2.5). Custom feature detectors are themselves object detectors and therefore must address each of the unconstrained factors. In contrast, general feature point techniques efficiently process

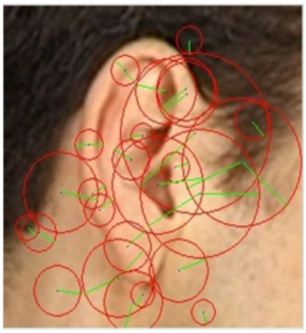


Figure 2.5: An example of the SIFT feature points detected on an ear image

an image producing many points with associated signatures. These signatures can then be matched between images to determine correspondences.

General feature points are detected by iterating over an image at multiple positions and scales, calculating a regional property. Local peaks in this property result in feature point detections. Different techniques use different properties. For example, SIFT uses peaks in a difference of Gaussian pyramid, whereas the Kadir Brady saliency detector uses peaks in local Shannon entropy [63]. Once these points are detected, the regions surrounding them can be processed to calculate a robust signature.

Occlusion In practice only a small subset of points are required for accurate detection and localisation, making the approach very robust to occlusion.

Pose Affine feature point detectors create signatures that are invariant to position, scale, rotation and pose variations. If the corresponding locations on an object are approximately planar, these signatures can be very accurately matched across multiple poses. For example, when tested with planar surfaces, the Harris-Affine detector was found to have a 38% repeatability of feature point detection across 70 degrees of variation [75].

Another approach, developed for general object recognition [20], is to approximate an object as a single planar surface. By matching multiple feature points between probe and enrolment images a homography transform can be calculated. This can be used to align the enrolled images with the probe and thus enables accurate comparison across affine transformations. This approach, however, is limited by the robustness of the feature point detector and the degree to which the object can be approximated by a plane.

2.3.1.10 Shape from X

There are a number of different approaches that use image properties to infer the 3-dimensional shape of a scene. The shading, texture, focus or even edge directions can be used to estimate shape. In the case of ears there is little variation due to focus or texture and no simple properties of edge directions, unlike buildings, for example. However, the ear is a smoothly varying ornate shape making it suitable for shape from shading techniques [26]. Such techniques work by making certain assumptions about the surface materials and the lighting of an object. Under these assumptions the surface can be reconstructed from a single image.

Lighting By comparing the reconstructed surfaces, objects can be recognised under a range of lighting conditions. One of the challenges of any shape from shading approach is the under-constrained nature of the reconstruction problem. For example, the intensity variations in an image may be caused by surface colour, lighting variation or surface shape. Work by Zhou*et al.* on the shape from shading of faces has reduced this ambiguity by imposing a facial symmetry constraint [118]. If the face is symmetric, pixels that appear on opposite sides of the face should have the same skin colour and reflected surface normals. This constraint enables a unique solution to be obtained. Other work by Smith [99] uses a model of face normals to constrain the results and improve reconstruction results.

Shape from shading has been applied to ear recognition by Cadavid and Abdel-Mottaleb [26](Figure 2.6). They achieved an 84% rank 1 recognition rate which although reasonably accurate, is lower than that for other published techniques. One explanation for the reduced performance is that ears have significant self occlusion. This introduces self-shadowing effects, which may introduce errors as they are ignored in most shape from shading techniques.

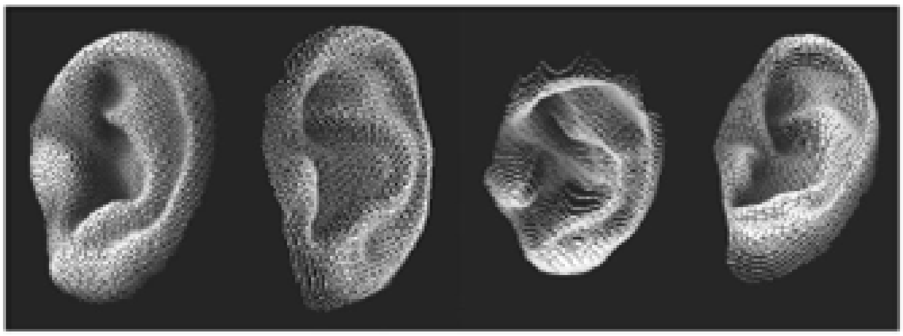


Figure 2.6: Examples of 3 dimensional ear shapes created using shape from shading [26]

Pose As with lighting variation, the reconstructed 3D shape of an object can be compared across subjects with varying pose. However, as the details of the interior of the object may be occluded at different poses, this shape will be incomplete. In practice, this limits the degree of pose variation to which compensation can be applied accurately.

2.3.2 Model based techniques

This section reviews a number of model based recognition techniques. The first three approaches, *template matching*, *the Hough transform* and *model fitting to interest points* are approaches for matching models to images or other features. These techniques are followed by a description of four increasingly complex models: *Subspaces*, *Active Shape Models*, *Active Appearance Models* and *Morphable Models*.

2.3.2.1 Template matching

Position and Scale The template matching algorithm iterates over an image at a fixed number of positions and scales comparing the local region with a template. In a simple algorithm this match may consist of a pixel comparison between the test image and an enrolled sample. Other variants of this technique use different distance measures and may include additional processing of the test image. For example, Abdel-Mottaleb *et al.* use Hausdorff edge template matching between an example ear helix edge and edges identified on skin coloured regions of an image [2].

The same approach of iterating over local regions is used in many other techniques. For example, the SIFT feature point detector iterates over a difference of Gaussian (DOG) pyramid [68]. The pyramid is created by applying progressively larger Gaussian blur operations to an image. These images are then subtracted to get the DOG values. Iterating over this pyramid is equivalent to scanning the image at multiple positions and scales. Regions where this value is a local maximum are used as interest point locations. A similar iteration approach is used in the Viola Jones detector [104]. The detector examines the image at multiple positions and scales applying a classifier to determine if the local region contains the object.

2.3.2.2 Hough transform

The Hough Transform is used to detect objects in images using features such as edge pixels [58]. Each feature pixel in an image is used to accumulate votes for possible object locations and shapes that could have generated it. These votes are held in a histogram, known as the accumulator space. If an object is present in an image this technique will produce a large peak in the space. The parameters of the peak define the location of the object. The approach is highly robust and accurate; however, its cost is exponential in the number of parameters that define the location and shape of the object. This makes it impractical for objects that have a large space of potential appearances.

Position, Scale and Rotation The Hough transform approach can be highly robust to background clutter. This robustness is achieved as the accumulator values of non object edges are unlikely to be correlated. Only those edges generated by the object will contribute consistent values resulting in a peak.

Occlusion The Hough transform represents one extreme of the recognition by parts approach. As the detection uses each edge within an image, only a small number of matches are required to produce a peak in the accumulator space. This technique is therefore highly robust to occlusion. The Hough transform was used by Arbab-Zavar *et al.* to detect the ear [7]. They recognised that the outer curve of the ear could be approximated by an ellipse. Ears therefore create peaks in the accumulation space of a Hough transform designed to detect this shape. As a result, in scenes with limited clutter and 30% occlusion, ears could be detected with 90% accuracy.

2.3.2.3 Model fitting to interest points

Some objects lack distinctive features that can be reliably and precisely detected across variations. In such situations, feature detection can be refined with a model fitting approach. One example is elastic bunch graph matching [109]. It proceeds by first estimating the location of a set of feature points using a Gabor jet based detector [109]. An average object model is then fitted to the detected points. The fitted model constrains the relative position of the points and provides an improved estimate for their location. The region surrounding the new estimates can then be searched using the feature point detector to improve their localisation. The identity of the object can then be determined based on the similarity between the feature point regions and the relative lengths of the edges of the fitted model.

A related approach is the SoftPOSIT algorithm [37], developed to localise 3D objects when feature point correspondences are not known. The algorithm alternates between estimating point correspondences and model fitting. The correspondences are calculated using the closest one to one mapping of detected and model points. This technique is particularly useful when the object being recognised has many locally similar points as, for example, in the case of windows on a large building.

2.3.2.4 Active shape models

An alternative to model fitting using feature points is to fit using edges. This can be achieved using an Active Shape Model [32] (Figure 2.7). These models represent the possible shapes the edges of an object can produce. They can be aligned by calculating the distance from edges of the model to the nearest edge pixels extracted from an image. An optimisation step can then be used to adjust the position, rotation, scale and shape parameters to minimise these distances.

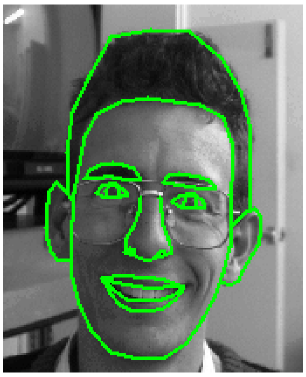


Figure 2.7: A fitted Active Shape Model [32]

Position, Scale and Rotation Provided they start with an approximately correct solution, Active Shape Models can converge to a very accurate estimate of the position, scale and orientation of an object within an image. They have been applied to ear recognition by Lu *et al.* [69]. With manual initialisation, on a gallery of 56 subjects they achieved a 93% rank 1 recognition rate.

2.3.2.5 Subspace

By treating each pixel as a separate dimension, an image can be viewed as a high dimensional vector. The set of all images of an object can then be expressed as a high dimensional subspace. Subspace techniques use linear algebra and kernel techniques to help estimate this space from sample vectors.

A popular example of this approach is the Eigenface recognition technique developed by Turk and Pentland [102]. This technique calculates a model of facial appearance by calculating the Principle Component Analysis of a large set of registered face images. The Eigenvectors produced by this calculation represent the main directions of variation between the training images (Figure 2.8). The distance of a probe image from this space gives an estimate of the likelihood it represents a face. This approach has been applied to ear recognition by Chang *et al.* [29]. In their work they used a dataset of 88 subjects' faces and ears. Under controlled recognition conditions, with gallery and test data recorded on different days, they achieved the same 70% rank 1 recognition rate for either ear or face.

Lighting The light field technique [14] can be used to recognise objects under varying lighting conditions. Such flexibility is achieved because the technique accurately models the space of possible appearances of an object due to light from different directions. The appearance of a non-emissive object is due to the sum of the reflected light from it. In addition, as the path of each light ray interacting with an object is independent of any other, the space of possible object appearances can be expressed as a linear sum of the

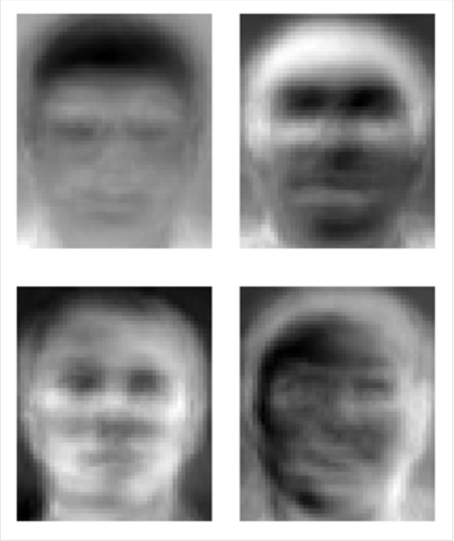


Figure 2.8: An example of some Eigenvectors for a face dataset produced by AT&T Laboratories Cambridge [96]

reflected light due to the incoming light rays (Figure 2.9). If the incoming light is of low frequency, such as daylight, the space of appearances can be accurately approximated by a low dimensional subspace (Figure 2.10). In the case of a class of objects, for example all ears, this subspace can be approximated by combining the lighting subspaces of many training samples. The combined subspace may be further reduced by finding the principal components of this set of spaces and using that to represent the class. This approach is known as the Eigen light field technique and has been used with facial light fields to achieve high recognition accuracy under highly varying lighting conditions.

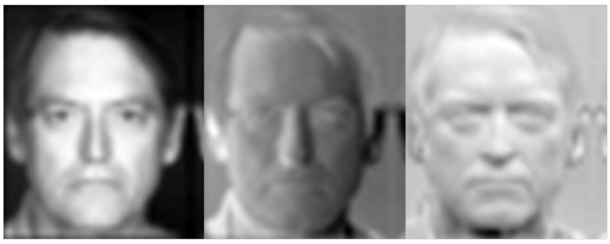


Figure 2.9: The mean and first two basis images of a face light field [14]

Pose As small pose variations create small variations in object appearance, the set of all such object appearances forms a smooth nonlinear shape in the possible image space. One way of achieving pose invariant object recognition is to estimate this shape and then determine whether a probe image is contained within it. With sufficient samples, this shape can be estimated using non-linear methods such as kernel support vector machines [34]. In addition, with sufficient training samples, tensor methods can be used to estimate the space of variation of a particular object from one example [105].

These approaches need very large datasets to estimate the spaces accurately. One ap-

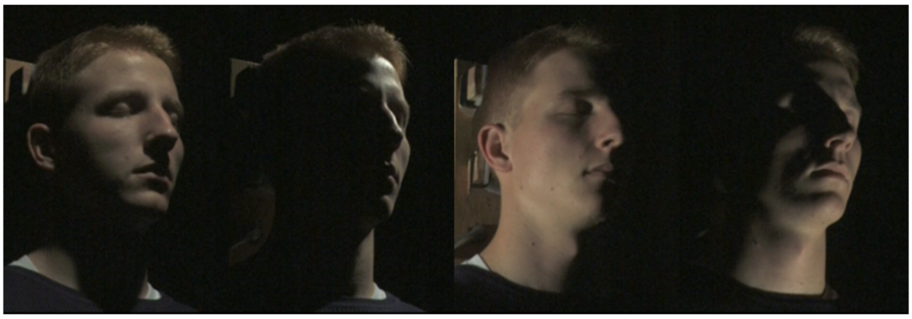


Figure 2.10: A set of some of the images that can be used to calculate the light field of a face [40]

proach is to fit a Morphable Model to a set of enrolled subjects and then use the model to synthesise their appearance with varying lighting conditions and poses. This approach was used by Huang*et al.* to create an SVM-based face detector [54]. Their system maintained a 90% recognition rate with a false acceptance rate of 10% when applied to a dataset containing small pose variations.

Noise Images can be projected into the subspace by calculating their distances from the mean along the largest Eigenvectors. Comparing enrolment and test images in this model space will produce more accurate results [102]. This is because the changes due to noise and variation are unlikely to be strongly correlated with the changes due to identity. Therefore projecting the image into the model space will increase the correlated identity-based variation while decreasing the uncorrelated noise differences.

2.3.2.6 Active appearance models

An Active Appearance Model is a generalisation of an Active Shape Model [32]. It combines a model of shape variation with a subspace model of surface texture variation (Figure 2.11). In this way the whole appearance of an object can contribute to its alignment. A model can be aligned with an image by adjusting the parameters of its appearance and location to obtain an optimal fit. This can be achieved using an optimisation algorithm that minimises the pixel difference between the rendered model and the image [71]. Active Appearance Models are commonly used for face recognition but have yet to be applied to ear recognition.

Position, Scale and Rotation As with Active Shape models, Active Appearance Models can achieve very accurate position, scale and rotation estimates. However, they also require a close initial estimate. Without this estimate the fitting process is likely to converge to an incorrect local minimum [71].

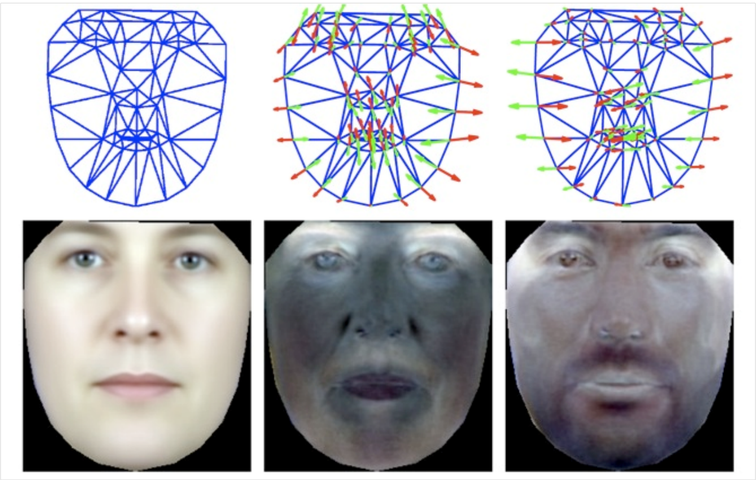


FIGURE 2.11: Images showing parts of the shape and texture models of an Active Appearance Model. The images to the left represent the mean shape and texture. The other 4 images represent the top 2 Eigenvectors of shape and texture variation [71]

Resolution Recent work by Dedeoglu *et al.* [49] has extended the model fitting of Active Appearance Models to include a model of the image formation process. This modelling enables much more accurate fitting to low resolution data sources (Figure 2.12).

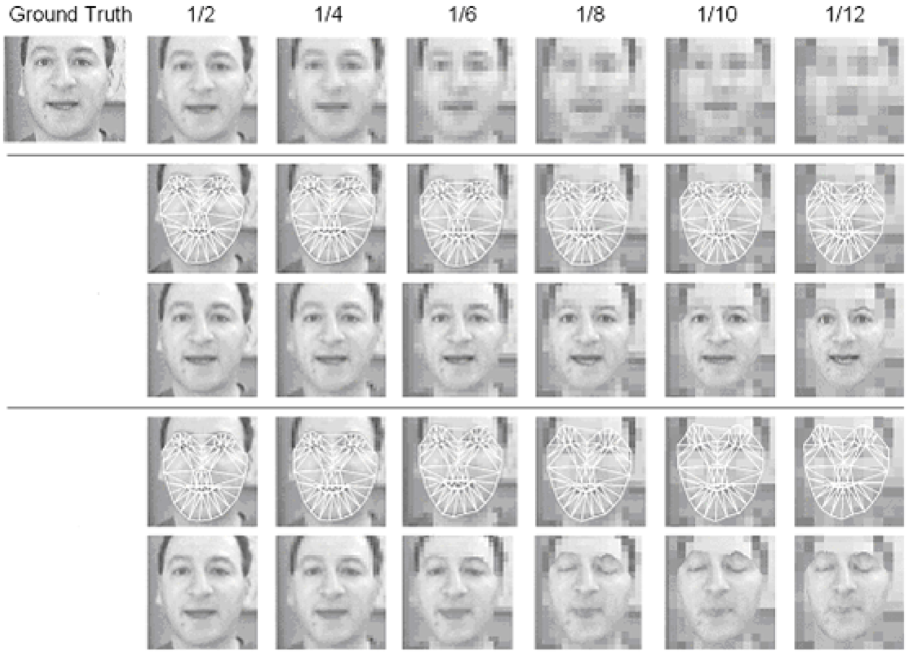


FIGURE 2.12: The improvement in fitting possible with a robust formulation [49] Top: Ground truth and reduced resolution input Middle: Results with robust formulation Bottom: Normal AAM fitting results

2.3.2.7 Morphable models

Morphable Models [17] are effectively a 3-dimensional generalisation of Active Appearance Models. They consist of a 3D morphable shape and a subspace texture colour model. As with Active Appearance Models the mesh and texture have a set of basis vectors that reflect their space of possible values. In addition to these shape and texture parameters a fitted instance of a model also includes an estimate of the light surrounding it and the properties of the camera used to produce the image. This information is sufficient to render the model accurately using computer graphics. The fitting process involves adjusting these parameters until the rendered model matches the image. The fitting is highly non-linear and early techniques required stochastic optimisation to achieve accurate alignments [17]. Recent work has smoother fitting constraints that achieve more efficient and robust model fitting [93]. Once a model has been aligned with an image the identity of the subject can be determined by comparing the parameters of shape and colour with those of subjects in the gallery. By modelling an objects appearance in this way, recognition is highly robust. For example, this technique has produced the most accurate recognition results on the highly challenging CMU PIE face dataset [98].

Lighting As the modelling process includes a simulation of the object's appearance due to lighting, Morphable Models are highly robust to these variations. Blanz and Vetter's work approximated the lighting using multiple directed lights [17]. More recent by Zhang *et al.* work has extended this approach using spherical harmonic lighting models for more accurate reconstructions [116].

Pose Morphable Models are highly robust to pose variation as they contain correlations between an object's appearance at multiple angles. By fitting a model to an image the most probable shape and surface texture of hidden object regions can be estimated. In this way, objects can be recognised across large pose variations even when there is relatively little overlap in the visible regions of an object (Figure 2.13).

