3D Face Recognition Using Rigid and Non-Rigid Surface Registration

Theodore Papatheodorou

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of the

University of London

July 2006

Department of Computing

Imperial College London

Abstract

3D Face Recognition is one of the most active biometric research areas and its applications range from security, content tagging, facial reconstruction to human-computer interaction. One of the challenges in face recognition in general, is the fact that the face is a complex three-dimensional structure and its appearance is affected by factors such as the viewing angle, the illumination and facial expressions. In this work 3D face recognition techniques are presented using surface registration in an effort to minimize the influence of these factors on face recognition.

Initially the facial surfaces are registered to each other using a variant of the Iterative Closest Point (ICP) algorithm and the 3D distance between them is used as a similarity metric for performing face recognition. 2D intensity values of the surface texture are also used in a combined metric and their effect under various poses and facial expressions is analyzed.

Furthermore, we describe a point registration technique using manually landmarked data. Free-form deformation is used to establish dense point correspondence between all faces and a base mesh. Once correspondence is established principal component analysis is used to generate face models for face recognition. This technique is compared with an ICP-based technique in which dense correspondence is established after the automatic rigid surface registration of the faces. The evaluation and comparison of the two techniques is done by comparing the face recognition results as well as the compactness, specificity and generalization ability of the two models.

Additionally, we examine the use of surface free-form registration for establishing more uniform correspondences between the faces.

In the final chapter we examine which parts of the face contribute the most to its recognition. We implement an eigenfeatures approach for comparing manually segmented anatomical components of the face. Moreover, we propose a method for automatically optimizing the model using information about the variability within the population.

Acknowledgements

Naturally, I want to thank my supervisor, Daniel Rueckert, who always finds time to discuss with his students and assist them (and even debugs their code). He has provided enormous help over the past four years. Furthermore, I would also like to thank Friedericke, Henriette and Julia, who had to do without him when he was devoting his days to correcting my thesis.

This thesis has been completed also due to the support of my family back in Greece who would encourage and keep me sane when I felt that there would be no end to it. They told me to “start with an A4 sheet of paper” and I did just that.

I would also like to thank various Imperial College employees who have been extremely patient with me over the past months while I was writing up. The cleaners, for not vacuuming in the morning while I was sleeping under my office desk, the security men for chatting with me during their patrols in the middle of the night and the catering staff for providing me with extra-large portions of food.

Finally, I would also like to give $\Theta o\lambda\eta\lambda\alpha\kappa\iota$ a slap on the back. I tested his limits for the past seven months... and he did not break.

Tierra y libertad!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The survival of an individual in a socially complex world depends greatly on his/her ability to read information about the age, sex, race, identity and current psychological state of another person based on that person's face. The face is controlled by multiple muscles, which are responsible for a wide range of facial expressions and movements. By manipulating these facial muscles (Figure 1.1) one is able to communicate a vast array of emotions that facilitate cooperation with other individuals. Furthermore, given our weak sense of smell as a species, our ability to identify people from their face has proven to be a particularly useful skill. Friend or foe can be distinguished with remarkable robustness without conscious effort.

Face recognition research using automatic or semiautomatic techniques was started in the 1960s and especially in the last two decades, with the wide availability of powerful computers, it has received significant attention from the research community. Various conferences dealing with face recognition emerged during the 1990s (AVBPA, AFGR) and several databases and testing protocols were established (FERET, FRVT, XM2VTS etc.) [228]. One of the reasons for this research interest is the wide range of possible applications for face recognition systems.

There are currently automatic systems that perform well when the face images are captured under uniform and controlled conditions. For a system, however, to be a viable solution it needs to meet many challenges (see Section 1.3).

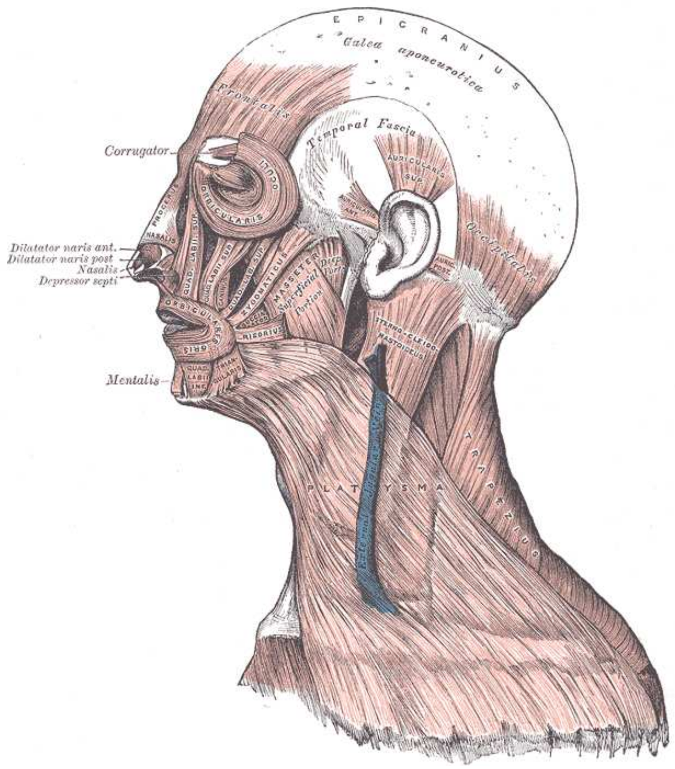


Figure 1.1: The facial muscles (from [77]).

There are various alternative biometric techniques that perform very well today, like fingerprint analysis and iris scans, but these methods require the participants to cooperate and follow a relatively strict data acquisition protocol. In a face recognition scenario participants are not necessarily required to cooperate or even be aware of being scanned and identified, which makes face recognition a less intrusive and potentially more effective identification technique. Finally, the public's perception of the face as a biometric modality is more positive compared to the other modalities [90].

1.1 Motivation

Correct identification of an individual based on their face would make many aspects of our life a lot simpler. First of all, in terms of the need to identify oneself, it would minimize the number of information one must carry or remember for one's day-to-day survival in the modern world. For example, it might be possible in the future to move in and out of a country without requiring a passport and voting might not require registration. Furthermore, in terms of protecting information, logging on to a network or a computer would be automated while access to a database or one's personal files would not require a password. When it comes to law enforcement and security applications, trained policemen would not have to go through thousands of hours of CCTV footage in order to identify suspects as a real-time identification system processing footage from various locations would be able to spot wanted persons and notify police of their whereabouts. Finally, the use of face recognition technology could be used for content tagging. Some online services already offer to tag member's pictures based on the identity of the subject portrayed using face recognition technology [166].

Progress made in face recognition would also provide important techniques for other disciplines and applications. Some statistical face models can already identify certain genetic syndromes which subtly affect the face [85]. In addition a model that can simulate the aging of a person [147] can be useful in the search for people who have been missing for many years. Finally, research in this area would also lead to advances in facial expression recognition which could have profound effects on human-computer-interaction. It is for these reasons that face recognition is such an inter-disciplinary research area with scientists from various fields contributing to the literature, from psychophysics and psychology, to mathematics and computer science.

Another reason for the growing interest in face recognition has been the emergence of affordable hardware, such as digital photography and video, which have made the goals of such research more feasible. Affordable systems that capture the 3D geometry of a surface, like the one used in our study and presented later in the thesis, make 3D face recognition a “tolerable” research interest, which combined with cheap computing power makes the intense calculations involved in surface processing realistic. Finally, the field of face recognition has inherited a wealth of algorithmic tools that had been traditionally developed for other disciplines, such as medical imaging, something which has allowed research to move forward relatively fast.

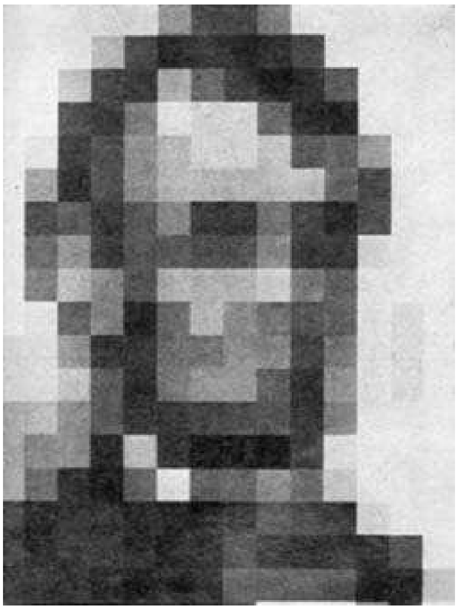


Figure 1.2: Exploring the limits of the human face recognition ability. Harmon and Julesz [86] used images like this to establish the amount of information that is necessary for successful face recognition.

1.2 Face recognition by the brain

The human brain is extremely good at classifying faces by identity even under adverse conditions such as low resolution or low illumination. Harmon and Julesz [86] showed that participants were able to identify people from images even when the latter contain only 16×16 pixels (Figure 1.2). The group of images participants had to identify, however, were of familiar faces which later research suggested is a significant experimental parameter, because the brain is performing particularly well with known faces [36]. Despite its great skill at recognizing faces one of the brain's drawbacks is its limitation as to the total number of faces it can recall and the even smaller number of familiar faces that it can recognize under extreme conditions. Furthermore, given that the brain is made up of interconnected systems, it is natural that one system can interfere with the functioning of another. A typical example is when a witness to a crime has particular difficulty recalling the characteristics of the perpetrator's face. The stressful situation affects the witness' ability to "store" the face in his memory.

Given the elusive nature of the "mind", scientists often try to draw conclusions about its underlying mechanisms by studying cases where the brain fails, often surprisingly, to perform a task. This kind of phenomena are often examined in depth in order to settle some of the most prominent debates in human face recognition research.

1.2.1 Face recognition as a dedicated brain process

Some evidence shows that face perception is a dedicated process in the brain, with a specialized region committed to it [16, 55, 66, 67]. The evidence for this comes from many sources. First of all, humans can remember faces much more easily than other objects [55]. The human brain can distinguish from memory hundreds of faces from each other, which it is unable to do for the vast majority of object classes. Furthermore, a rare disorder, prosopagnosia, caused by injury indicates that face recognition might be a special higher level process in the brain. Patients with prosopagnosia can not recognize familiar people from their faces, but have no problem spotting a face when asked to locate one in an image. They are able to describe the individual characteristics of a face and of other objects and suffer from no other profound agnosia, yet are unable to “unify” the information and make the identification [177].

Some researchers [102] have shown that newborns prefer images with face-like patterns, suggesting that humans are “pre-wired” for providing specialized treatment to faces, but their findings are not universally accepted [185]. Another evidence that faces receive specialized processing in the brain comes again from Burton and Wilson [36]. They conducted an experiment to show the effects of familiarity on identifying faces from poor quality CCTV video. After demonstrating that the group which outperformed the others was the one that was familiar with the subjects in the video, they occluded the faces of the people portrayed in order to see how much information comes from body posture and gait. In line with what was predicted the recognition rates were reduced dramatically. Despite the face being only a small part of the body, the subjects were able to extract significantly more cues as to the identity of the people in the video by seeing the face, which implies that the latter might be processed differently.

It is important to note that despite the aforementioned research suggesting that face recognition is a dedicated process, some researchers have shown that it is possible that other classes of objects might also be processed by the same neural circuits. For example Gauthier *et al.* [68] discuss an alternative hypothesis according to which face recognition is performed by mechanisms specialized for processing any object class in which an individual has expertise. Functional neuroimaging methods have been employed to see the response of face-selective areas to other stimuli in which the subject is an expert. Indeed these areas respond similarly suggesting that they might be specializing in more than just faces.

On the other hand, Xu *et al.* [220] recently reported that certain magnetoencephalographic responses do not occur for other classes of objects in which the participants had expertise. In particular, the authors study the effects on the face-selective M170, a component that occurs $170ms$ after the stimulus has been displayed and has been associated with the identification of individual faces. Other objects, such as cars, in which the participants had expertise, did not elicit a higher M170 response compared to control subjects. In a parallel experiment the M170 response was correlated with successful face identification but not with successful car identification by the car experts, prompting the authors to conclude that the early face processing mechanisms associated with the M170 are involved in face identification and not in the identification of just any object of expertise. Some researchers proposed a distributed neural system in which different face processing mechanisms are found in three different core brain areas, as well as four other areas that extend the system to related tasks [87]. In conclusion, the anatomical seat of face recognition circuitry, or whether such a specialized system even exists, is a matter of ongoing debate, but the fact that faces are processed differently than most physical objects is widely accepted.

1.2.2 Holistic versus feature-based processing

Researchers have also tried to analyze the way the brain encodes faces for identification. It has been suggested that both holistic and feature information is crucial for recognition [30, 32]. It has been shown, for example, that dependence on global descriptors is reduced if dominant local features are present and vice versa.

The importance of the global configuration of facial features was demonstrated in Yin [224] where an inverted face was shown to be hard to recognize despite the size and shape of individual features remaining intact. This was later on famously demonstrated by the Thatcher illusion [200]. This sensitivity to orientation is not displayed when recognizing other mono-oriented objects such as airplanes or houses and a significant amount of research has focused on the reasons behind this.

Further insights into how the brain encodes a face is provided by studying facial composites created by police based on witnesses accounts. In most cases the witness has to select the most appropriate facial features from a set, which, as is generally acknowledged, leads to poor images. The individual features might be accurately portrayed but the faces have only a “generic likeness” to the actual person they depict. The “global signature” of the face is difficult to capture with a feature-based approach. The performance of such systems is often so poor that it has prompted police to declare that the image demonstrates not what the suspect looks like but what the suspect does *not* look like. In other words it should be used not to incriminate people based on their likeness but to exonerate them based on their dissimilarity. Finally, Young *et al.* [225] and Robins and McKone [168] have shown that even though the top and bottom half of two faces might be enough for identifying the individual, combining the two makes the task significantly more difficult (Figure 1.3). It was suggested, in this case, that the holistic processing of the face affects the processing of the features on the two halves.

Trying to identify the components of the face that most assist in identification, Shepherd *et al.* [184] have shown that the upper part of the face is more useful for face recognition than the bottom part. Taking a step further, Sadr *et al.* [178] showed that sometimes

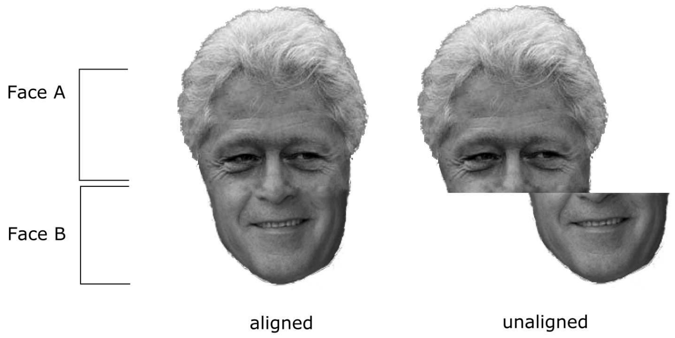


Figure 1.3: Subjects found it much harder to identify the people depicted in the top and bottom half in the aligned images (left) than in the unaligned (right). This suggests that holistic processing of the face disrupts on some level the processing based on the available features of the two halves. Face A is Bill Clinton and Face B is George W. Bush.

individual features such as eyes can provide enough cues for identification. Sinha and Poggio [187] split the face to its internal features (mouth, nose, eyes etc.) and its external ones (hairline, facial shape etc.) and showed that both contribute to our ability to recognize faces. When only the internal features are used the recognition rates are significantly worse than when the whole face is used. Furthermore, when the order of these internal features is rearranged the recognition rate drops further. Other studies have reached the same conclusions with regards to internal versus external features [30]. Schyns and Oliva [146] developed a technique where they superimpose the low spatial frequency components of an image on the high spatial frequency ones of another (coarse scale versus fine details). Faces are produced which if viewed from a close distance look like one person but if viewed from a greater distance another person is identified. In the local (internal) features domain, it seems that some features are more important than others. For example, the hair, eyebrows, mouth and face outline seem to be particularly instrumental for recognition [184, 30, 178].

1.2.3 On computational parallelisms

An interesting observation robustly demonstrated in the literature is that not all faces are equally easy to recognize. Distinctive faces (big ears, hooked nose etc.) are easier to recognize than faces with few distinguishing characteristics [122]. Valentine [208] suggests

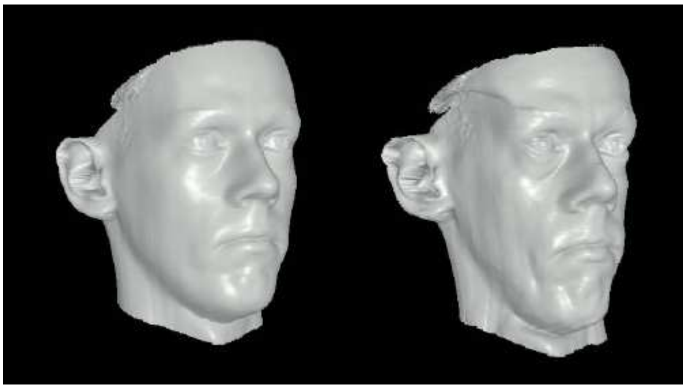


Figure 1.4: O'Toole *et al.* demonstrated that caricatures of the human face (right) are more easily identified than the original representation of the face (left) (from [147]).

that this might be because humans compare faces to a representation for an average face (prototype) stored in their brain. In other words humans store relative and not absolute information about the face. Under this interpretation faces in the extremes of the facespace distribution are easier to recognize because they are not surrounded by as many other possible candidates as a relatively average face would. Computational models such as the ones that use *principal component analysis* (PCA) [206], discussed later in the thesis, work along the same principles as the prototype theory for face recognition. A facespace of reduced dimensionality is created and all faces become samples in that facespace. Computational techniques based on creating such a facespace have been shown to validate the predictions made by the hypothetical mental facespace found in Valentine [208] and referred to above. For example, exaggerating the facial characteristics by manipulating the facial parameters in such a facespace using a model as in O'Toole *et al.* [148] confirmed that a caricatured face is more easily recognized by the brain than a face which depicts the exact proportions of the facial features. At the same time, faces generated between the face in question and the average, called anti-caricatures, have proven more difficult to identify than the original face, demonstrating that indeed distinctiveness of a face affects its recognition. If one assumes that the average face stored in the brain is tuned to a particular population then a person in Africa will not have the same facial prototype as a European. This might explain why people often report difficulty distinguishing persons of a different race from each other.

The study of the way the brain performs object recognition can provide some insight into how a machine-driven face recognition system should work. Conversely, psychologists must not draw conclusions without considering some of the computational possibilities. This does not mean however that there must necessarily be a link between the biological workings of the human visual system and an engineer's implementation. Nevertheless it is important to consider some of the challenges for face recognition which an agent, whether human or machine must overcome in order to successfully identify a face.

1.3 Challenges for face recognition

The face is a three-dimensional object which contains shape as well as texture (pigment) information. Unsurprisingly, systems that use these types of information, whether it is the human brain or a machine, are affected. Broadly speaking, the obstacles that a face recognition system must overcome are differences in appearance due to variations in illumination, viewing angle, facial expressions, occlusion and changes due to time.

Using 2D images for face recognition, pixel intensities represent all the information that is available and therefore, any algorithm needs to cope with variation due to illumination explicitly. Figure 1.5 shows such an example. The human brain seems also to be affected by illumination in performing face recognition tasks [92]. Johnston *et al.* [103] report that familiar faces were more difficult to recognize when lit from below than when lit from above and Hill *et al.* [91] demonstrate that matching facial surfaces to an identity is more difficult when the surfaces were lit from different directions. Furthermore, Bruce and Langton [33] reported that showing the images in photographic negatives had a detrimental effect on the identification of familiar faces and Liu *et al.* [125] demonstrated later on that the effects of negation are linked to the direction of lighting in the non-negated images. One explanation for these findings is that dramatic illumination or pigmentation changes interfere with the shape-from-shading processes involved in constructing representations of faces. If the brain reconstructs 3D shape from 2D images, it remains a question why face recognition by humans remains viewpoint-dependent to the

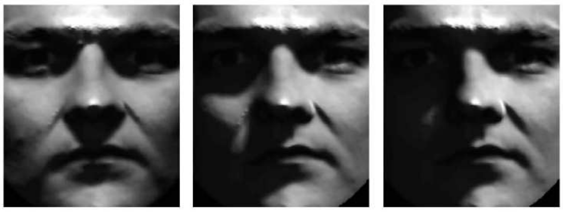


Figure 1.5: The effects of illumination, a challenge for face recognition systems.

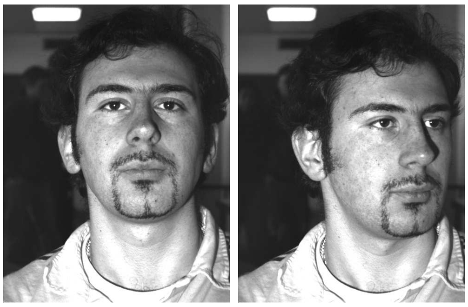


Figure 1.6: The effects of pose.

extend that it is.

