BIOLOGICALLY MOTIVATED 3D FACE RECOGNITION

by

Albert Ali Salah

B.S, in Computer Engineering, Boğaziçi University, 1998

M.S, in Computer Engineering, Boğaziçi University, 2000

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in
Boğaziçi University
2007

BIOLOGICALLY MOTIVATED 3D FACE RECOGNITION

APPROVED BY:

Prof. Lale Akarun . . . .
(Thesis Supervisor)

Prof. H. Levent Akın . . . .

Prof. Fikret Gürgen . . . .

Prof. Bülent Sankur . . . .

Assoc. Prof. Berrin Yanıkoğlu . . . .

DATE OF APPROVAL: 24.1.2007

This dissertation is dedicated to the memory of my grandfather, Dr. Hayim Molinas.

ACKNOWLEDGEMENTS

With gratitude to my PhD advisor Lale Akarun for her boundless support, her affection, her guidance and for making my work towards the dissertation a joyful experience.

With gratitude to Bülent Sankur for his efforts in maintaining the Boğaziçi University Multimedia Group (BUMM), for his unwavering principles of scientific conduct, and for sharing his vast experience.

With gratitude to Ethem Alpaydın, who supervised the development of the material presented in Chapter 4, and to my colleagues in the Machine Learning Group for fruitful discussions.

With gratitude to all my teachers: Dean Adams, Levent Akın, Kuban Altınel, Taner Bilgiç, Haluk Bingöl, Mine Nakipoğlu-Demiralp, Yağmur Denizhan, Jeff Elman, Susan Epstein, Greame Halford, Robert Goldstone, Tunga Güngör, Fikret Gürgen, Sibel Irzık, Ali Karatay, Josef Kittler, Ayhan Koç, Boicho Kokinov, Alice O'Toole, Sumru Özsoy, Csaba Pleh, James Rohlf, Cem Say, Dennis Slice, Ali Vahit Şahiner, Neslihan Şengör, Ali Tekcan, Michael Thomas, Oğuz Tosun, Bülent Yener.

With gratitude to Cem Ersoy and Tuna Tuğcu, for their great contributions to our department.

With gratitude to people and institutions who supported me: Matthew Casey, Aytül Erçil, Robert Glushko, Didem Gökçay, Aykut Kence, Boicho Kokinov, Richard Reilly, Bilge Say, Max Snijder, Oğuz Tanrıdağ, Massimo Tistarelli, TÜBİTAK project 104E080, DPT project 03K120250, BÜVAK, European Biometrics Forum, and European FP6 NoEs Biosecure and SIMILAR.

With gratitude to Patrick Flynn, Kevin Bowyer, Jonathon Phillips and the University of Notre Dame for their efforts in gathering the FRGC face database, and kindly allowing us to use it.

With gratitude to all my friends and colleagues: Hatice Çınar Akakın, Oya Aran, Koray Balcı, Manuele Bicego, İlker Demirkol, Çağlayan Dicle, Onur Dikmen, Cüneyt Argun Genç, Ulaş Başar Gezgin, Çiğdem Gündüz, Mirek Hamouz, Amaç Herdağdelen, Cristophe Labiouse, Okan İrfanoğlu, Itır Karaç, Cem Keskin, Rabun Koşar, Elif Kuş, Atay Özgövde, Rafael Tena, Burak Turhan, Aydın Ulaş, Olcay Taner Yıldız, and especially to Neşe Alyüz for her diligence, and to Berk Gökberk for bootstrapping my dissertation.

With gratitude to Fatih Köksal and Kerem Özkısacık, who were with me from the start.

With gratitude to my family for everything, and to my wife, who bore the brunt of it with patience, and supported me to the end.

ABSTRACT

BIOLOGICALLY MOTIVATED 3D FACE RECOGNITION

Face recognition has been an active area of study for both computer vision and image processing communities, not only for biometrics but also for human-computer interaction applications. The purpose of the present work is to evaluate the existing 3D face recognition techniques and seek biologically motivated methods to improve them. We especially look at findings in psychophysics and cognitive science for insights. We propose a biologically motivated computational model, and focus on the earlier stages of the model, whose performance is critical for the later stages. Our emphasis is on automatic localization of facial features. We first propose a strong unsupervised learning algorithm for flexible and automatic training of Gaussian mixture models and use it in a novel feature-based algorithm for facial fiducial point localization. We also propose a novel structural correction algorithm to evaluate the quality of landmarking and to localize fiducial points under adverse conditions. We test the effects of automatic landmarking under rigid and non-rigid registration methods. For the rigid registration approach, we implement the iterative closest point method (ICP). The most important drawback of ICP is the computational cost of registering a test scan to each scan in the gallery. By using an average face model in rigid registration, we show that the computation bottleneck can be eliminated. Following psychophysical arguments on the "other race effect", we reason that organizing faces into different gender and morphological groups will help us in designing more discriminative classifiers. We test this claim by employing different average face models for dense registration. We propose a shape-based clustering approach that assigns faces into groups with nondescript gender and race. Finally, we propose a regular re-sampling step that increases the speed and the accuracy significantly. These components make up a full 3D face recognition system.

ÖZET

BİYOLOJİK TABANLI ÜÇ BOYUTLU YÜZ TANIMA

Yüz tanıma biyometri araştırmalarına konu olmanın yanı sıra insan-bilgisayar iletişimi bağlamında da çok araştırılmış, üzerinde çok çalışılmış bir problemdir. Bu çalışmada amacımız üç boyutlu yüz tanıma tekniklerini değerlendirmek ve biyolojik tabanlı modeller yoluyla geliştirmektir. Bu amaçla öncelikle insanlarda yüz tanımanın nasıl olduğuna baktık. Varolan 3B yüz tanıma sistemlerini değerlendirdikten sonra, bilişsel bilim bulguları ışığında bir yüz tanıma modeli önerdik. Modelin ilk kısmı olan otomatik kayıtlama, ve kayıtlama için elzem saydığımız otomatik nirengi noktası bulma problemlerine yoğunlaştık. Öncelikle özniteliklerin öğrenmesini kolaylaştıracak güçlü bir gözetimsiz öğrenme algoritması geliştirdik. Bu algoritma bize faktör analizi yaklaşımıyla esnek veri modellemesi sağladı. Ardından yüzlerde tanımladığımız nirengi noktalarını otomatik olarak bulmak için bu yaklaşımı kullandık. Sonra, bulunan nirengi noktalarını yeni bir yapısal düzeltme algoritmasıyla düzelttik. Bu algoritmayla eksik ve hatalı imgelerde bile kayıtlama yapmak mümkün oldu. Otomatik nirengi noktası bulma metodumuzun başarısını deformasyonlu ve deformasyonsuz kayıtlama metodlarıyla ölçtük, deformasyonsuz kayıtlamada daha yüksek başarı elde ettik. Literatürde sıkça kullanılan “döngülü en yakın nokta” algoritmasının en büyük sorunu kayıtlamanın yüksek maliyetli olmasıdır. Bunu aşmak için ortalama yüz modeli kullanarak kayıtlama önerdik. Ayrıca, “diğer ırk efekti” üzerine yapılan araştırmalardan yola çıkarak, yüzlerin gruplanarak paralel sınıflandırıcılarla değerlendirilmesinin başarımı artırabileceği varsayımını denedik. Bir diğer yaklaşımda da şekil uzayında topaklama yaparak değişik gruplar elde ettik. Ortalama yüz modeli kullandığımız için derinlik değerlerinin eşit aralıklı örnekleme ile düzenlenmesi de mümkün oldu. Bu yöntem hem hız, hem de başarımı artırdı. Sonuç olarak tamamen otomatik ve başarılı bir 3B yüz tanıma sistemi elde ettik.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

a	Magnitude of Gabor wavelet
B	Basis images
$\mathcal{C}(X,Y)$	For each point in X, closest points on Y
d	Dimensionality of data
D	Diagonal matrix of eigenvalues
g_j	Prior probability for component j
g_{lj}	Given the class l, prior probability for component j of that class
$\mathcal{G}_j$	Parameter set of the model j
h	Gradient kernel
h_{kj}	Posterior probability of data item k belonging to component j
H	A matrix composed of eigenvectors as columns
$H(x,y,z)$	Spherical harmonics
I	Identity matrix
$I(n,p)$	Surface radiosity
j	Index for components
J	Number of components
J_l	Number of components for class l
k	Index for data items
$\boldsymbol{k}$	Gabor kernel
l	Index for classes
l_j	Likelihood under model j
$\mathcal{L}$	Log-likelihood
M	Number of classes
N	Number of samples
p	Dimensionality of data subspace or factors
P	A shape matrix, made up of landmarks
$P_{i,k}$	k^{th} landmark vector of shape P_i

$\vec{q}$	Registration state vector
$\vec{q_H}$	Quaternion
$\vec{q_T}$	Translation vector
S	Sample covariance
t	Intensity image
v	Eigenvector
$\boldsymbol{x}^t$	A data sample
Y	Consensus shape
$\boldsymbol{z}$	Factors in factor analysis

α	Scaling coefficient
β	Multivariate kurtosis
$\boldsymbol{\epsilon}_{kj}$	Residual term in the residual factor addition
γ	Translation vector
$\hat{\gamma}$	Geometric mean likelihood
Γ	Rotation matrix
λ	Surface albedo
$\boldsymbol{\Lambda}$	Factor loading matrix
$\boldsymbol{\mu}$	Mean vector
μ_x	Centre of mass of point cloud X
π_j	Prior probability of component j
ϕ	Gabor wavelet phase parameter
Ψ_j	Diagonal error variance matrix for component j
σ	Standard deviation
$\boldsymbol{\Sigma}$	Covariance matrix
τ	Likelihood threshold in GOLLUM
θ	A set of parameters
χ	A data set
∇^4	Biharmonic operator

AFM	Average Face Model

DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
EER	Equal Error Rate
EM	Expectation-Maximization
ERP	Event Related Potentials
FA	Factor Analysis
FFA	Fusiform Face Area
FJ	Figueiredo-Jain
FRGC	Face Recognition Grand Challenge
GOLLUM	Gaussian Outlier Localization with Likelihood Margins
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICA	Independent Component Analysis
ICP	Iterative Closest Point
IMoFA	Incremental Mixtures of Factor Analyzers
LDA	Linear Discriminant Analysis
LM	Levenberg-Marquardt
MoFA	Mixtures of Factor Analyzers
MoG	Mixtures of Gaussians
MoMoG	Mixtures of Mixtures of Gaussians
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
PSD	Point Set Difference
SGML	Soft Geometric Mean Likelihood
SVM	Support Vector Machines
TPS	Thin-plate Spline

1. INTRODUCTION

The face is considered to be a universal signifier for a person's unique identity. Consequently, letting computers recognize a person from his or her face has been the most important goal in human-computer interaction research. The purpose of this work is to advance the state-of-the-art for computer recognition of faces, and to propose novel methods of information processing that make use of new technologies and new scientific findings. Our theoretical contributions include a novel and powerful unsupervised learning algorithm for feature extraction that does model selection automatically, and a structural shape analysis algorithm that uses statistical properties of shapes to evaluate and correct shapes as represented by landmarks. We propose a novel biologically motivated computational framework for 2D-3D face recognition, and novel methods for automatic facial landmarking and dense registration. With our new algorithms, we obtain results that surpass the current state-of-the-art.

Humans learn faces after a very brief exposure to them. Then, these faces are recognized under different poses, with different illumination conditions, and through changing expressions. These three conditions are the traditional hurdles faced by researchers of face recognition. The statistical differences in face images due to pose, illumination or expression are often greater than differences caused by a change of identity. Researchers in computer science and signal processing have been working on the face recognition problem for a long time, but the problem has retained its difficulty. Several commercial systems were produced with recognition rates acceptable for limited environments, yet the deployment of biometrics systems of a greater scale has not been possible.

In our attempt to find ways to increase the accuracy of computers in face recognition, we turn to humans. How is it possible that the human brain is so successful in the face of these difficulties? The brain is a very powerful learning system, which can be trained to discriminate between highly similar entities. Such a training results in the acquisition of so-called *expertise*, i.e. better and faster recognition in this context.

But is face recognition an acquired expertise of a general learning system, or do we have innately programmed mechanisms specific for faces? Do we perceive the face as a whole, or do we process facial features separately? What are the transformations learned during processing of faces?

Cognitive science seeks answers to these questions, because its aim is to understand how the brain works. Researchers in face recognition in computers have looked at cognitive science from time to time, with the hope that some mechanism of the brain will be uncovered that would lend itself to be modeled in a computer algorithm, eventually leading to better face recognition systems. Abstraction is a powerful tool, and biological inspirations have often served the designers of computer algorithms in different levels.

It is true that the complexity and power of a computer is incomparably small next to the primate brain, but sometimes technology makes leaps that go beyond human capacities. When infrared and 3D sensors were developed, new avenues were opened for face recognition research. The human visual system recovers 3D information from the scene by taking simultaneous input from two eyes, and by solving a correspondence problem. In many ways, this resembles a system with two calibrated cameras. But if the aim is to recover the 3D shape of the scene, a computer system is better served by an accurate laser scanner. The massively parallel organization of the brain involves processing methods that will be impossible to duplicate in our essentially serial computers, even if we can figure out their exact workings. Thus, it is useful to understand the principles of human face recognition, but not always meaningful to duplicate them exactly in computers. When the technology allows, it makes sense to use means that are not available to the human brain. 3D is very promising for face biometrics, as it solves some of the more difficult problems, and finally makes a large-scale system possible.

In its simplest setting, face recognition involves comparing a test face to a gallery of faces, to determine the closest match. Using 3D information, as opposed to 2D information, eliminates the variability caused by illumination conditions. There are several technologies to recover the 3D surface from a face. Then the problem is to compare the facial surface with the 3D faces stored in the gallery, and to determine the best match. The pose and expression variations are not automatically dealt with in 3D face recognition: We need to *register* the faces by putting all their parts (e.g. eyes, chin, nose) into correspondence. The speed and accuracy with which we perform this operation are very important to the overall system. Specifically, there is a need to locate several anchor points in the face image to guide the registration, called the *landmarking*. The main contributions of the present work are novel methods for the landmarking and registration tasks. The method we propose for landmarking is based on statistical modeling of facial features. We have developed a novel unsupervised feature learning algorithm, and used it successfully for anchor point localization.

The unsupervised model we have proposed (IMoFA, short for incremental mixtures of factor analyzers) is essentially a mixture of Gaussians method, based on the factor analysis approach. We assume that the data distribution in a high-dimensional space has an intrinsic lower dimensionality. The data are then modeled in this lower dimensional manifold of *factors*, plus an isotropic sensor noise model in the high-dimensional data space. The algorithm is initialized with a one-dimensional manifold, and a single component. The complexity of the mixture is gradually increased by splitting components and adding factors, until it is not justified by a likelihood increase in a separate validation set. Thus, the algorithm automatically tunes model complexity to the data at hand. Model selection is a very important issue in unsupervised learning, and our own results and independent results obtained by other researchers confirm our findings that the IMoFA method is very successful.

Localization of anchor points (or *landmarks*) solely by using local feature information is tricky, no matter how powerful the feature learning model is. Sometimes the information is simply not there, but the anchor point is recoverable via structural information. Some models incorporate structural and local information in a single expression, and optimize it jointly. This approach is plagued by local minima. We have chosen to integrate the structural information at a later stage, and proposed a novel structural correction algorithm that evaluates and corrects landmark locations on similar objects. Our model (GOLLUM) is based on the assumption that landmarking errors are independent, and the structural correction is necessary to correct a few wrong landmarks each time. Under this assumption, GOLLUM tries to find a subset of landmarks that confirms with the expected configuration, and uses those to correct the rest. We have compared our algorithm to a very recent alternative, and demonstrated its superior performance in our application.

The geometric properties of landmarks are in themselves not sufficient for recognition purposes. Their main purpose is to guide registration. The most popular method of 3D face registration is the iterative closest point (ICP) algorithm, which also happens to be a very costly algorithm in terms of computational requirements [1]. In the traditional ICP-based approaches, a test scan needs to be iteratively registered in many steps to each and every gallery face prior to similarity calculation. We have proposed to employ an average face model (AFM), which requires the system to perform ICP only once for each test scan. Our novel AFM generation method aims at facilitating the subsequent classification. A further regular re-sampling step both speeds up the computation of similarity and the accuracy of the resulting system significantly.

Inspired by behavioural studies on the *other race effect*, we have proposed an approach that uses one AFM for each facial category. The classifiers that are tuned for within-category differences stand to benefit from such a two-tiered architecture. To determine the facial categories, we contrast a gender- and race-based discrimination with one that is based on shape-space clustering. We show that the unsupervised clustering indeed creates groups with race and gender differences, and achieves good results in an ethically sound, gender- and race-blind manner.

The thesis is organized as follows: In Chapter 2 we review the work on human face recognition, reporting the neurological and behavioural findings that might give us clues and pointers for guidance. Chapter 3 is a survey of the state-of-the-art in 3D face recognition. 3D information can be represented in different ways, each leading to systems with different characteristics. We organize Chapter 3 in a way to emphasize this dependence on representation. In Section 3.6, a biologically inspired model for

2D-3D face recognition is proposed. This model serves as a schematic guide for the rest of the dissertation.

In Chapter 4, we develop the IMoFA algorithm step by step, and demonstrate its effectiveness in different pattern recognition problems. This algorithm is then used in Chapter 5 for modeling and locating facial landmarks. Following recent developments in 3D-assisted 2D illumination correction, we implement a method to recover the facial albedo image in Section 5.5.3. We then compare 2D and 3D features with albedo features for automatic landmark localization. A novel structural correction algorithm is proposed for detecting and correcting the landmarking errors in Section 5.6. We report our results on the FRGC 2D+3D dataset.

In Chapter 6, we inspect two different registration paradigms, and assess the effect of landmarking on rigid and non-rigid registration. We propose improvements to the popular iterative closest point algorithm for dense registration, and describe a novel AFM construction method. Although each chapter has its own concluding section, we summarize our results in Chapter 7. A small, selective glossary on 3D face recognition is added as Appendix A.

2. HUMAN FACE RECOGNITION

Beginnings can be the narrative limits of the knowable, the margins of the meaningful.

Homi Bhabha

Most animals and humans have great perceptual capabilities that result from complex visual systems, which in turn are optimized by evolutionary processes. Indeed, such is the daunting complexity and integrity of the human eye, with a great amount of neural matter to support its input, that it was long held as an evidence against the Darwinian evolution: such a mechanism could not result from mere random drift, but it must have been designed by an architect beyond human capacities. Yet half a century of work in computer science and related fields, combined with physiological and neurological findings accumulated over more than a century had made its aim to mimick the capabilities induced by these systems in man-made hardware.

We have several aims in investigating the psychological and neurological findings about human face recognition. There are certain advantages to doing things the way humans do: For instance, humans learn faces one by one, whereas most of the computer systems learn discriminatively (as this is a simpler problem), and require samples from all different faces to be present. In an application area where the system will admit new users all the time (e.g. [2]), a biologically inspired incremental learning scheme might be preferable. A computer system that mimicks the human brain in its various computational aspects can reduce the space and time complexity of the resulting face recognition algorithm [3, 4]. It is only natural to turn towards the best working face recognition system for inspiration and guidance, and there are a number of biologically motivated approaches to face recognition.

Here is a short list of inter-related questions that we can ask and hope to find answers to what would ultimately guide us to implement better computer models for face recognition:

- What are the current theories of human object recognition in general and face recognition in particular?
- How much of face recognition is innate, and how much is learned? How should a machine learning system be biased to make the best use of the statistical information in the faces?
- How is the problem decomposed by the brain, and what constraints enforce such a decomposition? Should we mimick this decomposition for maximum efficiency? If we should, are the resources available to the computers up to the task?
- What is the evidence for feature based and holistic components of face recognition?
- Which features are employed by the humans, and to what degree? How are they combined? How does the system deal with invariances?
- How is the depth information employed? Can we use a similar mechanism to enhance 3D recognition?
- What can we learn from biologically motivated face recognition systems? Which aspects of the brain did they use, and what advantages do they offer?

In this chapter, we review the cognitive aspects of human face recognition and look for answers to these questions[1] .

2.1. Theories of human face recognition

Human face recognition is a widely researched area. One reason for this is that both low-level neurological studies and high-level behavioural studies point out to faces as having special status among other object recognition tasks. It has been long known that some cells in a monkey's brain fire only when the animal sees a face [6]. There are subjects with brain damage that have relatively intact object recognition, but poor face recognition (*prosopagnosia*) [7] but also subjects with intact face recognition and impaired general object recognition capacities (*agnosia*) [8]. Due to brain lesions, the capacity to learn new faces can be selectively impaired as well (*prosopamnesia*) [9]. A great number of neurological studies found face specific activity in the brain (e.g. [10,

[1] Some of the material presented in this chapter appeared in [5].

11, 12, 13, 14, 15]). These findings are supported by a number of behavioural effects specific to faces. For example face recognition is more effected by inversions than ordinary object recognition [16, 17], configural processing is more important for faces than for other objects [18], humans detect changes faster and more accurately in facial images (in comparsion to other images) following flicker [19].

There are also a number of studies that challenge the specificity of face recognition, stating that other object recognition tasks can show similar properties if sufficient expertise is acquired [20]. Expertise also brings some amount of configural processing with it [21]. In Gauthier and Tarr's famous experiment, the subjects were trained to be *Greeble* experts (see Fig. 2.1), and were able to distinguish 30 different sorts of them. Some of the results that were obtained with faces were replicated with Greebles; the configural changes in Greebles affect the experts more than it does the novices.

Tong *et al.* remark that expertise-related visual judgements involve enhanced within-category discrimination, as opposed to between-category discrimination, the former requiring a magnification of differences in similar objects, and the latter calling for a neglecting of differences to group similar items together[22]. They demonstrate that a neural network model trained for being experts on one category also learns the expertise of another object category easier, as the mechanism for difference amplification is common to both.

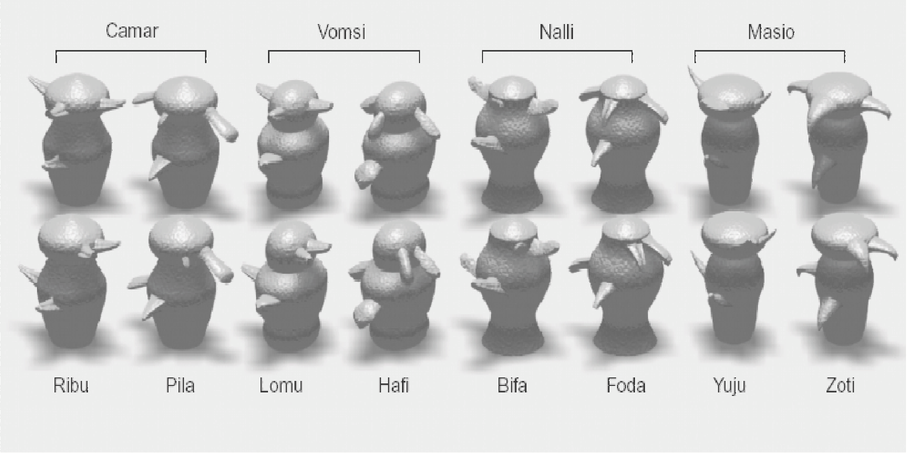


Figure 2.1. Some Greebles (From Tarr and Cheng, 2003).

It may well be that faces are one type of expertise, and their learning is also correlated with functions of communication. We elaborate on these points in Section 2.4. It is important to understand how much the brain is biased in learning the faces, and the nature of the input to the learner if the aim is to mimick the brain in this modality. A successful computer model in turn can validate which features and feature transforms are more informative, lending empirical support to cognitive hypotheses.

The statistical learning models have problems dealing with faces in recognition settings, mainly because the statistical differences between faces from a single person under different viewing conditions can have more variation than faces of two different persons under similar viewing conditions. The primary aim of studying human face recognition for computer scientists is to understand the representations and processing that result in correct identification of faces. O'Toole *et al.* note that "models of human face processing must be sensitive both to the statistical structure of faces and to the statistical structure of our experience with faces" [23]. The second part of this proposition is related to the fact that prolonged exposure to stimulus which calls for discriminatory judgement (in short, expertise) induces a higher resolution (for instance subordinate rather than entry level recognition), a sharpened sensitivity to informative dimensions and a change in the way the task is achieved.

Wallis and Bülthoff review the neurological findings on object recognition, and conclude that object recognition in the brain is hierarchical, distributed, and view-based [24]. The first two attributes actually apply to everything about the brain, but the proposition that object recognition is view-based leads to a number of assumptions about the face recognition capabilities as well. Although people can generalize to novel views of objects to a great extent, novel views are recognized slower than familiar views. That means there are ways of processing the input to generalize to novel views, but the processing comes with a computational cost. One group of theorists thus suggest that object recognition is achieved by storing different templates for objects and their different viewpoints. Experiments with computer generated 3D objects have shown that object representations in the brain are not viewpoint independent [25].

Another approach that dates to the 1980's is the structural description theory, which postulates that the human brain segments the images and the relations of these segments to each other are used for describing object classes [26, 27]. The two approaches can be treated as holistic and analytic, respectively. This is an important distinction that we elaborate upon in Section 2.6. Hummel notes that the viewpoint-dependence found in human object recognition does not necessarily mean that object recognition is explicitly template-based [28]. Structural relations remain important sources of information, both for object and face recognition, and were employed in computational face recognition models successfully [29].

2.2. Depth Perception

Three dimensional form of objects in the world are peceived by human viewers from a number of different cues. Motion of objects, oculomotor cues (which include *convergence*, the angle changes of the eyes to focus on nearby objects, and *accomodation*, which is the change in lens thickness for focusing on objects of different depth), and pictorial cues like perspective, texture gradients, shading, and contour information are found to be important for depth perception [30]. The binocular disparity that leads to stereoscopic vision doubtlessly plays a central role in depth perception at all levels. Although there are a lot of disparity tuned cells in V1, recent research indicates that stereoscopic depth perception reaches deeper into the higher visual areas [31]. The V1 cells tuned to binocular disparity may be working on a simple principle that leads to firing when the same feature is detected for a particular disparity and visual angle. Yet these neurons will only perform local feature matching, and by themselves, cannot solve the correspondence problem.

In [32] the authors use a hierarchical neural network model of the human visual system (VisNet) to demonstrate that for simple 3D objects, the network model can generalize to novel views, provided that the local features retain similarity to some degree.

The research on depth perception at the behavioural level faces some difficulties. Since the 2D information is very difficult to eliminate from the 3D information in an experiment conducted for human subjects, the relative importance of 2D and 3D information is not easily determined. The method followed by developmental psychologists is to look for earliest evidence of usage for each particular type of information. Thus, it has been found that motion-related cues predate other types of information.

There is another gap in visual cognition that relates to depth perception. To our knowledge, there are no studies that investigate attention modulation via true depth cues. In [33] a virtual reality environment is presented to the subjects, whose captured eye movements are used to interpolate points with attributed depth. In [34] the subjects eye movements on 3D model images were used to implement a saliency-guided mesh simplification algorithm. Similarly in [35], eye movement patterns of subjects viewing a 3D image were recorded to implement a saliency function, which was later used in 2D-3D registration. In all these cases, the subjects gaze into the computer screen to see the 3D object. Eye movement research focuses on computer displays and on flat surfaces, but the actual 3D object may cause a different saccadic pattern to emerge, one that relies partly on depth cues. This is an area where the computational models that integrate depth cues may provide valuable insight for biological systems.

Kavšek notes that the influence on depth perception on face processing has been neglected by developmental psychologists [30]. According to him, the most important class of 3D objects in the infant's visual world are human faces, and to recognize and discriminate familiar faces, the child needs to learn how to process the internal facial features, as well as to acquire a structural understanding of facial configuration. The depth information is relevant for the child, as it makes the 3D structure of the face explicit. However, two questions need to be answered: Which 3D features are relevant, and at what point of development are they used. These issues are not fully investigated yet.

2.3. Developmental Perspectives

Two questions are asked by the developmental psychologists about the face: Is face recognition domain specific, meaning that, is there a special mechanism just for faces; and if it is, how does this face-specific processing change during the course of development [11].

Development is a key concept in computational modeling of human cognitive functions. A cognitive skill is either innate (partly or totally), i.e. it is present in the brain prior to exposure to the sense data, or it is acquired over a time. The acquisition may involve a number of different stages and representations. It can also depend on the development of other skills, or classes of skills as Piaget implied in the stages he proposed (i.e. sensorimotor, pre-operational, concrete operational and formal operational) [36]. We would like to understand how much of human face recognition skills are present at birth, how the rest is developed after exposure to faces, which other systems (like attention, emotions, communication, etc.) it relates to, and how the representations of faces change. Taking a developmental perspective means tracing this change, and perhaps, understanding how the representations change incrementally as a function of older representations and new input (*representational redescription* [37]).

There are several reasons for believing that there are innate neural mechanisms that facilitate face recognition. In a biologically-motivated computer system, this innate part corresponds to a system bias, something that is not explicitly learned, but facilitates -or rather directs- learning.

Expertise in humans -and also in computers- is closely related to representation. Becoming an expert means storing and accessing more information, but it also means more efficient representations. The classical example is the chess master, who perceives patterns of offensive and defensive formation rather than individual pieces on the board. The memory of a chess master for meaningful game configurations is much better than a novice, but on configurations that cannot be related to a chess game masters and novices perform equally [38]. This means that while the master acquires an efficient way of representing information, this representation depends on and facilitates the task at hand. It should be useful at this point to keep in mind that expertise need not follow long exposure (e.g. sheep that are exposed to human faces for a long time reach a certain performance level by learning the facial features, but fail to develop a configurational encoding [39]).

We can easily say that humans are face experts. How they represent faces is correlated with what they do with the faces. Recognizing familiar people, detecting emotions, reading lips to facilitate communication, establishing eye-contact to initiate and maintain communication: These are some of the common tasks that involve face recognition. When we say that eyes contain discriminatory information, we rely on psychophysical data; we can validate experimentally that humans attend to eyes and corners of the mouth in a novel face (See Fig. 2.2 from [40]). All these habits are born out of expertise and necessity. It is the information content of the eyes that is continuously used in communication that drives us to attend to it. Therefore, it is useful to distinguish between top-down and bottom-up information content. The bottom-up information is of statistical nature, and the human brain makes use of it in time. The top-down information is contextual, and it probably has bottom-up roots: When we communicate, we learn that eyes matter. Later, this information is taken for granted, and we impose it on the saccadic system in a top-down manner.

There are two opposing hypotheses for face recognition. The *expertise hypothesis* holds that a single object recognition system with some initial bias (and lots of training) is sufficient to achieve human face recognition [41, 42]. The *domain specificity hypothesis*, on the other hand, claims that face recognition is unique, and supported by innate mechanisms and not the result of a general object learning system [43, 14, 44].

Whenever a human ability seems too good to be true, innateness becomes an issue. Face recognition and language are similar in this respect. A long-standing argument for the innateness of language has been the poverty of stimulus argument; the child does not receive what we would call supervised training in language during its first year, but constantly hears fragments. Tomasello argues that actually the stimulus

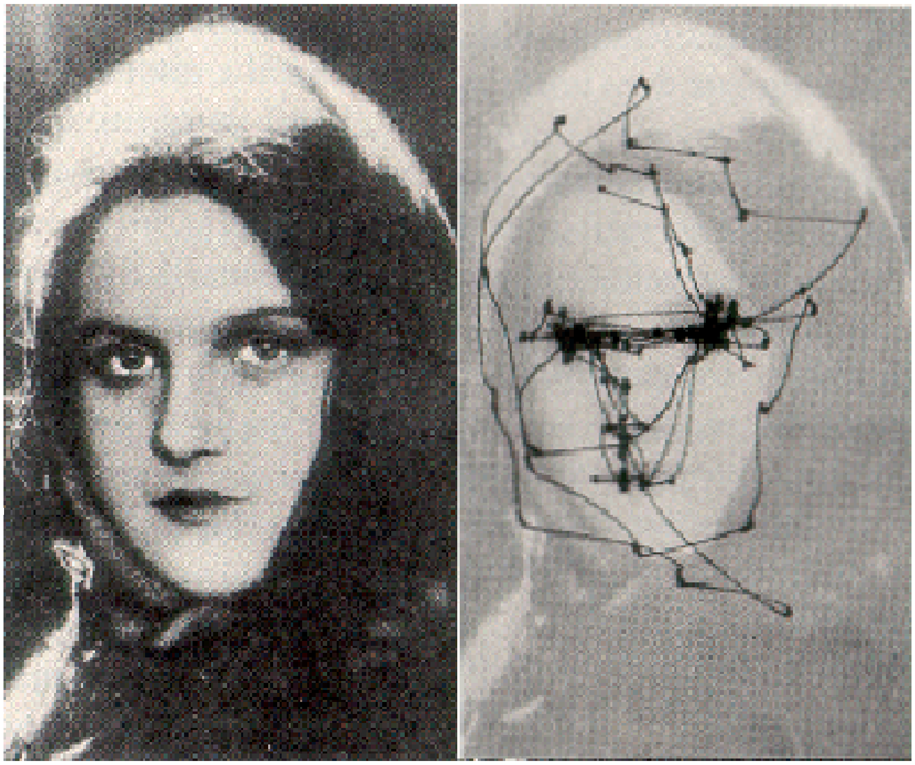


Figure 2.2. Selective attention on human faces. The saccades are focused on the eyes and other spots that contain discriminatory information (from Yarbus, 1976).

is far from being poor, the child hears some six million words up to its seventh month, and valuable information is present in the statistical distribution of these words [45]. A similar argument can be made for faces. The child sees millions of face-shapshots in various illumination, pose, scale, occlusion conditions. Any system that physically underlies the human face recognition is trained with enormous amount of data. Here, as in (and parallel to) linguistic communication, eye contact and imitation serve as means of bootstrapping (See Fig. 2.3), and are supported by innate mechanisms [46].

Infants are genetically biased to attend to faces. When they are presented with faces and other objects, they look at faces for longer periods, indicating their interest (*face preference*) [47, 48]. Actually, this should not be very surprising, as many animals are similarly biased for their species. There is a very pronounced evolutionary pressure for an early mechanism that binds the infant to its parents, who are potentially protectors and teachers. Thus the early skills of face perception might play an important role in developing language and communication skills [49].

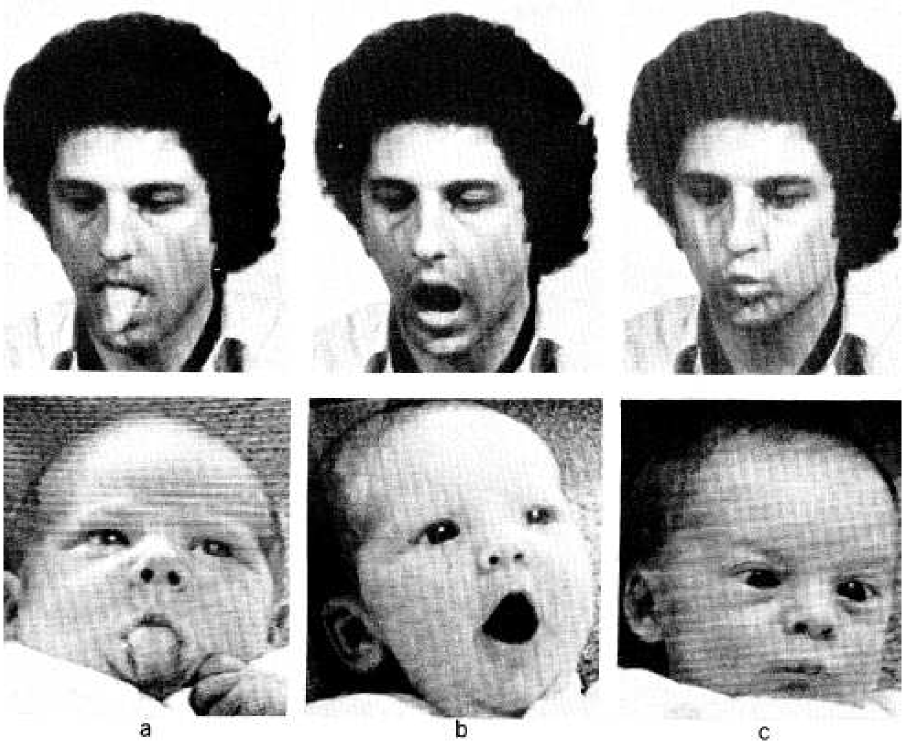


Figure 2.3. A three-week old baby can imitate some facial actions by the adult, which probably does not imply a conscious understanding of the gesture and a subsequent controlled action response, but provides evidence for face-related innate mechanisms to initiate communication and learning (from Meltzoff and Moore, 1977).

Xu and Carey found that 10-month-olds treat objects as general objects, and not as individual entities [50]. In the experiment they performed, a cow enters a box, and then a truck comes out of the box. The infants are not surprised, which means that for them, one object entered, and one object left. Only if they see the cow and the truck together do they show surprise at the mentioned setup. This experiment suggests that infants can detect the boundaries of objects, but do not distinguish them according to their features, when they are very young (Object First Hypothesis). Later, feature-based processing is added on top of this more general, holistic perception. Infants can recognize their mothers after only a couple of days, based on the general shape of the face, and the hair plays an important role as well. If the mother changes her hairstyle, the infant fails to recognize her. This means that the relative unimportance of hair is learned only later. For adults, hair is a semi-salient feature, much like a recognition heuristic. We use it for recognition, but we do not rely on it explicitly.

In an important early study by Goren, Sarty and Wu, a number of newborns with median age of nine minutes were shown moving schematic faces, faces with scrambled features, and just a blank head outline [51]. The infants attended to the schematic faces for longer periods. Later, Maurer and Barrera showed that one month old infants were no longer looking at the faces more; but the face preference returned at two months [52]. Failing to reconcile this pattern with a continuous learning hypothesis, Morton and Johnson came up with two distinct mechanisms operational in infants: CONSPEC, an innate structural bias for conspecifics that later gets inhibited, and CONLERN, which is trained by species-specific input (for instance human faces) and is probably regulated by CONSPEC [48]. The effect of CONLERN is felt only after some training time, which is used to explain the drop of interest to faces around one month.

The discontinuity in the type of processing is not acceptable to some. Bonatti *et al.* argue that infants start using features to distinguish between faces and other objects very early [53]. Their experiments show that infants are more sensitive to differences between a human face and a dogs face than the differences between, say, a metal car and a wooden box. Although their findings imply a feature-based component, the issue of innateness is not resolved, as the child is exposed to human faces for very long times after the birth. Any feature-based system has had ample training time by then. Since the child has poor visual acuity and sensitivity to contrast for some time after birth [54], the feature-based component would logically need some time to develop.

The most plausible account for the development of face-specificity comes from Elman and his colleagues [55, 37]. According to this account, very broad, genetically-tuned biases exist at birth, which, when combined with the structured input provided to the child by the more or less consistent environment, enables a quick development of neural tools to deal with this environment, including a face recognition system. Turati and colleagues performed experiments with faces and nonfaces that share featural properties with faces, and showed that newborns have no preference for faces, but they respond to facial inner features and features that resemble those [56, 57]. Another finding that supports this view is that six-month olds can discriminate between monkey faces, whereas nine-month olds cannot, having already specialized in human faces [58]. This type of specialization is typical in many domains including language, where one loses the ability to produce phonemes that do not exist in the native language.

2.4. Brain studies

There are a number of regions in the brain that respond to faces in a selective manner. We should mention especially the fusiform gyrus [49], and the superior temporal sulcus [59]. Experiments with Mooney images (See Fig. 2.4) show that when subjects perceive the Mooney images as faces, there is more activity in the fusiform gyrus [60]. This is an important finding, because in these experiments the stimulus does not change, although the perception does. The activation does not stem from a data-driven, bottom-up process.

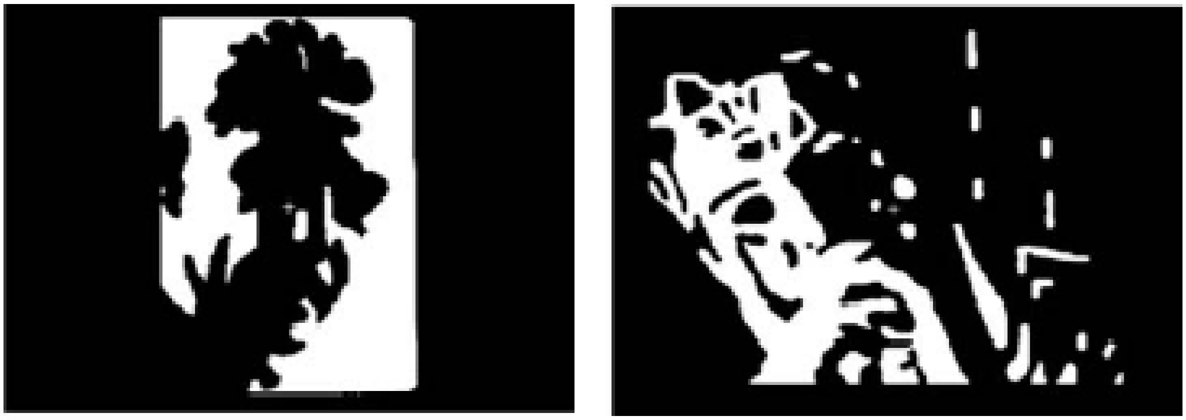


Figure 2.4. Mooney faces, from Andrews *et al.*, 2004.

These findings led Kanwisher *et al.* to name a part of the fusiform gyrus that shows strong activation in face recognition tasks as the *fusiform face area* (FFA for short) [14]. The face specificity of the FFA is a debated topic; Gauthier *et al.* claim that classifications involving expertise also activate the FFA [61]. Following their argument, Tong *et al.* humorously call the FFA the *fusiform fish area*, and claim that it would be employed in any expert within-category discrimination task [22]. These arguments are rejected by Kanwisher and her colleagues, on the grounds that the activation levels in FFA during general object expert judgements are much lower than activation related to face recognition [43, 44].

Additional evidence supporting the expertise view comes from studies with event related potentials (ERP). First, the existence of an enhanced early negative component called N170 was found to be specific to face perception [62]. Later, it was shown that N170 is also present in categorization of birds and dogs by their respective experts [63] and also in Greeble recognition of Greeble experts [64].

Face-specific effects have inspired many models of computer and human face recognition. For example, single cells in the human brain (as well as in the brain of lower primates) that respond to individual faces and specific face postures are found in the inferotemporal cortex [65, 66]. Based on these and other cells that respond to particular shape, colour, orientation, and speed percepts, Poggio and Hurlbert propose a face recognition system that uses radial basis functions with many templates to provide for various invariances [67]. Their claim is that the neurological evidence points out to many multidimensional units with Gaussian-like tuning in the brain.

The Capgras delusion is a very interesting disorder, where patients recognize familiar faces, but they lose the accompanying familiarity sensation [68, 69]. Thus, the patients claim that those people are impostors that replaced the familiar person. Experiments with Capgras patients led the researchers to believe that there are two routes of processing involved in face recognition. One of them is termed the affective route, and the other is the cognitive route [70, 71] (See Fig. 2.5).

2.5. Familiar faces

When humans see faces, they perceive each face individually, and not merely as specimens of a generic object category. This is obviously a justified property; recognizing six almost identical trees individually gives a human very little useful information, whereas recognizing faces individually is necessary to separate friend from foe.

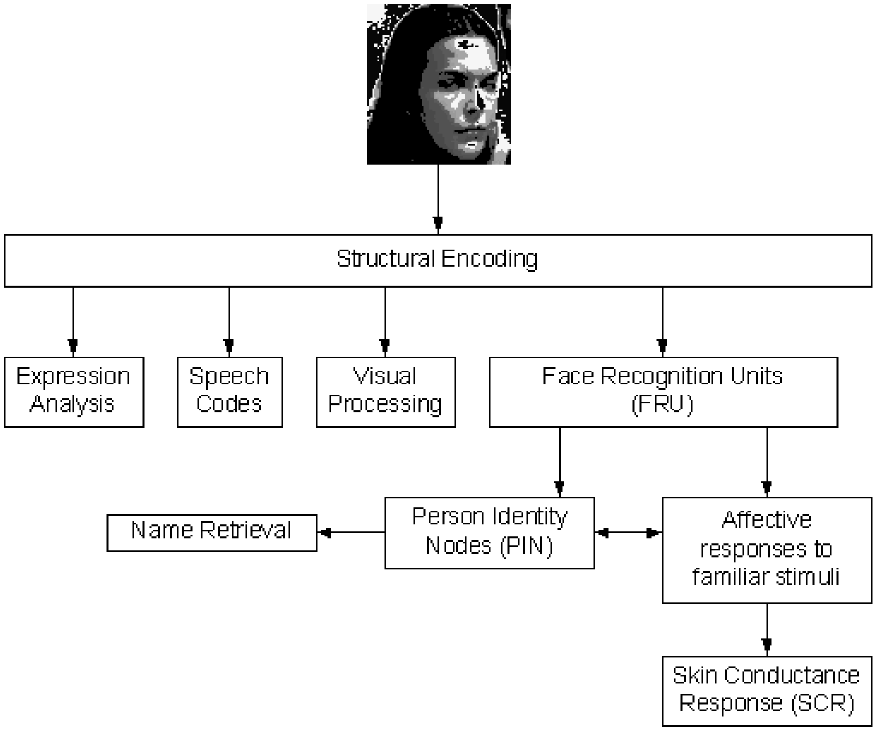

Structural Encoding
Expression Analysis
Speech Codes
Visual Processing
Face Recognition Units (FRU)
Name Retrieval
Person Identity Nodes (PIN)
Affective responses to familiar stimuli
Skin Conductance Response (SCR)

Figure 2.5. The dual-route model for face processing. Adapted from Breen *et al.* 2000.

Hancock *et al.* use principal components analysis (PCA) to predict a particular aspect of human performance in face recognition [72]. It is known that (in an experimental setup with limited number of target faces) familiarity judgments and distinctiveness ratings for a face are not correlated [73]. Their experiments suggest that familiarity judgment depends on the subject's previous experience, whereas distinctiveness is more or less consistent across subjects. If a face has strong average components (indicated by the PCA) it may be mistakenly classified as familiar. Distinctive faces are those that are distanced from the average, and PCA is able to predict distinctiveness judgements.

PCA was also applied by Valentine to predict the *other race effect* [74]. People have great difficulty recognizing faces of another race if they are not exposed to them for prolonged periods. This finding supports the importance of expertise in face recognition.

Although we tend to believe that the human face recognition capabilities are robust to affine transformations, occlusions, illumination and expression changes, this is not true. We need to distinguish between familiar face recognition and unfamiliar face recognition. Familiar faces are the ones that we have seen lots of times, that of friends and family for instance. For these faces, our recognition system is robust to changes. It is postulated that we have found out the most discriminative features of these faces, and use this feature-based information rather than the whole face in recognizing familiar faces.

Unfamiliar faces are the faces we have seen only once or twice, from a limited set of viewpoint and conditions. For these faces, illumination drastically effects the recognition performance. A face that is illuminated from below (which is very unnatural, because we are genetically biased to expect lighting sources to be above) is very hard to recognize. Familiar face recognition is also effected by such a change in illumination, but less so.

2.6. Holistic versus feature-based processing

In an early study, Yin showed that inverting faces impairs their recognition far more than it would impair the recognition of any other object (called the *inversion effect*) [17]. The individual features of the face are not affected too much by inversion. It is argued that inversion disrupts the configural processing, which is specific to faces. There are other findings that support this claim; features of the face are much easier recognized in the presence of other facial features (*face superiority effect*) [18], and inversion of isolated features affects their recognition rates relatively less [75].

Based on these studies, a distinction was made between *holistic* (also called *global, configural, relational, monolithic* and *coarse*) and *feature-based* (also called *analytic, local, piecemeal, part-based, componential* and *fine grained*) [76]. According to one extreme, face recognition is entirely holistic, and faces are perceived as units [18]. Martelli *et al.* challenge this view by showing that face superiority can arise from the combined effect of familiarity and crowding (i.e. presence of multiple features that need to be integrated); facial features are combined into faces just like letters are combined into words [77]. On the other hand, findings suggesting that face recognition has a right-hemisphere bias agree well with the holistic approach, as the right hemisphere of the brain is assumed to be more important for relational encoding [11].

Studies on the other race effect have also shown a differentiation between holistic and feature-based processing. Caucasians with little exposure to Asian faces have shown holistic processing for recognition of Caucasian faces, and featural processing for the recognition of Asian faces [78]. Holistic processing seems to be a more advanced form of processing, which requires more training, or innate support. The so-called *holistic own-race advantage* is the cognitive inspiration behind the alternative face-based registration algorithm we present in Chapter 6.

An interesting effect was demonstrated by Thompson in 1980 by inverting several features of Margaret Thatcher's face (See Fig. 2.6) [79]. This procedure is called *thatcherisation*, and the whole effect is termed the *Thatcher illusion* [80]. When one sees an inverted thatcherized face, one does not perceive the grotesqueness one would perceive had one seen the face upright. The most plausible explanation lies with the assumption that on the inverted faces, features are processed individually, and their outputs are integrated later. The configural processing is disrupted by inverting the face. Moscovitch *et al.* have tested various face recognition conditions (including isolated face features, fractured faces, faces made of other objects, inverted faces, caricatures, etc.) on a patient with object agnosia, and found out that whenever the configural information was lost, the performance of the subject decreased greatly [8]. This was seen as an indication that the configural processes were intact for the patient (hence the performance change), but a separate, feature-based system that co-exists with the configural process was impaired selectively.

As a side remark, we should note that inverting features in Thatcher's illusion does not completely eliminate configural processing. If we think about a hierarchy of processing elements, moving from purely feature-based processing units to purely-configural processing units, there must be units that encode only some configural information as well. For instance, the chin, the ears, the hair, the rough shape of the head are not changed in the scrambled Thatcher face. A partly-configural processing unit would find useful information in these features and their configurations. This claim can be tested by varying the degree of scrambling, and we should encounter a graceful degradation in performance if the claim is correct.

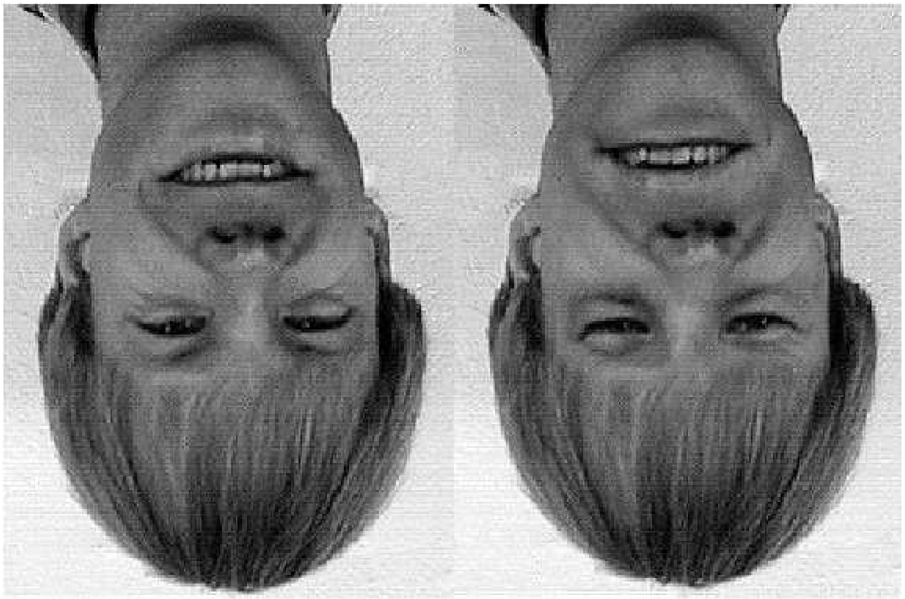


Figure 2.6. The Thatcher illusion, from Hancock *et al.* 2000.

One important assumption about human visual processing is that a scene is not perceived at a single glance, but parts of the scene are visited incrementally. The brain immediately starts processing the parts that are already seen, and this process influences to which locations of the scene the eye will attend next (top-down attention). The so called microgenetic face processing models were investigated by Carbon, who found that inverted Thatcher faces are recognized faster than ordinary inverted faces [81]. The global features (e.g. outline of the face) are recognized very fast and accurately, but the repeated interaction with a person leads to emphasized eye and mouth recognition, the parts that are the most involved in communication. Thus, for familiar faces, the early visual processing gathers local feature information and simple global information; the holistic, relational face processing (e.g. global template matching) comes at a later stage. Earlier studies using response time paradigm suggest that configural and component processing strategies can be usefully combined for face recognition [82]. We incorporate this perspective into our face recognition framework in Section 3.6.

Ullman and Sali note that small fragments of images corresponding to simple features processed in the early to intermediate stages of the visual system can be used for object detection in general, and face detection in particular [83]. Instead of using universal features that are simple enough to be used for all classes (e.g. Gabor wavelet features), the authors propose class-related features that maximize mutual information for each object class that indicate increased probability for the presence of that object in the image. These features are more complex than the lines, edges, and centre-surround receptive fields processed in the first stages of the visual system (retina, LGN, V1), but simpler than partial or complete representations of the objects postulated for the final stages (anterior regions of the inferotemporal cortex). The authors note that the success of the fragment-based approach is consistent with an object representation in the brain that consists of class-related fragments of intermediate complexity (See Fig. 2.7). A hierarchical organization of different resolution levels is also put forward, where bigger fragments are stored and matched with a low-resolution, as a large and high-resolution fragment has a very low probability of occuring. In this approach, the 3D information is not explicitly stored, but is assumed to be present in the fragments to some extent.

Heisele and Koshizen use fragments that are grown iteratively on fourteen predetermined feature points of the face for face recognition (See Fig. 2.8) [84]. Their findings confirm the promise of class-specific fragments for recognition, but they add that the method works for small databases (their database includes only six persons). In addition to class-specificity, the learned fragments are also viewpoint-specific.

Caricatures provide an interesting insight into face recognition. Caricaturizing a face is achieved by selecting features that deviate from the average, and by exaggerating them even more. Valentine proposed that the faces learned by the human brain constitute a multi-dimensional space, called the *face space*, where the directions represent meaningful consistent changes (expression changes, fattening, larger nose,

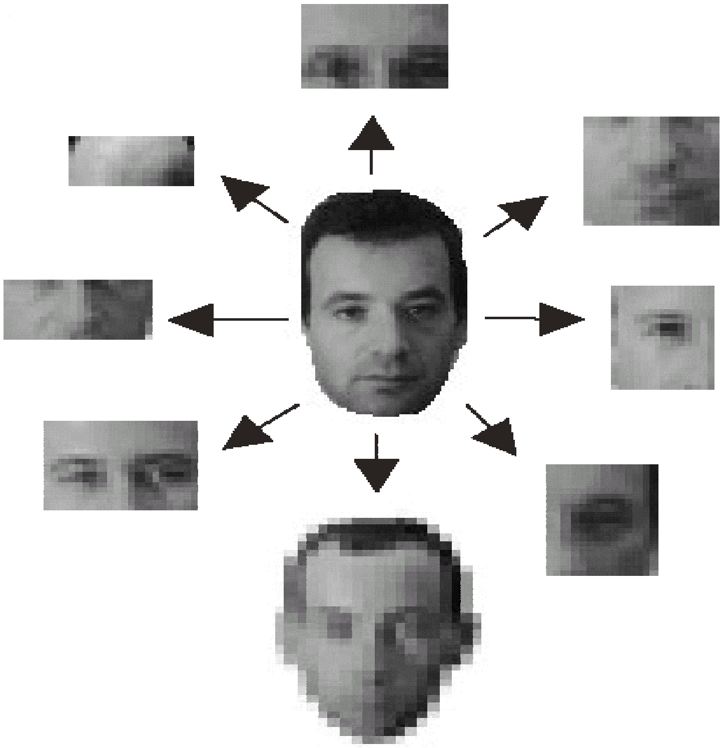


Figure 2.7. The best eight features (from top, moving counter-clockwise) in terms of mutual information proposed as visual features with intermediate complexity in Ullman *et al.* 2002.

etc.) [74]. If we assume that the face space hypothesis is correct, then caricaturizing a face means moving the face away from the average face on the direction vector found by substracting the average from the face itself. An *anti–caricature* is obtained by moving the face towards the mean.