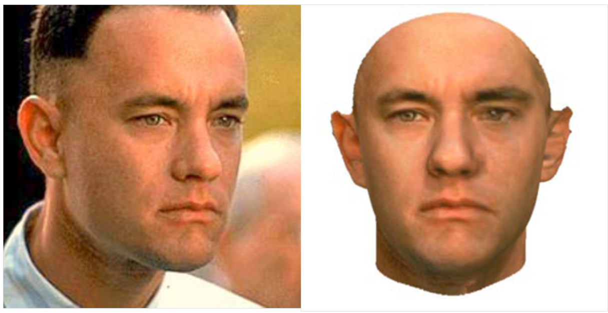


FIGURE 2.13: An image that has been fit with a Morphable Model, and the model viewed at a novel pose revealing the estimated shape and texture of hidden head regions [17]

Intra Subject Variation Morphable models can also be used to address subject variations such as expressions and aging. For example, Park *et al.* [85] used a simplified Morphable Model and multiple enrolment images to normalise subject ages before comparison. When these normalised images were used with a state of the art face recognition system, an improvement of 9% in rank 1 recognition was achieved (29% to 38%). These relatively low recognition rates reflect the extremely challenging nature of their dataset. In addition, Amberg*et al.* have used Morphable Models to parameterise 3D range scans of faces with varying expressions [4]. They make the assumption that variations in face shape due to expressions are similar across different subjects. With this assumption they calculated the principal components of mesh changes due to expression. These components were then combined with the principal components calculated from neutral poses to construct a complete Morphable Model. The range scans could then be normalised by fitting the model and setting the magnitude of the expression parameters to zero. Using this approach, with 61 subjects, they significantly improved their verification rates, achieving a zero false accept rate, with a 0.25% false reject rate.

2.3.3 Summary

An assessment of the robustness of each of the reviewed techniques is given in Tables 2.3, 2.4, 2.5 and 2.6. Each technique is rated based on its sensitivity to each of the unconstrained factors. The robustness of each approach has been broadly classified into one of five categories as summarised in Table 2.2.

Table 2.2: The relative levels of robustness

✓✓✓	As robust as possible. Near to complete invariance
✓✓	Works well under some constraints
✓	Slight improvement on image matching
✗	More sensitive than using image matching
NA	The technique has no effect on this factor

Tables 2.3 and 2.4 compare the robustness of the reviewed image processing and feature extraction techniques (Filtering, Super resolution, Normalisation, Robust distance measures, Edges, Sub regions Area transform, Histograms, Interest points, Shape from X) against the key unconstrained factors (Position and Scale, Rotation, Occlusion, Lighting, Pose, Field of View, Resolution, Colour Response, Noise). Some of the techniques apply to one particular issue, such as super resolution and filtering, while others are suitable for many factors, such as area transforms and interest points. Interest points provide one of the most comprehensive approaches to addressing each of the unconstrained factors. This stability is obtained because of the consistency of the signatures associated with the feature point approaches, which in turn are effective because they combine many of the other techniques.

Tables 2.5 and 2.6 compare the robustness of the model based techniques (Template matching, Hough transform, Model fitting to interest points, Active Shape Models, Subspace or light field, Active Appearance Models, Morphable Models) against the key unconstrained factors (Position and Scale, Rotation, Occlusion, Lighting, Pose, Field of View, Resolution, Colour Response, Noise). Generally the model based approaches have greater robustness than the feature extraction techniques; however, it should be noted that many of these techniques require initialisation to perform accurately and so are primarily a refinement used in conjunction with a robust detection algorithm. Of particular interest, are the subspace and Morphable Model approaches as they have the greatest robustness to pose and lighting variations. These factors are particularly important as they have a significant effect on recognition performance. It can also be seen from the tables that no technique is completely robust to all variations. It is therefore necessary to combine techniques to achieve the most effective approach overall.

TABLE 2.3: The location, occlusion, lighting and pose robustness of filtering and feature extraction techniques

Technique	**Position & Scale**	**Rotation**	**Occlusion**	**Lighting**	**Pose**
Filtering	NA	NA	NA	NA	NA
Super resolution	NA[a]	NA[a]	NA[b]	NA[b]	NA[b]
Normalisation	✓	✓✓	✗	✓	NA
Robust distance measures	NA	NA	NA	NA	NA
Edges	✗	✗	NA	✓✓	✗
Sub regions	NA	NA	✓✓✓	NA	NA
Area transform	✓[c]	✓✓✓[d]	✓	✓[e]	✓
Histograms	✓✓[c]	✓✓[c]	✓	✓✓[f]	✓✓
Interest points	✓✓✓	✓✓✓	✓✓✓	✓✓	✓✓
Shape from X	NA	NA	NA	✓✓	✓✓

[a] Provided tracking is accurate
[b] Provided the property doesn't vary while tracking
[c] For small misregistrations
[d] If invariant transform used
[e] If the resulting signature is normalised
[f] If edges are used

Table 2.4: The relative levels of robustness to field of view variations, resolution, colour response and noise of filtering and feature extraction techniques

Technique	Field of view	Resolution	Colour response	Noise
Filtering	NA	NA	NA	✓✓
Super resolution	NA	✓✓	NA	NA
Normalisation	NA	✓✓[a]	✓✓✓[b]	✗
Robust distance measures	NA	NA	NA	✓✓✓
Edges	✗	✓	✓✓✓	✓
Sub regions	NA	NA	NA	✓
Area transform	✓	✓✓	✓✓[c]	✓
Histograms	✓✓	✓	✓✓✓[d]	✓
Interest points	✓✓	✓✓	✓✓✓	✓✓
Shape from X	✓✓✓[e]	NA	✓✓[c]	✗

[a] With upscaling
[b] If unoccluded
[c] If grayscale is used
[d] If edges are used
[e] If field of view is known

2.4 Detection and Recognition using combined techniques

The previous section reviewed the techniques that can help achieve robust recognition. This section continues the analysis by examining the recognition accuracy that is possible by combining these techniques. It starts by summarising the top recognition performances for the most popular non-contact biometrics: Face, Iris and Gait. This is followed by a summary of evaluations which demonstrate how unconstrained factors affect the performance of these systems. The results highlight the need for further developments in robust recognition. A detailed review of existing ear recognition techniques is then presented, examining the combinations of techniques that have proved effective. The review also highlights problem areas that have yet to be resolved and discusses a number of approaches that combine both face and ear recognition. For each technique, the use of the ear is shown to improve recognition performance significantly.

Table 2.5: The location, occlusion, lighting and pose robustness of model based recognition techniques

Technique	Position & Scale	Rotation	Occlusion	Lighting	Pose
Template matching[a]	✓✓	NA	✓✓[b]	✓[c]	NA
Hough transform[d]	✓✓✓	✓✓✓	✓✓✓	✓✓	NA
Model fitting to interest points	✓✓✓	✓✓✓	✓✓✓	✓✓[e]	✓✓[e]
Active Shape Models	✓✓✓[f]	✓✓✓[f]	✓✓[b]	✓✓	✓✓[g]
Subspace or light field	NA	NA	✓✓[b]	✓✓✓[h]	✓✓✓[g]
Active Appearance Models	✓✓✓[f]	✓✓✓[f]	✓✓[b]	✓✓[h]	✓✓[g]
Morphable Models	✓✓✓[f]	✓✓✓[f]	✓✓[b]	✓✓✓	✓✓✓

[a] Assuming image matching is used
[b] With a robust distance measure
[c] If region is normalised
[d] Provided edges are used to detect primitives such as ellipses or lines
[e] Depending on the feature point technique
[f] With accurate initialisation
[g] If trained with pose variation
[h] If trained with light variation

2.4.1 Non-contact biometrics

Biometric systems require large scale studies to validate their reliability. For non-contact biometrics the most recent of these include the Face Recognition Vendor Test (FRVT)[89], the Iris Challenge Evaluation (ICE)[90] and the HumanID Gait Challenge [97]. The FRVT is a large scale evaluation of commercial and academic face recognition systems. It has been run in 2000, 2002 and 2006. Each time, the overall recognition performance improved significantly. Algorithms were primarily assessed using verification performance, which is a measure of how accurately a system can confirm that a person is who they claim to be. The algorithms were compared by fixing the false accept rate and measuring the false rejection rate. The false accept rate is the rate at which imposters are incorrectly allowed through. For the FRVT 2006 analysis this rate was set at one in a thousand (0.001). The false reject rate then refers to the rate at which valid users fail to be recognised. The systems were tested with high resolution frontal

Table 2.6: The relative levels of robustness to field of view variations, resolution, colour response and noise of model based recognition techniques

Technique	**Field of view**	**Resolution**	**Colour response**	**Noise**
Template matching[a]	NA	✓✓[b]	✓✓[c]	✓✓✓[d]
Hough transform[e]	NA	✓✓	✓✓✓	✓✓✓
Model fitting to interest points	✓✓[f]	✓✓[f]	✓✓✓	✓✓✓
Active Shape Models	✓	✓✓	✓✓✓	✓✓
Subspace or light field	✓	✓✓✓[g]	✓✓[c]	✓✓[d]
Active Appearance Models	✓	✓✓✓[g]	NA	✓✓[d]
Morphable Models	✓✓✓	✓✓✓[g]	✓✓	✓✓[d]

[a] Assuming image matching is used
[b] With normalisation
[c] If grayscale and normalised
[d] With a robust distance measure
[e] Provided edges are used to detect primitives such as ellipses or lines
[f] Dependant on the robustness of the interest points
[g] With resolution aware model fitting

still images under controlled lighting. In 2006 the best algorithm achieved an FRR of under 0.01. This is a substantial improvement in performance from 2002 when the best FRR was 0.2.

ICE was conducted in 2006 and is similar to FRVT. ICE showed that iris recognition can achieve a similar performance to that of face, with the best performing algorithm obtaining an FRR of 0.015 for the same FAR (0.001).

Existing work on gait recognition has achieved near perfect results on small subject studies in indoor environments. However, there is no equivalent study to FRVT and ICE, where evaluation is performed by an external group using data not previously available to the algorithm developers. Of the existing data sources, the HumanID Gait Challenge represents one of the more realistic and challenging datasets, involving 122 subjects recorded in outdoor video sequences. The study evaluated the variation in performance due to variations in shoes, surfaces and viewing angle. The 2005 version of the HumanID Challenge included a baseline recognition algorithm which under the most favourable conditions achieved an FRR of 0.17 with a FAR of 0.02. The FRVT, ICE and HumanID performance results are summarised in Table 2.7.

Table 2.7: Recognition rates for Non-Contact Biometrics

Evaluation	Number of Subjects	FRR	FAR
FRVT 2006 (Face)	263	0.01	0.001
ICE 2006 (Iris)	240	0.015	0.001
HumanID (Gait)	122	0.17	0.02

2.4.2 Unconstrained recognition

Existing commercial non-contact biometric solutions need cooperation with the user and require controlled conditions for accurate results. This section describes existing studies evaluating their sensitivity to more challenging situations.

For faces, the FRVT included an examination of the effects of pose, lighting and resolution on verification performance. The 2002 evaluation showed a very large drop in performance when subjects were recorded with pose variations or under unconstrained exterior lighting. In 2002, with 45 degrees of pose variation, the best performing system achieved an FRR of 0.6 with a FAR of 0.01. Also in 2002, the best performing system under varying outdoor lighting achieved an FRR of 0.5 with the same FAR (0.01). However, the 2006 study demonstrated that techniques had improved to a point where, with unconstrained lighting, the best algorithm achieved an FRR of around 0.1 with a FAR of 0.01. Also in 2006, the performance on a low resolution dataset (average between-eye distance of 75 pixels) was evaluated. The best performing algorithm in this case achieved an FRR of 0.02, much closer to the high resolution performance. This information is summarised in Table 2.8.

Recent work on face recognition has focused on improving robustness as well as addressing variations due to ageing, expression and image quality [118] [107].

Iris recognition typically requires significant user cooperation [70] and there are currently no systems that attempt recognition without such cooperation. The closest unconstrained approach is a system developed by Matey*et al.*, which identifies subjects as they walk through a controlled sensor [70]. However, this approach is still far from being practical in unconstrained environments.

In terms of gait recognition, techniques have been developed to recognise subjects walking at different angles to the camera. In addition, recent work has improved gait performance for subjects carrying objects [66]. However, gait recognition is still at an experimental stage and robust commercial systems may not be available for many years.

Table 2.8: Recognition results for unconstrained factors

Evaluation	**Factor**	**Number of Subjects**	**FRR**	**FAR**
FRVT 2002 (Face)	Pose variation 45°	63	0.6	0.01
FRVT 2002 (Face)	Unconstrained lighting different days	103	0.5	0.01
FRVT 2006 (Face)	Unconstrained lighting constrained gallery	257	0.1	0.01
FRVT 2006 (Face)	Low resolution 75 pixels between eyes	37,437	0.2	0.001

2.5 Ear Recognition

In comparison with the face, iris and gait, ear recognition is a relatively young field. Initial work has concentrated on demonstrating high accuracy using controlled datasets. Most of these datasets have minimal noise, little pose variation and uniform lighting. In addition, they constrain the probe image to a single profile head, removing the problem of detection within background clutter. In many cases, ear registration is performed manually, with the techniques focusing on the development of robust distance measures [56]. Manual registration has also been used in the comparison of ear and face recognition. These studies highlight the potential for combining the biometric information from both face and ear to improve recognition results [29]. Of particular significance is a study by Theoharis *et al.* which has shown that the face and ear shape have very low statistical correlation, with a Pearson correlation coefficient of 0.161 [100]. More recently, fully automated recognition systems have been produced [8][61]. These are more representative of true recognition performance as initial detection can be a significant source of error [8].

Recent work has also started to use less constrained datasets, which highlight a sensitivity to pose and lighting variation [39]. In particular, a study by Chang *et al.* shows a drop in recognition performance from 90% to 34% when the gallery and probe images have pose variations of 22.5 degrees [29]. This sensitivity has led to the development of techniques that use 3D laser scans of the ear shape. These 3D approaches have very accurate recognition results on datasets with small pose and lighting variations [59][112]. In addition, techniques have been developed to estimate 3D ear shape from video sequences and shape from shading approaches [26]. These techniques have the potential to improve pose and lighting robustness without requiring specialised 3D sensors.

In terms of evaluation, most research on ear biometrics has concentrated on recognition rates rather than the verification rates used in the more established biometrics. Recognition rates measure the percentage of subjects who are correctly identified from the subjects in the gallery. Most systems produce multiple candidate identities for each test image, which are then ranked according to the estimated likelihood that they match the image. If the true subject is in the top n returned identities the result is considered to be correct to rank n. The rank n recognition rate is then the percentage of test images that are correct to rank n.

Many techniques use different datasets for their evaluation and so accurate comparisons of performance are not possible. However, five datasets have been used in multiple studies: XM2VTS [73], USTB [111], UND_E [29], UND_J2 [112] and FRGC [91]. The XM2VTS dataset was created by the University of Surrey and includes 295 high quality and high resolution head profile images. The USTB ear dataset was produced by the University of Science and Technology Beijing. It contains 79 subjects recorded with pose variations and a subset of 77 with lighting variations. The UND_E, UND_J2 and FRGC datasets were all produced by the University of Notre Dame. The first set, UND_E, contains 114 2D profile ear images. UND_J2 is a larger dataset consisting of profile 3D colour and range scans of 415 subjects. Both datasets have small variations in pose and lighting. Finally, the FRGC is one of the largest 3D face datasets available. It includes scans of 324 people who have also been recorded in UND_J2 and can therefore be used for combined ear and face recognition experiments.

Tables 2.9, 2.10, 2.11 and 2.12 provide a summary of the main results in ear recognition. The tables show the relative degrees of robustness that have been obtained. The tables list research in 2D, 3D and combined face and ear recognition approaches. The research has been broadly sorted based on its robustness to pose and occlusion and on the difficulty of the evaluation datasets. For each algorithm the base recognition rate and the performance under pose variation and occlusion have been included. None of the other factors have been explicitly evaluated in the existing work. 15 techniques based on 2D recognition with manual registration have been compared, as well as an additional six fully automated 2D approaches. For both manual and fully automated techniques, recognition rates of over 90% have been achieved using relatively constrained datasets. However, on the more challenging UND_E dataset, recognition rates are generally within the 80%-90% range. The main exception is the technique of Naseem *et al.* [79] which achieves 98% recognition rate performance on a small subset of the UND_E images.

In addition to the 2D techniques, five 3D techniques have also been produced, one using manual registration, two of which are fully automated and a further two which combine both face and ear. Of these techniques, the combined face and ear approaches achieve the best results with 98% (and above) recognition rates on datasets consisting of over 300 subjects.

The next section examines each of the existing ear recognition techniques in detail.

2.5.1 Ear Recognition Robustness

2.5.1.1 Position, rotation and scale

Only the fully automated approaches directly address positioning issues. The other techniques involve manual preprocessing to locate the ear prior to recognition. Of the automated techniques, Abate *et al.* [1], Marsico *et al.* [39] and Islam *et al.* [60] all use a Viola Jones based classifier to detect the ear within an image. In the case of Abate *et al.* the classifier is a significant source of error. However, this may be a result of the size of their training set, as both Islam *et al.* and Marsico *et al.* have achieved very accurate results with the same technique. An alternate approach, used by both Abdel-Mottaleb *et al.* [2] and Arbab-Zavar *et al.* [7] is to detect the ear using the curved outer ear shape (helix). Abdel-Mottaleb *et al.* use a template matching procedure whereas Arbab-Zavar *et al.* use a Hough transform designed to detect ellipses. Both approaches ultimately produce similar recognition rates but the datasets are different so it is not clear which approach is superior. Jeges *et al.* [61] take the edge matching principle one stage further by using edge orientation templates to locate the ear. In contrast to the helix approach, this technique uses all of the edges of the ear for detection purposes. This is likely to make the detector more discriminating against background clutter; however, it may also reduce the techniques robustness to occlusion. In addition, two techniques have been developed for automated 3D recognition. The first is from Yan *et al.* [112]. It begins by detecting the point of the nose and then traces back from there to locate the ear pit. The second approach, developed by Theoharis *et al.* [100], uses an exhaustive Iterated Closest Point (ICP) based technique to align a general head and ear model to 3D range scans. Based on the recognition results, all of these approaches are effective. However, their evaluations are performed in relatively controlled environments. In particular, no existing technique has demonstrated that ears can be accurately detected amongst large amounts of clutter, with significant variations in scale or with in-plane rotations.

A general problem is that many of the existing ear detection algorithms have inaccuracies in localisation. To address this, subsequent more precise registration is often performed. One approach is to normalise the position using a centroid of all of the edges in the detected region. This normalisation is applied by Abate *et al.* [1] in their fully automated recognition system. In the case of 3D, all existing algorithms refine their positioning

Table 2.9: Recognition results for manually registered 2D ear recognition algorithms

2D Manual Registration	No. of Ears	Constrained	Pose Variation	Occlusion
Hurley *et al.*[56]	63 (XM2VTS)	99%	-	-
Guo *et al.*[50]	79 (USTB)	92%	-	-
Xiaoyun *et al.*[110]	79 (USTB)	100%	-	-
Nanni *et al.*[77]	114 (UND_E)	80%	-	-
Nanni *et al.*[78]	114 (UND_E)	84%	-	-
Zhao *et al.*[117]	79 (USTB)[1]	-	95% (5 Degrees)	-
Naseem *et al.*[79]	32 (UND_E)[a]	-	98% (5 Degrees)	-
Choras *et al.*[31]	94[b]	-	100% (+/- 5 Degrees)	-
Badrinath *et al.*[10]	105	-	95% (20 Degrees)	-
Xie *et al.*[111]	79 (USTB)	99%	70% (-20 Degrees) 92% (20 Degrees)	-
Yuan *et al.*[113]	73 (USTB)	100%	78% (-20 Degrees) 85% (20 Degrees)	-
Wang *et al.*[106]	79 (USTB)	100%	92% (20 Degrees)	-
Cadavid *et al.*[26]	49	84%	-	-
Cadavid *et al.*[26]	402[c]	95%	63% (20 Degrees)	-
Yuan *et al.*[113]	24 (USTB)	93%	-	85% (20% occlusion)

[a] In both of these studies the gallery consists of images at all other poses ensuring that there is never more than 5 degrees difference between a probe and gallery image.

[b] The dataset consists of 10 images per person covering five pose variations and two lighting conditions.

[c] There are only 60 probe videos tested against 402 unique ears in this dataset.

using ICP prior to measuring shape similarity. Robust signatures can also be used as an alternative to precise localisation. Such an approach is used by the force field method of Hurley *et al.* [56] and the robust feature point approaches used by Arbab-Zavar *et al.* [8] and Marsico *et al.* [39]

Table 2.10: Recognition results for fully automated 2D ear recognition algorithms

2D Fully Automated	No. of Ears	Constrained	Pose Variation	Occlusion
Abate *et al.*[1]	70	62%	-	-
Abdel-Mottaleb *et al.*[2]	29	88%	-	-
Jeges *et al.*[61]	28	94%[a]	-	-
ArbabZavar *et al.*[8]	189 (XM2VTS)	87%	-	Top: 80% (20% occlusion)
Marsico *et al.*[39]	114 (UND_E)	62%	-	59% (8% occlusion)
Marsico *et al.*[39]	100	93%	-	93% (8% occlusion)

[a] This technique has an equal error verification rate of 5.6% but no recognition rate. The recognition rate has been estimated to enable a more meaningful comparison.