The difference between two images of the same subject photographed from different angles is greater than the differences between two images of different subjects photographed from the same angle (Figure 1.6). Bruce [29] found recognition rates for unfamiliar faces dropping significantly when there was a difference in viewpoint between the training and test set. More recently, however, there has been debate about whether object recognition is viewpoint-dependent or not [197]. Generally speaking the brain is good at generalizing from one viewpoint to another as long as the change in angle is not extreme. For example, matching a profile viewpoint to a frontal image is poor, though matching of a three-quarter view to a frontal remains very good [92]. There have been suggestions that the brain might be storing a view-specific prototype abstraction of a face in order to deal with varying views [31]. Interpolation-based models [161], for example, support the idea that the brain identifies faces across different views by interpolating to the closest previously seen view of the face.

The challenges involving the face do not include only viewpoint variation which affects any rigid body. The face is a dynamic non-rigid structure that changes shape due to

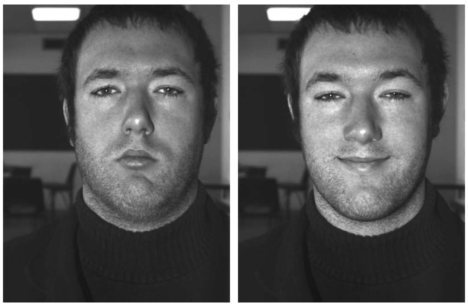


Figure 1.7: The effects of emotional expressions on the facial surface.

the muscles pulling the soft tissue and bones. Figure 1.7 shows the face changing due to the subject smiling. The modeling of dynamic objects introduces a type of problem that is different from the aforementioned ones as the face manifests itself as a nonrigid object. Neurophysiological studies have suggested that facial expression recognition happens in parallel to face identification [30]. Some case studies in prosopagnostic patients show that they are able to recognize expressions even though identifying the actor remains a near-impossible task. Similarly, patients who suffer from *organic brain syndrome* perform very poorly in analyzing expressions but have no problems in performing face recognition.

The shape of the face also changes due to aging and lifestyle choices people make. The skin becomes less elastic and more loose with age, the lip and hair-line often recedes, the skin color changes, people gain or lose weight, grow a beard, change hairstyle etc. Figure 1.8 shows three people over a time course of many years. The pictures were taken at two-year intervals from 1978 until 2004.

Finally, occlusion is a problem related to some of the above challenges and it involves cases when parts of the face are hidden from view. This can be for a number of reasons such as parts of the face itself (i.e. nose) hiding other parts when the image is taken from certain angles, or because the subject grew a beard, is wearing glasses or a hat.

The hypothesis throughout this work is that using 3D data, rather than 2D intensity images, would provide a lot more redundancy in order to deal with many of the traditional challenges for face recognition more effectively.

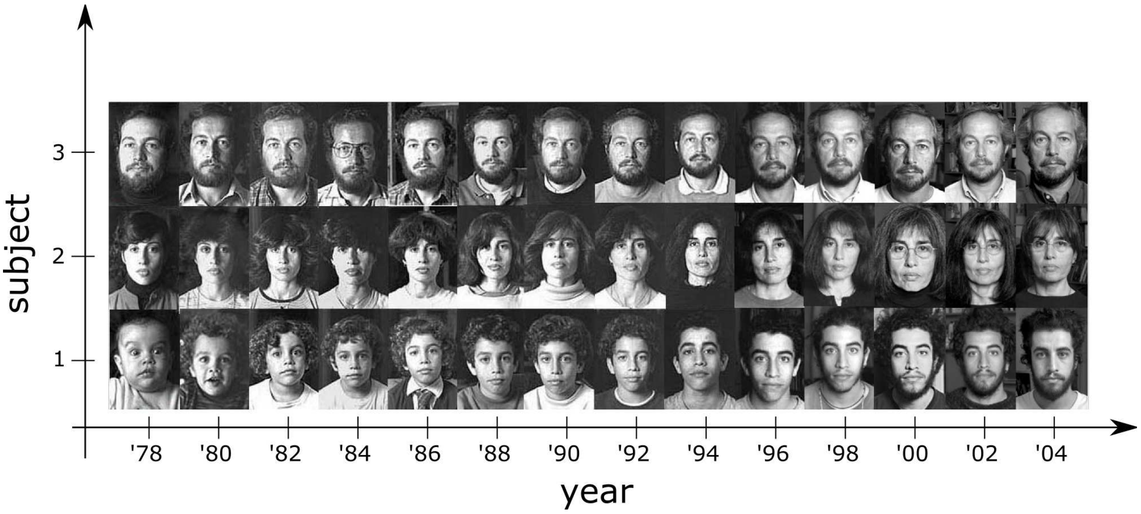


Figure 1.8: The effects of time on the face. The images were collected at two-year intervals from 1978 until 2004 (adapted from [72]).

1.4 Contributions

In this work we propose several novel 3D face recognition algorithms.

The first algorithm is a geometric technique where the average distance between corresponding surface points on the two faces is used as a similarity measure. To achieve that, an automatic rigid surface registration is used to align 3D faces to each other and compensate for pose differences. The technique achieves good results for frontal and non-frontal faces as well as faces with facial expressions. Furthermore, the relationship between pose, expression and illumination is investigated by fusing 3D and 2D information in a combined metric that results in further improvements in recognition under certain conditions.

In contrast to the geometric nature of the first algorithm, the second algorithm is based on the construction of a 3D statistical face model for recognition. For this purpose we propose several techniques for model construction: The first model-construction technique is employing landmarks to align features of the face together using non-rigid point-based registration. This method is compared and contrasted against another technique, which employs an automatic rigid surface registration using the ICP algorithm to align the facial surfaces to each other. Both techniques are shown to be strong classifiers under various recognition benchmarks. We also propose a modelling technique, which extends the rigid surface registration to a non-rigid surface registration in order to improve the point-correspondence across faces. Finally, a fourth technique is proposed in which a uniform synthetic surface (a sphere) is used to resample the surfaces to create point-correspondence across surfaces with even less noise. We explore the performance of the resulting models for face recognition. We also investigate the generic quality of the resulting models in terms of compactness, specificity and generalization ability.

We also propose a novel 3D eigenfeatures technique which involves the semi-automatic segmentation of faces. We investigate which parts of the face contribute most to classification. Furthermore, we combine different regions of the face into a single model which yields better recognition rates than the standard statistical face model. The scores from the recognition of individual features of the face are combined by classifier fusion leading to improved results. Finally, we examine the between- and within-class variability of 3D face databases. We show how statistical face models can be improved by reducing the amount of within variability in the model.

1.5 Structure of the thesis

Chapter 2 reviews the state of the art in machine face recognition focusing in particular on the area of 3D face recognition.

Chapter 3 describes the concept of surface registration and presents some of their applications.

Chapter 4 proposes and evaluates a face recognition technique based on a variant of the iterative closest point (ICP) surface registration algorithm [12]. The residual 3D distance between two faces after registration is used as a metric for assessing their similarity. 2D intensity values of the surface texture are also used in a combined metric and their effect under various poses and facial expressions is analyzed and discussed.

Chapter 5 describes a face recognition techniques based on principal component analysis which is used to construct a 3D statistical face model for classification. A technique for the construction of statistical shape models based on registration of manually landmarked 3D faces is introduced. An alternative technique based automatic rigid surface registration of 3D faces is also introduced. Both techniques are evaluated by comparing their respective face recognition rates as well as the compactness, specificity and generalization ability of the models they generate.

Chapter 6 presents an extension to the model-building techniques of Chapter 5. Automatic non-rigid surface registration is used create a more uniform dense correspondence between faces, which is shown to further increase the recognition rates.

Chapter 7 proposes an eigenfeatures approach for performing 3D face recognition by dividing the face into facial regions and using these regions to perform classification both individually and via classifier fusion. Furthermore, the within- and between-class variability of the face across the population is computed. Based on these measures areas of high variability are removed from the 3D face data in order to create optimized statistical models.

Chapter 8 describes the contributions of this work and proposes extensions and improvements of the presented techniques.

Chapter 2

Review of Machine Face Recognition

Heraclitus famously stated that *“one can not step into the same river twice”* meaning that everything in the world is continuously changing. This entails that scenes in the real world never repeat themselves in full detail. The challenge for a vision system is to bridge the differences between various scenes and to detect similarities despite variation within the class of objects. The human face is one of the many classes for which the identification problem is studied. Machine face recognition involves the automatic identification of a person from an input image or sequence of images achieved by comparing facial features in that input image to images stored in a face database. Recognizing a face can easily be broken down into a list of sub-problems. First of all, a machine would have to detect a face from an image, normalize it (correct for illumination, size, pose etc.), extract facial features and finally perform the identification against all faces in the database. Figure 2.1 shows such a processing pipeline. For identity-based classification to take place the aforementioned subtasks need to be addressed. Isolating these tasks is also necessary for tackling one problem at a time. Perhaps more importantly, this modularity encouraged collaboration between different research groups which have similar subtask requirements such as, for example, face detection and segmentation which is a common requirement for facial expression recognition as well as face identification.

In this chapter a brief outline of research in 2D face recognition is given, followed by a more extensive examination of techniques used for 3D data. For the purposes of

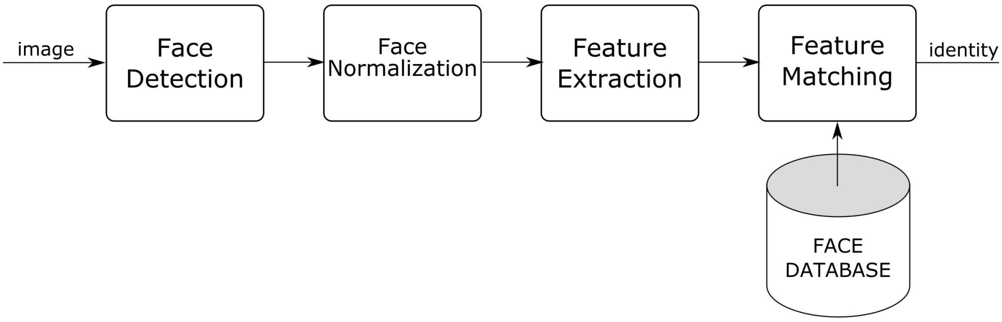


Figure 2.1: A face recognition pipeline (adapted from [119]).

this review the focus is going to be on the subtask of face recognition even though some of the methods reviewed here include automatic face detection and feature extraction. The sections that follow introduce some performance measures in order to examine these techniques and evaluate their results.

2.1 Performance measures

The performance of a biometric system can be described based on how well it performs three basic tasks; *verification*, *open-set identification* and *closed-set identification* [82]. In order to measure performance, the dataset, whether that is fingerprints, faces or any other biometric sample, can be divided into three sets of images. The first one is the set of images of the people that are known to the system, which form the database and are referred to as the *gallery* $\mathcal{G}$. The other two datasets are the *probe* sets which are presented to the system for verification or identification. Probe set $\mathcal{P}_\mathcal{G}$ contains different biometric samples of the same subjects as the ones contained in the gallery while probe set $\mathcal{P}_\mathcal{N}$ contains samples of people not in the gallery. Subjects in $\mathcal{P}_\mathcal{G}$ are also known as *clients* while subjects in $\mathcal{P}_\mathcal{N}$ are known as *imposters*.

2.1.1 Open-set identification

Open-set identification is when the system has to decide if a probe p_j is in gallery set $\mathcal{G}$ or not. When a biometric sample, p_j, is presented all samples in $\mathcal{G}$ are compared to it. When

p_j is compared to a gallery sample g_i it produces a similarity score s_{ij}, which is called a $matchscore$ if p_j and g_i belong to the same person and $nonmatchscore$ if they belong to different people. All similarity scores are ranked with the most similar (greatest score) at the top. A certain rank is denoted by rank$(p_j) = n$ which means that a probe face is the n^{th} largest score.

There are two performance statistics associated with open-set identification, the *correct detection and identification rate* and the *false alarm rate*. Initially, it is assumed that probe p_j is a probe which has a corresponding unique biometric sample g^* in $\mathcal{G}$ (i.e., $p_j \in P_G$) and their similarity score is s_{*j}. Probe p_j is considered correctly detected and identified if it is top ranked (rank$(p_j) = 1$) and the similarity score is greater than threshold τ ($s_{*j} \geq \tau$). The correct detection and identification rate P_{DI} for threshold τ is the percentage of probes in $\mathcal{P}_\mathcal{G}$ that are correctly detected and identified and it is denoted by:

$$P_{DI}(\tau, 1) = \frac{|\{p_j : \text{rank}(p_j) = 1, \quad \text{and} \quad s_{*j} \geq \tau\}|}{|\mathcal{P}_\mathcal{G}|} \tag{2.1}$$

In other words, it measures, the frequency with which clients are correctly identified as clients.

Sometimes, the general open-set identification case is reported, which examines the top n matches between a probe and a gallery. In this case, a probe is considered correctly identified if the correct match is above the operating threshold and its rank is n or less. The detection and identification rate at rank n is defined as:

$$P_{DI}(\tau, n) = \frac{|\{p_j : \text{rank}(p_j) \leq n, \text{ and } s_{*j} \geq \tau\}|}{|\mathcal{P}_\mathcal{G}|} \tag{2.2}$$

which is plotted along three axes: detection and identification rate, false acceptance rate and rank.

The second performance measure related to the open-set identification is the false alarm rate P_{FA} which measures the number of times the top match is not in $\mathcal{P}_\mathcal{G}$ and the similarity score is above the operating threshold. In this case it is assumed that p_j is in

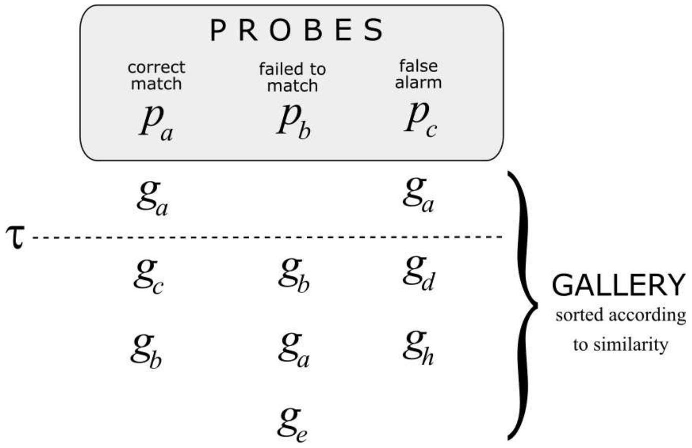


Figure 2.2: examples of a correct match, a failed match and a false alarm based on threshold τ all of which are associated with P_{DI} and P_{FA}.

$\mathcal{P}_{\mathcal{N}}$ (i.e., $p_j \in P_N$) and a false alarm occurs when the top match's score is greater than τ [82]:

$$\max_i s_{ij} \geq \tau \tag{2.3}$$

The false alarm rate P_{FA} reflects the fraction of probes in $\mathcal{P}_{\mathcal{N}}$ that are falsely thought of as being in $\mathcal{P}_{\mathcal{G}}$ and is computed by:

$$P_{FA}(\tau) = \frac{|\{p_j : \max_i s_{ij} \geq \tau\}|}{|\mathcal{P}_{\mathcal{N}}|} \tag{2.4}$$

Using the client/imposter terminology, P_{FA} measures the frequency with which imposters are thought to be clients. Figure 2.2 shows examples of a correct match, a failed match and a false alarm based on threshold τ, all of which are associated with P_{DI} and P_{FA}. Naturally, the false alarm rate P_{FA} is correlated with the detection and identification rate P_{DI}. As the threshold is lowered P_{FA} goes down (fewer imposters are taken for clients). At the same time however, making the criteria stricter decreases the detection and identification rate P_{DI}. By varying the threshold and measuring the two rates, P_{FA} and P_{DI}, one can plot this trade-off curve on a *receiver operator characteristic* (ROC) [111]. Figure 2.3 shows the two extreme cases along with a typical scenario.

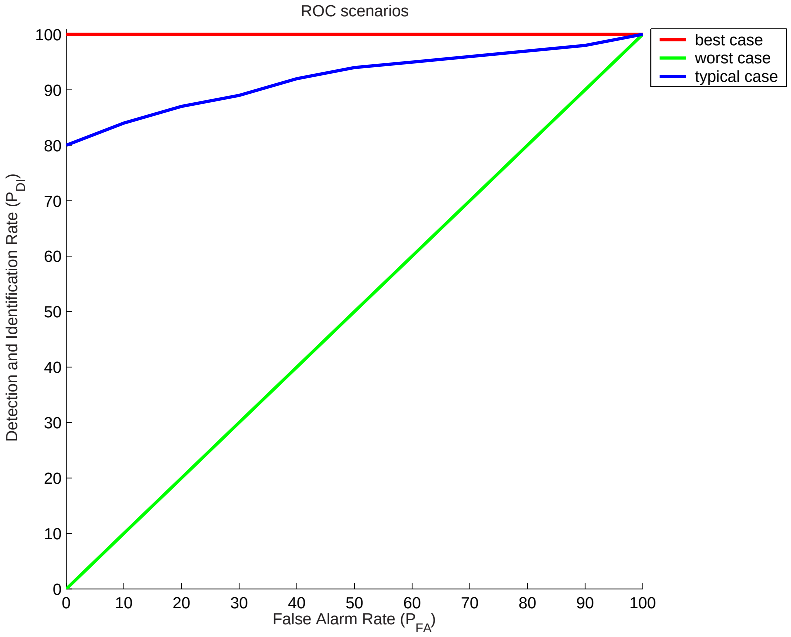


Figure 2.3: The best case, worst case and typical case scenario of an ROC curve.

2.1.2 Closed-set identification

The term closed-set identification is used when all probes belong to someone in the gallery set $\mathcal{G}$ and therefore the question asked is which gallery samples resemble the probe more closely. In other words the similarity scores between a probe p_j and all faces in $\mathcal{G}$ are sorted and the question is whether the correct match is in the top n matches. The cumulative count is calculated by:

$$C(n) = |\{p_j : \text{rank}(p_j) \leq n\}| \tag{2.5}$$

The closed-set identification rate for rank n, $P_I(n)$ is the fraction of probes at rank n or lower and it is described by:

$$P_I(n) = \frac{|C(n)|}{|\mathcal{P}_{\mathcal{G}}|} \tag{2.6}$$

2.1.3 Verification

The final performance measure reported is the verification rate. In a typical scenario the person claims to be a specific person in the gallery. The system will then compare the person's biometric sample with the gallery sample and based on a threshold will accept or reject the person's claim. There are two protocols for calculating the verification rate. The first one is called the round robin protocol in which both probe sets $\mathcal{P}_{\mathcal{G}}$ and $\mathcal{P}_{\mathcal{N}}$ are joined in one probe set $\mathcal{P}$. All scores between probe and gallery set are computed and the match scores are used to calculate the verification rate:

$$P_V(\tau) = \frac{|\{p_j : s_{ij} \geq \tau, \mathrm{id}(g_i) = \mathrm{id}(p_j)\}|}{|\mathcal{P}|} \tag{2.7}$$

while all nonmatch scores are used to calculate the false acceptance rate:

$$P_{FA}(\tau) = \frac{|\{s_{ij} : s_{ij} \geq \tau \text{ and } \mathrm{id}(g_i) \neq \mathrm{id}(p_j)\}|}{(|\mathcal{P}| - 1)|\mathcal{G}|} \tag{2.8}$$

This protocol is sometimes criticized for joining the two probe sets. Critics argue that in order to correctly assess the effectiveness of an algorithm one needs to use true (unseen) imposters. In the round robin protocol, mentioned above, all subjects in the probe set have a biometric sample in the gallery set too- they are not true imposters. An alternative solution is called the true imposter protocol in which there are still two probe sets, $\mathcal{P}_{\mathcal{G}}$ and $\mathcal{P}_{\mathcal{N}}$. The verification rate is computed from the match scores between the gallery and $\mathcal{P}_{\mathcal{G}}$:

$$P_V(\tau) = \frac{|\{p_j : s_{ij} \geq \tau, \mathrm{id}(g_i) = \mathrm{id}(p_j)\}|}{|\mathcal{P}_{\mathcal{G}}|} \tag{2.9}$$

and the false alarm rate from all the nonmatch scores between the gallery and $\mathcal{P}_{\mathcal{N}}$:

$$P_{FA}(\tau) = \frac{|\{s_{ij} : s_{ij} \geq \tau\}|}{|\mathcal{P}_{\mathcal{N}}||\mathcal{G}|} \tag{2.10}$$

2.2 Face recognition in 2D

The problem of face recognition from intensity images is the problem of identifying a three-dimensional object from its two-dimensional projection. Face recognition appeared in the engineering literature in the 1960s [21] but serious efforts in automatic face recognition research started in the early 1970s. The first works published dealt with the problem by measuring the distances between facial features in order to perform classification [106, 108]. Research remained relatively dormant during the 1980s but was revived during the early 1990s with a particular interest in making face recognition systems fully automatic. A face recognition system is often very complex as more than one technique can be used and it is therefore sometimes difficult to classify it into a category.

A brief historical overview of 2D face recognition methods is necessary to provide the historical foundations for many of the techniques that are nowadays used in 3D face recognition. In the next sections 2D techniques are divided into feature-based, holistic and hybrid methods. This is in line with the psychological models discussed in Chapter 1, which also recognized the importance of global and local features as well as their combination.