The first computer automated caricature generator that rests on this idea was created by Susan Brennan [85]. The caricatures created thus contain sufficient information for familiar face recognition, and the recognition performance is higher for the exaggerated drawings in comparison to drawings that are faithful to the face contours [86]. Humans identify caricatures more accurately, and anti–caricatures less so [87]. These effects were demonstrated to exist in computer-based face recognition systems as well [88]. O'Toole *et al.* found that 3D caricaturing increases the apparent age of the subject, and anti-caricatures are judged by humans to be younger and more attractive [89].

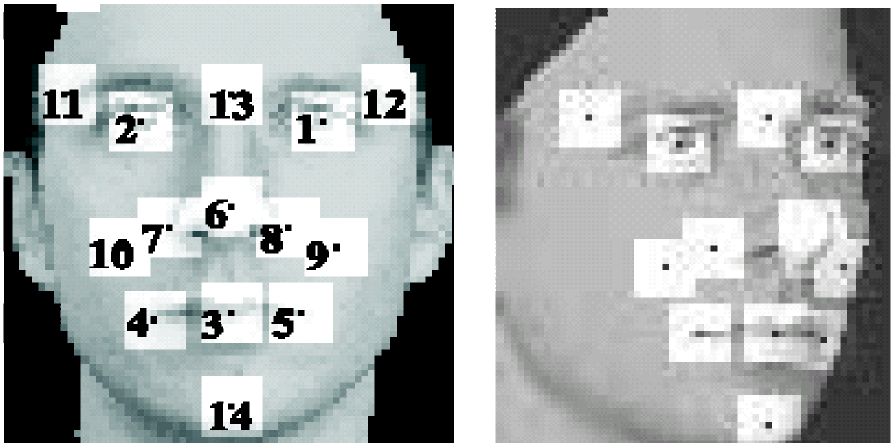


Figure 2.8. The initial component locations for the frontal and rotated faces (from Heisele and Koshizen, 2004).

The effects of caricaturization are not confined to drawings. Exaggerating features in a photograph helps recognition as well, although to a degree [90]. The effect in drawings is much more striking, particulary as the original drawing contains poor discriminatory information in the absence of shading cues. These findings are consistent with the view that a feature based component and a holistic component work together to achive recognition. Caricaturization helps the feature based system, and the improvement is more obvious in the absence of the holistic system.

2.7. Computational Models of Human Face Recognition

The earliest face recognition methods with a claim to biological plausibility are the eigenface method [91] and elastic graph matching [29]. The eigenface method employs PCA, and can be seen as a holistic approach to face recognition, where the coefficients of the first few eigenvectors that code the face are used for recognition. The resulting system is sensitive to scaling, translations, and pose. In order to alleviate some of the shortcomings, view-based variants of the algorithm were proposed, where a set of eigenfaces was stored for each view [92]. The similarity space obtained after PCA correlates well with human perception of similarity [93, 94][2] .

[2] The projection offered by the PCA also ensures a good encoding for a given training set, but still, it does not have to correspond to the projection realized in the human brain [95]. Indeed, to calculate the eigenfaces, one needs to have access to the whole training set at once.

In a related holistic approach, Ramasubramanian and Venkatesh suggest transforming face images to the frequency domain by a discrete cosine transform (DCT), and making use of the sensitivity of the human visual system to low-frequencies in conjunction with the observation that the energy of face images are primarily in the low-frequency regions, they implement a system that achieves good dimensionality reduction and high recognition accuracy [96].

In the elastic bunch graph matching method, features are derived from the faces, and matched to a bunch graph that is constructed from the features of the training images [29]. The Gabor wavelet filters that are used for obtaining the features behave like simple cells in the early stages of human visual system (i.e. V1), which lends the model some biological plausibility [97]. This model was also found to correlate well with the human recognition of faces [93]. The graph model encodes the structural relations between the features and performs better than the eigenface approach [98].

Wallraven *et al.* recently proposed a feature-based framework for object tracking and recognition, where the images are analysed at multiple levels for the presence of corners, and a feature similarity metric based on the location and content of each selected feature window is used to solve the correspondence problem for sequences of images [99]. They compared three different viewpoint-dependent approaches (aligning the image with a 3D model, using linear combinations of 2D images, and interpolation, respectively) with their model on a face recognition problem and find that psychophysical data support their model [100]. The configural processing breaks down when the facial parts are scrambled, and the feature based system breaks down when faces are blurred [101]. These experiments support the existence of a feature-based subsystem for object recognition in humans.

Kalocsai *et al.* compare a local feature based system that uses Gabor-wavelet responses and a global PCA + LDA system with the human performance on a dataset of 32 images on a similarity judgement task [102]. Their findings indicate that both systems show positive correlation with the human performance, but the local feature based system has better prediction.

2.8. Conclusions

Our survey on human face recognition is not complete. We have left out the temporal aspects of object recognition in general, and face recognition in particular. It has been argued that temporal association is the glue that binds the different viewpoints of the objects [103], and the motion cues are important in face recognition, especially for non-optimal viewing conditions, for familiar face recognition, and in predicting the shape-related features [104]. Selective attention and its effect on object recognition is an area left out from our survey as well. We have also left out the exact localization of related functions in the brain and neurophysiological theories of information processing. However, we hope that in this form, our survey covers enough ground to suggest good heuristics for better recognition systems.

Although our purpose is to build a good 3D face recognition system, the effort is not entirely irrelevant from the cognitive science perspective. Most of the research conducted with infants about face perception employ 2D images of facelike stimuli. However, Slater *et al.* have failed to obtain a transfer of experience between the perception of 3D objects and their 2D projections in the infants [105]. Using 3D information for a biologically-motivated approach might be more realistic than working on 2D images.

The second benefit of looking into the human visual system is grounded in the robustness of the evolutionary process that shaped it. We can safely assume that the brain is optimized to a certain extent for face recognition. Some of the resulting operations are derived from the scarcity of certain resources (e.g. limited foveal window, types of photo-receptive cells, etc.), but others may point out to statistically sound ways of going about face recognition. The dichotomy of holistic and analytic processes for recognition suggests that a combination of both approaches would be beneficial. However, since the types of processing required for these approaches are different, this may result in doubling the computational requirements for little gain. As Chellappa *et al.* note in their widely cited survey on face recognition [106]:

> “Designers of face recognition algorithms and systems should be aware of relevant psychophysics and neurophysiological studies but should be prudent in using only those that are applicable or relevant from a practical/implementation point of view.”

It is difficult to separate 3D face recognition from 2D face recognition. This will be apparent in the next chapter, which contains a stand-alone survey of the state-of-the-art in 3D face recognition, followed by the description of a model that represents an attempt at conjoining the intuitions and considerations derived from the present chapter, with the technical requirements of implementing a 3D face recognition system in computers.

3. STATE OF THE ART IN 3D FACE RECOGNITION

"I'm leaving for abroad, Zorba.
The old goat within me has still got a lot of papers to chew over."
Nikos Kazantzakis

Improvements in sensor technology, and the difficulty of implementing robust 2D face recognition systems have made 3D face recognition an attractive alternative, especially in biometrics applications [107, 108]. The cost and the accuracy of a 3D system depends mainly on the hardware, yet different physical settings call for different algorithms. Accurate sensors require less pre-processing, but they are more costly. Conversely, cheaper sensors call for much more robust algorithms. It also makes sense to explore the possibilities of using 3D and 2D together, as acquiring 2D simultaneously with 3D usually requires small or no increase in hardware and acquisition time, but also because there is a large body of research on 2D techniques.

This chapter deals with general and specific considerations for 3D face recognition, and aims at laying the groundwork for the rest of the thesis, as well as to serve as a reference to the state-of-the-art in 3D face recognition[3] .

There are a number of questions we would like to look at to gain a unified understanding of 3D face recognition. We will devote one section to each of these questions.

- What are the settings (i.e. scenarios) in which 3D recognition is used? (Section 3.1)
- How should we acquire and represent the 3D information? (Section 3.2)
- What types of preprocessing are necessary for reliable recognition? (Section 3.3)
- How do we bring faces into a common coordinate frame for fair comparison? (Section 3.4)

[3] Some of the material presented in this chapter will appear in the Handbook of Multimodal Biometrics [109].

- How should we use 3D information in recognition? (Section 3.5)
- What are the main research questions for 3D face recognition, and where is the present work situated? (Section 3.6)

A glossary of terms and techniques mentioned in this chapter is included in Appendix A.

3.1. Scenarios

Face recognition has different meanings in cognitive science and computer vision. For the latter, the problem of 3D face recognition can be conceptualized in two different scenarios:

1. **Person recognition:** In the ordinary recognition scenario, we have a *gallery* of faces. Their representation can be 2D, 3D, a combination of both, or it can consist of extracted features. It can even be in the form of a trained classification algorithm. However in that case, the extensions to the gallery may be problematic. The problem is to find the best matching face in the gallery for a novel sample. The sample can contain partial information. For example, the gallery may contain 3D meshes, but the new instance can be a 2D snapshot [110]. The alignment of a 2D image to a 3D model poses different problems.

   Once the matching is done, the next step is to decide whether the best match is good enough, i.e. whether the person is in the gallery or not. The recognition system can also return a ranked list of candidates for further inspection. In a *screening* scenario, a human expert is usually the final judge of the outcome, and the face recognition system is required to produce a small subset of the gallery faces for the human expert as candidates.
2. **Person verification (authentication):** In this scenario we have a person with a claim, and we test the novel instance against a single class. Lu and Jain count among the benefits of 3D systems the fact that they are more difficult to fake [111]. The verification problem can be simplified by assuming controlled illumination and pose conditions, although these assumptions are not always valid.

It is common to report receiver-operator curves (ROC) for the verification scenario, where the threshold for acceptance is changed over a range, and number of falsely accepted instances is plotted against falsely rejected instances. The equal error rate (EER) occurs where the plots intersect. The EER is actually a special case of weighted error rate (WER), where the false accept rate (FAR) and false reject rate (FRR) penalties are weighted depending on the application. One should also keep in mind that the experimental setup (e.g. number of impostor accesses in the evaluation scheme) has different effects on these rates, and ROC curves must be evaluated with respect to the experimental setup [112].

For a large gallery of faces, the recognition scenario is more challenging than the verification scenario, as comparing one sample to all the gallery produces more false targets. The computational complexity is a also a very important issue here. The registration of a new sample to each and every gallery face requires a lot of time. We will propose a way to deal with this problem in Chapter 6.

3.2. 3D Acquisition and Representation

3.2.1. Acquisition

We distinguish between a number of range data acquisition techniques. The usability, cost and performance of any 3D face recognition system greatly depends on the acquisition method.

1. **Acquisition with stereo cameras:** In this technique, two or more cameras that are positioned and calibrated are employed to acquire simultaneous snapshots of the subject. There are cases where the images are not acquired simultaneously, and we can still talk about correspondence matching, but this is not “true stereo” and poses a greater challenge [113]. Correct calibration of the cameras is important for the accuracy of the representation [114]. The depth information for each point can be computed from geometrical models and by solving a correspondence problem. If the exact camera locations are not known, it will be hard to solve the

correspondence problem, as the human face contains smooth areas like the cheek. In [115] the 3D face is constructed from two 2D images (a frontal and a profile, respectively) by warping the first to the second to establish the correspondence. Another approach is to deform a generic 3D model to fit a number of derived landmark locations [116, 117].

2. **Acquisition with structured light:** In this acquision scheme, the structural light patterns are obtained by using a projector in addition to the camera. The depth information is derived by analyzing the deformations in the patterns. This setup is more expensive than the previous approach, but it needs only a single camera for producing 3D information. The acquisition time is shorter than in laser scanners, and the 2D texture can be acquired almost simultaneously. There is some processing overhead to derive 3D information. As the 2D image is taken from a fixed perspective, the resulting depth information will not be fully 3D. This type of depth information is frequently called 2.5D, and it is possible to represent it like a 2D image. The light patterns can be colour coded and the texture can be recovered from the obtained colour, like in the HISCORE project [118]. This system was employed in [119].

   For increased resolution, one needs to increase the number of patterns that are projected on the image. This may lead to ambiguities in matching. If coloured stripes are used, the colour difference between adjacent stripes will decrease. Wong *et al.* solve this problem by repeating a set of coloured stripes across the image [120].

   Structured light based acquisition is used in some of the most important 3D face databases [121, 122, 123]. A cost-optimized solution with a single camera and a projector was outlined in [124]. Chang *et al.* use a range scanning camera (Minolta Vivid-900) that uses structured light and provides depth and intensity values for all points on the acquired image [125]. The derivation of the 3D information and its accuracy also depends on the chosen structured light pattern [126]. Batlle *et al.* present a survey of different coded structured light methods [127].

   The structured light approach can also be combined with stereo acquisition. A stereo system with three cameras and speckled light pattern projection that extracts range and texture information is described in [128]. In Fujimura *et al.* a

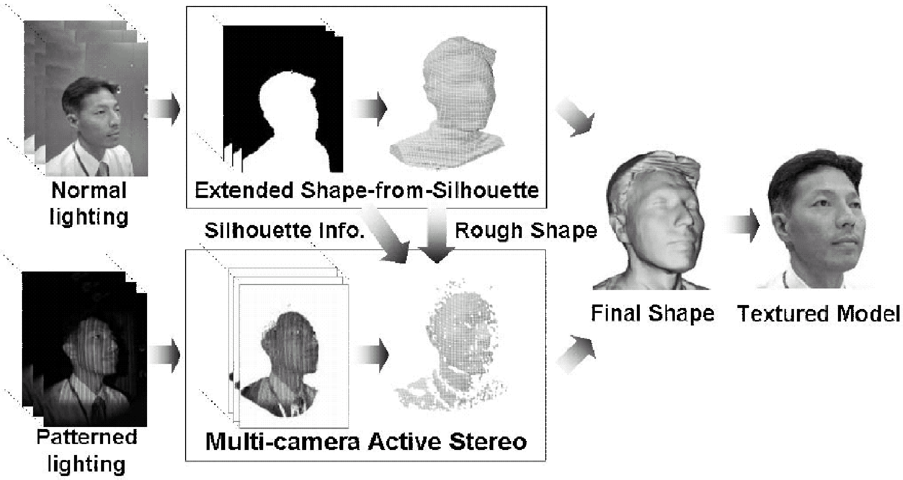


Figure 3.1. The hybrid modeling algorithm proposed in Fujimura *et al.*, 2004.

multiple camera system that acquires the depth image by feature-based stereo matching is detailed [129]. A rough shape is determined from the silhouette, and the depth data are acquired via structural light projection by multiple cameras (See Fig. 3.1). The integration is performed by voting. The obtained data are refined through interpolation and area based subpixel estimation. The system uses 16 cameras under normal lighting, and 12 cameras under structured light. The computation of a full 3D textured model takes about fifteen minutes on a 2.4GHz Pentium IV with 512Mb memory.

3. **Acquisition with laser sensors:** Laser sensors are more accurate, but also more expensive and usually it takes more time to scan a face with a laser scanner. They are especially useful in producing accurate head models for the gallery [130, 131]. In Lee *et al.*, silhouettes obtained from laser scans are used in establishing correspondence between vertices [132]. Tsutsumi *et al.* combine a fiber grating vision sensor (made up of a fiber grating, a laser diode and a CCD camera) with an infrared sensor to acquire range data with gray-scale image data [133]. The Minolta Vivid 700 uses a laser line in conjunction with a CCD camera. The positions of the light emitter and sensor are used to calculate the range to the surface through triangulation. In Zoller and Fröhlich LARA 25200 scanner the emitted light has a sine-wave power variation that is detected in the reflected signal [134]. The phase difference method suffers from the wrap around of the phase, and the range of overlapping phases is called the *ambiguity interval.* The

ambiguous points need special attention during preprocessing.

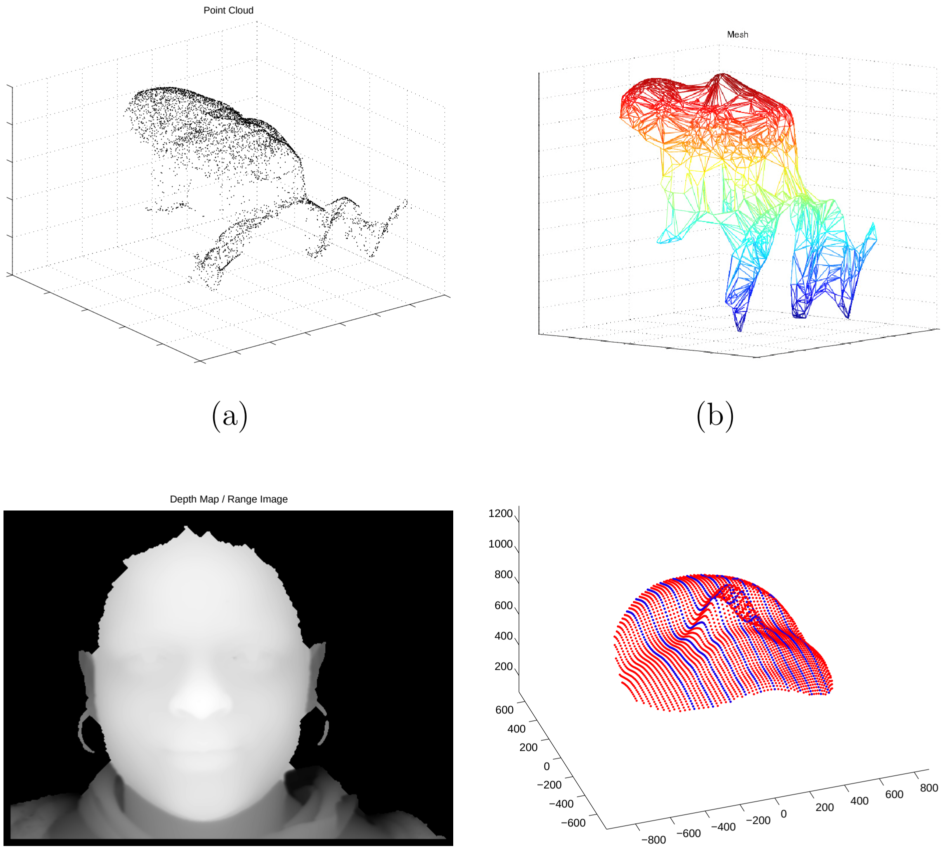
The FRGC dataset that we use in the subsequent chapters of the present work was acquired with a Minolta 910 scanner, which produces a 3D depth map with its registered 2D texture image [135]. The relatively slow scanning speed of the scanner gives rise to poor correspondences for some of the acquisitions. Blais gives a thorough survey of commercial sensors and their range and accuracy characteristics [136]. The calibration of these systems is discussed in [137]. See [110] for a survey of 3D assisted face recognition techniques that especially pertains to sensor technology.

4. **Acquisition with CT/MRI:** Magnetic resonance imaging (MRI) and computed tomography (CT) scans can also be employed to determine the head shape [138]. This method is very expensive, slow, intrusive, and should be supplemented with additional input if the texture is needed.

There are commercial systems for the all these approaches, although the trend with later structural light systems is to use multiple cameras as well. Genex Technologies produces 3D FaceCam, where three sensors acquire the face image simultaneously [139]. Their structural light-based Rainbow250 camera was used to collect a database in [140]. Similarly, Geometrix's FaceVision employs a two-camera stereo system [141]. In all these systems, high-resolution 2D images are used for constructing 3D shape. A4 Vision uses near-infrared light to produce a 3D model with accompanying standard 2D texture, but their internal representation is not available as an output [142]. The Minolta Vivid 910 [143] and Cyberware 3030 [144] are examples for the laser scanning approach, although the former is not specific to facial images. These acquisition methods can be extended to capture 3D video. In [145], a structural light system is described that can acquire 3D scene information with 30 frames per second.

3.2.2. Representation

The most frequently used representations for the acquired 3D data can be listed as:



(a) (b)

(c) (d)

Figure 3.2. Common 3D representations (a) Point cloud (b) Mesh (c) Range image or depth map (d) Profile curves.

- Point cloud: A large number of 3D points that are sampled from the surface of the face are stored (See Fig. 3.2 (a)).
- 3D mesh: Triangularization is used to produce a mesh from the point cloud (See Fig. 3.2 (b)).
- Range images: A range image is a 2D representation that has depth values instead of intensities (See Fig. 3.2 (c)). One or more 2D range images can be stored for a more complete representation, as the range data is taken from a single perspective.
- Feature sets: There are different features that one can derive and store for each face. Typical features are profile curves (See Fig. 3.2 (d)), landmark locations (nose tip, eyes, corners of the mouth, etc.), surface normals, curvatures, shape indices, depth and/or colour histograms, edges, and subspace projection coefficients (PCA and LDA are frequently used).

Usually, more than one representation is used in a single algorithm. Texture information, if available, is generally stored for each 3D point or triangle.

3.2.3. 3D Face Datasets

Research on 3D face recognition gained impetus with the availability of newer and larger 3D face datasets. In this section, we briefly list the most important 3D face datasets.

- **3DRMA:** A structural light based database of 120 individuals, where six images are collected in two different sessions for each individual [122]. Although used frequently in early 3D face research, the quality of acquisition is relatively low.
- **GavabDB:** Contains 427 facial meshes with pose and expression variations collected from 61 subjects [146].
- **University of Surrey, Extended M2VTS Database:** Four 3D head models of 295 subjects taken over a period of four months are included in the database [123].
- **FRGC dataset:** Collected at University of Notre Dame, version 1.0a of this dataset contains 943 near-frontal acquisitions from 277 subjects [135]. For each acquisition there is a range image and the corresponding registered 2D texture, taken under controlled indoor lighting conditions. Due to slow acquisition, the correspondence between 2D and 3D images are poor for some of the samples. Version 2.0 subsumes version 1.0a, and contains 4,007 frontal scans from 466 subjects recorded at 22 sessions with minor pose, but difficult illumination and expression variations. It comes with an infrastructure for evaluating biometrics experiments in an effort to evaluate the myriad of approaches under comparable training and testing conditions.
- **York University 3D Face Database:** Fifteen images with expression and pose variations are recorded from each of approximately 350 subjects [147].
- **MIT-CBCL Face Recognition Database:** Contains acquired and synthesized (324 images per subject) training sets for 10 subjects, and a test set of 200 images [131]. Illumination, pose and background changes are allowed.

- **BJUT-3D Large-scale Chinese Face Database:** Contains 3D laser scans of 250 male and 250 female Chinese subjects [148]. There is no hair (acquisition with a swim cap) and no facial accessories.

3.3. Preprocessing

Recognizing faces in natural or controlled environments is difficult, primarily because the illumination conditions change a lot. Other difficulties include expression changes, pose changes (including 3D rotations, scale differences and localization), background clutter, motion blur, facial aging, make-up, and face-specific changes like glasses, beard, etc. 3D acquisition bypasses some of these problems, but depending on the type of sensor, the range data may have holes and spikes (artifacts) that need filtering. The eyes and hair are problematic too, as they do not reflect the light appropriately. Illumination still has some effects on 3D acquisition, unless accurate laser scanners are employed [149]. We must also distinguish between 3D and 2.5D, i.e. depth information taken from a single viewpoint. In [150] five 2.5D images are combined to produce a true 3D model of the face, to be used in the model gallery. In the subsequent chapters of the present work, we will use the FRGC dataset, which contains a 2D image and its corresponding 2.5D range image. We will nonetheless refer to the latter as 3D information, in order to be consistent with the 3D face literature.

The preprocessing of 3D information consists of several steps. Missing points and holes can be filled by local interpolation or by making use of facial symmetry [119, 151]. However, it is a well-known fact that faces are not truly symmetrical. Gaussian smoothing and linear interpolation are used to eliminate irregularities in both texture and range images [121, 152, 125, 153, 154, 119, 155, 156]. The background clutter and hair artifacts are usually manually removed [157, 152, 154, 158, 111, 155, 128, 156].

If there is too much acquisition noise in a dataset, it will be difficult to assess the accuracy of algorithms, as the deterioration due to noise can be more pronounced than improvement due to a better algorithm. Datasets are frequently cleared of samples with very high noise levels [125, 111, 159]. Mean and median filters are also employed to reduce local noise [160, 125]. To help distance calculation after registration, the mesh representations can be regularized [161, 162], voxel discretization can be used [160], or the depth values can be resampled from a regular grid [163]. Similarly, the corresponding intensity values can be thresholded for robustness [164, 158].

3.4. Registration

Registration aims at bringing facial images into a close alignment to allow comparisons between them. For 2D images, face detection and segmentation is a preliminary step. However, depth information makes the detection much easier, and we will assume that the face is detected, and the background is absent. Most of the algorithms start by coarsely aligning the faces, either by their centres of mass [152, 128], nose tip [154, 158, 165, 119, 166, 156], the eyes [153, 164], or by fitting a plane to the face and aligning it with that of the camera [160]. Good registration of the images is important for all local similarity measures. In Chapter 6, we show that even the coarse initial alignment has a positive effect on fine registration and the subsequent classification.

Registration is about finding a transformation that maximizes the similarity between two faces. Consequently, one needs to define a similarity measure and a set of possible transformations. The similarity is usually measured by point-to-point or point-to-surface distances, or by computing cross correlation between more complex features. The transformations can be rigid, affine, elastic or liquid [167]. The rigid transformation of a 3D object involves a 3D rotation and translation, i.e. it has 6 degrees of freedom, but the nonlinearity of the problem calls for iterative methods [168]. Facial landmarks are frequently used to guide the registration process. We say more on finding facial landmarks in Chapter 5, where the automatic landmarking problem is inspected in detail, and a new method is proposed. The effect of landmarking on subsequent registration depends on the robustness of the registration algorithm. This issue is inspected in Chapter 6.

The most frequently used ([169, 170, 171, 161, 165, 111, 150, 119, 172, 128]) registration technique is the iterative closest point (ICP) algorithm, which establishes a dense correspondence between two point clouds in a rigid manner [1]. Typically, a test face is registered to each gallery face separately [169, 150, 172, 128], and a point set distance is adopted for classification. A number of variants are developed for ICP (see [173] for a review). For a review of other registration methods, especially for a number of 3D images in temporal sequence, see [174]. In the work of İrfanoğlu, an alternative and fast method was proposed to register faces [161]. An average face model (AFM) was employed to determine a single point-to-point correspondence. The gallery faces, previously used in generating the AFM, are already in dense correspondence with it. Thus, a single ICP is enough to register a test face, which is much faster for a reasonably sized gallery. This overcomes what has been reported as the major drawback of ICP [175], at a cost: Since the registration with an AFM is (hypothetically) poorer than one-to-all registration, it is expected that the method should suffer in terms of verification accuracy. We will propose a biologically motivated approach to 3D registration that builds on the AFM idea in Chapter 6.

Warping and deforming the models for better alignment helps in co-locating the landmarks. The Thin Plate Spline (TPS) algorithm [176, 177] establishes perfect correspondence between the landmarks of two registered surfaces [161, 178]. One should however keep in mind that the deformation may be detrimental to the recognition performance, as discriminatory information is lost proportional to the number of anchor points. Lu and Jain also distinguish between inter-subject and intra-subject deformations, which is found useful for classification [178].

To reduce the effect of deformations, Hutton *et al.* proposed an alternative usage of the TPS [179]. In their work, ten manually located landmarks guide a TPS deformation of the sample face to a canonical face (base mesh), followed by a resampling. Then, all points on the base mesh find a corresponding point on the sample face. The resampled face is unwarped, and returned to its original location. Thus, the registration is only used to densely label the points on the facial surface. Building on this idea, Mao *et al.* describe a method that iteratively improves the dense registration by jointly optimizing an energy function that is composed of a local feature similarity term and a global deformation term [180]. The local feature similarity is based on distance, curvature and surface normals. Tena *et al.* extend this approach event further by i) using an adaptive search in local matching, ii) taking a coarse-to-fine strategy, iii) enforcing bilateral symmetry [181]. Their results indicate that TPS-based registration can indeed be very good, provided that the guiding landmarks are correctly located.

3.5. Recognition

In this section, we look at the important 3D face recognition work in detail. We have classified each work according to the primary representation used in the recognition algorithm, much in the spirit of [149]. Although some of the proposed algorithms use more than one representation, they are listed only once, in the section where we think the contribution is most important.

3.5.1. Curvatures and Surface Features

In one of the early 3D face papers, Gordon proposed a curvature-based method for face recognition from 3D data, kept in a cylindrical coordinate system [153]. Since the curvatures involve second derivatives, they are very sensitive to noise. An adaptive Gaussian smoothing is applied so as not to destroy curvature information. The fact that skin is relatively smooth when compared to the hair and clothing is used in the segmentation process. The eyes and the nose are found, and used for alignment. The volume between two normalized surfaces is used to measure similarity.

In [182] two types of features based on the mean and Gaussian curvatures (represented with H and K, respectively) are extracted. *Ridge lines* are a set of maximal principal direction vectors, and *valley lines* are a set of minimum principal direction vectors. These are mapped to a unit sphere to produce the Extended Gaussian Image (EGI). The similarity between models and test images are found by looking at Fisher's spherical correlation coefficient between the EGI's of those. Alignment is performed by looking at the dense cells of the EGI's and determining the correlation between these cells only. The advantage of curvatures over surface normals is that they are applicable to free-form surfaces.

Moreno *et al.* extracted a number of features from 3D data, and found that curvature and line features perform better than area features [155]. Their dataset is acquired with a 3D digitizer, and the neck, ears, and hair are removed manually. Median and Gaussian filters were employed to smooth the meshes. The segmentation of the images was done with the HK-algorithm based on the signs of the median and Gaussian curvatures. Then, a threshold is applied to find regions of large curvature, and seven regions were identified based on the characteristics of the points they contain. A number of different features (including region areas, area relations, area means, center of mass distances, H and K means and variances, angles among centers of mass, line based features) are extracted. The feature extraction is followed by a simple feature selection based on discriminative power, indicated by Fisher information: the best 35 features were selected. Euclidean distance based nearest neighbour matching is used.

In [183] the mesh normals were stored as a pixel's RGB components, and the 3D information is processed like 2D. To provide robustness against expression variations, subject-specific mask are used to weight face locations. Their database was enhanced by synthetic samples with expression variations. Wang and Chua extract 3D Gabor features from the face surface to accomodate rotations in depth [184]. In the absence of good registration, template-based approaches fail under rotations as much as ±60°. The methods presented in this Section are summarized in Table 3.1.

Table 3.1. 3D face recognition systems that use surface features

Group	**Representation**	**Database**	**Algorithm**	**Notes**
Gordon [153]	curvatures	26 training 24 test	Euclidean nearest neighbour	Curvatures used for feature detection, they are sensitive to smoothing.
Tanaka *et al.* [182]	curvature based EGI	NRCC	Fisher's spherical correlation	Principal curvatures instead of surface normals for non-polyhedral objects.
Moreno *et al.* [155]	Curvature, line, region features	7 img.× 60 subj.	Euclidean nearest neighbour	Angle, distance and curvature features work better than area based features.
Wang and Chua [184]	3D Gabor features	12 img. × 80 subj.	Least Trimmed Square Hausdorff	Gabor responses are histogram normalized before comparison.
Abate *et al.* [183]	surface normals	11 img. × 133 subj.	Angular distance	Expression weighting mask for robustness against expression variations.

3.5.2. Point Clouds and Meshes

Point cloud is the most primitive 3D representation for faces, and it is difficult to work with. When the data are in point cloud representation, ICP is the most widely used registration technique. The similarity of two point sets that is calculated at each iteration of the ICP algorithm is frequently used in point cloud-based face recognizers.

Medioni and Waupotitsch present an authentication system that acquires the 3D image of the subject with two calibrated cameras [185] and ICP algorithm is used to define similarity between two face meshes. The details of the 3D reconstruction algorithm are given in Chen and Medioni [186]. Epipolar line alignment posed as a least squares optimization is followed by an image matching step that employs a normalized cross-correlation measure over a square neighbourhood as its similarity metric. Camera geometry is used for a projective reconstruction from a number of correspondences between the images.

Lu *et al.* use a hybrid-ICP based registration using Besl's method and Chen's method successively [165]. For each scan, around 100 control points are employed, mostly from around less malleable parts of the face. Shape indices are calculated for those points, and a cross correlation between shape indices of two different images is combined with the z-normalized sum-of-squares distances coming from the ICP to determine the final matching metric. Their automatic feature point selection is deemed to be robust, but sometimes it fails completely. The authors stress the importance of integrating colour and texture information with the 3D information for further performance increase.

In [162], features are extracted from around landmark points, and nearest neighbour after PCA is used for recognition. A regular mesh with fixed number of nodes is built from each point cloud. A four-triangle coarse mesh is fit by localizing the nose, and it is made finer by dividing each triangle into four new triangles at each refinement step. The final mesh contains 4^5 triangles and 545 nodes (details of the mesh construction are given in [187]). The pose parameters are obtained from the mesh, rather than from the point cloud. An average mesh is calculated, and all other meshes (and their corresponding point clouds) are aligned to it. The authors note that the z-coordinates of the aligned mesh nodes are not sufficiently informative for recognition purposes. Instead, local areas around mouth, nose, and the eyes are determined (i.e. 156 nodes are selected), and a shape vector made up of distances to neighbouring mesh nodes is derived for all nodes within those regions. The first and second Gaussian-Hermite moments are calculated for each selected node at its six neighbours [188]. PCA is employed to reduce dimensionality from 2,417 to 50, and then a nearest neighbour classifier is used.

Lu and Jain use ICP for rigid deformations, but they also propose to use TPS for intra-subject and inter-subject nonrigid deformations, with the purpose of handling expression variations [178]. Matching starts with ICP, and if the total distance for the manually located landmarks exceeds a threshold, non-rigid phase of registration is used. A SVM classifier is employed to classify the non-rigid deformations into intra-subject and inter-subject classes by looking at the concatenated displacement vector field values (See Fig. 3.3). Deformation analysis and combination with appearance based classifiers both increase the recognition accuracy [111].

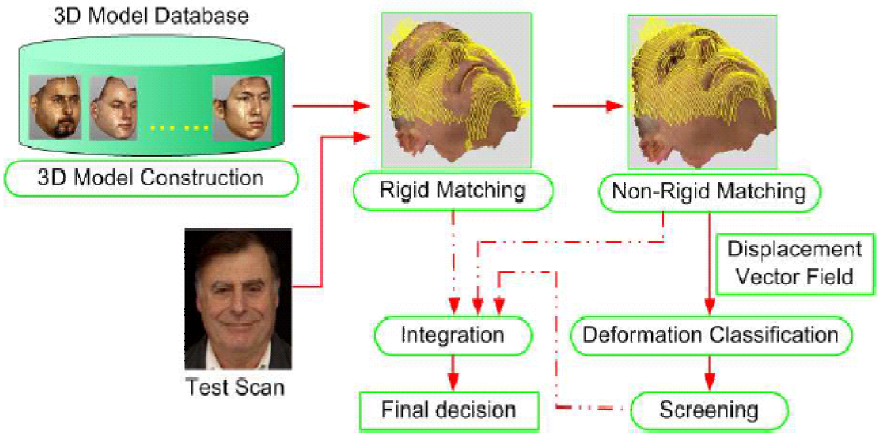


Figure 3.3. Matching with deformation analysis, proposed in Lu and Jain, 2005 [178].

Achermann and Bunke employ Hausdorff distance for matching the point clouds [160]. A 3D distance map is calculated (by producing a discrete voxel array) for each image offline to speed up the matching. This is a good idea, as the Hausdorff distance calculation for two sets with N and M points respectively is $O(NM)$, but some information is lost in thresholding. Alignment of models is achieved by first fitting a plane to the data, and then rotating the plane to align it with the focal plane of the camera, which is faster than searching through possible rotations.

In [189] a method for building a 3D facial model by employing line and curve segments and their relations in conjunction for stereo matching is proposed. The alignment is achieved by locating the irises and the center of the mouth. For iris detection, a number of rules specifying the shape and relation of the irises were determined. The reconstruction of the 3D model consists of vertices derived by fitting lines and circles to boundary segments, their arcs, and points that are sampled from boundary segments uniformly. The boundaries are represented by isoluminance lines in the thresholded intensity images (See Fig. 3.4). This work is extended in [164] to use 3D models to register the pose of 2D faces. Another extension is to generate 2D faces from the 3D model in different poses. The approach followed in the paper is to do template matching in 3D space: Shapes and models are aligned as described, and distances from each point in the matched shape to the model are found. If their distance is smaller than a threshold, it is used in average distance calculation; otherwise it is discarded as noise.

In a recent work, Dutağacı *et al.* contrast point cloud and depth images processed with signal processing methods [190]. DFT, DCT, ICA and NMF coefficients were computed on the depth images, and contrasted with DFT, ICA and NMF on 3D point clouds. The point cloud based ICA and NMF methods resulted in better accuracies, with smaller numbers of projection parameters.

It is possible to speed-up ICP based matching by an intermediate model. In [191] an annotated mesh is aligned with the test face with a deformable variant of ICP, and Haar coefficients from particular points are used for recognition. İrfanoğlu *et al.* also proposed a registration algorithm that constructs a base mesh (called an Average Face Model, or AFM for short) by Procrustes analysis and TPS warping for speeding up the registration [161]. In their work, each of the registered training and test faces will have corresponding points for all points in the AFM. The registration to the AFM

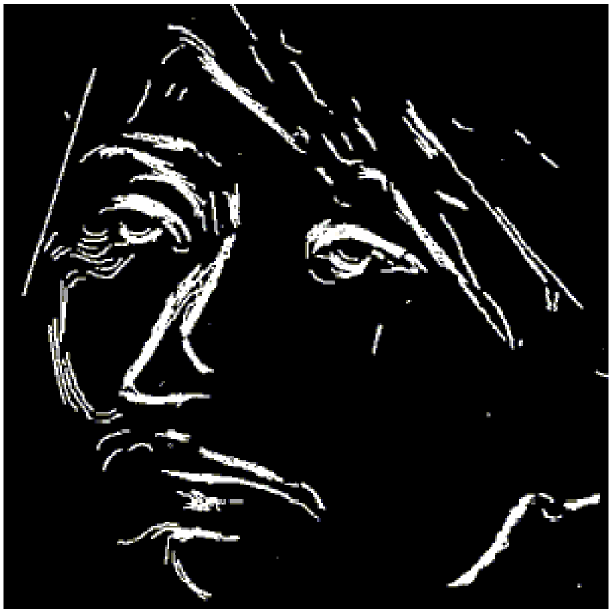


Figure 3.4. A 3D facial model reconstructed from boundary segments by 10 thresholds (from Lao *et al.*, 2000).

is achieved with ICP, and the nose tips are aligned by fine-tuning. Once the dense correspondence between all faces is established, point set difference with Euclidean norm (which approximates the volume between registered faces) is used for similiarity calculation and recognition. We extend this work in Chapter 6 by proposing a new AFM construction method, and by usign more than one AFM in the registration process [163]. The methods described in this Section are summarized in Table 3.2.

3.5.3. Depth Map

Depth maps are generally used in conjunction with subspace methods, although most of the existing 2D techniques are suitable for processing the depth maps. The baseline technique for this representation is the 3D version of the eigenface method [91], which is a PCA projection to some suitable dimensionality, and nearest neighbour-based matching. The depth map construction consists of selecting a viewpoint, and smoothing the sampled depth values.

Table 3.2. 3D face recognition systems that use point clouds and meshes

Group	**Representation**	**Database**	**Algorithm**	**Notes**
Achermann and Bunke [160]	point cloud	120 training 120 test	Hausdorff nearest neighbour	Hausdorff distance calculation can be speeded up by voxel discretization.
Lao *et al.* [164]	curve segments	36 img. × 10 subj.	Euclidean nearest neighbour	Points with bad correspondence are not used in distance calculation.
Medioni *et al.* [185]	mesh	7 img. × 100 subj.	normalized cross-correlation	After alignment, a distance map is found. Statistics of the map are used in similarity calculation.
İrfanoğlu *et al.* [161]	point cloud	3DRMA	Point set difference (PSD)	ICP used to align point clouds with a base mesh. PSD outperforms PCA on depth map.
Lu *et al.* [165]	mesh	90 training 113 test	hybrid ICP and cross-correlation	ICP distances and shape index based correlation can be usefully combined.
Xu *et al.* [162]	regular mesh	3DRMA	Feature extraction, PCA + NN	Feature derivation + PCA around landmarks worked better than aligned mesh distances.
Lu and Jain [178]	deformation points	500 training 196 test	ICP + TPS, nearest neighbour	Distinguishing between inter-subject and intra-subject deformations helps recognition.
Passalis *et al.* [191]	mesh	FRGC ver.2	ICP variant + Haar wavelets	Most errors are due to poor registration.
Dutağacı *et al.* [190]	point cloud, depth image, voxel	3DRMA	DCT, DFT, ICA, NMF	Point cloud + ICA or NMF works best.

In [154], the range images were derived from 3D meshes. PCA and ICA were compared on the depth maps with nearest neighbour classification. Since subspace projection methods are sensitive to registration and clutter, the nose tip was used in alignment, an elliptical mask is used to discard the possibly noisy periphery, and the missing pixels are filled with linear interpolation. ICA was found to perform better, but PCA degraded more gracefully with declining numbers of training samples. Srivastava *et al.* extend this work with expression variations in the database and a new subspace projection method [156]. In their experiments, clutter is removed manually, and a fixed mask (there is almost no variation in scale, so this method works) is employed to crop images. Holes are patched through interpolation. The range images are transformed to a subspace which optimizes the recognition performance on the training set. The Grassman manifold $\mathcal{G}_{n,k}$, which is the set of all k-dimensional subspaces of $\mathbb{R}^n$, is searched with a MCMC simulated annealing algorithm for the optimal linear subspace. Classification is either with nearest neighbour or 2-class SVM with polynomial kernels on the subspace. The optimal subspace method performs better than PCA, LDA or ICA.

Achermann *et al.* compare an eigenface method with a 5-state left-right HMM on a database of depth maps [121]. For the HMM, a window is passed over the image in overlapping steps, and the states of the HMM stand for forehead, eyes, nose, mouth and chin. They show that the eigenface method outperforms the HMM, and the Gaussian smoothing used in depth map construction effects the eigenface method positively, while its effect on the HMM is detrimental.

3D data are usually more suitable than 2D data for alignment, and should be preferred if available. In Lee *et al.* the 3D recognition algorithm starts by finding the nose tip, as the maximum point in the face [158]. This is not a very robust approach, as we show in Chapter 6. Then, the faces are aligned by panning, rotating and tilting the 3D model, in that order. The background is removed manually. A depth value is selected to threshold the 3D image to produce a 2D binary image. 5 × 5 windows are resampled from the area around the nose, and the means and variances of the depth values within those windows are used as features. Table 3.3 summarizes the methods based on the depth map.

Table 3.3. 3D face recognition systems that use depth map

Group	**Representation**	**Database**	**Algorithm**	**Notes**
Achermann *et al.* [121]	depth map	120 training 120 test	eigenface vs. HMM	Eigenface outperforms HMM. Smoothing is good for eigenface, bad for HMM.
Hesher *et al.* [154]	mesh	FSU	ICA or PCA + nearest neighbour	ICA outperforms PCA, PCA degrades more gracefully as training samples are decreased.
Lee *et al.* [158]	depth map	2 img. × 35 subj.	feat. extraction+ nearest neighbour	Mean and variance of depth from windows around the nose are used as features.
Srivastava *et al.* [156]	depth map	6 img. × 67 subj.	subspace projection + SVM	Optimal subspace found with MCMC simulated annealing outperforms PCA, ICA and LDA.

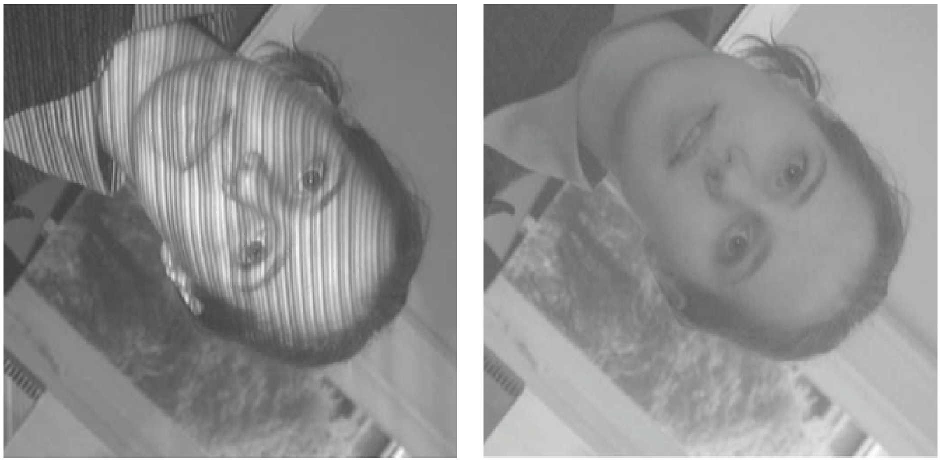
3.5.4. Profile

The most important issue for profile-based schemes is the extraction of the profile. In an early paper Cartoux *et al.* use an iterative scheme to find the symmetry plane that cuts the face into two similar parts [192]. The nose tip and a second point (the nasion) are used to extract the profiles. Nagamine *et al.* use various heuristics to find feature points and align the faces by looking at the symmetry [193]. Then the faces are intersected with different kinds of planes (vertical, horizontal or cylindrical around the nose tip), and the intersection curve is used in recognition. Vertical planes around ± 20mm. of the central region and selecting a cylinder with $20-30$mm. radius around the nose (crossing the inner corners of the eyes) produced the best results. In [122], Beumier and Acheroy detail the acquisition of the popular 3DRMA dataset with structural light (See Fig. 3.5) and report profile based recognition results. In addition to the central profile, they use the average of two lateral profiles in recognition. The profiles are aligned using the ICM (Iterative Conditional Mode) optimization procedure.

Once the profiles are obtained, there are several ways of matching them. In [192], corresponding points of two profiles are selected to maximize a matching coefficient that uses the curvature on the profile curve. Then a correlation coefficient and the mean quadratic distance is calculated between the coordinates of the aligned profile curves, as two alternative measures. In [122], the area between the profile curves is used. In [194] distances calculated with L_1 norm, L_2 norm, and generalized Hausdorff distance were compared for aligned profiles, and the L_1 norm is found to perform better.

Table 3.4. 3D face recognition systems that use profiles

Group	**Representation**	**Database**	**Algorithm**	**Notes**
Cartoux *et al.* [192]	profile	3/4 img. × 5 subj.	curvature based nearest neighbour	High quality images needed for principal curvatures.
Nagamine *et al.* [193]	vertical, horiz., circular profiles	10 img. × 16 subj.	Euclidean nearest neighbour	Central vertical profile and circular sections touching eye corners are most informative.
Beumier and Acheroy [122]	vertical profiles	3DRMA	area based nearest neighbour	Central profile and mean lateral profiles are fused by averaging.



(*a*) (*b*)

Figure 3.5. (a) An image with projected structural light and (b) its grayscale counterpart (from Beumier and Acheroy, 2001).

3.5.5. Analysis by Synthesis

This section deals with approaches that use 3D models to remove pose and illumination effects from 2D images. Recognition is performed with 2D methods, after several stages of correction. The interaction between 2D and 3D is more difficult to achieve when the input is only 2D. Registration is possible by synthesizing 2D images from 3D models, but it is time-consuming.

Blanz and Vetter use a generic 3D model as an intermediate representation to guide their analysis-by-synthesis approach [157, 195]. Their ultimate aim is to recognize 2D images. The 3D model is used to render synthetic images, which are used to determine the pose and illumination conditions for the 2D image. In [157], the analysis-by-synthesis approach that uses morphable models is detailed. A morphable model is defined as a convex combination of shape and texture vectors of a number of samples that are placed in dense correspondence. A single 3D face model is used to render an image similar to the test image, which leads to the estimation of viewpoint parameters (pose angles, 3D translation, focal length of the camera), illumination parameters (ambient and directed light intensities, direction angles of the light, colour contrast, gains and offsets of the colour channels), and deformation parameters (shape and texture).

The morphing of the model to fit the sample is achieved by locating seven landmark points and finding the parameters that are involved in the convex combination for the best fit by a stochastic Newton method. With this (rather computationally intensive) method, it is possible to estimate the 3D parameters for the face from a single 2D image.

A similar system is proposed by Lee and Ranganath, where a generic 3D model combines edge, colour and location information [196]. Zhao and Chellappa propose a method where a generic 3D model is used in a shape-from-shading approach with the same purpose [197].

In [119] a system is proposed to work with 2D colour images and corresponding 3D depth maps. The HISCORE system of [118] is employed to record 2,200 images of 20 persons, which contain different facial expressions, different illumination conditions, pose variations (±20°), images with or without glasses, and frontal images. Given a pair of 2D and 3D images, the algorithm synthesizes a pose and illumination corrected pair of images. The nose-nose ridge line and the symmetry axis around the nose is used for pose correction. In the verification setting, the pose corrected face is warped and aligned with the gallery face of the claimed person using ICP. The missing pixels from the depth map are found from symmetry and linear interpolation. For illumination correction, a number of single-light sources are simulated and low-dimensional projections of rendered images are used for computing the light source direction with a SVM-based regressor. Embedded Hidden Markov Models are used for colour images and depth images separately for baseline classification, and a combined similarity measure was obtained using the product rule. In all simulations, the depth images performed significantly better (by 4-7 per cent) than colour images, and the combination increased the accuracy as well (by 1-2 per cent). Pose correction is found to be more important than illumination correction.

In [198], a morphable model is used to recover 3D information from a 2D image. The illumination is assumed to be Lambertian, and the basis images for illumination space are found by spherical harmonics. During testing, faces are assigned the class for which there exists a weighted combination of basis images that is the closest to the test face image. Illumination invariance is also aimed in the method proposed by Weyrauch *et al.*, which combines morphable models with the component-based recognition paradigm [131]. A large number of faces are synthesized from the models to train a SVM classifier. The recognition starts with hierarchical component-based face detection based on linear and polynomial SVMs. The localized components are used for feature extraction by 14 different SVM classifiers to which a global component was added, and the outputs of these classifiers are fed to a geometrical SVM classifier that performs the final classification (See Fig. 3.6).

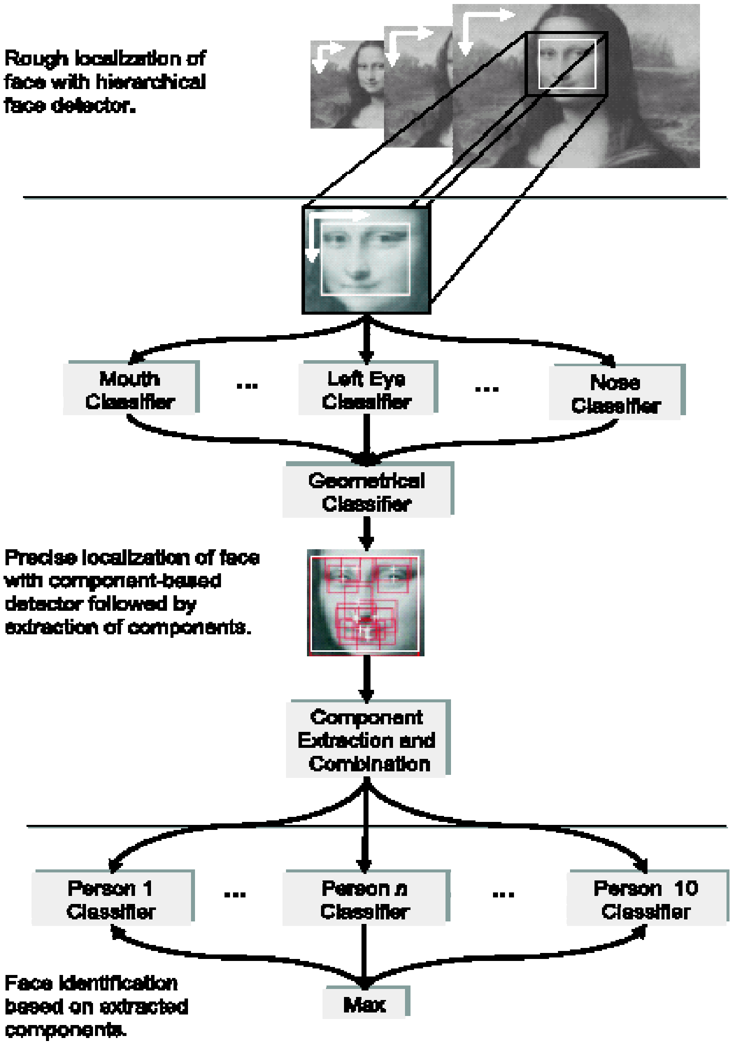

Rough localization of face with hierarchical face detector.
Mouth Classifier
...
Left Eye Classifier
...
Nose Classifier
Geometrical Classifier
Precise localization of face with component-based detector followed by extraction of components.
Component Extraction and Combination
Person 1 Classifier
...
Person n Classifier
...
Person 10 Classifier
Face identification based on extracted components.
Max

Figure 3.6. The component based face recognition system proposed in Weyrauch *et al.*, 2004.

Table 3.5. 2D face recognition systems that use 3D via analysis by synthesis

Group	**Representation**	**Database**	**Algorithm**	**Notes**
Blanz and Vetter [157]	2D + viewpoint parameters	CMU-PIE, FERET	analysis by synthesis	Using a generic 3D model, 2D viewpoint parameters are found.
Zhao and Chel-lappa [197]	illumination corrected 2D	FERET, Yale, Weizmann	Albedo & pose recovery + LDA	Prototype images are synthesized with the help of 3D.
Malassiotis & Strinzis [119]	texture + depth map	110 img. × 20 subj.	embedded HMM + fusion	Depth is better than colour, fusion is best. Prefer pose to illumination for correction.
Zhang and Samaras [198]	illumination basis images	CMU-PIE	morphable model+ spherical harmonics	Statistical illumination model helps removing illumination effects, even with a single training sample.
Lee and Ranganath [196]	edge + colour + mesh	44 img.× 15 subj.	Euclidean distance, PCA	Pose and illumination is estimated from edge and colour, respectively.
Weyrauch *et al.* [131]	2D patches	20 img. × 10 subj.	morphable model for synthesis + component SVM + fusion	Components are more robust to rotations, and histogram equalization helps for illumination invariance.

3.5.6. Combinations of Representations

Most of the work that uses 3D face data use a combination of representations. The enriched variety of features, when combined with classifiers with different statistical properties, produce more accurate and more robust results. In [133], surface normals and intensities are concatenated to form a single feature vector, and the dimensionality is reduced with PCA. Adding perturbed versions of training images reduces sensitivity of PCA. In [159], the 3D data are described by point signatures, and the 2D data by Gabor wavelet responses, respectively. 3D information may have missing elements around the eyes and the eyebrows, and the mouth area is sensitive to expressions. These are omitted for robustness. The combination of point signatures and Gabor features is achieved by concatenating the PCA-transformed feature vectors, and applying a similarity function. A Decision Directed Acyclic Graph (DDAG), which transforms the multi-class classification to a number of two-class decisions, is used in conjunction with a Support Vector classifier.

3D intensities and texture were combined to form the 4D representation in [128]. An ellipsoidal region on each face is extracted manually, and translated to align the centres of mass. ICP on the weighted 4D data are used for registration of the faces. The simulations show that the 3D information helps recognition performance for the profile views considerably, and the texture information is added with a preferably small weight (the weight coefficient for the Euclidean distance is in the range [0.01,0.1]). The fusion of 3D and texture information increases the accuracy for appropriate pose and expression variations. Smiling faces are better recognized with the addition of texture information, but frowning faces, tilted faces, and the profile views suffer.