Table 2.11: Recognition results for 3D ear recognition algorithms

3D Manual Registration	No. of Ears	Constrained	Pose Variation	Occlusion
Chen *et al.*[30]	30	93%	-	-
3D Fully Automated				
Islam *et al.*[59]	100 (UND_J2)	90%	-	-
Yan *et al.*[112]	415 (UND_J2)	97%	-	-

Table 2.12: Recognition results for combined ear and face recognition algorithms

Combined Ear and Face	No. of Ears	Ear	Face	Fused
2D Manual Registration				
Victor *et al.*[103]	75	51%	71%	-
Chang *et al.*[29]	88	70%	70%	90%
3D Fully Automated				
Islam *et al.*[60]	315 (UND_J2 & FRGC)	87%	84%	98%
Theoharis *et al.*[100]	324 (UND_J2 & FRGC)	95%	97%	99%

2.5.1.2 Occlusion

The detection technique developed by Islam *et al.* [60] can correctly detect ears even when up to 70% of their surface has been occluded. However, even with this accuracy, no distance measure has been developed which can correctly rank ears when they have been occluded to that extent. Of the fully automated techniques, both Arbab-Zavar *et al.* [8] and Marsico *et al.*'s [39] techniques measure the effects of occlusion on recognition performance. Both approaches are based on feature points detected in the centre of the ear, which provides a degree of robustness to partial occlusions along the ear's outer edge.

In the case of the manual techniques, only Yuan *et al.*'s [113] approach measures the effects of occlusion explicitly. Their technique splits the ear image into three separate regions. If the signature of any of these region is sufficiently far from a Gaussian model of gallery ear signatures, the region is considered to be occluded and thus does not contribute to the ranking. No detailed analysis of occlusion robustness has been performed for any of the 3D approaches. However, Yan *et al.* [112] do note that the presence of earrings causes only a minimal loss of accuracy for their technique.

2.5.1.3 Pose

An early evaluation by Chang *et al.* [29] demonstrated that PCA based ranking is very sensitive to pose variations. This was confirmed by Yuan *et al.* [113] who performed a thorough analysis of their technique's performance over a range of poses. Their work indicated that views from behind are worse than those from the front. This may be due to ears being pointed slightly in the direction of the face, leading to views from behind having greater self occlusion. In addition, their study also demonstrated that there was a rapid fall off in performance beyond 20 degrees.

The work of Choras *et al.* [31], Naseem *et al.* [79] and Badrinath *et al.* [10] all demonstrate that recording ears at multiple poses significantly improves recognition rates. Of these approaches, Badrinath *et al.*'s technique achieves the most significant pose robustness. Their technique enrolls ears at -40, 0 and +40 degrees. The SIFT points of each pose are used for recognition. The identity of a probe image is determined by calculating the number of pairs of SIFT points whose signatures match within a given threshold.

Of the other manually registered 2D techniques, three approaches have been developed which have pose robustness. The first, created by Yuan *et al.* [113] fits an active shape model to ear images. This model is used to normalise position and orientation. The resulting images are then processed using filters calculated using linear discriminant analysis. The second approach, developed by Xie *et al.* [111] uses local linear embedding to create a non-linear space in which variations due to pose are minimised and variations due to identity are maximised. Wang *et al.* [106] produced the third and most robust of these techniques. They used a combination of Haar wavelets and local binary patterns.

It should be noted, however, that all three techniques rely on manual registration, which, if automated, may be sensitive to pose changes. Similarly, all of the fully automated systems use detection techniques that depend on templates, classifiers or facial features whose shape is significantly altered by pose changes. Other than by enrolling subjects from multiple angles, no existing technique has demonstrated automatic ear detection under large pose variations. A potential solution is to use 3D shape. Cadavid *et al.* [26] offer two approaches which estimate the 3D shape from 2D video sequences. In terms of recognition, only the evaluation of the shape from shading technique explicitly measures robustness to pose changes. This evaluation demonstrates that the technique's performance drops from 95% to 63% at 20 degrees. This reduction indicates that robust distance measures may provide greater accuracy than the shape from shading technique. Other 3D approaches which use accurate laser scans of the ear may provide greater robustness, but none of the existing 3D techniques include evaluations with significant pose variation.

2.5.1.4 Lighting

Within existing published work on ear recognition, the effect of lighting direction is rarely analysed. Most authors uses datasets recorded under similar lighting conditions and without explicit lighting calibration. However, for 2D the techniques used typically include an intensity normalisation process to remove the effects of overall brightness and contrast changes. One of the most sophisticated of these normalisation techniques is that used by Badrinath *et al.* [10] Their normalisation algorithm alters pixel brightness to match the shape of image histograms prior to comparison. This compensates for complex non-linear, lighting effects. However, as with all global normalisation techniques, the algorithm can increase errors when the ear is occluded. This is because the darkness (or brightness) of the occluded regions will cause the remaining pixels within the image to be altered to compensate. The locally normalised SIFT signatures used by Arbab-Zavar *et al.* [8] may prove to be more robust in these cases.

2.5.1.5 Within class variation

Each technique uses a different method for ranking potential gallery subjects to determine an ear image's most likely identity. Hurley *et al.* [56] produced the first published ear biometric system that included an evaluation of recognition performance. Their system used an image processing technique based on a force field analogy. This technique focuses the information within an image into localised features which are then compared using template matching.

Other 2D recognition systems use similar area filtering techniques in order to calculate robust signatures. Xiaoyun *et al.* [110] tested a number of different approaches, including both Fourier and Gabor transforms. They found that low order moments produced the best results. Other work, by Wang *et al.* [106], used a combination of Haar wavelets and local binary patterns to obtain highly accurate results on the challenging USTB pose varying dataset.

Another approach is to use the gallery images to learn a more robust comparison process. For example, Yuan *et al.* [113] used linear discriminant analysis to maximise the differences due to identity; Xie *et al.* [111] constructed a local linear embedding to minimise the effects of pose variation and Nanni *et al.* [77] identified subwindows of the ear image that had the most discriminative power. Their subsequent improvement of the technique used the response from different colour representations to provide a further 4% improvement in recognition rate. This is notable as it is the only technique to use colour information [78].

Alternative 2D comparison approaches use the location and/or local appearance of feature points. Marsico *et al.* [39] calculated these points using a fractal encoding technique. Their approach also included a means to adjust the speed of matching to achieve greater or lesser robustness. Similarly, both Badrinath *et al.* [10] and Arbab-Zavar *et al.* [8] use SIFT points to provide a robust basis for comparison. The technique of Arbab-Zavar *et al.* is also notable as it used the gallery to create a model of feature point types which could then be used to create a signature for rapid ear comparison.

Jeges *et al.* [61] also used a model based technique. They fitted an active shape model to detected ear images. The resulting fitted shape was then used to calculate a signature for comparison. Choras *et al.* [31] also used the pattern of edges to determine an ear signature. However, this signature was obtained without a model fitting process, making the technique more sensitive to lighting changes and occlusion.

In the case of 3D recognition, both Yan *et al.* [112] and Islam *et al.* [59] calculated rankings of 3D scans using the ICP distance error. However, Islam *et al.* also added a measure based on the similarity of local surface patches. This improved face recognition performance significantly but actually reduced ear recognition performance by 5%.

The approach of Theoharis *et al.* [100] is the only model based 3D technique. The technique fits a rigid average model of the ear and face. Once the model has been fitted it is used to extract a normalised shape representation of both the face and ear. These shapes are processed using the Walsh transform to create a wavelet based signature. These signatures can then be used to provide very fast indexing into a large gallery database.

Finally, Yan *et al.*'s [112] 3D recognition approach is the only study which analyses the effect of dataset size on accuracy. Their work demonstrates a drop of 2% between 100 and 150 subjects, which then holds steady up to their maximum dataset size of 415. This relatively minor change indicates that larger datasets are needed to obtain a more accurate measure of how performance decreases with dataset size.

2.6 Conclusions

The objective of the work described in this thesis is to relax the constraints on ear biometric systems, through the introduction of more robust recognition techniques. This chapter identified five key constraints that are imposed on such systems and examined the range of techniques that have been developed to address them. The existing research in ear recognition was also reviewed and a number of areas for improvement identified. In particular, the analysis indicated that no existing system has demonstrated robust detection of ears in cluttered environments. The next chapter describes a new SIFT based ear recognition system that has been developed to address this issue. The system combines interest points, model fitting and robust distance measures to create a fully automated recognition approach that is robust to many unconstrained factors.

The main limitations of the SIFT based approach are that pose and/or lighting variations can significantly reduce its recognition accuracy. This is a common weakness of many ear and face recognition systems. As indicated in this chapter, the Morphable Model technique is one of the few approaches that can remain accurate despite these variations. Chapter 4 describes a technique that has been developed to make Morphable Model based ear recognition possible. This technique provides an important step towards fully robust ear recognition and is the first example of applying Morphable Models to ear recognition.

Chapter 3

SIFT Based Registration and Recognition

The main technical contribution of this chapter is to propose an improved ear registration and recognition technique based on the object detection algorithm of Brown *et al.* [20]. Their technique calculates a homography transform between a gallery object image and a probe image using SIFT (Scale-Invariant Feature Transform) point matches [68]. The probe is considered to be a potential match to a gallery image if a homography can be created using four corresponding SIFT points. The homography defines the registration between the gallery and the probe. This creates a very accurate registration. Brown *et al.* demonstrated good results for various objects but their approach is insufficiently discriminating to rank ear images. The work described in this chapter extends their technique with an image distance algorithm to obtain a precise ranking. To calculate the image distance accurately, gallery ears are segmented using a mask. These masks are semi-automatically created as a preprocessing step on the gallery.

Collectively, these developments produce an automated, accurate, ear recognition technique that is robust to location, rotation, scale, pose, lighting, background clutter and occlusion. The technique is an important step towards creating a fully automated, unconstrained ear recognition system. This chapter describes the technique that has been developed. Section 3.1 covers its stages, including the semi-automatic creation of gallery masks. The registration calculation and its theoretical justification are also presented, along with an overview of the distance measure which provides accurate ranking. The technique is evaluated in Section 3.2. This includes both a traditional, controlled environment recognition test, as well as more challenging datasets that evaluate the technique's robustness to occlusion, background clutter, resolution and pose variation.

3.1 Technique

The approach described here uses a combination of techniques to achieve robustness. The initial registration process uses SIFT feature point matching. These features are robust to many unconstrained factors [75]. By using feature points the registration is inherently robust to occlusion as any four point matches are sufficient to register the ear. Also, by modelling the ear as a planar surface and registering it using a homography transform, the ear can be recognised across small variations in pose and camera properties. To rank registered ears, a distance measure is used that performs both normalisation and outlier detection. This makes the ranking step robust to varying lighting levels and occlusion. In addition, the combination of feature matches, homography registration and image distance are sufficiently discriminating to detect and recognise ears within cluttered environments.

Before any probe images can be tested, the gallery images are processed to segment the ears. Each gallery image is then analysed to determine its SIFT feature points. Once this is complete a probe image can be recognised. The first step is to identify feature points in the probe. For each of these points the gallery is searched to find correspondences. If four points are matched between the probe and the gallery, they are used to calculate a perspective transformation that registers the probe. Once the two images are aligned, the distance between them is calculated. The nearest gallery image identifies the person. Each stage of this process is described in the subsections that follow.

3.1.1 Building the Gallery Database

Images of the same ear taken at different times can vary significantly due to changes in hair length and colour. This variation can create many false point matches and significantly reduces the accuracy of image distance measurements. For this reason, gallery ears are masked to segment the ear from the surrounding skin and hair, as illustrated in Figure 3.1. Ideally, these masks would be created automatically thus enabling the efficient enrolment of subjects from large existing data sources such as criminal mugshot databases. Unfortunately, without a model of all possible appearances, new ears cannot be detected automatically. However, the number of manually created

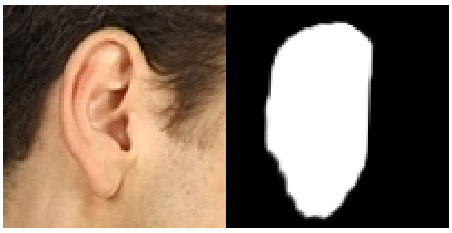


FIGURE 3.1: A gallery ear image and its associated mask

masks can be greatly reduced by using a bootstrapping process. This approach exploits the fact that while each person's ear is unique, specific regions may be very similar. One explanation for this similarity is that ear shapes are created through a set of independent local deformations. Some evidence for this hypothesis has been provided by Arbab-Zavar's model-based ear recognition algorithm [8], which describes six growth factors that define an ear's shape. Similar ear regions will have correspondingly similar SIFT points. If four matches between these points can be detected, the ears can be registered with one another (Figure 3.2). These registrations can then be used to transfer the masks to unlabelled gallery images (Figure 3.3).

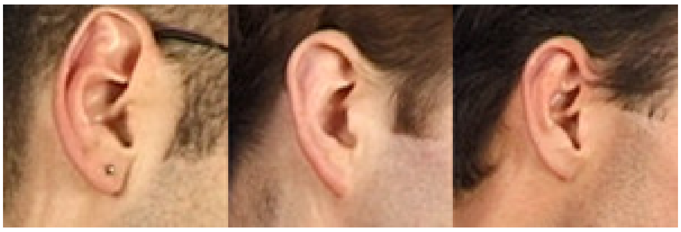


FIGURE 3.2: A set of gallery ears that partially match the seed ear

Each newly masked ear can then be matched against the rest of the gallery. These ears may introduce additional local regions that are similar, enabling more masks to be transferred. This process can then be repeated until no further matches can be made. At this point one of the unmasked ears must be selected and manually processed. This seed can then be used to bootstrap the remaining gallery.

In this way only a subset of the gallery needs to be manually masked. In addition, as the gallery size increases it becomes more likely that ears will form matches, thus reducing the percentage of manual masks that are required.

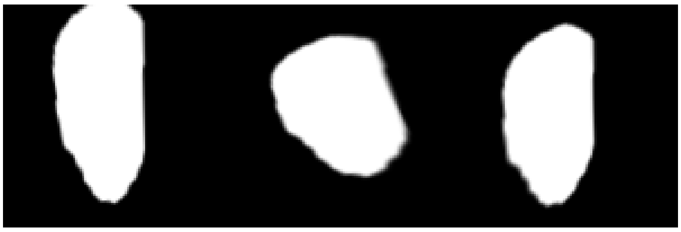


FIGURE 3.3: The masks automatically created from the homography registration of seed to gallery

3.1.2 Feature Detection

The technique uses SIFT [68] for the detection of features. It is robust to scale, in-plane rotation and to lighting, and with some robustness to pose. The main parameters of the SIFT algorithm define the resolution of the Gaussian image pyramid used to detect the feature point locations. Where possible, default values have been used, with the number of octaves determined by the image size; the lowest octave size was 8x8. Each octave had three intermediate Gaussian blurred versions. The features were also normalised to improve robustness to lighting variations.

The Approximate Nearest Neighbours algorithm [9] was used to make the matching of features against a large gallery more efficient. The algorithm enables 128 dimensional point matches in O(log(n)) where n is the number of feature points in the gallery. SIFT points were considered a potential match if the squared Euclidean distance of their signatures was less than 0.45, with a maximum of 1024 matches returned (closest first).

3.1.3 Registration Calculation

By making the simplification that an ear is a planar structure, ears can be registered accurately. If ears are enrolled with the ear plane facing the camera, the image produced can be used to approximate the ear appearance at varying poses. By finding four point matches between an enrolled gallery image and a probe, a homography can be calculated [51]. This homography can be used to transform the gallery image to match the position, rotation, scale and pose of the probe ear. The transformed image can then be used to compare the two images accurately. The homography is calculated as follows:

Let x be a homogeneous point in the probe image and x' be a homogeneous point in the gallery image, then the homography H is defined by

$$x' = Hx$$

where

$$x = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} x' = \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} H = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix}$$

This can be expressed as

$$x' \times Hx = 0$$

By considering H as a matrix of row vectors h^{jT}

$$H = \begin{bmatrix} h^{1T} \\ h^{2T} \\ h^{3T} \end{bmatrix}$$

the cross product can be expanded to give

$$x' \times Hx = \begin{pmatrix} y'h^{3T}x - h^{2T}x \\ h^{1T}x - x'h^{3T}x \\ x'h^{2T}x - y'h^{1T}x \end{pmatrix}$$

Since $h^{jT}x = x^T h^j$ this can be rewritten as:

$$\begin{bmatrix} 0^T & -x^T & y'x^T \\ x^T & 0^T & -x'x^T \\ -y'x^T & x'x^T & 0^T \end{bmatrix} \begin{pmatrix} h^1 \\ h^2 \\ h^3 \end{pmatrix} = 0$$

This is a linear equation in h of the form $Ah = 0$, where A is a 3x9 matrix and h is a 9 vector. A has only two linearly independent equations as the third row is the sum of $-x'$ times the first row and $-y'$ times the second. By omitting this equation, the remaining set becomes:

$$\begin{bmatrix} 0^T & -x^T & y'x^T \\ x^T & 0^T & -x'x^T \end{bmatrix} \begin{pmatrix} h^1 \\ h^2 \\ h^3 \end{pmatrix} = 0$$

This shows that each point correspondence adds two independent equations in the entries of H. By combining these equations into a single matrix, four point correspondences create a matrix with a size 8x9 and rank 8. This matrix has a one dimensional null-space which provides a solution to H up to a non-zero scale. As these points are homogeneous, if the transformed points are normalised by dividing through by their third component, this scale factor will be removed.

Homographies define changes that are caused by camera perspective and relative position, orientation and scale if matched points. This is in contrast to the more commonly used affine transform that can represent skewed images but does not accurately recreate perspective changes. This enables the technique to be used with the ear is relatively large within the image or when significant perspective distortion is present. For example when a wide angle lens is used. However, the extra flexibility of the homography means that invalid matches can produce transformations with implausible perspective to address this there is a threshold on the magnitude of the sum of the squares of the two homography parameters that model perspective variations.

The SIFT-matching distance is quite generous to enable large variations in pose and lighting. However, this will result in a significant number of false positives in the point correspondences. To reduce such errors, an affine consistency constraint was applied. This constraint groups the SIFT matches into sets of points that have an approximately equal in-plane affine transform. This constraint is reasonable under small pose variations where the homography will be close to affine.

As part of the SIFT detection process, there is a search for interest points across locations and scales. When an interest point is detected, the region surrounding it is used to calculate a canonical orientation. By comparing these values between the probe and the gallery, each point can be used to calculate an approximate affine transform between the two images. By grouping points into bins based on their affine transform, many false positives can be excluded.

The potential space of affine transforms was subdivided into four dimensions: two for position, one for logarithm of the scale, and one for rotation. Each of these dimensions was then partitioned into bins: eight for scale and rotation and one for every 128 pixels in width and height. A low resolution of bins was used to ensure that the matching is robust to pose variation. Each point match is placed in the appropriate bin and its closest neighbor (16 bin entries per point). If any bin contains four or more point matches, its points are passed to the next stage. This process greatly reduces false positives, but some invalid point matches remain. To address this problem, a RANdom SAmple Consensus (RANSAC) algorithm was used. Random sets of four points were selected from the list of point correspondences and a homography calculated. The homography that matches the most points within some threshold, i.e., in this case, 1% of the ear mask size, was selected as the best match. Gallery images that have four affine matching feature points are then passed to the distance measure. The combination of SIFT-matching and affine constraints greatly reduces the set of potential gallery matches. This process is sufficiently accurate to prevent false matches, both with the majority of incorrect ears and with background clutter.

3.1.4 Distance Measure

Once the gallery images have a good registration they can be matched against the probe. The distance is calculated as the robust sum of the squared pixel error after normalisation. The distance measure is made robust to occlusion by thresholding the error. Pixels that differ by more than half the maximum brightness variation are considered to be occluded and thus do not contribute to the distance value.

Once occluded regions have been excluded, normalisation is applied which adjusts the scale and offset of the remaining intensity values so they have a zero mean and unit standard deviation:

$$\begin{aligned}&G(I,x,y) = \\ &\quad (r(I(x,y)) + g(I(x,y)) + b(I(x,y)))/3 \\ &\quad \forall x \in [1..w] \quad \forall y \in [1..h]\end{aligned}$$

$$mean(G,I) = \left(\sum_{y=1}^{h}\sum_{x=1}^{w} G(I,x,y)\right)/(wh)$$

$$\begin{aligned}
&scale(G, I) = \\
&\quad \sqrt{\textstyle\sum_{y=1}^{h} \sum_{x=1}^{w} (G(I, x, y))^2 - mean(G, I)^2}/(wh)
\end{aligned}$$

$$\begin{aligned}
&N(I, x, y) = \\
&\quad (G(I, x, y) - mean(G, I)/scale(G, I)) \\
&\quad \forall x \in [1..w] \quad \forall y \in [1..h]
\end{aligned}$$

$$\begin{aligned}
&notoutlier(I_1, I_2, x, y) = \\
&\quad \left\{ \begin{array}{cc} 0 & \|G(I_1, x, y) - G(I_2, x, y)\| >= 0.5 \\ 1 & \|G(I_1, x, y) - G(I_2, x, y)\| < 0.5 \end{array} \right\}
\end{aligned}$$

$$\begin{aligned}
&ndistance(I_1, I_2) = \\
&\quad \textstyle\sum_{y=1}^{h} \sum_{x=1}^{w} notoutlier(x, y) * (N(I_1, x, y) - N(I_2, x, y))^2
\end{aligned}$$

where G is a function that returns the gray scale values of an image, N is a function that returns the normalised values of an image, w and h are the width and height of those images, and $r()$ $g()$ and $b()$ are functions that return the magnitude of the red, green and blue components, respectively. Each component returns values in the range 0 to 1. This normalisation removes variations due to changing lighting magnitude and camera sensitivities.