2.2.1 Feature-based approaches

Feature-based or structural techniques were used early on in face recognition and involve methods that try to extract information about a subset of the total information that the image offered, usually involving anatomical landmarks. Kelly [108] used distances between features, while in his seminal work Kanade [106] used distances and angles between anatomically distinct features such as eye corners, nostrils and mouth to identify an individual. Many years later Cox *et al.* [49] introduced a mixture-distance technique in which each face is represented by a collection of thirty manually measured distances. Conducting experiments on a database of 685 people the authors managed to achieve 95% rank 1 recognition, while a simple nearest neighbor search in Euclidean space yielded 84%. Nefian and Hayes [143] as well as Samaria [179] did not attempt to find the exact location of facial features, but instead implemented a method based on *hidden Markov models* (HMM), using strips of pixels across the forehead, eye, mouth, etc. Finally, a particularly successful system in structural matching methods has been the graph matching system [145, 216] using the *dynamic link architecture* [35]. Gabor wavelets are the building blocks of the face representation in these graph matching methods. Local features are represented by wavelet coefficients for different scales and rotations based on fixed wavelet bases (called jets), which are particularly robust to illumination change, translation, rotation, distortion and scaling. After the facial features have been located, only jets visible in both faces are used when comparing a set.

All of the above techniques depend on careful landmarking of the faces (whether automatic or manual) in order to accurately measure differences between faces, as the authors themselves readily admit [145]. Given that up to today there are no reliable enough methods to landmark the face automatically, these techniques have serious limitations. Furthermore, particularly in earlier techniques, which simply measured geometric distances between anatomical features, some parts of the textural information is discarded and the focus is placed on specific features. The holistic approaches presented in the next section are trying to explicitly exploit the information in the global appearance of the face.

2.2.2 Holistic approaches

Early techniques [106, 21] focused on detecting individual features and exploiting the relationship between these features. However, research in how humans perform face recognition has shown that the use of these features and their relationship is not enough to account for the face recognition ability of humans [205]. Ruderman [173] showed that images of the natural world are easily distinguished from images generated randomly by a computer because of their particular structure. Furthermore, naturally occurring objects such as faces are much more structurally regular than man-made objects groups [201]. This structural regularity implies that there is a great amount of redundancy in the human face that could potentially be exploited in order to describe a face with fewer parameters. In intensity images the dimensionality of the space depends on the number of pixels in the input, while in 3D data it depends on the number of points on the surface. Expressing the faces in a pixel-by-pixel or point-by-point basis is not only unnecessarily dense but also computationally very expensive.

Kirby and Sirovich [110, 188] presented a technique for a low dimensional reconstruction of the face using the *Karhunen-Loéve procedure* (KL). The dimensions of the pixel-space are reduced to a small set in which images are linear combinations of basis vectors called *eigenpictures*. By projecting each face onto each eigenpicture a parameter (weight) is obtained for each dimension which can be used to describe the face. After Kirby and Sirovich showed that it is possible to reduce the dimensionality of the face using the KL procedure many techniques sprung up using this type of projection in order to perform classification. Motivated by the aforementioned achievement, Turk and Pentland [205] used principal component analysis (PCA), a technique closely related to the KL expansion, to encode the face into a set of weights and used these weights to compare the similarity between faces. Instead on focusing on local characteristics, the *eigenfaces* technique processes the faces holistically and the global features it encodes might not be the same as a human observer's notion of features that are pivotal for identification. The approach presented in Turk and Pentland [205] takes a training set of images Γ_1, Γ_2, Γ_3,...,Γ_M and calculates the average of the set by:

$$\overline{\Gamma} = \frac{1}{M}\sum_{n=1}^{M}\Gamma_n \tag{2.11}$$

A vector of differences from the mean, $\boldsymbol{\gamma}_i$, is given by:

$$\boldsymbol{\gamma}_i = \Gamma_i - \overline{\Gamma} \tag{2.12}$$

PCA is used on this set of vectors seeking a set of M orthonormal vectors $\boldsymbol{u}_n$ and their associated eigenvalues λ_k which best describe the data distribution. The vectors $\boldsymbol{u}_k$ and scalars λ_k are the eigenvectors and eigenvalues of the covariance matrix $\boldsymbol{C}$ given by:

$$\boldsymbol{C} = \frac{1}{M}\sum_{n=1}^{M}\boldsymbol{\gamma}_n\boldsymbol{\gamma}_n^T \tag{2.13}$$

which can be represented as $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^T$ where $\boldsymbol{A} = [\boldsymbol{\gamma}_1\boldsymbol{\gamma}_2...\boldsymbol{\gamma}_M]$. Using intensity images of size $N \times N$ pixels, $\boldsymbol{C}$ becomes computationally prohibitively expensive since it is of

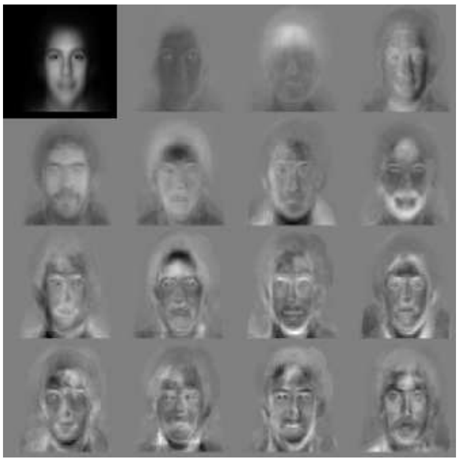


Figure 2.4: Examples of the principal components of a 2D face image population (from [83]).

size $N^2 \times N^2$. Since the number of data points is less than the dimension of the space ($M \leq N^2$) there are only $M-1$ meaningful eigenvectors and thus one could solve for the N^2 eigenvectors by first solving for the eigenvalues of an $M \times M$ matrix $\boldsymbol{L} = \boldsymbol{A}^T\boldsymbol{A}$ and then taking linear combination of the resulting vectors. With the calculation cost greatly reduced, the eigenvectors are ranked according to the degree to which they characterized the variation among the images. Out of M components the M' most significant ones are heuristically selected and retained. A new face Γ_{new} is then projected into the *facespace* by:

$$\boldsymbol{\beta}_k = \boldsymbol{u}_k^T(\Gamma_{new} - \overline{\Gamma}) \tag{2.14}$$

where $k = 1, ...M'$. In other words every face is described by a vector of weights $\boldsymbol{\beta}^T = [\beta_1, \beta_2, ...\beta_{M'}]$ that describes how much each of the principal eigenfaces contributes to describe the input face image. An example of 2D eigenfaces can be seen in Figure 2.4. The vector of each face projected on the facespace was used to compare a face $\boldsymbol{\beta}_A$ to face $\boldsymbol{\beta}_B$ by computing the Euclidean distance between them:

$$d_E(\boldsymbol{\beta}_A, \boldsymbol{\beta}_B) = ||\boldsymbol{\beta}_A - \boldsymbol{\beta}_B|| \tag{2.15}$$

The search for the closest match in a face recognition scenario involves finding the face that minimizes d_E. The logic of the eigenface approach and the reason it works better than simply comparing pixel intensities is that it uses a model that contains prior shape information. The former simply compares pixels treats all bit of information equally and does not take advantage of the high degree or regularity associated with the face. Furthermore, a simple pixel-to-pixel comparison of two images does not easily allow for meaningful conclusions about the way that the objects are different. A comparison of the principal components of two faces, however, can lead to conclusions about nature of the differences between two faces if what knows what the specific components encode. In the original publication of Turk and Pentland a database of 2,500 faces (16 subject) was used, which contained subjects digitized under three head orientations, three head sizes and three lighting conditions. That system managed to achieve 96% rank 1 classification rate averaged over lighting variation, 85% under orientation variation and 64% under head size variation [206].

Moghaddam and Pentland [139] extended the eigenface technique to a Bayesian approach. Two classes were defined, one representing the intrapersonal variation between an individual's image (V_I) and another class representing the interpersonal variation due to different identity (V_E). Given an intensity difference between two images $\Delta = \Gamma_1 - \Gamma_2$ (assuming that both classes have a Gaussian distribution), two probability functions $P(\Delta|V_I)$ and $P(\Delta|V_E)$ were calculated. An image belongs to the same individual if $P(\Delta|V_I) > P(\Delta|V_E)$. A comparative study [157] of many face recognition techniques reported a significant improvement in tests using this method over the traditional nearest neighbor classification of Turk and Pentland. Li and Lu [120] presented a different way of making decisions in an eigenspace. At least two prototype images are projected into the facespace (under different illumination or pose). A line passes through the two images and forms a *feature line* (FL) as in Figure 2.5. An input image is identified with a corresponding class, based on the distance between the *feature point* of the image and the FL of the prototype images. In comparative tests using four well-known databases this technique returned 40% error rates across varying illuminations and poses, compared to the 60% standard eigenfaces technique [205]. Despite these improvements, in the chapters

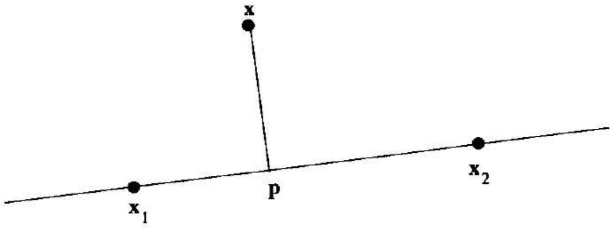


Figure 2.5: Two projections of faces x_1 and x_2 have a line pass through them forming a feature line (FL). An input image is associated with a corresponding class, based on the distance between the feature point of the image and the FL of the prototype images (from [120]).

that follow, the standard nearest neighbour classification will be used. That is for two reasons. Firstly, the latter is a more popular way of performing classification in PCA-based approaches and will therefore make comparisons between techniques more valid. More importantly, however, the method above needs at least three images; two for the gallery set and one to probe with. The experiments presented in later chapters were performed using only two images; one for the gallery set and one for the probe set.

Another way to compensate for pose differences within the subject class is the approach of Prince *et al.* [162]. They proposed a mapping from a traditional feature space to a space constructed so that each feature vector is associated with an individual regardless of pose. Based on this, they introduce some observation features that are particularly good for face classification across large pose variations. Combined with a probabilistic distance metric their experiments on the FERET database yield higher recognition rates than other contemporary methods.

Linear Discriminant Analysis (LDA) is another dimensionality reduction technique for dealing with an M-class problem and has been used extensively [9, 57, 194, 226, 227]. It is often called Fisher's Linear Discriminant to which it is closely related. PCA tries to maximize the scatter between all classes making no distinction between variation due to identity or illumination. Belhumeur [9] tried to avoid that by using LDA which creates such a reduced-dimensionality space where the ratio of the between- and within-class scatter is maximized. In other words, when using PCA the facespace is constructed so that the face *object* is represented "optimally" while when using LDA, a discriminant subspace is created to "optimally" distinguish faces of *different* people. This could potentially make classification more effective as the within-class variation is taken into account. Figure 2.6 demonstrates the effect of Fisher's linear discriminant analysis compared to a PCA-generated space. Since the number of pixels in images is larger than the number of images in the training set it is possible for the within-class scatter of the projected samples to be zero. To avoid this problem PCA is first carried out on the data (W_{pca}) to reduce the dimensions of the dataset and Fisher's linear discriminant W_{fld} [61] is then applied obtaining an optimal facespace W_{opt} by:

$$W_{opt} = W_{fld} W_{pca} \tag{2.16}$$

Comparative studies show that this technique, known as *fisherfaces* produces lower error rates than a space generated using only PCA [9]. In conclusion, these linear techniques avoid the pitfalls associated with the early geometric feature-based techniques, yet are not accurate enough to deal with nonlinearity in face modeling. For that reason these linear techniques have been extended to using nonlinear kernel techniques such as kernel PCA[181] and kernel LDA [138]. These generally perform better on the training data but may perform worse in unseen faces due to their flexibility and possible over-fitting to the training set.

Another derivative of PCA that has been used, is the *independent components analysis* (ICA) which utilizes higher order statistics to achieve greater classification power than PCA [8]. ICA effectively separates a multivariate signal into additive subcomponents on the assumption that the non-Gaussian source signals are mutually statistically independent.

Support vector machines (SVM) are another way in which the classification problem has been dealt with. Given that one wants to classify points in a multidimensional space, one is interested in splitting them by a hyperplane that separates the data points, with the maximum distance to the closest data point from both classes. Phillips [158] used an SVM-based algorithm on a difficult set of images from the FERET database and compared

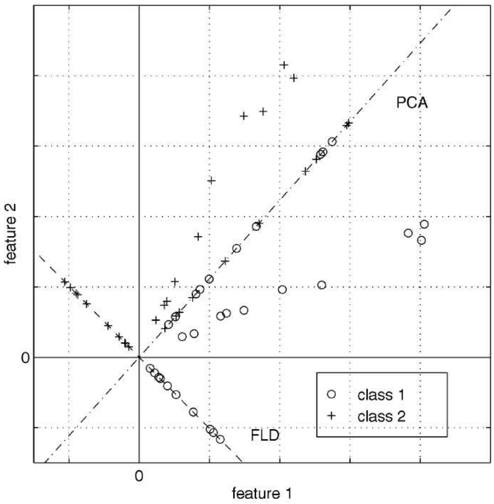


Figure 2.6: A comparison of the subspaces created with PCA and Fisher's linear discriminant (from [9]).

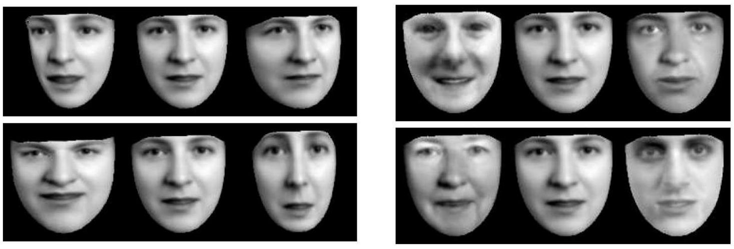
it to a PCA-based one (400 images, 200 subjects). The rank 1 rate for SVM was $77 - 78\%$ versus 54% for PCA. For verification, the equal error rate was 7% for SVM and 13% for PCA. Finally, neural networks have also been used in the holistic approach, the idea being that greater generalization can be achieved through learning. Lin *et al.* [124] used a *probabilistic desision-based* neural network while Liu and Wechsler used the *evolution pursuit* method [126].

2.2.3 Hybrid approaches

As discussed in Chapter 1 the human brain is affected by both global and local features for identifying faces. Based on that, Pentland *et al.* [154] used PCA on the whole face in combination with a modular technique which involved the generation of "eigeneyes", "eigennose" and "eigenmouth". This modular representation yields higher recognition rates and provides a more robust statistical model in lower dimensions of the eigenspace, but when the combined set is used the improvement is marginal. It was argued, imitating the brain once again, that it might be optimal to assign weights to local and global features in order to shift between them depending on the amount of gross variations that are present in the input images. The authors hypothesize that the potential advantage of the modular representation is that it can overcome some of the shortcomings of the standard eigenface approach. For example, the latter can be fooled by gross variations in the input image such as facial hair, glasses, etc.

Local feature analysis (LFA) is another technique that has been proposed in order to describe objects in a low-dimensional space in terms of the local feature characteristics and their positions [153]. A more robust system could be built by estimating eigenfaces that have large eigenvalues, while for higher-order eigenmodes LFA might be more appropriate.

Statistical model-based techniques for intensity images have also proven popular. A flexible appearance model-based method for automatic face recognition was proposed by Lanitis *et al.* [113] who used both shape and greylevel information (texture). An *active shape model* (ASM), which iteratively deforms to fit an unseen example was used to describe shape. Initially the ASM is trained using PCA on the coordinates of selected landmark points of the training set. In order to perform classification, discriminant analysis is used in order to separate interclass from within-class shape variation, caused by small changes in orientation and facial expression. Using the mean model shape, PCA is used again to construct a global shape-free model (Figure 2.7). To further strengthen this method, local greylevel models are constructed on the shape model points using simple local profiles perpendicular to the shape boundary. When given an input image all three types of information, shape parameters, shape-free global image parameters and the extracted profiles on model points, are used for classification by computing the nearest neighbor in a Mahalanobis space. Using 10 images to train and 13 images from each of the 30 subjects a rank 1 rate of 92% for 10 normal images was reached and 48% for the three remaining more challenging ones. A similar approach for modeling appearance has been developed by Jones and Poggio [104] called *multidimensional morphable model*. A 3D extension of the aforementioned models is the *3D morphable model* [18, 20, 94], which is based on similar principles. Its application has been in 2D face recognition, but since a 3D model is constructed this work is going to be reviewed in more detail in the next section.



(a) Shape modes (b) Appearance modes

Figure 2.7: The first two principal modes of shape (a) and appearance (b). The face in the middle of each mode shows the mean (from [53]).

2.3 From 2D to 3D face recognition

2D face recognition is a much older research area than 3D face recognition research and broadly speaking, presently the former still outperforms the latter. Three-dimensional techniques might in the future take over classical ones, because 3D data offers a wealth of information that 2D images do not. The next section examines some of the inherent differences between 2D and 3D face recognition.

2.3.1 Advantages and disadvantages of 3D face recognition

As previously discussed, 2D images are sensitive to illumination changes. The light collected from a face is a function of the geometry of the face, albedo, the properties of the light source and even the specification of the capturing equipment. Given this complexity, it is difficult to develop statistical models taking all these variations into account. Training over different illumination scenarios as well as normalization of 2D images has been used, but with limited success. In 3D images, variations in illumination are irrelevant as the captured shape remains intact [89]. Another factor affecting comparability has been variation in pose. Effort has been put into transforming an image into a canonical position [109] but this relies on accurate landmark placement and does not tackle the issue of occlusion. Moreover, this particular task is nearly impossible due to the projective nature of 2D images. Li and Lu [121] proposed an alternative design using a SVM-based multi-view face detection and recognition framework, which stores and models different views of the face. Many face detectors are employed, which are trained on specific views. The appropriate one is chosen using support vector regression and this helps improve the accuracy and reduce the computations required. This framework, however, requires a large number of data from many views to be collected. Statistical models [47] have addressed the pose variation problem but have not completely solved it. One of the most promising techniques has been presented more recently by Blanz *et al.* [17]. They used their 3D morphable model to estimate the 3D shape of novel faces from the non-frontal 2D input images and to generate frontal views of the reconstructed faces. These frontal views are then used for recognition instead of the original non-frontal ones. Using five images of 87 subjects they managed to reach rank 1 rates of 85.6% for non-frontal views (more on the 3D morphable model in Section 2.4.3). Higher rates of up to 92% are reached in Gross *et al.* [79] using Eigen Light-Fields on near-profile images of the CMU PIE and FERET databases. Prince and Elder [163] proposed a generative model that creates a one-to-many mapping from and ideal identity-space to the observed (input) space. In the ideal identity space the representation for each individual does not vary with pose. Using seven different poses for each of the 320 individuals they extracted from the FERET database (220 for training, 100 for testing) they managed to reach rank 1 rates of 100% for a $22.5°$. Performance at $67.5°$ is at 94% and at $90°$ 86% correct first choice matches. Both of the latter techniques, however, require the algorithm to be trained across an array of poses for each subject. Pose variation in a 3D data scenario can be minimized (if not rendered irrelevant) depending on the pre-processing or capturing method used. Another fundamental problem with 2D sensors, which is also related to pose, is that the physical dimensions of the face are unknown. The size in this case is a function of the distance from the sensor and it is therefore imperative for all 2D systems, independently of the technique is used, to standardize the faces before processing them. In 3D images the physical dimensions of the face are known and are inherently encoded in the data.

Apart from overcoming the above shortcomings of 2D data, 3D images are better at capturing surface-based events and more appropriate to describe certain properties of the face that 2D images can not. Traditional 2D-based face recognition focuses on high-contrast areas of the face such as eyes, mouth, nose and face boundary because low contrast areas such as the jaw boundary and cheeks are difficult to describe from intensity images [75]. 3D images, on the other hand, make no distinction between high- and low-contrast areas.

3D face recognition, however, is not without its problems. Illumination, for example, may not be an issue during the processing of 3D data, but it is still a problem during capturing. Depending on the sensor technology used, oily parts of the face with high reflectance may introduce artifacts under certain lighting on the surface (Figure 2.8). The overall quality of 3D data collected using a range camera is perhaps not as reliable as 2D intensity data, because 3D sensor technology is currently not as mature as 2D sensors. Another disadvantage of 3D face recognition techniques is the cost of the hardware. 3D capturing equipment is getting cheaper and more widely available but its price most often can not be compared to a high resolution digital camera. Moreover, the current computational cost of processing 3D data is higher than for 2D data. The processing involved makes a lot of the 3D techniques impractical for real-time environments. Both of the above hardware issues are problems today, but given the trend in pricing, one can expect the capturing and computational hardware to become an increasingly smaller cost. A more important disadvantage of 3D capturing technology is the fact that capturing 3D data requires cooperation from a subject. As mentioned above, lens or laser-based scanners require the subject to be at a certain distance from the sensor. Furthermore, a laser scanner requires a few seconds of complete immobility, while a traditional camera can capture images from far away with no cooperation from the subjects. A final disadvantage of 3D technology is data related. There are still not as many publicly available, high-quality 3D face databases as for 2D data and moreover, the vast 2D legacy databases build by law-enforcement agencies world-wide could be rendered unusable if 3D scans become a standard biometric modality. It is partly because of this 3D data shortage, especially in the early stages of 3D face research, that a lot of researchers reported very promising identification rates, which drop sharply when applied to larger databases.