Bronstein *et al.* point to the non-rigid nature of the face, and to the necessity of using a suitable similarity metric that takes this deformability into account [152]. A *bending-invariant canoncial form* is obtained by multi-dimensional scaling (MDS) to a three-dimensional subspace. The transformed representations are aligned by carrying the centre of mass to the origin (translation) and by setting the second moments to zero (rotation). The texture information, which is in one-to-one correspondence with the depth information, is flattened by projection to two dimensions. This flattened texture and the canonical image (see Fig. 3.7) are mapped to their respective eigenspaces, and the resulting vectors are concatenated to obtain an *eigenform.* The proposed classification method is nearest neighbour with Euclidean distance on these eigenforms. Apart from techniques that fuse the representations at the feature level, there are a number of systems that employ combination at the decision level. Chang *et al.* propose in [125] to use Mahalanobis distance-based nearest-neighbor classifiers on the

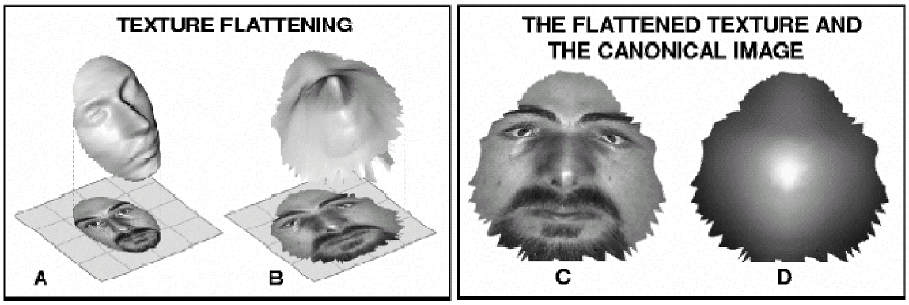


Figure 3.7. Texture flattening on the facial surface (A) and on the canonical surface (B). The flattened texture (C) and the canonical image (D) are used for recognition (from Bronstein *et al.*, 2003).

2D intensity and 3D range images separately, and to fuse the decisions with a rank-based approach at the decision level. Tsalakanidou *et al.* introduce a method, where the depth map and colour maps (one for each YUV channel) are projected via PCA and the distances in four subspaces are combined by multiplication [199]. In [151] the 3D data are segmented as head and torso with a mixture of two Gaussians. Eyes are used in face localization by making use of the symmetry of the face. For classification, embedded hidden Markov models (EHMM) are used, one for the depth maps, and one for intensity images, respectively. The scores obtained from the EHMMs (S_d for depth map and S_i for intensity) are normalized to the range $[0, 1]$. The combined scores are calculated with the following formula [200]:

$$S_{id} = w_i * log(S_i + 1) + w_d * log(S_d + 1) \tag{3.1}$$

The registration of the images is achieved by manually locating the centres of the eyes and the chin, producing the depth map, warping the depth map to aling the determined points, and patching the holes on the depth map by linear interpolation and with the help of symmetrical correspondences. Their experiments indicate that the depth data are not as discriminatory as the intensity data.

Lu and Jain combine texture (LDA) and surface (point-to-plane distance) with weighted sum rule [111, 150]. Coarse alignment is performed via manually located landmark points, and ICP is used for fine alignment, as in [165]. For the appearance based test sample classification, additional samples are synthesized from the 3D model, and LDA is performed on a subset of possible target classes (called *constrained appearance–based matching*). A point-to-plane distance metric is used for surface matching. The different scores are combined with weighted voting, but only the difficult samples are classified via the combined system (See Fig. 3.8 for the recognition scheme overview).

In [140] texture and depth maps are fused in the feature level for classification with LDA, but a decision-level fusion scheme that relies on local feature analysis was found to be more successful. In [175] hierarchical graph matching is applied on 2D and 3D separately. A score-level fusion was shown to be of marginal use. In [194] a number

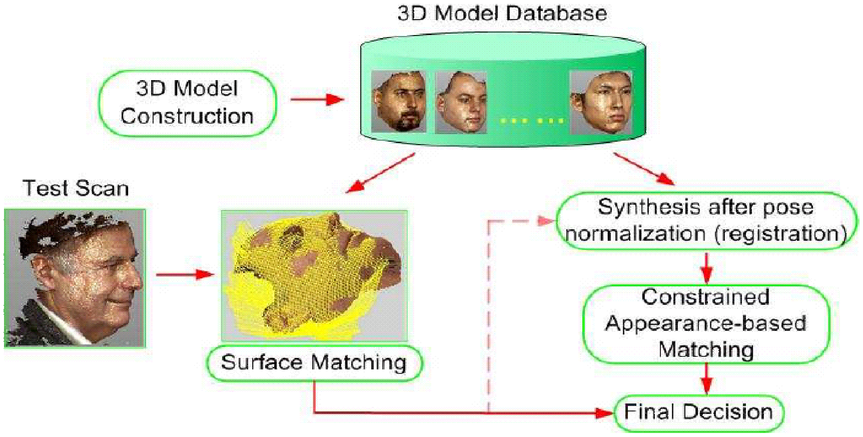
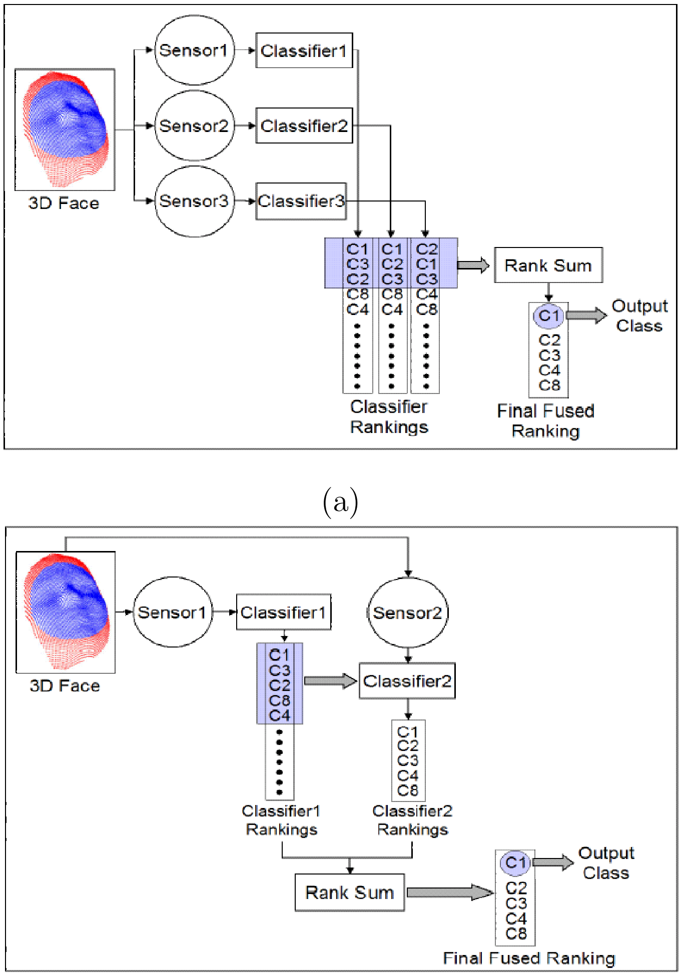

3D Model Database
3D Model Construction
Test Scan
Synthesis after pose normalization (registration)
Constrained Appearance-based Matching
Surface Matching
Final Decision

Figure 3.8. Shape and appearance based face recognition system proposed in Lu and Jain, 2005 [111].

of representations are derived from the 3D range data, and the extracted features are used in separate nearest neighbour classifiers that are fused with a rank-based decision level combination scheme. The following features are used for each face: a point cloud, surface normals at each point, the central and six lateral facial profile (three on either side), a PCA projection of the depth map, and an LDA projection of the depth map. Two classifier combination schemes are proposed (See Fig. 3.9). In the parallel scheme the ranked outputs of individual classifiers are combined with consensus voting, rank-sum, nonlinear rank-sum and highest-rank majority rule. In the hierarchical scheme, the best classes found by one classifier are used to restrict the search space of the second classifier. On a 106 class subset of the 3D_RMA dataset, the best single classifier is the LDA projection of the depth map (96.27 per cent accuracy). The hierarchical fusion of point cloud and LDA produces 98.13 per cent, whereas the parallel fusion of surface-normal, Depth-LDA, and profile-based classifiers through nonlinear rank-sum results in 99.07 per cent accuracy.

In their later work [171], Gökberk *et al.* extend their experiments by employing more base classifiers (point clouds, surface normals, shape index values, facial profiles, LDA of depth images, and LDA of surface normals), and by employing different fusion schemes (sum/product rules, consensus voting, Borda count method, improved consensus voting, and highest confidence rule) for both identification and verification



(a)

(b)

Figure 3.9. (a) Parallel and (b) hierarchical fusion of classifiers for 3D face recognition, proposed in Gökberk *et al.*, 2005.

scenarios. They also propose a confidence-assisted serial fusion scheme, where the second classifier is consulted only if the confidence of the first classifier is below a certain threshold. This scheme is significantly faster than forwarding the nearest classes found by the first classifier.

Profiles are used in conjunction with other features in several other papers. In [201], 3D central and lateral profiles, gray level central and lateral profiles were evaluated separately, and then fused with Fisher's method. Fusion generally increases accuracy, but here, the authors also tested combining data from several acquisitions (temporal fusion), which was even better. In [166] a surface-based recognizer and a profile-based recognizer are combined at the decision level. Surface-matcher's similarity is based on a point cloud distance approach, and profile similarity is calculated using Hausdorff distance. In [172], a number of methods are tested on the depth map (Eigenface, Fisherface, and kernel Fisherface), and the depth map expert is fused with three profile experts with Max, Min, Sum, Product, Median and Majority Vote rules, out of which the Sum rule was selected.

3D representations have different strengths and weaknesses. The point cloud is easy to obtain, useful in dense registration, but cumbersome. Meshes are more structured, but require fast and good triangulation. A mesh is also useful in energy-based guiding of facial deformations. Depth maps are straightforward representations, which enable the use of many 2D methods, yet they are sensitive to pose and scale variations. The representation(s) used by any application must necessarily be selected according to the computational cost and accuracy requirements of the system. There are several algorithms developed for each of the representations, and some that use novel representations. There is a significant effort to find representations that are complementary, and fuse well. Table 3.6 gives a summary of methods detailed in this Section.

Table 3.6. 2D face recognition systems that use combinations of representations

Group	**Representation**	**Database**	**Algorithm**	**Notes**
Tsutsumi *et al.* [133]	texture +depth map	35 img.× 24 subj.	concatenated features + PCA	Adding perturbed versions of training images reduces sensitivity of PCA.
Beumier and Acheroy [201]	2D and 3D vertical profiles	3DRMA	nearest neighbour + fusion	Combination of 2D and 3D helps. Temporal fusion (snapshots taken in time) helps too.
Wang [159]	Gabor features+ point signatures	6 img.× 50 subj.	PCA+SVM	Accurately located points are assigned greater weights in recognition.
Bronstein *et al.* [152]	texture+depth map	157 subj.	concatenated PCA + nearest neighbour	Bending-invariant canonical representation is robust to facial expressions.
Chang *et al.* [125]	texture+depth map	278 train, 166 test	Mahalanobis based near.neigh.+fusion	Pose correction through 3D is not better than rotation-corrected 2D.
Pan *et al.* [166]	profile + point cloud	3DRMA	ICP+Hausdorff + fusion	Surface and profile combined usefully. Discard worst points during registration.
Tsalakanidou *et al.* [199]	texture+depth map	XM2VTS	nearest neighbour+fusion	Fusion of frontal colour and depth images with colour faces from profile.
Papatheodorou *et al.* [128]	dense mesh+ texture	12 img.× 62 subj.	nearest neighbour+ fusion	3D helps 2D especially for profile views. Texture has small relative weight.
Tsalakanidou *et al.* [151]	texture+depth map	60 img.× 50 subj.	embedded HMM+ fusion	Processed texture is more informative than warped depth maps.
Benabdelkader *et al.* [140]	texture+depth map	4 img.× 185 subj.	Faceit vs. LDA	Decision-level fusion is better than feature level fusion.
Gökberk *et al.* [194]	surface normals, profiles,depth map, point cloud	3DRMA	PCA, LDA, nearest neighbour, rank based fusion	Best single classifier is depth-LDA. Combining it with surface normals and profiles increases accuracy.
Hüsken *et al.* [175]	texture & shape models	FRGC	Hierarchical graph matching	Independence of scores is important in fusion.
Pan and Wu [172]	depth map + profile	6 img.× 120 subj.	kernel Fisherface+ Eigenface+fusion	Sum rule is preferred to max, min, product, median and majority vote for fusion.
Lu and Jain [150]	mesh+texture	598 test scans	ICP(3D), LDA(2D) + fusion	Difficult samples are evaluated by the combined scheme.
Gökberk *et al.* [171]	surface normals, profiles, point clouds, shape indices, LDA	3DRMA	confidence-assisted serial fusion	Selecting only the most confident classifier is as good as fusing all base classifiers.

3.6. A Biologically Motivated Framework for 3D Face Recognition

For future directions in 3D face recognition research, Bowyer *et al.* remark that 3D data acquisition is far from perfect, and call for better sensors and better algorithms that handle expression variations more robustly [149]. They also stress the importance that there is a need for larger, more challenging datasets and more rigorous experimental evaluation methodology. The corpus prepared for the face recognition grand challenge (FRGC) contains about 4000 3D scans, which partly alleviates this need [135].

Research into 3D face recognition, either standalone or as a means of supporting the more thoroughly researched 2D approaches, has intensified in the past ten years. The FRGC call puts emphasis on the following research questions that are relevant to 3D researchers:

- **Development of 3D face recognition algorithms**
- Comparison of human and machine performance, psycho-physics
- Performance analysis and statistics
- Morphable models
- Estimation of 3D structure from 2D images
- **Face and eye localization**
- **Sex, pose, age, etc. estimation**

The present work contributes to several items on this list, indicated in boldface. Based on the survey of human face recognition studies, we set a course with the following considerations for a general 2D-3D face recognition system:

- **Object-centered representation:** A uniform, object-centered representation facilitates subsequent recognition greatly. This approach is in accordance with biological models. It is postulated that the superior colliculus of the human brain is responsible for bringing sensory information from different modalities into a single coordinate frame to drive motor actions [202]. Bringing faces to the center of a coordinate system requires the detection of the face area and the localization of a few landmark points on facial images. As it should be apparent from the literature survey presented in this chapter, there is a lot of manual intervention in the 3D literature, particularly in the early stages of the systems. Clutter removal and landmarking are rarely fully-automatic. We have identified this as the major weakness of the present approaches, and focused on automatical landmarking and registration in particular. The concept of feature saliency is instrumental in our approach to landmark localization. For a uniform treatment of different landmark locations, we will use an unsupervised learning approach instead of heuristics. A powerful unsupervised learning algorithm will be developed in Chapter 4, and applied to landmark localization in Chapter 5.

- **Analytic and holistic information:** As it is evident from the psychophysical studies, feature-based systems and holistic systems have different strengths. On the one hand, two separate subsystems may work in conjunction, making use of different types of information for mutual benefit. On the other hand, the computational requirements for such a scheme is higher. This issue needs to be explored for particular settings.
- **Feature selection:** It is difficult to guide the selection of features according to biological considerations, as the representation of depth in the brain is not very well known. For intensity information, Gabor wavelets have long been used as biologically plausible features. 3D features can be selected according to their robustness, cost and informativeness.
- **Illumination:** Texture information can be used after illumination correction. The analysis-by-synthesis methods are very succesful for pose and illumination estimation and correction. However, they are very costly in terms of computation time. In Chapter 5 we explore a recent method that uses 3D information to correct 2D illumination.
- **Structural information:** Structural information should be used in registering and recognizing faces. There is a strong innate knowledge of facial structure in humans. We propose a novel algorithm for structural analysis of (facial) landmarks in Section 5.6.
- **Hierarchical classification:** The recognition of faces proceeds in a hierarchical manner. The "other-race effect" suggests that using an entry-level system for

gender and apparent race discrimination, combined with separate subsystems, may be beneficial. Special care should be given to the ethical considerations involving the apparent race. The registration approach proposed in Chapter 6 is based on this idea. Apart from exploring the speed-accuracy trade-off, it also helps recovering meta-data from the faces, like gender and morphological group of a face.

- **Scalability:** It is desirable for a real-time system that new classes can be added incrementally, and the final stored representations are as small as possible. Scalability directly relates to the usefulness of the system. This property suggests using a generative model or template matching instead of a discriminative model. For the classification problem, we take the template matching approach in Chapter 6, as our experimental setting allows very few samples per class.
- **Robustness:** Irregularities, noise and occlusions should be taken into account. The existence of a number of subpaths to recognition will allow the system to disregard noisy channels, or to diminish their effect. In the human brain, sensory information from different modalities is fused with multimodal neurons that reach their highest activation levels when the information content of each individual modality is poor [203]. Classifier fusion seems to be an essential component of successful face recognition systems [204].

Putting these considerations into a complete system (as depicted in Fig. 3.10) results in a complex framework that contains a large number of subsystems, touching upon almost all the problems of face recognition. The first part is *Preprocessing*, which includes landmark localization, registration, smoothing, illumination and affine corrections. The resulting representations are fed to the recognition subsystems. Only one such system is shown in Fig. 3.10, denoted with *Recognition Subsystem i.* The meta classifier judges apparent race and gender. All classifier outputs are fused to produce a ranked recognition result.

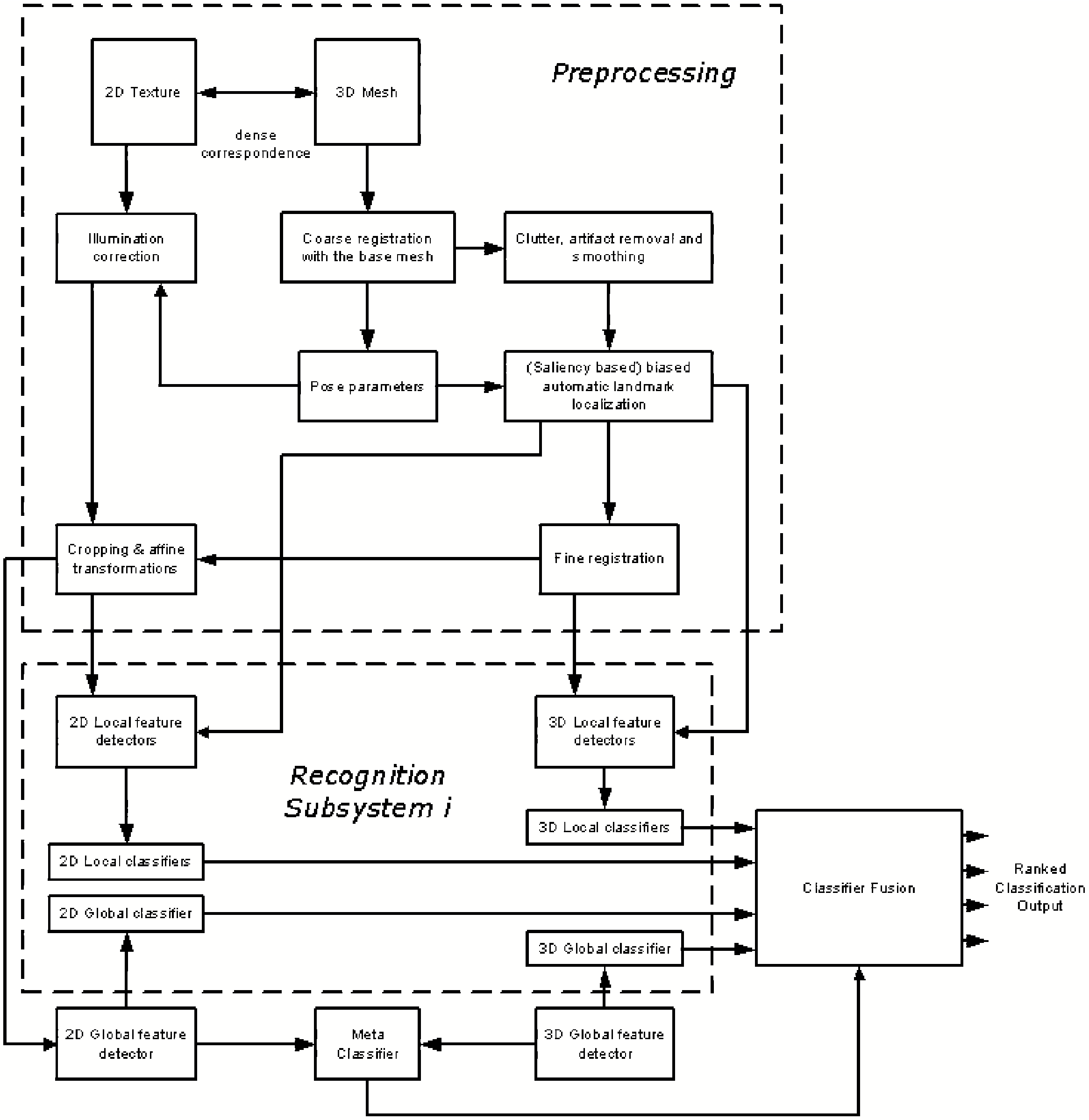

Preprocessing
2D Texture
3D Mesh
dense correspondence
Illumination correction
Coarse registration with the base mesh
Clutter, artifact removal and smoothing
Pose parameters
(Saliency based) biased automatic landmark localization
Cropping & affine transformations
Fine registration
2D Local feature detectors
3D Local feature detectors
Recognition Subsystem i
3D Local classifiers
2D Local classifiers
2D Global classifier
3D Global classifier
Classifier Fusion
Ranked Classification Output
2D Global feature detector
Meta Classifier
3D Global feature detector

Figure 3.10. The 2D + 3D face recognition model. See text for the details.

4. UNSUPERVISED FEATURE LEARNING

On they went, raising muffled echoes, and Trurl looked and grimaced, as did Klapaucius, for though there was plenty of authentic and top-quality information lying about, wherever the eye fell was nothing but must, dust and clutter.

Stanislaw Lem

We will face the problem of local feature learning several times in the remaining part of this work. We have used a particular algorithm for statistical feature modeling, called Incremental Mixtures of Factor Analyzers (IMoFA), which we detail in this chapter. This algorithm is used successfully in automatical localization of facial features, which is the topic of the next chapter. The IMoFA algorithm is not specific to the problem of face or facial feature recognition. In this chapter, we explain the workings of the algorithm, and justify its steps on different pattern recognition tasks.

4.1. Basics

In probabilistic modeling of data, the complexity of data distribution often does not allow accurate modeling with a single probability density expression. It will then be useful to conceptualize the data as made up of groups (or clusters), each generated by a different process. The overall probability model will then be a mixture of densities of simpler nature.

A *mixture model* is written as

$$p(\boldsymbol{x}) = \sum_{j=1}^{J} p(\boldsymbol{x}|\mathcal{G}_j)P(\mathcal{G}_j) \tag{4.1}$$

where $\mathcal{G}_j$ stand for the components, $P(\mathcal{G}_j)$ is the prior probability, and $p(\boldsymbol{x}|\mathcal{G}_j)$ is the probability that the data point is generated by component j.

In classification, each class-conditional density is written as a separate mixture model and the input is then a *mixture of mixtures* (MoM):

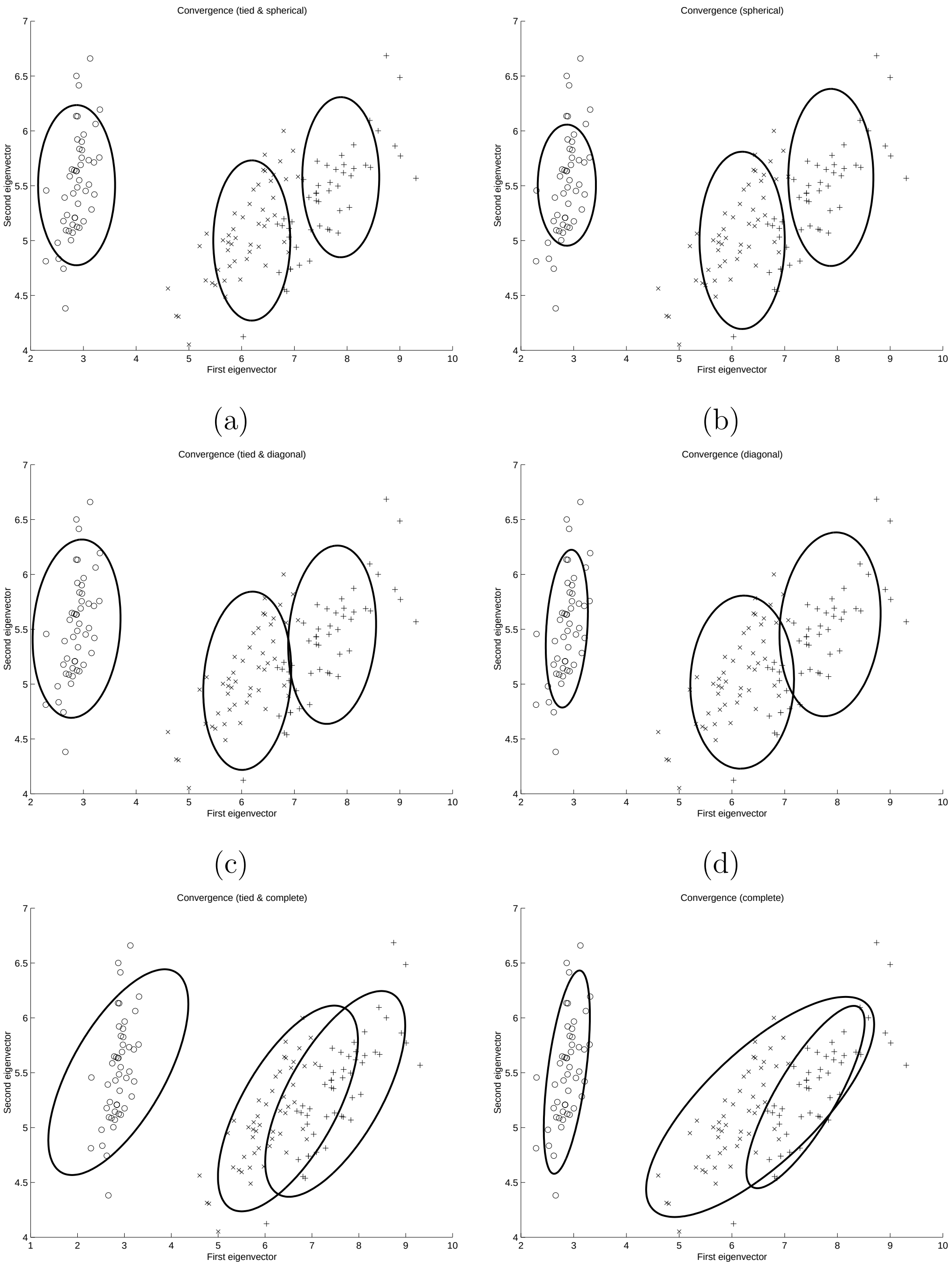
$$p(\boldsymbol{x}|\mathcal{C}_i) = \sum_{j=1}^{J_i} p(\boldsymbol{x}|\mathcal{G}_{ij})P(\mathcal{G}_{ij}) \tag{4.2}$$

$$p(\boldsymbol{x}) = \sum_{i=1}^{K} p(\boldsymbol{x}|\mathcal{C}_i)P(\mathcal{C}_i) \tag{4.3}$$

During testing, all $p(x|\mathcal{C}_i)$ are calculated and the class with the highest posterior is chosen using Bayes' rule, where the posterior is defined as $P(\mathcal{C}_i|x) = P(\mathcal{C}_i)p(\boldsymbol{x}|\mathcal{C}_i)/p(\boldsymbol{x})$.

In a *mixture of Gaussians* (MoG), each component is a Gaussian: $p(x|\mathcal{G}_j) \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Using complete covariance matrices $\boldsymbol{\Sigma}_j$ means that the number of parameters scales quadratically with the number of dimensions, and this may cause overfitting. With tied covariances (i.e. shared across components), the number of parameters decreases, but this model assumes the same input distribution in different components and may be a source of bias. Another approach is to constrain the covariances to be diagonal or spherical, but these discard valuable correlation information. When all combinations are considered, the MoG model offers six different covariance shapes: Tied spherical, spherical, tied diagonal, diagonal, tied, and unrestricted, respectively. Fig. 4.1 demonstrates all possible covariance shape selections for a three-component MoG solution on the simple Iris dataset with three classes, four dimensions, and 150 samples. The mixture plots are projected onto the first two principal components.

Decreasing the number of parameters while still modeling the covariances is possible with a *factor analysis* (FA) model. While MoG models span the parameter space unevenly, factor analysers are more flexible and allow a better control over the number of parameters. In FA, we assume that a small number of low-dimensional latent variables (factors) $\boldsymbol{z}$ cause the correlation in component j. The p-dimensional latent variable is carried over to the d-dimensional space by multiplying it with a $p \times d$-dimensional *factor loading matrix* $\boldsymbol{\Lambda}$. Furthermore, we assume the existence of additional isotropic



(a) (b)

(c) (d)

(e) (f)

Figure 4.1. MoG models with different covariance shapes on the Iris dataset. The ranges of x and y axes are different, therefore the spherical covariances that are circles in reality are shown as ovals. (a)tied & spherical (b) spherical (c) tied & diagonal (d) diagonal (e) tied & complete (f) complete.

sensor noise $\boldsymbol{\epsilon}$ in the d-dimensional space. This conveniently breaks the singularity that will result from modeling the high-dimensional data in the low-dimensional manifold.

If we denote the mean of the data distribution $\boldsymbol{x}$ with $\boldsymbol{\mu}$, the factor analysis equation for component j is written as:

$$\boldsymbol{x} - \boldsymbol{\mu}_j = \boldsymbol{\Lambda}_j \boldsymbol{z} + \boldsymbol{\epsilon}_j \tag{4.4}$$

where the covariance in the data manifold is expressed as $\boldsymbol{\Sigma}_j = \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^T + \Psi$. In this equation, $\boldsymbol{\Lambda}_j$ is the factor loading matrix for component j, and shows the dependence of data points to each factor. $\boldsymbol{\epsilon}_j$ is the Gaussian noise and is assumed to be distributed $\mathcal{N}(0, \Psi)$, where Ψ is a diagonal matrix, interpreted as sensor noise common to all components. When $\boldsymbol{x}$ is d-dimensional, $\boldsymbol{\Sigma}_j$ is $d \times d$, whereas with $p < d$ factors, $\boldsymbol{\Lambda}_j$ is $d \times p$. By indexing with j, we indicate that there are multiple components, and we have a *mixture of factor analyzers* (MoFA). The factor analysis model is depicted in Fig. 4.2.

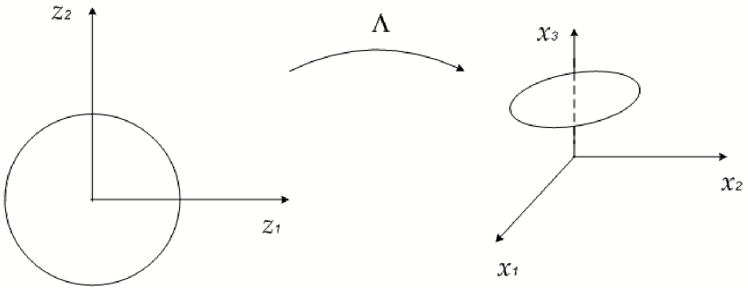


Figure 4.2. Factor analysis model, adapted from [205].

Given a training set, the maximum likelihood estimates can be calculated using the Expectation-Maximization (EM) algorithm, which simultaneously places the components in the input space ($\boldsymbol{\mu}_j$) and also finds the factors in each component, performing dimensionality reduction in each component ($\boldsymbol{\Lambda}_j$) [206, 207, 208]. However, this requires that the number of components and the factors in each component be specified in advance. The IMoFA model we use is an *incremental algorithm* where components and factors are added iteratively as needed, without requiring them to be specified in advance, thereby better matching the complexity of the mixture model to that of the data.

4.2. EM for a General Mixture of Factor Analysers

In this section, we will look at the derivation of a two-step EM algorithm for a general mixture of factor analysers. The difference of the IMoFA model with the MFA model proposed by Ghahramani and Hinton [206] is that IMoFA allows the components to have different number of factors p_j, and each component j will have a -possibly-different uniqueness Ψ_j. The argument for a single Ψ is the interpretation of Ψ as the common sensor noise. The IMoFA formulation is more general and can incorporate the common sensor noise assumption with the addition of a single weighted sum, given in Eq. 4.17. The complete model will be specified by the component priors π_j, factor loadings $\mathbf{\Lambda}_j$, component means $\boldsymbol{\mu}_j$, and the diagonal uniquenesses Ψ_j.

We define $\tilde{\boldsymbol{z}}_j^t$ and $\tilde{\mathbf{\Lambda}}_j$ as:

$$\tilde{\boldsymbol{z}}_j^t = [\boldsymbol{z}_j^t \;\; 1]^T \tag{4.5}$$

$$\tilde{\mathbf{\Lambda}}_j = [\mathbf{\Lambda}_j \;\; \boldsymbol{\mu}_j] \tag{4.6}$$

Let p_j denote the probability density function for a single component:

$$p_j(\boldsymbol{x}^t) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Psi_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}^t - \tilde{\mathbf{\Lambda}}_j\tilde{\boldsymbol{z}}_j^t)^T\Psi_j^{-1}(\boldsymbol{x}^t - \tilde{\mathbf{\Lambda}}_j\tilde{\boldsymbol{z}}_j^t)} \tag{4.7}$$

Furthermore, let us define h_j^t to be the hidden variable, which indicates the membership probability of sample $\boldsymbol{x}^t$ to a component $\mathcal{G}_j$.

$$\sum_{j=1}^{J} h_j^t = 1 \tag{4.8}$$

Then the expected log-likelihood is

$$Q = E\left[\sum_{t=1}^{N} \log \sum_{j=1}^{J} h_j^t \pi_j p_j(\boldsymbol{x}^t)\right] \tag{4.9}$$

E-step In the expectation step, we calculate the responsibilities h_j^t, and the expected values $E[\boldsymbol{z}_j^t|\boldsymbol{x}^t, \mathcal{G}_j]$ and $E[\boldsymbol{z}_j^t(\boldsymbol{z}_j^t)^T|\boldsymbol{x}^t, \mathcal{G}_j]$ for all data points and mixture components:

$$h_j^t = \frac{\pi_j p_j(\boldsymbol{x}^t)}{\sum_{l=1}^{J} \pi_l p_l(\boldsymbol{x}^t)} \tag{4.10}$$

$$E[\boldsymbol{z}_j^t|\boldsymbol{x}^t, \mathcal{G}_j] = \beta_j(\boldsymbol{x}^t - \boldsymbol{\mu}_j) \tag{4.11}$$

where $\beta_j = \boldsymbol{\Lambda}_j^T(\boldsymbol{\Lambda}_j\boldsymbol{\Lambda}_j^T + \Psi_j)^{-1}$ as in FA. Similarly,

$$E[\boldsymbol{z}_j^t(\boldsymbol{z}_j^t)^T|\boldsymbol{x}^t, \mathcal{G}_j] = I_j - \beta_j\boldsymbol{\Lambda}_j + \beta_j(\boldsymbol{x}^t - \boldsymbol{\mu}_j)(\boldsymbol{x}^t - \boldsymbol{\mu}_j)^T\beta_j^T \tag{4.12}$$

I_j is indexed with j, as it is the square identity matrix of rank p_j.

M-step The maximization step attempts to minimize the expected error function with respect to the parameters of the mixture model.

$$\begin{aligned}\frac{\partial Q}{\partial \tilde{\boldsymbol{\Lambda}}_j} \quad = \quad & -\sum_{t=1}^{N} h_j^t {\Psi_j^{(i)}}^{-1} \boldsymbol{x}^t E[\tilde{\boldsymbol{z}}_j^t|\boldsymbol{x}^t, \mathcal{G}_j]^T \\ & +\sum_{l=1}^{N} h_j^l {\Psi_j^{(l)}}^{-1} \tilde{\boldsymbol{\Lambda}}_j^{(i+1)} E[\tilde{\boldsymbol{z}}_j^l(\tilde{\boldsymbol{z}}_j^l)^T|\boldsymbol{x}^l, \mathcal{G}_j]^T = 0\end{aligned} \tag{4.13}$$

Then

$$\begin{aligned}\tilde{\boldsymbol{\Lambda}}_j^{(i+1)} \quad = \quad & [\boldsymbol{\Lambda}_j^{(i+1)}\boldsymbol{\mu}_j^{(i+1)}] \\ =& \left(\sum_{t=1}^{N} h_j^t \boldsymbol{x}^t E[\tilde{\boldsymbol{z}}_j^t|\boldsymbol{x}^t, \mathcal{G}_j]^T\right)\left(\sum_{l=1}^{N} h_l^t E[\tilde{\boldsymbol{z}}_j^l(\tilde{\boldsymbol{z}}_l^t)^T|\boldsymbol{x}^l, \mathcal{G}_j]^T\right)^{-1}\end{aligned} \tag{4.14}$$

Again we estimate $\Psi_j^{(i+1)}$ through the partial derivative of Q with respect to its inverse.

$$\begin{aligned}\frac{\partial Q}{\partial \Psi_j^{-1}} &= \frac{N\pi_j}{2}\Psi_j^{(i+1)} - \sum_{t=1}^{N}(-h_j^t \boldsymbol{\Lambda}_j^{(i+1)} E[\tilde{\boldsymbol{z}}^t|\boldsymbol{x}^t, \mathcal{G}_j]^T (\boldsymbol{x}^t)^T \\ &+ \frac{1}{2}h_j^t \boldsymbol{x}^t (\boldsymbol{x}^t)^T + \frac{1}{2}h_j^t \boldsymbol{\Lambda}_j^{(i+1)} E[\tilde{\boldsymbol{z}}^t (\tilde{\boldsymbol{z}}^t)^T|\boldsymbol{x}^t, \mathcal{G}_j]^T \boldsymbol{\Lambda}_j^{(i+1)^T}) = 0 \qquad (4.15)\end{aligned}$$

which gives

$$\Psi_j^{(i+1)} = \frac{1}{N\pi_j} diag\left(\sum_{t=1}^{N} h_j^t (\boldsymbol{x}^t - \boldsymbol{\Lambda}_j^{i+1} E[\tilde{\boldsymbol{z}}^t|\boldsymbol{x}^t, \mathcal{G}_j])(\boldsymbol{x}^t)^T\right) \qquad (4.16)$$

Note that the estimate for Ψ_j and the estimate of Ψ in the common noise MFA are simply related by the following equation:

$$\Psi = \sum_{j=1}^{J} \pi_j \Psi_j \qquad (4.17)$$

The estimation of $\pi_j^{(i+1)}$ is not changed:

$$\pi_j^{(i+1)} = \frac{1}{N}\sum_{t=1}^{N} h_j^t \qquad (4.18)$$

4.3. Incremental Mixture of Factor Analysers

Our simulations on different pattern recognition problems show that using mixtures of factor analyzers presents a better choice over mixture of Gaussians for a suitably chosen set of parameters. Even though a clear separation of dimensions into groups related to different factors does not happen in image processing applications (as dimensions correspond to pixels, and are usually correlated), a good coverage of the parameter space, and the validity of independent noise assumption ensures more successful models.

Two parameters are of great importance to the mixture of factor analyzers: The number of components in the mixture, and the number of factors in a given component, respectively. As the number of factors is increased from one to the number of data dimensions, the sample covariance matrix that is used in the training is approximated better and better. Conversely, adding more components per class devotes more parameters for modeling smaller groups of data. Both methods increase the accuracy on the training set. Both methods result in overlearning. Furthermore, increasing the number of parameters both increases the computational complexity, and causes regularization problems due to singularities.

We could try to optimize both parameters separately. Selecting the number of factors would be trivial, if we were dealing with a noise-free model and also had access to the true data covariance matrix. The number of positive eigenvalues of the covariance matrix would give us the required number of factors. However, we cannot directly use this value, because we only have access to the sample covariance matrix, and of course, there is noise. What we can do as a simple heuristic is to set a variance threshold to determine the number of significant eigenvalues, and to use this number as the number of factors across the mixture model. However, if a separate process is responsible from the generation of each component, it is possible that these processes operate in manifolds of different dimensionality. Thus, each component may require a different number of factors. Consequently, these two parameters should be determined jointly.

Two primary approaches to this problem are incremental and decremental; we can start with a very simple model, and increase its complexity, or we can start with a very complex model and decrease the complexity. From a computational point of view, the latter is not very attractive, because Gaussian models in higher dimensions are very cumbersome, and the amount of training samples required scales quadratically with increasing dimensionality.

Suppose we decide on an incremental algorithm, and at some point in training, we decide to increase the model complexity by adding a component or a factor. The costs of adding a factor and adding a component are not similar. On a trained model, adding a component represents a broader re-tuning of the whole parameter set, whereas a factor addition is mostly a local change of a single component. Algorithmically, the implementation of both can be achieved by a small modification to the EM algorithm. Each requires an initial point in the parameter space to be determined, and EM will be used until the new model converges.

One question we can ask at this point is whether adding factors to an already converged factor analysis model is useful or not. We know that a FA model with p-factors does not necessarily incorporate the factors of a $(p-1)$-factor FA model. We have experimentally shown that adding a new factor to a trained model is computationally much more efficient than training a new model from scratch (See Section 4.4.4).

In [209], the authors describe a method to learn a Gaussian mixture model in an incremental manner. The basic idea is that the maximum likelihood parameters for the one-component mixture can be directly derived from the data. Then, adding a new component can be achieved by selecting the optimum insertion point from among a number of carefully chosen candidate data points for the placement of a new component, and applying EM until convergence to obtain the new mixture. This insertion procedure can be repeated until a maximum number of components is added, or until a stopping criterion is met.

In a similar fashion, we will introduce the incremental MoFA algorithm that starts with a one-factor, one-component mixture and proceeds by adding new factors or new components until some stopping condition is satisfied. The *Incremental Mixture of Factor Analyzers* (IMoFA) algorithm relies on fast heuristic metrics to single out one component for splitting and another for factor addition. The pseudocode of the algorithm is given in Fig. 4.3. To check complexity and alleviate overfitting, the split and factor addition are tested on a validation set, separate from the training set over which the parameters are calculated, and the action that causes the greatest increase in the validation likelihood is chosen. The algorithm terminates when there is no additional improvement on the validation likelihood.

IMoFA algorithm given in Fig. 4.3 is unsupervised. In the case of a classification problem, we can use it to fit a MoFA to examples of each class seperately. This is a *likelihood-based* approach (IMoFA-L) where the likelihood of each class is maximized separately. There is also a *discriminative variant* of this algorithm (IMoFA-A) that adds factors and components while monitoring classification accuracy on the validation set, instead of the likelihood [210]. All class models are needed to calculate accuracy, thus they can no longer be trained separately. From here onwards, we will focus on the likelihood based version of this model, and use IMoFA-L and IMoFA interchangeably. For another variant that uses *alternating expectation, conditional maximization* (AECM) algorithm instead of EM for training, see [211].

4.4. Component and Factor Addition

For the incremental algorithm to be accurate and efficient, the candidate component and factor should be placed and initialized intelligently. Adding new components by splitting an existing one involves two decisions: which component to split, and how to split it. We have tested various schemes for each question.

4.4.1. Kurtosis-based Component Selection

The first method to select the component relies on univariate kurtosis of the data. A univariate Gaussian component has a kurtosis close to 3. Smaller values are indicative of non-Gaussianity. Hence, one way of choosing a component to split is to select the component with the smallest average (taken over data dimensions) kurtosis:

$$\gamma_1(j) = \frac{1}{\pi_j} \sum_{t=1}^{N} h_j^t \left(\sum_{l=1}^{d} \left| \frac{||\boldsymbol{x}_l^t - \boldsymbol{\mu}_{jl}||^4}{\sigma_{jl}^4} - 3 \right| \right) \tag{4.19}$$

Mardia proposed a multivariate kurtosis metric, which can be used to test the non-Gaussianity of the data under each component [212]. For a multinormal distribution, the multivariate kurtosis takes the value

$$\beta_{2,d} = d(d+2) \tag{4.20}$$

and if the underlying population is multivariate normal with mean $\boldsymbol{\mu}$, the sample counterpart of $\beta_{2,d}$, namely $b_{2,d}$, has the following asymptotic distribution as the number of samples $N \to \infty$:

$$\frac{b_{2,d} - d(d+2)}{\left[\frac{8d(d+2)}{N}\right]^{\frac{1}{2}}} \sim \mathcal{N}(0,1) \tag{4.21}$$

with

$$b_{2,d} = \frac{1}{N}\sum_{t=1}^{N}\left[(\boldsymbol{x}^t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^t - \boldsymbol{\mu})\right]^2 \tag{4.22}$$

This metric can be adapted to the mixture model by using the"soft count" $\sum h_j^t \equiv E[\mathcal{G}_j|\boldsymbol{x}^t]$:

$$\gamma_2(j) = \{b_{2,d}^j - d(d+2)\}\left[\frac{8d(d+2)}{\sum_{t=1}^{N} h_j^t}\right]^{-\frac{1}{2}} \tag{4.23}$$

$$b_{2,d}^j = \frac{1}{\sum_{l=1}^{N} h_j^l}\sum_{t=1}^{N} h_j^t\left[(\boldsymbol{x}^t - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}^t - \boldsymbol{\mu}_j)\right]^2 \tag{4.24}$$

The component with the greatest γ_j is the one that looks least unimodal and is selected for splitting.

Kurtosis-based measures of non-Gaussianity are intended to discover the discrepancy between the data and the model fitted. We can also look at the likelihood of the data under the model to decide the component to split. For this purpose, a method called the *soft geometric mean likelihood* is inspected.

4.4.2. Soft Geometric Mean Likelihood

In assessing the fit of a model on some data set, one often refers to the likelihood. By itself, the likelihood is not immediately indicative of goodness-of-fit, as it depends on N, the number of samples. Denoting the converged parameters of model j with $\hat{\theta}_j$, and the likelihood of the data set χ under this model with $l_j(\chi;\hat{\theta}_j)$, we can use the following scaled value to remove the dependence on N:

$$\hat{\gamma}_j = \left[l_j(\chi;\hat{\theta}_j)\right]^{\frac{1}{N}} = \left[\prod_{t=1}^{N} l_j(\boldsymbol{x}^t;\hat{\theta}_j)\right]^{\frac{1}{N}} \tag{4.25}$$

$\hat{\gamma}_j$ is called the *geometric mean likelihood* for model j [213]. If we are dealing with a mixture model, we might want to assess the goodness-of-fit for each component of the mixture. In the IMoFA algorithm, this kind of a metric can be useful in determining the candidate component for splitting. The following modification in this metric is proposed to take into account the membership probabilities for each data point, i.e. the posteriors h_j^t:

$$\hat{\gamma}_j = \left[\prod_{t=1}^{N} l_j(\boldsymbol{x}^t;\hat{\theta}_j)^{h_j^t}\right]^{\frac{1}{\sum_{t=1}^{N} h_j^t}} \tag{4.26}$$

We can use this *soft geometric mean likelihood* and the following log version in assessing the goodness-of-fit of the components of our mixture model:

$$\log(\hat{\gamma}_j) = \frac{\left[\sum_{t=1}^{N} h_j^t log(l_j(\boldsymbol{x}^t;\hat{\theta}_j))\right]}{\sum_{l=1}^{N} h_j^l} \tag{4.27}$$

As expected, in the extreme case of $h_j^t \in \{0, 1\}$, the proposed metric reduces to the original geometric mean likelihood of the respective model under the hard clustered data belonging solely to the component.

However, there is a problem with this metric: It is based on the assumption that there exists a single data set, drawn from a Gaussian distribution with unknown parameters. If we want to compare different fits on different sets, we also need to correct for the assumed shape of the Gaussian. This is indeed the case in our mixture model, since different portions of the data are modeled by different components, as indicated by the posteriors. In other words, the assumption that there is a single Gaussian model underlying the data must be relaxed.

To correct the geometric mean likelihood, we need to scale it by its expected value for any sample drawn from the distribution specified by the model parameters $\mathbf{\Sigma}_j$ and $\boldsymbol{\mu}_j$. Using $E[\boldsymbol{x}] = \boldsymbol{\mu}_j$, with $\boldsymbol{\mu}_j$ being the mean parameter of the model, we get:

$$\begin{aligned} E[\hat{\gamma}_j|\Sigma_j] &= \left[\prod_{t=1}^{N} \left(\frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}_j|^{\frac{1}{2}}} \right)^{h_j^t} \right]^{\frac{1}{\Sigma_{t=1}^{N} h_j^t}} \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}_j|^{\frac{1}{2}}} \end{aligned} \tag{4.28}$$

which is independent of the sample size, just like $\hat{\gamma}_j$.

Hence, our final metric is

$$\gamma_3(j) = \frac{\hat{\gamma}_j}{E[\hat{\gamma}_j|\Sigma_j]} = \hat{\gamma}_j (2\pi)^{\frac{d}{2}} |\Sigma_j|^{\frac{1}{2}} \tag{4.29}$$

and a good fit is indicated by a value close to one.

These methods were compared with a statistical procedure. One-tailed t-test results for the IMoFA-L component selection metrics on the Pendigits set reveals that the multivariate kurtosis based method has significantly higher accuracy than univariate-kurtosis and SGML methods.

4.4.3. Component Splitting

Once we choose the component, we must decide how we are going to split it. In the greedy EM scheme, Verbeek *et al.* generated a number of candidate points for the new component means and let the EM procedure take the model to the nearest local optimum from those initial points [209].

We will also proceed in a similar way, but we will choose a single pair of points and pass it to EM. Two different schemes are tested for deciding on the two data points $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ to initialize the EM algorithm. These points will evolve into the two components that replace the original split component. The first method looks at the principal direction of the data that are associated with the component, and places $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ symmetrically along the principal axis:

$$\hat{\boldsymbol{\mu}}_{1,2} = \boldsymbol{\mu} \pm \boldsymbol{v} \tag{4.30}$$

where $\boldsymbol{\mu}$ is the mean vector of the split component, and $\boldsymbol{v}$ is the principal eigenvector (*PCA-based split*).

The second method places $\hat{\boldsymbol{\mu}}_1$ to the data point of the component that has the greatest Mahalanobis distance to $\boldsymbol{\mu}$ (*Distance-based split*). $\hat{\boldsymbol{\mu}}_2$ is again selected to make Eq. 4.30 hold. These two schemes are contrasted with a third scheme that places the new means randomly. As expected, both schemes outperform the random placement in terms of the final log-likelihood of the split (See Figs. 4.4, 4.5 and 4.6).

4.4.4. Factor Addition

The basic factor analysis uses the factor loading matrix $\boldsymbol{\Lambda}$ to model all the covariances, and part of the variances. We will look at the difference between the sample covariance and the modeled covariance for each component, and consider the one with the greatest difference for factor addition. Once the component is selected, the new factor can be randomly initialized, in which case EM is expected to take it to a local optimum. A method of initializing the new factor called the *residual factor addition* is shown to work better.

Given a component $\mathcal{G}_j$ and a data point $\boldsymbol{x}^t$, we can find the expected value of $\boldsymbol{z}^t$. Actually, $\mathrm{E}[\boldsymbol{z}^t|\boldsymbol{x}^t, \mathcal{G}_j]$ is the p-dimensional data point if we intend to employ the factor analyser as a dimensionality reduction method. If we wish to restore the original data point from the dimensionality-reduced $\boldsymbol{z}^t$, we will multiply it with the factor loading matrix:

$$\tilde{\boldsymbol{x}}_j^t = \boldsymbol{\Lambda}_j E[\boldsymbol{z}^t|\boldsymbol{x}^t, \mathcal{G}_j] \tag{4.31}$$

However, unless we use the complete eigenvectors in $\boldsymbol{\Sigma}_j$ (or some equivalent basis), $\tilde{\boldsymbol{x}}_j^t$ will not exactly coincide with $\boldsymbol{x}^t$. The residual error is:

$$\boldsymbol{e}_j^t = \boldsymbol{x}^t - \tilde{\boldsymbol{x}}_j^t \tag{4.32}$$

When we add the new factor column $\boldsymbol{\Lambda}_{j,p+1}$ to $\boldsymbol{\Lambda}_j$, it will make a contribution to the re-estimated value of each $\boldsymbol{x}^t$ in the direction of $\boldsymbol{\Lambda}_{j,p+1}$ with the magnitude of $\boldsymbol{z}_{p+1}^t$. The residual factor addition aims at minimizing these residuals by selecting $\boldsymbol{\Lambda}_{j,p+1}$ to be the principal direction (the eigenvector with the largest eigenvalue) of the residual vectors. This new factor is used in bootstrapping EM. We also correct the estimate of the common noise component according to this new factor.

Suppose $\boldsymbol{\Lambda}$ is the maximum likelihood estimator of the factor loading matrix of a component with p factors, and Ψ contains the variances that are left unexplained by $\boldsymbol{\Lambda}$. The sample covariance $\boldsymbol{\Sigma}$ is estimated through the model parameters by

$$\boldsymbol{\Sigma} \sim \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \Psi \tag{4.33}$$

Assume we add a new factor $\mathbf{\Lambda}_{p+1}$ to $\mathbf{\Lambda}$. This new term will cause an increase in the variances estimated through the factor loading matrix:

$$tr([\mathbf{\Lambda}\mathbf{\Lambda}_{p+1}][\mathbf{\Lambda}\mathbf{\Lambda}_{p+1}]^T) = tr(\mathbf{\Lambda}\mathbf{\Lambda}^T) + [\mathbf{\Lambda}^2_{p+1,1}\mathbf{\Lambda}^2_{p+1,2}\dots\mathbf{\Lambda}^2_{p+1,d}]^T \tag{4.34}$$

Plugging Eq. 4.34 into Eq. 4.33, we can see that in order to preserve the variance estimates, Ψ shall be updated by subtracting the positive vector obtained by squaring the values of the new factor:

$$\Psi^{new} = \Psi - diag([\mathbf{\Lambda}^2_{p+1,1}\mathbf{\Lambda}^2_{p+1,2}\dots\mathbf{\Lambda}^2_{p+1,d}]) \tag{4.35}$$

Fig. 4.7 shows the comparison of the residual factor addition method with the random factor addition method on a sample subset from the Optdigits dataset. Adding a residual factor leads to shorter EM runs, and better log-likelihood values.

For the residual factor addition method, we are going to calculate the principal direction of the data distributed under the parameter set of a certain component. The data points will contribute to the principal direction only by an amount proportional to their posteriors. Since the assignment of data points to components will be weighted, this is a *soft* PCA scheme.

With a single component, the $(i,l)^{th}$ entry of the unbiased sample covariance matrix S can be expressed as:

$$S(i,l) = \frac{\sum_{t=1}^{N} \boldsymbol{x}_i^t \boldsymbol{x}_l^t}{N-1} \tag{4.36}$$

where the data samples $\boldsymbol{x}^t$ are centered by subtracting the data mean.

If we are dealing with multiple components $j = 1..J$, we need to take the posteriors h_j^t into account. The soft covariance S_j for component j will be calculated using the following formula:

$$S_j(i,l) = \frac{\sum_{t=1}^{N} h_j^t \boldsymbol{x}_i^t \boldsymbol{x}_l^t}{(\sum_{m=1}^{N} h_j^m) - 1} \tag{4.37}$$

where the data are centered by the soft component mean $\boldsymbol{\mu}_j$:

$$\boldsymbol{\mu}_j = \frac{\sum_{t=1}^{N} h_j^t \boldsymbol{x}^t}{\sum_{m=1}^{N} h_j^m} \tag{4.38}$$

For a fast calculation of S_j, we can pre-multiply each centered data point by the square-root of the corresponding posterior and proceed normally. This soft covariance will be used both for the residual factor addition method, and for the PCA-based initialization of the new means of a split component.

Fig. 4.8 shows the training of an FA model from scratch versus training it through factor addition to an already converged model. The initial log-likelihood is naturally much higher in factor addition. Fig. 4.9 shows the number of EM iterations needed to train a FA model with p factors on the 16-dimensional Pendigits data set (See Section 4.5 for information on datasets used in this section). Especially when p increases, the factor addition method becomes more feasible.