3.2 Evaluation

Eight datasets were used for evaluation of this technique. The first provided a straight test of recognition accuracy on a relatively constrained dataset. For this, a subset of the XM2VTS [73] face-profile database was chosen. It consists of 63 subjects with relatively unoccluded ears. This is the same dataset used by Hurley *et al.* [56] and Arbab-Zavar *et al.* [8].

The second dataset was created by recording both ears of 20 subjects from a range of angles to test the technique's robustness to pose variation. The remaining datasets were synthesised from these XM2VTS images to test the effects of occlusion, background clutter, resolution, noise, contrast and brightness.

3.2.1 Recognition

For the constrained gallery set, two comparison implementations were created. The first used ear images which had been manually registered using the technique described by Yan *et al* [112]. This involved hand labelling the Triangular Fossa and Incisure Intertragica of each ear. These landmarks were then used to standardise the scale and

Table 3.1: Recognition rate for different registration techniques

Registration	Technique	% Rank 1
Manual	PCA	96%
Automatic using outer ellipse	PCA	75%
Automatic using homography	Image distance	96%

Table 3.2: Number of features at each stage XM2VTS dataset

Feature	Count
Number of gallery images	251
Number of gallery SIFT points	14,234
Average number of SIFT points on XM2VTS image (720x576)	4,659
Average number of SIFT matches	20,834
Average number of images with SIFT matches	250
Average number of images with affine constrained homographies	4

rotation of all gallery and pose images. The resulting normalised images were then segmented using a rectangular mask applied to the centre of the image. This mask excluded the variation due to hair while retaining the inner ear features.

The second technique applied the algorithm described by Arbab-Zavar *et al.* [7] to register the ear automatically, using the outer ear ellipse. In both cases, the intensity values had their mean and standard deviation normalised. These registered images were ranked using the Eigen Ear technique [29] giving the results shown in Table 3.1. Each technique used the 'leave one out' strategy, with each image removed from the gallery and tested against the rest of the dataset in turn.

The bootstrapping process, using the first ear, matches over 75% of the gallery. In total, 22 masks were created manually to cover 252 gallery images.

Generally, the masks are not a precise fit for the ears but the accuracy is sufficient to obtain enough feature points for the registration and distance measures.

It can be seen from Table 3.2 that the homography registration is the primary point at which the ears are recognised, going from almost the entire gallery down to four candidate images. The registration calculation is also the cause of 4% of the probe images remaining unclassified. These ears failed to produce a valid homography because of insufficient SIFT point matches.

3.2.2 Robustness

3.2.2.1 Gallery

The clutter dataset was created by randomly placing XM2VTS masked ear images on a set of complex background images. These images more closely represent the type of unconstrained environment present with covert biometrics. The occlusion dataset was built by adding varying sized solid black rectangles over the top or side of the original gallery images. These rectangles correspond to the areas of the ear that are most frequently occluded by hair. To determine the percentage of occlusion that each rectangle represents, the occlusion of each mask was calculated and then averaged across the gallery. The resolution dataset was created by linearly down-sampling and then bicubicly up-sampling the probe images. The contrast dataset was constructed by subtracting the mean pixel colour, scaling the result and then adding back the mean. Similarly, the brightness dataset added an offset to each pixel's channel.

Finally, to generate the pose dataset, 20 subjects were recorded turning in front of a camera. Both sides of the head were recorded to obtain 40 unique ears. For the purpose of evaluation, each ear was treated as an independent subject. Each person had a camera calibration grid attached to a hat they wore when photographed. This grid enabled the camera intrinsics and pose angles to be calculated accurately. The calculations were performed using the standard camera calibration algorithms provided with the OpenCV[1] libraries. Figure 3.4 shows examples from some of these datasets.

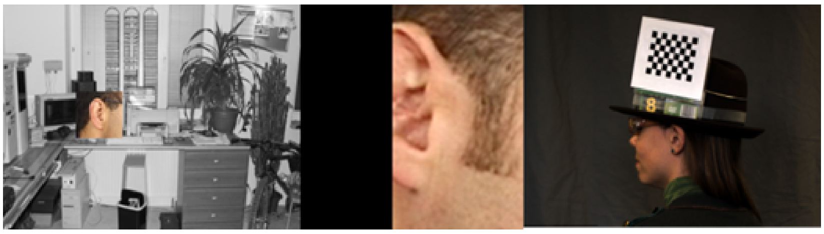


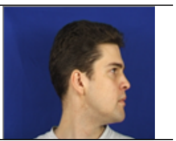
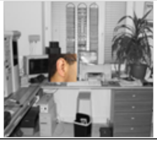
FIGURE 3.4: Examples of more challenging probe images. From left to right background clutter, occlusion and pose variation

3.2.2.2 Results

Tables 3.3, 3.4, 3.5 and 3.6 summarise the results of the recognition tests. Background clutter was found to have little effect on the recognition rate as was up to 30% occlusion from above and 18% occlusion from the side. However, any greater occlusion significantly reduced the technique's accuracy. Once again, this was due to finding insufficient SIFT matches to calculate the homography. With resolution changes, images remained

[1] The OpenCV library is available at http://opencv.willowgarage.com/wiki/

Table 3.3: Average recognition rates for controlled and cluttered

Technique	% Rank 1 Recognition	Examples
Base recognition rate	96%	
Background clutter	93%	

recognisable at 50% of their original size (i.e. when reduced from 40x70 to 20x35, depending on mask size). The contrast results show that the approach maintains 90% recognition accuracy with 80% of the contrast. The approach is sensitive to brightness, however, with a 20% increase almost halving the recognition rate. In both cases, recognition failures are primarily due to failing to find SIFT matches. Figure 3.5 shows the average recognition rate for 40 ears with varying pose. The technique maintains a 100% recognition rate up to 13 degrees. However, this performance is dependent on enrolled gallery ears being recorded with the ear plane facing the camera. If gallery ears are protruding, pose invariance is reduced. For example, if subjects are recorded with the ear plane tilted 30 degrees from the camera the pose invariance is reduced to 10 degrees.

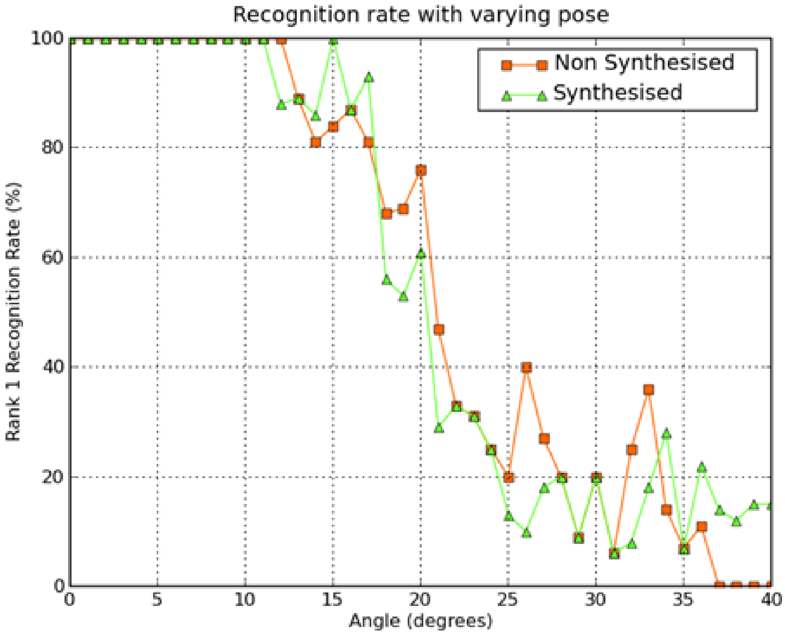
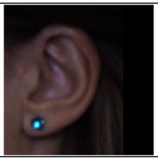
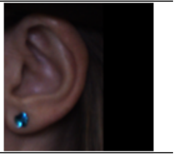
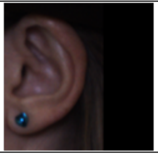
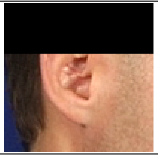
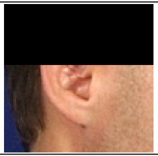
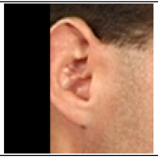
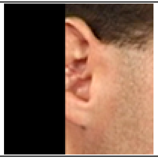


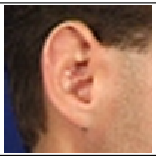
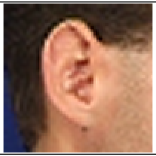
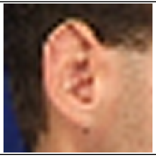
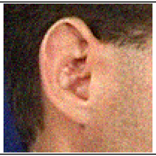
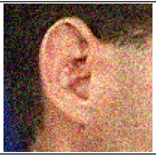
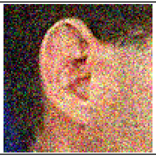
Figure 3.5: Recognition rate with varying pose, with and without synthesised ear images

Table 3.4: Average recognition rates for pose and occlusion

Technique	**% Rank 1 Recognition**	**Examples**
0 degrees pose variation	100%	
13 degrees pose variation	100%	
22 degrees pose variation	33%	
30% occlusion from above	92%	
40% occlusion from above	74%	
18% occlusion from the side	92%	
37% occlusion from the side	66%	

As an experiment to improve the technique's robustness to pose variation, additional gallery images were synthesised at novel poses. These were created by treating the ear image as a plane photographed at an estimated distance with an approximated field of view. The plane was then rotated in the image plane x and y axes and re-rendered to simulate different poses. This increased the number of SIFT matches but also the number of false positives. As the ears are not completely planar, the image distance increases with angle, resulting in incorrect ears having a shorter image distance and so no significant increase in robustness was observed.

TABLE 3.5: Average recognition rates for resolution and noise

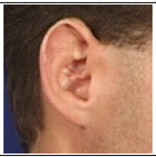
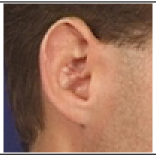
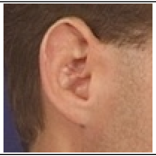
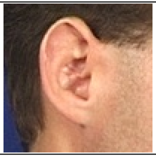
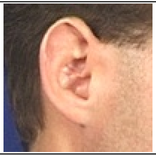
Technique	**% Rank 1 Recognition**	**Examples**
50% of original resolution	93%	
40% of original resolution	87%	
30% of original resolution	66%	
Gaussian noise with sd. at 10% max. chan. mag.	93%	
Gaussian noise with sd. at 20% max. chan. mag.	75%	
Gaussian noise with sd. at 30% max. chan. mag.	54%	

3.3 Conclusions

The approach described is relatively successful in identifying ears under different conditions but as is evident from Table 3.4 it would be desirable to increase the degree of pose variation over which recognition can be achieved. One strategy would be to record subjects at multiple angles, either at gallery creation or as probes. Alternatively, if this were not possible, the synthesis algorithm could be improved through the use of a Morphable Model [17].

Another area for improvement is the computation time of the algorithm. Despite the use of the ANN library, the processing of each 720x576 probe image takes over four minutes on a 2.4Ghz PC. The majority of this time is spent in calculating and measuring the image distance and the RANSAC homographies. Each image requires over 10,000 of these calculations on average. In total this takes over three and a half minutes. The next most expensive stage is the SIFT matching process which takes around a minute. The remaining calculations, such as the detection of SIFT points in the probe

TABLE 3.6: Average recognition rates contrast and brightness

Technique	**% Rank 1 Recognition**	**Examples**
10% reduction in contrast	94%	
20% reduction in contrast	90%	
30% reduction in contrast	81%	
10% max chan. mag. increase in brightness	92%	
20% max chan. mag. increase in brightness	58%	
30% max chan. mag. increase in brightness	31%	

take seconds and have a relatively small impact on performance. Efficiency could be improved through using a pyramid hash matching technique. The hashing algorithm enables comparisons between sets of high dimensional features and can be scaled to very large datasets. Another area for improvement is the use of image pixel differences as a distance measure. This could be made more robust with a less sensitive measure, such as Hausdorff edge distances [43].

This chapter has described the first of two novel approaches to ear recognition. It is the first robust and accurate ear registration algorithm suitable for recognition within background clutter. The next chapter introduces the second recognition approach which has the potential to be more robust and fully automated.

Chapter 4

Ear Model Construction

Chapter 3 presented a first step towards the detection and recognition of individuals by their ears in unconstrained settings. Three main areas were identified as requiring further attention namely:

1. Lighting variation
2. Large pose variation
3. Efficient recognition with large datasets

These are challenging problems and it can be seen from the overview of techniques in Chapter 2 that few approaches are available for addressing them. In particular, there is no simple way of extending the technique described in Chapter 3 to improve its robustness to these variations. For this reason the second recognition technique takes a different approach based on the use of a Morphable Model [17].

A Morphable Model is a parameterised 3D representation from which all possible variations of an object's shape and appearance can be constructed. Such a model can be used to calculate the shape of an ear from measured data, accurately and efficiently. By constraining any estimate of ear shape to the space of the model, a more accurate measurement of ear structure can be obtained. In addition, when the model is used as part of a fitting process, its low dimensions greatly improve performance as fewer parameters are needed to match the ear shape. Also, if the model is accurate, computer graphics techniques can be used to synthesise all possible ear appearances. When applied to 2D images, model fitting can be achieved by adjusting the model parameters until a synthesised image matches the probe ear image to be recognised. Once achieved, the shape parameters provide a unique signature for the ear.

Existing Morphable Models of the head have focused on the face and implicitly or explicitly avoided accurate ear reconstruction [13] [3]. This neglect of the ear is partly due to the challenge of modelling its more detailed and self occluding structure. In particular, range scan data of the ear is generally of lower quality and less complete than that available for the face [112]. In addition, ears have not been a priority in existing Morphable Model work as they are not generally used by humans for recognition.

The main contribution of this chapter is a novel technique for the construction of a Morphable Model of the face profile and ear using noisy, partial and occluded data. The model is constructed by registering a generic head mesh with 160 range scans of head profiles. Regions of the scans that are occluded or have significant noise artifacts are identified automatically using a classifier. The remaining valid regions are then used to register the mesh using a robust non-linear optimisation algorithm. Once registered, the scan's orientations are normalised and then used to construct a linear model of all head shapes. The next section summarises relevant existing work on Morphable Model construction and the representation of ears in those models. This is followed by a discussion of how the fitting problem can be formalised. The technique is then described in detail. The chapter concludes with an evaluation of the fitting technique and the resulting model using three new metrics.

4.1 Related Work

In 1999, Blanz and Vetter created the first 3D Morphable Model [17]. It was constructed from over 200 cylindrical range and colour scans of male and female heads. These heads were registered with each other using an optical flow algorithm. The model was then constructed using the mean of these values and their first 90 eigenvectors calculated using PCA (principal components analysis). This produced a highly realistic model of face appearances. The technique was used to create photorealistic 3D meshes from single photographs.

Further work has concentrated on improving the registration process. For example, in 2006, Basso *et al.* extended the optic flow approach to handle expression varying datasets [13]. In the same year, Vlasic *et al.* produced a multilinear Morphable Model that enabled independent adjustment of identity and expressions [105]. This used an optimisation-based fitting technique. Then in 2008 Amberg *et al.* [4] replaced the general optimisation framework with a non-rigid iterated closest point algorithm. With their approach they were able to construct models using partial range scans. Another contribution from Patel and Smith was to demonstrate the importance of Procrustes normalisation of scans prior to calculating the subspace. Their work also included an alternative fitting algorithm that used thin-plate splines calculated from manually labelled feature point correspondences [86].

Other researchers have applied the Morphable Model approach to different objects. For example, in 2005, Allen *et al.* constructed Morphable Models of human body shapes [3].

In similar work, Anguelov *et al.* [5] developed a novel automated registration algorithm for bodies with significant pose variation. Their technique enabled the construction of a model to estimate a subject's body shape under different poses.

These existing approaches have concentrated on face or body shape and have, in general, avoided constructing accurate ear models. For example, in the work of Allen *et al.* [3] the ear region is manually excluded from the fitting constraints. This produces heads on which all ears have the same shape. Similarly, the fitting results of Amberg *et al.* [4] and Basso *et al.* [13] show little influence of the ear on the fitted meshes, as indicated in Figure 4.1 and Figure 4.2.

Of the existing research, some of the most accurate ear models have been produced by the original Blanz and Vetter model (Figure 4.3). Their work approximated the head by sampling points uniformly over its surface and then connecting them to form a complete manifold. This technique creates ear models that lack folds and self occlusion but as the surface texture includes some shadow information the visual results appear reasonably accurate.

In addition, a recent paper by Paysan *et al.* [87] has demonstrated high quality head models (Figure 4.4). These were constructed using a more precise scanning process and registered using the algorithm of Amberg *et al.* [4]. This is in contrast to the work described here, which focuses on achieving accurate results using a less constrained dataset, through the use of an automated occlusion and noise classifier. All of the above techniques involve some degree of manual labelling to achieve accurate results. However, early work by Brand [19] demonstrated the potential for automatically constructing Morphable Models directly from video sequences. This approach used optical flow, structure from motion, and a novel matrix factorisation technique. The resulting models are visually convincing but are only demonstrated on single individuals and lack the detail of models produced by registered range scans.

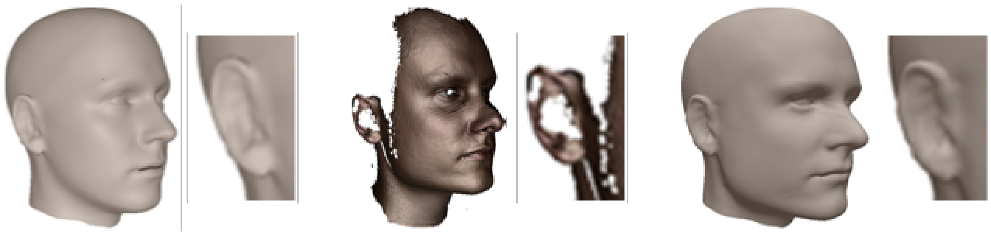


Figure 4.1: This image shows the base mesh, cleaned scan and fitted model for the technique used by Amberg *et al.* [4] It shows that the ear is not affected by the range scan and retains the shape of the base mesh.

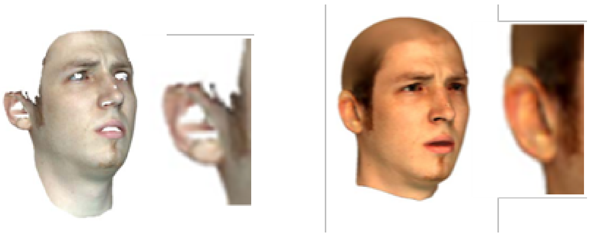


FIGURE 4.2: Results from the work by Basso *et al.* [13] showing the base mesh and range scans and the resulting registered models. Note that the registered models have less protruding ears than the original scans and lack internal detail.

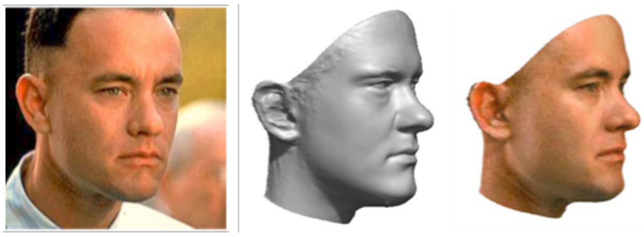


FIGURE 4.3: Image of a fitted Blanz and Vetter [17] ear, the inner detail of which is missing but is compensated by the information in the texture.

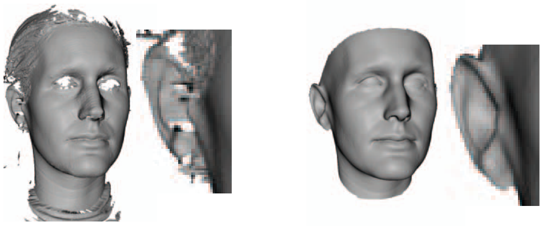


FIGURE 4.4: Image of a fitted Paysan *et al.* [87] ear. Much of the ear is accurately modelled but the top left outer curve and middle ear hole covering are not fully reconstructed.