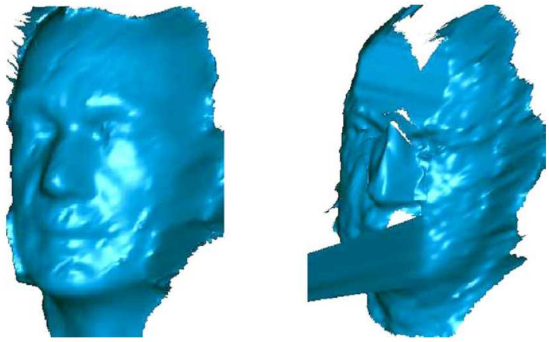


Figure 2.8: 3D data capture errors due to illumination. The image on the left is captured with light that is appropriate for the sensor used while the image on the right was captured when an additional studio spotlight was used (from [24]).

2.4 An overview of 3D face recognition

The earliest research in 3D face recognition was presented in 1989 [38], but for most of the 1990s there was little work done in the area. By the end of the last decade interest in 3D face recognition was revived and has increased rapidly since then. In the sections that follow, an overview of relevant research in techniques using 3D data is presented. Once again, many approaches are difficult to classify as they often combine many techniques. In this review the techniques have been broadly divided into three categories: surface-based, statistical and model-based approaches. Within each category an attempt was made to group similar techniques together, rather than to report findings in simple chronological order.

2.4.1 Surface-based approaches

Surface-based approaches are approaches that describe faces by explicitly using the surface geometry, whether by relying on local or global curvature, profile lines or distance-based metrics.

2.4.1.1 Local methods

More specifically, *local* surface-based methods are methods where local features such as eyes, nose and mouth are extracted from each face and their respective characteristics compared.

With the availability of 3D data a lot of researchers early on tried to harness the discriminatory power in curvature information. Lee and Milios [115] tried to match similar local, facial characteristics together in order to perform comparisons. Facial features correspond to convex regions of the face scan and based on the sign of the mean and Gaussian curvature an *Extended Gaussian Image* (EGI) is created for each region. An EGI is a 1-1 mapping between all points in a feature region and points on the unit sphere with the same normals. By correlating the EGIs of regions a similarity metric is established, which is used in concurrence with a graph matching algorithm with relational constraints to establish optimal correspondence between convex regions. This proposal is able, up to a certain extent, to deal with facial expressions, but is insensitive to change in object size. The reason for this improvement in dealing with facial expressions is because convex regions of the face that this technique uses, do not change as much as other regions between facial expressions [115]. The authors did not report face recognition results. Years later Gordon [75] used the surface curvature in order to segment the face into its features. The features are defined in terms of a high level set of relationships of depth and curvature values and the extraction is implemented as a constrained search on the surface. For each point $\boldsymbol{p}$ on the surface, a curve is formed by intersecting the surface and the normal plane in a tangent direction $\boldsymbol{t}$. The curvature of this planar curve is the normal curvature κ_n at point $\boldsymbol{p}$ in the direction $\boldsymbol{t}$. The principal curvatures, K_{max} and K_{min}, are defined by the maximum and minimum normal curvatures at each point. The Gaussian curvature K at each point is then defined by the product $K_{max}K_{min}$ and the mean curvature H by $(K_{max} + K_{min})/2$. Using these curvature maps features such as eyelids and noses are extracted and each face is represented by a vector of feature descriptors. The comparison to another face takes place in that feature space and as long as the features are correctly detected and there is no variation due to facial expressions, the descriptors belonging to the same person are very similar allowing for good classification by identity. Using three views of eight faces, rank 1 rates between 80% and 100% were reported. Moreno *et al.* [141] also used three-dimensional descriptors to compare facial surfaces. Once again, using the signs of the mean and Gaussian curvatures, the facial features with pronounced curvature are segmented. Eighty-six non-independent features from the segmented regions are then obtained and for each image a feature vector is created. Using the Fisher coefficient [84] the 35 more powerful descriptors out of the whole set of 86 are selected and the authors managed to achieve rank 1 rates of 78% and 92% in rank 5 experiments on datasets of 60 individuals which included small rotations and facial expressions. The great difference between rank 1 and rank 5 rates is not surprising since it is much easier for the correct match to be in one of the five most similar matches. A further use of curves for local feature analysis was proposed by Lee *et al.* [118] who extracted three curvatures, eight invariant feature points and their relative features using the geometric characteristics of the face. By relative features the authors refer to the distances and the ratios between feature points and the angles between feature points extracted previously. Using two classification techniques, a feature-based SVM and a depth-based *dynamic programming* (DP), Lee *et al.* reported rank 1 rates of 96% using 100 subjects.

Another locally-oriented technique is based on using *point signatures*, an attempt to describe complex free-form surfaces, such as the face. It was proposed by Chua and Jarvis [45] as a form of representation of the structural neighborhood of a point in a more complete manner than just using its 3D coordinates. These point signatures could then be used for surface comparisons by matching the signatures of data points of a “sensed” surface to the signatures of data points representing the model’s surface. Figure 2.9 shows some example point signatures for various types of surfaces. Point signatures like these are used in a later publication by Chua *et al.* [44] for comparing faces. Firstly these signatures are employed for rigidly registering the faces to each other and then they are used for face-to-face comparisons. In order to make the algorithm more robust to variation due to facial expressions the part of the face that is particularly non-rigid (mouth and chin) is automatically discarded and just the rigid parts (forehead, eyes, nose) are used for comparison (Figure 2.10).

Extracted shape features from 3D feature points and texture features from 2D feature points are first projected into their own subspace using PCA.

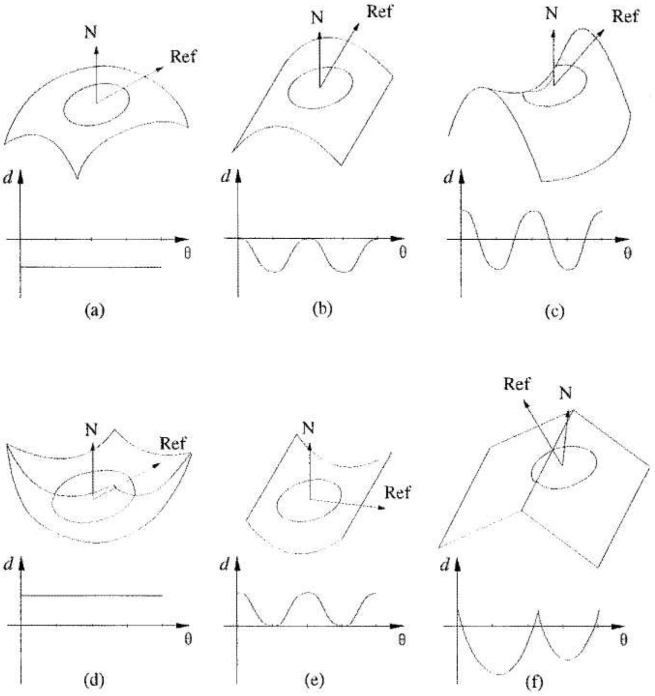


Figure 2.9: Point signature examples: (a) peak, (b) ridge, (c) saddle, (d) pit, (e) valley, (f) roof edge (from [45]).

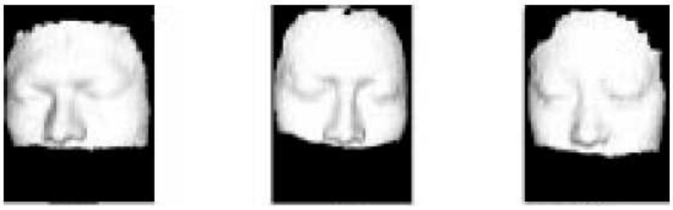


Figure 2.10: Removing the non-rigid part of the face to perform surface-based recognition (from [44]).

Point signatures were also used in Wang *et al.* [213] but previous work was extended by fusing extracted 3D shape and 2D texture features. In the 2D domain Gabor filter responses are used to get feature points. Both the 2D and 3D features are projected into their own subspaces using PCA. The two vectors are then normalized to form a combined vector to represent each facial image. The classification is done using SVM and a *decision directed acyclic graph* (DDAG) and the rank 1 rate involving 50 people with various facial expressions taken from different viewpoints exceeded 90%. What is notable is that the recognition rate using this combined feature vector is significantly higher when using either modality by itself.

A hybrid technique using both local and global information was developed by Xu *et al.* [219]. A scattered 3D point cloud is first represented with a regular mesh using hierarchical mesh fitting. Then local shape variation information, in the form of *Gaussian-Hermite moments* along with a 3D mesh representing the whole facial surface, is used to describe the individual. Both global (surface mesh) and local (Gaussian-Hermite moments) information is encoded as a combined vector in a low-dimensional PCA space and matching is based on minimum distance in that space. It was noted that variation near the mouth, nose and eyes is particularly important for characterizing the individual. Using 30 faces Xu *et al.* demonstrated that taking into account the local shape variation can improve the rank 1 rate allowing it to reach 92%. These rates, however, decrease to 72% when 120 faces are used. This sharp difference in score is discussed later in this chapter were the importance of using large enough datasets for testing and evaluation is discussed.

2.4.1.2 Global methods

Global surface-based methods are considered methods that use the whole face as the input to a recognition system. The earliest method that uses both global and local features is the one presented by Cartoux *et al.* [38]. In that work the face's plane of bilateral symmetry is located by segmenting the range image based on the principal curvature. This plane is then used to align the faces to each other in order to compare facial profiles extracted as well as the whole surface using the distance between them as a similarity metric. Conducting experiments on a limited 5-person database, 100% rank 1 rate was reached.

Similarly to Lee and Milios [115], Tanaka *et al.* [196] represent the face based on the analysis of maximum and minimum principal curvatures and their directions. But contrary to the former, this technique does not require face feature extraction or surface segmentation. Each face is represented as an EGI by mapping the curvatures at each surface point onto two unit spheres, representing ridge and valley lines respectively. Fisher's spherical correlation is then employed on the EGIs of faces to evaluate the similarity between faces. 100% rank 1 rates are reached on 37 subjects. EGIs are also used in Wong *et al.* [217] to summarize surface normal orientation statistics. The authors recognized the inadequacies of the original EGI representation in distinguishing between subtly different classes of head models and instead used genetic optimization algorithms to search for an optimal transformation for these surface normal orientations. The transformed distributions for these random variables are then used as the modified classifier input. Two classification methods based on minimum distance were also suggested, which had to be trained, initially, to maximize correct classification before they could be tried on test data. This technique was tried on a dataset of five subjects and a rank 1 rate of 80.08% was achieved. On experiments using synthetic data, however, when the set is increased from 6 to 21 subjects there is a decrease of 10% in the recognition rate.

Various distance-based techniques have also produced good results. The *Hausdorf distance* has been used extensively for measuring the similarity between two sets of points [172]. It is a general measurement and it can be applied to a variety of problems. The *undirected* Hausdorff distance between two point sets A and B is defined as:

$$H(A, B) = max(h(A, B), h(B, A)) \tag{2.17}$$

where $h(A, B)$ represents the *directed* Hausdorff distance:

$$h(A, B) = \underset{\boldsymbol{a} \in A}{max} \; \underset{\boldsymbol{b} \in B}{min} ||\boldsymbol{a} - \boldsymbol{b}|| \tag{2.18}$$

and $||x||$ is a norm. However, if the surfaces are very close to each other, $h(A, B)$ and $h(B, A)$ are small, which results in $H(A, B)$ being small. To tackle this issue Ackermann and Bunke [1] used the *partial* Hausdorff distance defined as:

$$H_{LK}(A, B) = max(h_L(A, B), h_K(B, A)) \tag{2.19}$$

where only the L and K closest points in sets A and B respectively are taken into consideration. The directed partial Hausdorff distance in this case is:

$$h_L(A, B) = L^{th}_{\boldsymbol{a} \in A} \underset{\boldsymbol{b} \in B}{min} ||\boldsymbol{a} - \boldsymbol{b}|| \tag{2.20}$$

where L is the L^{th} ranked distance from any point in A to B, where distances are ranked in increasing order. Before comparing one face to another, the faces had to be aligned to each other. To do that a plane is fitted into a given set of data and is rotated around the $x-$ and $y-$axis until it becomes parallel to the focal plane of the camera. The 3D version of the partial Hausdorff distance is used to measure the similarity between probe and gallery images. This method was applied on 240 images (10 images for each of the 24 people) and rank 1 rates of up to 100% were achieved. Pan *et al.* [150] use a priori knowledge about the structure of the face and its features, such as the prominence of the nose, to align the input data with a face stored in the database by minimizing the partial directed Hausdorff distance and solving for $\boldsymbol{T}$:

$$\underset{\boldsymbol{T}}{min}\ h_L(A, \boldsymbol{T}(B)) \tag{2.21}$$

where $\boldsymbol{T}$ is a transformation group (translation, rotation, scaling). After the faces are aligned the directed partial Hausforff distance is used as a similarity metric once again and on a database of 30 individuals an equal error rate (EER) as low as 3.24% was reported. For comparative reasons an approach using PCA was also implemented and the best EER achieved was 5%. Lee and Shim [117] used a *depth-weighted* Hausdorff distance combined with the surface curvature information in order to measure the similarity between faces. This version of the Hausdorf distance reports rank 1 rates of up to 98% while the traditional implementation of the Hausdorff distance yields below 90% in experiments on a database of 42 people. Further optimizations of the Hausdorff distance approach have been proposed such as in Russ *et al.* [176] where a way to reduce the space and time complexity of the search for the closest match was proposed by modifying the standard 3D formulation of the Hausdorff matching algorithm to operate on a 2D range image.

Medioni *et al.* [135] used the *iterative closest point algorithm* (ICP) (see Section 3.2.2) to register two surfaces and to generate a distance map based on the distances between pairs of points. Comparison of these maps yielded a recognition rate of 90% using 100 subjects. What is also interesting is that instead of using a structured light sensor for data collection a passive sensor is used, illustrating their eligibility as a hardware option. A passive sensor is a system in which two cameras with known geometric relationship collect images from the subject, the system finds correspondences between the two images and the 3D location of the points can be calculated. In our 2004 publication [151] a face recognition technique was introduced and evaluated, which uses the iterative closest point algorithm to register two facial surfaces and calculate their similarity. The similarity between faces is computed by measuring the average point-to-point distance from one surface to the other. Furthermore, the differences in texture intensity between sets of corresponding points is included in the metric (see Chapter 4). A similar technique was presented a year later by Maurer *et al.* [133] using a more complex score fusion technique. Lu *et al.* [127] extended our method on two levels. The authors made the rigid registration more robust by introducing a coarse initial rigid registration step by performing automatic detection of a few easily locatable landmark points. A finer surface registration is then performed using *hybrid* ICP. In the surface evaluation stage the same metrics are used as in our publication (surface+texture), but the shape index at each point is also compared. The shape index has been derived from the maximum and minimum local curvature. In experiments using 18 faces that included some semi-profile faces and some with facial expressions, a rank 1 rate of 92% was achieved. In 2005, Lu and Jain [128] proposed a

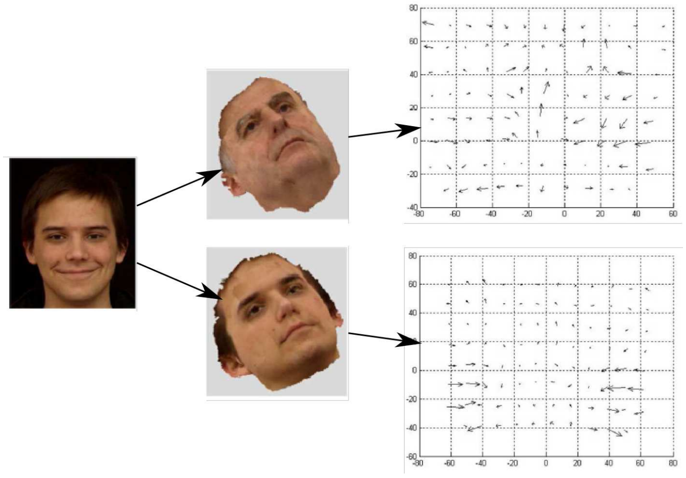


Figure 2.11: Examples of displacement vectors generated between facial surfaces of the same subject (bottom) and of different subjects (top) (adapted from [128]).

model in which after rigidly registering the surfaces using ICP, thin plate splines (TPS) are used to estimate the non-rigid transformation from one surface to another. In order to make the classification more robust, the intra- and inter-subject variation in deformation using a small training set are measured and compared to each other. This provided a general scheme to distinguish deformations due to identity from deformations due to facial expressions. Figure 2.11 shows the inter- and intra-subject deformation when a smiling face is compared to neutral ones. These deformation fields are fused with the average point-to-point distance map and SVM is used to perform the classification. Using a 100-person dataset with neutral and smiling data Lu and Jain found that most of the incorrect identifications are due to smiling. After non-rigid deformation these are reduced substantially and 89% rank 1 rates were reached using 3D data and 91% using 3D+2D.

Another sub-family of techniques employed is the one using profiles extracted from the face. In early work by Nagamine *et al.* [142] five feature points are used to align the range data to each other. Curves of intersections, called “sections”, which were considered important for recognition are then extracted across the central region of the face. These were the horizontal plane across the eyes, the vertical plane splitting the face in half and a cylindrical section around the nose (Figure 2.12). The range data along each section forms a feature vector which can then be compared to other vectors using the Euclidean

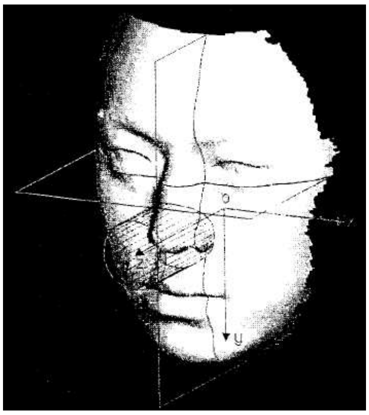


Figure 2.12: Curves of intersections used for feature-based recognition (from [142]).

distance. Rank 1 rates of up to 100% were reported using this technique on 16 subjects with 10 images per subject. This technique, however, is reliant on good landmark selection for the initial dataset alignment. The authors underline that “correct” registration of the datasets is pivotal for achieving good recognition rates.

More recently Beumier and Acheroy [13] employed vertical profiles of 3D models. The central profile and an average of the two lateral profiles are used (Figure 2.13) for comparisons, both of which are extracted automatically from the surface. After aligning the profiles to each other the similarity between them is calculated by measuring the average nearest neighbor distance and an error rate of 9% on a dataset of 30 people was reported. This technique was extented further by using 2D information [14]. A weighted sum of four classifiers (2 shape profiles, 2 texture profiles) is used for classification. The coefficients of this linear combination are estimated using Fisher’s method, which searches for the hyperplane that best separates client and imposter scores. The error rate with the fused scores was reduced to 1.4% using 27 faces, which was significantly better than either modality by itself. Others have used a different set of profiles to reach similar conclusions. Wu *et al.* [218] used the central profile and two *horizontal* crossing profiles across the nose and the forehead. Profiles in the sensed data are aligned to the database data by using an automatic technique minimizing the partial Hausdorff distance which is also used as a similarity metric. Using the same dataset as Beumier and Acheroy [13, 14], an error rate between 1.1% and 5.5% was achieved.

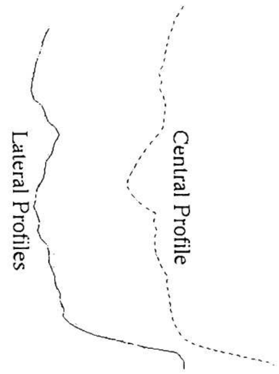


Figure 2.13: Vertical profile lines used for recognition. The central profile is used along with the average of two lateral profiles (from [13]).

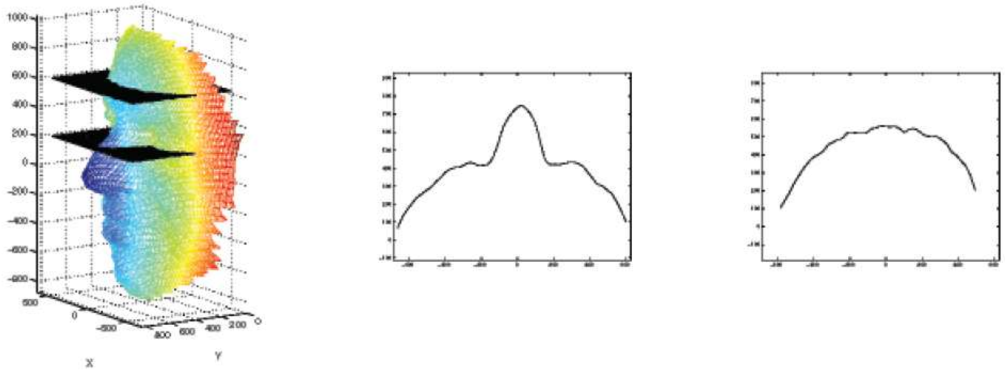


Figure 2.14: Horizontal profile lines used for recognition (from [218]).