4.5. Simulation Results

The IMoFA algorithm was tested on ten datasets. Pendigits (PEN) and Optdigits (OPT) datasets are both optical character recognition sets, developed in Boğaziçi University. The PEN, OPT, the Waveform dataset (WAVE) and University of Massachusetts, Vision Group's Image Segmentation Data (SEG) can be found at the UCI Machine Learning Repository [214]. We have used the ORL face data for male-female classification [215] and a pre-processed 10-class selection of Vistex texture database [216]. We have the Yeast microarray gene expression data [217] (YEAST), and a 20-class phoneme data distributed with LVQ package of the Helsinki University of Technology [218]. Two synthetic datasets for the *threshold max* problem were used,

Table 4.1. Datasets

Dataset	**Training**	**Test**	**Dimensions**	**Classes**
PEN	5,494	3,498	16	10
OPT	2,880	1,797	64	10
ORL	400	cv10	256	2
VIS	2,700	910	169	10
LVQ	1,929	1,929	20	16
SEG	700	1,610	14	7
YEAST	208	cv7	79	5
WAVE	300	4,700	21	3
TM 50	2,000	1,000	50	2
TM 150	2,000	1,000	150	2

one with 50 relevant features, and another with 50 relevant, 50 irrelevant, and 50 redundant (i.e. copied) features [219]. The latter two datasets are selected to see how the algorithms behave in face of redundant and irrelevant information. See Table 4.1 for a summary of datasets used. Cross-validation with k samples is indicated as cv$k$ in the column that lists the number of test samples.

We report experimental results with five different methods:

- MoG (Mixtures of Gaussians): Each class is modeled with one Gaussian component. Shared, diagonal and full covariance models are trained, and only the best result is reported.
- MoMoG (Mixtures of mixtures of Gaussians): Each class is modeled with a mixture of Gaussians, with one to ten components per class. Shared covariances are defined for all the components in the final mixture. Shared, diagonal and full covariance models are trained, and the best result is reported.
- MoFA (Mixtures of factor analyzers): Each class is modeled with a single Gaussian component, derived from a FA formulation. Models of increasing complexity are trained for increasing numbers of factors. The best result is reported.

- MoMoFA (Mixtures of mixtures of factor analyzers): Similar to the previous methods, but each class is modeled with a mixture of FAs. A large range of component and factor parameters is systematically explored, and the best result is reported.
- IMoFA-L (Likelihood-based incremental mixtures of factor analyzers): Multivariate kurtosis-based component selection, and PCA-based splitting is used. Each class is modeled with IMoFA-L separately. As opposed to the previous results, which are selected from a number of models with different complexity, we directly report the IMoFA-L result.

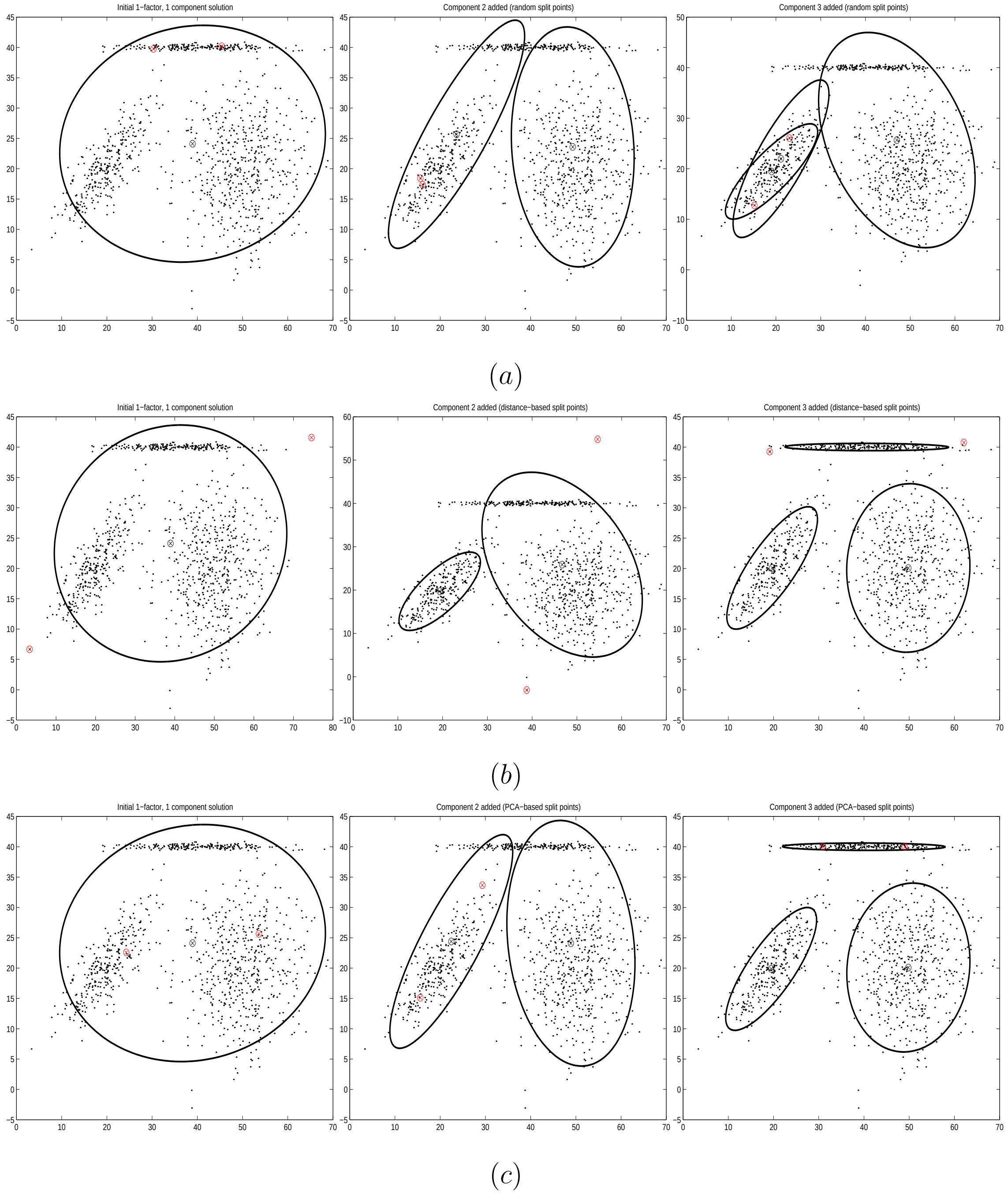
Tables 4.2 and 4.3 report the simulation results in terms of log-likelihood, classification accuracy, and the number of parameters for the final model. These indicate that the incremental mixture of factor analyzers is capable of finding reasonable points in the complexity versus accuracy curve. Compared to fixed-parameter MoG and MoFA models, IMoFA-L finds high-likelihood solutions with moderate number of parameters.

Modeling a class with more than one component is useful if samples of the class are generated by different processes, like the allomorphs of digits. The incremental algorithm allows automatic allocation of more components to classes with many alternative structures. In a dataset where this is not the case, increasing the number of components quickly leads to overlearning, as the components impose superficial clusters on the data. The incremental algorithm checks for improvement after each component addition on a separate validation set and does not permit such overfitting. This property will be useful in modeling facial features in Chapter 5.

Shi and Xu have recently compared IMoFA to Akaike's information criterion (AIC) [220], Bozdoğan's consistent Akaike's information criterion (CAIC) [221], Schwarz's Bayesian inference criterion (BIC) [222], which is equivalent to Rissanen's minimum description length (MDL) criterion [223], and Bayesian Yin-Yang (BYY) harmony learning [224]. Their results show that using IMoFA or BYY reduces the computation time about five times, when compared to AIC, CAIC and BIC, with comparable or superior accuracies. They also observe that "BIC, IMoFA, and BYY have the highest correct rates, while AIC has a risk of overestimating both the number of Gaussian components and the number of local factors, and CAIC has a risk of underestimating the number of components."

```
algorithm **IMoFA**(train, validation)
    $[\Lambda, \mu, \Psi] \leftarrow$ train a 1-component, 1-factor model
    oldLikelihood $\leftarrow$ -Infinity
    /*Likelihoods are calculated on validation set*/
    newLikelihood $\leftarrow$ likelihood$(\Lambda, \mu, \Psi)$
    **while** newLikelihood $>$ oldLikelihood
    /*Perform a single split*/
        x $\leftarrow$ Select a component for splitting
        $[\Lambda_1, \mu_1, \Psi_1, \pi_1] \leftarrow$ EM(split x).
        actionL(1) $\leftarrow$ likelihood$(\Lambda_1, \mu_1, \Psi_1, \pi_1)$
    /*Perform a single factor addition*/
        y $\leftarrow$ Select a component to add a factor
        $[\Lambda_2, \mu_2, \Psi_2, \pi_2] \leftarrow$ EM(add factor to y).
        actionL(2) $\leftarrow$ likelihood$(\Lambda_2, \mu_2, \Psi_2, \pi_2)$
    /*Select the best action*/
        z $\leftarrow$ max(actionL)
    /*Update the parameters*/
        $[\Lambda, \mu, \Psi, \pi] \leftarrow [\Lambda_z, \mu_z, \Psi_z, \pi_z]$
        oldLikelihood $\leftarrow$ newLikelihood
        newLikelihood $\leftarrow$ likelihood$(\Lambda, \mu, \Psi, \pi)$
    **end**
    **return** $[\Lambda, \mu, \Psi, \pi]$
**end**
```

Figure 4.3. IMoFA Algorithm

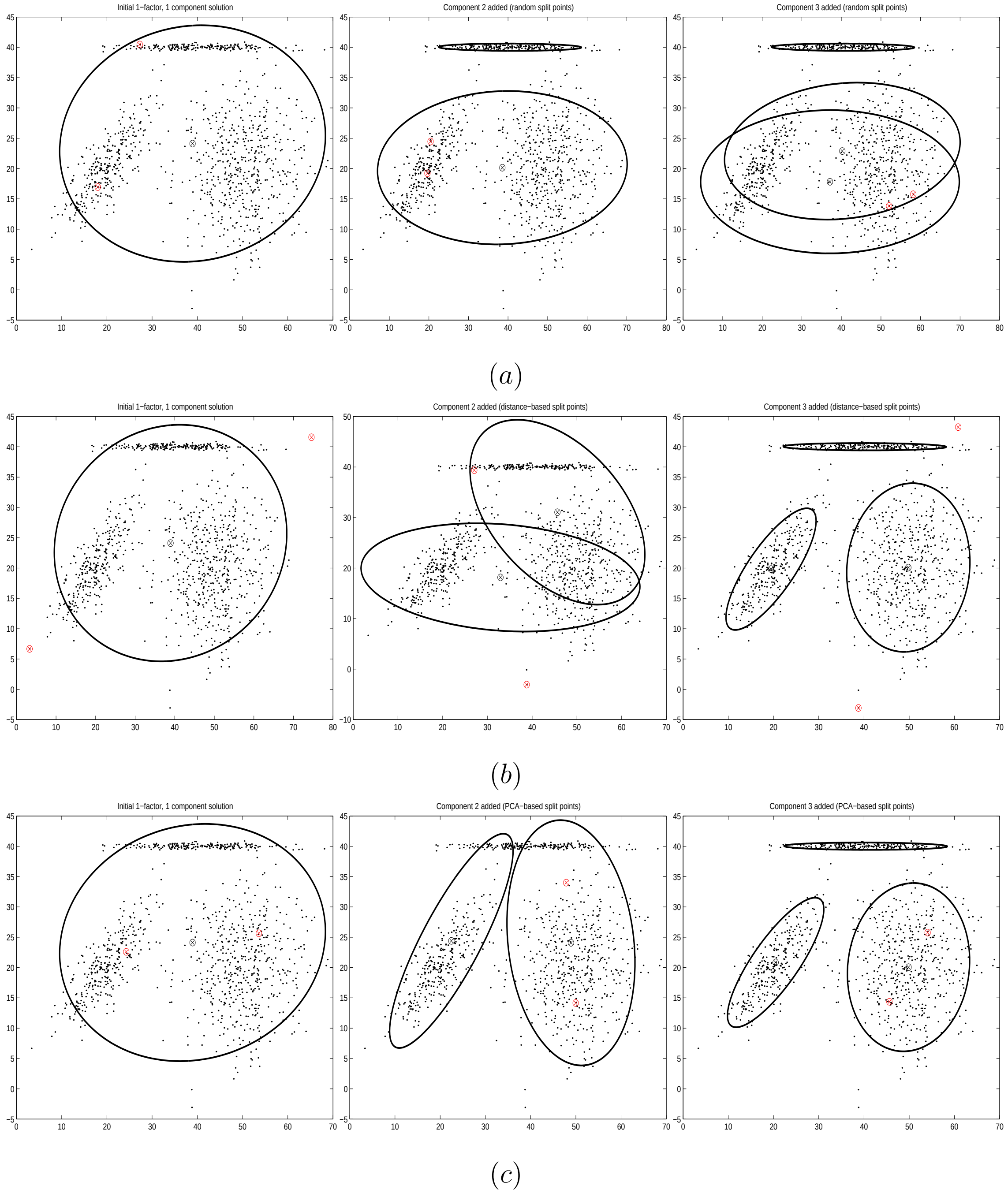


(*a*)

(*b*)

(*c*)

Figure 4.4. Addition of components to the artificial dataset. The components are selected by looking at the **univariate kurtosis**. The initial means for the next split are shown on the figures. They also indicate which component is split. Three methods for splitting are shown successively: (a) Random split. (b) Distance-based split. (c) PCA-based split.



(*a*)

(*b*)

(*c*)

Figure 4.5. Addition of components to the artificial dataset. The components are selected by looking at the **average likelihood**. As in the previous figure, the three methods for splitting are shown successively: (a) Random split. (b) Distance-based split. (c) PCA-based split.

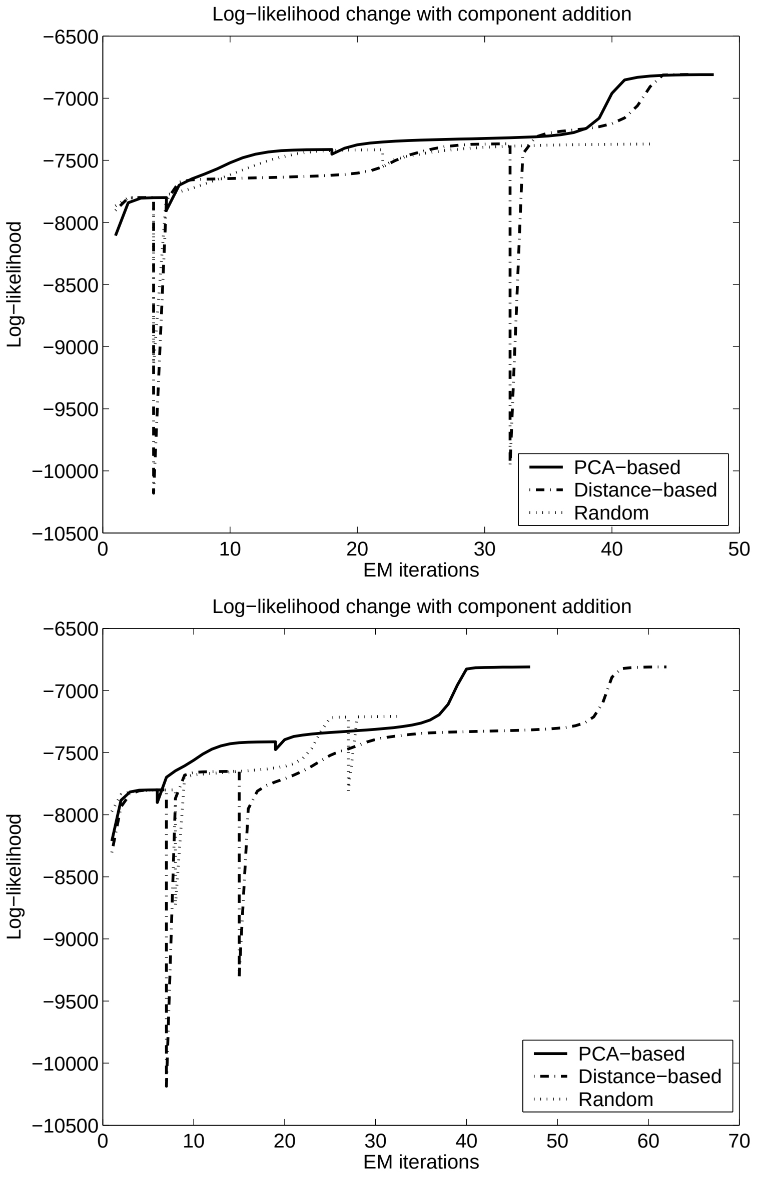


Figure 4.6. The change in log-likelihood versus EM iterations. In the first figure component selection is done by looking at the kurtosis, in the second it is done by looking at the average likelihoods. Each figure shows the progress of the three splitting methods. The locations with log-likelihood decrease indicate component addition. As expected, distance-based addition greatly lowers the likelihood initially, as the new means are as far away from the component mean as possible. In general, PCA-based splitting is more promising.

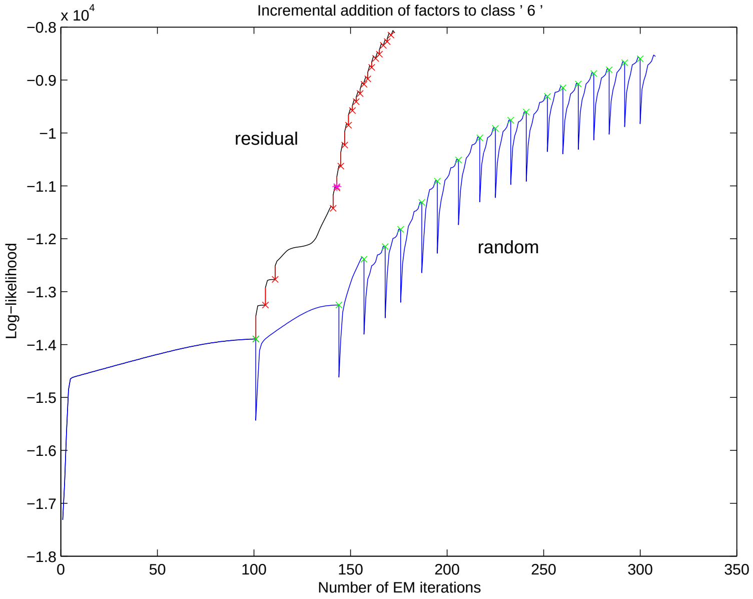


Figure 4.7. Log-likelihood change upon addition of new factors with residual factor addition method and random addition in Pendigits dataset. '×' symbols indicate that a new factor is added. The residual scheme reaches higher likelihood values, and requires less EM iterations. Usually the likelihood increases right after the factor addition, before EM iterations start.

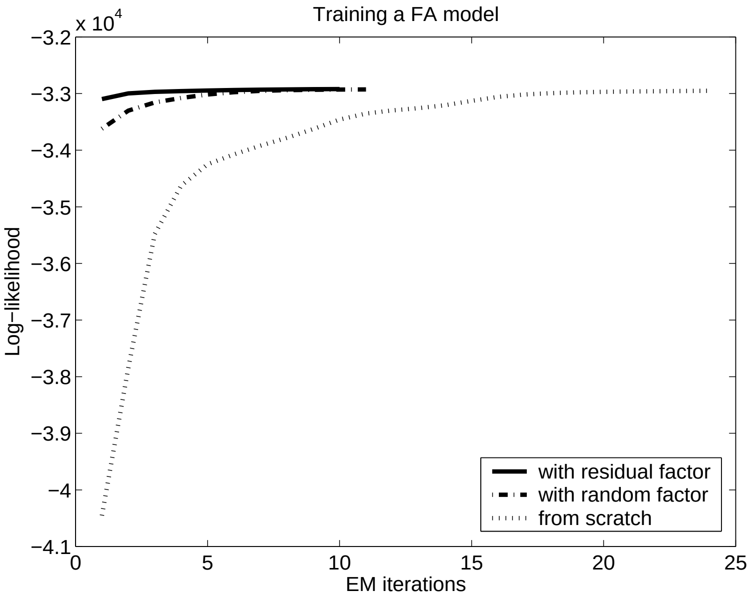


Figure 4.8. Training a FA model for the Pendigits dataset. The factor addition methods require significantly fewer EM iterations than training the model from scratch.

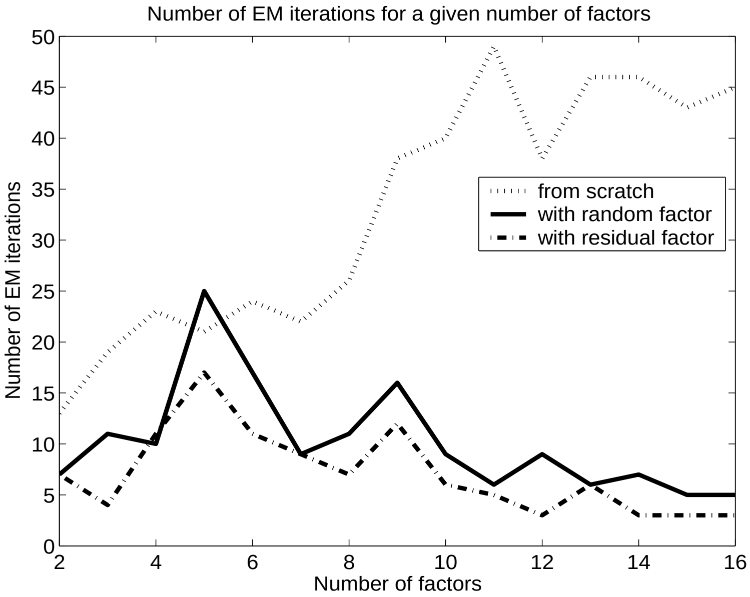


Figure 4.9. Number of EM iterations necessary to train a FA model for the Pendigits dataset. As the number of factors increase, the factor addition method becomes more plausible, as EM needs roughly the same number of iterations each time.

Table 4.2. Simulation Results

	Method	**Log-Likelihood**	**Accuracy**	**Number of Parameters**
PEN	MoG	-117804	98.50(±0.00)	1529
	MoMoG	-116495	98.72(±0.13)	4589
	MoFA	-142120	95.20(±0.00)	1625
	MoMoFA	-118203	99.35(±0.22)	8865
	IMoFA-L	-114188	97.93(±0.35)	2699
OPT	MoG	-68462	95.76(±0.00)	18909
	MoMoG	-90758	92.98(±0.54)	6049
	MoFA	-110719	96.92(±0.00)	6669
	MoMoFA	-93558	97.94(±0.38)	30109
	IMoFA-L	-78994	92.92(±0.56)	7629
ORL	MoG	-6918	87.25(±4.30)	66305
	MoMoG	4967	98.25(±1.70)	5129
	MoFA	7661	99.00(±1.30)	5889
	MoMoFA	8493	99.25(±1.20)	28425
	IMoFA-L	6702	98.50(±1.30)	8068
VIS	MoG	65318	35.60(±0.00)	145349
	MoMoG	93108	73.82(±1.10)	8549
	MoFA	76770	62.53(±0.00)	18768
	MoMoFA	80699	58.29(±1.20)	93168
	IMoFA-L	126801	70.40(±1.70)	29493
YEAST	MoG	-11393	63.78(±4.30)	16199
	MoMoG	-2932	93.37(±6.00)	2384
	MoFA	-2536	92.35(±6.00)	2453
	MoMoFA	-3245	93.88(±7.40)	7203
	IMoFA-L	-3028	91.33(±6.80)	3151

Table 4.3. Simulation Results - Continued

	Method	**Log-Likelihood**	**Accuracy**	**Number of Parameters**
	MoG	−110147	88.96(±0.00)	3695
L	MoMoG	−122773	86.15(±1.10)	14783
V	MoFA	−114128	87.56(±0.00)	3555
Q	MoMoFA	−121208	89.24(±0.59)	17699
	IMoFA-L	−108905	89.44(±0.62)	1749
W	MoG	−160899	77.89(±0.00)	758
A	MoMoG	−172130	74.57(±1.70)	3035
V	MoFA	−161938	75.85(±0.00)	716
E	MoMoFA	−206572	73.53(±0.83)	3500
	IMoFA-L	−154295	82.55(±0.53)	195
	MoG	−33896	88.57(±0.00)	839
S	MoMoG	−43611	89.65(±2.10)	3359
E	MoFA	−63726	65.78(±0.00)	1098
G	MoMoFA	−54001	88.89(±1.30)	419
	IMoFA-L	−39454	85.13(±2.40)	603
T	MoG	−45011	50.30(±0.00)	2651
M	MoMoG	−45945	50.04(±1.70)	13259
5	MoFA	−44403	54.90(±0.00)	251
0	MoMoFA	−45572	50.54(±1.40)	5559
	IMoFA-L	−44306	55.81(±1.10)	300
T	MoG	191557	52.10(±0.00)	22951
M	MoMoG	−88722	51.02(±0.40)	114759
1	MoFA	−131162	50.70(±0.00)	3451
5	MoMoFA	−129142	50.58(±1.60)	16659
0	IMoFA-L	−124944	51.40(±1.50)	3333

5. AUTOMATIC LANDMARK LOCALIZATION

Marco Polo describes a bridge, stone by stone.
"But which is the stone that supports the bridge?" Kublai Khan asks.
"The bridge is not supported by one stone or another,"
Marco answers, "but by the line of arch that they form."
Kublai Khan remains silent, reflecting. Then he adds:
"Why do you speak to me of the stones? It is only the arch that matters to me."
Polo answers: "Without stones there is no arch."
Italo Calvino

5.1. Introduction

Facial feature localization is an important component of applications like facial feature tracking, facial modeling and animation, expression analysis, face recognition and biometric applications that rely on 2D and 3D face data. There are a number of problems that need to be solved when using 3D information for registration: The scale differences change local features (e.g. curvatures), and the number of sampled points within the facial region. The acquisition is usually problematic for the eyes and mouth regions. The shoulders, the hair and non-facial clutter in the scene are rife with features that can be misleading. Finally, the amount of data to be processed is large. Conversely, if the facial landmarks are located, many of these issues will be easier to deal with.

Robust and automatic detection of facial features is a difficult problem, suffering from all the known problems of face recognition such as illumination, pose and expression variations, and clutter. With the emergence of 3D face recognition as a stand-alone or supporting modality for 2D face recognition, automatic facial feature detection needs to be accomplished for robust 3D registration prior to recognition. We distinguish this problem from face detection, which is the localization of a bounding box for the face: The aim in landmark detection is locating selected facial points with the greatest possible accuracy.

There is no universal set of landmark points accepted for the use of registration. Fred Bookstein defines landmarks as “points in one form for which objectively meaningful and reproducible biological counterparts exist in all the other forms of a data set” [176]. Most frequently used landmarks are the nose tip, eye and mouth corners, centre of the iris, tip of the chin, the nostrils, the eyebrows and the nasion. A small number of landmarks (3-5) are deemed to be sufficient for a good initial transformation prior to registration. Some registration methods require as many landmarks as possible, and are sensitive to errors in landmark locations. We comment more on this issue in Chapter 6.

This part of our work deals with a saliency-based facial feature landmarking model. We would like to implement a general system that can learn arbitrary features without explicity stated heuristics. We will employ the IMoFA-L method detailed in Chapter 4 to treat the problem of automatic landmark localization uniformly [225, 226, 227, 228]. Statistical features are harnessed to locate each landmark independently, and mixture models are used to accomodate various types of landmarks (e.g. open and closed eyes). We would also like our system to be robust in the face of acquisition noise, i.e. we would like to be able to landmark a face without a nose, or with eyes hidden behind sunglasses. For this purpose, a structural correction algorithm is developed in Section 5.6. Finally, we would like our system to be accurate and efficient, in terms of computational resources it uses. For this purpose, we explore a 2D method based on Gabor wavelets, a 3D method based on the depth map and a 3D assisted-2D method based on a recent illumination correction technique.

5.2. Related Work in Landmarking

As Brunelli and Poggio state in [229], “features are only as good as they can be computed.” The correct localization of the landmarks is crucial to many algorithms, and it is usually not possible to judge the sensitivity of an algorithm to localization errors from its description. Most of the facial recognition approaches opt for a manual localization of the facial landmarks on a given training set, committing a fully automatic recognition system to future work [140, 157, 125, 164, 165, 150, 193, 151, 184]. In the few systems that propose automatic methods, the detection of facial landmarks or anchor points are usually accomplished with heuristics that are experimentally determined to work under particular conditions [230, 231, 232, 233, 161, 165, 234, 120, 235, 236]. While this approach produces excellent results for some cases, it requires the designer of the system to come up with different solutions for each different landmark, and can fail as soon as the assumptions are violated. For instance taking the closest point to the camera as the tip of the nose (as in [237] or [238]) may work for the majority of the frontal images, but a streak of hair can be detected as the nose tip in some of the images [239].

In 2D, vertical projection histograms of intensity values can be used to localize the eye and mouth regions [240, 241, 130, 242, 243, 234, 244, 245]. Also the contrast differences in the eye region were employed to train classifiers for eye detection [246, 247, 248, 238]. However, one has to take into account that landmarks can change appearance, depending on the expression. The assumption that the eyes are open for detection can easily be violated, especially when there is simultaneous 3D acquisition with a laser scanner.

Another typical characteristic of facial landmarking is the serial search approach, where the localization of one landmark depends on the localization of other landmarks [230, 231, 130, 249]. For example, one often starts with one prominent landmark, say tip of the nose in 3D, and based on this ground-truth, proceeds to identify the other features [237]. This approach is not robust, because an erroneous detection in the chain will cause errors in the rest of the landmarks as well.

Next to employing heuristics, a second approach to landmark localization is the joint optimization of structural relationships between landmark locations and local feature constraints, which are frequently conceived as distances to feature templates [250, 29]. The landmark locations are modeled with graphs, where vertices are placed on each landmark and the arcs characterize pairwise distances. In [29] local features are modeled with Gabor jets, and a template library (called *the bunch*) is exhaustively searched for the best match at each feature location. A large number of facial landmarks (typically 30-40) are used for graph based methods. Fewer and sparsely distributed landmarks arguably do not produce a sufficient number of structural constraints. We give more detail on the Gabor jets in Section 5.5.

A third and recent approach in facial feature localization is the adaptation of the popular Viola-Jones face detector to this problem [251, 252]. In this approach, patches around facial landmarks are detected in the face area with a boosted cascade of simple classifiers based on Haar wavelet features [253]. This approach is used for the coarse-scale detection, as a substitute for manual initialization. Face detection techniques are sometimes used for landmark detection. We postpone a review of these methods to Section 5.4, to put them in the context of our proposed model. We summarize 2D approaches to facial landmark localization in Table 5.1.

3D information is not commonly used in finding facial fiducial points, since 3D face imaging and handling of the resulting data volume are still not mainstream techniques. Furthermore outlier noise makes reliable processing difficult. In [159] the bunch graph method that uses 2D Gabor jet features introduced in [29] is extended to a 324-dimensional 3D jet method that simultaneously locates facial landmarks. Colbry *et al.* employ surface curvature-based shape indices under geometrical constraints to locate features on frontal 3D faces [237]. Their method has been generalized to the multi-pose case with the aid of 2D information, such as the output of Harris corner detector on the gray-level information and related geometrical constraints. In [260] curvature- and geometry-based heuristics were combined to locate a large number of landmarks. In a supporting study, these landmarks are used to find further landmark points marked on a template, which is used in registration [261].

Conde *et al.* use SVM classifiers trained on spin images for a purely 3D approach [262]. As their proposed method requires great computational resources, they constrain the search for the landmarks by using apriori knowledge about the face. In [231], 3D information plays a secondary or support role, in filtering out the back-

Table 5.1. 2D Landmarking Methods

Reference	**Coarse Localization**	**Fine Localization**
Chen *et al.* [240]	Gaussian mixture based feature model + 3D shape model	
Cristinacce *et al.* [252]	Assumed given	Boosted Haar wavelet-like features and classifiers
Smeraldi, Bigun [254]	30 dimensional Gabor response of each point + SVM	Gabor responses of the complete retinal field + SVM
Feris *et al.* [255]	Template matching using Hierarchical Gabor Wavelet Network (GWN) representation of faces	Template matching using GWN representation of features
Lai *et al.* [242]	Color segmentation for skin and lip + edge map	Vertical projection of thresholded coarse image
Shakunaga *et al.* [256]	PCA on canonical positions of features + structural matching	PCA
Ryu, Oh [243]	Vertical and horizontal proj. of face edge map	PCA on edge coordinates+ MLP for template matching
Shih, Chuang [234]	Edge projections + geometric model of facial features	Not present
Arca *et al.* [230]	Color segmentation for skin and lip + SVM	Geometrical heuristics
Zobel *et al.* [257]	DCT + Geometrical heuristics + Model of feature locations	Not present
Gourier *et al.* [258]	1st and 2nd Gaussian derivatives + clustering to 10 centroids	Not present
Antonini *et al.* [259]	Corner detection	PCA and ICA of windows at the corner points + SVM

ground, and to compute intra-feature distances in geometry-based heuristics. In [225], 3D information is used to assist 2D in filtering out the background, and a comparison between 2D and 3D landmarking methods under controlled illumination conditions indicates superiority of the 2D approaches. However, 3D methods are more robust under adverse illumination conditions [227]. Table 5.2 summarizes 3D approaches to facial landmark localization.

Table 5.2. 3D Landmarking Methods

Reference	**Coarse Localization**	**Fine Localization**
Li, Corner, Paquette [260]	Curvature and geometry based heuristics	
Boehnen and Russ [231]	Cascaded smoothing, minimum and z-filtering+ 2D and 3D geometry	
Colbry, Stockman, Jain [237]	Interpoint statistics + heuristics	Shape index+Harris edge detector
Conde *et al.* [262]	Curvature analysis + heuristics	Spin images + SVM
İrfanoğlu, Gökberk, Akarun [161]	ICP based registration	Curvature- and surface normal-based heuristics
Çınar Akakın *et al.* [225]	2D IMoFA-L + GOLLUM	IMoFA-L projection vs. DCT coefficients on 2D and 3D + SVM
Salah, Akarun [227]	3D IMoFA-L + GOLLUM	Not present

5.3. Description of the Model

In our model, we follow a saliency based scheme that works on a coarse to fine scale for efficient use of computational resources. Following [3], we prepare the input to the coarse system by downsampling the high-resolution input image. The advantage of downsampling is two-fold: We decrease the computational burden, and we make the feature detection easier. Patches are cropped from the downsampled images, and modeled with a statistical, unsupervised model. During testing, the statistical models are used to produce *conspicuity maps* from the downsampled image that indicate probable locations for the landmark. If there is more than a single feature channel, the conspicuity maps are summed up to a *saliency master map* for each landmark. This is the case in our 2D-based system, detailed in Section 5.5.1.

The unsupervised feature model represents the top-down part of the system. Each landmark has its own distribution, and the saliency maps are processed for each landmark separately. This relates to Sahbi and Boujema's face recognition system, where the idea is to pass overlapping windows over the face images to find salient features, indicated by high entropy [263]. There, each salient feature is expressed as a Gaussian distribution around a mean feature. Then for matching two images, a dynamic space warping scheme is employed, where each feature matched constrains the rest of the features by imposing an ordering. We have used the mixture of factor analysers model, described in Chapter 4, to reduce the number of parameters and to exploit the flexibility due to employing a mixture model. Yang *et al.* have shown previously that a mixture of factor analyzers outperforms PCA in a face detection application [264].

The detection of local features is a difficult problem, and there will be mistakes. We propose a structural analysis subsystem that detects correctly localized landmarks, and interpolates others if necessary. The interpolation is complemented with a local search on the saliency master map. Once the coarse features are in place, the high-resolution image is used for fine-tuning the landmark locations. The complete system is schematized in Fig. 5.1.

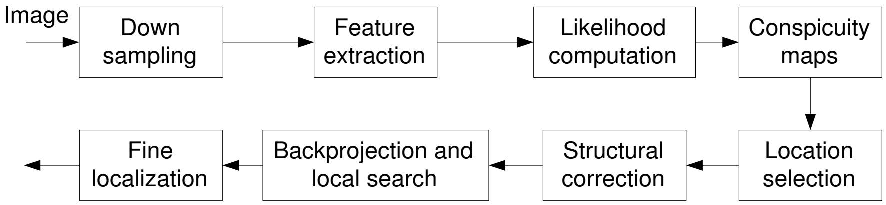


Figure 5.1. The proposed saliency based landmarking model

Before describing each step of this model individually, we take a moment to look at the related work in face detection. We focus on the use of feature saliency in particular. We hope that some of the design decisions we take (e.g. Gabor wavelets) will be more evident in the presence of related and competing schemes.

5.4. Related Work in Saliency Based Face Detection

The landmarking method we develop relies on feature saliency as computed by a feature learning system. There is a body of work, closely related to ours, which aims at automatic detection of faces in images. In these methods, saliency measures are used to constrain a fine-grained search, pointing out to regions of interest. There are also applications of face detection methods to automatic facial landmark detection. These algorithms usually employ face colour models and Gabor features in conjunction.

In a coarse-to-fine approach, Senior proposes to use a colour thresholding scheme for initial face localization, followed by a Fisher linear discriminant that evaluates rectangular candidate regions for faceness [265]. A multi-resolution pyramid is employed to search across different scales, and the mean intensity is subtracted from each pixel to increase robustness under varying illumination. The features are then searched in expected ellipsoidal regions, and the highest valued points are retained. The distribution of normalized pairwise feature distances are modeled with Gaussian densities on the training set, and the collection of feature points is pruned while monitoring the log-likelihood change in the joint density as features are removed one at a time.

In spatial visual attention, an influential hypothesis is that the feature-specific representations are modulated by the top-down component of attention [266]. In [267], the edge, colour and symmetry information sampled in six different scales contribute to a bottom-up saliency map. Adapting this model to face detection, Ban and Lee enhance it by a top-down intensity ratio model based on the relatively stable average intensity ratios of face regions [268]. A face colour model is added to the final map as a third component (See Fig. 5.2). For the bottom-up stage, the contributions of different feature maps in different scales are found by independent component analysis (ICA) in both papers. There is also a biological justification of using ICA, in that it makes the detected features as independent as possible.

In an automatic landmark localization scheme, Herpers *et al.* use first and second derivative Gaussian steerable filters for edge and corner detection in a number of orientations [233]. Elaborate models of the facial features are used to guide the search. In a later work, an attentive scheme is employed to constrain the detailed search to smaller areas [269]. The top-down influence is implemented by emphasizing the portion of the bottom-up saliency map that hopefully contains the next feature each time a

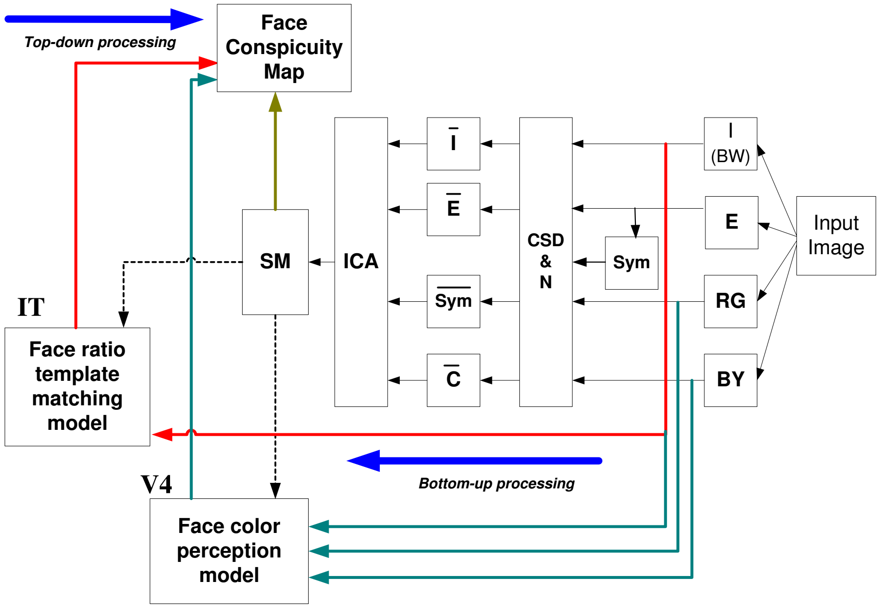


Figure 5.2. Ban and Lee's saliency based face detection model, adapted from [268].

new feature is searched. A feature graph is generated from the feature point candidates and a simulated annealing scheme is used to find the distortion relative to the canonic graph that results in the best match. The authors caution that "the large number of features that are derivable by the filtering scheme does not allow a *classical* computational detection strategy."

Gabor filter outputs are similar to simple cell outputs in V1 region of the human visual cortex. In [270], a contrast filter that resembles the profiles of retinal ganglion cells is used before the Gabor filter. The output values of Gabor filters (probability distribution) are estimated from a pooled 9×9 fragment database, obtained from 1.500 TV images. The models for faces are averages of 20 pictures per person. The ensuing saliency maps are compared for classification. Liu *et al.* also employ Gabor filters in 3 frequencies and 8 orientations for classification, but a fine sampling is only applied at landmark locations found by PCA on the training images [271]. Other points used in assisting the decision are sampled at a coarse interval.

Smeraldi and Bigün used support vector machine (SVM) classifiers to describe the Gabor responses around facial features [254]. A retinotropic grid is designed by placing 50 points in concentric circles, and 30 Gabor filter responses (in 5 frequencies and 6 orientations) are computed for each such point. Either the Gabor responses on the facial feature spot (local model), or the whole response to the retinotropic grid (extended model) is used for classification. The feature vectors are normalized for contrast, and since their distribution is said to be difficult to model statistically, SVM is selected as the choice classifier. An EER of 0.3 per cent is reported on eyes and mouth localization for the authentication task on the M2VTS database.

We have proposed a generative approach for modeling landmark features. In contrast, a number of face detection models employ supervised, discriminative techniques. For these models it is necessary to use negative samples along with positive samples in training [240, 272, 273, 254]. Serre *et al.* use a much greater number of non-face patterns in training their face detection system [272]. They also note that features learned from faces work better than features learned from the whole data; supervised guidance helps detection, as one would expect. In [273] the authors Z-normalize the intensity before applying Gabor filters to localize salient edges. For each landmark, 200 positive and 200 negative samples are used in training a neural-network classifier.

Chen *et al.* also use a great number of negative samples to train the classifiers in a chained boosting scheme [240]. A Gaussian mixture model is used to model the thresholded map that indicates possible feature locations. This is found useful, as it was observed that the maxima in the feature map come with a number of other maximum points in their proximity. The 3D shape constraints obtained during the training is used to select a subset of feature points consistent with the generic face. Even with a very good local feature model, we expect interfering local maxima in the feature map. The structural correction in our model assumes that this interference is relatively rare, and the true feature is correctly localized in most of the cases. It also deals with missing landmarks under this assumption. In another work that takes the missing landmarks into account, Sobottka and Pitas use the intensity minima at each row of the image as potential face feature candidates and use a fuzzy membership function to search for the best feature constellation [244]. The authors note that poorly detected landmarks correlate with each other; e.g. if the eye is difficult to detect, so is the corresponding eyebrow.

In Craw *et al.*'s FindFace system, a number of single and compound feature experts are designed and consulted in a dynamic order based on expert success [274]. A model expert stores a canonical model of the object to be recognized (i.e. the face), and a context expert affinely matches the current candidate features to the model to restrict search area for new features. In our model, restricting the search area is achieved by the coarse-to-fine search.

The depth data are rarely employed in conjunction with saliency based models. In [275] and [276] the authors describe a saliency-based 3D object recognition system mounted on the Kurt3D autonomous mobile robot. The 3D and 2D data are processed similarly, in an image pyramid with 5 different scales, from which 6 intensity maps and 4 orientation maps are derived. These maps are combined to conspicuity maps in their respective modalities. Finally, the mode specific maps are combined to produce a master saliency map (See Fig. 5.3). For the combination, a weighted sum is adopted, where the weight of a map is inversely proportional to the number of local maxima it contains. This scheme credits unusual features, and allows pop-out effects.

The idea of using feature saliency is a recurring theme in face detection literature, as it is evident from the short survey presented in this section. Gabor wavelets are the most frequently used features for intensity (or 2D) images. 3D features are relatively unexplored, primarily because of the drawbacks of 3D sensors, as discussed in Section 3.2. In the next section, we give a detailed report on the feature extraction part of our model.

5.5. Feature Extraction

One of the assumptions of the present work is that we have access to 2D and 3D facial information simultaneously. In this section we describe the methods we have adapted to exploit these types of information for coarse-scale feature extraction. Typically, the images are downsampled by a factor of eight before feature extraction.

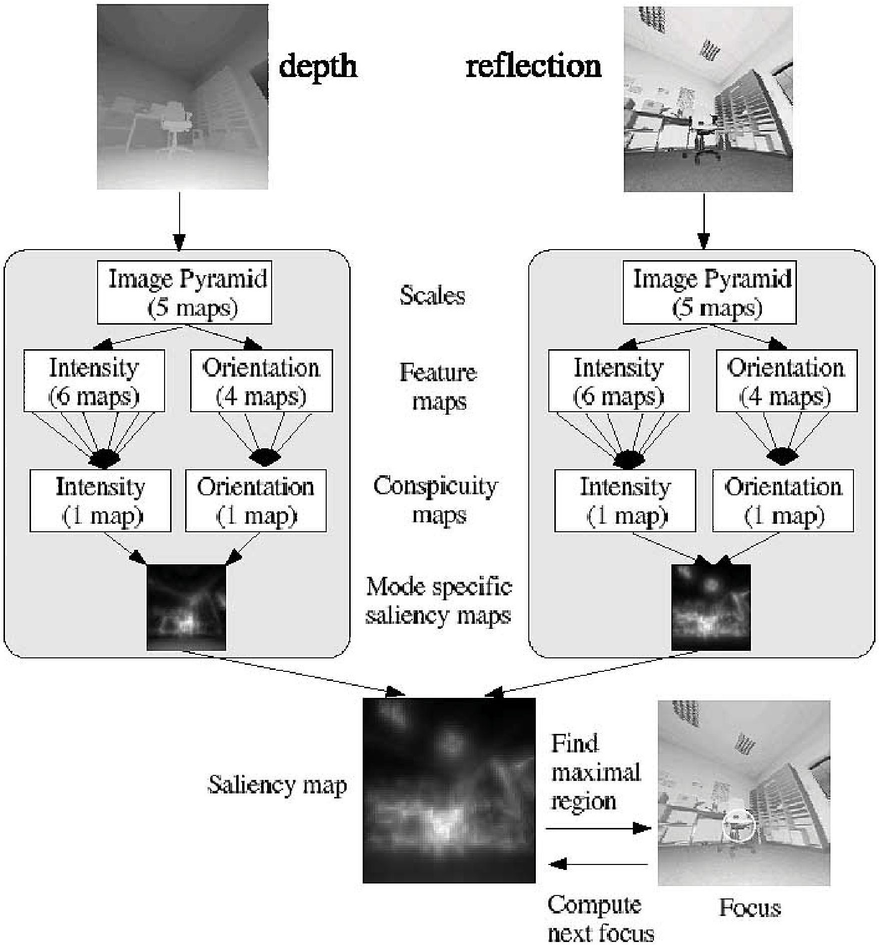


Figure 5.3. Frintrop *et al.*'s model, from [275].

For the FRGC dataset, theis brings the images from 480 × 640 to 60 × 80.

5.5.1. 2D Scheme

We have used Gabor wavelets for the 2D scheme, which have a biological counterpart [97]. It is also known that they have good performance in face detection and facial feature localization [270, 277, 278, 254, 279]. From a point $\boldsymbol{x}$ on the image, the following feature is computed using a Gabor kernel $\boldsymbol{k}_j$:

$$\Psi_j(\boldsymbol{x}) = \frac{\boldsymbol{k}_j \boldsymbol{k}_j^T}{\sigma^2} \exp\left(-\frac{\boldsymbol{k}_j \boldsymbol{k}_j^T \boldsymbol{x}\boldsymbol{x}^T}{2\sigma^2}\right) \left[e^{i\boldsymbol{k}_j\boldsymbol{x}} - e^{\frac{\sigma^2}{2}}\right] \tag{5.1}$$

The Gabor kernel in this expression is determined by two parameters, the orientation and the scale, respectively:

$$\boldsymbol{k}_j = (\boldsymbol{k}_{jx}, \boldsymbol{k}_{jy}) = (\boldsymbol{k}_v \cos\phi_w, \boldsymbol{k}_v \sin\phi_w) \tag{5.2}$$

$$\boldsymbol{k}_v = 2^{-\frac{v+2}{2}}\pi \tag{5.3}$$

$$\phi_w = w\frac{\pi}{8} \tag{5.4}$$

The training samples were manually landmarked. Around each landmark, a 7×7 window was cropped for modeling. In our 2D localization scheme, for each cropped landmark patch, Gabor wavelets in eight orientations ($w \in \{0, 1, 2, 3, 4, 5, 6, 7\}$) and a single scale ($v \in 3$) are applied. Using more scales or neighbourhoods larger than 7×7 did not increase the success rate. 49-dimensional vectors obtained from each Gabor channel are min-max normalized and separately modeled with IMoFA-L.

During testing, the likelihood scores are computed for each feature window of the face area, for each mixture model separately. These are the conspicuity maps, which are then summed up to a saliency master map to determine the most likely location for each landmark. For the FRGC dataset, the 3D mask is used to eliminate the background in the 2D model. For the BANCA dataset, we assumed that the face is roughly localized in a rectangular bounding box.

Fig. 5.4 shows the Gabor wavelet outputs in eight orientations for a sample image. The windows sampled around each location are normalized and used to calculate the likelihood of that location with respect to a particular orientation and a particular landmark. The likelihood-based saliency maps for the first landmark are shown in Fig. 5.5. All maps of different orientation are summed up to produce the saliency map for the given landmark. The saliency maps for seven landmarks, their maxima and the locations on the corresponding intensity image are shown in Fig. 5.6.

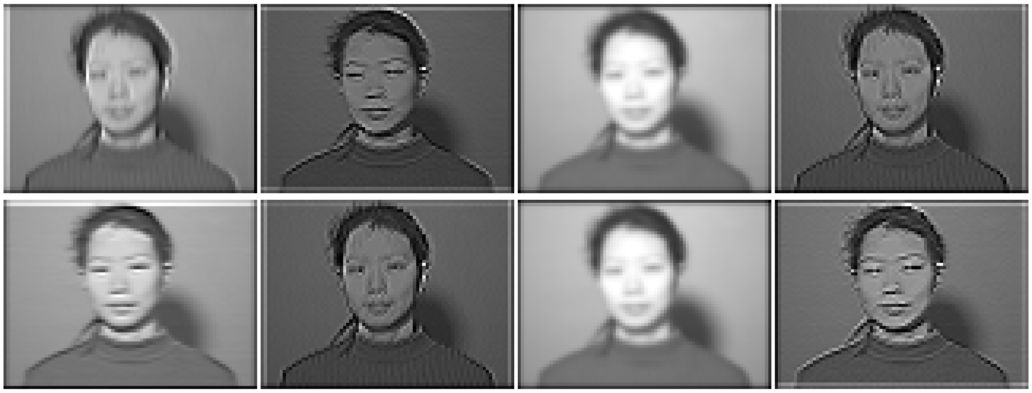


Figure 5.4. Gabor wavelet filter outputs in eight orientations

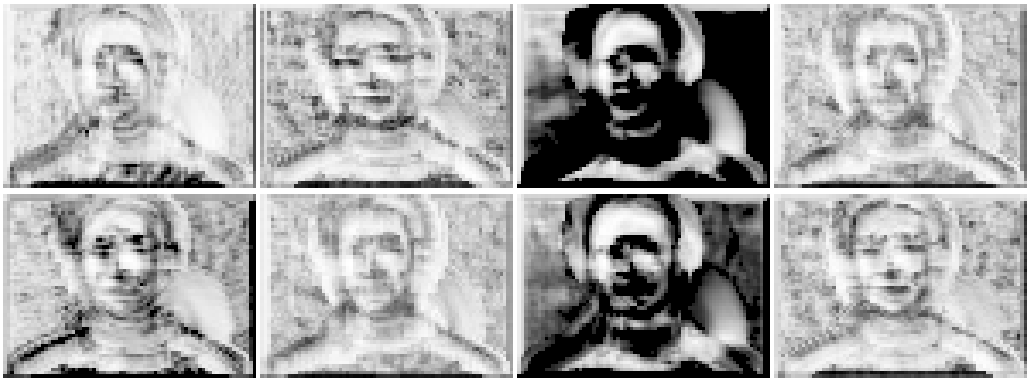


Figure 5.5. Conspicuity maps obtained from IMoFA-L outputs for the first landmark in eight orientations

The trained IMoFA-L models are depicted graphically in Fig. 5.7. There is one mixture for each facial landmark at each orientation. The landmarks are denoted by their abbreviations: left eye outer corner (leo), left eye inner corner (lei), right eye inner corner (rei), right eye outer corner (reo), nose (n), mouth left corner (ml), and mouth right corner (mr), respectively. For a given mixture, the number of boxes in the figure gives the number of components in the mixture, and the sizes of the boxes are proportional to the number of factors for a given component. Complex patterns in the data are modeled with more components, and with more factors per component (e.g. mouth corners), whereas simple patterns (e.g. outer eye corners) are modelled with a smaller number of parameters.

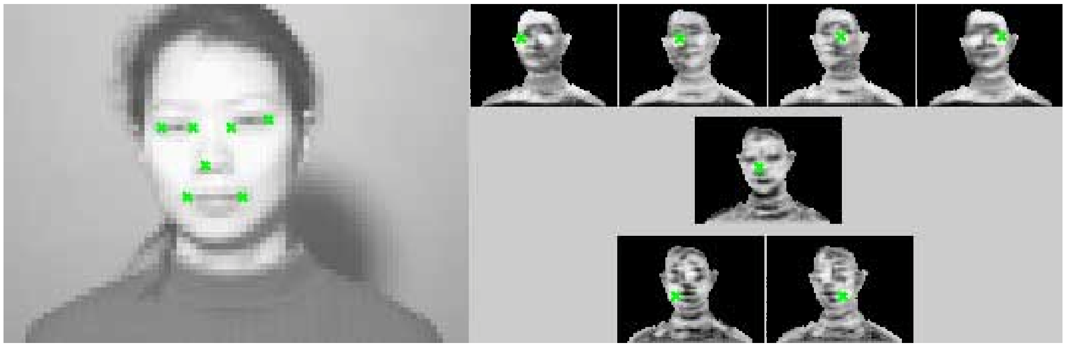


Figure 5.6. The correctly located landmarks on the original texture and the summed conspicuity map for each landmark. The crosses on the conspicuity maps stand for the location with the highest salience.

If we look at the second orientation channel (second row of squares), the mixture for the outer corner of the left eye (leo) is a depicted as a single, large square, which means there is a single component with a large number of factors. The number of factors stands for the intrinsic dimensionality of the feature distribution, and a large number means there is variation in many directions. Conversely, if the number of components is large, as in the left mouth corner (ml) for the same orientation, there exist a number of different structures, modeled as clusters in the feature space. IMoFA-L allocates these components and factors automatically.

Regarding Fig. 5.7, there is another observation we shall make. The models allocated to similar landmarks (i.e. ml-mr, leo-reo, and lei-rei) have similar parameter distributions. The training of each mixture is performed independently, and there is a certain amount of randomness involved in the training procedure due to initialization. However, we can see from the figure that the method is robust enough to converge to similar outcomes for similar feature sets.