4.2 Defining the Problem

One difficulty in constructing a Morphable Model is that there are no recognised definitions for an optimal model, in particular, there is no agreement on the metrics to be used to optimise the model parameters. Initial Morphable Model construction work used qualitative rather than quantitative evaluations of accuracy. The precision of these techniques was shown visually through a number of rendered example registrations [17]. One exception to this, however, is the work of Amberg *et al.* [4], which includes two measures of the quality of their created models. The first is based on the accuracy with which excluded training samples can be reconstructed using a model built from the remaining data. The second measure evaluates the average angle between the normals of corresponding points on separate registered models. This quality value is determined by calculating the difference between every possible pair of models in the training set and averaging the results. The rationale behind this measure is that correctly registered model parts should have correspondingly similar surface normals. This error measure can be formalised as:

$$error(i,j) = \frac{1}{l}\sum_{k=0}^{l}(cos^{-1}(vn(i,k).vn(j,k)))^2$$

where i and j are the two aligned models, $vn(i,k)$ is a function that returns the normal of the vertex k of the model i, and l is the number of vertices in the model.

Another evaluation was provided by Patel and Smith [86] in which they calculated how accurately the model could infer the shape of the face from a sparse set of hand labelled feature points.

Building on this work, two desirable Morphable Model properties are explored here: *generality* and *total variance*. Model generality refers to the degree to which the model can generalise to samples beyond its training set and model total variance refers to the amount the model can change shape. Unfortunately, optimising these properties directly is ill-posed and computationally expensive. This is due to the extremely large number of parameters involved and the relatively small quantity of training data available. In the case of Morphable Models of a class of objects, such as all heads, existing approaches use heuristic techniques and assumptions to make the problem tractable. The primary underlying assumption of existing Morphable Model work is that by creating an accurate registration between a large set of head models, a linear subspace can be created that will accurately represent all head shapes. A difficulty here, however, is that the registration is not unique, as there is no agreement on which parts of a modelled object should be considered the same between different subjects. Existing approaches have tackled this issue in different ways.

With the exception of the automated technique of Brand [19], all current approaches initialise their registrations using a set of manually labelled features, followed by the application of an algorithm that calculates a dense correspondence. In the work of Patel and Smith [86] a thin plate spline is used to interpolate a cylindicral mapping calculated from the manually labelled points. In contrast, Vlasic *et al.* [105] use an iterative approach based on a similar principle to the iterative closest point algorithm, progressively adjusting the fitting using closest points as an estimate of the correspondence between mesh and range data. This was refined by Amberg *et al.* [4] to calculate the optimal least deformation solution to the correspondence at each iteration. In addtion, Blanz and Vetter [17] and Anguelov *et al.* [5] include local appearance similarity measures in their fitting techniques. Blanz and Vetter's work formulated the problem as an optic flow calculation between cylindrical range scans. In contrast, Anguelov *et al.* calculated local mesh signatures using Spin images [62]. By using this feature their approach could be applied to less constrained range meshes. In addition, the technique included a correlated correspondence algorithm which calculates the global minimum for both the local similarity and the local smoothness of each part of a mesh using loopy belief propagation. Using this technique they were able to construct accurate full body Morphable Models from range scans with significant variations in body pose.

Overall, the accuracy and consistency of these techniques is dependent on the quality of the initial labelling and the smoothness of their cost functions. Their relative performance can be evaluated by analysing the resulting Morphable Models or through direct metrics that evaluate each fitting independently. In the work described here, the quality of each registration has been evaluated using the consistency of the resulting mesh. This consistency is calculated by comparing the similarity of meshes produced when they are fitted to multiple range images of the same person. If multiple scans of the same person result in meshes that are more similar than those of any of the other subjects then it is likely that the fitting process and the resulting model can be used for recognition purposes.

One approach to improve the minimalism of the constructed model is to normalise the pose of training images prior to calculating the model subspace. The benefit of this approach is most obvious when a number of identical meshes at different poses are used to construct a model. In such a case, the minimal subspace is only achieved when all the scans are perfectly aligned.

The details of the technique used in this work and its evaluation are outlined in the next section.

4.3 Technique

This section starts with an overview of the problem of constructing an accurate Morphable Model of the face and ear. The data and its preprocessing to remove occluders and noise are then described. Next, the stages of the registration are outlined, including the feature point placement, initialisation and precise fitting. This also covers the robust sparse optimisation algorithm used in the fitting process. The section concludes by describing how the fitted scans are used to construct the model.

4.3.1 Base Mesh

In many existing approaches to Morphable Model construction, a base mesh created by an artist is used to register the range scans. One source for such a mesh is the MakeHuman project[1]. This provides an open source application that simplifies the creation of 3D models of human bodies. It includes a mesh that has been carefully produced through consulting with some of the most respected 3D human modelling artists. The mesh can create an accurate approximation to the ear shape using relatively

[1] MakeHuman http://www.makehuman.org/blog/index.php

few vertices. More detailed meshes can be obtained by using subdivision algorithms [44] (Figure 4.5).

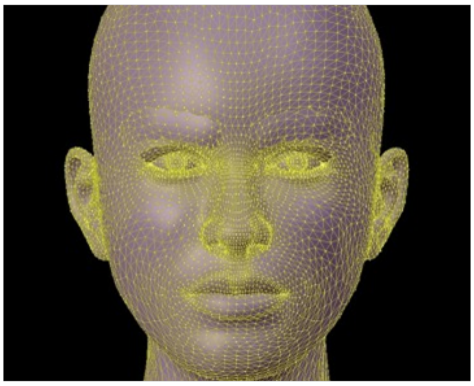


FIGURE 4.5: An image of the MakeHuman head mesh after 1 subdivision.

4.3.2 Training Data

Generally, it is desirable to use existing datasets so that model construction techniques can be replicated and compared. One of the largest of these datasets is the one provided by the Face Recognition Grand Challenge [91]. This data is in the form of front facing range data and colour values. The range data is estimated to be accurate to within 0.4 mm. As the images are front facing, however, the ear detail is restricted. A smaller, but more relevant source is the Notre Dame Biometric Database J [112]. This dataset was obtained with the same laser scanner as the FRGC data and was designed specifically for ear recognition. It contains 324 profile head images and has been used to evaluate a number of existing 3D ear recognition algorithms [100] [60] [112].

The original work by Blanz and Vetter [17] used complete head scans with minimal noise and uniform lighting. As a result, they could perform an optic flow calculation based on the surface appearance. In contrast, the Notre Dame data is less constrained and less complete. In particular, it does not cover the entire head surface and is recorded with varying poses and under differing lighting conditions. This makes an optic flow calculation less reliable. More recent work, such as that of Allen *et al.* [3], has focused on deforming a generic base mesh to align it with the range data. The aligned mesh parameterises the range data and registers the samples. This is the approach that has been used in the work described here.

4.3.3 Preprocessing

The scans used by Blanz and Vetter [17] were preprocessed manually to remove hair and provide an initial head alignment. They also made their subjects wear caps to minimise hair occlusion. Allen *et al.* also used caps [3] and marked regions of the base mesh as being scanned inaccurately. These points then had no influence on the fitting process. The range scan data used in the work described here only covers a part of the head volume. The ear regions also contain significant self-occlusion. General hole filling algorithms have been developed that could complete these scans [53]. However, the internal complexity of the ear is likely to mean that these algorithms will produce incorrect results as they make smoothness and convexity assumptions that are not valid for an ear shape. For this reason, the mesh registration and model construction steps described here have been designed to work with partial data.

The Notre Dame data also contains partial occlusions (Figure 4.6) as well as noise generated by the range scanning process. This is particularly noticeable near the Intertragic Notch where there is a sudden change in depth. This region is circled in Figure 4.7.

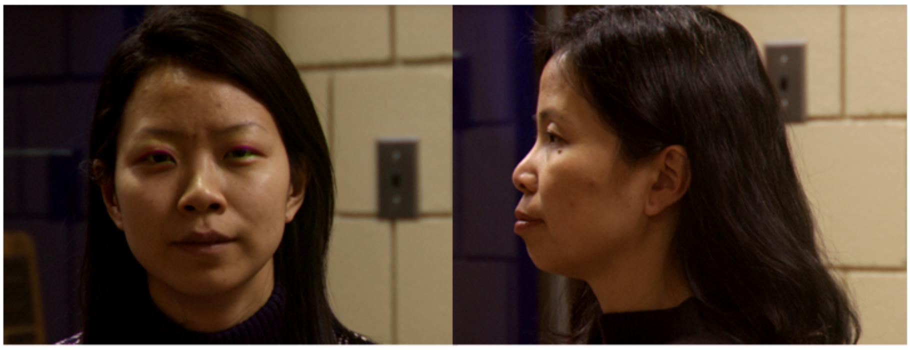


Figure 4.6: Images of minor hair occlusion

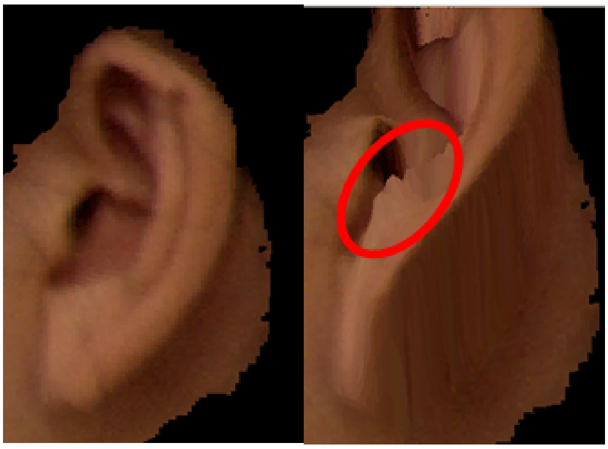


Figure 4.7: Left: A range scan of the ear Right: A tilted view revealing inaccurate mesh regions

4.3.4 Automated surface classification

To address these issues, the model construction process described here has been adapted to be robust to both occlusion and noise. This is achieved by automatically labelling parts of the mesh as invalid. These invalid regions do not contribute to the fitting

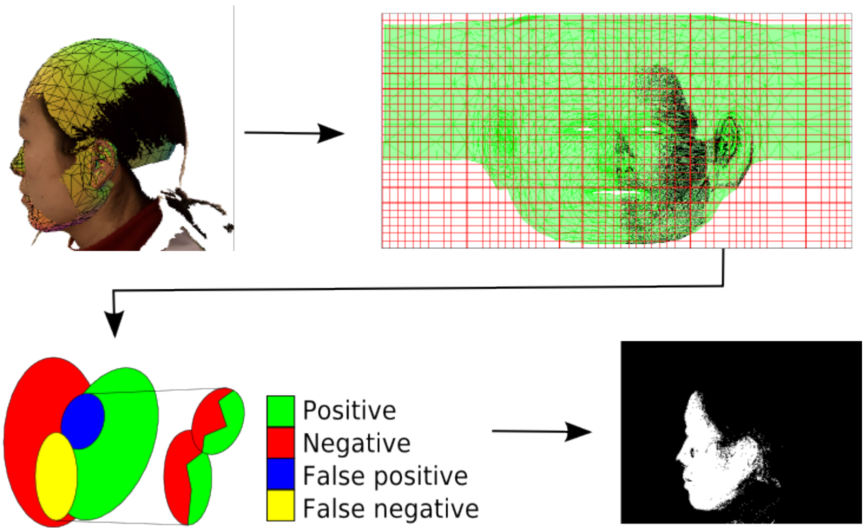


FIGURE 4.8: A flow chart showing the classification algorithm

process. The labelling process uses a classifier trained on 30 manually labelled images. Figure 4.8 shows a flow chart summarising this process.

To classify the surface, the model was first aligned and deformed using a set of labelled feature points. The range scan was then given an approximate registration by calculating the closest mesh surface points to range scan pixels. Using this registration, each point on the mesh surface was assigned a two dimensional uv location. This uv-map of points was then split into a number of overlapping regions. Each region was used to train a separate classifier built from the Spin image of each pixel of the manually labelled examples. The classifier approximates the space of positive and negative samples with a pair of Gaussian models. These models are used to estimate the class of each training sample, with the closer model providing an estimate of validity. The values that are incorrectly estimated by this process are grouped into sets of false positive and false negative samples. Each of these samples is used to calculate an additional pair of Gaussian models representing their regions. Within each error region, memory efficient kdtrees are constructed for classification. The kdtrees are used to evaluate the k nearest neighbours' estimate of each region efficiently. In this way most of the potential range samples are classified with the efficiency of a simple Gaussian classifier while still maintaining the accuracy of a detailed decision boundary between classes. Once classified, each overlapping region contributes votes to the validity of its pixels. The estimate with the most votes is returned as the classification. In cases where the votes are equal, a negative classification is returned. The accuracy of the labelling process is analysed in the evaluation section.

4.3.5 Feature Points

Similar to the work of Blanz and Vetter, Allen *et al.*, and Anguelov *et al.*, the base mesh is initialised using manually placed feature points, as indicated in Figure 4.9. The feature points used in this work are based on the Mpeg 4 facial feature points plus an additional 8 points at locations over the ear. These points provide the needed initialisation so that the latter fitting stages can converge accurately. The Mpeg 4 facial feature points include the ear lobe point which initialises the ear position. Two additional points were then added at the Intertragic Notch and the Anterior Notch to enable the opening to the ear canal to be aligned accurately. The remaining six points are needed to ensure that the ear helix is correctly registered. Without these points, the outer ear shape can become stuck in local minima, incorrectly deforming to fit inner ear regions.

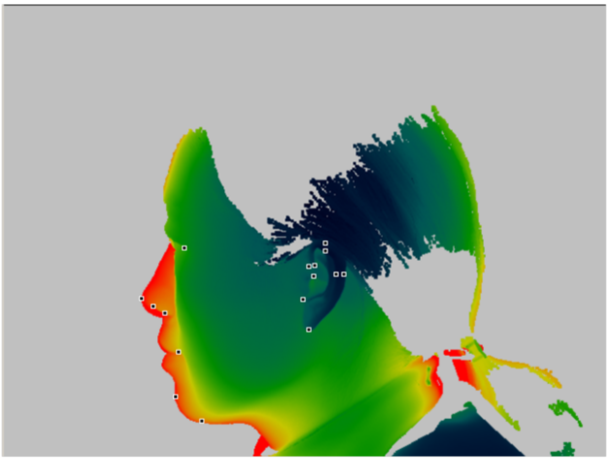


FIGURE 4.9: The location of the manually labelled feature points

4.3.6 Initial Registration

Using the 3D position of the marked feature points, the Procrustes algorithm is used to calculate the least-squared error rigid transformation between the base mesh feature points and those in the range scan (Figure 4.10). Once initialised a more precise fit is obtained.

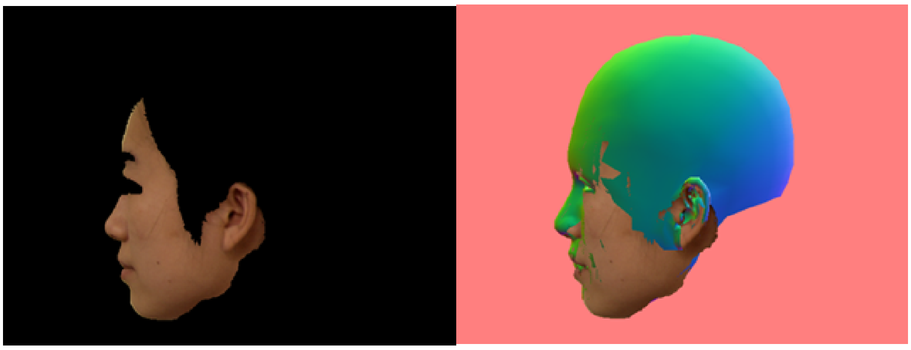


FIGURE 4.10: Image of least squared rigid registration of base mesh to feature points

Another issue is that the Procrustes alignment can result in a transform that includes a reflection due to the approximately planar nature of the ear and its feature points. This can be prevented by adjusting the rotation calculation within the Procrustes algorithm, as follows:

Calculate the optimal translation using the relative position of the centroids (M_a and M_b) of the two pointsets (P_a and P_b).

$$t = M_b - M_a$$

Calculate the correlation matrix of the two pointset's with respect to their relative offsets from their centroids.

$$A = P_a - M_a$$

$$B = P_b - M_b$$

$$C = B^T A$$

Decompose this matrix using singular value decomposition.

$$C = UDV^T$$

Calculate the rotation. A reflection is present if the determinant of this rotation matrix is negative.

$$R = VU^T$$

Determine the closest non-reflecting transform by inverting the smallest singular value of the U matrix (producing a new matrix W)and then recalculating the rotation.

$$R = VW^T$$

4.3.7 Precise Registration

The original work by Blanz and Vetter had complete head scans with minimal noise and uniform lighting. As a result, they could perform an optic flow calculation based on the surface appearance. However, the Notre Dame and FRGC data are less constrained and incomplete. In particular, they do not cover the entire head surface and are recorded with variations in pose and lighting. This makes the optic flow calculation less reliable. More recent work, such as that of Allen *et al.* [3], has instead focused on deforming a generic base mesh to align it with the range data. The aligned mesh parameterises the range data and registers the samples. This is the framework that has been used for the approach described here.

4.3.8 Optimisation Based Fitting

The fitting optimisation problem can be formulated in a number of different ways. In all cases, however, the goal is to define the optimisation such that when the error values are minimised the resulting registration will lead to an optimal model. The ideal error values are those that smoothly and monotonically decrease towards a solution. Under these conditions, existing non-linear least squares based optimisations can be used to find a solution. In addition, it is advantageous if each parameter affects only a small number of error values. If this is the case, the majority of constraint matrix entries will be zero, allowing efficient sparse variants of linear algebra methods to be used. The technique described here is based on the approach developed by Vlasic *et al.* [105]. By using a general non-linear optimisation framework, constraints can easily be adapted and incorporated without a major change in the underlying algorithm.

One difficulty with Morphable Model construction is that it is not obvious how best to define fitting constraints so that a minimal value corresponds to an optimal result. The technique described here uses similar constraints to existing approaches. In particular, the following properties are minimised:

- *Distance to feature points* The features points are labelled on the range image by selecting the range point that most closely represents the feature. The error is a three dimensional value representing the relative displacements of the Mesh features in the X, Y and Z directions. The feature points are defined on the mesh as points on the triangles of the surface of the mesh. Changes in the vertices of the triangle are then the only parameters that influence these error values.

  This can be formalised as:

  $$Ep = FP - (a * T0 * P0 + b * T1 * P1 + c * T2 * P2)$$

  Where Ep is a 3 dimensional vector of errors, FP is the labelled world position, a, b and c are the scalar barycentric coordinates of the feature point on the mesh. $P0$, $P1$ and $P2$ are the 4 dimensional untransformed positions of the mesh with an additional value of 1 in the fourth dimension to enable homogeneous transformations. $T0$, $T1$ and $T2$ are the $3x4$ transformation matrices which are associated with each mesh point.

  Each feature point adds 3 error terms corresponding to the difference in X, Y and Z dimensions of the mesh point and the labelled point. Each point is influenced by 36 parameters which define the transformation matrices.

- *Smooth deformation* In the work of Allen *et al.* [3] and Vlasic *et al.* [105] the smoothness error is formulated by associating a homogeneous transformation matrix with each vertex of the mesh to be aligned. Smoothness is then maintained by minimising the Frobenius norm between the matrices of connected vertices, where connection is determined by the edges of the mesh. This causes a preference for regions of consistent local affine transformation, namely rotation, scaling and shearing. These operations maintain the surface continuity and preserve local detail, resulting in a smooth deformation. In this way the parameters of the fitting process are the elements of the matrices of each vertex.

This can be formalised as:

$$Et = (T0 - T1)$$

Where Et is a $3x4$ matrix of error terms associated with the differences between each $3x4$ transformation matrix associated with vertices that are connected by edges over the mesh.

- *Alignment of mesh surface to range points* Each valid range point has the closest mesh surface point estimated. These are parameterised using barycentric coordinates of the closest triangle. Each point places a constraint on the three vertices defining the triangle. This is in contrast to existing work which uses the distance between mesh vertices and their closest range points as an error term. The approach proposed here enables a more accurate alignment of any given resolution of mesh.

  This is formalised in a similar manner to the feature points.

$$Erp = RP - (a * T0 * P0 + b * T1 * P1 + c * T2 * P2)$$

The total error is a a large vector of all the other terms concatenated together, with a different weight for each type of constraint. In total the algorithm has 2 configuration parameters which determine the relative error contribution of the feature points, the smoothness regularisation term and the closest range point fitting values.