2.4.2 Statistical approaches

PCA-based techniques are widely used for 2D facial images. More recently, PCA-based techniques have also been applied to 3D data. Mavridis *et al.* [134] presented a technique using PCA for face recognition, but it was only in a publication a year later [203] that PCA is used to evaluate three modalities of face recognition: color, depth and a combination of color and depth. Using 40 subjects, a rank 1 rate of almost 99% was achieved with the multi-modal algorithm performing significantly better than 3D or 2D alone. In a later publication Tsalakanidou *et al.* used *embedded* hidden Markov models (EHMM) instead of eigenfaces in order to combine depth and intensity images. This approach resulted in an error rate between 7% and 9%[202].

Hesher *et al.* [89] also presented a method using PCA. One of the reoccurring issues discussed later in this thesis is the need, when using PCA, to align the data “correctly”. Hesher *et al.* achieve this by converting the 3D data to 2D depth maps, thus simplifying the problem computationally and using a feature line on the nose to perform rigid registration. What they then assume is that pixels on the 2D images with the same index value correspond to each other. The depth maps are later trimmed by fitting an ellipse around them, thus reducing the effects of noise on PCA classification. In their paper the gallery used contained six images for each of the 37 subjects. This is known to improve the results significantly. The authors explored the effects that the gallery and testing size have on the recognition rates by manipulating these parameters across experiments. The best rank 1 rate achieved was 90% using the largest number of training samples (185) and 37 images for testing. In contrast using 37 images for training (1 for each of the 37 subjects) and 185 for testing reduced the rank 1 rates to 83%.

Chang *et al.* [39] presented a very extensive study of PCA-based face recognition using 2D and 3D facial data. Once again the range images collected are converted into depth maps which are normalized for pose using manually selected landmarks. Experiments are conducted to evaluate the effects of reducing the spatial and depth resolution of images giving insights into the sensor accuracy level needed to meet the requirements of face recognition tasks. Their results with 3D and 2D data provide a valuable insight into the fidelity of depth representation that is required for face recognition. Their experiments indicate that 2D-based techniques are less sensitive to a degradation in quality while 3D are more sensitive. Starting with an initial 130×150 2D image they got 89% using the whole image in 2D. The rank 1 rates were reduced only marginally until the image was reduced to 25% of its original size when they dropped to 79%. On the other hand using 130×150 3D range images the rank 1 rate dropped to 61% when the image was reduced to 35% of the original size. The authors warn the reader, however, not to jump to conclusion about this disparity in sensitivity between 2D and 3D images until the 3D capture technology matures enough. What is also important in this study, is that the extensive database of 676 probes used, includes images of people that are taken over many weeks, introducing variability due to time, which is generally absent from 3D datasets. Multimodal rank 1 recognition using nearest neighbor reached 99% while 3D and 2D reached 94% and 89% respectively.

Just as with 2D data, LDA has also been applied to 3D data. Gökberk *et al.* [71] made use of various approaches to 3D face recognition on the same dataset. ICP-based point cloud representations, normal-based representations, PCA and LDA-based depth map techniques and profile-based approaches are compared. The LDA-based technique performs best, but point cloud and normal-based classifiers came close seconds. It was also concluded that PCA-based techniques are sensitive to alignment issues. LDA is not so sensitive since it takes into account within-class variability, which also includes registration errors (see Section 2.2.2). Three classifiers are finally combined (normal-based, LDA-based, profile-based) using nonlinear rank-sum method, and rank 1 rates of 99.07% were reported while using the LDA-based classifier by itself achieved 96.27%.

Some of the more interesting variations of statistical approaches have been developed in an effort to minimize the effects of facial expressions. Empirical observations show that facial expressions can be modeled as length-preserving (isometric) transformations which do not stretch or tear the face and thus preserve the surface metric. Schwartz *et al.* used *multidimensional scaling* (MDS) to flatten complex surfaces of the brain onto a plane in

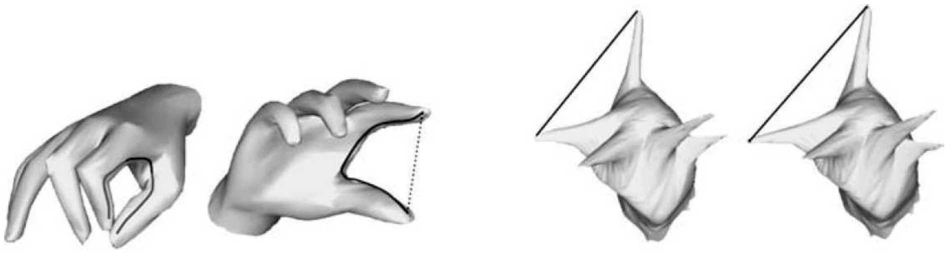


Figure 2.15: A hand undergoing an isometric transformation converting the geodesic distance between points into Euclidean. Despite the differences in the hand gestures the geodesic distances between the two points after the transformations remain similar [28].

order to study their functional architecture [182]. Zigelman *et al.* [229] and Grossman *et al.* [81] extended these ideas to texture and voxel-based cortex flattening. Elad and Kimmel [54] introduced a generalization of this approach in the field of object recognition. Figure 2.15 shows a hand undergoing isometric transformations which convert the geodesic distance between points into Euclidean. Under the assumption that the geodesic distance between parts of the face changes very little due to facial expression Bronstein *et al.* [27, 28] applied this to facial surfaces so that they are invariant to isometric deformations. It is noted that deformations due to facial expressions can be modeled as isometries while maintaining the intrinsic geometric properties of the face intact. Figure 2.16 shows three faces of the same person with strong facial expressions. Nevertheless, the canonical representations of the facial surfaces do not appear to change significantly. All images in the database are flattened as shown in Figure 2.17 and the flattened texture together with the canonical image are used with PCA on a database of 30 participants reaching 100% recognition rates. Furthermore, the authors claimed that using this technique, it was possible to distinguish between identical twins as well.

A similar technique involving the mapping of a 3D facial surface to an isomorphic planar space was proposed by Pan *et al.* [149]. The authors claimed that this mapping scheme transforming data “from 3D spatial space to an isomorphic planar space provides a trade-off among different features of the surface while trying to preserve them”. This approach is more flexible and adaptive than the Bronstein solution. A *region of interest* (ROI) from the face is initially extracted by detecting the facial bilateral symmetry plane

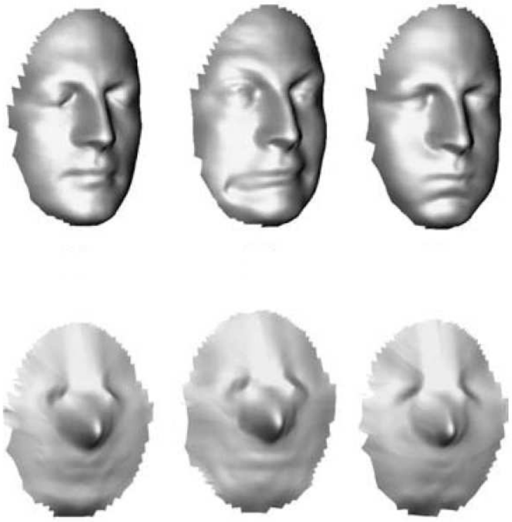


Figure 2.16: Faces with strong facial expressions (top) and their canonical forms (bottom). Notice how despite the apparent change in the 3D structure the canonical forms of the faces remain remarkably similar [28].

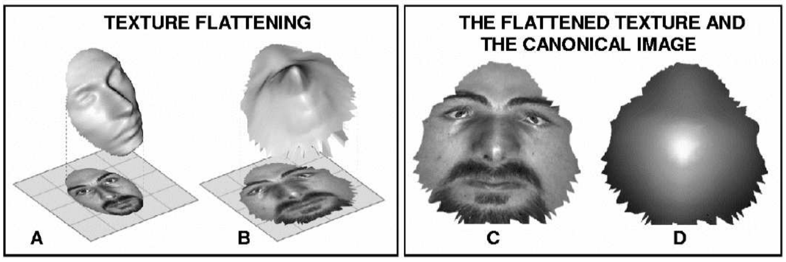


Figure 2.17: Texture mapping of the facial surface (A), in canonical form (B), the resulting flattened texture (C), and the canonical image (D) [27].

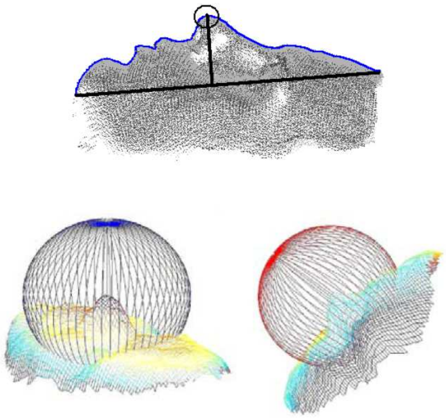


Figure 2.18: The nose is automatically detected as the vertex on the central profile curve with the maximum distance to the line through both ends of the curve. The region of interest is extracted using a sphere on the nose tip (from [149]).

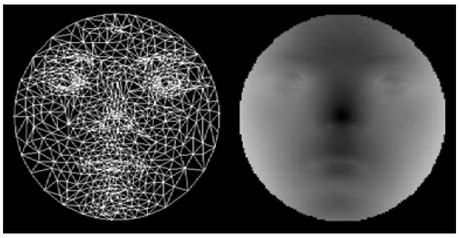


Figure 2.19: After the region of interest (ROI) is extracted using a sphere placed on the nose tip, the mapped ROI (left) is generated along with the mapped relative depth image (right) (from [149]).

and finding the nose tip as in Figure 2.18 (top). A reference plane is then built through the nose tip for calculating the relative depth values (Figure 2.18 (bottom)). The ROI is triangulated and parameterized into an isomorphic 2D planar circle, trying to preserve the geometric properties of the face (Figure 2.19). The depth values at each point are mapped on the planar circle and PCA is performed on it. It is claimed that this technique is insensitive to pose variation and 95% rank 1 rates is reported, 5% more than the baseline PCA technique it was compared to.

Generally speaking, PCA is a powerful technique for dimensionality reduction of the feature space. It has been noted, however, that the performance of this technique deteriorates with a larger database as it starts getting affected by outliers which appear more often as the database size increases. In order to deal with a skewed facespace some researchers have focused on optimizing the reduced dimensionality projection. Srivastava *et al.* [192] used PCA, *independent component analysis* (ICA), *Fisher's discriminant analysis* (FDA) and other techniques for classification with the primary aim to find the optimal k-dimensional subspaces of R^n where n is the size of the images used and k is the desired dimensionality of the feature space. In other words, they used an optimal component analysis in order to learn the subspace (i.e. a stochastic optimization algorithm) to find a subspace that maximizes the performance of the classifier on the training image set. This way, the problem is reduced to an optimization task and once the optimal projection is found the recognition rates are significantly higher than the ones obtained from an non-optimized feature space. Using nearest neighbor classification on a “optimal” PCA-space increased the face recognition (rank 1) to 99% from 77% in a sub-optimal facespace. Similar increases are also observed using the other statistical methods. However, this technique requires a significant amount of computation since there is no closed-form solution to the optimization problem dealt with.

2.4.3 Model-based approaches

Blanz and Vetter [19] proposed a technique that uses a statistical 3D model of the face, based on high resolution laser scans. This 3D model is employed to assist face recognition of 2D images. This 3D representation of a face tries to accurately model the illumination and pose variation and separate these from the variation caused from the face itself due to identity and facial expression. In theory this would make a face recognition method more robust to pose and illumination changes. The goal is to be able to use this statistical 3D model to synthesize a 2D image of the face that is as similar as possible to the input image. The parameters of the synthetic face can then be used for identifying the individual.

The work involving the 3D morphable model can be divided in two parts; generating the statistical model and morphing it to fit the input data. As discussed in Chapter 5 in more detail, establishing good correspondence between features is important for creating good models. Past techniques ([50, 113, 211]) build a statistical face model from separate shape and texture vector spaces.

Instead of using pixel intensities from 2D images, Blanz and Vetter used 3D data to extract shape information and build statistical models. This image synthesis method is based on the assumption that 3D object classes are linear and that this linearity extends to 2D projections of 3D objects [210]. The linear shape model is extended from a feature-based representation to full images of the objects and is in principle very similar to the active shape model of Cootes *et al.* [48].

For building the 3D morphable model, 200 face scans consisting of 70,000 points were used. The data is stored in cylindrical coordinates relative to a vertical axis. The scanner measures the radius r and the red, green and blue components of the surface texture (RGB) in angular steps ϕ and vertical steps h. The head is parameterized as a cylindrical representation:

$$\mathbf{I}(h,\phi) = [r(h,\phi), R(h,\phi), G(h,\phi), B(h,\phi)] \tag{2.22}$$

At the core of the model-building process lies the dense correspondence calculation using optical flow to find the dense vector field $\boldsymbol{v}(h,\phi) = [\Delta h(h,\phi), \Delta\phi(h,\phi)]$ so that all points in two face scans, $\mathbf{I}_1(h,\phi)$ and $\mathbf{I}_2(h+\Delta h, \phi+\Delta\phi)$ correspond to each other. The pairing of points is achieved by minimizing the following cost function:

$$E = \sum_{h,\phi \in R} ||v_h \frac{\partial \mathbf{I}(h,\phi)}{\partial h} + v_\phi \frac{\partial \mathbf{I}(h,\phi)}{\partial \phi} + \Delta \mathbf{I}||^2 \tag{2.23}$$

which is a modified version of a typical optical flow algorithm. In eq. 2.23 **I** is normalized by:

$$||\mathbf{I}||^2 = w_r r^2 + w_R R^2 + w_G G^2 + w_B B^2 \tag{2.24}$$

The weights, w_r, w_R, w_G, w_B are chosen heuristically to compensate for the variation of different components and to control the overall weighting of texture and shape information. After correspondence has been established, PCA is performed on the shape and texture vectors. Using the resulting model it is possible to generate novel faces as a linear combination of the shape and texture principal components. In order to generate a larger variety of different faces these linear combinations of shape and texture are computed separately for different areas of the face. These areas are defined on the reference face manually once and are then propagated across all scans, given that the correspondence is established. These areas are the eyes, the nose, the mouth, and the surrounding area. The algorithm then manipulates the global shape and appearance parameters and the parameters of the segmented regions separately, enabling it greater flexibility.

The second part of this work involves the calculation of the shape and appearance parameters that fit the 2D projection of the 3D model on the 2D input image (Figure 2.20). The algorithm requires the 2D coordinates of seven facial features for initialization in order to reduce the size of the search space. Figure 2.21 shows some examples of fitting. After fitting is achieved, every 2D face in a database can be described by the shape and texture parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ of the 3D morphable model in a unified vector $\boldsymbol{c}$ which contains 99 heuristically chosen parameters that were deemed most relevant for describing facial variation of the statistical model.

$$\mathbf{c} = \left[\frac{\alpha_1}{\sigma_{S,1}}, ..., \frac{\alpha_9 9}{\sigma_{S,99}}, \frac{\beta_1}{\sigma_{T,1}}, ..., \frac{\beta_9 9}{\sigma_{T,99}}\right] \tag{2.25}$$

The similarity between two faces can then be computed by comparing their combined vectors $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. Blanz *et al.* [18] used a gallery of 68 subjects illuminated from the same direction which was queried with 4,420 images of the same subjects in 3 poses under 22 different illumination directions. Using images that the model fitted to correctly (80% of all images) a 92.8% recognition rate (rank 1) was reached despite such extreme variations in pose and illumination. When those with poor model fits were included 82.6% were correctly identified.

Two disadvantages of the above system are its computational cost, which was in the order of minutes per image, and the need for manually placed landmarks to initialize the pose of the 3D morphable model. To tackle these, Huang *et al.* [94] implemented a component-based technique using the 3D morphable model of Blanz and Vetter [19] and

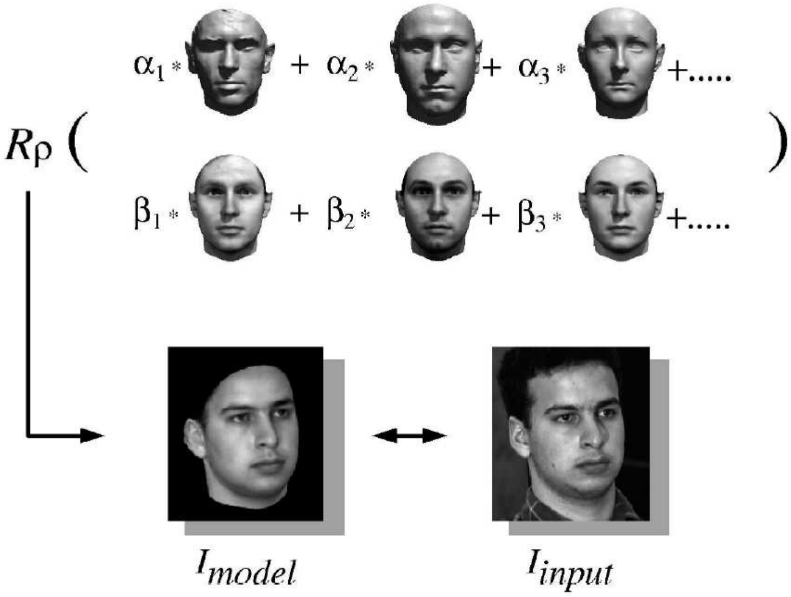

R_ρ (
$\alpha_1 *$ + $\alpha_2 *$ + $\alpha_3 *$ +.....
$\beta_1 *$ + $\beta_2 *$ + $\beta_3 *$ +.....
)
I_{model}
I_{input}

Figure 2.20: Generating synthetic images by manipulating the parameters of a 3D morphable model (from [19]).

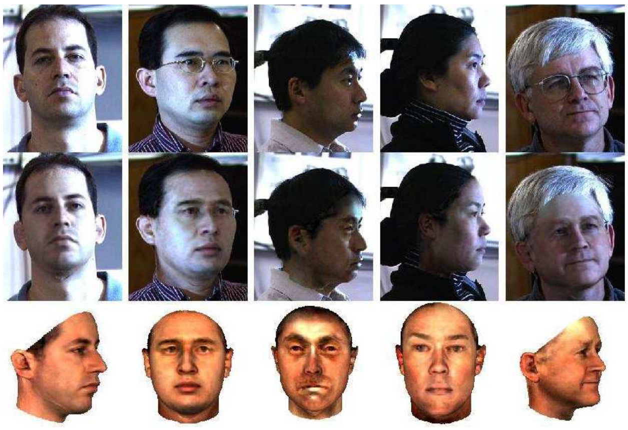


Figure 2.21: Examples of synthesized faces using the 3D morphable model. The top row shows the input images, the middle row the fitting of the synthetic 2D face image on the original input image while the bottom row shows a different rendering of the input images using the morphable model (from [20]).

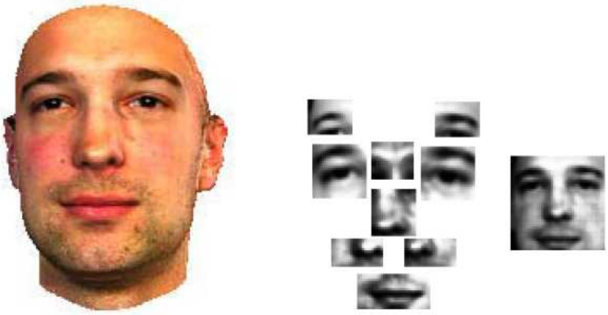


Figure 2.22: Component-based face recognition using the 3D morphable model. The image on the right shows the 2D components extracted from the face that are related to each other via a geometrical model and used for classification (from [94]).

the component-based detection and recognition of Heisele *et al.* [88]. The component-based method of Heisele *et al.* decomposes a face into the set of anatomical facial features which are related to each other via a geometrical model. In order to train such a model one needs a large number of images. The 3D morphable model of Blanz and Vetter is used to synthesize arbitrary images under varying pose and illumination. Using three images of each person in the gallery a 3D face model is computed and synthetic images are generated in order to train both the features detector and the classifier. Ten facial components are extracted for each face (Figure 2.22) and are combined into a single vector which is trained by a SVM classifier. Using a testing set of 10 subjects with 200 images per subject across various poses it was possible to reach 88% rank 1 rate for the component-based technique which was significantly higher than the global method. It was also possible to bring up the processing speed to four faces per second.