5.5.1.1. Baseline Methods. We have implemented two baseline methods from the literature for local feature evaluation. In 2D landmark localization, Lades *et al.* have employed a similarity metric that is based on the comparison of Gabor wavelet jets, sampled in five different frequencies and eight different orientations [280]. Each jet is thus a 40-dimensional feature set. Jets obtained from the landmark locations of about

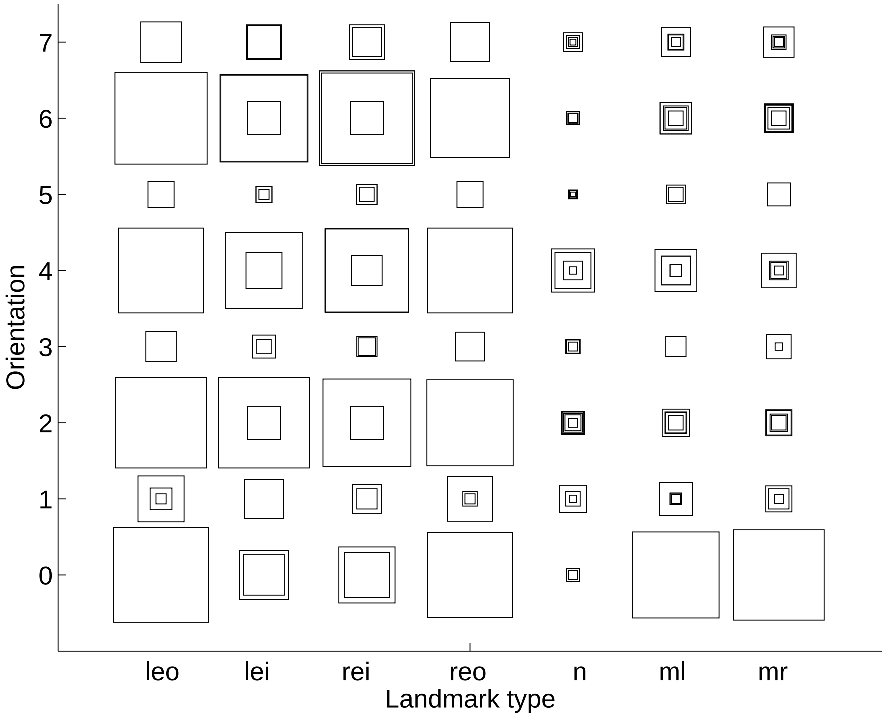


Figure 5.7. The trained IMoFA-L models, for each landmark type and each Gabor orientation. Each box stands for a component, and the size of the box is proportional to the number of factors in that component.

70 training samples are stored in a *bunch*. When analysing a new face, Gabor jets are extracted from each candidate point, and compared to all the stored jets in the bunch, using a special jet similarity measure. The smallest distance to any of the jets in the bunch is taken as the distance value for that location, and the local search is driven to minimize the distance along with a structural constraint.

We have constructed a bunch from the training samples, and used the proposed jet similarity measure to localize the best candidate:

$$S(J', J) = \frac{\sum_j a_j a'_j}{\sqrt{\sum_j a_j^2 \sum_j {a'_j}^2}} \tag{5.5}$$

where a_j denotes the magnitude of the Gabor wavelet for a particular orientation and frequency, and the subscript j indexes all 40 Gabor features. Wiskott *et al.* extend this in [29] by considering the phase information as well:

$$S_{\phi}(J', J) = \frac{\sum_j a_j a'_j \cos(\phi_j - \phi'_j)}{\sqrt{\sum_j a_j^2 \sum_j {a'_j}^2}} \tag{5.6}$$

where ϕ_j is the phase.

In the elastic bunch graph matching approach, an elastic face graph that relates the positions of located landmarks to the stored templates is combined with the local similarity scores obtained by comparing the jet extracted from the face point to all the jets in the bunch and retaining the highest similarity value. In our comparisons, the bunch graph was learned from the training set, but since our purpose is to evaluate the goodness of the local analysis metric, we have left out the coefficient due to the elastic graph, and evaluated the metrics exhaustively within a maximum neighbourhood. This means that the bunch-based method does not search in a particular direction in our case, but automatically finds the best location. Thus, we have an upper bound on the usefulness of the jet-similarity measure. The elastic graph (i.e. the structural information) is not integrated at this point, because we need many more landmarks (about 30-40 is used) for the training of the elastic graph.

When comparing the methods, we have centered a search window of variable size on the true landmark, and gradually increased the search window size. The search begins with a window of size 3 × 3, where the maximum possible error is $\sqrt{2}$ pixels. As the search window is enlarged, new candidates far from the true landmark become available. Thus, the plot of neighbourhood size versus pixel error is necessarily monotonically increasing. A flat and low curve means that the error does not increase, even we search among more and more candidate points. Figure 5.8 shows that for coarse-level search, our local features have much wider basins of attraction compared to baseline methods around the true landmarks. We should also note that the bunch methods are slower in general.

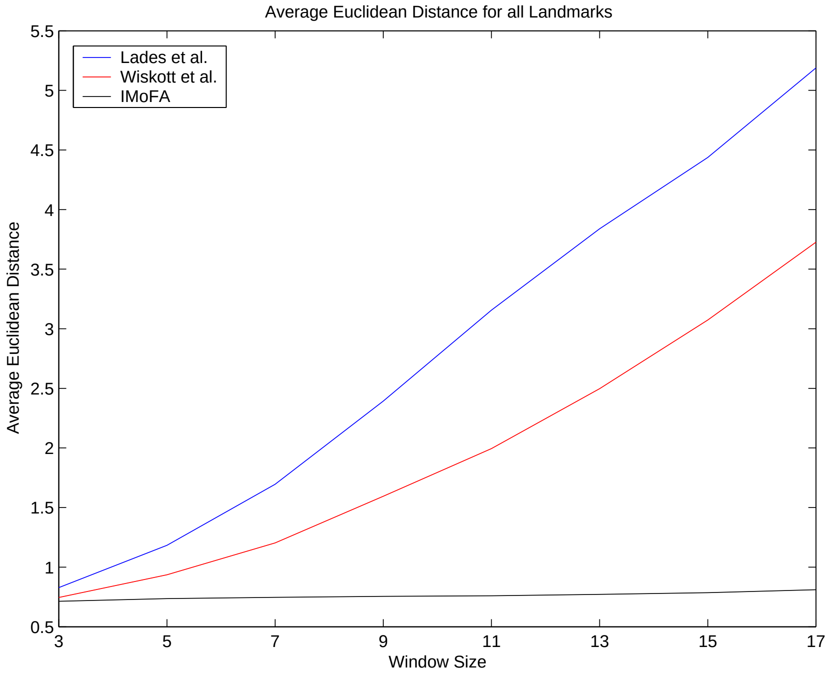


Figure 5.8. Coarse localization errors (in pixels) for increasing sized search basins around true landmarks. When the search basin is large, IMoFA-L based similarity does not deviate from the landmark as the Gabor jet similarity metrics proposed in [280] and [29].

The most plausible explanation for the relative success of our method is that both bunch-based methods are template matching methods, while the IMoFA-L is a generative method that models the feature distribution probabilistically. Also, it should be noted that increasing the number of training samples is beneficial for the IMoFA-L method, but not recommended for the bunch methods, as the increased number of comparisons linearly increase computation time for only marginal improvement [29]. This observation was validated by our simulations.

5.5.2. 3D Scheme

The FRGC dataset contains noisy 3D data acquired with a Minolta 910 scanner. The provided images contain a flag that indicates the pixels for which depth information exists. We would like to fill in the missing depth values within the face region by interpolation, and eliminate spike artifacts. If we directly attempt to use the flag, all the background would be tagged as lacking depth information. We process this flag to obtain a mask for the texture information by eliminating the gaps and irregularities with a closing operation, followed by an opening operation. A 21×21 diamond-shaped structuring element was used (See Fig. 5.9). This second flag is used to control where we need to fill in depth information for the missing points.

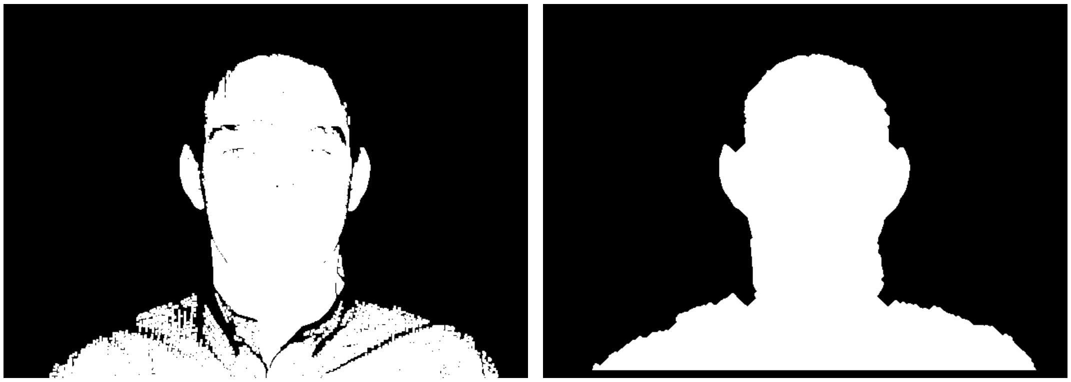


Figure 5.9. Closing and opening applied to the flag to eliminate gaps and irregularities.

For the missing points, all three dimensions are missing, although the x and y dimensions are more or less regularly sampled. Therefore we use different interpolation methods for each dimension. Noticing that for the FRGC data, the x values are regularly sampled along horizontal lines, and the y values similarly on vertical lines, we collect valid points along these dimensions to fit a line for interpolation (See Fig. 5.10). A higher degree polynomial can also be employed to accomodate slight bends in the data.

The depth values pose a more difficult problem, as the locations of artifacts are not trivially indicated by the flag information (see Fig. 5.11). A sharp decrease (or increase) in the depth values does not necessarily indicate an artifact, as the facial boundary will induce sharp changes as well. We have modeled the depth values with Gaussian distributions to detect artifacts as outliers. Our aim was to avoid median or mean filtering that would introduce unnecessary smoothing into the real values. This procedure did not result in successful removal, as the depth values are only locally

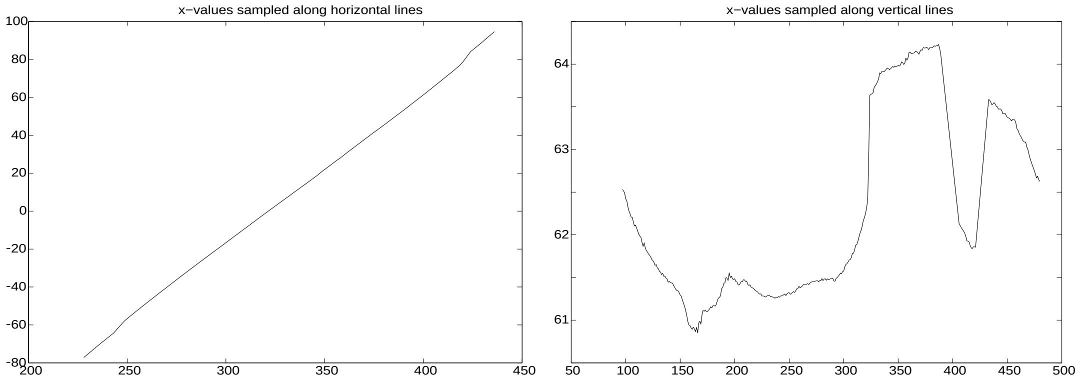


Figure 5.10. Interpolation of x values for missing points. The sampling direction is relevant, as the sampled values are regular in the horizontal direction only.

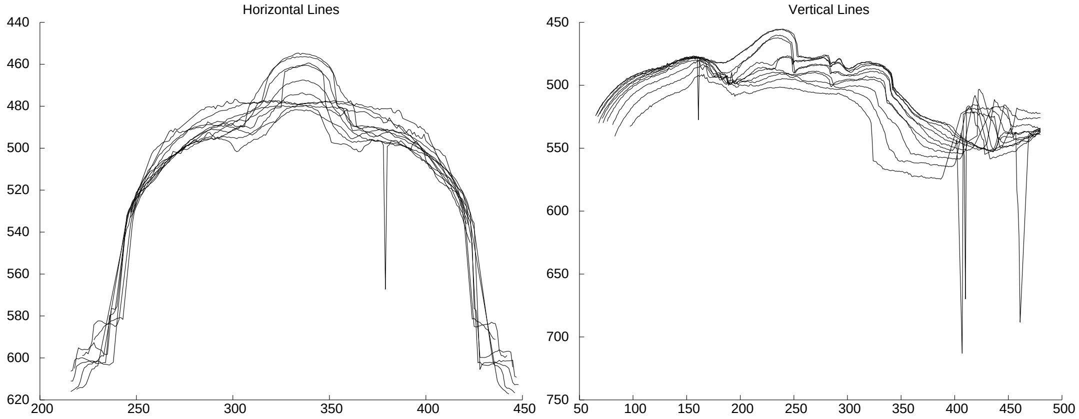


Figure 5.11. Even after we eliminate the points with no depth value, the depth values sampled in vertical and horizontal lines show artifacts that should be determined locally.

meaningful. Consequently we have reverted to median and mean filters. We have tried different mean and median filters for the correction of depth information. Windows of sizes 3, 5, 7, and 9 were tested for each type of filter. Fig. 5.12 show the resulting corrections along vertical and horizontal lines. We select the 9×9 median filter, followed by a 9×9 mean filter in our final implementation. Whenever the depth values are missing in greater quantities, the window size is automatically enlarged.

After the preprocessing, we extract features from the 3D depth map. The most simple feature is cropped patches of the range image itself. In the coarse search,

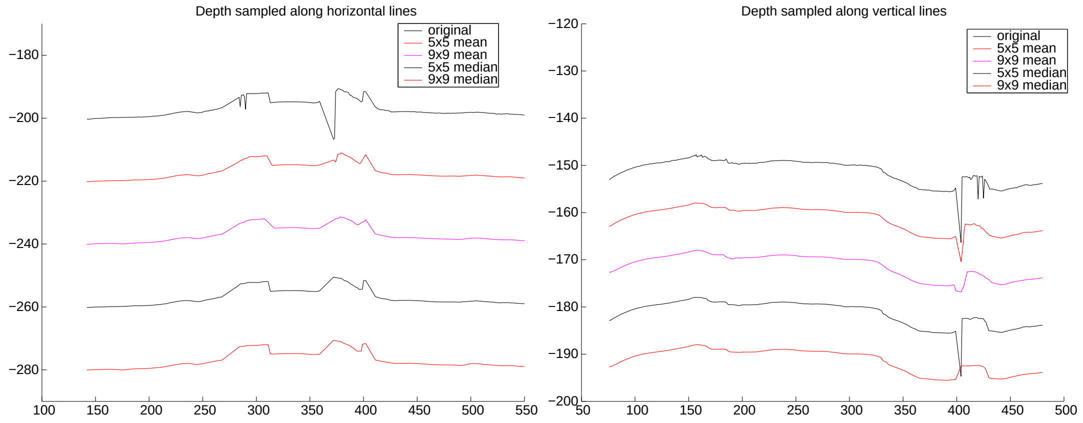


Figure 5.12. Mean and median filters of various window sizes used for artifact elimination in depth data are shown along horizontally and vertically sampled points.

employing surface normals, curvatures, or shape indices does not pay off, as these features come with a high computational cost. Our experiments have shown that the curvature computation is sensitive to surface noise, and requires a very robust preprocessing. We model 3D depth patches with the IMoFA-L model, as in the 2D scheme, to obtain a single conspicuity map. The patch size is fixed at 7 × 7.

5.5.3. 3D-Assisted 2D Scheme

The 3D information can also be employed indirectly, to alleviate the effect of illumination from the 2D images. Recent results indicate that the set of appearances for convex Lambertian objects under different illumination conditions can be approximated by a low-dimensional manifold [281, 282]. This manifold is 9-dimensional, and spanned by the first nine spherical harmonics. Since the human face is assumed to be a convex Lambertian object, this model can be employed for illumination correction in human faces [283, 284, 198, 285].

5.5.3.1. Spherical Harmonics. Spherical harmonics are an orthogonal set of solutions to Laplace's equation, represented in spherical coordinates:

$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2\frac{\partial f}{\partial r}) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}(\sin\theta\frac{\partial f}{\partial \theta}) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 f}{\partial \psi^2} = 0 \tag{5.7}$$

where the relation between the spherical coordinates (r, θ, ψ) and the Cartesian coordinates is expressed as:

$$\begin{aligned} x &= r \sin\theta \cos\psi \\ y &= r \sin\theta \sin\psi \\ z &= r \cos\theta \end{aligned} \tag{5.8}$$

In Cartesian coordinates, Eq. 5.7 is equivalent to

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = 0 \tag{5.9}$$

and the solutions are twice-differentiable real-valued functions.

5.5.3.2. Illumination Approximation by Spherical Harmonics. The following analysis is from [285]. Assuming a Lambertian distant illumination model $L(w)$ and ignoring cast shadows and near-field illumination, the irradiance E of the face is a function of surface normals n. According to [282], $E(n)$ can be expressed as an integral over the upper hemisphere:

$$E(n) = \int L(w)(n.w)dw \tag{5.10}$$

E is scaled by the surface albedo $\lambda(p)$ to produce the radiosity I, which corresponds to the image intensity:

$$I(n, p) = \lambda(p)E(n) \tag{5.11}$$

Here, p denotes a point on the surface. It is shown in [281] and [282] that E can be approximated by a combination of the first nine spherical harmonics $H(x, y, z)$ for Lambertian surfaces:

$$h_{00} = \frac{1}{\sqrt{4\pi}} \tag{5.12}$$

$$h_{10} = z\sqrt{\frac{3}{4\pi}} \tag{5.13}$$

$$h_{11}^{e} = x\sqrt{\frac{3}{4\pi}} \tag{5.14}$$

$$h_{11}^{o} = y\sqrt{\frac{3}{4\pi}} \tag{5.15}$$

$$h_{20} = \frac{1}{2}(2z^2 - x^2 - y^2)\sqrt{\frac{5}{4\pi}} \tag{5.16}$$

$$h_{21}^{e} = 3xz\sqrt{\frac{5}{12\pi}} \tag{5.17}$$

$$h_{21}^{o} = 3yz\sqrt{\frac{5}{12\pi}} \tag{5.18}$$

$$h_{22}^{e} = \frac{3}{2}(x^2 - y^2)\sqrt{\frac{5}{12\pi}} \tag{5.19}$$

$$h_{22}^{o} = 3xy\sqrt{\frac{5}{12\pi}} \tag{5.20}$$

To compute the intensity image under the approximation given by the spherical harmonics for a point p with surface normal $n = (n_x, n_y, n_z)$ and albedo λ, one can replace x, y, z with n_x, n_y, n_z in Eq. 5.11.

5.5.3.3. Texture Recovery. For texture information recovery, the following method is proposed in [285]. Under the assumptions of the previous sections, the intensity image t of a face can be expressed as

$$t = B * l \tag{5.21}$$

where B are the basis images used to synthesize the texture, and l is the vector of illumination coefficients. B can be expressed as

$$B = H(n_x, n_y, n_z).\lambda \tag{5.22}$$

where H is the approximation to the reflectance function in Eq 5.11. We are looking for the texture λ from which the image intensity can be synthesized under a given illumination model. The iterative algorithm used to compute λ is summarized here:

Basis computation: The initial albedo λ for each vertex is set to the average intensity of the image. The surface normals n are estimated at all vertices. Then we compute the first nine basis images B and the spherical harmonics $H(n)$ for reflectance function using Eq. 5.21, Eq. 5.22 and the spherical harmonics given in Eqs. 5.12-5.20.

Illumination Coefficients Estimation: The set of illumination coefficients l is updated by solving a linear system of equations:

$$t = Bl \tag{5.23}$$

At each step of the iterative algorithm, this equation is solved using the fixed image intensity t and the current basis B. The basis changes in the second step, with each updeate of the albedo.

Texture Recovery: In this step, we update the albedo by solving

$$t = H(n_x, n_y, n_z)l.\lambda \tag{5.24}$$

When we solve this equation, we obtain a new estimate of λ. As λ depends on both the value of the albedo prior to this calculation (λ_{cur}) and the illumination parameters, the update of the current albedo is done by taking a linear combination:

$$\lambda = (1 - \eta)\lambda_{cur} + \eta(t/(H(n_x, n_y, n_z)l)) \tag{5.25}$$

Eq. 5.24 and Eq. 5.24 are solved iteratively by setting η to 0.5 and increasing it by 0.1 at each iteration. Thus, the convergence is obtained with five iterations.

Figure 5.13 shows a sample image from the FRGC database for which the albedo image is estimated with this method.

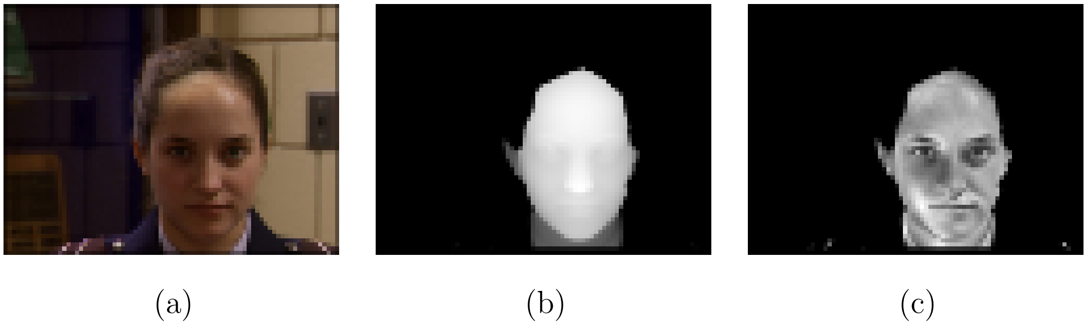


Figure 5.13. (a)The original 2D image. (b)Corresponding depth map. (c)Recovered albedo on pixels with valid depth.

5.6. Structural Analysis Subsystem

In a typical affine-invariant shape normalization scenario, the center of mass of the landmarks is translated to the origin, the landmarks are scaled to a fixed average distance to the origin, and rotated to satisfy some direction criterion. This type of normalization makes the stored facial images directly comparable. However, errors in landmark localization also corrupt the normalization: an outlier will shift the center of mass, change the scale, and the normalized landmarks will not conform to the general shape. Furthermore, if these transformation parameters are determined from an erroneous set of landmarks, the resulting registration will not be usable. As we will show in Chapter 6, 3D registration methods are sensitive to landmarking errors. In this section, we describe a novel structural correction algorithm that makes normalization robust to individual landmarking errors. Additionally, the algorithm we propose allows the estimation of a reliability measure for landmark detection, and helps us correct mistakes automatically.

5.6.1. The GOLLUM Algorithm

In this section we describe our structural analysis method that detects and corrects erroneous landmarks. In an earlier study, Burl and Perona have assumed that false alarms are distributed independently from each other and are independent from the feature location [286]. If this assumption is correct, a structural model that searches features at their expected locations will be able to single out false alarms. However, this is a simplifying assumption, and a more truthful model can take the correlations between false alarms (e.g. moustaches cause failures at both mouth corners) at the expense of being more complex. Burl and Perona have modeled the joint distribution of the landmark coordinates with a single multivariate Gaussian, and they base the affine correction of the face solely on the eye landmarks. However, their scheme fails if the eyes are incorrectly detected in the first place. It is possible to make the system more robust by not assuming the correctness of any particular landmark.

In our system, the configuration of the landmarks are assumed to be related by an affine transformation, i.e. if we could register the face with a canonical face, they would be placed consistently, within a small margin of error due to individual variation. The structural correction method we propose uses this property of consistent landmark configurations to identify a reliable subset of landmarks, called the *support set*.

When we inspected the performance of the saliency-based landmarking model on the validation set of FRGC ver.1, we have seen that the probability of finding a single landmark correctly is about 0.9. This means that the probability of finding at least three landmarks out of seven correctly is close to unity. Our design relies on the idea of using the landmarks jointly to remove the errors, instead of chaining the landmarks to propagate errors. During the training phase, we model the position of a landmark given the other landmarks, with a bivariate Gaussian distribution, making the complete model a mixture of bivariate Gaussians.

Traditional face graphs incorporate structural information in the framework of an energy minimization problem, where any deviation from the nominal distance between landmarks (conceptualized as edges of the face graph) is penalized. These approaches bring in a non-uniform energy gradient around the landmark, as perturbations of landmark locations affect the graph edge lengths in ways dependent upon the direction of the perturbation [237, 280, 287, 288]. Emphasizing a directionality for perturbations is meaningful if a large number of landmarks that follow common contour segments are modeled, as displacing a landmark in one direction makes the next landmark more likely to be displaced in that direction. However, for a small set of landmarks, this sort of constraint imposed by the face graph is not justified. Our simulations indicate that the Gaussian mixture model is very successful in modeling the landmark distributions.

In our proposed scheme, subsets of located landmarks take turns as *support sets.* For each such support set, we perform normalization involving translation, rotation and scaling with respect to the subset coordinates. The subsets are taken three at a time and then the rest of the landmarks are compared to their relative expected locations. If the ensemble of landmarks gives a high structural fitting score after normalization, the support set is validated. Any incorrectly localized landmark in the support set will badly distort the positioning of the other landmarks during normalization, and result in a poor fitting score.

We learn the spatial distribution of the normalized landmarks for each possible support set, as each support set corresponds to a different normalization. If we have a support set size of i, and l landmarks in all, the number of possible support sets is $\mathcal{C}(l, i)$, where $\mathcal{C}(.)$ is the combination operator. For example, a support set of size three from within seven landmarks results in 35 support combinations. For each such combination, we model the distribution of the remaining landmark positions (after normalization) with a mixture of Gaussians. In the testing phase of a support set, the likelihood of the non-support feature locations is calculated.

For a small number of landmarks, the number of possible support sets is not big. For a large number of landmarks, it will be sufficient to consider only a subset of all possible support sets, chosen according to landmark reliability. For each such set, we model the distribution of the landmarks after normalization with a mixture of bivariate Gaussian distributions. Then for any test sample, under a particular support set hypothesis, we are able to find the likelihood easily.

The normalization of the landmarks involves a translation that brings the centroid of the support set to the origin, followed by a scaling that sets the average distance of the landmarks in the support set to the origin to $\sqrt{2}$ and a rotation that aligns the first landmark in the support set with the y-axis. Since the first landmark of the support set varies on a single dimension only, it is not usable in likelihood calculations in the test phase. In order to remedy this situation, we use a further minimization step that rotates the landmarks to maximize the joint likelihood under the model. If we denote the rotation function with $r(\boldsymbol{x}, \theta)$, the Gaussian distribution for landmark j with $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, and the normalized position of the landmark with $\boldsymbol{x}_j = [x_j, y_j]^T$, the joint likelihood to be maximized is expressed as

$$\mathcal{L} = \prod_{j=1}^{L} \frac{1}{2\pi|\Sigma_j|^{1/2}} e^{-\frac{1}{2}(r(\boldsymbol{x}_j,\theta)-\boldsymbol{\mu}_j)^T \Sigma_j^{-1} (r(\boldsymbol{x}_j,\theta)-\boldsymbol{\mu}_j)} \tag{5.26}$$

Maximizing this expression is equivalent to minimizing the following:

$$\min_{\theta} \sum_{j=1}^{L} (r(\boldsymbol{x}_j, \theta) - \boldsymbol{\mu}_j)^T \Sigma_j^{-1} (r(\boldsymbol{x}_j, \theta) - \boldsymbol{\mu}_j) \tag{5.27}$$

with

$$r(\boldsymbol{x}_j, \theta) = \begin{bmatrix} \cos\theta x_j - \sin\theta y_j \\ \sin\theta x_j + \cos\theta y_j \end{bmatrix} \tag{5.28}$$

In the 3D case, we only need to change the rotation expression accordingly:

$$r(\begin{vmatrix} x \\ y \\ z \end{vmatrix}, \theta, \phi) = \begin{vmatrix} \cos(\phi)\cos(-\theta) & -\cos(\phi)\sin(-\theta) & \sin(\phi) \\ \sin(-\theta) & \cos(-\theta) & 0 \\ -\sin(\phi)\cos(-\theta) & \sin(\phi)\sin(-\theta) & \cos(\phi) \end{vmatrix} \begin{vmatrix} x \\ y \\ z \end{vmatrix} \tag{5.29}$$

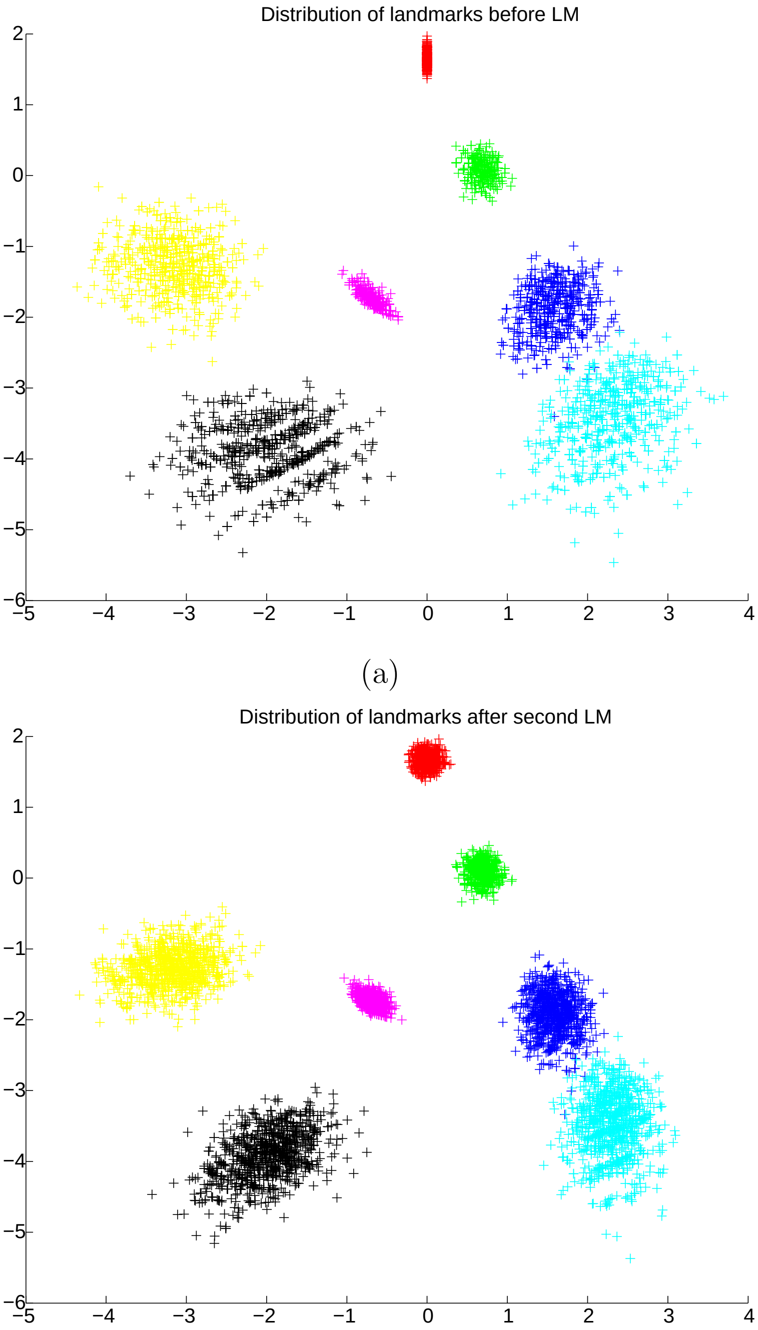
where θ is the azimuth, and ϕ is the elevation angle.

We have used a Levenberg-Marquardt procedure to minimize this expression. The first landmark in the support set is excluded from the minimization, as it varies in one dimension only, and consequently its covariance expression singular. However, since the initial location is very close to the global minimum, and the function is smooth, two or three iterations are sufficient to find the solution. Then the distributions of the landmarks are re-estimated, and the minimization procedure is applied again. If we continue iterating between estimation and maximization, we converge to the maximum likelihood solution for the joint data distribution under the support set. We only iterate twice to remove singularity in the distribution of the first landmark and to pack the remaining distributions tightly together. Fig. 5.14 shows the distribution of landmarks under a particular support set before and after the Levenberg-Marquardt step. The particular landmarks belonging to the support set can be discerned by their smaller variance.

During testing of a support set, if at least one landmark outside the support set is *acceptable*, then the corresponding support set is accepted. A non-support landmark l_j is assumed to be acceptable, if its likelihood under the model is higher than a threshold:

$$\mathcal{L}(l_j, \boldsymbol{\mu}_j, \Sigma_j) > \tau(k) \tag{5.30}$$

Incorrect landmarks are replaced by new the backprojection of the expected landmark location according to the saliency (see Fig. 5.15). Since each support set corresponds to a different normalization, the expected location of the missing landmark is potentially different for each support set.



(a)

(b)

Figure 5.14. The distribution of landmarks after normalization (a) without Levenberg-Marquardt step (b) with Levenberg-Marquardt step

We conceptualize the threshold $\tau(k)$ as isodensity lines around its expected location, as illustrated in Fig. 5.16. We heuristically determine the threshold in the following manner. The covariance can be visualised as an ellipsoid around the data distribution. We scale the covariance matrix by a scalar k, and obtain a larger ellipsoid. To obtain the likelihood value on any point of this ellipsoid, we select an

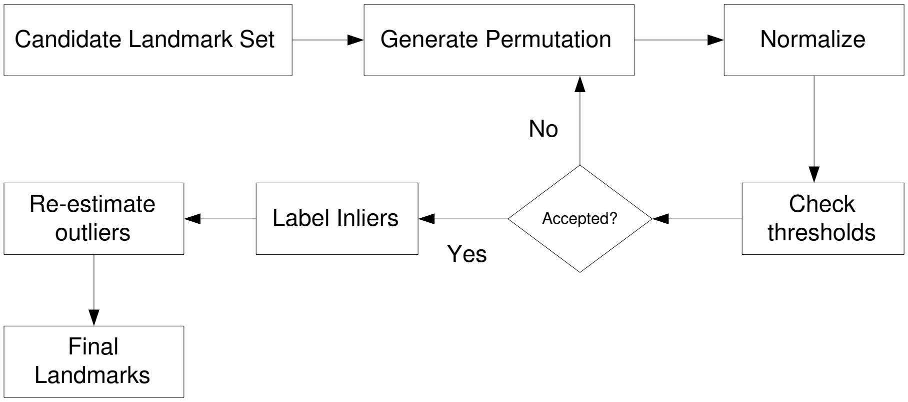


Figure 5.15. Structural analysis subsystem

eigenvector of the covariance matrix, and displace the mean in that direction by an amount proportional to the square-root of the corresponding eigenvalue. This gives us the threshold:

$$\tau(k) = \mathcal{L}(\boldsymbol{\mu}_j + \boldsymbol{v}'_j\sqrt{\boldsymbol{u}'_j}, \boldsymbol{\mu}_j, \Sigma'_j) \tag{5.31}$$

where Σ'_j is the scaled covariance matrix $\Sigma'_j = k^2\Sigma_j$, $\boldsymbol{u}'_j$ is an eigenvalue of Σ'_j, $\boldsymbol{v}'_j$ is the corresponding eigenvector, and $\mathcal{L}$ is the likelihood function, as defined previously. For our simulations we have chosen $k = 4$, although larger (but not smaller) values can be considered. Setting k to a smaller value means that the structural subsystem will label more points as outliers, and will be forced to re-estimate them. Conversely, if k is too large, some of the close outliers will be missed. Fig. 5.16 shows the threshold boundaries across each landmark for a given support set. The likelihood is equal on all points of the threshold line for a given landmark, but not across landmarks. We name our method GOLLUM, short for Gaussian Outlier Localization with Likelihood Margins.

The non-support landmarks are labeled as acceptable or unacceptable according to their agreement with learned models under that support set. An unacceptable landmark, say the n^{th} landmark that yields a likelihood score below threshold, is replaced by the *backprojection* (i.e. reverse rotation, scaling and translation with respect to the local coordinate system) of the expected landmark location, $\boldsymbol{\mu}_n$. The support sets

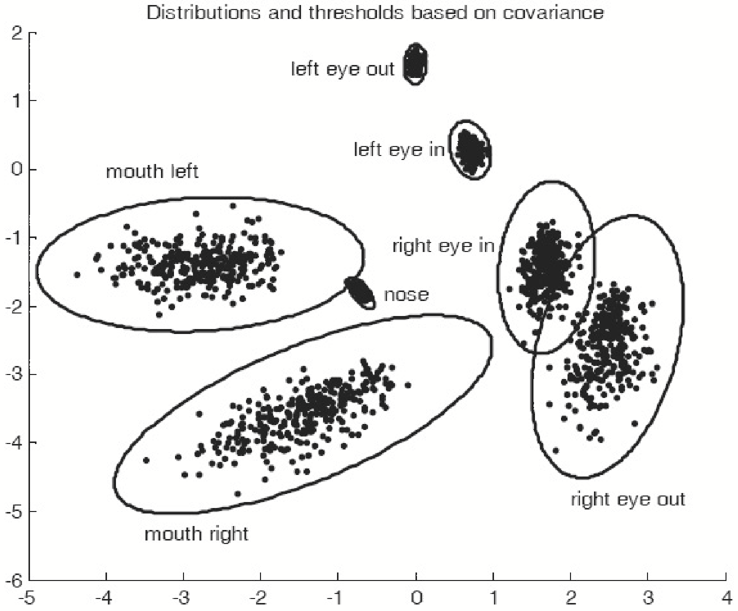
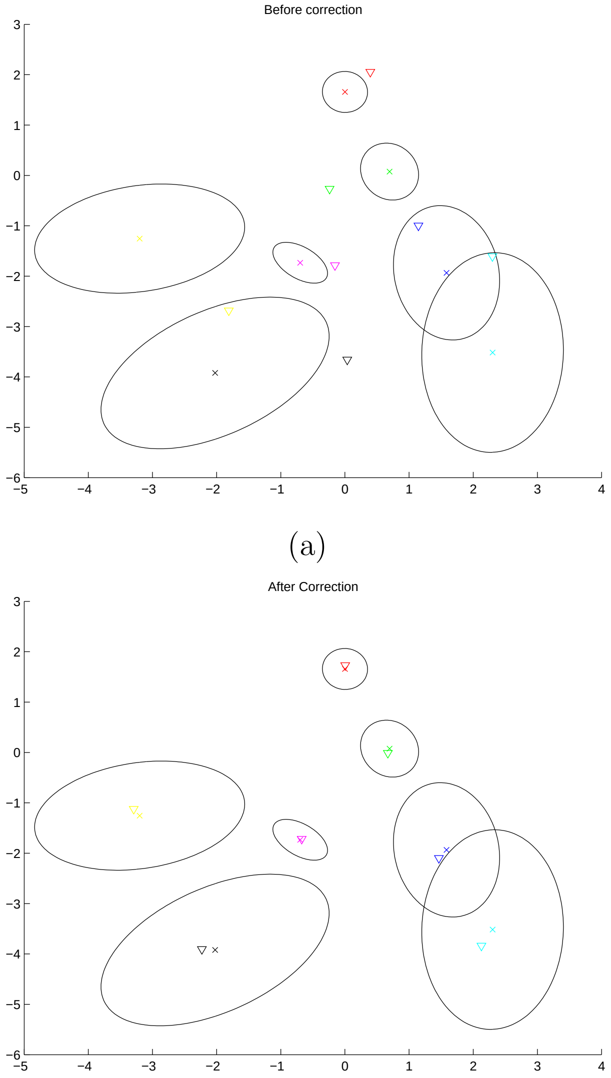


Figure 5.16. The thresholds for outlier detection are shown as ellipsoids around landmark clusters. The support set consists of the corners of the left eye and the nose. As the normalization is based on the landmarks of the support set, they have smaller variations.

are ordered according to their relative reliability, and searched until the non-support landmarks validate one of them. For example the first permutation to be tested in our simulations has the corners of the first eye and the nose tip in its support set. When the mislocated landmark is in the support set, the remaining landmarks are off the mark by a large margin, and the support set is changed. Fig. 5.17 illustrates this case.

GOLLUM works on the locations of landmarks, and it is independent from the actual image properties. This allows landmark detection in absence of features (See Fig. 5.18). An additional benefit of the GOLLUM is that the landmarks that do not conform to any of the stored permutation patterns are labeled as such. These are the cases where the saliency scheme failed for some reason, and it is important to be able to detect such failures as well.



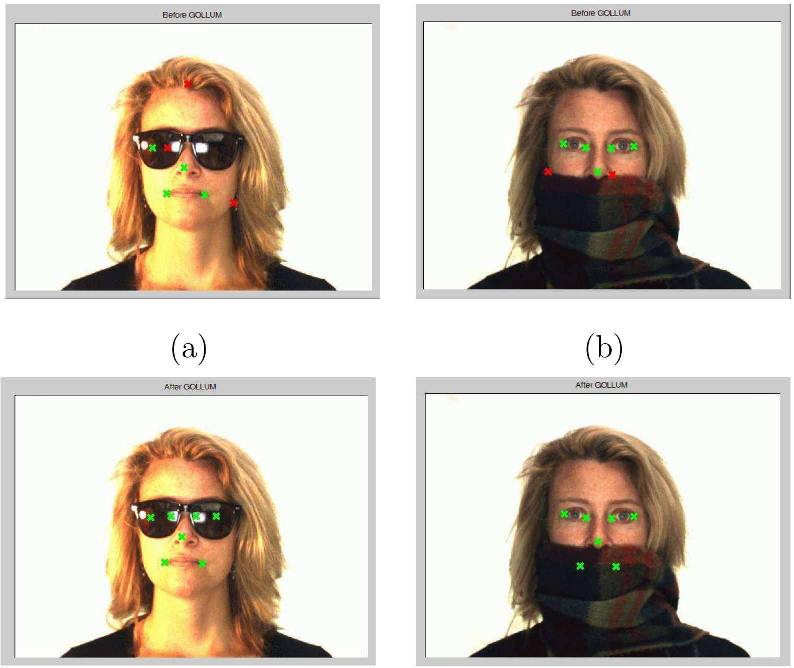
(a)

(b)

Figure 5.17. (a) The support set contains a wrong landmark, most of the landmarks are unacceptable. (b) The permutations are tested until a good support set is found. For this sample, the corrected location agrees perfectly with ground truth.

5.6.2. The BILBO Algorithm

In a recent paper, Beumer *et al.* proposed an iterative structural correction scheme with a similar purpose [251]. The proposed algorithm (BILBO) first registers landmark locations to an average shape. During training, the registered landmark locations are perturbed with small rotations, translations and scalings. Then a singular value decomposition is used to compute a lower dimensional subspace. During testing, the landmark locations are projected to this subspace and back. Deviations



(a) (b)

(c) (d)

Figure 5.18. Correction of landmarks in the absence of features with GOLLUM. These images are best viewed in colour.

from the average shape are corrected when passing through the bottleneck created by the subspace projection. A threshold value is monitored to detect the change due to backprojection. This threshold is increased at each iteration, and the algorithm stops once the change is smaller than the threshold.

We have constrasted our structural correction method GOLLUM with BILBO. We have used the parameter settings indicated in [251]. The results are reported in Section 5.8.

5.7. Fine Landmark Localization

For the fine level (i.e. 480 × 640 images in FRGC), a window around the coarse landmark location is searched for the best candidate. We have used another batch of IMoFA-L mixtures to compute the best candidate, trained on patches of fine-resolution images. For the FRGC database, we have selected a 9×9 search window. Note that since the upsampling factor is eight, a 9×9 window essentially corresponds to searching an area 10 per cent larger than the corresponding coarse-level pixel found by the coarse landmarking scheme. We have evaluated window sizes up to 41×41; and observed that window sizes larger than 9×9 deteriorate performance. For the more challenging BANCA dataset, the accuracy peaked for a larger search window.

Fig. 5.19 shows the application of the IMoFA-L method to fine-level images. We have tested 2D Gabor features, 3D depth features, and a combination of these features. The score-level 2D+3D fusion was accomplished with the SUM rule [289]. The generative approach does not perform sufficiently well for the fine-level images for a simple reason. The patches cropped from the fine-level images are close to uniform textures when they are small, hence they are not discriminating enough. Conversely, cropping larger patches increases the feature dimensionality quadratically, necessitating a training set with much more samples for proper feature learning. Since the number of training samples are fixed, the scheme performs sub-optimally. In [228], a 2D scheme based on DCT coefficient templates is successfully applied for the fine level landmark detection on FRGC ver.1 dataset. However, our simulations show that 3D is more robust than 2D. Subsequently, we have implemented a simple 3D method based on the first and second depth gradients.

5.7.1. Gradient-based Approach

The gradient information can be used for a cheap and discriminative way to determine the fine positions of landmarks. We consider the first (Δ_x and Δ_y) and the second gradients (Δ_{xx}, Δ_{xy}, Δ_{yx} and Δ_{yy}) of the depth map in x and y directions for this purpose. The gradient maps are thresholded to produce a binary map. The threshold value θ depends on the scale of the depth information, for our purposes we have considered $\theta \in \{0, \pm 0.1, \pm 0.3, \pm 0.5, \pm 1\}$ for the first gradient, and $\theta \in \{0, \pm 0.03, \pm 0.06, \pm 0.09, \pm 0.15\}$ for the second gradient. Three types of thresholding were tested: an upper bound (with a negative threshold value), a lower bound (with a positive threshold value) or a double threshold where only the values above θ or below $-\theta$ were retained. Binary kernels were used to compute feature value at each location. We tested odd kernel sizes of 7, 9, 11, and 13.

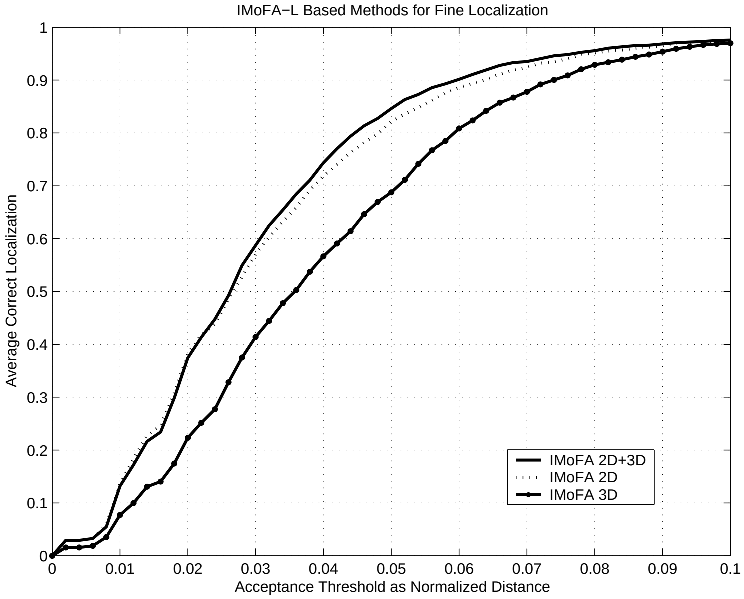
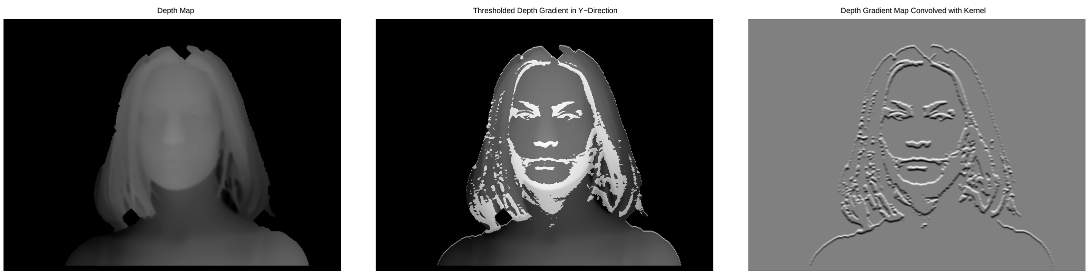


Figure 5.19. Application of IMoFA-L on fine level images. 2D, 3D and their fusion are shown.

To give an example, consider the following kernel that promotes horizontal edges:

$$h = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{5.32}$$

We demonstrate the effect of applying this kernel on the range image depicted in Fig. 5.20 (a). When we compute the depth gradient Δ_y in the vertical direction, and apply a threshold, we highlight locations where the decrease in depth values in the vertical direction is greater than the threshold, as shown in Fig. 5.20 (b). Applying the kernel to this thresholded map produces the edges with sharply decreasing depth, as in Fig. 5.20 (c). The output of the final convolution will be maximized for different features depending on the kernel shape.


Depth Map
Thresholded Depth Gradient in Y−Direction
Depth Gradient Map Convolved with Kernel

(a) (b) (c)

Figure 5.20. (a)The depth map. (b)Thresholded gradient map. (c)Convolution with the kernel.

For each parameter setting, the optimum kernel (in the sense that the response is maximized on the training set) is selected by cropping patches from the ground truth landmark locations of the thresholded gradient images of the training set, computing their average, and thresholding to assign either $+1$ or -1 to each location of the patch. This kernel gives the greatest activation over the training set for that particular parameter setting (i.e. kernel size, gradient type, threshold type, threshold value, and landmark). If we look at the parameter space, we have six types of gradient maps, 14 different threshold settings, and four kernel sizes for each landmark. This means there are 336 kernels to evaluate on the training set.

To speed up the training, we monitor the success of each kernel during training and eliminate candidates as the training proceeds. The success of the kernel is measured by the average localization error in a 41×41 neighbourhood around the true landmark (i.e. the basin of attraction). For each landmark we select the best kernel, and convolve a neighbourhood around the coarse landmark with this kernel. The location with the maximum response is selected as the best feature. For a more reliable estimation, we compute the second-best kernel in a cascade fashion, where we add the convolution result to the previous result. The second kernel is selected to maximize the total response of the ground truth locations in the training set.

5.8. Simulation Results

For the majority of our experiments, we have used the first part of the Notre Dame University 2D+3D face database (FRGC ver.1) [135]. There were 943 images, of which half were used for training, one quarter for validation, and the rest for the test sets. Samples with poor 2D-3D correspondence were left out to treat all methods fairly. In this section, the results are reported separately for each different landmark type. The same structural subsystem corrections are applied to landmarks located with 2D, 3D and 2D+3D (ALBEDO) methods. Table 5.3 shows the coarse localization accuracies for each landmark type when the acceptable distance to ground truth is less than or equal to three pixels on the downsampled image.

Table 5.3. Localization results for the first experiment

Method	*Outer Eye Corners*	*Inner Eye Corners*	*Nose*	*Mouth Corners*
2D	96.9 %	98.0 %	98.7 %	94.6 %
2D+BILBO	98.0 %	97.1 %	98.7 %	94.0 %
2D+GOLLUM	**99.3** %	**99.6** %	**100.0** %	**99.3** %
3D	87.9 %	98.4 %	96.7 %	85.4 %
3D+BILBO	89.7 %	98.2 %	96.9 %	88.8 %
3D+GOLLUM	95.7 %	99.3 %	98.2 %	88.1 %
ALBEDO	37.0 %	84.8 %	59.2 %	58.8 %
ALBEDO+BILBO	43.1 %	84.3 %	60.1 %	59.6 %
ALBEDO+GOLLUM	72.7 %	87.7 %	78.9 %	72.4 %

It is observed that 2D performs better in localizing outer eye corners and mouth corners. When coupled with the structural correction subsystem, the performance of 2D and 3D systems are close. Since the 2D information is richer, we expect it to produce a more accurate system when the training and test conditions are similar. Our simulations show that the proposed GOLLUM scheme outperforms it competitor BILBO.

The albedo corrected images lose their discriminative power, and perform suboptimally. We have observed that illumination effects are not completely removed by the albedo correction technique. Very poorly illuminated images do not lend themselves to a full recovery. We shall note that the surface normal calculation is also a potential source of error. The resolution of the laser scanner is high, and the 3D surface is not very smooth if inspected closely. Consequently, the surface normals are not very reliable either. An additional disadvantage is the computational cost of surface normal computation.

We have used the FRGC ver.2, Fall 2003 dataset for a more challenging experiment. This dataset contains 1893 2D+3D images from the same set of subjects, acquired six months later under expression variations and different lighting conditions, some of them so challenging that even the manual landmarking is difficult. Without suitable illumination compensation, the 2D statistical model is not expected to generalize correctly. However, 3D information is expected to be robust to illumination changes. We have directly applied the IMoFA-L models previously learned on ver.1 to this new dataset (denoted with "ver.1" in the Method column). Table 5.4 gives the localization results at an acceptance threshold equal to three pixels.

The system based on 2D features fails in the absence of adequate illumination compensation, whereas 3D depth features produce good results. The left and right ends of the horizontal crevice between the lower lip and the chin produce false positives for the mouth corners in 3D, and since this pattern conforms to the general face configuration it is very difficult to detect. This is the source of most of the mouth corner errors. The decrease in the mouth corner detection accuracy is partly due open-mouthed expressions in ver.2. The albedo correction increases the recognition accuracy for some landmarks, but there is no overall improvement. Fig. 5.21 shows some samples of the 3D-based system on FRGC ver.2.

Perhaps the illumination conditions in FRGC ver.2 are too poor to allow any robust learning? To test this hypothesis, we have randomly divided ver.2 into training, validation and test sets. The last three rows of Table 5.4 report the accuracies of the

Table 5.4. Localization results for the second experiment

Method	*Outer Eye Corners*	*Inner Eye Corners*	*Nose*	*Mouth Corners*
2D ver.1	18.4 %	9.9 %	0.2 %	31.8 %
2D ver.1+BILBO	17.0 %	15.5 %	1.4 %	29.9 %
2D ver.1+GOLLUM	18.4 %	10.8 %	1.8 %	31.7 %
3D ver.1	78.3 %	97.2 %	96.7 %	20.1 %
3D ver.1+BILBO	79.3 %	96.3 %	96.8 %	37.8 %
3D ver.1+GOLLUM	83.4 %	97.1 %	98.0 %	29.3 %
ALBEDO ver.1	3.5 %	12.9 %	1.5 %	21.5 %
ALBEDO ver.1+BILBO	4.1 %	15.1 %	2.6 %	20.6 %
ALBEDO ver.1+GOLLUM	3.9 %	12.8 %	2.3 %	21.2 %
2D ver.2	92.5 %	97.0 %	96.7 %	84.4 %
2D ver.2+BILBO	92.0 %	97.3 %	96.6 %	82.3 %
2D ver.2+GOLLUM	98.8 %	98.9 %	99.6 %	94.2 %

models trained and tested on ver.2, and these are higher than the reported 3D scheme results. It is clear that the 2D model is powerful enough to learn the features, even in difficult illumination conditions, as long as training conditions prevail during testing.