4.3.9 Optimisation Algorithm

The equations described above are linear equations of the form $\mathbf{A}x = \mathbf{b}.$ The optimisation algorithm solves this set of equations to minimise the least squared error.

$$S(\mathbf{v}) = \sum_{i=1}^{m} e_i(\mathbf{v})^2$$

In the context of the fitting algorithm, the value of v is constructed by concatenating all of the parameters of the transformation matrices together to produce a single large vector that defines the parameters of the fitting algorithm. Each of the e_i values (from 1 to m) represent the dimensions of the error vector. This vector has been constructed by concatenating the Ep, Et and Erp terms together to get the total error value for a given mesh fitting.

In their original formulation, and that used by most 'black box' optimisation algorithms, the solution to this problem requires that A is non-singular and therefore has a unique inverse. This restricts the use of these algorithms to problems that have more error terms than parameters. In addition, it means that the rate of change of the error terms due to any parameter must be linearly independent. Even for problems where these properties hold, numerical inaccuracies can lead to approximately singular matrices and thus cause the optimisation to fail. However, this restriction can be avoided by using Singular Value Decomposition to calculate the pseudo inverse of A. This will simultaneously find the least squared and least norm solution. Thus, for under-constrained problems, where there are multiple solutions to the local quadratic error function, the step of the solution with the smallest change will be chosen. Using this technique, a solution can be found reliably for the majority of error functions.

4.3.10 Sparsity

Due to the large number of parameters in the optimisation problem, there is a need to use sparse linear algebra routines to calculate the results efficiently. However, calculating the pseudo inverse directly destroys matrix sparsity and makes the calculations impractically slow. In addition, alternative approaches using sparse QR, LU or Cholesky factorisation will fail due to the potentially singular matrix. One way to address this difficulty is to apply the MinRes algorithm developed by Paige and Saunders [82]. The MinRes algorithm uses a conjugate gradient method to solve singular $\mathbf{A}x = \mathbf{b}$ problems directly, while only requiring the repeated evaluation of $\mathbf{A}x$. This calculation can be performed efficiently with sparse matrices and thus enables results to be calculated quickly and with stability.

4.3.11 Robustness

The optimisation algorithm has also been adapted to increase its robustness by excluding some of the constraints that contribute the most significant errors in each iteration. Distance constraints are occluded if they are greater than a threshold distance. This distance is calculated by approximating the distances using a Gaussian model and excluding values that are greater than three standard deviations from this model. The model's mean and standard deviation are robustly estimated using the median and median absolute deviation of the constraint distances.

4.3.12 Data Normalisation

Once a number of range images have been registered, the model can be created. This involves calculating the mean and principal components of the shape. However, the range meshes may contain variations due to the pose of the subjects in the range scan. These variations are partially addressed by the initialisation using feature points but may not represent the optimal normalisation necessary to construct a minimal model. To address this problem, models are normalised by applying the Procrustes transform. An added complication is that fitted meshes are only valid for part of the surface due to noise and occlusion of the range data. In each case, only this valid subset of the data is used for normalisation. Each fitted mesh is registered to the original base model shape using the vertices that were constrained by the valid range data. The position of the remaining vertices are then estimated using the smoothness constraints. This creates a smooth head model with no visible artifacts and exploits the implicit anatomical information contained within the base mesh. As all of the head samples cover a similar surface region this normalised head shape remains similar between multiple scans of the same subject. Once normalised the PCA algorithm is used to construct the model.

4.4 Morphable Model Evaluation

The work here builds on existing approaches by examining a number of metrics for determining the quality of the registration and the resulting model. The model has been constructed using 160 training images taken from the Notre Dame Biometric Database J [112].

4.4.1 Automated classification accuracy

To evaluate the accuracy of the noise and occlusion classifiers, ten-fold cross validation was applied to 30 hand labelled training samples. The distribution of accuracies can be seen in Figure 4.11 where the percentage of regions within a given error range has been identified. Over half of the head surface is either not visible within the training set or consists exclusively of skin or occluder samples. The majority of the remaining regions have been correctly classified, but there are still some surface areas with many errors. The robust fitting process compensates for these errors and in many cases can correctly identify all of the occluding and noise regions without the need of the additional classification process.

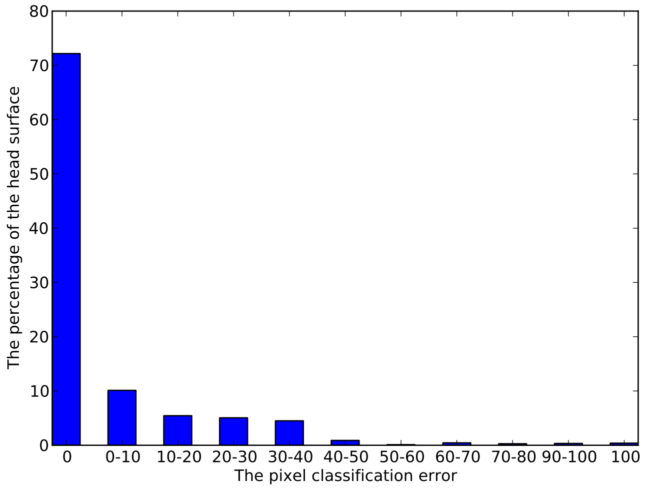


FIGURE 4.11: Graph showing the percentage of the head region that can be classified within a given error range

4.4.2 Fitting consistency

The fitting consistency evaluates the uniqueness of each registration by comparing the similarity of meshes produced when two range images of the same person are fitted independently. It is calculated as the average distance between vertices of triangles that have been constrained by both range images. To compensate for variations in the pose of heads in each scan, the meshes are normalised to the base mesh using the Procrustes transform applied to the shared vertices. Figure 4.12 shows the relative distances between the closest registered head and the next closest nine scans. For all 160 samples within the training set, the closest head is the scan of the same person taken at a later date. The significant difference between the closest and the other scans indicates that the registration process is effective in extracting a consistent shape that can be used for recognition. Examples of these registrations are shown in Figure 4.13.

4.4.3 Model metrics

The model has been evaluated using the criteria of generality and total variance as defined in Section 4.2. The results are presented in Figure 4.14 and Figure 4.15.

4.4.3.1 Generalisation

Generalisation is measured using a ten-fold cross validation technique, implemented by excluding a set of training images from the model construction process and then measuring the accuracy with which the fitted images can be reconstructed using the

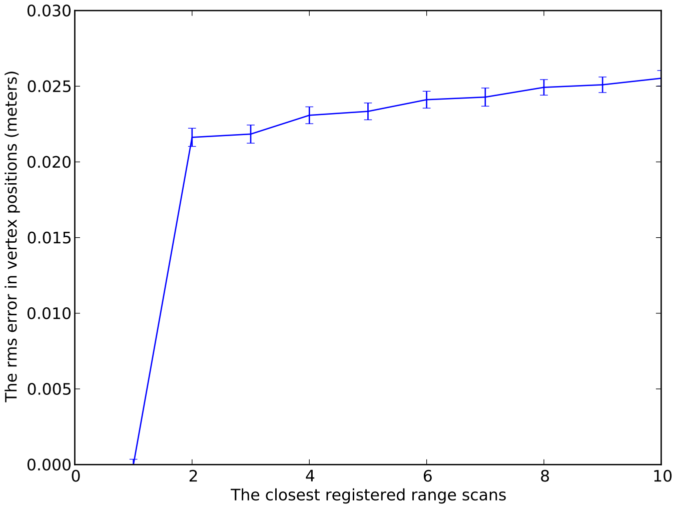


FIGURE 4.12: Graph of fitting consistency

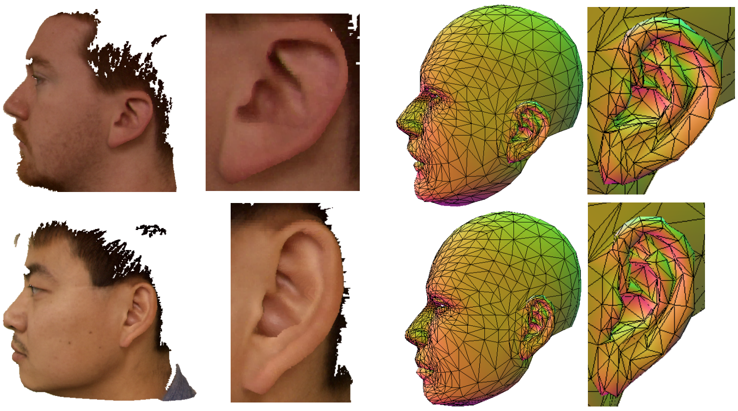


FIGURE 4.13: Two examples of registered head models

model. This is evaluated by projecting the fitted meshes into the space of the model and measuring the root mean square (rms) of the difference in vertex positions between the fitted meshes and their representation using the model. The results can be seen in Figure 4.14.

4.4.3.2 Total Variance

It is not possible to calculate the redundancy in the model directly. However, a measure of the total variance within a given model can be produced. Models with a smaller total variance have a closer fit to the data and thus are likely to have less redundancy. Total variation smoothly varies with the number and magnitude of the Eigenvalues. This is a simple measure that is sufficient for comparing models. Alternative metrics, such as

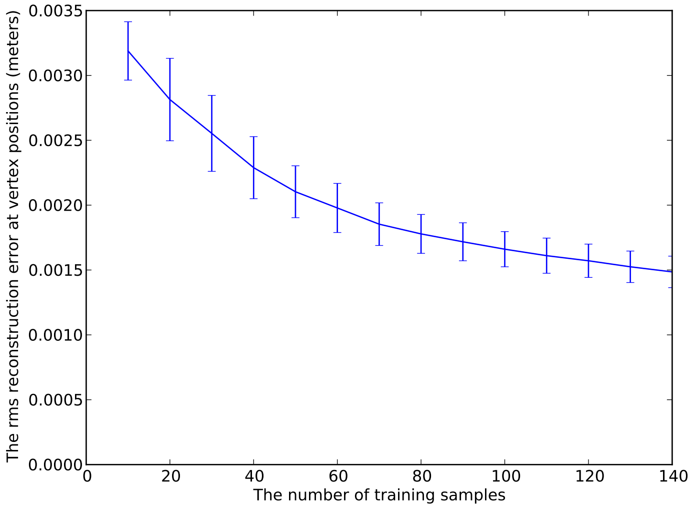


FIGURE 4.14: Graph of generality

the model volume or the largest displacement for a given probability, have undesirable properties. For example, the volume is reduced when an additional small eigenvalue is added to the model. Also, if the largest displacement is used, all dimensions smaller than the largest eigenvalue are ignored. Variance can thus be calculated using the equation:

$$variance(\mathbf{e}) = \sum_{k=0}^{l} e_k$$

where **e** represents the model's eigenvalues. As with generality, the error in these values has been calculated using 10-fold cross validation. Figure 4.15 shows the results. The graph of generality has a much faster rate of convergence than that of variance. This may be explained by errors within the registration process increasing the variance of the model without significantly improving its generality.

4.5 Conclusions

This work demonstrates the first complete Morphable Model of the head that is explicitly designed to recreate both the face and ear shape accurately. This approach includes the identification of a key set of ear feature points to enable accurate registration. In addition, it improves the robustness of existing Morphable Model construction techniques by using classifiers trained to detect occluding and high noise areas. It also provides a framework for evaluating the quality of the registration of training scans, and metrics to assess the resulting model quality. These metrics are used to indicate the improvement in the generalisation capabilities of the model as the training set size is increased. The

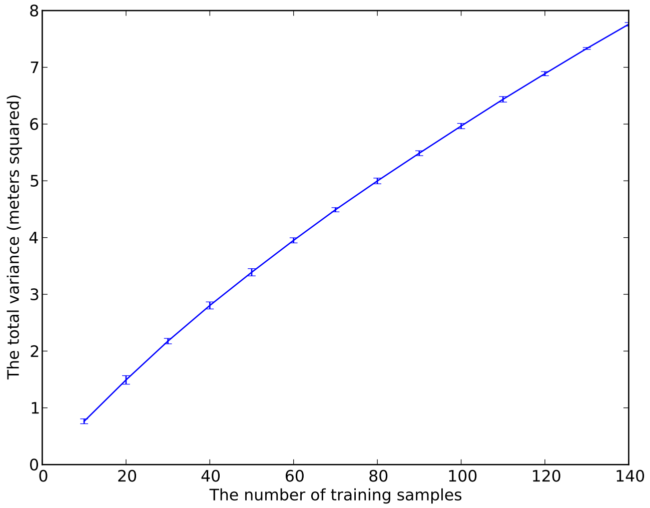


FIGURE 4.15: Graph of total variance

evaluation shows that the described technique extracts a consistent shape for a given individual and that within the error margins of the registration process, 160 training samples are close to achieving convergence of the model.

Further work in this area will involve the use of the evaluation framework to examine other Morphable Model techniques, such as the correlated correspondence algorithm of Anguelov *et al.* [5] and the non-rigid iterated closest point algorithm developed by Amberg *et al.* [4]. The framework can also be expanded to evaluate the utility of the model in inferring identity from partial data such as detected features within 2D images.

There is also scope for creating an improved surface classifier, through additional distance constraints, and providing improved speed and precision using support vector machines or boosting techniques. If this classifier could be combined with an automatic feature point detection process, fully automated model construction from relatively unconstrained training data would be possible. This offers the potential for more widespread application of Morphable Models within object recognition.

To complete the construction of the Morphable Model, the surface texture variation also needs to be calculated. First the surface textures associated with each range scan are extracted. These are calculated by mapping the colour values of the range points onto the uv-map of the fitted mesh. The resulting variation in appearance is then modelled by applying PCA to the extracted textures. One difficulty, however, is that the range scans are recorded with uncalibrated lighting. This is in contrast to the original Morphable Model technique developed by Blanz and Vetter [17] which used uniformly lit 3D scans. By using uniformly lit textures the effect of varying light direction can be simulated using a rendering algorithm. Further work is needed to investigate whether it is possible to reconstruct uniformly lit textures by applying a shape from shading technique to the unconstrained textures. When combined with the shape modelling approach described in this chapter, such a lighting normalisation algorithm would enable the construction of complete Morphable Models from largely unconstrained range scans.

Chapter 5

Conclusions and Discussion

The work described in this thesis examined the hypothesis that

The constraints on the use of ear-based biometric systems can be relaxed significantly through the introduction of robust recognition algorithms.

It began by clarifying the main obstacles to achieving unconstrained recognition and after examining the existing approaches to addressing these issues, proposed two new techniques. This chapter provides a critical analysis of these techniques followed by a discussion of possible future work.

5.1 SIFT Based Ear Recognition

The first new technique presented was a robust and automated 2D ear registration and recognition algorithm. By taking a SIFT point based approach it provides the first reliable solution to identifying the position and rotation of ears even when they are occluded. The evaluation of the technique demonstrated that the approach is practical for small datasets and relatively controlled pose and lighting conditions. A detailed evaluation, however, highlighted a sensitivity to pose variation with recognition rates dropping significantly beyond 20 degrees and falling to zero at 40 degrees or more. Another limitation is that the time taken to recognise a new ear is directly proportional to the size of the gallery. This means that the approach is only practical when applied to verification, or when recognition is across small datasets. Recent developments in internet scale image matching could be applied to this technique to make it practical for large scale datasets. However, the additional issue of pose variation requires a new approach.

5.2 Morphable Model of the Face and Ear

The second technique presented contributes to a more complex and potentially much more robust solution to ear recognition. It uses a 3D Morphable Model based approach to address the variations in pose and lighting. This technique also enables efficient large scale recognition using the parameters of a fitted model.

The first step in developing the approach was to create an algorithm for constructing a Morphable Model that included an accurate measurement of the ear. This is difficult, because without very accurate initialisation, an ear mesh will not converge to a tight fit with range scan training data. Through manual experimentation, a set of feature points have been identified that significantly improve on existing fitting quality. The evaluation of the technique showed that this process was sufficiently accurate for recognition purposes. However, when the registrations are examined in detail it is evident that further work is needed to obtain a perfect alignment with the scans. One possibility is to use a very different registration approach, such as the correlated correspondence algorithm. This algorithm uses local surface signatures and graph matching to produce a globally optimised alignment between the mesh and the scan.

Another goal in developing this technique was to improve the robustness of the model construction process. A consistent theme throughout this work has been the development of algorithms that reduce the need for manual input from users. As a small step towards fully automating the Morphable Model construction process, an automated surface classification technique was developed. This technique supports model construction with noisy and partially occluded range scans. Further developments of this technique have revealed that the median absolute deviation of the constraints can be used to compensate for any classification errors. The measurement provides a basis for a threshold which excludes occluders and noise. This has proved to be a very effective approach and, for many training scans, makes classification unnecessary.

The work also includes a new set of metrics for analysing the quality of the construction process and the accuracy of the resulting model. These metrics served not only to validate the technique but also to provide a basis for further automating the model construction process. Using the metrics it should be possible to evaluate the quality of an individual training sample and thus exclude or reduce the influence of unsuitable scans.

5.3 Further Work

5.3.1 Completion of the Morphable Model

The first area for improvement is the Morphable Model itself. To render accurate ear appearances, the model needs to be able to recreate the ear's surface colour and reflectance properties. One approach is to use a dataset with calibrated lighting variations to calculate the surface properties for each scan. These properties can then be modelled using a subspace in a similar manner to the shape variations. It may also be possible to use shape from shading techniques to estimate the lighting and surface properties directly from uncalibrated training images. As a result, relatively unconstrained scans could be used enabling a step towards the development of a robust and fully automated Morphable Model construction algorithm.

In addition, if the face is to be combined with the ear, the model needs to be robust to varying facial expressions. This can be achieved by creating a tensor Morphable Model built using an expression varying training set. Such a model would enable the identity and facial expressions to be altered independently. Once the model had been fitted to an image the identity signature could then be separated from the expression parameters.

5.3.2 Morphable Model Fitting

To create a complete recognition algorithm based on the Morphable Model approach three areas require further development. These are:

1. Robust ear detection
2. Model initialisation to a detected ear
3. Detailed Morphable Model fitting

Significant progress has been made in addressing the detection and model initialisation stages of this further work. The details of this technique are outlined in Appendix A. The technique is designed to provide a means to detect an ear and initialise the Morphable Model so that its appearance approximately matches the ear within the image. The technique uses a Viola Jones detector to locate the ear. Key feature points are then identified within the detected ear region. These points are used to align and parameterise the model. The focus of the work has been precise localisation of the feature points so that the initialisation can be made as accurate as possible.

The technique models the variation of each feature point using a Gaussian Eigen space. The inverse compositional alignment algorithm (IC) can be applied to fit these models accurately.

The technique includes three refinements of the original IC algorithm to improve precision. Firstly, the fitting algorithm uses the model likelihood to regularise the fitting. This ensures that the feature localisation process doesn't overfit to points with improbable appearances. The second contribution is to include the expected variation in normalised brightness and contrast within the regularisation term. This addresses a weakness of the original IC algorithm, where the technique can incorrectly converge to smooth uniform parts of the image. In the original algorithm no restriction was placed on the contrast and brightness values of the model. As a result, the fitting algorithm would set the contrast to zero for smooth regions, at which point the fitting algorithm would then stop as no further improvement could be made. The final refinement is to include a robust model learning process to compensate for any inprecision in the labelling of the training set.

Once the points have been identified, a Levenberg-Marquart based fitting algorithm is used to calculate the optimal positioning and model parameter values to align the model. Existing work addressing Morphable Model initialisation has focused on creating fast approximate solutions to this fitting process. However, as the number of feature points is relatively small, a full perspective camera model has been used as the optimisation is not particularly slow relative to the time taken by the feature detection stage.

On a small evaluation set this approach has produced reasonably accurate results. However, further investigation is needed to determine if it is suitable for locating feature points across large pose and lighting variations.

To complete the development of the model initialisation algorithm it is necessary to train a Viola Jones detector using ears with varying poses and lighting conditions. One approach is to use the Morphable Model to render a very large training set of ear appearances. In practice this would enable use of a much larger dataset than could be obtained through recording and labelling test subjects manually.

Another advantage of using synthesised data is that it can be used to determine the regions of the ear that are most reliably and efficiently detected under varying conditions. These regions can be calculated by creating custom feature point detectors at uniform points over the ear surface. The accuracy of each detector can then be compared using an additional test set of synthesised ear images. The most reliable detectors are selected to maximise the reliability and efficiency of the initialisation process.

The initialisation stage may not be precise enough for the model parameters to provide accurate recognition results. Therefore, it may be necessary to include an additional detailed fitting step. In Blanz and Vetter's original work, model fitting was performed by minimising the difference between a rendered instance of the model and the image. More recent work fits the model in stages, matching the image edges and specular highlights prior to complete fitting. Further work is needed to evaluate this approach and to determine to what extent accurate feature point localisation can improve the process.