Ansari *et al.* [4, 5] presented a much simpler model generation using just a selection of points on a 2D image rather than the whole face. This technique involves the automatic extraction of feature points from a frontal and profile view of intensity images and using these features to deform a 3D generic face model in order to obtain a 3D face structure for each person. *Procrustes analysis* is used in order to globally minimize the distance between the features of the 3D model and the 2D points obtained from the images. After the global transformation is completed a local deformation around the feature vertices is performed in order to create a more realistic 3D representation of each person. Figure 2.23

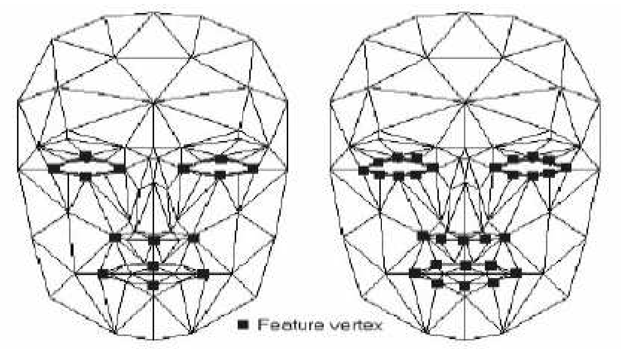


Figure 2.23: A generic model used for modeling distances between global and local features of the face (from [4, 5])..

shows the global deformation vertices (left) and the local ones (right). The coordinates of these 29 feature points are used for distance-based comparisons to other faces. The database model that has the largest number of vertices close to the test model is returned as a match. The algorithm was tested on 26 faces and 96.2% of the people were classified correctly. The automatic recognition of feature points is a positive thing but it is doubtful if higher rates are possible using only a limited set of feature points.

Similar work has been done by other groups on developing 3D models for the face, but contrary to the 3D morphable model mentioned above, Ansari *et al.* [4, 5] do not use the generic model as a “bridge”. In other words the model is not employed in order to generate multiple 2D images across many views and illuminations to assist with 2D recognition of images. Instead, it is used for comparisons to other 3D faces.

Lu *et al.* [129] attempted to develop a face recognition system that is also more robust to differences in lighting and facial appearance. For that purpose a model is constructed, integrating several scans from different viewpoints (Figure 2.24). Keeping a more “complete” model in the face database allows them to perform more thorough comparisons with a sensed face that needs identification. This technique integrates shape matching with a constrained appearance-based method. Because 3D sensed models from various view points are used and fully automatic registration is problematic with dramatically different surfaces, landmark points are used to help the initial alignment. Further improvements in surface matching are achieved using a hybrid ICP implementation and a mean

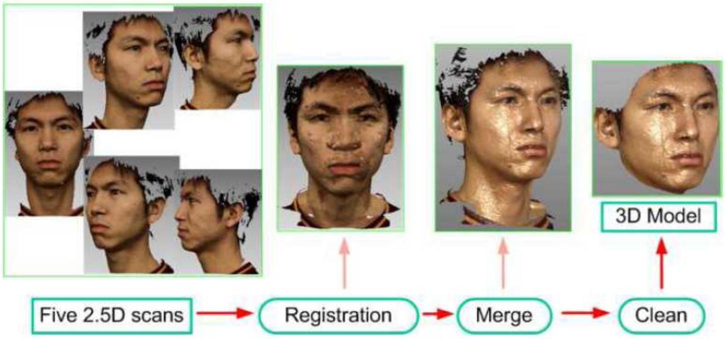


Figure 2.24: Using different surface scans to create a more complete gallery face scan (from [129]).

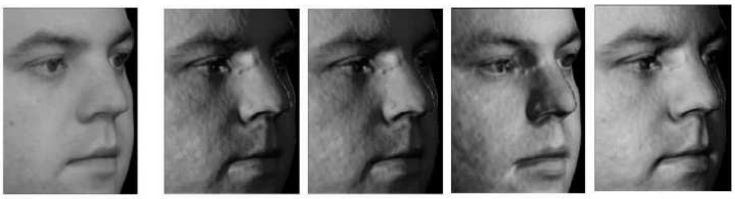


Figure 2.25: Cropped synthesized training samples for discriminant subspace analysis. Some of the images synthesized were ones with lighting changes in order to make the recognition more robust (from [129]).

point-to-point distance is used as a similarity metric. The top 30 matches from the surface classification are retained and the 3D model is utilized to synthesize training samples with facial appearance variations on which a discriminant subspace analysis is performed (Figure 2.25). The two scores of the matching components, the ICP and LDA score, are combined in a weighted sum to make the final decision. A 98% rank 1 rate is reported for the combined metric, but this drops to 91% when faces with emotional expressions are included in the population. The authors are exploring the possibility of using AAMs in order to model variation due to expression and aging and not just illumination.

Passalis *et al.* [152] described another approach for 3D face recognition using an annotated face model (AFM). The model is built based on an average 3D mesh constructed using statistical data. Anatomical landmarks are then placed on its vertices. Different areas of the face are annotated as seen in Figure 2.26. The AFM is registered to each input mesh and fitted onto it by using ICP for the global registration and solving a *finite element method* approximation to compute the local deformation. A deformation image

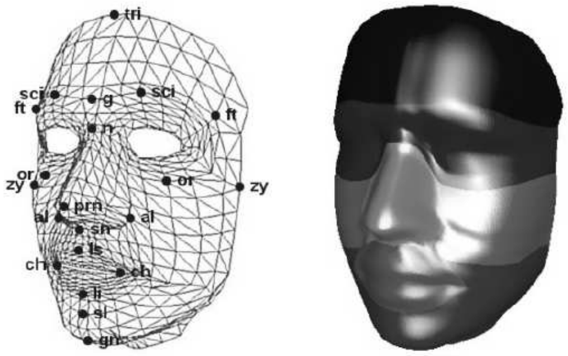


Figure 2.26: The AFM model with anatomical landmarks (left) and the segmented annotated areas (right)from [152].

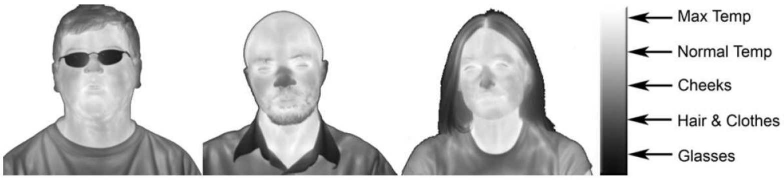


Figure 2.27: Using infra-red information in conjunction to 2D/3D (from [105]).

encoding the shape information of the AFM is generated and compressed using wavelet transform. The metadata extracted from it constitutes its biometric signature which during enrollment is stored in the database. During testing, a gallery signature is compared to a probe one. Using 446 subjects in the database and 3,541 different data of the same subjects to probe the database, a 90% rank 1 rate is achieved. Kakadiaris *et al.* [105] used the same model in combination with information from both visible spectrum and thermal infrared sensors in a combined metric and it was possible to further improve the scores (Figure 2.27). Others researchers have also used thermal data in addition to 2D and/or 3D data but are not explicitly reviewed in this chapter as its focus is on different techniques, rather than new modalities.

In Blanz *et al.* [18] both the gallery and the probe set consist of intensity images and the 3D morphable model, as discussed above, assists in 2D-to-2D matching. In Lu *et al.* [129] and Passalis *et al.* [152] the gallery and probes are 3D images and recognition involves explicit 3D-to-3D matching. Yin and Yourst [223] explored a scenario that is somewhere in between. In this work the comparison between faces is taking place in 3D

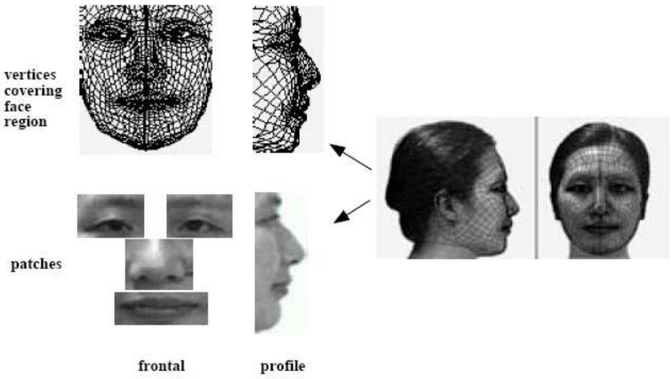


Figure 2.28: The 3D model and the texture patches used for comparisons from[223].

space having extracted the necessary geometry from frontal and profile view 2D intensity images and videos. This technique is fully automatic and works by initially using profile face analysis to recover feature points and obtain the curve of the facial outline. This helps the locating of the specific features in the next step as well as to adjust for pose and orientation. The facial features are then identified and their shape is estimated and modeled by the generic 3D model. The location of the 3D vertices are then used in combination with a weighted difference between texture patches around facial features. The authors report a rank 3 rate of 91.2%. No subspace analysis is carried out.

2.5 Conclusions

2.5.1 Issues in comparative evaluation

The comparison of different techniques based on their evaluation scores is very problematic and perilous to say the least. A major reason for that is that most researchers, report experimental results based on different datasets. The size of the datasets varies a lot and as Xu *et al.* [219] demonstrated it can significantly affect the performance of an algorithm. For example, 96.1% rank 1 recognition rate was reached using a 30-person dataset but that decreased to 72.4% when 120 datasets were used. Other researchers have reported similar but less dramatic decreases in performance [40]. It is also for that reason that early experiments in the field reported 100% rank 1 rates. Nowadays, with the increasing

Table 2.1: Overview Of Techniques

Method	Modality	Reference	Number of subjects	Dataset size	Core matching algorithm	Reported performance
Surface-based Approaches						
Local Methods						
EGI	3D	Lee & Milios[115]	6	6	Correlation	N/A
Feature Vector	3D	Gordon[75]	26 train 8 test	26 train 24 test	Closest vector	80-100%
Feature Vector	3D	Moreno *et al.*[141]	60	420	Closest vector	78%
Feature Vector	3D	Lee *et al.*[118]	100	200	SVM	96%
Point set	3D	Chua *et al.*[44]	6	24	Point signature	100%
Feature Vector	2D+3D	Wang *et al.*[213]	50	300	SVM, DDAG	$>$ 90%
Point set +feature vector	3D	Xu *et al.*[219]	30 / 120	720	Min. distance	96% / 72%
Global Methods						
Profile+surface	3D	Cartoux *et al.*[38]	5	18	Min. distance	100%
EGI	3D	Tanaka *et al.*[196]	37	37	Correlation	100%
EGI	3D	Wong *et al.*[217]	5	n/a	Min. Distance +Evolutionary optimization	80.08%
Point set	3D	Ackermann & Bunke[1]	24	240	Hausdorff distance	100%
Point set / range image	3D	Pan *et al.*[150]	30	360	Hausdorff / PCA	3-5%EER / 5-7%EER
Range+curvature	3D	Lee & Shim[117]	42	84	Weighted Hausforff	98%
Point set	3D+2D	Lu *et al.*[127]	10	63	ICP	96%
Point set	3D+2D	Lu & Jain[128]	100	196 probes	ICP+TPS	91%
Point set	3D	Medioni *et al.*[135]	100	700	ICP	91%
Surface mesh	3D+2D	Maurer *et al.*[133]	466	4,007	ICP+Neven	87% verification at 0.01 FAR
Multiple profiles	3D	Nagamine *et al.*[142]	16	160	Closest vector	100%
Multiple profiles	3D+2D	Beumier & Acheroy[14]	27 gallery, 29 probes	81 gallery, 87 probes	Min. distance	1.4% EER
Multiple profiles	3D	Wu *et al.*[218]	30	90	Min. distance	1.1-5.5% EER
Statistical Approaches						
Range images	3D+2D	Tsalakanidou *et al.*[203]	40	80	PCA	99% 3D+2D / 93% 3D only
Range images	3D+2D	Tsalakanidou *et al.*[202]	50	3,000	EHMM	4% EER
Range images	3D	Hesher *et al.*[89]	37	222	PCA	90%
Range images	3D	Chang *et al.*[39]	200 (275 train)	951	PCA	99% 3D+2D / 93% 3D only
Various	3D	Gökberk *et al.*[71]	106	579	Various	99%
Point set	3D+2D	Bronstein *et al.*[27],	30	220	“canonical forms”	100%
“Isomorphic” range image	3D	Pan *et al.*[149]	276	943	PCA	95%, 3% EER
Model-based Approaches						
2D for testing, 3D for training	2D+3D	Blanz *et al.*[18]	68	4,420	3D Morphable Model	92.8% when correctly fit
2D for testing, 3D for training	2D+3D	Huang *et al.*[94]	10	200	Component-based 3D Morphable Model	88%
Feature points extr. from 2D	3D	Ansari *et al.*[4, 5]	26	104	Generic model	96%
Point set	3D+2D	Lu *et al.*[129]	100	598	ICP+LDA	96%
2D probes, 3D gallery	3D+2D	Yin & Yourst[223]	60	240	Flexible model	91.2% rank 3
Surface mesh	3D	Passalis *et al.*[152]	446	4,007	Deformable model	90%

availability of 3D data it is expected that these scores would drop.

The actual type and size of data is also different from publication to publication. Some range images have a high number of points, such as the ones taken with laser scanners while others, like the ones taken with structured light, are particularly poor compared to the former. When it comes to multimodal techniques, the 2D data used may be color or greylevel images. Another factor is the heterogeneity of the data. Some report results based on neutral frontal images, which is the best case scenario in face recognition. Others, on the other hand, use databases containing faces under varying illumination, with facial expressions, with glasses, facial hair and even collected over a period of months. It is difficult therefore to compare a technique, which uses data with a lot of extraneous variables, to a technique which uses a more limited data collection protocol. Additionally, some databases contain more than one sample image per subject, which can affect the evaluation scores. As previously stated that generally improves the ability of the algorithm to identify faces especially when one exploits the relationship between the intra- and inter-subject variability. Moreover, the statistics reported by many researchers can also inhibit any attempt to compare methodologies. Most report rank 1 rate while others report rank 5 rate, which are impossible to compare. As previously stated, it is known that a rank 5 score can be dramatically different from a rank 1. Finally, differences in experimental design make a comparative evaluation of current techniques particularly difficult. For example, some groups separate the data into a training set and a testing set while others use the same population for both stages. A few researchers, such as Gökberk *et al.* [71] conducted experiments using many available techniques under identical conditions in order to be able to safely draw conclusions based on the strengths of each technique.

Differences in the dataset and experimental setup may also be the reason that researchers report contradictory findings. Godil *et al.* [69] used PCA on 3D and 2D data and found that a combined metric performs significantly better than any single modality, but it was noted that the 2D modality performs slightly better than 3D. Chang *et al.* [41] reach a similar conclusion, which contradicts other publications [39, 133]. The improved score of the combined modalities is also questionable as it might be a result of using two image samples to represent each person. Chang *et al.* again tried to study the effect of using two image samples. It was found that using PCA with a 3D scan and a 2D image combined, yields 95% rank 1 rate while 3D and 2D alone yield 89% and 91% respectively [41]. When the same experiment was conducted using two 2D images a 93% rank 1 rate was reached which suggests that at least half of the improved performance is due to having more than one sample per person irrespective of its modality (3D scan or 2D intensity image). The standardized evaluations discussed in Section 2.5.2 are an attempt to deal with these differences in experimental design and conditions in order to enable comparisons between various techniques. The literature examined seems to be divided between proponents of 2D and 3D analysis. Some researchers believe that 3D data is presently a more powerful discriminant [39, 133] while others believe the opposite [202, 95]. The balance of opinions might change as more time and money is invested in developing more acccurate and affordable 3D sensors and the use of 3D data becomes more widespread.

2.5.2 The FERET and FRVT 2002 evaluations

There is a need to compare the strength of each technique in a controlled setting where they would be subjected to the same evaluation protocol on a large dataset. This need for objective evaluation prompted the design of the FERET and FRVT 2002 evaluation protocols (and the upcoming FRVT 2006) [159, 160]. Both protocols followed the principles of biometric evaluation laid down by Phillips *et al.* [160].

A requirement of the latter is that all evaluations are to be designed and administered by people that have no affiliation to the participants being tested. This ensures that the test is not going to favor one participant over another. Secondly, the data on which the participants are going to be benchmarked on, is not disclosed before the evaluation. If the participants are tested on known data there is the risk that the algorithms will be tuned to a particular dataset. Thirdly, the details of the evaluation test design, protocol, methodology and a representative example of the test data have to be published. This will allow other groups to assess and repeat the evaluation in their own setting. Finally, the evaluation should not be too hard or too easy. Very high or very low recognition results will make the assessment of the capabilities of the systems and the comparison between them very difficult. In an ideal evaluation the performance scores are spread widely so that the strengths and weaknesses of each approach can be observed. By seeing differences in performance of each algorithm across experiments one can assess which problems are solved and which still pose challenges. In order to achieve a good spread in the performance, different levels of image difficulty are presented as input. Faces included in the aforementioned protocols contained variations in illumination, location with respect to camera, different levels of background complexity and some were taken in sessions on different dates.

Additionally, there are three rules governing the evaluation protocols. The input to the algorithms is separated in two groups. The target set T and the query set Q. The galleries and probe sets are constructed from these two groups and according to the first rule, all the similarities between each face t_i in the target set and q_j in the probe must be calculated and returned in a similarity matrix. This way detailed statistics can be computed for all algorithms. Furthermore, multiple biometric samples of each person are placed in both the target and query set and each sample is unique. Finally, training must be completed before the evaluation in order to ensure that each algorithm does not have gallery-specific information during the evaluation. The FERET evaluations were conducted three times, in 1994, 1995 and 1996 on a database of faces collected in 15 sessions between 1993 and 1996. Each session lasted between one and two days and the setup did not change within each session. Each subject was photographed in sets containing 5-11 images which included two frontal views f_a and f_b. The latter also contained a different facial expression. For 200 individuals a third frontal set was collected f_c and the remaining sets were non-frontal images, such as full, half and quarter profiles. By 1996 the database contained 1,564 sets of images totaling 14,126. That meant that it consists of 1,199 individuals and 365 duplicates. Duplicates were images taken on different days. The algorithms that performed the best were a probabilistic PCA, a subspace LDA and an elastic branch graph algorithm. Figure 2.29 shows the verification results for all the participating algorithms.

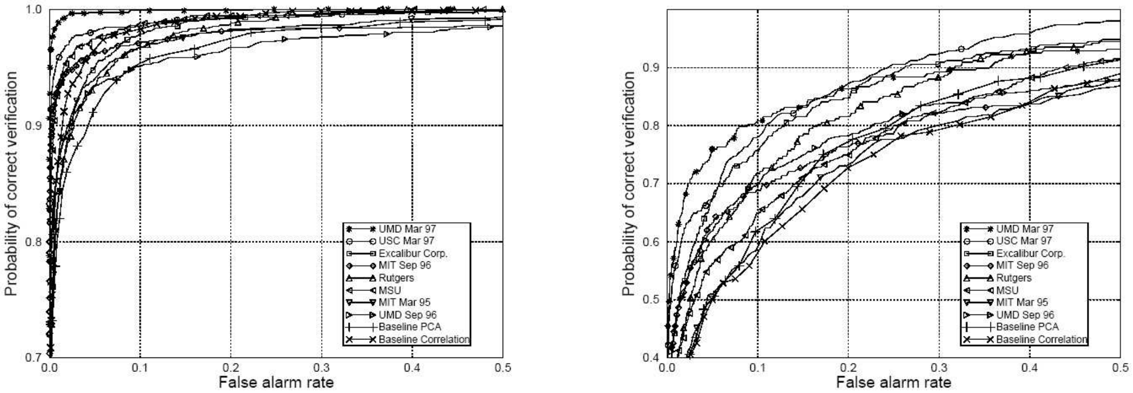


Figure 2.29: The effect of time lapse between data collection sessions on the verification rate (FERET database). The graph on the left shows the verification rates of various techniques with images taken on the same day. The graph on the right shows a clear decrease in the verification rates because duplicate images of subjects were collected on a different day (from [167]).

It also demonstrates the great differences in performance between images taken on the same session and images taken during different sessions.

The FRVT 2002 evaluation was build on top of the FERET and FRVT that preceded it. By the the year 2000 there were a few commercially available systems and there was a need to evaluate them in a controlled environment. The FRVT 2002 provided a demanding setting simulating a real world scenario with a very large database of images of 37,437 people totaling 121,589 images for testing commercial off-the-shelf (COTS) systems. Performances measured were closed- and open-set identification as well as verification. A second set of images tested the ability of the systems to perform face recognition tasks across a wide rage of image types such as images taken outside, non-frontal images etc. Changes in indoor lighting were shown to have little effect on the rates achieved with the best system reporting 90% verification rate with 1% false acceptance rate. On the other hand, the best rate achieved using outdoor images at 1% false acceptance rate was only 50%. Furthermore, using images taken over many years it was found that the best systems performed 5% worse for every year between the images sessions. Another point that was investigated is how the database size affects the rank 1 rates. The best system, once again, achieved 85% using a database of 800 people, 83% on 1,600 and 73% on all

37,437 images. During the FRVT 2002 it was possible for the first time to investigate the effects of different demographics on the recognition rates. It was found that men were more easily identified than women (by 6%-9%) and so were older people, showing that for every 10 years of age there is an increase of 5% in the recognition rate until the age of 63. Finally, FRVT 2002 also investigated the effectiveness of the 3D morphable model technique of Blanz and Vetter for generating synthetic views of faces and it was found to significantly improve the recognition rates.