5.8.1. Justifying IMoFA-L

The mixture model we use is an elaborate statistical model, and its use should be justified. We have contrasted the IMoFA-L model with Gaussian distributions with full covariance matrices on FRGC ver.1, to see whether a single Gaussian could perform as well as the mixture. Fig. 5.22 shows the resulting correct identification rates. Since IMoFA-L model uses a separate validation set, we have trained two Gaussian models; One with the training set, and another with the training set enhanced with validation samples, respectively. IMoFA-L outperforms both, for less than half the number of parameters. In [240] the right mouth corner and nose images are mirrored to enhance the training set. Such an extension of the training set by including symmetry images

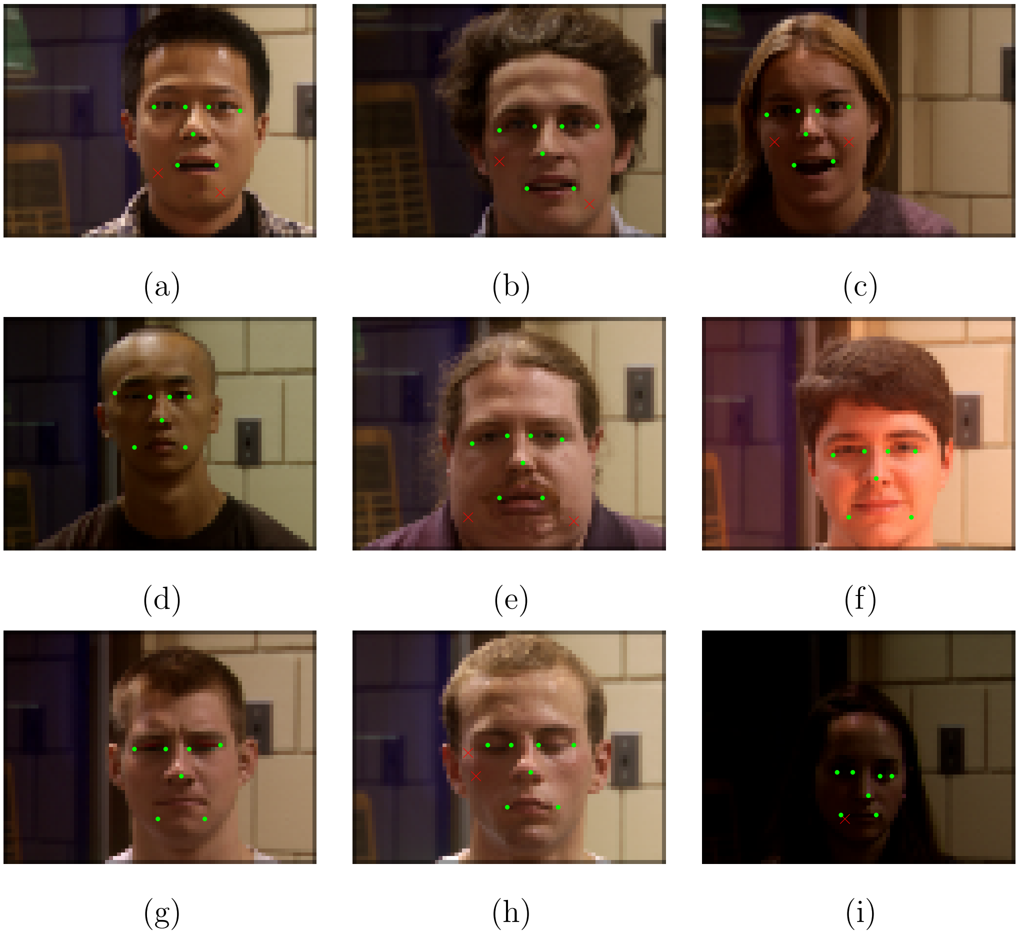

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 5.21. Samples from FRGC ver.2 Fall 2003 dataset, landmarked with the 3D system. The green dots are detected landmark positions after structural correction. The red crosses are initial mouth corner locations, corrected by the structural analysis subsystem.

did not lead to better results in our simulations, although we should take into account the fact that the training set was already of a moderate size to begin with. We have also considered the probabilistic nature of the IMoFA-L algorithm by training ten models for the same training set to assess the effect of initial conditions. The model that produced the highest likelihood on the validation set was considered as the final model, but the detection results were similar for all models.

An additional experiment was performed to assess the contribution of IMoFA-L by comparing it with a popular Gaussian mixture model (GMM) approach, proposed by Figueiredo and Jain (FJ) [290]. In this model the user still specifies the covariance

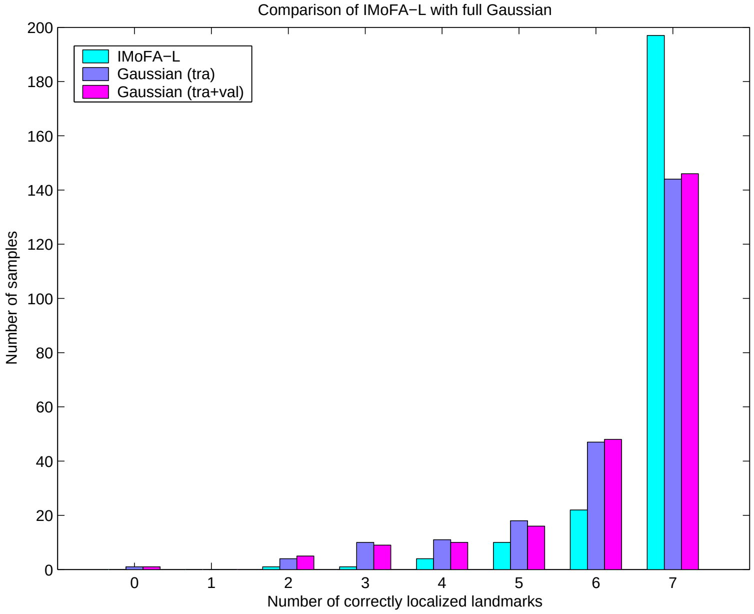


Figure 5.22. Comparison of IMoFA-L and Gaussian distributions for saliency maps. For 235 test samples, and out of seven landmarks, the IMoFA-L model correctly locates all landmarks in about 200 cases.

shape (full, diagonal or shared), but the number of components is determined automatically. The proposed method starts by initializing a large, user-specified number of components randomly. The GMM is trained with the component-wise EM algorithm, and the components with vanishing support are annihilated as the algorithm proceeds. Models with different number of components are tested, down to a minimum number of components. The algorithm then selects one model among these by looking at a minimum description length (MDL) based criterion. The FJ model was used in [291] for facial feature detection. Our reported results are for the 2D model trained with the ver.2 data. As Table 5.5 shows, the IMoFA-L model that is trained and tested with the same data (denoted by 2D v2) outperforms all the FJ models consistently. The full-covariance Gaussian models have higher accuracies than the diagonal covariance model, showing that the covariances between features are useful to the learning algorithm.

Table 5.5. Contrasting IMoFA-L with Figueiredo and Jain's method

Method	*Outer Eye Corners*	*Inner Eye Corners*	*Nose*	*Mouth Corners*
2D v2 IMoFA-L	92.5	97.0	96.7	84.4
2D v2 FJ-Full	90.6	96.5	95.9	80.7
2D v2 FJ-Shared	90.7	96.5	94.9	76.9
2D v2 FJ-Diag.	77.7	92.2	88.3	79.2

5.8.2. Parameter Selection and Extensions for GOLLUM

We have performed further experiments to determine the relevant parameters for GOLLUM, i.e. the support set size, and the number of landmarks used for validating the support set. In Table 5.6 we report our simulation results on the FRGC ver.1 [135] and BANCA [292] face image datasets. The columns of the table show localization accuracies at acceptance thresholds of 2 and 3 pixels, and the average number of tested support sets. The first row is the result without structural correction. The rest of the rows start with a header that indicates the number of landmarks used in the support set (3 or 2, respectively) and the number of landmarks used for validation. The method denoted with "best subset" evaluates all the support sets and selects the one with the highest likelihood and greatest number of inliers for re-estimation. Note that the number of average support sets is less than the maximum (35) for this method, as we allow early stopping in the case that a support set labels all the other landmarks as inliers. This is especially useful for the FRGC set, where the high-quality of images lead to better landmark localization.

The results reported in Table 5.6 confirm that using 3 landmarks for the support set is better than using 2 landmarks only. Selecting a single landmark for validation seems to be better than using two landmarks. With this method, the first landmark set that labels another landmark as inlier is used for re-estimation. In this case, a maximum of 3 landmarks (as opposed to 2 in the 2-val scheme) can be re-estimated. Apparently, our concern that a single inlier could arise randomly and mislead the algorithm was not grounded. Selecting the best subset leads to further improvements in accuracy. Although the BANCA result deteriorates at 3 pixel threshold, it improves at 2 pixel threshold. The more difficult BANCA set experiences a much greater increase in the average number of support sets, as the early stopping criterion is usually not satisfied.

Table 5.6. Experiments with the structural correction scheme

	FRGC			*BANCA*		
	2 pix. thr.	3 pix. thr.	avg. # sup. sets	2 pix. thr.	3 pix. thr.	avg. # sup. sets
Without correction.	95.99	96.78	0.00	80.38	81.62	0.00
3-supp, 2-val	96.84	98.24	1.78	89.24	92.67	4.91
3-supp, 1-val	97.08	98.72	1.22	89.81	**94.00**	2.98
3-supp, 0-val	97.14	98.97	1.12	88.00	92.38	1.81
3-supp, best subset	98.30	**99.45**	5.72	90.67	93.33	24.13
2-supp, 3 val	96.84	98.05	1.37	86.48	89.24	3.02
2-supp, 2 val	97.26	98.72	1.08	86.10	89.24	2.19
2-supp, 1 val	96.66	98.12	1.06	84.29	87.71	1.67
2-supp, 0 val	93.98	95.32	1.00	68.29	70.76	1.00
2-supp, best subset	97.57	98.72	3.62	85.43	88.76	11.67

We tested GOLLUM further by supplying it with systematically disrupted landmark information. We have deleted one or two landmarks from the FRGC ver.1 images, and supplied the system with erroneously locations that deviated from the ground truth at least by 10 pixels. Table 5.7 shows the percentage with which the landmarks were recovered by the structural analysis subsystem.

The structural correction can be complemented with a local search following the backprojection. There are two types of errors in the landmark localization. In the first type of error, the local feature is not prominent for some reason, and the landmark is missed. Mouth corners masked by a moustache are typical examples. In the second type of error, some other point of the image, usually a piece of clothing or hair, both rich sources of noise, is detected as global maximum, and the true landmark is only

Table 5.7. Recovery from one or two missing landmarks

	leo	**lei**	**rei**	**reo**	**n**	**ml**	**mr**
leo	99.45	99.21	99.21	99.27	99.33	98.60	99.09
lei		99.88	99.64	99.70	99.76	98.97	99.45
rei			99.76	99.51	99.64	98.97	99.27
reo				99.76	99.70	98.97	99.27
n					99.88	98.97	99.39
ml						99.21	98.72
mr							99.51

a local maximum. This second type or error stands a chance of correction via local search, whereas for the first type, the local search is not meaningful, and is likely to deteriorate the accuracy by taking the landmark away from the location that maximizes structural likelihood.

The simulation studies agree with this suggested behaviour. We have experimented with the 3D-based model on FRGC ver.2, and searched a local neighbourhood after backprojection. Table 5.8 shows that for increasing neighbourhood sizes, the localization accuracy improves with all landmarks except for mouth corners, with diminishing returns. The accuracy decreases for the mouth corners, as the local feature is inaccurate to guide the search. Note that the structural correction all by itself accounts for one third of the correct localizations for the mouth corners. This is an indication that mouth corner features are not reliably modeled with the depth features. The results for this table are slightly different than the results in Table 5.4, because the whole ver.1 is used in the training, with a smaller validation set made up of one sample per class.

Another possible enhancement is tested by making use of the facial symmetry. If we can determine the facial symmetry axis, the missing landmarks may be estimated by projecting their counterpart on the other side of the face. We use the landmarks labeled as inliers to estimate the facial symmetry axis. The nose and the centers of

Table 5.8. Experiments with the structural correction scheme

Landmark	Without corr.	GOLLUM	GOLLUM 3 × 3	GOLLUM 5 × 5	GOLLUM 7 × 7	GOLLUM 9 × 9
Eye out	71.15	79.72	82.30	84.93	85.77	85.89
Eye in	95.53	94.80	94.91	95.33	95.67	95.76
Nose	90.38	94.57	95.76	95.87	96.04	96.10
Mouth	20.79	30.23	29.19	27.60	26.30	25.42
Average	69.46	74.83	75.54	75.93	75.95	75.79
Symmetry	69.46	73.95	74.76	75.21	75.33	75.26

outer eye, inner eye and mouth corners are four points that are supposed to be on the facial symmetry axis. If we are confident about at least two of these points (but not only the eye corners) we can fit a line to these points and use it as our best guess of the symmetry axis. When the symmetry axis is not determined, the algorithm proceeds as before. The last two rows of Table 5.8 report the average localization success for the normal GOLLUM and its symmetry enhanced version. Although we have same improvement in the inner and outer eye corners for smaller local search sizes, the overall accuracy decreases. The symmetry axis is not too reliable for samples where we have to rely on structural correction.

5.8.3. 2D Experiments on BANCA

We have done a series of experiments with the 2D scheme on the English part of the BANCA dataset [292]. Of twelve sessions of the dataset, three sessions were selected as representative of the different illumination and background conditions. These are the first, fifth and the tenth sessions. In the tenth session, the camera is positioned lower than the first two selected sessions. The first session contains a uniform background. The dataset is difficult for landmarking, because a great number of subjects have eyeglasses that are specifically positioned to corrupt eye localization. However, the illumination conditions do not change as drastically as they do when switching from FRGC ver.1 to FRGC ver.2.

We have trained the 2D IMoFA-L models on a manually landmarked subset of this dataset. The first 50 subjects were used from each session. We used three images per subject per session, from which one was used for training, one for validation and one for testing. The one-to-eight downsampling ratio was retained, producing 72 × 90 images. A bounding box that offset the true face area (as described by the true landmarks) by five pixels on either size was used on the downsampled images to eliminate the background interference. Figure 5.23 shows the detection results with and without structural subsystem correction.

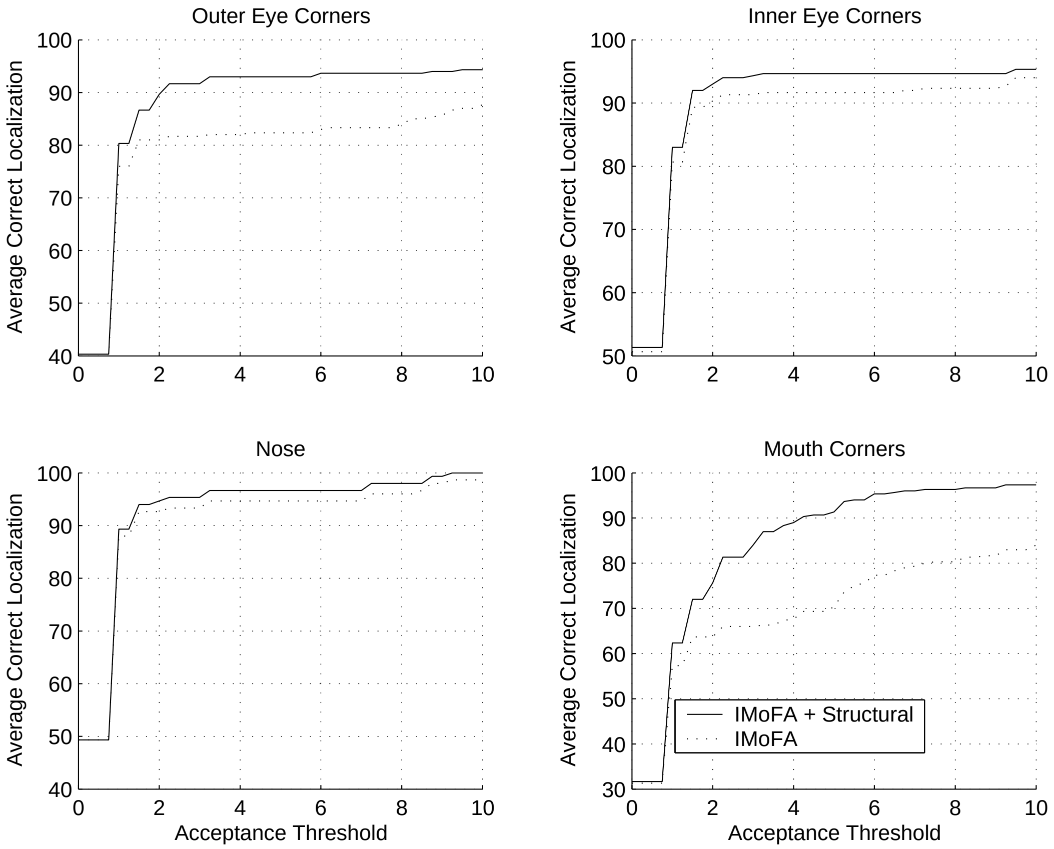


Figure 5.23. The application of the IMoFA-L model on the BANCA dataset with 150 test images.

Figure 5.24 shows samples from three sessions. In (a) the mouth corners are incorrectly found, but corrected to some extent by the structural analysis subsystem. In (b) the eyeglasses hide the outer left eye, and the mouth corner gets detected as the eye instead. Similarly in (c) facial hair masks the mouth corners, and eyebrows

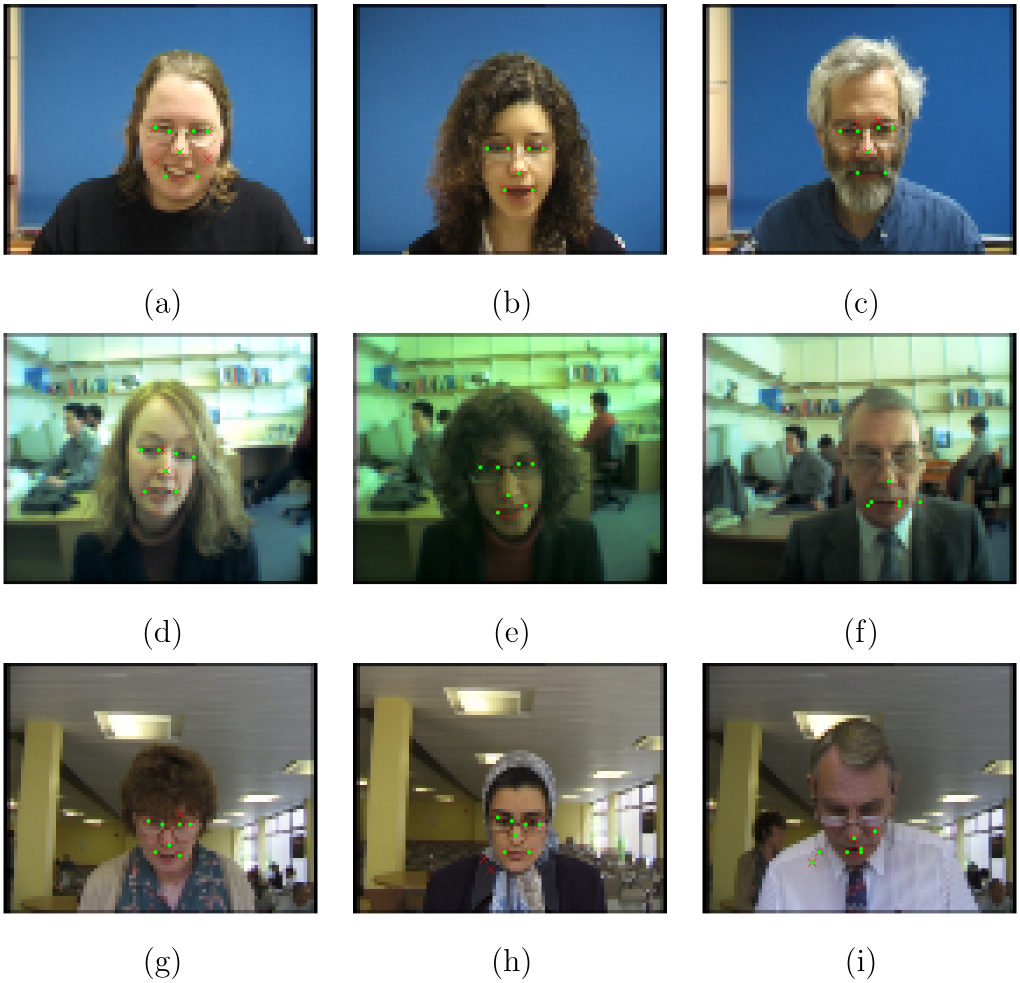


Figure 5.24. Samples from BANCA dataset sessions 1 (a-c), 5 (d-f) and 10 (g-i). The red crosses are detected landmark positions, and the green dots are final positions after structural correction.

are detected as mouth corners instead. In all these cases, the structural correction is able to interpolate the wrong landmarks. However in (e) and (i) the eye region is heavily blurred and masked by the eyeglasses, and the structural correction is unable to interpolate more than half the landmarks. In (i), as the bounding box is larger than the actual face region, the background interferes with localization. Overall, the 2D method performs well in this dataset.

5.8.4. Fine Localization

The fine localization results obtained with the gradient-based method are shown in Fig. 5.25 for FRGC ver.1 and in Fig. 5.26 for FRGC ver.2, Fall 2003 datasets. The

x-axis is the acceptance threshold of a localization, given as a percentage of the inter-eye distance. The mouth corners are the poorest localized landmarks in the coarse model, and they benefit the most from the fine search, although even then they don't quite reach a very high level of accuracy. The reason for this is the lack of adequate learning for the expression variations for ver.2, and the presence of facial hair for both versions.

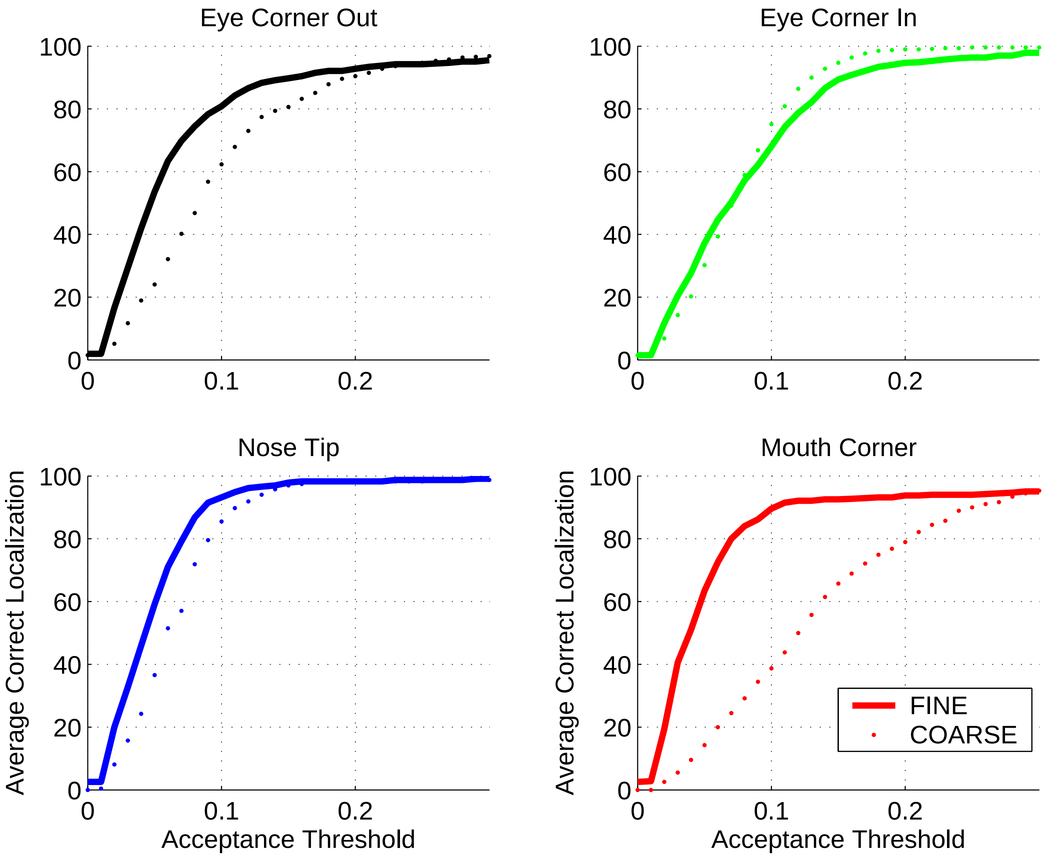


Figure 5.25. Kernel based fine landmarking on FRGC Spring 2003 dataset.

5.9. Space and Time Complexity Analysis

We present an analysis of the system in terms of space required to store the parameters, the offline training time, and the online landmarking time for a test sample. Some of the parameters are fixed in the implemented system, nevertheless we present those as variables in the complexity equations and note their values separately.

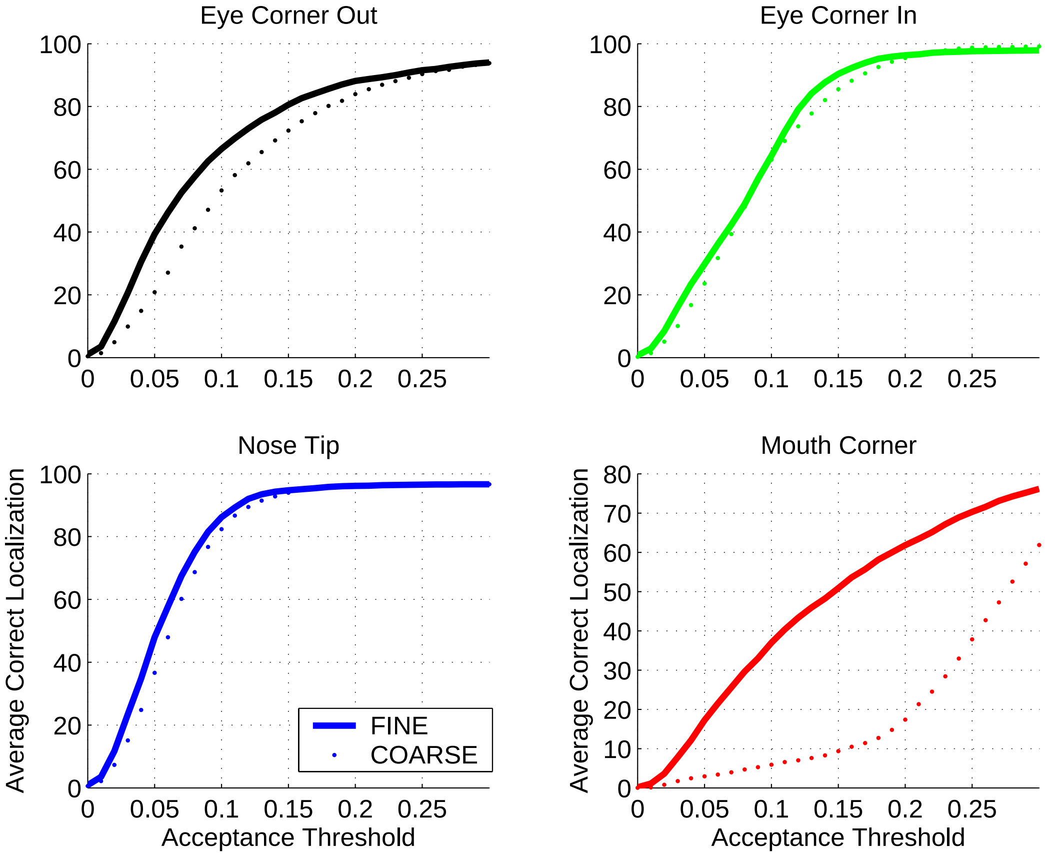


Figure 5.26. Kernel based fine landmarking on FRGC Fall 2003 dataset.

5.9.1. Space Complexity

The space complexity is made up of the parameters stored for the Gaussian distributions for each landmark, and the parameters stored for the structural analysis subsystem. If we denote the number of landmarks with L, and the number of features for each landmark with d, then the number of parameters for the Gaussian distributions with full covariance will be

$$S_G = O(L \times d^2). \tag{5.33}$$

If we use a IMoFA-L model with c components and p factors per component on the average, the parameters for the Gaussians can be expressed as

$$S_G = O(L \times cdp). \tag{5.34}$$

The IMoFA-L model is not correctly described by the average numbers, as the number of components and factors greatly differ depending on the shape of the distribution. In our simulations we used 7×7 feature windows, thus d was selected to be 49. We have observed that the IMoFA-L model required less than half the number of parameters used for a single full Gaussian.

For the structural analysis subsystem, we learn the distribution parameters of L 2D Gaussians for each combination that is supported by the system. The mean and covariance of a 2D Gaussian is represented with five parameters. If we denote the set of combinations supported by the system with $\boldsymbol{M}$, the number of parameters is expressed as

$$S_S = 5 \times \sum_{i \in \boldsymbol{M}} \binom{L}{i} = 5 \times \sum_{i \in \boldsymbol{M}} \frac{L!}{i!(L-i)!}. \tag{5.35}$$

For our simulations, L is equal to seven, and the structural analysis is performed for three or four landmarks in the fixed set. Therefore we need only 350 parameters for this part. Similarly, the number of parameters is small for the fine level search. There are L landmarks; two kernels are stored per landmark, up to 169 parameters per kernel, plus a threshold value for each kernel.

5.9.2. Offline Training Time Complexity

The most elaborate part of the training is the manual landmarking of the samples. Once the landmarking is performed, one IMoFA-L model is learned per landmark.

Learning an IMoFA-L model is slower than computing the mean and covariance of a single Gaussian; for this application, 500–1000 EM iterations for the training of the complete model were common. The mixture models typically contained 3–5 components, with 10–15 factors.

The locations of the landmarks are used to learn 2D Gaussian models for a number of support sets. Although this procedure includes two LM-optimization steps per model, the initial conditions and the low dimensionality ensure that these are learned very fast.

The fine level requires the evaluation of 336 kernels for the training set. This part is slower than the training for the IMoFA-L model, especially if the training set is large.

5.9.3. Online Landmarking Time Complexity

For a given sample, the steps of the algorithm can be summarized as:

1. Apply closing and opening to the flag.
2. Use polynomial interpolation to determine missing x and y values.
3. Use a 9×9 median filter to smooth depth values and to remove the artifacts.
4. Downsample the corrected flag and the grayscale intensity image by a factor of eight.
5. Extract windows from each spot, calculate likelihood under the IMoFA-L model.
6. Find the highest saliency spots, and form the candidate landmark set.
7. Start testing permutations of landmarks for support sets. Once an appropriate support set is found, label the rest of the landmarks as inliers and outliers.
8. Conduct a local search on the saliency map for the re-estimated outliers, select the best locations.
9. Compute necessary depth gradients around the coarsely located landmarks.
10. Apply the pre-stored threshold, and convolve with stored kernels.
11. Take the location with the highest value for each landmark.

Steps 1, 3, and 4 depend on the size of the input image, and are $O(n)$, if we denote the image size with a single parameter n. The polynomial interpolation in Step 2 is performed on a very small subset of the image, and thus it is negligible. The calculation of the likelihood is $O(cd^2)$ as the inverses of the covariance matrices are calculated in the training phase. In fact it is only necessary to store the inverses. Steps 5 and 7 are performed on the downsampled images, and are further constrained by the depth points. Although this reduces the computation time by a factor of about 15, it still scales linearly with n. The structural analysis part is performed with 2D Gaussian distributions, and on a handful of landmark locations. The extra time complexity required for this part is also negligible. The fine level search is performed on a fixed-sized neighbourhood, usually much smaller than n.

5.10. Conclusions

We have presented a biologically motivated method for automatically finding fiducial points on face images. The 2D scheme we have proposed relies on Gabor wavelet filters, and performs well under controlled illumination conditions. The 3D system based on range images has performed close to the 2D system in our experiments with FRGC ver.1, which contains illumination controlled 2D images. In the more challenging experiment with FRGC ver.2, 3D has performed remarkably good at nose tip and eye corners; but has failed at mouth corners, while the 2D system and 3D-assisted 2D system have very low detection rate. Our additional experiments where FRGC ver.2 supplied both the training and test samples demonstrated that the IMoFA-L model is capable of learning under the difficult illumination conditions of ver.2. However, the model is not able to generalize across different illumination conditions. This was not an unexpected result, but it is useful to see the extent of failure in this case.

Our simulations show that the simple albedo correction scheme improves 2D on some points, but the illumination effects still deteriorate recognition. More elaborate albedo correction schemes use synthetic images to find suitable bases and iteratively estimate the illumination coefficients, at much greater computational cost than the method we have employed [293].

The local features of the faces provide reliable cues to identify facial landmarks independently. This is particularly useful when some of the landmarks are not available for detection. There may be acquisition noise that we frequently see in the laser-scanned eye regions, the subject may have a scar or deformity that renders some of the landmarks unrecognizable, there may be partial occlusions by facial hair. In this case, an optimization approach that attempts to locate all landmarks simultaneously may not converge to the correct solution. We proposed an alternative approach that treats each landmark individually, and uses the structural relations between landmarks separately. Our structural correction scheme is shown to be superior to a recent competing technique.

Employing mixtures of factor analyzers allows us to strike a balance between temporal and spatial model complexity and accuracy. Although a full-covariance Gaussian mixture model has more representational power, it requires much more training samples than the model presently employed. Our model is able to represent the data with a smaller number of parameters. We have also shown that our mixture model is more successful than a recent and popular approach to Gaussian mixture models.

Once the landmarks are located in the coarse scale, a fine-resolution search can be employed to refine these locations. The methods employed for the coarse scale are available in fine scale as well. However, larger windows need to be sampled in order to do justice to the local statistical information. In [228] a discriminatory approach that uses 2D DCT coefficients was successfully used for large scale refinement, but further experimentation is necessary to assess its illumination dependence. Taking into consideration the increased computational cost of searching in the fine scale, we have proposed a very simple and straightforward approach for the fine level localization.

Automatic landmark localization is usually an early step in an application with other aims. We will now turn to the problem of 3D registration, and look at the effect of landmarking on subsequent registration.

6. DENSE 3D REGISTRATION AND CLASSIFICATION

In the vast Library there are no two identical books.
Jorge Luis Borges

The speed and accuracy of a 3D face recognition system depends on fast and correct registration for aligning the facial surfaces, thus making a fair comparison possible. The best results obtained so far use a one-to-all registration approach, which means each new facial surface is registered to all faces in the gallery, at a great computational cost [175]. In Chapter 3, we have mentioned that this computational cost can be greatly reduced by using an average face model (AFM), which automatically establishes correspondence to the pre-registered gallery faces. In this chapter, we focus on registration approaches that use AFMs.

We first propose a novel AFM generation method that aims at facilitating the subsequent classification (Section 6.2). As a baseline method, the point set difference (PSD) measure is adopted for nearest neighbour classification after registration. We contrast this measure with the Eigenface approach, as applied to depth values (Section 6.6). We show that the classification accuracy can be greatly improved by sampling the depth values from a regular grid.

We evaluate the quality of the AFM under rigid (iterative closest point, ICP) and non-rigid registration (thin-plate spline, TPS) methods. Details for these algorithms are provided in Section 6.1. The coarse registration in ICP and the non-rigid TPS-based registration both require a couple of fiducial points for guidance. We evaluate the effect of errors in landmark detection by using 3D ground-truth versus automatically located landmarks of Chapter 5. This permits us to analyze the algorithms under realistic assumptions, as automatic landmarking errors are not uniformly distributed. We evaluate several approaches for the coarse initialization of ICP, and show that using simple heuristics is not enough to ensure high accuracy (Section 6.7.1).

In accordance with the face recognition model proposed in Section 3.6, we assess the use of different subsystems for different face categories. What constitutes a facial category is an open issue; we propose and contrast an approach based on cognitive justifications (Section 6.3) with one that is based on clustering on the shape space (Section 6.4). We then explore whether an approach that uses one AFM per face category can trade-off computation time with accuracy or not. As in Chapter 5, we report our results on the FRGC face database.

6.1. Basic Algorithms

In this section, we provide descriptions of three algorithms that are mentioned frequently throughout this chapter. We start with Procrustes Analysis, which is used to find a consensus shape for a collection of shapes. Then we describe ICP and TPS, which are used for dense registration, in which two shapes are put into point-to-point correspondence.

6.1.1. Procrustes Analysis

Procrustes analysis is a statistical tool for the analysis of geometrical shapes [294, 295]. A shape (or equivalently a *figure*) P in $\mathbb{R}^p$ is represented by l landmarks. Two figures $P : l \times p$ and $P' : l \times p$ are said to have the same shape, if they are related by a special similarity transformation:

$$P' = \alpha P\Gamma + \mathbf{1}_l\gamma^T, \tag{6.1}$$

where the parameters of the similarity transformation are a rotation matrix $\Gamma : p \times p, |\Gamma| = 1$, a translation vector $\gamma : p \times 1$, and a positive scaling factor α. By using the generalized Procrustes analysis, it is possible to derive a *consensus shape* for a collection of figures [295]. This consensus shape is then used in registering new shapes into alignment with the collection by an affine transformation.

Here we give Gower's step-by-step description of the generalized Procrustes algorithm for a collection of N shapes [295], extending it with Rohlf and Slice's suggestions [296]:

1. Center all shapes P_i:

$$M = \frac{1}{N}\sum_{i=1}^{N} P_i \tag{6.2}$$

$$P_i = P_i - M \tag{6.3}$$

2. Bring the shapes to a common scale. In Gower's method [295], the mean of the squared interlandmark distances is set to unity:

$$P_i = \frac{lP_i}{\sum_{k=1}^{l} ||P_{i,k}||} \tag{6.4}$$

In [296], scaling is performed by setting the median of the squared interlandmark distance to unity. Representing the k^{th} landmark of shape P_i with $P_{i,k}$:

$$D_i = \{d_{j,k} = ||P_{i,j} - P_{i,k}|| \ \ |j,k = 1\ldots l\} \tag{6.5}$$

$$P_i = \frac{P_i}{m(D_i)} \tag{6.6}$$

where $m(x)$ is the median operator.

3. Set the consensus shape Y equal to the first shape P_1, as an initialization.
4. For $i = 2, 3, \ldots, N$, rotate P_i to fit Y. In Gower's method, Y is re-evaluated after each update of P_i as

$$Y = \frac{1}{i}\sum_{j=1}^{i} P_j \tag{6.7}$$

In Rohlf and Slice's method, the Y is updated once, after each P_i is rotated. The rotation matrix H in two dimensions is expressed as:

$$H = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{6.8}$$

The best rotation can be found with an Eckart-Young singular-value decomposition:

$$H = VSU^T \tag{6.9}$$

where U and V are such that

$$P_i^T Y = U\Sigma V^T \tag{6.10}$$

and Σ is diagonal. In Eq. 6.9, the matrix S is used instead of Σ, where S is a diagonal matrix with $s_{ii} = \pm 1$. The signs are taken to be equal to the signs of the corresponding diagonal elements of Σ. This takes into account reflections that would lead to a better fit. Using S instead of Σ constrains H to be a rotation, rather than a shear. As a practical implementation issue, we suggest monitoring the sign of the determinant of H for incorrect reflections.

5. Repeat updating the P_i and Y, while monitoring the residual sum-of-squares:

$$S_r = N(1 - tr(Y_t Y_t^T - Y_{t-1} Y_{t-1}^T)) \tag{6.11}$$

where Y_t is the consensus at iteration t, and Y_{t-1} is the consensus at iteration $t-1$. When S_r is below a threshold (e.g. 0.0001, as suggested by Gower) stop the iterations, and output the consensus shape.

In [297] a *refined Procrustes distance* was proposed by removing the correlation of landmarks, mapping the shapes onto a unit sphere. This projection is used to analyze the shapes, and to reduce the gallery size by weeding out unlikely shapes. However, the geometrical distribution of facial landmarks resemble each other, and a great number of accurate landmarks (29 were used) is needed to use it for classification.

6.1.2. Iterative Closest Point

In this section we describe the ICP method to register a point set P with a model shape Y [1]. Although the computational cost of ICP is very high, its straightforward implementation and accuracy makes it the most frequently used registration tool in 3D face recognition research.

Define the *unit quaternion* as a four vector $\vec{q_H} = [q_0 q_1 q_2 q_3]^T$, with $q_0 \geq 0$, and $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$. The 3×3 rotation matrix H generated by this quaternion is:

$$H = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 - q_0q_3) & 2(q_1q_3 + q_0q_2) \\ 2(q_1q_2 + q_0q_3) & q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_2q_3 - q_0q_1) \\ 2(q_1q_3 - q_0q_2) & 2(q_2q_3 + q_0q_1) & q_0^2 + q_1^2 - q_2^2 - q_3^2 \end{bmatrix} \tag{6.12}$$

Let $\vec{q_T} = [q_4 q_5 q_6]^T$ be a translation vector. Together with q_H, they make up the complete registration state vector $\vec{q} = [\vec{q_H} | \vec{q_T}]^T$. Let $P = \{\vec{p_i}\}$ be a data point set to be aligned with the model point set $Y = \{\vec{y_i}\}$. The two models will have the same number of points, and ICP will put the points with the same indices into one-to-one correspondence. Denoting the number of points in each model with N, the objective function minimized by the ICP procedure is:

$$f(\vec{q}) = \frac{1}{N} \sum_{i=1}^{N} ||\vec{y_i} - H(\vec{q_h})\vec{p_i} - \vec{q_T}||^2 \tag{6.13}$$

Denoting the centre of mass of the point set P with μ_p, and that of the model set with μ_y, the cross-covariance matrix Σ_{py} is given by:

$$\begin{aligned} \Sigma_{py} &= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} [(\vec{p_i} - \vec{\mu_p})(\vec{y_i} - \vec{\mu_y})^T] \\ &= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} [\vec{p_i}\vec{y_i}^T] - \vec{\mu_p}\vec{\mu_y}^T] \end{aligned} \tag{6.14}$$

The cyclic components of the matrix $A_{ij} = (\Sigma_{py} - \Sigma_{py}^T)_{ij}$ are used to form a column vector $\Delta = [A_{23} \;\; A_{31} \;\; A_{12}]^T$, which in turn is used to form a symmetric 4×4 matrix $Q(\Sigma_{py})$:

$$Q(\Sigma_{py}) = \begin{bmatrix} tr(\Sigma_{py}) & \Delta^T \\ \Delta & \Sigma_{py} + \Sigma_{py}^T - tr(\Sigma_{py})I \end{bmatrix} \tag{6.15}$$

where I is the 3×3 identity matrix. The optimum rotation is given by the unit eigenvector $\vec{q_H} = [q_0 q_1 q_2 q_3]^T$ corresponding to the maximum eigenvalue of the matrix $Q(\Sigma_{py})$. The optimum translation vector is given by:

$$\vec{q_T} = \vec{\mu_y} - H(\vec{q_H})\vec{\mu_p} \tag{6.16}$$

We use $\vec{q}(P)$ to denote the point set P after the application of the transformation represented by $\vec{q}$. The ICP algorithm computes and applies these transformations iteratively. A step-by-step description of the algorithm follows:

1. Initialize ICP by setting $P_0 = P$, $\vec{q_0} = [1, 0, 0, 0, 0, 0, 0]^T$ and $k = 0$. The registration is defined relative to P_0, which requires a coarse registration. Steps 2-5 are applied iteratively, until convergence is achieved within a tolerance τ.
2. Compute the closest points: $Y_k = \mathcal{C}(P_k, Y)$. The computational cost of this step is $O(N_p N_y)$ at the worst case, where N_p is the number of points on the registered point cloud, and N_y is the number of points on the model shape.
3. Compute the registration: $(\vec{q_k}, d_k) = \mathcal{Q}(P_0, Y_k)$. The computational cost is $O(N_p)$.
4. Apply the registration: $P_{k+1} = \vec{q_k}(P_0)$. The computational cost is $O(N_p)$.
5. Terminate the iteration if the change in the mean square error is below pre-set threshold τ. A heuristic value for τ is a multiple of $\sqrt{tr(\Sigma_y)}$, where Σ_x is the covariance matrix of the model shape, and the square root of its trace is a rough indicator of model shape size.

In 3D face recognition practice, the number of points N_Y of the model shape and the number of points N_P of the registered surface usually do not agree. The gallery faces (or the average face model) are cropped beforehand, and have fewer points. In our simulations, the gallery point clouds contained about 30.000 points, whereas the test scans contained 80.000-130.000 points. For this reason, the test face acts as the model shape, and the cropped gallery face is aligned to it. The points of the test scan that are put into one-to-one correspondence with the model are retained, and the rest are discarded. If the registration is correct, this procedure automatically gives a good cropping, possibly including hair and clutter removal. In the rare cases, where a streak of hair extends over the face centre, the registration and the subsequent classification will be inaccurate.

The greatest computational burden in ICP is the phase of computing the closest points. In a recent paper, Yan and Bowyer propose to build a voxel discretization of the gallery in an offline manner to reduce the computation time during online comparison [298]. Their results also suggest that point-to-surface distance is more accurate than point-to-point distance in guiding comparisons, as long as its additional computational burden can be shifted to offline computation. Ayyagari *et al.* have proposed an improvement that introduces a differentiable cost function and the fast Gauss transform to overcome the initialization problem in ICP. They observe that random sampling may create problems due to uneven point distributions, whereas uniform sampling may lose some representation power, which we confirm through our simulations in Section 6.7.4.

6.1.3. Thin-plate Splines

The thin-plate spline (TPS) model expresses the bending energy of a thin metal plate fixed at certain points [176]. At the heart of the model is a special surface function:

$$z(x, y) = -U(r) = -r^2 \log(r^2) \tag{6.17}$$

with $r = \sqrt{x^2 + y^2}$ equal to the Euclidean distance of point (x, y) to the origin. This function defines the surface demonstrated in Fig. 6.1.

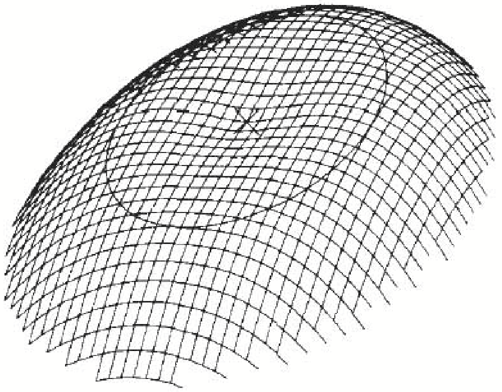


Figure 6.1. The special surface function $U(x, y)$ used in TPS model (from Bookstein, 1989).

For a set of anchor points $P_i = (x_i, y_i), i = 1 \dots n$, the thin-plate spline interpolation is a vector-valued function $f(x, y) = [f_x(x, y), f_y(x, y)]$ that maps the anchor points to their specified homologues $P_i' = (x_i', y_i'), i = 1 \dots n$, and specifies a surface which has the least possible bending, as measured by an *integral bending norm.* We will give a mathematical specification of the model here, and refer the reader to Bookstein's paper for further details [176].

Define $r_{ij} = |P_i - P_j|$ to be the distance between the points i and j. Also define the following matrices:

$$K = \begin{bmatrix} 0 & U(r_{12}) & \dots & U(r_{1n}) \\ U(r_{21}) & 0 & \dots & U(r_{2n}) \\ \dots & \dots & \dots & \dots \\ U(r_{n1}) & U(r_{n2}) & \dots & 0 \end{bmatrix} \tag{6.18}$$

$$P = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \dots & \dots & \dots \\ 1 & x_n & y_n \end{bmatrix} \tag{6.19}$$

and

$$L = \left[\begin{array}{c|c} K & P \\ \hline P^T & O \end{array}\right] \tag{6.20}$$

where O is a 3×3 matrix of zeros. Let V be a matrix made up of the homologues of the anchor points:

$$V = \begin{bmatrix} x'_1 & x'_2 & \dots & x'_n \\ y'_1 & y'_2 & \dots & y'_n \end{bmatrix} \tag{6.21}$$

Define w_i and the coefficients a_1, a_x, and a_y as:

$$L^{-1}(V|\mathbf{0},\mathbf{0},\mathbf{0}) = (w_1, w_2, \dots, w_n, a_1, a_x, a_y)^T \tag{6.22}$$

The function $f(x, y)$ is defined as:

$$f(x,y) = a_1 + a_x x + a_y y + \sum_{i=1}^{n} w_i U(|P_i - (x,y)|) \tag{6.23}$$

$f(x, y)$ minimizes the nonnegative integral bending norm I_f over all such interpolants:

$$I_f = \iint_{\mathbb{R}^2} \left(\left(\frac{\partial^2 f}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}\right)^2 \right) \mathrm{d}x\, \mathrm{d}y \tag{6.24}$$

The thin-plate spline function $f(x, y$ is invariant under rotations and translations. It maps the landmarks P_i to their homologues P'_i, and defines a smooth interpolation for the rest of the points on the surface. P_i and P'_i taken together exactly specify the function $f(x, y)$, and are therefore crucial to the accuracy of the deformation. In the TPS-based registration method we use, P_i are the landmarks defined on the average face model, and P'_i are the corresponding landmarks of the test scan.

6.2. Average Face Model Construction

In *dense registration*, the points on the test surface and the points on the gallery surface are put into one-to-one correspondence. ICP achieves this by iteratively locating the closest point on the test surface for each point on the gallery surface, and rigidly moving the aligned surface to minimize the total point-to-point distances [1]. Upon convergence, the distances between the points can be summed up to find a total distance to the gallery face. This is the *point set difference* (PSD) measure, which is easily obtained in ICP. Usually the gallery face is cropped and cleansed from all clutter, and the number of correspondences equals the number of points on the gallery surface. For the TPS-based non-rigid registration method used in this chapter, seven landmarks are identified on the test face, and used to drive the registration. The TPS method is much faster than the ICP alignment.

We propose to use an average face model (AFM) in dense registration. The idea in AFM-based registration is to register all gallery surfaces to a single AFM, which acts as an index file. The test surface is registered once and for all with the AFM, which associates one point on the test surface for each point of the AFM, and consequently, for each point of any given gallery surface.

In [161], a method for generating the AFM was described: The method involves a TPS-based registration of the training faces to a consensus shape. Then, one of the faces is selected as the AFM candidate, and its vertices are trimmed if their distance to an other training image exceeds a threshold. This procedure creates a very smooth facial surface.

In this thesis, we generate a more pronounced AFM by using a set of landmarked training faces, with the following procedure:

- Using Procrustes analysis, a consensus distribution of landmarks (`consensus shape`) is found on the training set.
- The landmarks of the consensus shape are rectified to present a fully frontal face.

This heuristic is used to facilitate the use of the transformed range image in later stages. Rectification is achieved by rotating the face so that the eye and mouth planes are parallel to the x-axis and the z-axis as much as possible.

- TPS deformation is computed for the training faces, which warps the landmarks of each face to the consensus shape perfectly, and interpolates the rest of the points.
- The depth values of the interpolated face are re-sampled from a regular x-y grid. This ensures that all added faces have points with corresponding x and y values, and the depth values are given for matching points. For the simple range image representation, this extra offline computation leads to much faster online model comparison.
- Faces are cropped before they are added to the average face model. One face is used to define a cropping mask used for the rest of the faces. First we calculate the maximum distance from the nose tip to any landmark in the consensus shape. We add a ten per cent margin to this distance to take into account landmark variations, and retain all points closer than this value to the nose.
- After all the training faces are added, depth values are simply averaged.

Samples of AFMs generated with this method can be seen in Figure 6.2. Any irregularity in the surfaces is due to poor pre-processing of the depth data. The FRGC database we use in this study was collected with a laser sensor that typically generates holes (especially at the eyes and the mouth) or other artifacts [135]. The pre-processing for these files (9×9 median filtering, followed by 9×9 mean filtering, followed by polynomial interpolation of missing points at each dimension) sometimes falls short of repairing larger errors. More elaborate pre-processing methods with better interpolation that use facial symmetry or an AFM to fill the holes are conceivable, but these will only work after registration.

6.3. Cognitive Justification for Multiple AFMs

As we have noted in Section 2.5, the *other race effect* occurs when people try and fail at distinguishing faces from another race group. However, members of a race can recall unfamiliar faces of another race more easily, if that other race is represented better in the environment of the subject [299, 300]. Following the expertise hypothesis in its explanation of the other race effect, Tong *et al.* conjecture that enhanced within-category discrimination is a property of visual expertise, and support their view by a neural network model of the fusiform face area that agrees with psychophysical data [22]. In the light of their experiments, it seems reasonable that during the acquisition of face expertise, the transformations learned by the visual cortex serve to magnify the differences between individual faces, as indicated by the statistical distribution of the encountered facial features. By this reasoning, the other race effect suggests that different races exhibit different statistical distributions of distinguishing facial characteristics.

Our aim is not to detect the *race* of a person; therefore, we will use the term *morphology* to denote racially similar facial surface characteristics. Based on cognitive cues, we predict better recognition rates if the faces are clustered into morphological or gender groups that exhibit greater intra-group similarity and the discriminative features are learned within each group separately. This is not trivially true for all pattern recognition applications, as the grouping reduces the number of training samples, and consequently runs the risk of impairing learning conditions.

6.4. Shape Space Clustering

If the hypothesis of meta-classification is correct, we expect morphology and gender to be discriminating dimensions of the face space. However, we do not wish to categorize faces into races explicitly, as this approach has ethical consequences. Can the gender and race determination during the training (and possibly, in the testing) stage be evaded? For simulation purposes, we have roughly assigned facial images into African, Asian and Caucasian morphological face classes. The other-race effect suggests that racial-morphology based clusters exist in the face space, and an unsupervised clustering method can recover those clusters, among other meaningful structure. Thus, it is not necessary to recover the race and gender of a person; the clustering will hopefully provide us with a useful set of average faces to serve in meta-classification with increased discrimination within clusters.

We propose to take a straightforward race- and gender-blind clustering approach with the k-means algorithm. The clustering is performed on the shape space, as represented by the aligned coordinates of seven facial landmarks. We specify the number of clusters for the shapes, and initialize the cluster consensus shapes by random selection from the training samples. At each iteration, we align the training samples to the consensus shapes of the clusters via Procrustes analysis, and assign each sample to the cluster with minimum average distance to the cluster consensus. We then re-estimate cluster consensus shapes from the samples assigned to the cluster, and iterate until the total distance stabilizes.

The clustering gives us a number of cluster consensus shapes, and assigns each training face to one of these clusters. We apply our AFM generation algorithm to these reduced training sets separately, and obtain one AFM for each cluster. These models can be seen in Section 6.7.3.

6.5. Registration Methods

We test two different registration methods. In the first method (termed *TPS-based* in the experiments section), the test face is aligned to the average face with the TPS method, and the points in correspondence with the AFM are cropped [161]. This method deforms the test face to fit the AFM, and the amount of deformation is proportional to the number (and spread) of the landmarks. At the limit of using all facial points as landmarks, the face deforms into the AFM, losing the discriminative information completely. However, with a few landmarks, corresponding facial structures are aligned.

In the second method, we use the iterative closest point method to align the test face with the AFM. ICP is a rigid registration method, hence the test face is not deformed at all. TPS-based methods are completely guided by the landmarks, whereas ICP needs a coarse initialization. Previous work on ICP show that a good initialization is necessary for fast convergence and an accurate end-result. We compare several approaches for the coarse registration.

In our first approach, the point with the greatest depth value is assumed to be the tip of the nose [237, 238], and a translation is found to align it to the nose tip of the AFM. This is the fastest and simplest heuristic we can use, and we expect it to perform well in near-frontal faces. In the second approach, we use the manually determined nose tip (i.e. ground truth) in the coarse alignment. In the third approach, we use Procrustes analysis to bring seven manually determined landmark points (inner and outer eye corners, nose tip, and the mouth corners) into alignment with the average face model. Finally, we use Procrustes analysis to align automatically determined landmarks with the average face model. The automatic landmarking errors are not random, and cannot be simulated by injecting noise to the manually determined landmarks, except by modeling the specific landmarking procedure.

Intuitively, ICP will benefit from using category-specific AFMs, as the rigid registration is not able to cope with shape differences very well. A more similar average face will ensure that the dense correspondence will be established between points that have better structural correspondence. The TPS-based method will also benefit from category-specific AFMs, albeit for another reason. A more similar average face means that the test surface will be less deformed, and discriminatory information will not be lost.

6.6. The Eigenface Approach

We have contrasted the PSD method for classification with the Eigenface method [91]. The presentation of the material in this section follows closely the original Eigenface paper of Turk and Pentland. In this approach, the face images are assumed to span a so-called *face space*, a manifold of the high-dimensional space of images, populated by the face images. Each face image can be imagined as a d-dimensional vector, or a point in the d-dimensional space. Quite similar to the main idea behind factor analysis, we hypothesize the existence of a much lower-dimensional space that describes the facial variations.

Since we have a training set of faces, our learning system will only be able to describe variations that are manifested in this set. If the variations of faces in the face space are adequately described by our training set, the resulting system will be accurate and frugal, as a projection to this subspace will allow us to represent the facial images with much fewer parameters. Furthermore, if the vectors that span this subspace are ordered in terms of variation along their direction, we will have a natural ordering of projected features that reflects the relative importance of each feature. In general, features that show high variability allow a better separation of samples. There can be pathological cases where this is not true, but it is still a good and widely used heuristic, motivated by biological considerations detailed in Section 2.7.