5.3.3 Comprehensive Evaluation

Once a recognition system is complete, a large scale evaluation is needed to demonstrate that recognition can be performed reliably. As with the evaluation of the SIFT based recognition system, a detailed evaluation of unconstrained factors can provide a means to identify the technique's weaknesses and help prioritise future developments. However, there are practical limitations on the size of evaluation and training datasets. As such, it would be valuable if the performance of large scale evaluations could be predicted from smaller tests. The analysis of the generality and variance of the Morphable Model shows that these metrics can vary smoothly with the number of training and testing samples. The implication is that there is a predictable consistency in how recognition performance varies with training size. Further work could investigate this consistency and make estimates of the optimal number of training scans and the expected large scale recognition performance under varying conditions. Similar approaches could also be used to evaluate whether synthesised training data can achieve similar detection and feature localisation accuracy to that of manually recorded training samples.

An additional area for improvement is that both the Morphable Model construction and feature point classifiers use single linear Gaussian models. Further work could investigate to what extent the actual distribution of training samples matches this model. More complex models may be more accurate and lead to improved recognition performance.

5.3.4 Optimised Configuration Parameters

Both the SIFT and Morphable Model approaches have configuration parameters that have been manually adjusted to produce accurate results. By having a computable measure of quality it is possible to optimise these parameters automatically. One simple approach is to perform a systematic grid search of the parameter values. However, when there are many parameters and a large training set this search may be impractically slow. An alternative strategy is to calculate configuration parameters directly from the training data. For example, in the case of the SIFT technique, thresholds on the number of point comparisons can be estimated by calculating the number of matches that are needed for all of the training images to include four correctly detected points. Further work might extend these ideas to develop algorithms in which all of the configuration parameters are optimised automatically.

5.3.5 Further Modelling Techniques

One interesting future direction is to adapt the initial detection algorithm so that it returns an estimate for the pose, shape and lighting parameters as well as identifying a potential face or ear. This could be achieved by synthesising a large database of potential appearances and then using large database indexing techniques, such as Locality Sensitive Minimal Perfect Hashing to provide a very rapid means of testing whether an image region was close to one of the synthesised images. The image parameters would then provide an initialisation for the model.

Another approach is to use real time 3D sensors, such as stereo or structured light cameras, to estimate the 3D face and ear shape. Pose and illumination invariant signatures could then be calculated from the scans and used for detection and model initialisation. In principle, such an approach is more robust than 2D images as it can deal with variations such as makeup and complex lighting conditions that may be difficult to model. However, further analysis is needed to determine if the output from such sensors is sufficiently precise for accurate recognition.

A more ambitious goal is to fully automate the Morphable Model construction process as part of a general online vision system. Such a system would construct robust models of novel objects and then use the models to recognise those objects in unconstrained scenarios. If the resulting system could be made reliable, it would represent a significant step towards addressing the long term challenge of general robust object recognition and would have numerous practical applications.

5.4 Concluding Remarks

There are many challenges that must be overcome in order to recognise ears in unconstrained environments. This thesis has described a number of important steps in addressing these problems. It is hoped that through this and further work, robust recognition can become an established feature of future biometric systems and make an important contribution to a general vision system.

Appendix A

Model Initialisation

This appendix outlines ideas on a new approach to the initialisation of a Morphable Model. Existing Morphable Model fitting algorithms initialise a model using feature points. However, these fitting algorithms are often evaluated with manually specified feature points [18]. This can lead to misleading recognition performance as accurate initial detection and feature point localisation are difficult to achieve and errors at this stage can cause subsequent fitting steps to fail. It is therefore critical that the feature detection process be as robust and accurate as possible.

A.1 Existing Research

A significant volume of detection research is devoted to finding faces within images. As a result this section refers almost exclusively to faces. However, in each case the approaches described can be adapted for use in ear detection.

A.1.1 Occlusion Robust Detection

One of the most popular detection algorithms is the Viola Jones face detector. This algorithm can achieve real-time performance under a wide range of conditions. In addition, this technique can also be applied to the localisation of sub-features. For example, both Medioni *et al.* [72] and Everingham *et al.* [45] have used this technique to detect facial feature points in unconstrained settings. In both approaches, a model of the probabilities of relative feature locations was used to improve feature point localisation.

While the Viola Jones approach can be fast and reliable, the robustness of the technique is strongly dependent on the training set. In particular, its robustness to occlusion is dependent on training data that includes expected occluders, such as, for example images of subjects wearing glasses. This greatly increases the training set size and limits the robustness to occlusion that can be achieved.

To address this sensitivity, other approaches have been developed. These directly detect face components, such as the eyes, nose and mouth and then combine the results of these classifiers to determine if an object has been found. One example of this approach is that developed by Heisele *et al.* [52]. In their work they used a Morphable Model to train a set of support vector machine (SVM) feature detectors. The output from these detectors was then used as input to a final SVM classifier which determined if a face was present. Given a test image, this system of detectors was evaluated at multiple positions and scales to detect and locate faces. When compared against a standard implementation of the Viola Jones detector[1] this technique had significantly improved ROC curves, resulting in a much higher percentage of correct detections for any given false detection rate. More recent work by Caunce et al. [28] uses view-based texture patches learnt from the average texture calculated by registering 923 head meshes. Their approach could automatically detect front facing feature points and track them with up to 70 degrees of pose variation.

An alternative to the custom feature detector approach, is to use general interest point detectors such as SIFT. These points can then be classified to determine if they represent object parts. This was the approach used by Dorko and Schmid [42] to detect cars. They explored a range of interest point detectors and classifiers to determine the best approach. Classifiers were created by manually labelling detected interest point regions on objects. Each labelled region had an associated classifier constructed for it which was then reduced using subset selection. For their dataset they found that the most accurate feature detection results were obtained by using a Gaussian Mixture Model to detect each part, and maximum likelihood for subset selection.

Yet another approach is to split the test region into a number of separate sections that are tested independently. Each section that returns a positive result is considered to be unoccluded. These sections can then be combined and tested to detect if an object is present [80] [64]. Such approaches indicate no loss of detection accuracy with up to a sixth of the object being occluded.

A.1.2 Precise Localisation

The majority of existing work is devoted to detection accuracy, with relatively few papers evaluating the precision of each object's localisation. One difficulty with addressing both precise localisation and accurate detection is the conflicting requirements on the properties of an accurate detector. To be reliable, a classifier needs to be robust, in that it must detect an object with a large range of appearances. However, the classifier must also

[1] The OpenCV library is available at http://opencv.willowgarage.com/wiki/

be *sensitive*, meaning that it must return a negative result when an object is not precisely located. This causes a problem for many object detection systems. In particular, it is a issue for systems which iterate over an image at a fixed set of positions and scales. The object may not be precisely located at any of these locations. At each location the classifier is then used to determine whether the local region represents the object or not. Because these discrete regions may not exactly align with the object in question, the classifiers are trained using multiple offset samples (Figure A.1). This increases the robustness of the detector but at the price of reducing the precision of the object's localisation. One way of improving precision is to take the output of an imprecise robust classifier and refine the localisation with a more sensitive technique, such as the Inverse Compositional Alignment algorithm [11]. This was originally developed for aligning Active Appearance Models and performs a non-linear optimisation to find the best affine transform and parameters to minimise the difference between a model and an image. The algorithm can be applied with models as complex as an AAM or may represent something simpler such as a single Eigen image model. Recent work by Papandreou and Maragos has extended this approach to incorporate priors of likely model parameters [83]. By incorporating these priors, the optimisation algorithm fitting iterations are reduced, fitting converges more frequently, and fitting becomes more precise.

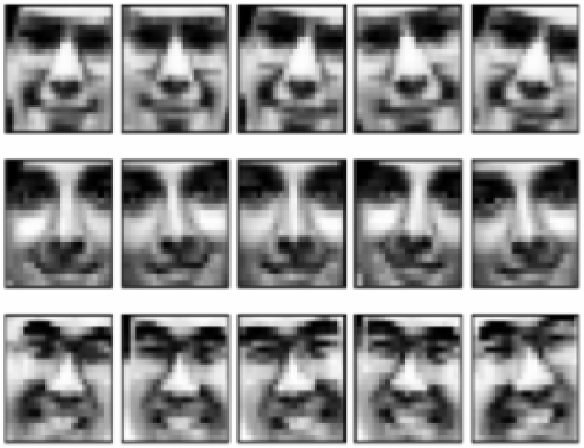


Figure A.1: Randomly mirrored, rotated translated and scaled faces [21]

An alternative approach is to train the detector to produce low values for marginally misregistered features so that it can be applied at multiple points around the detected location to find the best alignment. This approach was used by Brunelli and Poggio to achieve precise localisation of eye features [21].

A related issue is that training data may also be localised inaccurately. This has a similar effect to that of adding offset samples, limiting the precision of feature localisation. This problem has been examined by De la Torre and Black [38]. In their work they calculated an Eigen model for face features by aligning each training image so that resulting subspace was minimal in size (Figure A.2). Their approach was to construct an optimisation problem solved using a genetic algorithm. Significant improvements were achieved with their reduced dataset. One issue here, however, is that the genetic algorithm is likely to be impractically slow for large datasets. An alternative is to use a greedy model construction process. By building the model iteratively from a subset of samples and then aligning it to new data, an estimate of the minimal subspace can be obtained efficiently.

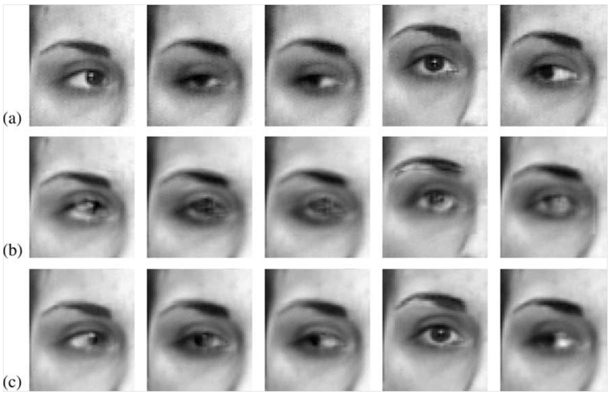


FIGURE A.2: A comparison showing the effect of optimised subspace model construction (a)Training Samples (b)Reconstruction using a PCA model (c)Reconstruction using the optimised subspace technique

A.1.3 Model Fitting

Once features are detected, model fitting can be applied to detect outliers in the detected points and improve existing feature localisation. This approach was used in the 2D ear registration algorithm described in Chapter 3. In addition, if only a small number of feature points are detected, a fitted model can guide the search for additional or more accurate features.

There has been a substantial amount of research into solving the model registration problem for objects with a rigid shape. In the literature this is commonly referred to as the Perspective-n-point problem. These problems assume the object is to be placed at the origin and the feature points then used to define a camera transform that produces the image. One of the most common approaches is to formulate the problem as a non-linear optimisation problem that is solved with Gauss-Newton optimisation. This requires a good initial estimate to avoid local minima. In addition, it can be expensive for large point sets. This has led to the development of a range of approximate solutions, such as constraining the camera model to be orthogonal [41]. By applying these constraints, the Jacobian matrix can be calculated more efficiently, which leads to improved performance on large point sets. Although these algorithms are very efficient, their results can be inaccurate when their assumptions are invalid, as, for example, when an orthographic camera model is used on images of objects close to the camera. Further developments have reformulated the problem to calculate both the depth of the feature points and the camera transform. This removes the non-linearity due to camera perspective and enables efficient and accurate calibration [67].

These algorithms produce good results for rigid objects but few of them have been generalised to flexible objects such as Morphable Models. The main approach in this case was developed by Blanz *et al.* [16]. It uses an orthogonal camera approximation and solves the calculation using an iterative optimisation approach.

One weakness with the model fitting techniques is that they generally assume the feature points have well defined, unique locations and that their values do not contain outliers. This makes them sensitive to errors in the feature detection process and prevents the use of additional information when feature points are incorrectly detected in multiple locations. Under these circumstances the registration problem has more in common with the more challenging task of fitting an object where the correspondences of points are not known.

One approach to achieve fitting without correspondences is to use the Iterated Closest Point (ICP) algorithm [94]. This proceeds by first finding the closest points between the features of the model and the features in the image. The model is then aligned to minimise the distance between these points. At each step of this fitting, the correspondences are recalculated. In this way, the model is gradually 'pulled' into position. However, one limitation of the approach is that it needs to be started close to the correct solution to avoid local minima.

A related approach is that used by the SoftPosit technique [37]. It uses the Soft Assign algorithm to estimate outliers and calculate one to one point correspondences efficiently. Similar to ICP, the technique uses the proximity of points to the model as an estimate of correspondence. It then uses the Posit algorithm to solve the Perspective-n-point problem very efficiently. This process is repeated until the algorithm converges or an iteration limit is reached.

Finally, for small point sets, a Random Sample Consensus (RANSAC) algorithm can be used [47]. This is a general technique for dealing with data containing many outliers. In the context of model fitting it involves randomly assigning model features to points in the image and then calculating the best alignment to minimise their displacement. The quality of the fit can then be evaluated to assess if the sample points are likely to be valid. With many points, this measure can be based on the number of points that conform to the registration. Other measures of fitting quality can also be applied. For example, as part of the model construction process the relative likelihood of different values of model parameters can be calculated. This likelihood can then be included as an error term in the fitting process.

A.2 Technique

The previous section has outlined the existing work in the key areas of object detection, feature localisation and model fitting. This section describes a possible new technique that addresses each of these stages in turn. The system is based on the approach used by Medioni *et al.* [72] and Everingham *et al.* [45] for detecting features in unconstrained environments. It starts by using a Viola Jones detector to locate the ear. This is followed by a feature localisation stage. In the work of Medioni *et al.* and Everingham *et al.* this was performed using more Viola Jones detectors trained for facial features. However, this approach has limited precision. In contrast, the new approach described here uses an enhanced version of the Inverse Compositional Algorithm for feature localisation. Once the feature points have been detected they are used to align a Morphable Model. This fitting process is performed by a robust Morphable Model registration algorithm that uses a full perspective camera model. This more complex model fitting process has been used to maximise the precision of initialisation. The details of this system are described in the sub-sections that follow.

A.2.1 Inverse Compositional Alignment with Model Likelihood

The Inverse Compositional Alignment algorithm (IC) adjusts a model's parameters iteratively to align it progressively with an image. The IC algorithm can be applied to a variety of model structures. However, in the work described here the model is a single square image patch that represents a feature to be recognised. It is constructed from the mean and Eigen vectors of a set of training images (Figure A.3). The algorithm treats the patch as fixed and adjusts the affine transform of the test image to minimise their sum of squared pixel differences. In addition, the parameters of the patch appearance are updated to bring its appearance closer to that of the transformed test image. The parameters represent the magnitude of the deviations from the mean of the image patch appearance. The transform is updated by multiplying the current transform with a refinement transform to improve its alignment. Both updates, the appearance and refinement transform, are calculated using the Gauss-Newton optimisation algorithm. This algorithm is structured as follows:

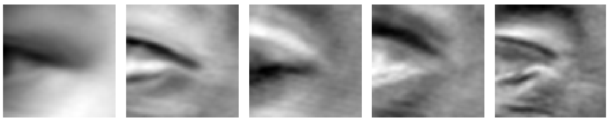


FIGURE A.3: Image of a mean and first four Eigen vectors of an eye corner model

Iterate:

1. Warp the test image I using the current warp $W(\mathbf{p})$ derived from the parameters

p

2. Construct the model image using the appearance parameters λ

3. Compute the error image **e** as the difference between the warp and the current model

4. Calculate the Jacobian **J** of the error image with respect to the appearance parameters and a composed warp (i.e. a warp of the current warped image)

$$\delta\mathbf{v}{=}[\delta\mathbf{p}^{\mathbf{T}},\delta\lambda^{\mathbf{T}}]^{\mathbf{T}}$$

$$\frac{\partial e_i}{\partial v_j}{=}\ \mathrm{J_{ij}}$$

5. Solve the local linear approximation to obtain the update parameters

$$\left(\mathbf{J}^{\mathbf{T}}\mathbf{J}\right)\delta\mathbf{v}{=}\ \ -\mathbf{J}^{\mathbf{T}}\mathbf{e}$$

6. Apply the update by composing the warp and adjusting the appearance parameters

$$\mathbf{W}(\mathbf{p})\overset{update}{\longleftarrow}\mathbf{W}\left(\delta\mathbf{p}\right)\mathbf{W}(\mathbf{p})$$

$$\lambda\overset{update}{\longleftarrow}\lambda{+}\ \delta\lambda$$

Repeat until convergence:

$$\|\delta\mathbf{v}\|\leq e$$

In its original formulation this error term is the pixel difference between the current model and the transformed image. This has been extended to include a model likelihood error term. This simultaneously minimises the appearance difference while maximising the probability of the patch parameters.

One of the strengths of the IC algorithm is that it is formulated such that the Jacobian matrix can be kept constant. This significantly improves performance over a general optimisation approach. However, this optimisation is not possible when the technique is adapted to ignore outliers. As the focus of this work is primarily on precision, this slower but more robust approach is used as part of the technique. Further evaluation may show that this robustness is not necessary as the feature points are small and so are less likely to be partially occluded.

A.2.2 Training Data

A critical factor in the accuracy of any detection system is the size and coverage of its training set. There are many databases of frontal faces that can be used for training face and face feature detectors. However, there are relatively few profile datasets suitable for ear detection training. In addition, most large datasets consist of subjects recorded in well lit environments with studio lighting or flash cameras. To address the shortage of data, some existing work has used synthesised images [52]. It is intended that a similar process be used in the proposed approach. In particular, by using the Morphable Model trained from the previous section a large space of potential ear shapes can be synthesised. In addition, by adjusting the pose and lighting of these ear shapes a large and comprehensive set of samples can be created. Only limited evaluations have been performed so far using a small collection of images from the FRGC dataset. In addition, the algorithm has been evaluated using faces rather than ears to enable the use of the pre-trained OpenCV face detector. This enables the algorithm to be evaluated without the significant time investment of training a Viola Jones detector for ears.

A.2.3 Robust Learning

Bootstrapping was used to compensate for the potentially inaccurate feature point labelling. A bootstrap update involves aligning the current model with all of the remaining training samples. The closest aligned sample is then used to improve the model. This process is repeated until all samples are used. To initialise this process the median sample is estimated. This sample was calculated by measuring the distance from each image to the rest of the training set. The median of these distances was used to estimate the sample's proximity to the other samples. Normalisation was applied to compensate for the potential variation in contrast and brightness between samples. This was achieved by adjusting the pixel intensities so that the initial model alignment region has a zero mean and unit variance. However, as patches are iteratively moved into place, this normalisation may become inaccurate. As the patch is moved the new region it covers may contain more bright or dark regions which would have resulted in a different appearance had they been included when the patch was initially normalised. To address this, training samples were included multiple times with differing brightness and contrast values. These values represent the normalisation that could have been applied had the patch been initially misaligned (Figure A.4). In addition to learning the variation in feature point appearance, the variation in feature point location was also calculated. This location space was approximated using a Gaussian model calculated from the aligned training samples. The resulting model defines ellipse regions relative to the detected object. To localise the associated features, these regions can be searched using the IC algorithm. Analysis on an additional training set revealed that the IC algorithm can correctly align features offset by as much as 4 pixels in both x and y directions. By

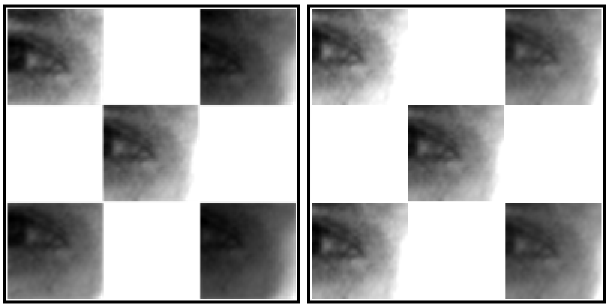


Figure A.4: The samples used to calculate the expected change in normalisation Left: Image of patch and 4 surrounding offset samples Right: resulting normalised samples used to create the model

sampling these ellipse regions at 4 pixel steps, the search space is reduced considerably, greatly improving detection efficiency.

A.2.4 Model Registration

Once feature points have been detected, a Morphable Model can be aligned with them. If the position and perspective of the camera is known, the 3D position of the head can be calculated using the detected face region. This can be calculated directly using the camera matrices (Figure A.5).