The availability of the above evaluation techniques has had a significant impact on the development of face recognition technology [156]. Apart from providing an objective benchmark for comparing different methodologies the FERET and FRVT 2002 generated a plethora of new research questions that need to be addressed. Why are men more easily recognized than women and why are younger people more difficult than older people? Why did indoor lighting not affect the participants in FRVT 2002 significantly but outdoor illumination changes did? How can one model the effects of age in order to provide a more accurate face recognition system and avoid the rate reduction when images are taken in sessions more than a year apart?

In terms of 2D face recognition the state of the art techniques seem to focus on subspace analysis of the population. Eigenfaces, Fischerfaces, Support Vector Machines and other subspace techniques seem to dominate the field. The rates today for frontal images seem to range between $80-95\%$ when reported by independent sources [190, 159, 160] (on a small database). Similar techniques that perform well in 2D are been used in 3D as well, achieving similar rates. However, other techniques such as the ones dealing with surface-based events (point signatures etc.) have also been employed with great success. When the author initially started working on this technology many researchers were still employing geometric techniques for recognition. These techniques as previously discussed lack the scalability that is required in such endeavours that PCA-based techniques offer and as a results the focus has shifted. Despite some impressive steps having being made face recognition technology is not mature enough to be employed in critical settings as the iris recognition systems that have been built and are being installed in airports worldwide. This might change dramatically in a few years. In the next chapters we are going to investigate both geometric and the more popular PCA-based techniques mirroring the general progress of 3D face recognition.

Chapter 3

Review of Surface Registration

Generally speaking, the purpose of a registration is to find the optimal transformation $\boldsymbol{T}$ that will map the points $\boldsymbol{a}$ in one coordinate system to the *corresponding* points $\boldsymbol{b}$ of another such that:

$$\boldsymbol{b} = \boldsymbol{T}(\boldsymbol{a}) \tag{3.1}$$

where $\boldsymbol{a} = (a_x, a_y, a_z)$ and $\boldsymbol{b} = (b_x, b_y, b_z)$ are points in the two coordinate systems which correspond to each other based on some similarity metric. A two-dimensional illustration is shown in Figure 3.1. There are many types of registration techniques and the choice of the appropriate technique depends on the type of surface used, which is often linked to the way it was captured.

Before comparing and interpreting the information that exist on two images or surfaces, one must align them so that anatomical points in one object are related to points corresponding to the same anatomical location on the second object. Comparing homologous areas is a core requirement in many applications. In medical imaging, registration can be used to align similar structures and fuse the information from different modalities. In face recognition, aligning facial surfaces is of tantamount importance as faces need to be registered in order to compare their features. As already discussed, the face of the same person under two different poses can vary so much that it makes its identification impossible. In that case, a rigid alignment can correct for pose variation. A face can also change in non-rigid ways, with a facial expression. This can be compensated for by using

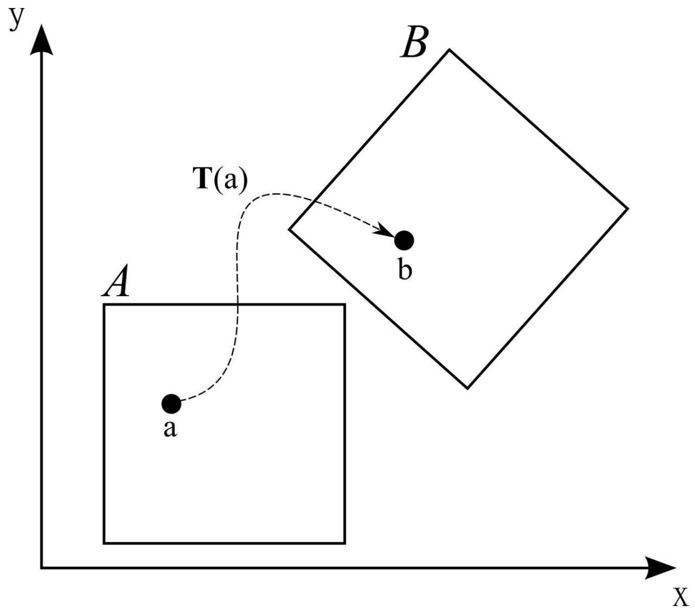


Figure 3.1: The transformation $\boldsymbol{T}(a)$ transforms a point a in image A into its corresponding location in image B.

a non-rigid registration to align corresponding areas to each other.

Any registration technique can be split into two principal design components: The type of transformation used for aligning the surfaces and the similarity metric associated with the specific surface representation. The next sections describe these components in more detail. Within the section about surface representation and similarity metrics, various techniques for optimizing the search for corresponding points, as well as the search for the optimal transformation in the parametric space, are also briefly discussed.

3.1 Transformation types

Transformations can be classified into two different types; those that preserve the straightness of lines and others that do not. Rigid and affine transformations, which preserve the straightness of lines, are more appropriate for objects that do not deform, such as bones. Non-affine transformations, which do not preserve the straightness of lines, are usually reserved for objects that can deform, such as the heart or the face.

3.1.1 Rigid transformation

A rigid transformation is a geometrical transformation, which when applied to an object, for example a surface $A = \{\boldsymbol{a}_i\}$, it maintains all the distances and internal angles. It can be formally expressed as a combination of rotation $\boldsymbol{R}$ and translation $\boldsymbol{t}$:

$$\boldsymbol{b} = \boldsymbol{R}_{AB}\boldsymbol{a} + \boldsymbol{t}_{AB} \tag{3.2}$$

and has six degrees of freedom. In matrix form it can be described as:

$$\mathbf{T}_{rigid}(a_x, a_y, a_z) = \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{3.3}$$

where $\boldsymbol{R}_{AB} = \{r_{ij}\}$, $i, j \in \{1, 2, 3\}$ is a 3×3 orthogonal rotation matrix describing the rotational component of the transformation, and $\boldsymbol{t}_{AB} = [t_x, t_y, t_z]^T$, which is a displacement vector describing the translational component of the transformation.

3.1.1.1 Affine transformation

A more general class of transformations is the affine transformation which is expressed by:

$$\boldsymbol{b} = \boldsymbol{A}_{AB}\boldsymbol{a} + \boldsymbol{t}_{AB} \tag{3.4}$$

where matrix $\boldsymbol{A}_{AB}$ is a rotation matrix. Lengths and angles are not preserved, but parallel lines are. In matrix form it is represented as:

$$\mathbf{T}_{affine}(a_x, a_y, a_z) = \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix} \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{3.5}$$

where $\boldsymbol{A} = \{\alpha_{ij}\}$, $i, j \in \{1, 2, 3\}$ is a 3×3 matrix describing the scale, shear and rotation components of the transformation while $\boldsymbol{t}$ is a displacement vector describing

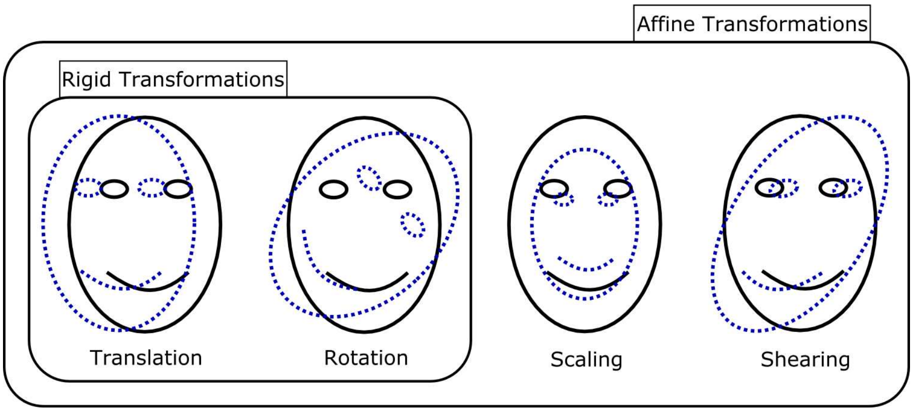


Figure 3.2: An example of a linear transformation.

the translational component. Figure 3.2 shows schematic examples of rigid and affine transformations.

3.1.2 Non-affine transformations

Sometimes the ideal mapping between two images is not affine and therefore a non-affine deformation is needed, as global functions are not adequate to capture local deformations. For example, in the case of faces, an affine transformation would be unable to model (or simulate) on a neutral face the local deformations caused by a smiling face and a non-affine transformation would be more appropriate to use (Figure 3.3). In practice, the transformation is often defined by so-called control points or landmarks and the deformation is smoothly interpolated at intermediate points. Spline-based transformations have, at the local level, a simple form, yet they maintain their global flexibility and smoothness. For a more detailed look into how splines are used in computer graphics and geometric modeling see [7].

Originally, splines were devised to be used for aircraft modeling where engineers were using long, flexible strips of metal or wood which they would deform with the use of weights at selected points. A set of control points, need to be identified where spline-based transformations approximate (or interpolate) the required displacement to map one set of control points to the other while providing a smooth displacement field between

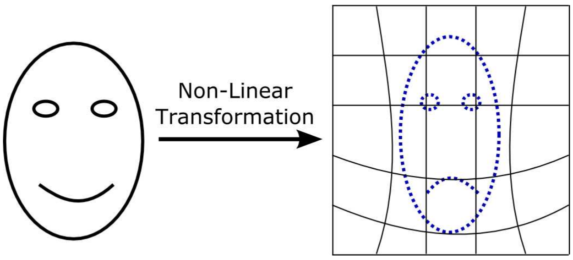


Figure 3.3: An example of a non-linear transformation.

them. The interpolation condition can be written as:

$$\mathbf{T}(\boldsymbol{a}_i) = \boldsymbol{b}_i \qquad i = 1, ..., n \tag{3.6}$$

where $\boldsymbol{a}_i$ represents the location of a control point in the target surface and $\boldsymbol{b}_i$ the location of the corresponding control point in the source surface.

3.1.2.1 Thin-plate splines

Thin-plane splines (TPS) are a family of splines based on radial-basis functions. They were originally formulated by Duchon [52] and Meinguet [136] for surface interpolation of scattered data and have been used extensively in registration [76, 23, 22]. Given a set of n landmarks, they can be defined as a combination of n radial basis functions $\theta(s)$:

$$t(x, y, z) = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 z + \sum_{j=1}^{n} \beta_j \theta(|\phi_j - (x, y, z)|) \tag{3.7}$$

Defining the transformation in three separate thin-plate spline functions $\mathbf{T} = (t_1, t_2, t_3)^T$ returns a mapping between the two surfaces where the coefficients α describe the affine part of the spline-based transformation and the coefficients β describe the non-affine part of the transformation. The interpolation conditions in eq. 3.6 form a set of $3n$ linear equations and in order to determine the $3(n + 4)$ coefficients uniquely, twelve additional equations are needed. These guarantee that the non-affine coefficients β sum to zero and that their crossproducts with the x, y, z coordinates of the control points are likewise zero. The crossproduct is required to ensure that the second part of equation 3.7 contains only non-affine transformations. This can be expressed as:

$$\begin{pmatrix} \boldsymbol{\Theta} & \boldsymbol{\Phi} \\ \boldsymbol{\Phi}^{\mathbf{T}} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Phi}' \\ 0 \end{pmatrix} \tag{3.8}$$

Here $\boldsymbol{\beta}$ is a $n \times 3$ matrix of the non-affine coefficients β, $\boldsymbol{\alpha}$ is a 4×3 matrix of the affine coefficients α, $\boldsymbol{\Theta}$ is the kernel matrix with $\Theta_{ij} = \theta(|\phi_i - \phi_j|)$ and $\boldsymbol{\Phi}$ is the control point matrix of the transformation who's i-th row is $(1, x_i, y_i, z_i)$. The solution for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is a thin-plate spline transformation which interpolates the displacements at the control points. The radial basis functions of thin-plate splines is defined as:

$$\theta(s) = \begin{cases} |s|^2 \, log(|s|), & \text{in 2D} \\ |s|\,, & \text{in 3D} \end{cases} \tag{3.9}$$

3.1.2.2 Free-form deformations

TPS are based on radial basis functions which have infinite support and therefore each control point has a global effect on the entire transformation. The global influence of the control points is undesirable since it has the potential of making the modeling of local transformations difficult. Furthermore, the TPS calculation is relatively inefficient and given a large number of control points, it can be prohibitive. Instead of TPS, *free-form deformations* (FFDs) [183] can be used which have local control. Contrary to thin-plate spline functions which can handle an arbitrary configuration of control points, FFDs define the displacements on a regular mesh.

The spatial domain of the surface or points to be deformed is defined as follows: $\Omega_I = \{(x, y, z) \mid 0 \leq\ x < X, 0 \leq\ y < Y, 0 \leq\ z < Z\}$ where Φ denotes a $n_x \times n_y \times n_z$ grid of control points $\phi_{i,j,k}$ with uniform spacing δ. The displacement field defined by the FFD can be expressed, in this case, as the 3D tensor product of 1D cubic *B-splines* [63]:

$$\mathbf{T}_{local}(x, y, z) = \sum_{l=0}^{3} \sum_{m=0}^{3} \sum_{n=0}^{3} B_l(u) B_m(v) B_n(w) \phi_{i+l, j+m, k+n} \tag{3.10}$$

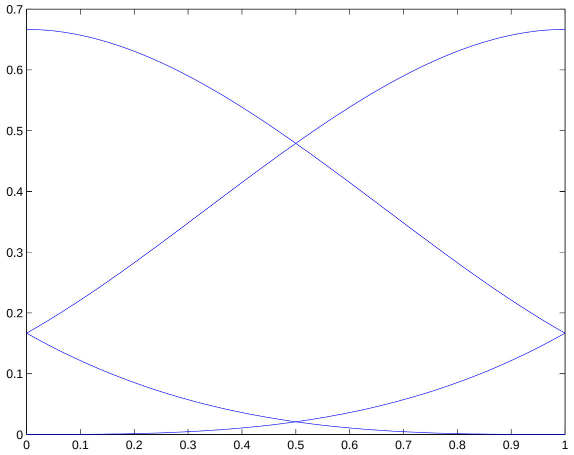


Figure 3.4: Graphical representation of B-Splines.

where $i = \lfloor\frac{x}{\delta}\rfloor - 1, j = \lfloor\frac{y}{\delta}\rfloor - 1, k = \lfloor\frac{z}{\delta}\rfloor - 1, u = \frac{x}{\delta} - \lfloor\frac{x}{\delta}\rfloor, v = \frac{y}{\delta} - \lfloor\frac{y}{\delta}\rfloor, w = \frac{z}{\delta} - \lfloor\frac{z}{\delta}\rfloor$ and where B_l represents the l-th basis function of the B-Spline (Figure 3.4):

$$\begin{aligned} B_0(u) &= (1-u)^3/6 \\ B_1(u) &= (3u^3 - 6u^2 + 4)/6 \\ B_2(u) &= (-3u^3 + 3u^2 + 3u + 1)/6 \\ B_3(u) &= u^3/6 \end{aligned}$$

The basis functions of cubic B-Splines have limited support and therefore changing a control point in the grid affects only a 4×4×4 region around that control point.

3.2 Surface similarity

The second choice one needs to make in the design of a registration algorithm is the type of surface representation to select and consequently, what similarity criterion to use. The similarity criterion should be one that allows for the unimportant differences between the two surfaces to be ignored while the important similarities to be used for aligning them to each other. In other words, the similarity measurement should discriminate between homologous points optimally to allow correct (and efficient) registrations. There are generally four techniques for representing surfaces and surface similarity in registration: feature-, point- and model-based and global similarity-based. The choice of technique depends primarily on the type of data used and the type of transformation that needs to be computed.

3.2.1 Feature-based methods

In feature-based representations the surface is pre-processed in order to extract a subset of the total information, using a discriminating measurement, which provides a more compact description of the surface. The idea behind this approach is to find points on the surfaces that match to each other and compute the distance between them. These matched features are usually used for calculating rigid transformations to bring the entire surface into alignment. Significant effort has been put into eliminating incorrect matches [180]. The features typically used for this representation are point features, curves and regions.

Feature points are loci of geometric significance such as peaks or pits. Thirion [199] computed the extrema of principal curvatures which were matched with the corresponding ones on the other surface by finding points which, among other similarities, had the same sign of principal curvature and similar curvature values. Goldgof *et al.* [73], on the other hand, used the extremum of the Gaussian curvature K, which they located by thresholding the value K.

When it comes to the registration of facial surfaces, feature-based representations have been used extensively, not only as a descriptor but as a way of establishing correspondence to align the faces before full-surface comparisons. Nagamine *et al.* [142] used various heuristics to extract five feature points from the human face, assuming that the faces were frontal. More interestingly, Lu *et al.* [127] calculated a shape index for each point $\boldsymbol{a}$ on the facial surface using the maximum and minimum local curvature, κ_1 and κ_2 as in:

$$ShapeIndex(\boldsymbol{a}) = \frac{1}{2} - \frac{1}{\pi} tan^{-1} \frac{\kappa_1(\boldsymbol{a}) + \kappa_2(\boldsymbol{a})}{\kappa_1(\boldsymbol{a}) - \kappa_2(\boldsymbol{a})} \tag{3.11}$$

in order to find an initial bootstrapping alignment for the surfaces, before using ICP (dis-

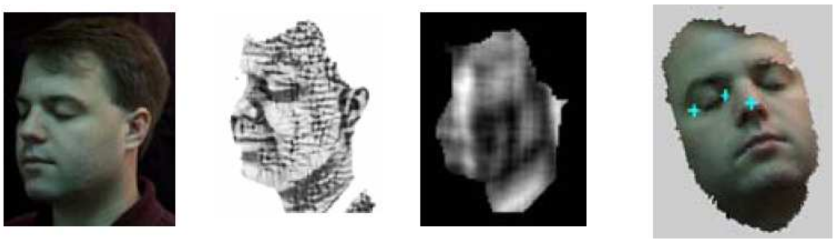


Figure 3.5: Finding feature points using local curvature information of the face. (a) the texture image (b) the shape index (c) after averaging mask is applied (d) the points located (from [127]).

cussed in 3.2.2) to more finely adjust the registration. Given this coordinate-independent way of locating features, features that were on faces with different poses could be aligned. They located an easily identifiable extreme point, the inside corner of the eye, and from there they began their search for nearby feature points such as the nose. Figure 3.5 shows the curvature information used for locating the features. Alternatively, Chua and Jarvis [44] used point signatures (described in Chapter 2) to register surfaces.

The second type of feature corresponds to continuous lines or curves, typically consisting of differential structures such as ridges or region boundaries and they are less compact than feature points in describing the surface shape. Of particular interest in medical imaging is the detection of ridges which are usually defined as long, narrow, raised strips of points on the surface. Monga and Benayoun [140] searched for a contiguous set of loci on a surface where the largest principal curvature $\kappa_{\mathbf{1}}$ is locally maximal. These loci correspond to the zero-crossings of the extremality function $e_1 = \nabla\kappa_{\mathbf{1}} \cdot \mathbf{t_1}$ where $\nabla\kappa_{\mathbf{1}}$ is the directional derivative of the largest principal curvature and $\mathbf{t_1}$ is the principal direction that corresponds to $\kappa_{\mathbf{1}}$.

Early in face recognition research, Cartoux *et al.* [38] segmented the face based on the principal curvature, which helped them establish a plane of bilateral symmetry which they used to normalize for pose before calculating the similarity between the faces. Years later, Hesher *et al.* [89] used a very simple technique which involved locating the nose tip and then the bridge of the nose for aligning the faces. They first locate the nose tip, as the most protruding point of a straight-looking face and then they search all the points

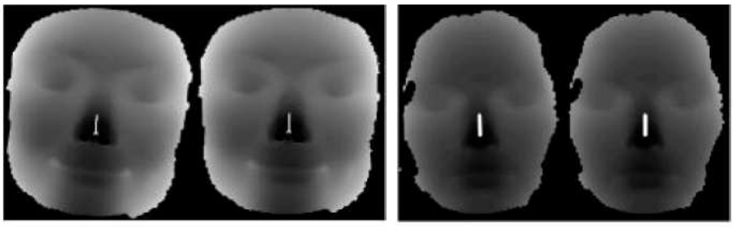


Figure 3.6: Using the nose bridge for registration (from [89]).

above the tip to locate the points that belong to a ridge, effectively locating the ones on the bridge of the nose. Registration of the surfaces involves a simple translation on the 2D plane to bring the nose tips into alignment. The rotational component was calculated by feeding the nose bridge points to a line-fitting algorithm which returned the appropriate rotation. Finally, depth adjustments are made by adding a constant to all pixels so that the tip of the nose pixel has the same value across all images. Figure 3.6 shows the depth images they used. Notice the highlighted feature line on the nose bridge that was used to correct the pose of the faces.