For a training set $\chi = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ of p samples, let $\boldsymbol{\mu}$ denote the average of the set. In the original Eigenface paper, $\boldsymbol{x}_i$ are facial images. In this work, they are employed generically, as we apply this method to patches as well as whole faces. The variation from the average is denoted by $\Phi_j = \boldsymbol{x}_j - \boldsymbol{\mu}$. We will use principal component analysis (PCA) to determine a number of orthonormal vectors that describe the distribution optimally. The optimality is mentioned in the mean squared error sense: The first n eigenvectors, sorted in decreasing order by their corresponding eigenvalues, describe a projection $f(\boldsymbol{x})$: $\mathbb{R}^d \rightarrow \mathbb{R}^n$ that can be used for compression. The inverse projection recovers the original dimensionality, but not the original data. However, $f(\boldsymbol{x})$ is the projection that gives the minimum mean squared error between the original image and the recovered image for the training set χ, among all projections to n dimensions.

The vectors $\boldsymbol{u}_i$ that make up the projection are the orthonormal vectors that maximize the value of corresponding scalars $\boldsymbol{v}_i$, computed as:

$$\boldsymbol{v}_i = \frac{1}{p} \sum_{k=1}^{p} (\boldsymbol{u}_i^T \Phi_k)^2 \tag{6.25}$$

with the additional condition that each such vector should satisfy the orthogonality condition:

$$\boldsymbol{u}_i^T \boldsymbol{u}_k = \begin{cases} 1, & \text{i=k} \\ 0, & o/w \end{cases} \tag{6.26}$$

The vectors $\boldsymbol{u}_i$ and the scalars $\boldsymbol{v}_i$ are the eigenvectors and the eigenvalues of the covariance matrix Σ, respectively:

$$\begin{aligned} \Sigma &= \frac{1}{p}\sum_{k=1}^{p}(\Phi_k \Phi_k^T) \\ &= AA^T \end{aligned} \tag{6.27}$$

where $A = [\Phi_1 \Phi_2 \ldots \Phi_p]$ is a $d \times p$ matrix. The covariance matrix Σ is $d \times d$ in this case, and it is not tractable to compute it directly. Usually, the number of training samples p is much less than d, and the sample covariance matrix will be singular, even if we can compute it. For instance, in the case of densely-registered and cropped 3D faces that we work with, d is about 30.000, whereas p is around 50. However, we can compute the first $p-1$ eigenvectors (-1, because we have subtracted the mean from each vector), which are the only ones that can have positive eigenvectors. This is done by computing the eigenvectors of $\boldsymbol{u}_k$ of the matrix $A^T A$ first, and the eigenvectors of AA^T then correspond to $A\boldsymbol{u}_k$.

A new data sample $\boldsymbol{x}$ is projected to the subspace described by $\boldsymbol{u}_i$ with the following computation:

$$\omega_i = \boldsymbol{u}_i^T(\boldsymbol{x} - \boldsymbol{\mu}) \tag{6.28}$$

The sample is projected onto a fixed number of eigenvectors, which represents a trade-off between computational cost and accuracy. If the whole set of eigenvectors are used in the projection, it is possible to recover the sample perfectly. Since the eigenvectors are ordered according to their importance (as measured by variability) it makes sense to use only the first few eigenvectors in the projection. The most frequently employed heuristic in determining the subspace dimensionality is to look at the explained variance, and take enough eigenvectors to explain a fixed percentage of the variance.

Once the subspace projection is computed for new samples, the comparison with the stored gallery samples can simply be performed by selecting the sample that minimizes the Euclidean distance in the low-dimensional subspace. Since the eigenvectors are orthonormal, using the Euclidean distance in the subspace is equivalent to using Mahalanobis distance in the high-dimensional space. One can also determine a threshold for this distance, beyond which the samples are not accepted as face images.

6.7. Experimental Results

We use a subset of the FRGC 2D-3D ver.1 face database in our registration experiments [135]. We use images from 195 subjects, with one training face in the gallery and 1-4 test faces (for a total of 659 test faces). We only work with 3D information; 2D is not used at all. We design a number of experiments to answer various questions. Each subsection deals with a particular question, and reports relevant simulation results. The overall system has many dimensions, ruling out a factorial experimentation protocol.

For each method, we run a recognition experiment and a verification experiment. For the recognition experiments, the rank-1 recognition rate (R1) is reported. In the verification experiments, each of the 659 test faces is used for one genuine and 194 false claims, thus the number of false claims is two orders of magnitude higher. The equal error rate (EER) is reported under these conditions.

6.7.1. Coarse registration

Table 6.1 shows the effect of coarse alignment methods on ICP-based registration. As we have suspected, the nose-tip heuristic performs worst. Automatic localization of seven landmarks and using Procrustes alignment works better than the nose-tip heuristic. For ICP, using the nose ground truth works well in this dataset, because the faces we deal with are mostly upright and frontal. Ideally three landmarks should be used to accomodate greater pose differences in 3D. Finally, the manual landmarks with the Procrustes analysis give us an upper-bound on the performance of the ICP-PSD method.

We have also contrasted our AFM construction method with the method of Irfanoğlu *et al.* [161] on ICP. Manual landmarks were used, and initialization was by Procrustes alignment. With their smoother AFM, a rank-1 recognition rate of 86.34 per cent was achieved, as opposed to our 92.11 per cent. Similarly, the EER was higher with their AFM by more than two per cent.

Table 6.1. Effect of coarse alignment on ICP

	Nose-tip heuristic	Automatic landmarks + Procrustes	Nose ground truth	Manual landmarks + Procrustes
R1	82.85	87.86	90.60	92.11
EER	14.25	8.12	6.60	6.20

6.7.2. Meta-classification

Does meta-classification and more specialized individual experts increase discrimination? We have tested this hypothesis by employing the average faces that are generated from groups of training faces. We have grouped the training samples for gender and morphology, and generated average face models (AFM) for each group. Figure 6.2 shows range images obtained for the average face, for male and female averages, for three morphological group averages (roughly Caucasian, Asian, and African), and for all combinations, respectively. The morphology does not correspond to a clear-cut distinction; the morphological group of a given face will be determined by its proximity to the average face of the group.

In Table 6.2, the authentication experiment results with or without specific average faces are shown for TPS-based registration. We have supplied both the generic-

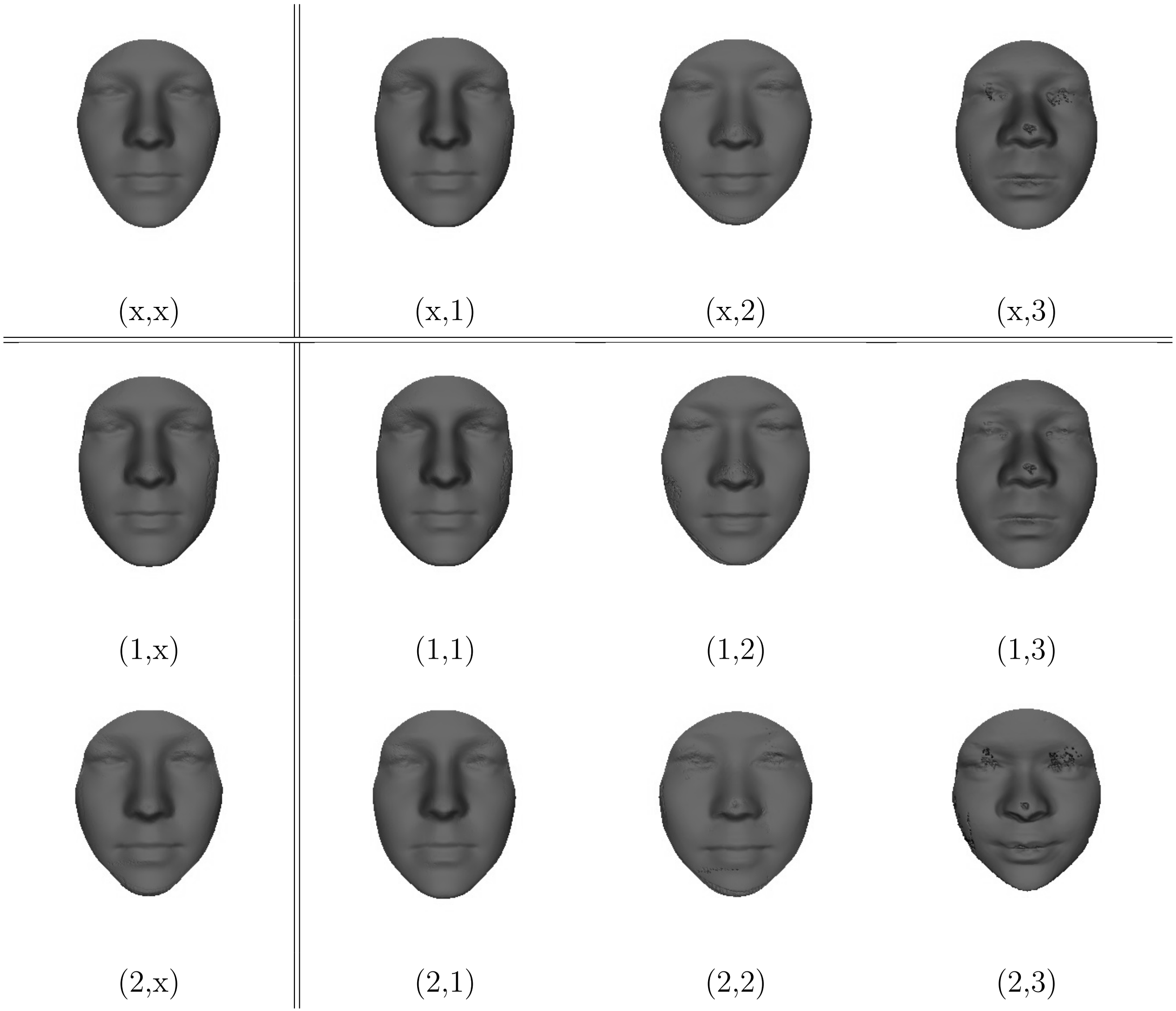


Figure 6.2. Average faces for different morphology and gender combinations. In (i, j) notation i is 1 (male) or 2 (female), and j is 1 (Caucasian), 2 (Asian) or 3 (African). Generic averages are denoted with x.

AFM based system and the specific-AFM based system with the categorical information. Then, any improvement in the specific system is strictly due to better registration. We have computed distances between the test face and the gallery faces with an L_1 distance metric, and trimmed the worst two per cent of corresponding points. The trimming is used to deal with outliers. The EER results reported under these conditions show that specific AFM usage is beneficial in this case. We stress that the main purpose of the PSD is to evaluate the effect of AFM usage; the EER is generally too high for these experiments because of deformations in the registration.

Table 6.2. Simulation of TPS registration with deformation, EER

	None	Gender	Morphology	Gender & Morphology
Generic AFM	20.10	16.79	18.50	13.87
Specific AFM	20.10	14.64	16.97	11.47

The point set distance after aligning the face to the female and male averages can be used in gender classification. This simple method works with 80 per cent accuracy.

6.7.3. Shape Space Clustering

To confirm our cognitively motivated subspace hypothesis, the shape space clustering should automatically create clusters with a dominant gender, or put samples from one race into a single group. This is more or less what happens, although the non-uniformity of the training set (many more Caucasian males than, say African females) introduces a bias. We have specified six clusters, as in the full morphology-gender combination case, and ran our algorithm on the training part of the FRGC ver.1 dataset. Figure 6.3 shows the cluster means and Figure 6.4 shows the distribution of morphology and gender in each group as pie charts. Out of six clusters, two are dominantly male and two are dominantly female. One of the mixed gender clusters has dominantly faces of Asian morphology. Almost all males labeled as African are put into a single cluster.

The number of training samples are evenly divided into clusters for this shape-space method. For the morphology-gender approach, the distribution was uneven, as we have too few samples from the some of the categories. However, this sort of handicap is expected with a limited dataset, and it is necessary to see how the algorithms cope with it.

Table 6.3 shows recognition rates for ICP and TPS based systems with manual or automatic landmarks. The first row shows the results obtained with a single generic AFM. The next three rows show results with gender-, morphology-, and gender +

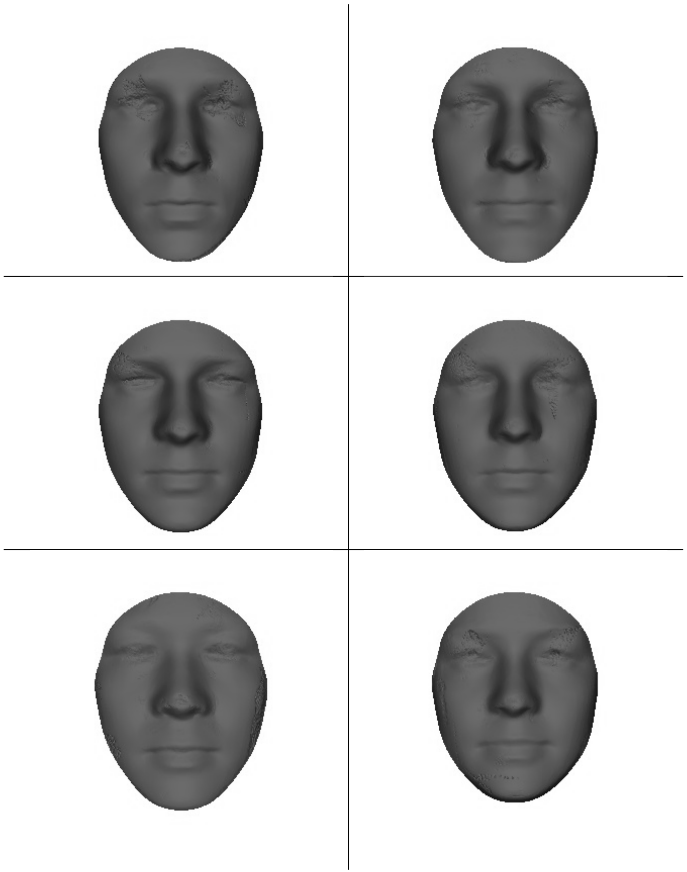


Figure 6.3. Shape space cluster means.

morphology-based specific AFMs, respectively. The results for the last row are obtained with six shape-space derived clusters. For this last case, the registration does not benefit from the injection of categorical information, and each test sample is compared with all the training samples. The best result is obtained with shape-space derived specific AFM and ICP. We comment more on these results in the next section.

The rigid registration method performs significantly better than the non-rigid registration method. TPS warping partly destroys discriminatory information. Increasing the number of landmarks will make this effect more pronounced, and in the limit where all points on the face are located and treated as landmarks the classification accuracy will drop to zero, as TPS will warp each test scan into the AFM.

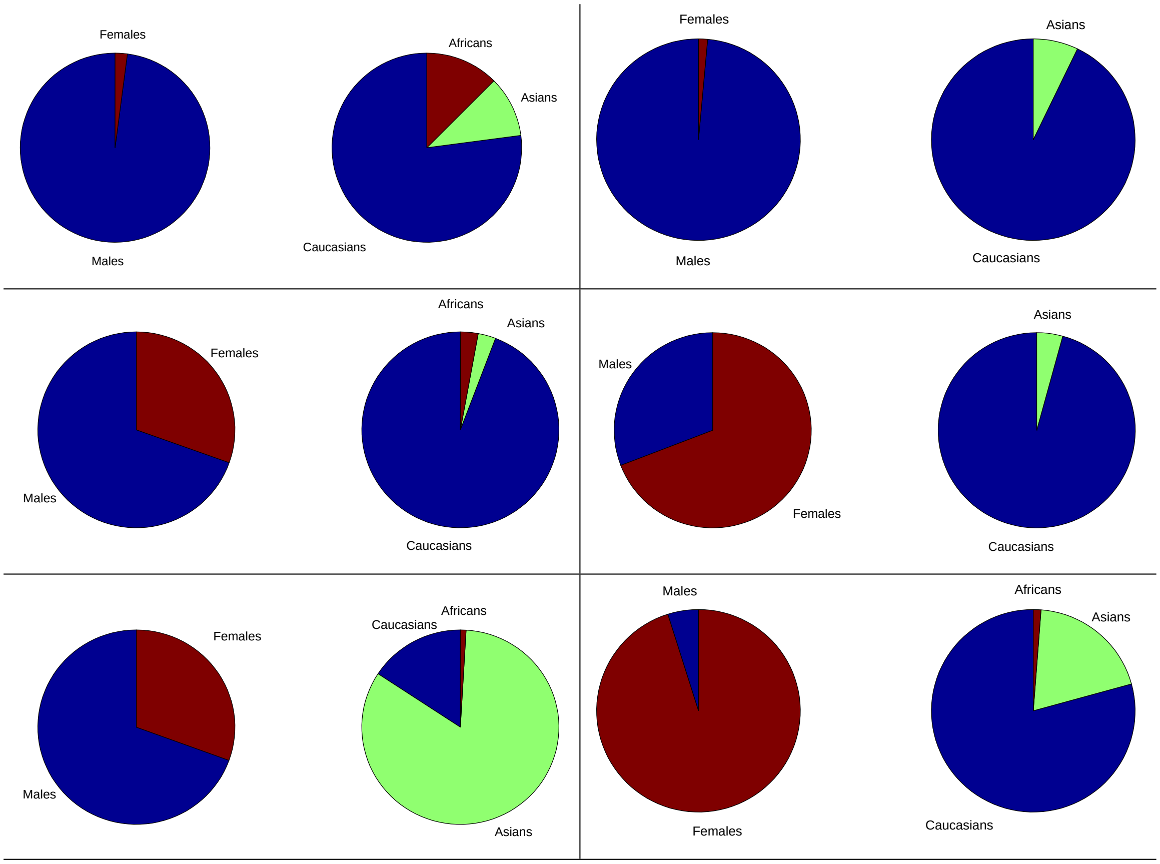


Figure 6.4. Shape space clustering distributions. For each cluster, the gender and morphology distributions are shown in separate pie charts. For six clusters, we have two dominantly male, and two dominantly female clusters, one dominantly Asian cluster, and almost all the males labeled as African are clustered into a single group.

6.7.4. Subspace Projection and Regular Re-sampling

In this section we apply the Eigenface method to 3D range images registered to different AFMs with ICP. The depth values of the range images are not sampled regularly from an $x - y$ grid, and so far we have computed 3D point distances, instead of a simpler 1D comparison based on the depth. Since we deal with aligned shapes in dense correspondence, applying a regular re-sampling will make it possible to discard two dimensions from the point cloud, making the subsequent comparison and subspace projection easier. We will also show that the regular re-sampling helps classification by making the distance measurement more accurate. We have used a simple triangle-based nearest-neighbour interpolation for this purpose [301].

Table 6.3. Comparison of specific AFMs, rank-1 recognition rates

	Manual lm. + ICP	Automatic lm. + ICP	Manual lm. + TPS	Automatic lm. + TPS
Generic	92.11	87.86	52.20	42.64
Gender	90.14	86.65	54.63	45.52
Morphology	89.98	86.80	53.87	44.92
Gender & morphology	91.05	86.49	56.90	47.95
Shape space derived	93.78	91.20	47.65	41.58

The dimensionality of the subspace is determined heuristically. We set the number of eigenvectors so that at least 95 per cent of the variance is accounted for. Additional experiments have shown that increasing the subspace dimensionality results in diminishing returns. The Eigenface method is contrasted with the PSD method, which, due to regular re-sampling, uses the sum of squared distances of depth values only.

The regular re-sampling lightens the computational burden of comparing the test sample with gallery images. Furthermore, the computed projection allows us to store much smaller gallery faces. For example in the experiments with the generic AFM, roughly 32.000-dimensional face vectors are represented with 50-dimensional vectors after the subspace projection. This means the comparison with gallery faces will be much faster. Table 6.4 shows that the accuracy loss due to subspace projection does not exceed one per cent, if there are sufficient training samples. For the gender & morphology combination, the training set is very limited, and consequently there are only 15 eigenvectors with non-zero eigenvalues in one of the groups (i.e. the results were obtained with $p = 15$). This number is too small to represent the facial variation, and the accuracy decrease is about three per cent. The morphology results were obtained with $p = 33$, and the gender results with $p = 49$. For the results in this section, we have grouped African faces and Caucasian faces into a single category, as we had too few samples from the African category.

Table 6.4. Subspace projection after ICP, rank-1 recognition rates

	Manual lm. PSD	Manual lm. Eigenface	Automatic lm. PSD	Automatic lm. Eigenface
Generic	98.18	98.03	98.03	97.88
Gender	96.81	96.21	95.30	95.90
Morphology	96.51	96.05	95.60	94.84
Gender & morphology	96.97	93.78	94.99	91.96
Shape space derived	98.18	97.72	98.03	96.81

The ICP results reported in Table 6.4 are much better than the results reported in Table 6.3. With the generic AFM and the automatic landmarks (i.e. the fully automatic system) the PSD method without re-sampling has a rank-1 recognition rate of 87.86 per cent, whereas after re-sampling, it has 98.03 per cent accuracy. The reason is graphically depicted in Fig. 6.5. The facial surface (shown symbolically as a dotted line in the figure) is irregularly sampled by the laser scanner (two points per facial surface, shown with black triangles). The ICP registration brings these surfaces into alignment by global rigid matching. Hence, the corresponding points may not be in close alignment locally, although the sum of all displacement vectors is at a local minimum. Regular re-sampling produces depth values at regular x and y intervals (shown with black square points). These points give a more realistic indication of the distance between the two surfaces, unless the absolute depth gradient is very high. In the latter case, small displacements in the $x - y$ plane will result in big changes in depth, making an irregular, point-to-point 3D comparison the logical choice. However, the facial surface as represented by a range image has few points with sharp depth changes (i.e. the nose ridge, mouth and eye corners, and the face boundary). Our cropping procedure eliminates the face boundary, and greatly reduces the number of these points. Consequently, the regular re-sampling is indispensable for AFM-based registration.

When we inspect the samples that are classified correctly after re-sampling, but not before, we see that the error due point-irregularities is large enough to disturb

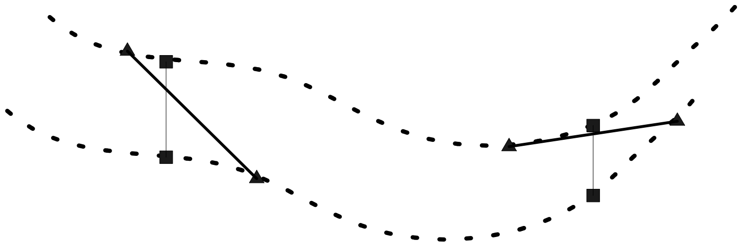
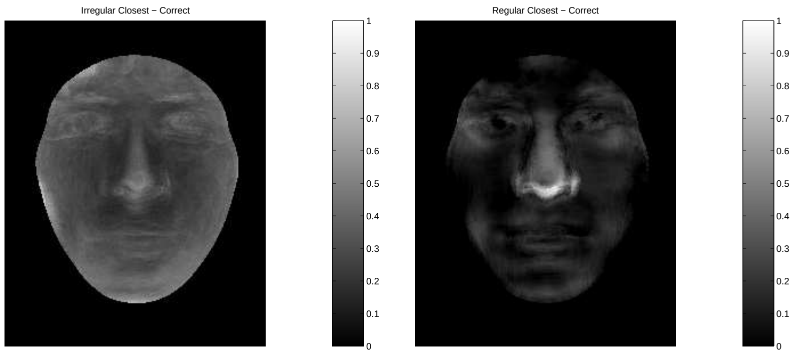


Figure 6.5. Regular re-sampling for ICP is beneficial. Registered surfaces are shown with dotted lines. Surface points are depicted with triangles before re-sampling, and with squares after re-sampling.

classification. Fig. 6.6 depicts the mean distance differences between the correct class and the incorrect class for these samples. For the irregular point distributions, the distance terms that add up to total face-to-face distance have large values (shown with light colour) all over the face. However, for regular re-sampled point distributions, large areas on the face have very low (shown with dark colour) error. As we predicted, error in re-sampled faces peaks for locations with greater depth gradient, and especially for the nose ridge. Since the nose ridge is a relatively small area of the face, increased error here is compensated by decreased error on the larger facial surfaces. Furthermore, the nose ridge error is increased for competing classes as well.

Looking at a test sample which is classified incorrectly before re-sampling, but correctly afterwards, we can see the effect of re-sampling more easily. In Fig. 6.7 we look at such a test sample more closely. The experimental setup is graphically depicted, followed by the difference images. We register the test scan T with the generic AFM, and measure PSD distances to the gallery samples. The sample with the smallest sample is labeled as A. This sample does not belong to T's class; The correct gallery sample (we have one gallery sample per class) is C. The first two distance images shown in Fig. 6.7 are the distances (T-A) and (T-C). These images show the local error for each point on the surface for the irregularly sampled faces. The third and the fourth images are the distances between the regularly re-sampled versions, (T'-A') and (T'-C'). The re-sampled faces have larger areas of low-error. This is not because



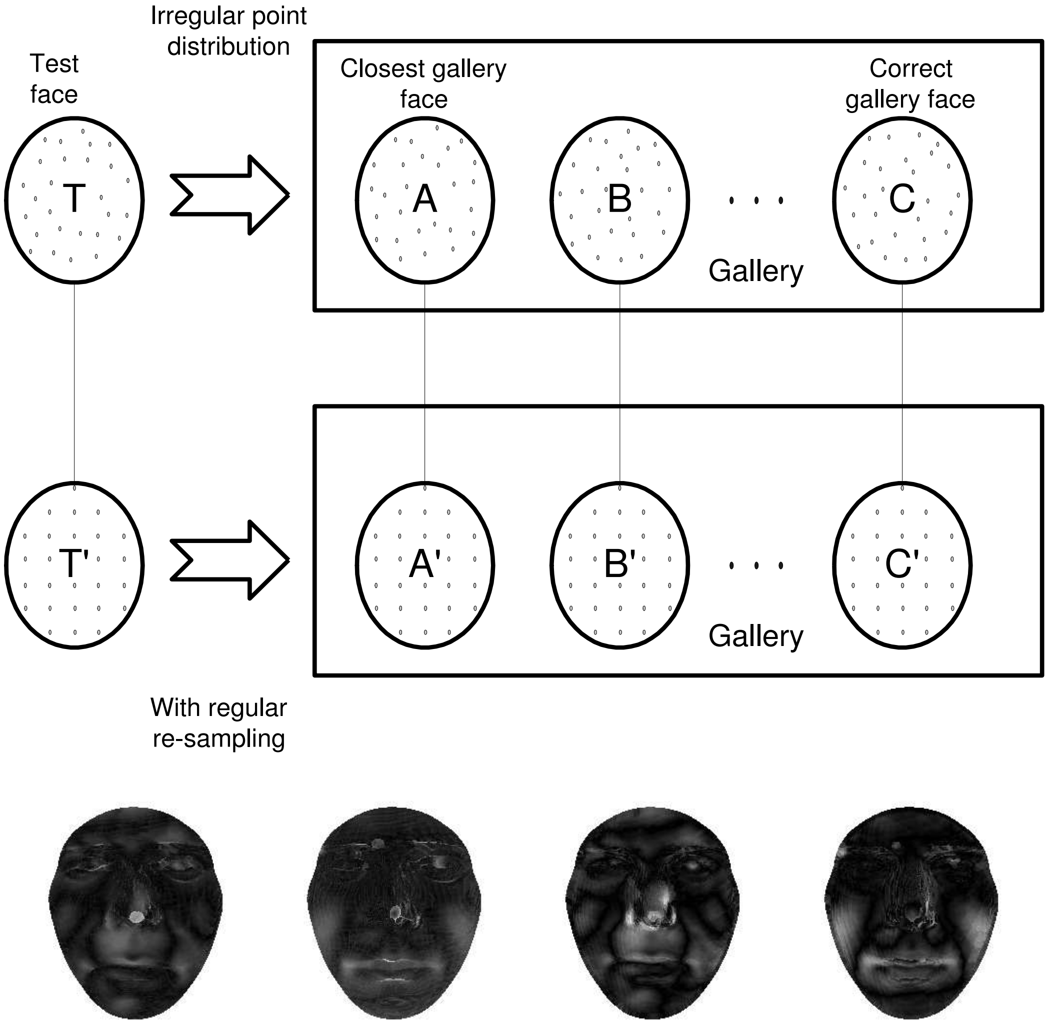
(a) (b)

Figure 6.6. Mean point-to-point distance differences in classification. The distances to the correct sample are subtracted from the distances to the closest sample. (a) Irregularly sampled points create errors uniformly on the face surface. (b) Regular sampling reduces the errors on the inner face, close to the nose where the registration is most accurate.

they are better registered, but the distance measure is more accurate. The nose ridge area in (T'-A') registers a higher error than (T'-C'), and this is frequently observed in other gallery images. Consequently, the nose area becomes useful in discrimination, even though sometimes it is the highest error area during registration with the correct gallery sample.

The re-sampling does not have a high computational cost, as the points are already ordered in the range image. For one-to-all ICP, it is possible to perform a similar re-sampling. However, if the gallery faces are not in alignment, the re-sampling has to be performed online for each gallery face separately. Another benefit of using the AFM is that the re-sampling is just performed once for each test face, and the computation is offline for gallery samples.

When we experiment with race and gender information, we make sure that for each comparison, the training and test samples are registered to the same AFM. For example, in the simulations where the gender is available, we register a male test face with the male AFM before comparing it to the male faces in the gallery, but we use the female AFM for comparisons with female gallery faces. Thus, we have two options



(T-A) (T-C) (T'-A') (T'-C')

Figure 6.7. The sketch shows the experimental setup for error analysis. The images depict point-to-point distances between the test face and a gallery face for a single sample, classified correctly only after re-sampling. T is the irregularly sampled test face, A is the closest gallery face to T, C is the gallery face of the correct class. T', A', and C' are the regularly re-sampled versions.

when using categorical AFMs: We can either inject ground truth information (for instance by setting intra-group distances to infinity), or we can let the system decide on the face category by picking the match with the smallest distance, like we do for shape space clustering based categories. The results reported in Table 6.4 are obtained with the latter method. Our simulations show that injecting the ground truth increases the accuracy only by 0.5–1 per cent. This means that for this dataset, cross-gender and cross-morphology errors are relatively rare, and we obtain categorical information with great accuracy by simply selecting the best gallery face.

Another potentially problematic issue we have considered was the number of points in different AFMs. The AFM for the females contains roughly 20 per cent fewer points that the AFM for the males. In our first experiments, we have used a single mask to crop faces in all categories. This procedure gives faces with equal numbers of points. The results of Table 6.4 are obtained by allowing each category to have a different number of points after cropping, and the Euclidean distances are normalized by dividing them to the square root of the number of points in the AFM. For the gender case, this procedure increased the accuracy by a half per cent.

6.7.5. Face Fragments

The registration of facial surfaces is performed holistically in our model. We have conducted an additional experiment to determine the relative contributions and discriminativeness of face fragments in 3D. Five patches were selected on the generic AFM as candidate fragments (See Fig. 6.8). The PSD and Eigenface methods are applied to patches after regular re-sampling.

The simulation results are reported in Table 6.5. The experiments suggest that the eyes are not as discriminative in 3D as they are in 2D. A very likely reason for this is that the eye regions are initially very noisy (e.g. there are often large holes to be filled) in the FRGC dataset, and it is not possible to recover discriminative information even with the pre-processing.

Our previous simulations suggested that the nose plays a key role in accurate 3D face registration. Experiments with face fragments show that the nose region is more robust than the eye and mouth regions in recognition. The inclusion of the rigid patch between the eyes is also helpful, as demonstrated by the high accuracy of face-centre fragment.

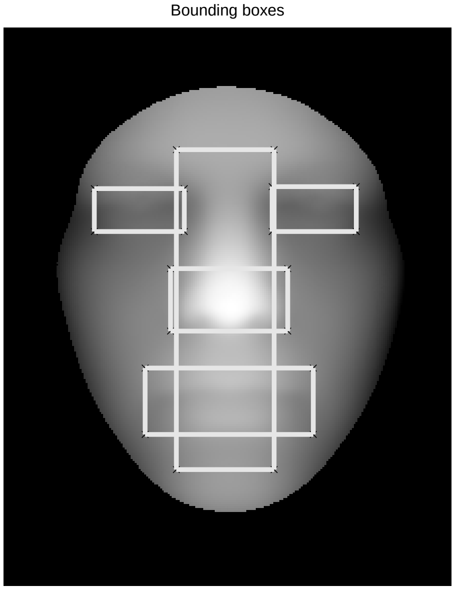


Figure 6.8. Five patches are selected on the generic AFM for the component-based recognition experiment.

6.8. Conclusions

We have evaluated ICP and TPS based registration of 3D faces with generic and specific average face models. Our proposed AFM generation method produces good models, and speeds up registration. The shape space clustering method revealed natural groups depending on morphology and gender in the face space, but also incorporated other factors that were useful for registration. Subsequently, the specific AFMs obtained with shape space clustering increased the accuracy of ICP.

Our experimental results have also confirmed that ICP is sensitive to initialization, and automatical landmarking as a pre-processing step is beneficial to ICP. The nose-tip heuristic may be useful in frontal faces, but the hair, clothing and sometimes the chin can be erroneously detected as the nose tip. The error due to incorrect nose localization can be gauged by looking at the results of the simulations that use the ground-truth for the nose in initialization. We should also keep in mind that the database we use is made up of near-frontal faces. The nose-tip heuristic will perform worse in any other pose settings.

Table 6.5. Comparison of face fragments

Landmarks	Distance	Left eye	Right eye	Nose	Mouth	Face centre
Auto	PSD	54.63	52.81	76.63	74.66	95.14
Auto	Eigenface	54.17	52.05	76.33	73.60	95.14
Manual	PSD	55.08	54.02	79.97	77.69	96.66
Manual	Eigenface	54.93	53.87	79.21	76.18	96.51
Dimensionality		$d = 1.081$ $p = 36$	$d = 1.056$ $p = 40$	$d = 2.013$ $p = 50$	$d = 3.045$ $p = 34$	$d = 8.415$ $p = 50$

The results show that the TPS based method is much inferior to ICP in accuracy. The beneficial effect from specific AFMs is evident in TPS methods that use either automatic or manual landmarks. Improvement is observed in the ICP method for the shape space clustering method, but not for the gender & morphology-specific AFMs. It is interesting to note that although the gender & morphology-based specific AFM method reduces the number of candidates for classification (whereas shape space derived AFM method does not), there is no accuracy improvement.

There may be several reasons for this lack of improvement. The categorical information is apparently not as beneficial as one would hope; closer inspection reveals that cross-gender and cross-morphology confusions are relatively rare. We can also argue that the distribution of categories in the training and test sets were uneven, and this reduced the quality of some of the specific AFMs. The TPS based method suffers less, because it only uses the landmark locations during registration, and not the actual AFM surface. With greater gallery sizes, we can conjecture that registration with category-specific AFMs can act as a filter, and reduce the number of gallery faces compared with the test scan.

A final observation is that the specific AFM models have different numbers of points. A male face usually contains 20 per cent more points than a female face. When we align a face to the female and the male AFMs, the distribution of distances is different in the centre of the face and at the periphery. Using a smaller AFM (the one for the females, or Asians, for instance) will effectively remove the points close to the periphery from the distance calculation. This can be an issue for one-to-all ICP approaches as well.

With the introduction of regular re-sampling, we have obtained 98.03 per cent rank-1 recognition rate on the FRGC ver.1 dataset. If we use manual landmark locations instead of using the automatic landmarking, the accuracy increases to 98.18 per cent. These are the highest results reported in the literature for this dataset. In [204], Gökberk *et al.* compare a wealth of classification methods on the same dataset: Point set difference, non-negative matrix factorization (NMF) and independent component analysis (ICA) coefficients for point clouds, DCT, DFT, PCA, LDA, and ICA projections on depth images, shape indices, mean and principal curvatures, 3D voxel DFT coefficients and 2D Gabor wavelet coefficients. Manual landmarks were used for registration. Their best classification methods are based on shape indices (90.06 per cent), principal directions (91.88 per cent) and surface normals (89.07 per cent). The best accuracy after classifier fusion is 93.63 per cent, obtained with modified plurality voting [204]. Our results are significantly better than these results. A direct comparison is possible, because we use the same set of landmarks and the same evaluation protocol (designated with E_1 in their paper). These results confirm our claim that the quality of landmarking and the subsequent registration is essential to the performance of the face recognition system.

We have conducted a pilot experiment on a 200 example subset of FRGC ver.2, where the 943 images of ver.1 are used as the gallery. We have obtained 87 per cent accuracy with manual landmarks and the generic AFM, and 89 per cent accuracy with the shape-space clustering approach. With the full ver.2 for testing (2010 samples that have corresponding subjects in the gallery), we have 86.02 per cent rank-1 accuracy with manual landmarks, and 86.37 per cent accuracy with automatic landmarks. In [204], it is reported that the best single expert (point cloud + ICA) has 88.31 per cent accuracy on FRGC ver.2. The slight accuracy increase in automatic landmarking is probably due to more consistent nose localization. The ground truth shows some variation, depending on the researcher doing the manual landmarking.

7. CONCLUSIONS

7.1. Contributions and Discussion

The conditions that lead humans to become face experts are significantly different than the conditions under which we develop computer models. Furthermore, with advances in sensor technology, storage capacities, and processing power, the latter conditions change in time. Compared to ten or even five years ago, we have access to larger databases (on the order of 5.000 images for 3D face recognition), and more powerful computers that are capable of computing more demanding image transformations. As the hardware becomes more capable, the human brain becomes a more lucrative source of ideas to guide computing.

In this thesis, we have reviewed human face recognition, and proposed a model for biologically motivated 3D face recognition, based on our findings. Human face recognition has a number of properties that lend themselves to computer modeling. For instance, faces are perceived more holistically than analytically, but features are also taken into account on their own. Structural information is used in addition to feature information. Discriminative features of faces are learned in time, and different races exhibit different facial characteristics, ideally represented with separate sets of projection features. The face space hypothesis suggests that faces reside in a multi-dimensional manifold, whose directions coincide with *meaningful* changes in face images. These hidden factors can be the facial identity, expression, illumination, and the pose of the face. Thorough exploration of the face space potentially makes possible to lay these factors bare. These considerations were incorporated into our model.

One can simulate a holistic face model in a computer, provided that the face images reside in a consistent coordinate frame. This in turn requires a good registration of face images. We have identified the problem of registration as a key issue in face recognition research. Registration in computers, as is in humans, is driven by facial features. Additionally, the analytical part of face recognition relies on recognition of facial features. A powerful feature detection method was necessary for these reasons. The IMoFA algorithm developed in Chapter 4 serves this purpose.

The IMoFA model has two major benefits. First of all, it allows fully automatic modeling, without user intervention. Many learning algorithms require that the user specify some parameters. Neural networks are especially notorious in this respect, since the model architecture and the learning regime have a great influence on the final accuracy of the system. Automatic tuning of the model complexity is a very desirable property in learning systems: Complex models are more expressive, but more difficult to train properly.

The second benefit of IMoFA is the flexibility it offers in fitting a Gaussian mixture to a dataset. Full-covariance Gaussian models are very clumsy in higher dimensions, and their training sample requirements scale quadratically with the number of dimensions. On the other hand, diagonal-covariance Gaussians only use variances in each dimension, and discard covariance information. IMoFA finds a good trade-off between these cases, performing better than its competitors in many different pattern recognition tasks.

The automatic landmarking system we have proposed in Chapter 5 is a statistical approach to landmark localization, purposefully avoiding heuristics for robust feature localization. Contrasting 2D Gabor, 3D depth, and 3D-assisted 2D albedo features, we confirmed that generalization across illumination conditions is very difficult, whereas 3D information is much more robust. A recent 3D-assisted illumination recovery algorithm was tested. Although the results looked promising for some of the cases, the algorithm did not perform nearly as well as expected. More elaborate schemes for texture recovery were found to be too costly for landmark localization, which is supposed to be a pre-registration step, and consequently needs to be performed as fast as possible. For this reason, we have proposed a coarse-to-fine model and avoided curvature-based features.

To incorporate the structural information, we have developed the GOLLUM algorithm. Like IMoFA, GOLLUM is not specific to face recognition: It can be used to recover errors from any shape descriptor. It models the distribution of shapes with a simple statistical model, and evaluates different interpretations of a landmark configuration. The most plausible configuration is used to correct wrong landmarks. We have contrasted GOLLUM with a recent competing structural correction algorithm called BILBO, and shown that for shapes with relatively few landmarks, GOLLUM performed significantly better. Separating local feature similarity from structural information allowed us to avoid complex optimization functions with many local minima, and we have shown that it is possible to detect 3D face features robustly with statistical methods.

The landmarks are especially useful in guiding the registration. We have contrasted the rigid ICP approach with the nonrigid TPS approach for registration. We have proposed a novel way of creating an average face model, which results in better registration when compared to a previously proposed AFM generation method. Our simulations have shown the superiority of ICP in registration. The deformation based methods do not perform as well, primarily because the errors in the landmarking have a large effect on them. ICP is able to recover from landmarking errors, because it uses them only for initialization. However, ICP is a costly method, especially when it is used to register a test scan to all faces in the gallery. Our proposed AFM-based registration method limits the registration to a single ICP run per test scan, and paves the way for very large gallery sizes. Based on our results, we would commend the point set representation for 3D faces for registration, and regularized depth maps for classification.

Our purpose is to evaluate the quality of landmarks and the subsequent registration. Methods of classification are beyond our scope. We have adopted a PSD measure, and proposed a modification via regular re-sampling to make it faster and significantly more accurate. On FRGC ver.1 and a subset of FRGC ver.2, our PSD-based method reached higher accuracies than results with more elaborate classification schemes reported in the literature. Our simulations were run on the most difficult FRGC ver.1

protocol, with a single training image per person. We have shown that the accuracy loss due to automatic landmarking is very small (from 98.18 per cent to 98.03 per cent) with a generic AFM-based registration. We have also demonstrated that the template sizes can be reduced to 50 dimensions from nearly 30.000 dimensions with minor loss in accuracy. Using a 50 dimensional PCA projection with automatic landmarks gave 97.88 per cent accuracy. As our cognitive model of human face recognition suggested, it is important to have the faces brought to a common frame of reference. We have confirmed that the quality of registration is as important as the choice of classification method.

Finally, we have proposed a biologically motivated way of clustering faces into groups for within group registration and classification. We have obtained marked increase of accuracy with the TPS-based method, but the accuracy did not change for the ICP-based method. The shape-space clustering technique we have proposed reached exactly the same accuracy with the single-AFM approach on FRGC ver.1. However, our pilot study on FRGC ver.2 suggests that this is a ceiling effect, and we have a full two per cent accuracy increase for FRGC ver.2 with the multiple-AFM approach.

It is clear that the representation and processing of depth information opens new doors in face recognition research. We believe that an increased understanding of human cognitive facilities will lead to better computer models. It is our hope that the present work serves as a demonstration of possibilities, and raises more questions than it answers.

7.2. Future Directions

We have emphasized the importance of locating facial features for the purposes of registration, and sought to locate a few features with discriminating local properties. Other salient features of the face include tip of the chin, the nasion, the nostrils, the eyebrows, and the centres of the irises. The present set of seven landmarks seems to be sufficient for registration, but there are other applications that would benefit from more landmarks. Automatic facial animation, or expression recognition are important research areas with industrial applications. These would benefit tremendously from the automatic localization of 50-60 fiducial points. Our system would serve as a good starting points for this type of extensions.

We have shown that 2D, although potentially more rich in features, has poor generalization across illumination conditions. The 3D-assisted 2D schemes need to be developed further to be useful for registration. More illumination-invariant features, particularly on the spectral domain, can be explored. The computer algorithms do not have this particular ability of the human visual system to perceive objects to be the same under different illumination conditions, partly because this sort of constancy is very difficult to obtain with purely statistical methods. Illumination compensation is an active area of research.

The classification part of the system was not fine-tuned, because there are so many different approaches to classification. Even with the simple point set difference-based classification, our results are excellent when compared to the latest results reported in the literature. We have identified the reason as robust and accurate registration. A natural extension is to implement different and more elaborate classifiers to improve the overall system.

Another open issue in classification is how to make use of parts and wholes. Face parts are correlated with face wholes, and simple combination strategies do not promise much improvement for highly-correlated classifiers. The best combination results are obtained when the experts have different error characteristics. However, more intelligent ways of combining classifiers are available. The GOLLUM structural correction scheme is able to provide reliability measures for the final registration, and for each individual landmark. The reliability can be used for guiding a more intuitive combination scheme. Again, there are numerous possibilities, and these are research issues that need to be inspected more thoroughly.

We have separated local feature information from the structural information. There are popular approaches to face recognition that jointly optimize local feature similarity and structural similarity. Our argument in separating the two was to keep the complexity low, and to escape incorrect local maxima during optimization. We have introduced a likelihood-based local similarity measure, and shown that it is better than the existing jet-based methods. One thing that needs to be investigated is, whether our new feature similarity measure can make joint-optimization approaches more robust or not. With several approaches to investiagate, and with many parameters to tune, this is a separate research subject on its own.

Our results with the alternative face models for registration were promising, but not conclusive. We believe that there is room for great improvement following that line of argument. The problem of uneven distribution of samples in the training set was alleviated by using a shape-space clustering approach, but learning discriminative manifolds within each subset remains an open issue.

Finally, the IMoFA and GOLLUM methods that were developed are not specific to face images, but applicable to other objects as well. We were careful to include no face image- or database-specific heuristics in our methods, and used the statistical properties of the data in a general manner. IMoFA was tested with ten different pattern recognition problems, and was found to be successful. Their application to other problems is left as a future direction.

APPENDIX A: GLOSSARY

- **2.5D-image:** Bowyer et al. refer to the face surface represented by a range image as 2.5D, as opposed to 3D, which is the whole head [149]. Lu et al. define it as "a simplified 3D (x, y, z) surface representation that contains at most one depth value (z direction) for every point in the (x, y) plane" [165]. Among the papers that use 2.5D are [122, 169, 153, 165, 111, 166, 227, 163, 182, 199, 151].
- **3D morphable face model:** A model of faces represented in 3D, where shape information is separated from texture information [293, 157]. Building a 3D morphable face model requires to transform the shape and texture spaces into vector spaces for which any convex combination of exemplar shapes and textures describes a realistic human face. This is achieved by setting the exemplar faces in full correspondence with respect to a reference shape. Correspondences between all exemplar faces and the reference face are established by an optical flow algorithm. This produces a 3D deformation field for each exemplar face which is used to warp the textures onto the reference shape yielding a shape-free texture. This scheme introduces a consistent labeling of vertices across the whole set of exemplar faces.
- **3D transformations:** On a three dimensional coordinate system (X-Y-Z), the pose variation around the X-, Y-, and Z-directions are called the *roll*, *pitch*, and *yaw*, respectively.
- **4D data:** As used in [128], 4D data consists of a 3D mesh and related 2D texture map.
- **Cumulative Match Characteristic (CMC) curve:** In the recognition setting, it is useful not only to report the cases where the person is correctly identified, but to indicate how close the misses were. Therefore the rank-n recognition rate is reported, which gives the percentage of cases where the correct match was within the n best-matching faces in the gallery. Naturally, as n increases, the rank-n recognition rate increases as well. When n equals the number of gallery faces, the rank-n recognition rate reaches 1.0. The CMC curve plots this rate against n, and steepness is indicative of greater success. Also called Cumulative

Match Score (CMS) curve.

- **Extended Gaussian Image (EGI):** The EGI of a 3D object is a translation invariant histogram of changes in the surface area for all possible orientations [302] (See Fig. A.1). For any two different convex objects, the EGIs are different as well. The complex EGI (CEGI) is introduced by adding a phase component that records distances from the origin to the surface for each point [303].

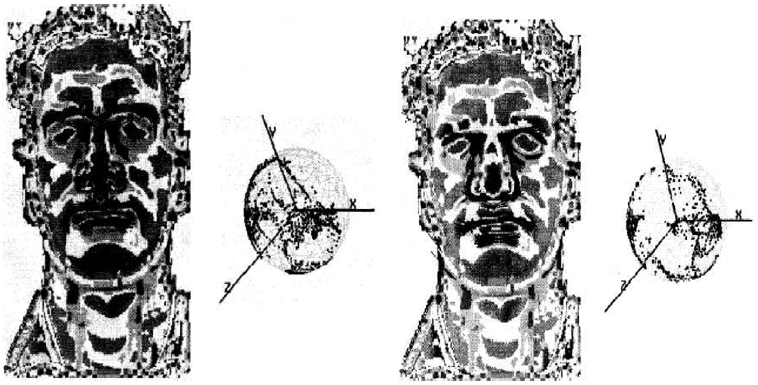


Figure A.1. Ridge and valley lines of surface curvatures, and their EGI representations (from Tanaka et al., 1998).

- **Fundamental matrix:** The fundamental matrix is a transformation that defines a mapping between the points in one image and their duals (defined by the conjugate of a homography M) in another image. These are also called the *epipolar lines.*
- **Gaussian-Hermitian moments:** The signal is convolved with a Gaussian kernel and a Hermitian polynomial function, and then integrated. It is used as a noise insensitive local feature [188, 162].
- **Geometry image:** A 2D array of quantized points that is obtained by mapping a mesh of a 3D object to an array and treating 3D features (coordinates or surface normals) as colour codes [304]. A 3D object stored this way can be treated as a regular 2D image, and compressed as such.
- **Hausdorff distance:** Used to measure similarity of two sets of points, applied to image comparison in [305]. It is small if every point in the first set is near some point in the other, and vice versa. If only the best ranked points are taken into account, outliers will not affect the metric much. It is asymmetrical. Used in [160, 166].

- HK **segmentation method:** H stands for the mean curvature, K stands for the Gaussian curvature. In this method, points are labeled according to the signs of H and K. If $K < 0$, the point is *hyperbolic*. If $K > 0$ and $H < 0$ it is *elliptical convex*. If both are positive, it is *elliptical concave*. It is difficult to obtain zero curvature values, therefore the *cylindrical* points ($K = 0$) and *planar* points (both equal to zero) are usually not found in the model. For segmentation, H and K values are thresholded from both sides, with possibly different values. Regions are selected according to the characteristics of the points they contain. Used in [155, 182].
- **Iterative Closest Point (ICP) Algorithm:** ICP is first proposed in [1]. It iteratively registers a point cloud to a model, which can be a point cloud, a curve, a surface, parametric or in implicit representation. A quaternion based 3D rotation and 3D translation is found as the resulting transformation. A number of initial registrations for coarse alignment are necessary, but the ICP registration itself is fast. All the points in the matched shape must have correspondences in the model (global or local alignment).
- **Point Signature:** This is a feature extraction technique proposed in [306]. For a given point on the 3D object, a sphere of fixed radius is placed on this point. The perpendicular projection of the intersection curve between the object and the sphere to a reference plane is found. The projection distances can be treated as a translation and rotation invariant, 1-dimensional, signed distance profile (See Fig. A.2).
- **Procrustes Analysis:** In this technique the shapes are aligned to minimize the sum of distances to the mean [294]. This can be achieved iteratively, by translating samples so that their centre of gravity will coincide, and then by proceeding like a 1-means clustering; by computing a mean, aligning the shapes and re-computing the mean. The scale and/or orientation can be varied during alignment, which has an influence in the final result. Another approach is to project the samples to the tangent space of the mean. The tangent space of $\mathbf{x}$ contains those vectors that are orthogonal to $\mathbf{x}$ and pass through it. This transformation helps to maintain a linearity in the shape variation.

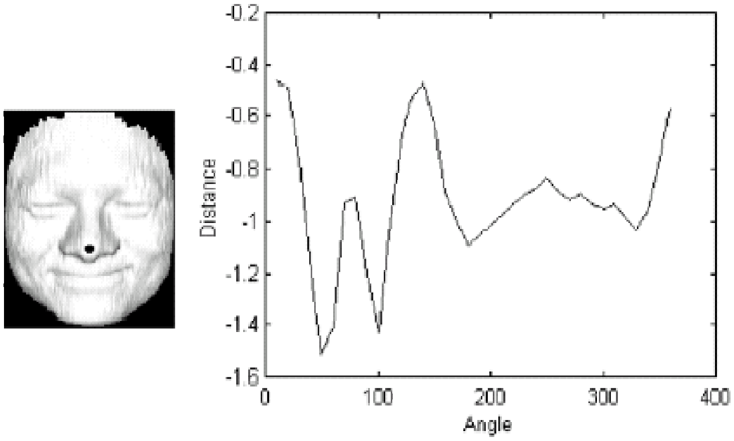


Figure A.2. Point signature of a 3D point (from Wang et al., 2002).

- **Receiver Operator Characteristic (ROC) curve:** The costs of mistakenly identifying a person (false positive) and failing to identify a person (false negative) usually depend on the application and not necessarily equal. The ROC curve plots false negatives versus false positives for the whole range of a given parameter, and indicates how one can change the conservativeness of the system. If the only performance metric is the correct recognition rate, this is equivalent to treating false positives and negatives as equal, and we operate on the point where the ROC curve intersects the 45° line.
- **Texture map:** The texture map is the reflectance component of a 3D scan. O'Toole remarks that this definition of the texture map by computer scientists is unrelated to the use of the same term by vision scientists [307].
- **Thin-Plate Spline (TPS) Algorithm:** TPS is proposed in [176]. It is a method for mapping a set of point to another set of point by a parametric deformation model.
- **z-buffering Algorithm:** An algorithm used to derive depth maps from 3D surfaces (proposed by [308], employed and summarized in [199]). It starts by defining an array (of depth values) for each pixel on the 2D result, and iterates through all polygons in the 3D representation to update the depth values of the pixels.

REFERENCES

1. Besl, P., and N. McKay, “A Method for Registration of 3-D Shapes,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.14, no.2, pp.239-256, 1992.

2. Sim, T., R. Sukthankar, M. Mullin, and S. Baluja, “Memory-based face recognition for visitor identification,” in *Proc. IEEE Inf. Conf. on Automatic Face and Gesture Recognition*, 2000.

3. Salah, A.A., E. Alpaydın, and L. Akarun, “A Selective Attention Based Method for Visual Pattern Recognition with Application to Handwritten Digit Recognition and Face Recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.24, no.3, pp.420-425, 2002.

4. Salah, A.A., M. Bicego, L. Akarun, E. Grosso, and M. Tistarelli, “Hidden Markov Model-based face recognition using selective attention,” in *SPIE Conf. on Human Vision and Electronic Imaging*, San Jose, 2007.

5. Salah, A.A., “Face Recognition in Humans and Computers,” (in Turkish) in *Proc.* 3^{rd} *Symposium on Brain as a Computing Machine*, Istanbul, 2005.

6. Gross, C.G., D.B. Bender, and C.E. Rocha-Miranda, “Visual Receptive Fields of Neurons in Inferotemporal Cortex of the Monkey,” *Science*, vol.166, pp.1303-1306, 1969.

7. Farah, M.J., C. Rabinowitz, G.E. Quinn, and G.T. Liu, “Early Commitment Of Neural Substrates For Face Recognition,” *Cognitive Neuropsychology*, vol.17, pp.117-123, 2000.

8. Moscovitch, M., G. Winocur, and M. Behrmann, “What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition,” *Journal of Cognitive Neuroscience*, vol.9,

pp.555-604, 1997.

9. Tippett, J.L., L.A. Miller, and M.J. Farah, “Prosopamnesia: A Selective Impairment In Face Learning,” *Cognitive Neuropsychology*, vol.17, pp.241-255, 2000.

10. Blonder, L.X., C.D. Smith, C.E. Davis, M.L. Kesler/West, T.F. Garritya, M.J. Avison, and A.H. Andersen, “Regional brain response to faces of humans and dogs,” *Cognitive Brain Research*, vol.20, pp.384-394, 2004.

11. de Haan, M., M.H. Johnson, and H. Halit, “Development of face-sensitive event-related potentials during infancy: a review,” *International Journal of Psychophysiology*, vol.51, pp.45-58, 2003.

12. Grill-Spector, K., “The neural basis of object perception,” *Current Opinion in Neurobiology*, vol.13, pp.159-166, 2003.