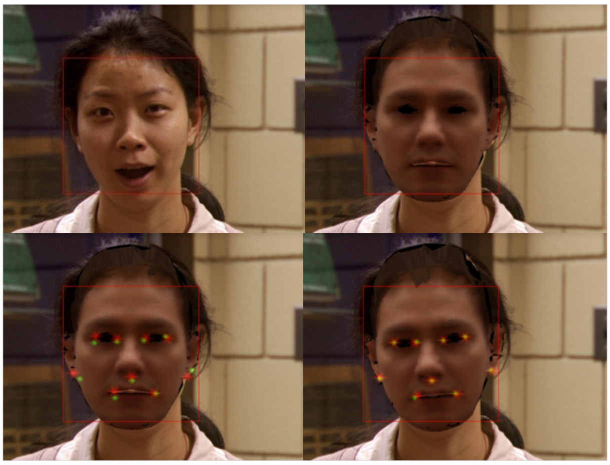


Figure A.5: Images showing alignment to feature points Top left: probe face with detected face region Top right: model initialised using detected region Bottom left: feature points on image (green) and model (red) Bottom right: aligned model after one step of registration

A.2.5 Robust Non-Linear Optimisation Fitting

The existing Morphable Model alignment algorithm, developed by Blanz, makes an orthonormal approximation to the camera to improve performance. However, with a good initial alignment, a more precise model fitting process can produce more accurate results. In addition, as the number of feature points detected is small , performance is not affected significantly. The current process uses the best estimate of the detected feature points for the model alignment. Further work is likely to expand this approach to search for the points which simultaneously achieve the best image match, the closest model to point alignment, and most likely model parameters. This is expected to produce an accurate estimate of the object properties. In addition, the current fitting process is made partially robust to occlusion by excluding the least accurately fitted point as an outlier.

The fitting process works in a similar fashion to the Inverse Compositional Alignment algorithm. In each iteration, the model's rigid transform and its principal component parameters are updated to minimise the alignment error. This error is calculated using the projected distance between the model's points and their corresponding detected image features. In addition, a prior parameter likelihood error term is included. In this way, the error function is minimised when the feature points are well aligned and the resulting object forms a probable shape.

Finally, it should be noted that this work assumes that the calibration parameters of the camera are known. It is possible to extend this fitting process to calculate the camera field of view as part of the fitting process but this is left as further work.

A.3 Evaluation

As the approach is incomplete, detailed evaluation is not possible. However, initial evaluations using the IC alignment algorithm have produced some promising results. This analysis is based on a very small training set of 5 subjects. Figure A.6 shows some examples of the features that have been detected by this process. In both cases the subject is not included in the training set.

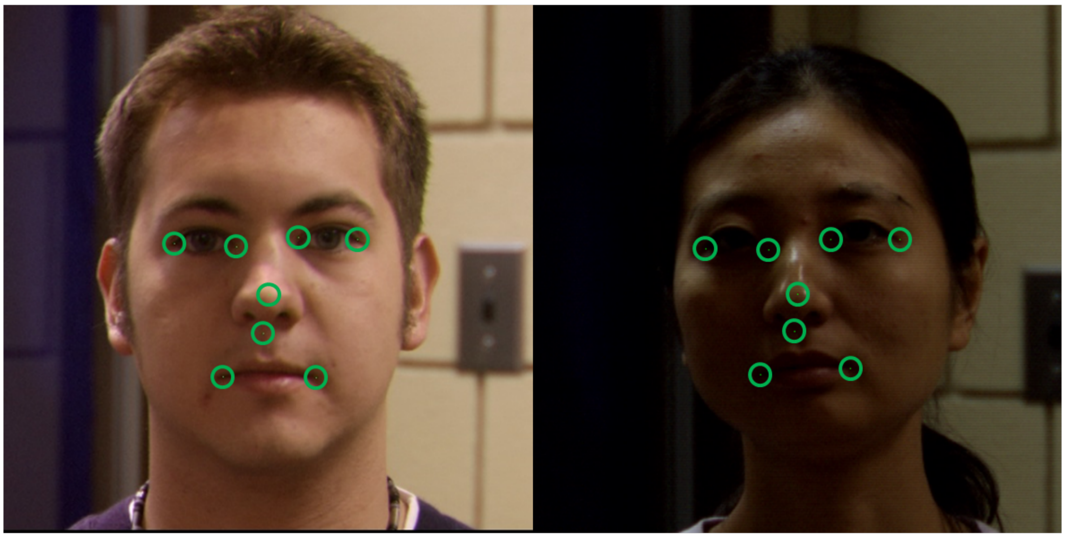


Figure A.6: Some randomly selected examples used to evaluate the new technique. Many points are very accurate although outliers are still present in both images

References

[1] A. F. Abate, M. Nappi, D. Riccio, and S. Ricciardi. Ear recognition by means of a rotation invariant descriptor. In *Proceedings of International Conference on Pattern Recognition*, volume 4, pages 437–440, 2006.

[2] M. Abdel-Mottaleb and J. Zhou. *Human Ear Recognition from Face Profile*, volume 3832, pages 786–792. Springer-Verlag, 2005.

[3] B. Allen, B. Curless, and Z. Popovic. The space of human body shapes: Reconstruction and parameterization from range scans. In *Proceedings of ACM SIGGRAPH*, volume 22, pages 587–594, 2003.

[4] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[5] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: Shape completion and animation of people. In *Proceedings of ACM SIGGRAPH*, volume 24, pages 408–416, 2005.

[6] S. Ansari and P. Gupta. Localization of ear using outer helix curve of the ear. In *Conference on Computing: Theory and Applications*, pages 688–692, 2007.

[7] B. Arbab-Zavar, M. S. Nixon, and J. N. Carter. *On shape-mediated enrolment in ear biometrics*, volume 4842, pages 549–558. Springer, 2007.

[8] B. Arbab-Zavar, M. S. Nixon, and Carter J. N. On model-based analysis of ear biometrics. In *Proceedings of Biometrics: Theory, Applications, and Systems*, pages 1–5, 2007.

[9] S. Arya, M. Mount, D., S. Netanyahu, N., R. Silverman, and Y. Wu, A. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[10] G. S. Badrinath and P. Gupta. Feature level fused ear biometric system. In *Advances in Pattern Recognition*, pages 197–200, feb 2009.

[11] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[12] B. A. Barsky and T. J. Kosloff. Algorithms for rendering depth of field effects in computer graphics. In *Proceedings of WSEAS international conference on Computers*, pages 999–1010. World Scientific and Engineering Academy and Society (WSEAS), 2008.

[13] C. Basso, P. Paysan, and T. Vetter. Registration of expressions data using a 3d morphable model. In *Proceedings of Automatic Face and Gesture Recognition*, pages 205–210, 2006.

[14] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, 28(3):245–260, 1998.

[15] A. Bertillon. *La photographie judiciaire, avec un appendice sur la classification et l'identification anthropometriques.* Gauthier-Villars, 1890.

[16] V. Blanz, A. Mehl, T. Vetter, and H. Seidel. A statistical method for robust 3d surface reconstruction from sparse data. In *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 293–300, 2004.

[17] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of ACM SIGGRAPH*, volume 25, pages 187–194, 1999.

[18] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[19] W. Brand. Morphable 3d models from video. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, volume 2, pages 456–463, 2001.

[20] M. Brown and G. Lowe, D. Invariant features from interest point groups. In *Proceedings of the 13th British Machine Vision Conference*, pages 253–262, 2002.

[21] R. Brunelli and T. Poggio. Template matching: Matched spatial filters and beyond. *International Journal on Pattern Recognition*, 30(5):751–768, 1997.

[22] M. Burge and W. Burger. *Ear Biometrics*, pages 271–286. Kluwer Academic Publishers, 1998.

[23] J. D. Bustard and M. S. Nixon. Robust 2d ear registration and recognition based on sift point matching. In *Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, 2008.

[24] J. D. Bustard and M. S. Nixon. 3d morphable model construction for robust ear and face recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 2582 – 2589, June 2010.

[25] J. D. Bustard and M. S. Nixon. Toward unconstrained ear recognition from two-dimensional images. *IEEE Transactions on Systems, Man and Cybernetics (A)*, 40(3):486–494, 2010.

[26] S. Cadavid and M. Abdel-Mottaleb. Human identification based on 3d ear models. In *Proceedings of Biometrics: Theory, Applications, and Systems*, pages 1–6, 2007.

[27] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 679–714, 1986.

[28] A. Caunce, D. Cristinacce, C. Taylor, and T. Cootes. Locating facial features and pose estimation using a 3d shape model. In *Proceedings of the 5th International Symposium on Advances in Visual Computing: Part I*, ISVC '09, pages 750–761. Springer-Verlag, 2009.

[29] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1160–1165, 2003.

[30] H. Chen and B. Bhanu. Human ear recognition in 3d. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 29(4):718–737, 2007.

[31] M Choras. Ear biometrics based on geometrical feature extraction. *Electronic Letters on Computer Vision and Image Analysis*, pages 84–95, 2005.

[32] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models their training and application. *Computer Vision and Image Understanding*, 61:38–59, January 1995.

[33] T. F. Cootes, K. Walker, and C. J. Taylor. View-based active appearance models. In *Proceedings of Automatic Face and Gesture Recognition*, pages 227–232, 2000.

[34] C. Cortes and V. Vapnik. Support-vector networks. *Journal of Machine Learning*, 20(3):273–297, 1995.

[35] J. Daugman. United states patent 5291560: Biometric personal identification system based on iris analysis., March 1994.

[36] J. Daugman. Demodulation by complex-valued wavelets for stochastic pattern recognition. *Journal of Wavelets, Multi-resolution and Information Processing*, 1(1):1–17, 2003.

[37] P. David, D. DeMenthon, R. Duraiswami, and H. Samet. Softposit: Simultaneous pose and correspondence determination. *International Journal of Computer Vision*, 49(3):259–284, 2004.

[38] F. De la Torre and M. J Black. Robust parameterized component analysis: Theory and applications to 2d facial appearance models. *The Journal of Computer Vision and Image Understanding*, 91(1):53–71, 2003.

[39] M. De Marsico, N. Michele, and D. Riccio. Hero: Human ear recognition against occlusions. In *Conference on Computer Vision and Pattern Recognition*, pages 178 –183, 2010.

[40] P. Debevec, T. Hawkins, C. Tchou, H. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *Proceedings of SIGGRAPH*, volume 27, pages 145–156, 2000.

[41] Daniel F. Dementhon and Larry S. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15:123–141, 1995.

[42] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Conference on Computer Vision*, volume 1, pages 634–639, 2003.

[43] P. Dubuisson, M. and K. Jain, A. A modified hausdorff distance for object matching. In *Conference on Pattern Recognition*, volume 1, pages 566–568, 1994.

[44] N. Dyn, D. Levine, and A. Gregory, J. A butterfly subdivision scheme for surface interpolation with tension control. *ACM Transactions on Graphics*, 9(2):160–169, 1990.

[45] M. Everingham, J. Sivic, and A. Zisserman. "hello! my name is... buffy" - automatic naming of characters in tv video. In *17th British Machine Vision Conference*, pages 889–908, 2006.

[46] G. Finlayson, S. Hordley, G. Schaefer, and G. Yui Tian. Illuminant and device invariant colour using histogram equalisation. *Pattern Recognition*, 38(2), 2005.

[47] A. Fischler, M. and C. Bolles, R. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395, 1981.

[48] L. Flom and S. Aran. Iris recognition system, February 3 1987.

[49] Dedeoglu G., Baker S., and Kanade T. Resolution-aware fitting of active appearance models to low-resolution images. In *European Conference on Computer Vision*, volume 3952, pages 83–97, 2006.

[50] Y Guo, G. Zhao, J. Chen, M. Pietikäinen, and Z. Xu. A new gabor phase difference pattern for face and ear recognition. In *Conference on Computer Analysis of Images and Patterns*, pages 41–49, Berlin, Heidelberg, 2009. Springer-Verlag.

[51] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision.* Cambridge University Press, Cambridge, U.K, 2000.

[52] B. Heisele, T. Serre, and T. Poggio. A component-based framework for face detection and identification. *International Journal of Computer Vision*, 74(2):167–181, 2007.

[53] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Surface reconstruction from unorganized points. In *SIGGRAPH '92: Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, pages 71–78, New York, NY, USA, 1992. ACM.

[54] J. Huang, B. Heisele, and V. Blanz. Component-based face recognition with 3d morphable models. *IEEE Workshop on Face processing in Video*, 2004.

[55] P. Huber. *Robust Statistics.* Wiley, 2008.

[56] D. J. Hurley, M. S. Nixon, and J. N. Carter. Force field feature extraction for ear biometrics. In *Computer Vision and Image Understanding*, volume 98, pages 491–512, 2005.

[57] A. Iannarelli. *Ear Identification.* Paramount Publishing Company, 1989.

[58] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer Vision, Graphics, and Image Processing*, 44(1):87–116, 1988.

[59] S. M. Islam, R. Davies, A. S. Mian, and M. Bennamoun. A fast and fully automatic ear recognition approach based on 3d local surface features. In *ACIVS '08: Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 1081–1092, Berlin, Heidelberg, 2008. Springer-Verlag.

[60] S. M. S. Islam, M. Bennamoun, A. S. Mian, and R. Davies. A fully automatic approach for human recognition from profile images using 2d and 3d ear data. In *Proceedings of 3DPVT'08 the Fourth International Symposium on 3D Data Processing, Visualization and Transmission*, 2008.

[61] E. Jeges and L. Mt. Model-based human ear localization and feature extraction. *International Journal of Intelligent Computing in Medical Sciences and Image Processing*, 1:101–112, 2007.

[62] A. Johnson. *Spin-Images: A Representation for 3-D Surface Matching.* PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, August 1997.

[63] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *European Conference on Computer Vision*, volume 3021, pages 228–241, 2004.

[64] T. Kim, K. Lee, S. Lee, and C. Yim. Occlusion invariant face recognition using two-dimensional pca. In *Advances in Computer Graphics and Computer Vision*, volume 4 of *Communications in Computer and Information Science*, pages 305–315. Springer Berlin Heidelberg, 2007.

[65] S. King, G. Yun Tian, D. Taylor, and S. Ward. Cross-channel histogram equalisation for colour face recognition. In *Audio- and Video-Based Biometric Person Authentication*, volume 2688, pages 1055–1056, 2003.

[66] S. Lee, Y. Liu, and R. Collins. Shape variation-based frieze pattern for robust gait recognition. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1–8, 2007.

[67] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal of Computer Vision*, 81(2):155–166, 2009.

[68] D. G. Lowe. Object recognition from local scale-invariant features. In *Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[69] L. Lu, X. Zhang, Y. Zhao, and Y. Jia. Ear recognition based on statistical shape model. In *Proceedings of the First International Conference on Innovative Computing, Information and Control*, volume 3, pages 353–356, 2006.

[70] R. Matey, J., O. Naroditsky, K. Hanna, R. Kolczynski, D. J. LoIacono, S. Mangru, M. Tinker, and M. Zappia, T. Iris on the move: Acquisition of images for iris recognition in less constrained environments. *Proceedings of the IEEE*, 94(11):1936–1947, 2006.

[71] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[72] G. Medioni, J. Choi, C. Kuo, A. Choudhury, L. Zhang, and D. Fidaleo. Non-cooperative persons identification at a distance with 3d face modeling. In *Conference on Biometrics Theory, Applications and Systems*, pages 1–6, 2007.

[73] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[74] K. Mikolajczyk and C Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63 – 86, 2004.

[75] K Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:16151630, 2005.

[76] B. Moreno and A. Sanchez. On the use of outer ear images for personal identification in security applications. In *Conference on Security Technology*, pages 496–476, 1999.

[77] L. Nanni and A. Lumini. A multi-matcher for ear authentication. *Pattern Recogn. Lett.*, 28(16):2219–2226, 2007.

[78] L. Nanni and A. Lumini. Fusion of color spaces for ear authentication. *Pattern Recogn.*, 42(9):1906–1913, 2009.

[79] I. Naseem, R. Togneri, and M. Bennamoun. Sparse representation for ear biometrics. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Paolo Remagnino, Fatih Porikli, Jrg Peters, James Klosowski, Laura Arns, Yu Chun, Theresa-Marie Rhyne, and Laura Monroe, editors, *Advances in Visual Computing*, volume 5359 of *Lecture Notes in Computer Science*, pages 336–345. Springer Berlin / Heidelberg, 2008.

[80] H. Oh, K. Lee, S. Lee, and C. Yim. Occlusion invariant face recognition using selective lnmf basis images. In *Proceeding of the Asian Conference on Computer Vision*, volume 3852, pages 120–129, 2006.

[81] J. W. Osterburgh. *Crime Laboratory*. Paramount Publishing Company, 1968.

[82] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[83] G. Papandreou and P. Maragos. Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[84] S. Paris and F. Durand. A fast approximation of the bilateral filter using a signal processing approach. In *European Conference on Computer Vision*, volume 81, pages 568–580, 2006.

[85] J. Park and S. Lee. Stepwise reconstruction of high-resolution facial image based on interpolated morphable face model. In *Conference on Audio and Video-Based Biometric Person Authentication*, volume 3546, pages 102–111, 2005.

[86] A. Patel and W.A.P. Smith. 3d morphable face models revisited. *Conference on Computer Vision and Pattern Recognition*, 0:1327–1334, 2009.

[87] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Conference on Advanced Video and Signal based Surveillance for Security, Safety and Monitoring in Smart Environments*, pages 296–301, 2009.

[88] A. P. Pentland. A new sense for depth of field. *Transactions on Pattern Analysis and Machine Intelligence*, 9(4):523–531, 1987.

[89] P. J. Philips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone. Technical report nistir 6965, national institute of standards and technology, frvt 2002: Evaluation report, March 2003.

[90] P. J. Philips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. Technical report nistir 7408, national institute of standards and technology, frvt 2006 and ice 2006 large-scale results, March 2007.

[91] J. P. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Conference on Computer Vision and Pattern Recognition*, pages 947–954, 2005.

[92] Z. Rahman, D. J. Jobson, and G. A. Woodell. Multi-scale retinex for color image enhancement. In *International Conference on Image Processing*, volume 3, pages 1003–1006, 1996.

[93] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 986–993, 2005.

[94] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *conference on 3-D Digital Imaging and Modeling*, pages 145–152, 2001.

[95] S. Samangooei, J. D. Bustard, R. D. Seely, M. S. Nixon, and J. N. Carter. On acquisition and analysis of a dataset comprising of gait, ear and semantic data. In *Multibiometrics for Human Identification*, 2011.

[96] F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Workshop on Applications of Computer Vision*, pages 138 –142, 1994.

[97] S. Sarkar, P. J. Philips, Z. Liu, I. R. Vega, P. Grother, and K. Bowyer. The humanid gait challenge problem:data sets, performance and analysis. *Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[98] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Conference on Automatic Face and Gesture Recognition*, pages 53–58, 2002.

[99] W. A. P. Smith and E. R. Hancock. Recovering facial shape using a statistical model of surface normal direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1914–1930, 2006.

[100] T. Theoharis, G. Passalis, G. Toderici, and I. A. Kakadiaris. Unified 3d face and ear recognition using wavelets on geometry images. In *Conference on Pattern Recognition*, volume 41, pages 796–804, 2008.

[101] D. M. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, 1985.

[102] M. A. Turk and A. Pentland. Face recognition using eigenfaces. In *Proceedings of Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[103] B. Victor, K. Bowyer, and S. Sarkar. An evaluation of face and ear biometrics. In *Conference on Pattern Recognition*, volume 1, pages 429–432, 2002.

[104] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[105] M. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. In *Conference on Computer Graphics and Interactive Techniques*, volume 24, pages 426–433, 2005.

[106] Y. Wang, Z. Mu, and H. Zeng. Block-based and multi-resolution methods for ear recognition using wavelet transform and uniform local binary patterns. In *Conference on Pattern Recognition*, pages 1–4, 2008.

[107] Y. Wang, G. Pan, and Z. Wu. 3d face recognition in the presence of expression: A guidance-based constraint deformation approach. In *Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007.

[108] F. W. Wheeler, L. Xiaoming, and P. H. Tu. Multi-frame super-resolution for face recognition. In *Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, 2007.

[109] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[110] W. Xiaoyun and Y. Weiqi. Human ear recognition based on block segmentation. In *Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 262–266, 2009.

[111] Z. Xie and Z. Mu. Ear recognition with variant poses using locally linear embedding. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, volume 5226, pages 460–465. Springer Berlin / Heidelberg, 2008.

[112] P. Yan and W. Bowyer, K. Biometric recognition using three-dimensional ear shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1297–1308, 2007.

[113] L. Yuan, Z. Mu, Y. Zhang, and K. Liu. Ear recognition using improved non-negative matrix factorization. In *International Conference on Pattern Recognition*, pages 501–504, 2006.

[114] L. Yuan, J. Sun, L. Quan, and H. Shum. Image deblurring with blurred/noisy image pairs. *Transactions on Graphics*, 26(3):1–9, 2007.

[115] T. Yuizono, Y. Wang, K. Satoh, and S. Nakayama. Study on individual recognition for ear images by using genetic local search. In *Congress on Evolutionary Computation*, volume 1, pages 237–242, 2002.

[116] L. Zhang, S. Wang, and D. Samaras. Face synthesis and recognition under arbitrary unknown lighting using a spherical harmonic basis morphable model. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 209–216, 2005.

[117] H. Zhao, Z. Mu, X. Zhang, and W. Dun. Ear recognition based on wavelet transform and discriminative common vectors. In *Conference on Intelligent System and Knowledge Engineering*, volume 1, pages 713–716, 2008.

[118] K. Zhou, S., R. Chellappa, and W. Zhao. *Unconstrained Face Recognition.* Springer, 2006.