13. Halit, H., G. Csibra, A. Volein, and M.H. Johnson, “Face-sensitive cortical processing in early infancy,” *Journal of Child Psychology and Psychiatry*, vol.45, no.7, pp.1228-1234, 2004.

14. Kanwisher, N., J. McDermott, and M.M. Chun, “The fusiform face area: A module in human extrastriate cortex specialized for face perception,” *Journal of Neuroscience*, vol.17, pp.4302-4311, 1997.

15. Puce, A., A. Smith, and T. Allison, “ERPs Evoked By Viewing Facial Movements,” *Cognitive Neuropsychology*, vol.17, pp.221-239, 2000.

16. Carey, S., and R. Diamond, “From piecemeal to configurational representation of faces,” *Science*, vol.195, pp.312-314, 1977.

17. Yin, R.K., “Looking at upside-down faces,” *Journal of Experimental Psychology*, vol.81, pp.141-145, 1969.

18. Tanaka, J.W., and M.J. Farah, “Parts and wholes in face recognition,” *Quarterly*

Journal of Experimental Psychology: Human Experimental Psychology, vol.46A, 225-245, 1993.

19. Ro, T., C. Russell, and N. Lavie, "Changing Faces: A Detection Advantage in the Flicker Paradigm," *Psychological Science,* vol.12, no.1, pp.94-99, 2001.

20. Gauthier, I., and N.K. Logothetis, "Is face recognition not so unique after all?," *Cognitive Neuropsychology*, vol.17, pp.125-142, 2000.

21. Gauthier, I., and M.J. Tarr, "Becoming a "Greeble expert": Exploring the face recognition mechanism," *Vision Research* vol.37, pp.1673-1682, 1997.

22. Tong, M.H., C.A. Joyce, and G.W. Cottrell, "Are Greebles special? Or, why the Fusiform Fish Area would be recruited for sword expertise (if we had one)," in *Proc.* 27th *Annual Cognitive Science Conference*, La Stresa, Italy. Mahwah: Lawrence Erlbaum, 2005.

23. O'Toole, A.J., S. Edelman, and H. Bülthoff, "Stimulus-specific effects in face recognition over changes in viewpoint," *Vision Research*, vol.38, pp.2351-2363, 1998.

24. Wallis, G., and H. Bülthoff, "A Brief Introduction to Cortical Representations of Objects," Technical Report 097, Max Planck Institute for Biological Cybernetics, 2002.

25. Bülthoff, H.H., S.Y. Edelman, and M.J. Tarr, "How are three-dimensional objects represented in the brain?," CogSci Memo no.5, Max Planck Institute for Biological Cybernetics, Tübingen, 1994.

26. Biederman, I., "Recognition by components: A theory of human image understanding," *Psychological Review*, vol.94, pp.115-145, 1987.

27. Marr, D., *Vision,* W. H. Freeman, San Francisco, CA, 1982.

28. Hummel, J.E., “Where view-based theories break down: The role of structure in shape perception and object recognition,” in E. Dietrich and A. Markman (eds.), *Cognitive Dynamics: Conceptual Change in Humans and Machines*, pp.157-185, Hillsdale, NJ: Erlbaum, 2000.

29. Wiskott, L., J.-M Fellous, N. Krüger, and C. von der Malsburg, “Face recognition by elastic bunch graph matching,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.19, no.7, pp.775-779, 1997.

30. Kavšek, M. “Development of depth and object perception in infancy,” in G. Schwarzer and H. Leder (eds.) *The Development of Face Processing,* Hogrefe and Huber Pub., 2003.

31. Read, J., “Early computational processing in binocular vision and depth perception,” *Progress in Biophysics and Molecular Biology*, vol.87, pp.77-108, 2005.

32. Stringer, S.M., and E.T. Rolls, “Invariant Object Recognition in the Visual System with Novel Views of 3D Objects,” *Neural Computation*, vol.14, no.11, pp.2585-2596, 2002.

33. Duchowski, A.T., E. Medlin, N. Cournia, A. Gramopadhye, B. Melloy, and S. Nair, “3D eye movement analysis for VR visual inspection training,” in *Proc. 2002 Symposium on Eye Tracking Research and Applications*, pp.103-110,New Orleans, Louisiana, 2002.

34. Howlett, S., J. Hamill, C. and O’Sullivan, “Predicting and Evaluating Saliency for Simplified Polygonal Models,” *ACM Trans. Applied Perception*, vol.2, no.3, pp.286-308, 2005.

35. Chung, A.J., F. Deligianni, X.-P. Hu, and G.-Z.Yang, “Extraction of visual features with eye tracking for saliency driven 2D/3D registration,” *Image and Vision Computing*, vol.23, no.11, pp.999-1008, 2005.

36. Piaget, J., “Piaget’s theory,” in P.H. Mussen (ed.), *Carmichael’s Manual of Child Psychology*, New York: Wiley, 1970.

37. Karmiloff-Smith, A., *Beyond modularity: A developmental perspective on cognitive science*, Cambridge MA: MIT Press, 1992.

38. Reingold, E.M., N. Charness, M. Pomplun, and D.M. Stampe, “Visual Span In Expert Chess Players: Evidence From Eye Movements,” *Psychological Science*, vol.12, no.1, pp.48-55, 2001.

39. Peirce, J.W., A.E. Leigh, A.P.C. daCosta, and K.M. Kendrick, “Human face recognition in sheep: lack of configurational coding and right hemisphere advantage,” *Behavioural Processes*, vol.55, pp.13-26, 2001.

40. Yarbus, A.L., *Eye Movement and Vision,* Plenum Press, New York, NY, 1976.

41. Diamond, R., and S. Carey, “Why faces are and are not special: an effect of expertise,” *Journal of Experimental Psychology:General*, vol.115, no.3, pp.107-117, 1986.

42. Tarr, M.J., and Y.D. Cheng, “Learning to see faces and objects,” *Trends in Cognitive Sciences*, vol.7, no.1, pp.23-30, 2003.

43. Grill-Spector, K., N. Knouf, and N. Kanwisher “The fusiform face area subserves face perception, not generic within-category identification,” *Nature Neuroscience*, vol.7, no.5, pp.555-562, 2004.

44. McKone, E., N. Kanwisher, and B.C. Duchaine, “Can generic expertise explain special processing for faces?,” *Trends in Cognitive Sciences*, vol.11, no.1 , pp.8-15, 2007.

45. Tomasello, M., *The cultural origins of human cognition*, Harvard University Press, Cambridge, Mass., 1999.

46. Meltzoff, A.N., and N.K. Moore, "Imitation of facial and manual gestures by human neonates," *Science*, vol.198, pp.75-78, 1977.

47. Johnson, M.H., and J. Morton, *Biology and cognitive development: The case of face recognition,* Oxford, UK: Blackwell, 1991.

48. Morton, J., and M.H. Johnson, "CONSPEC and CONLERN: A two-process theory of infant face recognition," *Psychological Review*, vol.98, no.2, pp.164-181, 1991.

49. Haxby, J.V., E.A. Hoffman, and M.I. Gobbini, "Human Neural Systems for Face Recognition and Social Communication," *Biological Psychiatry*, vol.51, pp.59-67, 2002.

50. Xu, F., and S. Carey, "Infants' metaphysics: The case of numerical identity." *Cognitive Psychology*, vol.30, no.2, pp.111-153, 1996.

51. Goren, C.C., M. Sarty, and P.Y.K. Wu, "Visual following and pattern discrimination of face-like stimuli by newborn infants," *Pediatrics*, vol.56, pp.544-549, 1975.

52. Maurer, D., and M. Barrera, "Infants' perception of natural and distorted arrangements of a schematic face," *Child Development*, vol.47, pp.523-527, 1981.

53. Bonatti, L., E. Frot, R. Zangl, and J. Mehler, "The Human First Hypothesis: Identification of Conspecifics and Individuation of Objects in the Young Infant," *Cognitive Psychology,* vol.44, pp.388-426, 2002.

54. Atkinson, J., O. Braddick, and K. Moar, "Development of contrast sensitivity over the first 3 months of life in the human infant," *Vision Research*, vol.17, pp.1037-1044, 1977.

55. Elman, J.L., E.A. Bates, M.H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett, *Rethinking innateness: A connectionist perspective on development*,

Cambridge, MA: MIT Press, 1996.

56. Turati, C., "Why Faces Are Not Special to Newborns: An Alternative Account of the Face Preference," *Current Directions In Psychological Science*, vol.13, no.1, pp.5-8, 2004.

57. Turati, C., and F. Simion, "Newborns' recognition of changing and unchanging aspects of schematic faces," *Journal of Experimental Child Psychology*, vol.83, pp.239-261, 2002.

58. Pascalis, O., M. de Haan, and C.A. Nelson, "Is face processing species-specific during the first year of life?," *Science*, vol.296, pp.1321-1323, 2002.

59. Haxby, J.V., E.A. Hoffman, and M.I. Gobbini, "The distributed human neural system for face perception," *Trends in Cognitive Sciences*, vol.4, no.6, pp.223-233, 2000.

60. Andrews, T.J., and D. Schluppeck, "Neural responses to Mooney images reveal a modular representation of faces in human visual cortex," *NeuroImage*, vol.21, pp.91-98, 2004.

61. Gauthier, I., M.J. Tarr, J. Moylan, A.W. Anderson, P. Skudlarski, and J.C. Gore, "Does visual subordinate-level categorisation engage the functionally-defined fusiform face area?," *Cognitive Neuropsychology*, vol.17, pp.143-163, 2000.

62. Bentin, S., T. Allison, A. Puce, E. Perez, E., and G. McCarthy, "Electrophysiological studies of face perceptions in humans," *Journal of Cognitive Neuroscience*, vol.8, pp.551-565, 1996.

63. Tanaka, J.W., and T. Curran, "A Neural Basis for Expert Object Recognition," *Psychological Science*, vol.12, no.1, 2001.

64. Rossion, B., I. Gauthier, V. Goffaux, M.J. Tarr, and M. Crommelinck, "Expertise Training with Novel Objects Leads to Left-Lateralized Facelike Electrophysiolog-

ical Responses," *Psychological Science*, vol.13, no.3, pp.250-257, 2002.

65. Perrett, D.I., E.T. Rolls, and W. Caan, "Visual neurones responsive to faces in the monkey temporal cortex," *Experimental Brain Research,* vol.47, pp.329-342, 1982.

66. Perrett, D.I., A.J. Mistlin, and A.J. Chitty, "Visual neurones responsive to faces," *Trends in Neurosciences*, vol.10, pp.358-364, 1989.

67. Poggio, T., and A. Hurlbert, "Observations on cortical mechanisms for object recognition and learning," AI Memo No.1404, MIT, 1993.

68. Ellis, H.D., A.W. Young, A.H. Quayle, and K.W. de Pauw, "Reduced autonomic responses to faces in Capgras delusion," *Proc. Royal Society of London Series B*, vol.264, pp.1085-1092, 1997.

69. Hirstein, W., and V.S. Ramachandran, "Capgras syndrome: a novel probe for understanding the neural representation of the identity and familiarity of persons," *Proc. Royal Society of London Series B*, vol.264, pp.437-444, 1997.

70. Breen, N., D. Caine, and M. Coltheart "Models of face recognition and delusional misidentification: a critical review," *Cognitive Neuropsychology*, vol.17, pp.55-71, 2000.

71. Ellis, H.D., and M.B. Lewis, "Capgras delusion: a window on face recognition," *Trends in Cognitive Sciences*, vol.5, pp.149-156, 2001.

72. Hancock, P.J.B., V. Bruce, and A.M. Burton, "Face processing: human perception and principal components analysis," *Memory & Cognition*, vol.24, pp.26-40, 1996.

73. Vokey, J.R., and J.D. Read, "Familiarity, Memorability, and the effect of typicality on the recognition of faces," *Memory and Cognition*, vol.20, pp.291-302, 1992.

74. Valentine, T., "A unified account of the effects of distinctiveness, inversion and

race in face recognition," *Quarterly Journal of Experimental Psychology*, vol.43A, pp.161-204, 1991.

75. Rakover, S.S., and B. Teucher, "Facial inversion effects: Parts and whole relationship," *Perception & Psychophysics*, vol.59, pp.752-761, 1997.

76. Peterson, M.A., and G. Rhodes, (eds.) *Perception of Faces, Objects, and Scenes,* Oxford University Press, New York, 2003.

77. Martelli, M., N.J. Majaj, and D.G. Pelli, "Are faces processed like words? A diagnostic test for recognition by parts," *Journal of Vision*, vol.5, no.1, pp.58-70, 2005.

78. Tanaka, J.W., M. Kiefer, and C.M. Bukach, "A holistic account of the own-race effect in face recognition: evidence from a cross-cultural study," *Cognition*, vol.93, pp.B1-B9, 2004.

79. Thompson, P., "Margaret Thatcher – A new illusion," *Perception*, vol.9, pp.483-484, 1980.

80. Lewis, M.B., and R.A. Johnston, "The Thatcher illusion as a test of configural disruption," *Perception*, vol.26, pp.225-227, 1997.

81. Carbon, C.C., "Face processing: Early processing in the recognition of faces," unpublished PhD Thesis, Freie Universität Berlin, 2003.

82. Sergent, J., "An investigation into component and configural processes underlying face perception," *The British Journal of Psychology*, vol.75, pp.221-242, 1984.

83. Ullman, S., and E. Sali, "Object classification using fragment-based representation," in S.-W. Lee, H.H. Bülthoff, T. Poggio (eds.), *Biologically Motivated Computer Vision*, LNCS 1811, pp.73-87, Berlin: Springer Verlag, 2000.

84. Heisele, B., and T. Koshizen, "Components for face recognition," in *Proc. 6th*

IEEE Inf. Conf. on Automatic Face and Gesture Recognition, 2004.

85. Brennan, S.E., "Caricature Generator: The Dynamic Exaggeration of Faces by Computer," *Leonardo*, vol.18, pp.170-178, 1985.

86. Rhodes, G., S. Brennan, and S. Carey, "Identification and ratings of caricatures: implications for mental representation of faces," *Cognitive Psychology,* vol.19, pp.473-497, 1987.

87. Lee, K., G. Byatt, and G. Rhodes, "Caricature Effects, Distinctiveness, And Identification: Testing the Face-Space Framework," *Psychological Science*, vol.11, no.5, pp.379-385, 2000.

88. Costen, N., I. Craw, T. Kato, G. Robertson, and S. Akamatsu, "Manifold Caricatures: on the Psychological Consistency of Computer Face Recognition," in *Proc. Inf. Conf. on Automatic Face and Gesture Recognition*, 1996.

89. O'Toole, A.J., T. Price, T. Vetter, J.C. Bartlett, and V. Blanz, "3D shape and 2D surface textures of human faces: the role of "averages" in attractiveness and age," *Image and Vision Computing*, vol.18, pp.9-19, 1999.

90. Benson, P.J., and D.I. Perrett, "Perception and recognition of photographic quality facial caricatures: implications for the recognition of natural images," *European Journal of Cognitive Psychology,* vol.3, pp.105-135, 1991.

91. Turk, M., and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol.3, no.1, pp.71-86, 1991.

92. Pentland, A., B. Moghaddam, T. Starner, O. Oliyide, and M. Turk, "View-Based and Modular Eigenspaces for Face Recognition," Technical Report 245, M.I.T Media Lab, 1993.

93. Bruce, V., P.J.B. Hancock, and A.M. Burton, "Comparisons between Human and Computer Recognition of Faces," in *Proc. Inf. Conf. on Automatic Face and*

Gesture Recognition, 1998.

94. O'Toole, A.J., P.J. Phillips, Y. Cheng, B. Ross, and H.A. Wild, "Face recognition algorithms as models of human face processing," in *Proc. 4th Int. Workshop on Automatic Face and Gesture Recognition*, pp.552-557, 2000.

95. Sukthankar, G., "Face recognition: A critical look at biologically-inspired approaches," Technical Report, CMU-RI-TR-00-04, The Robotics Institute, Carnegie Mellon University, 2000.

96. Ramasubramanian, D., and Y.V. Venkatesh, "Encoding and recognition of faces based on the human visual model and DCT," *Pattern Recognition*, vol.34, pp.2447-2458, 2001.

97. Daugman, J., "Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A*, vol.2, pp.1160-1169, 1985.

98. Tseng, S., "Comparison of Holistic and Feature Based Approaches to Face Recognition," unpublished thesis, School of Computer Science and Information Technology, Royal Melbourne Institute of Technology University, Australia, 2003.

99. Wallraven, C., and H.H. Bülthoff, "Automatic acquisition of exemplar-based representations for recognition from image sequences," in *Proc. CVPR*, Workshop on Models vs. Exemplars, 2001.

100. Wallraven, C., A. Schwaninger, S. Schuhmacher, and H.H. Bülthoff, "View-Based Recognition of Faces in Man and Machine: Re-visiting Inter-Extra-Ortho," in *LNCS*, vol.2525, pp.651-660, 2002.

101. Wallraven, C., A. Schwaninger, and H.H. Bülthoff, "Learning from humans: computational modeling of face recognition," *Network: Computation in Neural Systems*, vol.16, no.4, pp.401-418, 2005.

102. Kalocsai, P., W. Zhao, and E. Elagin, “Face similarity space as perceived by humans and artificial systems,” in *Proc. IEEE Inf. Conf. on Automatic Face and Gesture Recognition*, 1998.

103. Wallis, G., and H. Bülthoff, “Learning to recognize objects,” *Trends in Cognitive Sciences*, vol.3, pp.22-31, 1999.

104. O’Toole, A.J., D.A. Roark, and H. Abdi, “Recognizing moving faces: a psychological and neural synthesis,” *Trends in Cognitive Sciences*, vol.6, no.6, pp.261-266, 2002.

105. Slater, A., D. Rose, and V. Morison, “New-born infants’ perception of similarities and differences between two- and three-dimensional stimuli,” *British Journal of Developmental Psychology*, vol.2, pp.287-295, 1984.

106. Chellappa, R., C.L. Wilson, and S. Sirohey, “Human and machine recognition of faces: a survey,” *Proc. IEEE*, vol.83, no.5, 1995.

107. Akarun, L., B. Gökberk, and A.A. Salah, “3D Face Recognition for Biometric Applications,” in *Proc. European Signal Processing Conference*, Antalya, Turkey, 2005.

108. Salah, A.A., B. Gökberk, and L. Akarun, “3D Face Recognition,” (in Turkish) in *Proc. 5th GAP Engineering Congress*, Şanlıurfa, 2006.

109. Gökberk, B., A.A. Salah, and L. Akarun, “3D Face Recognition,” chapter in *Handbook of Multimodal Biometrics,* Springer Verlag, to appear.

110. Kittler, J., A. Hilton, M. Hamouz, and J. Illingworth, “3D Assisted Face Recognition: A Survey of 3D Imaging, Modelling and Recognition Approaches,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

111. Lu, X., and A.K. Jain, “Integrating Range and Texture Information for 3D Face Recognition,” *Proc. 7th IEEE Workshop on Applications of Computer Vision*

(WACV'05), pp. 156-163, Breckenridge, CO, 2005.

112. Poh, N., and S. Bengio, "Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication," Technical Report, IDIAP-RR 04-44, IDIAP, 2004.

113. Lane, R., and N. Thacker, "Stereo vision research: An algorithmic survey," 1996.

114. Tsai, R.Y., "An efficient and accurate camera calibration technique for 3D machine vision," *IEEE Computer Vision and Pattern Recognition,* pp.364-374, 1987.

115. Zhang, X., Y. Gao, and M.K.H. Leung, "3D face modeling using image warping in pose-invariant face recognition," in *Proc. 7th Inf. Conf. on Control, Automation, Robotics and Vision*, pp.497-501, 2002.

116. Ansari, A.N., and M. Abdel-Mottaleb, "Automatic facial feature extraction and 3D face modeling using two orthogonal views with application to 3D face recognition," *Pattern Recognition*, vol.38, pp.2549-2563, 2005.

117. Tang, L., and T.S. Huang, "Automatic construction of 3D human face models based on 2D images," in *Proc. Inf. Conf. on Image Processing*, vol.3, pp.467-470, 1996.

118. Forster, F., P. Rummel, M. Lang, and B. Radig, "The HISCORE camera: a real time three dimensional and color camera," in *Proc. IEEE Inf. Conf. on Image Processing,* vol.2, pp.598-601, 2001.

119. Malassiotis, S., and M.G. Strintzis, "Pose And Illumination Compensation For 3D Face Recognition," in *Proc. Inf. Conf. on Image Processing*, 2004.

120. Wong, A.K.C., P. Niu, and X. He, "Fast acquisition of dense depth data by a new structured light scheme," *Computer Vision and Image Understanding*, vol.98, no.3, pp.398-422, 2005.

121. Achermann, B., X. Jiang, and H. Bunke, “Face recognition using range images,” in *Proc. Inf. Conf. on Virtual Systems and MultiMedia*, pp.129-136, 1997.

122. Beumier, C., and M. Acheroy, “Automatic 3D face authentication,” *Image and Vision Computing,* vol.18, no.4, pp.315-321, 2000.

123. Messer, K., J. Matas, J. Kittler, J. Luettin, and G. Maitre, “XM2VTSDB: The extended M2VTS database,” in *Proc.* 2^{nd} *Inf. Conf. on Audio and Video-based Biometric Person Authentication*, 1999.

124. Tibbalds, A.D., “Three dimensional human face acquisition for recognition,” unpublished PhD Thesis, Signal Processing and Communications Laboratory, Dept. of Engineering, Univ. of Cambridge, 1998.

125. Chang, K., K. Bowyer, and P. Flynn, “Multi-modal 2D and 3D biometrics for face recognition,” in *Proc. IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures*, 2003.

126. Horn, E., and N. Kiryati, “Towards optimal structured light patterns,” *Image and Vision Computing*, vol.17, pp.87-97, 1999.

127. Batlle, J., E. Mouaddib, and J. Salvi, “Recent progress in coded structured light as a technique to solve the correspondence problem: a survey,” *Pattern Recognition,* vol.31, no.7, pp.963-982, 1998.

128. Papatheodorou, T., and D. Rueckert, “Evaluation of automatic 4D face recognition using surface and texture registration,” in *Proc. Inf. Conf. on Automatic Face and Gesture Recognition*, pp.321-326, 2004.

129. Fujimura, K., Y. Oue, and T. Terauchi, “Improved 3D Head Reconstruction System based on Combining Shape-From-Silhouette with Two-Stage Stereo Algorithm,” in *Proc. Inf. Conf. on Pattern Recognition,* 2004.

130. Gu, C. B. Yin, Y. Hu, and S. Cheng, “Resampling based method for pixel-wise

correspondence between 3D faces," in *Proc. Inf. Conf. on Information Technology: Coding and Computing*, vol.1, pp.614-619, 2004.

131. Weyrauch, B., J. Huang, B. Heisele, and V. Blanz, "Component-based Face Recognition with 3D Morphable Models," in *Proc. First IEEE Workshop on Face Processing in Video*, 2004.

132. Lee, J., B. Moghaddam, H. Pfister, and R. Machiraju, "Silhouette-Based 3D Face Shape Recovery," Technical Report, TR2003-081, *Graphics Interface*, June 2003.

133. Tsutsumi, S., S. Kikuchi, and M. Nakajima, "Face identification using a 3D gray-scale image-a method for lessening restrictions on facial directions," in *Proc.* 3^{rd} *IEEE Inf. Conf. on Automatic Face and Gesture Recognition*, pp.306-311, 1998.

134. Huber, D.F., "Automatic Three-dimensional Modeling from Reality," unpublished PhD thesis, CMU-RI-TR-02-35, The Robotics Institute, Carnegie Mellon University, 2002.

135. Phillips, P.J., P.J. Flynn, W.T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W.J. Worek, "Overview of the Face Recognition Grand Challenge," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol.1, pp.947-954, 2005.

136. Blais, F. "Review of 20 Years of Range Sensor Development," *Journal of Electronic Imaging,* vol.13, no.1, pp.231-240, 2004.

137. Forest, J., and J. Salvi, "A review of laser scanning three dimensional digitisers," *Intelligent Robots and Systems*, pp.73-78, 2002.

138. Haider, A.M., and T. Kaneko, "Realistic 3D head modeling from video captured images and CT data," in *Proc. IEEE Inf. Conf. on Information Technology Applications in Biomedicine*, pp.238-243, 2000.

139. Genex Technologies Inc., http://www.genextech.com/

140. BenAbdelkader, C., and P.A. Griffin, "Comparing and combining depth and texture cues for face recognition," *Image and Vision Computing*, vol.23, no.3, pp.339-352, 2005.

141. Geometrix Inc., http://www.geometrics.com/

142. A4 Vision Inc., http://www.a4vision.com/

143. Minolta Vivid 910 non-contact 3D laser scanner, http://www.minoltausa.com/vivid/

144. Cyberware Inc., http://www.cyberware.com/products/ scanners/ps.html

145. Vieira, M.B., L. Velho, A. Sa, and P.C. Carvalho, "A Camera-Projector System for Real-Time 3D Video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp.96, 2005.

146. Moreno, A.B., and Á. Sánchez, "GavabDB: A 3D Face Database," in *Proc.* 2nd *COST275 Workshop on Biometrics on the Internet*, Vigo (Spain), 2004.

147. University of York 3D Face Database, http://www-users.cs.york.ac.uk /˜tomh/3DFaceDatabase.html

148. The BJUT-3D Large-Scale Chinese Face Database, MISKL-TR-05-FMFR-001, Multimedia and Intelligent Software Technology Beijing Municipal Key Laboratory, Beijing University of Technology, 2005.

149. Bowyer, K., Chang K., and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *Computer Vision and Image Understanding*, vol.101, pp.1-15, 2006.

150. Lu, X., A.K. Jain, and D. Colbry, "Matching 2.5D Face Scans to 3D Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.28, no.1, 2006.

151. Tsalakanidou, F., S. Malassiotis, and M. Strinzis, “Integration of 2D and 3D images for enhanced face authentication,” in *Proc. Inf. Conf. on Automatic Face and Gesture Recognition,* pp.266-271, 2004.

152. Bronstein, A.M., M.M. Bronstein, and R. Kimmel, “Expression-invariant 3D face recognition,” in J. Kittler, M.S. Nixon (eds.) *Audio- and Video-Based Person Authentication*, pp.62-70, 2003.

153. Gordon, G. “Face recognition based on depth and curvature features,” in *SPIE Proc.: Geometric Methods in Computer Vision*, vol.1570, pp.234-247, 1991.

154. Hesher, C., A. Srivastava, and G. Erlebacher, “A novel technique for face recognition using range imaging,” in *Proc.* 7th *Int. Symposium on Signal Processing and Its Applications*, pp.201-204, 2003.

155. Moreno, A.B., Á. Sánchez, J.F. Vélez, and F.J. Díaz, “Face recognition using 3D surface-extracted descriptors,” in *Proc. Irish Machine Vision and Image Processing Conference,* 2003.

156. Srivastava, A., X. Liu, and C. Hesher, “Face recognition using optimal linear components of range images,” *Image and Vision Computing*, vol.24 pp.291-299, 2006.

157. Blanz, V., and T. Vetter, “Face Recognition Based on Fitting a 3D Morphable Model,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.25, no.9, pp.1063-1074, 2003.

158. Lee, Y., K. Park, J. Shim, and T. Yi, “3D face recognition using statistical multiple features for the local depth information,” in *Proc.* 16th *Inf. Conf. on Vision Interface*, 2003.

159. Wang, Y., C. Chua, and Y. Ho, “Facial feature detection and face recognition from 2D and 3D images,” *Pattern Recognition Letters*, vol.23, pp.1191-1202, 2002.

160. Achermann, B., and H. Bunke, "Classifying range images of human faces with Hausdorff distance," in *Proc. Inf. Conf. on Pattern Recognition*, pp.809-813, 2000.

161. İrfanoğlu, M.O., B. Gökberk, and L. Akarun, "3D Shape-Based Face Recognition Using Automatically Registered Facial Surfaces," in *Proc. Inf. Conf. on Pattern Recognition,* vol.4, pp.183-186, 2004.

162. Xu, C., Y. Wang, T. Tan, and L. Quan, "Automatic 3D face recognition combining global geometric features with local shape variation information," in *Proc.* 6^{th} *Inf. Conf. on Automatic Face and Gesture Recognition*, pp. 308-313, 2004.

163. Salah, A.A., N. Alyüz, and L. Akarun, "Alternative face models for 3D face registration," in *SPIE Conf. on Electronic Imaging, Vision Geometry*, San Jose, 2007.

164. Lao, S., Y. Sumi, M. Kawade, and F. Tomita, "3D template matching for pose invariant face recognition using 3D facial model built with iso-luminance line based stereo vision," in *Proc. Inf. Conf. on Pattern Recognition*, vol.2, pp.911-916, 2000.

165. Lu, X., D. Colbry, and A.K. Jain, "Three-Dimensional Model Based Face Recognition," in *Proc. Inf. Conf. on Pattern Recognition*, 2004.

166. Pan, G., Y. Wu, Z. Wu, and W. Liu, "3D Face recognition by profile and surface matching," in *Proc. of the Int. Joint Conference on Neural Networks*, vol.3, pp.2169-2174, 2003.

167. Lester, H., and S.R. Arridge, "A Survey of Hierarchical Non-Linear Medical Image Registration," *Pattern Recognition*, vol.32, pp.129-149, 1999.

168. Chen, Y., and G. Medioni, "Object Modeling by Registration of Multiple Range Images," *Image and Vision Computing*, vol.10, no.3, pp.145-155, 1992.

169. Chua, C.S., F. Han, and Y.K. Ho, "3D human face recognition using point sig-

nature," in *Proc. IEEE Inf. Conf. on Automatic Face and Gesture Recognition*, pp.233-238, 2000.

170. Garcia, E., J.L. Dugelay, and H. Delingette, "Low cost 3D face acquisition and modeling," in *Proc. Inf. Conf. on Information Technology: Coding and Computing*, pp.657-661, 2001.

171. Gökberk, B., M.O. İrfanoğlu, and L. Akarun, "3D Shape-Based Face Representation and Feature Extraction for Face Recognition," *Image and Vision Computing,* 2006.

172. Pan, G., and Z. Wu, "3D Face Recognition From Range Data," *International Journal of Image and Graphics*, vol.5, no.3, pp.573-593, 2005.

173. Rusinkiewicz, S., and M. Levoy, "Efficient Variants of the ICP Algorithm," in *Proc. of 3DIM01*, pp.145-152, 2001.

174. Matabosch, C., J. Salvi, X. Pinsach, and R. Garcia, "Surface registration from range image fusion," in *Proc. IEEE Inf. Conf. on Robotics and Automation*, pp.678-683, 2004.

175. Hüsken, M., M. Brauckmann, S. Gehlen, and C. von der Malsburg, "Strategies and Benefits of Fusion of 2D and 3D Face Recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

176. Bookstein, F.L., "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.11, pp.567-585, 1989.

177. Bookstein, F.L., *Morphometric Tools for Landmark Data: Geometry and Biology*, Cambridge University Press, 1991.

178. Lu, X., and A.K. Jain, "Deformation Analysis for 3D Face Matching," *Proc. 7^{th} IEEE Workshop on Applications of Computer Vision* (WACV'05), pp. 99-104,

Breckenridge, CO, 2005.

179. Hutton, T.J., B.F. Buxton, and P. Hammond, "Automated registration of 3d faces using dense surface models," in *Proc. British Machine Vision Conference*, 2003.

180. Mao, Z., P. Siebert, P. Cockshott, and A. Ayoub, "Constructing dense correspondences to analyze 3d facial change," in *Proc. 17th Int. Conf. on Pattern Recognition*, pp.144-148, 2004.

181. Tena, J.R., M. Hamouz, A. Hilton, and J. Illingworth, "A Validated Method for Dense Non-rigid 3D Face Registration," in *IEEE Int. Conf. on Video and Signal Based Surveillance*, p.81, 2006.

182. Tanaka, H., M. Ikeda, and H. Chiaki, "Curvature-based face surface recognition using spherical correlation," in *Proc. ICFG,* pp.372-377, 1998.

183. Abate, A., M. Nappi, S. Ricciardi, and G. Sabatino, "Fast 3D face recognition based on normal map," in *IEEE Int. Conf. Image Processing*, vol.2, pp.946-949, 2005.

184. Wang, Y., and C.-S. Chua, "Face recognition from 2D and 3D images using 3D Gabor filters," *Image and Vision Computing*, vol.23, no.11, pp.1018-1028, 2005.

185. Medioni, G., and R. Waupotitsch, "Face recognition and modeling in 3D," *IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures*, pp.232-233, 2003.

186. Chen, Q., and G. Medioni, "Building Human Face Models from Two Images," in *Proc. 2nd IEEE Workshop on Multimedia Signal Processing*, pp.117-122, 1998.

187. Xu, C., Y. Wang, T. Tan, and L. Quan, "A New Attempt to Face Recognition Using 3D Eigenfaces," in *Proc. ACCV*, pp.884-889, 2004.

188. Shen, J., W. Shen,and D. Shen, "On Geometric and Orthogonal Moments," *Inter-*

national Journal of Pattern Recognition and Artificial Intelligence, vol.14, no.7, pp.875-894, 2000.

189. Lao, S., Y. Sumi, M. Kawade, and F. Tomita, "Building 3D facial models and detecting face pose in 3D space," in *Proc.* 2nd *Inf. Conf. on 3-D Digital Imaging and Modeling*, pp.398-404, 1999.

190. Dutağacı, H., B. Sankur, and Y. Yemez, "3D face recognition by projection-based features," in *Proc. SPIE Conf. on Electronic Imaging: Security, Steganography, and Watermarking of Multimedia Contents*, 2006.

191. Passalis, G., I.A. Kakadiaris, T. Theoharis, G. Toderici, and N. Murtuza, "Evaluation of 3D Face Recognition in the presence of facial expressions: an Annotated Deformable Model approach," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

192. Cartoux, J.Y., J.T. LaPreste, and M. Richetin, "Face authentication or recognition by profile extraction from range images," in *Proc. of the Workshop on Interpretation of 3D Scenes*, pp.194-199, 1989.

193. Nagamine, T., T. Uemura, and I. Masuda, "3D facial image analysis for human identification," in *Proc. Inf. Conf. on Pattern Recognition*, pp.324-327, 1992.

194. Gökberk, B., A.A. Salah, and L. Akarun, "Rank-based Decision Fusion for 3D Shape-based Face Recognition," in T. Kanade, A. Jain, N.K. Ratha (eds.) *LNCS*, Vol.3546, pp.1019-1028, AVBPA 2005.

195. Blanz, V., S. Romdhani, and T. Vetter, "Face Identification across Different Poses and Illuminations with a 3D Morphable Model," in *Proc. Fifth IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2002.

196. Lee, M.W., and S. Ranganath, "Pose-invariant face recognition using a 3D deformable model," *Pattern Recognition*, vol.36, pp.1835-1846, 2003.

197. Zhao, W., and R. Chellappa, “SFS Based View Synthesis for Robust Face Recognition,” in *Proc. Inf. Conf. on Automatic Face and Gesture Recognition*, pp.285-292, 2000.

198. Zhang, L., and D. Samaras, “Face Recognition Under Variable Lighting using Harmonic Image Exemplars,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol.1, pp.19-25, 2003.

199. Tsalakanidou, F., D. Tzovaras, and M. Strinzis, “Use of depth and colour eigenfaces for face recognition,” *Pattern Recognition Letters,* vol.24, pp.1427-1435, 2003.

200. Kittler, J., M. Ballette, J. Czyz, F. Roli, and L. Vandendorpe, “Enhancing the performance of personal identity authentication systems by fusion of face verification experts,” in *Proc. IEEE Inf. Conf. on Multimedia and Expo*, vol.2, pp.581-584, 2002.

201. Beumier, C., and M. Acheroy, “Face verification from 3D and grey level cues,” *Pattern Recognition Letters,* vol.22, pp.1321-1329, 2001.

202. Stein, B.E., T.R. Stanford, M.T. Wallace, J.W. Vaughan, and W. Jiang, “Crossmodal spatial interactions in subcortical and cortical circuits,” in C. Spence, and J. Driver (eds.), *Crossmodal Space and Crossmodal Attention*, Oxford Univ. Press, 2004.

203. Salah, A.A., “Perceptual Information Fusion in Humans and Machines,” (in Turkish) appears in *Cognitive Neuroscience Forum*, 2007.

204. Gökberk, B., H. Dutağacı, L. Akarun, and B. Sankur, “Representation Plurality and Decision Level Fusion for 3D Face Recognition,” submitted to *IEEE Trans. Systems, Man, and Cybernetics.*

205. Moerland, P., “Mixture Models for Unsupervised and Supervised Learning,” PhD

thesis, Ecole Polytechnique Federale de Lausanne, Computer Science Department, 2000.

206. Ghahramani, Z., and G.E. Hinton, "The EM algorithm for mixtures of factor analyzers," Technical Report CRG-TR-96-1, University of Toronto, (revised), 1997.

207. McLachlan, G.J., and D. Peel, *Finite Mixture Models,* Wiley-Interscience, 2000.

208. Tipping, M., and C. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation,* vol.11(2), pp.443-482, 1999.

209. Verbeek, J.J., N. Vlassis, and B. Kröse, "Efficient Greedy Learning of Gaussian Mixture Models," *Neural Computation,* vol.15, pp.469-485, 2003.

210. Salah, A.A., and E. Alpaydın, "Incremental Mixtures of Factor Analyzers," in *Proc. Int. Conf. on Pattern Recognition*, vol.1, pp.276-279, 2004.

211. Salah, A.A., and E. Alpaydın, "Incremental Mixtures of Factor Analyzers," Technical Report, FBE/CMPE-01/2004-7, Institute of Graduate Studies in Science and Engineering, Dept. of Computer Engineering, Boğaziçi University, 2004.

212. Mardia K.V., J.T. Kent, and S.M. Bibby, *Multivariate Analysis*, Academic Press, 1979.

213. Agresti, A., and B. Caffo, "Measures of Relative Model Fit," *Computational Statistics & Data Analysis,* vol.39, pp.127-136, 2002.

214. Blake, C.L., and C.J. Mertz, *UCI repository of machine learning databases*, http://www.ics.uci.edu/∼mlearn/ MLRepository.html, UCI: Dept. of Info. and Comp. Sci., 1998.

215. The Olivetti Research Laboratory database of faces, http://www.cam-orl.co.uk/facedatabase.html, 1994.

216. MIT Media Lab Vistex database, http://www.white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html, 2002.

217. Brown, M.P.S. *et al.*, “Knowledge-based analysis of microarray gene expression data using support vector machines,” *Proc. National Academy of Sciences*, 97:262-267, 2000.

218. Kohonen, T., J. Hynninen, J. Kangas, J. Laaksonen, and K. Torkkola, *LVQ-PAK, The learning vector quantization program package*, Helsinki University of Technology, 1995.

219. Perkins, S., and J. Theiler, “Online Feature Selection using Grafting,” in *Proc.* 20^{th} *Int. Conf. on Machine Learning,* Washington, 2003.

220. Akaike, H., “A new look at statistical model identification,” *IEEE Trans. Automatic Control*, vol.19, no.6, pp.716-723, 1974.

221. Bozdoğan, H., “Model selection and Akaikes information criterion (AIC): the general theory and its analytical extensions,” *Psychometrika*, vol.52, no.3, pp.345-370, 1987.

222. Schwarz, G, “Estimating the dimension of a model,” *The Annals of Statistics,* vol.6, no.2, pp.461-464, 1978.

223. Barron, A., J. and Rissanen, “The minimum description length priciple in coding and modeling,” *IEEE Trans. Information Theory*, vol.44, pp.2743-2760, 1998.

224. Shi, L., and L. Xu, “Local Factor Analysis with Automatic Model Selection: A Comparative Study and Digits Recognition Application,” in S. Kollias *et al.* (eds.), *ICANN, Part II, LNCS 4132*, pp.260-269, 2006.

225. Çınar Akakın, H., A.A. Salah, L. Akarun, and B. Sankur, “2D/3D Facial Feature Extraction,” in *Proc. SPIE Conference on Electronic Imaging*, 2006.

226. Salah, A.A., and L. Akarun, "Gabor Factor Analysis for 2D+3D Facial Landmark Localization," (in Turkish) in *Proc. IEEE* 14th *Signal Processing and Communications Applications Conference*, 2006.

227. Salah, A.A., and L. Akarun, "3D Facial Feature Localization for Registration," in B. Günsel *et al.* (eds.), *LNCS,* vol. 4105/2006, Int. Workshop on Multimedia Content Representation, Classification and Security, pp. 338-345, 2006.

228. Salah, A.A., H. Çınar, L. Akarun, and B. Sankur, "Robust Facial Landmarking for Registration," *Annals of Telecommunications*, vol.62, no.1-2, pp.1608-1633, 2007.

229. Brunelli, R., and T. Poggio, "Face recognition: features versus templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.15, pp.1042-1052, 1993.

230. Arca, S., P. Campadelli, and R. Lanzarotti, "A face recognition system based on automatically determined facial fiducial points," *Pattern Recognition*, vol.39, no.3, pp.432-443, 2006.

231. Boehnen, C., and T. Russ, "A Fast Multi-Modal Approach to Facial Feature Detection," in *Proc.* 7th *IEEE Workshop on Applications of Computer Vision*, pp.135-142, 2005.

232. Colbry, D., X. Lu, A. Jain, and G. Stockman, "3D face feature extraction for recognition," Technical Report MSU-CSE-04-39, Computer Science and Engineering, Michigan State University, 2004.

233. Herpers, R., M. Michaelis, K.-H. Lichtenauer, and G. Sommer, "Edge and Keypoint Detection in Facial Regions," in *Proc.* 2nd *Int. Conf. on Automatic Face and Gesture Recognition*, pp.212-217, 1996.

234. Shih, F.Y., and C. Chuang, "Automatic Extraction of Head and Face Boundaries and Facial Features," *Information Sciences*, vol.158, pp.117-130, 2004.

235. Xu, C., T. Tan, Y. Wang, and L. Quan, "Combining local features for robust nose location in 3D facial data," *Pattern Recognition Letters*, vol.27, no.13 , pp.1487-1494, 2006.

236. Yacoob, Y., and L.S. Davis, "Labeling of human face components from range data," *CVGIP: Image Understanding*, vol.60, no.2, pp.168-178, 1994.

237. Colbry, D., G. Stockman, and A.K. Jain, "Detection of Anchor Points for 3D Face Verification," in *Proc. IEEE Workshop on Advanced 3D Imaging for Safety and Security*, 2005.

238. Yan, Y., and K. Challapali, "A system for the automatic extraction of 3-D facial feature points for face model calibration," in *Proc. Int. Conf. on Image Processing*, vol.2, pp.223-226, 2000.

239. Lu, X., and A.K. Jain, "Multimodal Facial Feature Extraction for Automatic 3D Face Recognition," Technical Report MSU-CSE-05-22, Michigan State University, 2005.

240. Chen, L., L. Zhang, H. Zhang, and M. Abdel-Mottaleb, "3D Shape Constraint for Facial Feature Localization Using Probabilistic-like Output," in Proc. 6^{th} IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2004.

241. Chen, L., L. Zhang, L. Zhu, M. Li, and H. Zhang, "A Novel Facial Feature Localization Method Using Probabilistic-like Output," in Asian Conference on Computer Vision, Korea, 2004.

242. Lai, J.H., P.C. Yuen, W.S. Chen, S. Lao, and M. Kawade, "Robust Facial Feature Point Detection under Nonlinear Illuminations," in *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp.168-174, 2001.

243. Ryu, Y.S., and S.Y. Oh, "Automatic Extraction of Eye and Mouth Fields from

a Face Image Using Eigenfeatures and Ensemble Networks," *Applied Intelligence*, vol.17, pp.171-185, 2002.

244. Sobottka, K., and I. Pitas, "A fully automatic approach to facial feature detection and tracking," in J. Bigün, G. Chollet, G. Borgefors (eds.), *Audio- and Video-based Biometric Person Authentication*, LNCS, vol.1206, pp.77-84, Springer Verlag, 1997.

245. Zhu, X., J. Fan, and A.K. Elmagarmid, "Towards facial feature extraction and verification for omni-face detection in video images," *Image Processing,* vol.2, pp.113-116, 2002.

246. Gargesha, M., and S. Panchanathan, "A hybrid technique for facial feature point detection," in 5th *IEEE Southwest Symposium on Image Analysis and Interpretation,* pp.134-138, 2002.

247. Ioannou, S., M. Wallace, K. Karpouzis, A. Raouzaiou, and S. Kollias, "Combination Of Multiple Extraction Algorithms in the Detection of Facial Features," in *IEEE Inf. Conf. on Image Processing*, vol.2, pp.378-381, 2005.

248. Liao, C.-T., Y.-K. Wu, and S.-H. Lai, "Locating facial feature points using support vector machines," in *Proc.* 9th *Int. Workshop on Cellular Neural Networks and Their Applications*, pp.296-299, 2005.

249. Gündüz, A., and H. Krim, "Facial Feature Extraction Using Topological Methods," in *IEEE Int. Conf. on Image Processing*, vol.1, pp.673-676, Barcelona, Spain, 2003.

250. Senaratne, R., and S. Halgamuge, "Optimised Landmark Model Matching for Face Recognition," *Proc.* 7th *Int. Conf. on Automatic Face and Gesture Recognition*, pp.120-125, 2006.

251. Beumer, G.M., Q. Tao, A.M. Bazen, and R.N.J. Veldhuis, "A Landmark Paper

in Face Recognition," in *Proc.* 7th *Inf. Conf. on on Automatic Face and Gesture Recognition*, pp.73-78, 2006.

252. Cristinacce, D., T. Cootes, and I. Scott, "A multi-stage approach to facial feature detection," in *Proc. British Machine Vision Conference,* pp.231-240, 2004.

253. Viola, P., and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol.1, pp.511-518, 2001.

254. Smeraldi, F., and J. Bigün, "Retinal Vision applied to Facial Features Detection and Face Authentication," *Pattern Recognition Letters*, vol.23, pp.463-475, 2002.

255. Feris, R.S., J. Gemmell, K. Toyama, and V. Krüger, "Hierarchical Wavelet Networks for Facial Feature Localization," in 5th *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 125-130, 2002.

256. Shakunaga, T., K. Ogawa, and S. Oki, "Integration of Eigentemplate and Structure Matching for Automatic Facial Feature Detection," in 3rd *Int. Conf. on Automatic Face and Gesture Recognition*, pp.94-98, 1998.

257. Zobel, M., A. Gebhard, D. Paulus, J. Denzler, and H. Niemann, "Robust Facial Feature Localization by Coupled Features," in 4th *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Grenoble, France, 2000.

258. Gourier, N., D. Hall, and J.L. Crowley, "Facial Features Detection Robust to Pose, Illumination and Identity," in *IEEE Int. Conf. on Systems, Man and Cybernetics*, vol.1, no.10-13, pp.617-622, 2004.

259. Antonini, G., V. Popovici, and J.P. Thiran, "Independent Component Analysis and Support Vector Machine for Face Feature Extraction," in *Proc.* 4th *Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, pp.111-118, 2003.

260. Li, P., B.D. Corner, and S. Paquette, "Automatic landmark extraction from three-

dimensional head scan data," in *Proc. of SPIE*, vol.4661, pp.169-176, 2002.

261. Corner, B.D., P. Li , B. Beecher,and S. Paquette, "Two Methods for Locating Feducial Points on Three-Dimensional Scans of the Human Face," in *Proc. of SPIE*, vol.4661, pp.157-168, 2002.

262. Conde, C., A. Serrano, L.J. Rodríguez-Aragón, and E. Cabello, "3D Facial Normalization with Spin Images and Influence of Range Data Calculation over Face Verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

263. Sahbi, H., and N. Boujema, "Robust Face Recognition Using Dynamic Space Warping," in M. Tistarelli, J. Bigun, A.K. Jain (eds.), *Biometric Authentication*, LNCS 2359, pp.121-132, Springer-Verlag, Berlin, Heidelberg, 2002.

264. Yang, M., N. Ahuja, and D. Kriegman, "Face Detection Using Mixtures of Linear Subspaces," in *Proc.* 4th *Int. Conf on Automatic Face and Gesture Recognition,* pp.70-76, 2000.

265. Senior, A.W., "Face and feature finding for a face recognition system," in *Proc.* 2nd *Int. Conf. on Audio- and Video-based Biometric Person Authentication*, pp.154-159, 1999.

266. Heslenfeld D.J., J.L. Kenemans, A. Kok, and P.C. Molenaar, "Feature processing and attention in the human visual system: an overview," *Biological Psychology*, vol.45, pp.183-215, 1997.

267. Park, S.J., J.K. Shin, and M. Lee, "Biologically Inspired Saliency Map Model for Bottom-up Visual Attention," *BMCV, LNCS*, vol.2525, pp.418-426, Springer-Verlag, Heidelberg, 2002.

268. Ban, S.W., and M. Lee, "Biologically Motivated Visual Selective Attention for Face Localization," in L. Paletta *et al.* (eds.) WAPCV, LNCS 3368, pp.196-205,

Springer-Verlag, 2005.

269. Herpers, R., and G. Sommer, "An attentive processing strategy for the analysis of facial features," in H. Wechsler *et al.* (eds.), *Face Recognition: From Theory to Applications*, pp.457-468, Springer, ASI Series, 1998.

270. Hotta, K., T. Mishima, T. Kurita, and S. Umeyama, "Face Matching through Information Theoretical Attention Points and Its Applications to Face Detection and Classification," in *Proc. 4th IEEE Inf. Conf. on Automatic Face and Gesture Recognition* pp.34-39, 2000.

271. Liu, D.H, K.M. Lam, and L.S. Shen, "Optimal sampling of Gabor features for face recognition," *Pattern Recognition Letters*, vol.25, 267-276, 2004.

272. Serre, T., M. Riesenhuber, J. Louie, and T. Poggio, "On the role of object-specific features for real world object recognition in biological vision," BMCV, LNCS, vol.2525, pp.387-397, Springer-Verlag, Heidelberg, 2002.

273. Siagian, C., and L. Itti, "Biologically-inspired face detection: non-brute-force-search approach," in *Proc. 1st IEEE Workshop on Face Processing in Video*, 2004.

274. Craw, I., D. Tock, and A. Bennet, "Finding Face Features," in *Proc. European Conf. Computer Vision*, pp.92-96, 1992.

275. Frintrop, S., A. Nüchter, H. Surmann, and J. Hertzberg, "Saliency-based Object Recognition in 3D Data," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp.2167-2172, Sendai, Japan, 2004.

276. Frintrop, S., E. Rome, A. Nüchter, and H. Surmann, "A bimodal laser-based attention system," *Computer Vision and Image Understanding*, vol.100, pp.124-151, 2005.

277. Huang, L., A. Shimizu, and H. Kobatake, "Robust face detection using Gabor filter features," *Pattern Recognition Letters*, vol.26, no.11, pp.1641-1649, 2005.

278. Lim, R., and Reinders, M.J.T., "Facial landmarks localization based on fuzzy and Gabor wavelet graph matching," in 10th *IEEE Int. Conf. on Fuzzy Systems*, vol.3, pp.683-686, 2001.

279. Vukadinovic, D., and M. Pantic, "Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers," in *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics*, vol.2, pp.1692-1698, 2005.

280. Lades, M., J. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Würtz, and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Trans. Computers*, vol.42, 1993.

281. Basri, R., and D.W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.25, no.2, pp.218-233, 2003.

282. Ramamoorthi, R., "Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian object," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.24, no.10, 2002.

283. Lee, K.-C., J. Ho, and D. Kriegman, "Nine Points of Light: Acquiring Subspaces for Face Recognition under Variable Lighting," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol.I, pp.519-526, 2001.

284. Smith, W.A.P., and E.R. Hancock, "Estimating the Albedo Map of a Face from a Single Image," in *Proc. IEEE Int. Conf. on Image Processing*, vol.3, pp.780-783, 2005.

285. Zhang, L., and D. Samaras, "Pose Invariant Face Recognition under Arbitrary Unknown Lighting using Spherical Harmonics," in D. Maltoni, A.K. Jain (eds.) *Proc. ECCV 2004 Int. Workshop on Biometric Authentication*, LNCS 3087, pp.10-23, 2004.

286. Burl, C., and P. Perona, "Recognition of Planar Object Classes," in *Proc. IEEE*

Conf. Computer Vision and Pattern Recognition, pp.223-230, San Francisco, USA, 1996.

287. Wiskott, L., J.-M Fellous, N. Krüger, and C. von der Malsburg, “Face recognition by elastic bunch graph matching,” in L.C. Jain *et al.* (eds.)*Intelligent Biometric Techniques in Fingerprint and Face Recognition*, pp.355-396, CRC Press, 1999.

288. Xue, Z., S.Z. Li, and E.K. Teoh, “Bayesian Shape Model for Facial Feature Extraction and Recognition,” *Pattern Recognition*, vol.36, pp.2819-2833, 2003.

289. Kuncheva, L.I., *Combining Pattern Classifiers: Methods and Algorithms,* Wiley, NJ, New Jersey, 2004.

290. Figueiredo, M.A.T., and A.K. Jain, “Unsupervised learning of finite mixture models,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol.24, pp.381-396, 2002.

291. Hamouz, M., J. Kittler, J.-K. Kamarainen, P. Paalanen, H. Kälviäinen, and J. Matas, “Feature-Based Affine-Invariant Localization of Faces,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.27, no.9, pp.1490-1495, 2005.

292. Bailly-Baillière, E., S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, “The BANCA database and evaluation protocol,” in*Proc. Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, pp.625-638, Springer-Verlag, 2003.

293. Vetter, T., and T. Poggio, “Linear object classes and image synthesis from a single example image,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.19, no.7, pp.733-742, 1997.

294. Goodall, C., “Procrustes methods in the statistical analysis of shape,” *Journal of the Royal Statistical Society B,* vol.53, no.2, pp.285-339, 1991.

295. Gower, J.C., “Generalized Procrustes Analysis,” *Psychometrika*, vol.40, no.1,

pp.33-51, 1975.

296. Rohlf, F.J., and D. Slice, "Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks," *Systematic Zoology,* vol.39, no.1., pp.40-59, 1990.

297. Shi, J., A. Samal, and D. Marx, "How effective are landmarks and their geometry for face recognition?," *Computer Vision and Image Understanding*, vol.102, no.2, pp.117-133, 2006.

298. Yan, P., and K.W. Bowyer, "A fast algorithm for ICP-based 3D shape biometrics," *Computer Vision and Image Understanding,* in press, 2007.

299. Furl, N., P.J. Phillips, and A.J. O'Toole, "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis," *Cognitive Science*, vol.26, pp.797-815, 2002.

300. Goldstone, R.L., "Do we all look alike to computers?," *Trends in Cognitive Sciences*, vol.7, no.2, pp.55-57, 2003.

301. Watson, D.F., *Contouring: A guide to the analysis and display of spacial data*, Pergamon, 1994.

302. Ikeuchi, K., "Recognition of 3-D Objects Using the Extended Gaussian Image," in *Proc. of Seventh IJCAI*, pp.595-600, 1981.

303. Kang, S.B., and K. Ikeuchi, "3-D Object Pose Determination Using Complex EGI," Technical Report, CMU-RI-TR-90-18, Carnegie Mellon University, 1990.

304. Gu, X., S. Gortler, and H. Hoppe, "Geometry images," in *Proc. ACM SIGGRAPH*, pp.355-361, 2002.

305. Huttenlocher, D.P., G.A. Klanderman, and W.J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.15, no.9, pp.850-863, 1993.

306. Chua, C.S., and R. Jarvis, “Point signature: a new representation for 3D object recognition,” *International Journal of Computer Vision*, vol.25, no.1, pp.63-85, 1997.

307. O’Toole, A.J., T. Vetter, and V. Blanz, “Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: an application of three-dimensional morphing,” *Vision Research*, vol.39, pp.3145-3155, 1999.

308. Catmull, E., “A subdivision algorithm for computer display of curved surfaces,” Ph.D. Thesis, Department of Computer Science, University of Utah, Salt Lake City, Utah, USA, 1974.