Finally, regions are areas that have some common characteristics such as consistent curvature sign or are surrounded by some boundary and they are denser surface descriptors than curves and points. Regions are a natural descriptor for surfaces and they allow for the characterization of a surface as an adjacency graph. Matching can then be performed by looking for the regions that have most compatible subgraphs in common. This way, matching is performed not just on the characteristics of each area but also on the extended neighborhood of each region. Besl and Jain [11, 107] employ the sign of the mean H and Gaussian K curvature (K/H) in order to measure homogeneity, and use it to identify and segment regions of interest. They identify surface types with which they could characterize patches of the surface such as elliptic and outward bulging, elliptic and inwardly bulging, planar, hyperbolic, saddle shaped etc. and they used them to segment structures in medical images. A very similar approach was adopted by Moreno *et al.* [141] to identify and segment regions of the face. In Figure 3.7 the brighter points are those where the curvature value is positive while the darker patches is where the curvature is negative. With the appropriate thresholding they managed to locate the features they

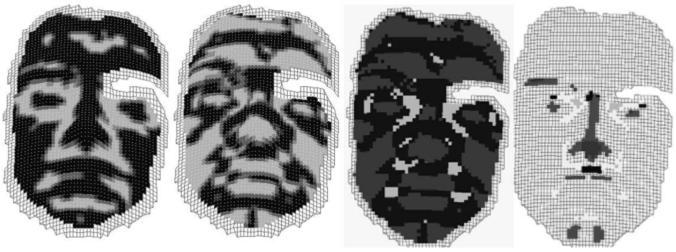


Figure 3.7: (a) The sign of the median and (b) Gaussian curvature. Point HK classification (c) before and (d) after thresholding (from [141]).

needed.

Based on the surface representation and transformation type chosen, one needs to search for the optimal matches between two surfaces. This generally happens either by successively comparing a set of candidates or by iteratively minimizing an objective function. As discussed earlier, because features on a surface can be relatively sparse, matching them generally determines a rigid transformation. It usually involves a comparison of various shape parameters such as curvature at extrema, shape type, etc. One of the early techniques for matching features has been the *generalized Hough transform* which matches similar structures and derives a transformation from each pair. The transformation space is quantized and the matches increment a corresponding cell [193]. Another early technique developed for finding the optimal matches is called *geometric hashing* [112] and it originally stood out for its efficiency. It works by pre-computing local matching information which does not vary with rotations and translations and which is stored in a hash table for each surface. The hash table has entries each of which is associated with a feature to which a local coordinate system (basis) can be assigned. Given a transformation between two bases, the consistency of the mapped non-basis features is evaluated and consistent feature pairs “vote” for the ideal transformation.

3.2.2 Point-based methods

Point-based methods register surfaces by calculating correspondences between a dense set of points making up the surface. Initially, the point sets need to be roughly aligned to each other and subsequently, they are finely registered by iteratively minimizing a distance metric. The latter is usually the sum of squared distances between closest points on the two surfaces.

Using the whole surface rather than just a manually selected array of landmarks or automatically extracted features, can be significantly more computationally intensive, but has many advantages. One of the advantages is the substantial increase in information provided by the surface and the redundancy that it entails. Early face recognition techniques that used only a small selection of points perform poorly on large datasets because there is simply not enough information to distinguish one face from another. Furthermore, manually selected landmarks, contrary to automatically extracted feature-based landmarks, can be very tedious to determine, prone to random error and can make registration procedures more difficult to repeat under the same conditions.

Besl and McKay [12] formulated the *iterative closest point* algorithm (ICP) for registering 3D shapes. Based on two sets of points where closest point correspondence has been established, they calculate the ideal rotation and translation using a quaternion-based method that would bring them into close alignment.

Let $A = \{\boldsymbol{a}_i\}$ be a measured point set (aka source) to be aligned with a model point set (aka target) $B = \{\boldsymbol{b}_i\}$, where $|B| = |A|$ and points on each point set with the same index correspond to each other. The authors first define a distance metric d between an individual source point $\boldsymbol{a}$ and a target (model) shape B:

$$d(\boldsymbol{a}, B) = \min_{\boldsymbol{b} \in B} ||\boldsymbol{b} - \boldsymbol{a}|| \tag{3.12}$$

Using this distance metric they propose looping over all points in A and finding the closest point in B. Let Y denote the resulting set of closest points and $\mathcal{C}$ the closest point operator:

$$Y = \mathcal{C}(A, B) \tag{3.13}$$

The ideal transformation is then calculated using a quaternion-based method and applied to the source set. The process is repeated until a certain threshold is reached.

The unit quaternion is a four vector $\boldsymbol{q}_R = [q_0, q_1, q_2, q_3]^t$ where $q_0 \geq 0$ and ${q_0}^2 + {q_1}^2 + {q_2}^2 + {q_3}^2 = 1$. $\boldsymbol{q}_T = [q_4, q_5, q_6]^t$ defines a translation vector and the complete registration state vector is $\boldsymbol{q} = [\boldsymbol{q}_R|\boldsymbol{q}_T]^t$. The mean square similarity function to be minimized is:

$$f(\boldsymbol{q}) = \frac{1}{|A|}\sum_{i=1}^{|A|} ||\boldsymbol{b}_i - \boldsymbol{R}(\boldsymbol{q}_R)\boldsymbol{a}_i - \boldsymbol{q}_T||^2. \tag{3.14}$$

The center of mass $\overline{\boldsymbol{a}}$ of point set A and the center of mass $\overline{\boldsymbol{b}}$ of point set B can be derived by:

$$\overline{\boldsymbol{a}} = \frac{1}{|A|}\sum_{i=1}^{|A|} \boldsymbol{a}_i \text{ and } \overline{\boldsymbol{b}} = \frac{1}{|B|}\sum_{i=1}^{|B|} \boldsymbol{b}_i. \tag{3.15}$$

The cross-covariance matrix Σ_{ab} of the sets A and B is given by:

$$\Sigma_{ab} = \frac{1}{|A|}\sum_{i=1}^{|A|} [(\boldsymbol{a}_i - \overline{\boldsymbol{a}})(\boldsymbol{b}_i - \overline{\boldsymbol{b}})^t] \tag{3.16}$$

The column vector $\boldsymbol{\Delta} = [A_{23} \;\; A_{31} \;\; A_{12}]^T$ is formed using the cyclic components of the anti-symmetric matrix $\boldsymbol{A}_{ij} = (\Sigma_{ab} - \Sigma_{ab}^T)_{ij}$. This vector is used to form the symmetric 4×4 matrix $\boldsymbol{Q}(\Sigma_{ab})$

$$\boldsymbol{Q}(\Sigma_{ab}) = \begin{pmatrix} tr(\Sigma_{ab}) & \boldsymbol{\Delta}^T \\ \boldsymbol{\Delta} & \Sigma_{ab} + \Sigma_{ab}^T - tr(\Sigma_{ab})\boldsymbol{I}_3 \end{pmatrix} \tag{3.17}$$

where $\boldsymbol{I}_3$ is the 3×3 identity matrix. The optimal rotation is the unit eigenvector $\boldsymbol{q}_R = [q_0, q_1, q_2, q_3]^t$ corresponding to the maximum eigenvalue of the matrix $\boldsymbol{Q}(\Sigma_{ab})$. The optimal translation vector can then be calculated by subtracting the rotated centroid of the source from the centroid of the target as in:

$$\boldsymbol{q}_T = \overline{\boldsymbol{b}} - \boldsymbol{R}(\boldsymbol{q}_R)\overline{\boldsymbol{a}} \tag{3.18}$$

This least squares quaternion operation to calculate the transformation vector **q** is denoted as:

$$(\boldsymbol{q}, d_{ms}) = \mathcal{Q}(A, B) \tag{3.19}$$

where d_{ms} is the mean square point matching error. Listing 1 shows the ICP algorithm more formally.

Listing 1 The Iterative Closest Point algorithm

```
Start with source set $A$ and target set $B$.
Set iteration counter $k = 0$, $A_0 = A$ and $\boldsymbol{q}_0 = [1, 0, 0, 0, 0, 0, 0]^t$.
**repeat**
    **Find** the closest points between $A$ and $B$ by: $Y_k = \mathcal{C}(A_k, B)$
    **Compute** the ideal transformation to align $Y_k$ and $A_0$ by: $(\boldsymbol{q}_k, d_k) = \mathcal{Q}(A_0, Y_k)$.
    **Apply** the transformation: $A_{k+1} = \boldsymbol{q}_k(A_0)$
**until** change in mean square error is smaller than a threshold $\tau$ as in: $d_k - d_{k+1} < \tau$
```

Their technique was proven particularly successful in registering comparable surface patches. A smaller subpatch can also be aligned with a bigger patch by considering several possible initial transformation states and using the one that returns the smallest mean square error. They also propose an accelerated version of ICP which involves a more optimized search for the ideal transformation. There have been since then many suggestions how to speed up the selection of correspondences and how to deal with outliers [12, 204, 215, 70].

Some of the modifications have to do with the selection of points from each surface to build point pairs. For example, instead of using all points on the other surface to compute corresponding pairs, some authors have proposed using a random subset in order to reduce the number of vertices that have to be processed [132]. Others, like Turk and Levoy [204], proposed a uniform subsampling of the surfaces.

Further modifications are related to the searching for the closest point (matching) on the other surface. Finding the closest point can be a very time consuming step. Simon [186] proposed the use of k-D trees [10, 65] as part of the shape alignment. K-D trees are a sequence of bisections in a k-dimensional space. A k-D tree uses planes that are perpendicular to one of the coordinate axis to split the space into two (Figure 3.8). By moving down the tree, one cycles through the axes used to choose the splitting plane.

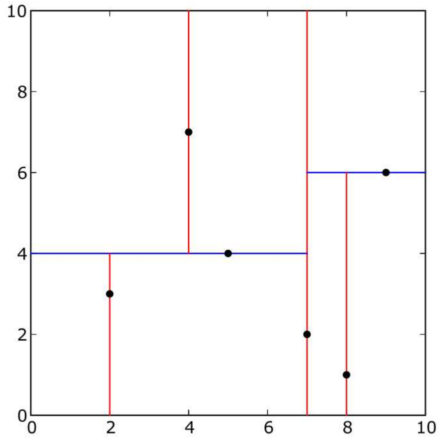


Figure 3.8: A k-D tree representation (from [3]).

At each step, the splitting plane is created on the median of the points being put into the kd-tree (with respect to the point coordinates in the axis being used). The process stops when no more subdivisions can take place and all points have been allocated to a node in the tree. The result of the algorithm is a balanced k-D tree. The search for the closest point in a k-D structure involves a recursive procedure that begins at the root of this tree, traversing it to the leaves until the point has been found, exploiting the structure. Searching a k-D tree has the potential to significantly reduce the time required to locate the closest point from a cloud of points as it reduces dramatically the number of point-to-point comparisons in the closest-point search. Other ways of finding the closest point on the other surface include locating the intersection of the ray originating at the surface point in the direction of the surface normal with the other surface [43]. Another technique, used by Szeliski and Lavallée [195] involves the use of the distance transform. They pre-computed offline the distances and closest points for discrete locations. This technique is discussed in more detail later in this section.

Other modifications involve the weighting of the corresponding pairs appropriately. Godin *et al.* [70] assigned lower weights to pairs with greater point-to-point distances. In the same work they also assigned weights based on the compatibility of normals.

Finally, some have opted for rejecting pairs of points based on some statistics. For example Pulli [164] rejected the worst $n\%$ of pairs based on their point-to-point distance. Masuda *et al.* [132] reject pairs with distances more than 2.5 times the standard deviation.

Rangarajan *et al.* [165] proposed a solution to the registration of surfaces by using a technique called *robust point matching* (RPM) algorithm. Their implementation is particularly robust in simultaneously finding correspondences between the two objects and computing the transformation parameters. Areas that contain cuts and tears in the surfaces often confuse traditional ICP implementations, but are accounted for in RPM. RPM works by first defining a set of corresponding variables $\{\boldsymbol{M}_{jk}\}$, the *match matrix*, such that $\{\boldsymbol{M}_{jk}\} = 1$, if point x_j corresponds to point y_k and $\{\boldsymbol{M}_{jk}\} = 0$, if it does not. Given two sets of points, they search simultaneously for the rotation $\boldsymbol{R}$, translation $\boldsymbol{t}$ and match matrix $\boldsymbol{M}$ that minimizes some distance function.

The ICP algorithm has been used extensively for aligning human faces [151, 133]. It has also been extended to the non-rigid alignment of surfaces. One of the seminal works in non-rigid registration was proposed in Szeliski and Lavallée [195] where a new method for determining the minimal non-rigid deformation was presented. To minimize the similarity function (eq. 3.14) they used B-splines, avoiding the higher order polynomials that tend to introduce artifacts such as oscillations. Their algorithm finds correspondences between points in each surface using the closest-point approach and the matching pairs were subsequently aligned using a B-spline model of deformations. In order to perform fast nonlinear least squares minimization they used the *Levenberg-Marquardt* algorithm (LM) because it works well on uncorrelated noisy measurements with a Gaussian distribution. In order to calculate the optimal transformation LM requires the evaluation of the distance function $d(\boldsymbol{a}, B)$ along with its derivative with respect to all of the unknown parameters. In order to avoid getting stuck in local minima in a high-dimensional parameter space, they initially estimated a simple rigid transformation before estimating more complex deformations at the global level. The global level is modelled with a low-resolution spline and the optimal spline parameters are used to bootstrap the estimates at finer levels. The least squares minimization requires a fast computation of the distance $d(\boldsymbol{a}, B)$ and its gradient.

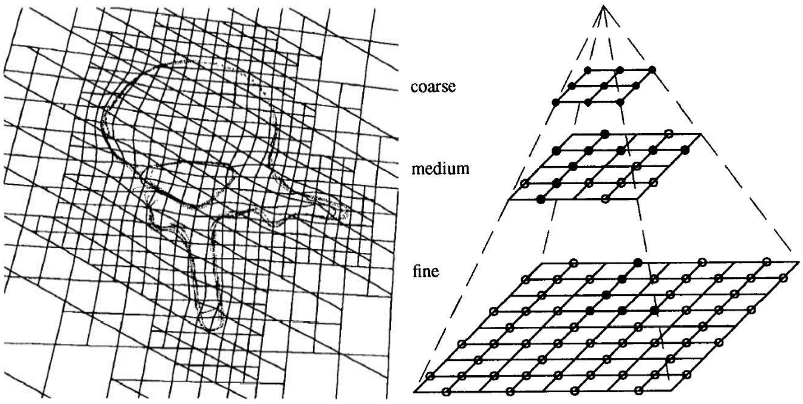


Figure 3.9: A hierarchical octree example (from [195]).

In order to speed this process up, the authors introduced a distance map which stores the distances that were pre-computed offline. A voxel-based volume is set up to contain surface B and for each voxel it stores the identity of the closest point to that surface, as well as the distance from it. Each point in surface A that enters the volume inherits the closest point precomputed for the voxel in which it falls, reducing the closest-point search to a simple table lookup. To further optimize for speed, space and accuracy, they developed a new kind of distance map which they called *octree spline* [114, 34]. The main benefit behind this implementation is that it allows greater accuracy near the surface than far away from it. As the registration improves so does the resolution of the distance map, allowing for initially fast, robust registrations followed by finer adjustments at later stages. More importantly, the octree spline was extended to represent the local deformation by storing the displacements of the B-spline model.

The use of the octree splines as a distance map representation implies that the distance from the surface determines at which resolution the spline coefficients are interpolated (Figure 3.9(a)). Finally, to further accelerate convergence, they used a hierarchical basis representation for the octree spline (Figure 3.9(b)) where displacement values at finer levels are added to the displacements interpolated from the parents and thus all finer levels contain a relative or offset representation. This made convergence not only faster but also made the final rendering significantly smoother. Their technique was used extensively for the registration of medical images.

Other techniques have been proposed that do not rely on the closest-point-based correspondence. For example, Chen and Medioni [43] used a subset of the measured surface A to look for corresponding points in the model surface B. For a point $\boldsymbol{a}$ on a smooth patch of A they located point $\boldsymbol{b}$ which is where the normal of $\boldsymbol{a}$ intersects surface B. Next, they defined a plane tangent to B at point $\boldsymbol{b}$ and they computed the transformation that would minimize the sum of squared distances between the transformed point $\boldsymbol{a}$ and its corresponding tangent plane on B. The authors report that this point to plane distance makes the convergence process less susceptible to local minima than the point-to-point distance of the classic ICP. Lu *et al.* [127] found that the point-to-plane distance reflects the true distance between the two surfaces better, but was significantly slower to calculate. To combine speed with a more robust registration they used a "hybrid ICP" in which they use Besl and McKay's system to calculate a coarse estimation of the alignment and Chen and Medioni's approach for a finer alignment.

Feldmar and Ayache [59, 58] searched for closest points in a feature space rather than Euclidean space. Instead of using the coordinates of a point in 3D, they compared eight coordinates for each point; the x, y, z coordinates, its normal (n_x, n_y, n_z) and its principal curvatures κ_1 and κ_2. They then used a weighted distance function, in order to find a compromise between the components of the feature vector used for registration, to determine the optimal transformation.

There have been various optimizations for accelerating non-rigid surface registration. The minimization of equation 3.14 can be, computationally speaking, a very expensive process, given the degrees of freedom that a transformation can have. Furthermore, there is the risk of running into local minima that do not allow the "correct" alignment of the surfaces. Finding the ideal transformation can be seen as an optimization problem for which various techniques can be used. Szeliski and Lavallée [195], for example, used an unconstrained nonlinear optimization technique. Other techniques that can be used include *steepest descent, conjugate-gradient descent,* etc. [130]. The version of ICP that we used has a few optimizations similar in nature to some of the above. More details on which optimizations were used, which were not used and why are provided in the end of

Section 4.3.1.

3.2.3 Model-based representations

Deformable surface modeling has been used extensively for segmenting medical images to identify a structure in a volume and/or track it in a sequence of volumes. Based on physical or surface evolution expressions, these methods compute the surface motion by modeling virtual forces that manipulate the object rather than trying to explicitly match the two surfaces.

Some researchers have proposed a *finite-element model* (FEM) approach for solving deformable surface models numerically [155, 198]. A 3D shape is thought of as the result of forces acting on a deformable material such as clay. They use the FEM mathematical formulation in order to simulate dynamically changing objects:

$$\boldsymbol{M}\ddot{\boldsymbol{u}} + \boldsymbol{C}\dot{\boldsymbol{u}} + \boldsymbol{K}\boldsymbol{u} = \boldsymbol{R} \tag{3.20}$$

where $\boldsymbol{u}$ is a $3n \times 1$ vector of the displacements of the n nodal points relative to the object's center of mass, while $\boldsymbol{M}$, $\boldsymbol{C}$ and $\boldsymbol{K}$ are matrices describing the mass, damping and material stiffness between each point within the body. $\boldsymbol{R}$ is a $3n \times 1$ vector describing the x, y, z components of the forces acting on the nodes. Using FEM a displacement function can be estimated which can be grafted upon a superquadric ellipsoid which represents the difference between a simple superquadric shape and the final more general shape.

More native to 3D surfaces, the model of Amini and Duncan [2] uses 3D points as the input data and represents all shapes as bending energy from a zero energy flat plane $\epsilon_{be}(\boldsymbol{u}, \boldsymbol{v}) = \kappa_1^2 + \kappa_2^2$. Matching between surfaces becomes the matching of points with similar energy and principal curvature.

Hutton [96] developed a *dense surface model* (DSM) which is a hybrid of ICP and ASM. A DSM is build by running PCA analysis on a series of faces for which dense, closest-point-based correspondence has been established with the help of manually placed

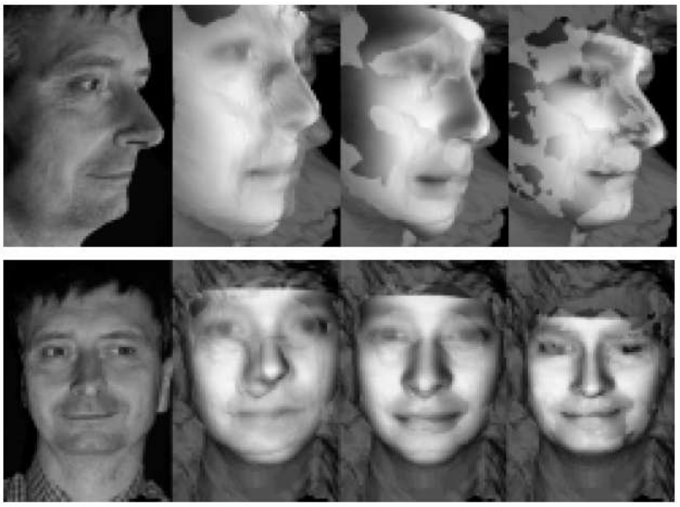


Figure 3.10: The fitting of the DSM to an unseen example (from [96]).

landmarks. Once the model is build a new face instance $\boldsymbol{x}_{new}$ can be generated by:

$$\boldsymbol{x}_{new} = \overline{\boldsymbol{x}} + \boldsymbol{\omega} b \tag{3.21}$$

where, as in the eigenfaces technique, $\boldsymbol{\omega} = [\omega_1|\omega_2|...|\omega_t]$ is the matrix of the first $\boldsymbol{t}$ eigenvectors and $\boldsymbol{b} = [b_1, b_2, ...b_t]$ is a set of $\boldsymbol{t}$ parameters controlling the shape. This part of shape manipulation is interchanged with ICP in order to fit the model to an unseen, non-landmarked face. The idea behind DMS is to iteratively manipulate the shape parameters within “legal” constraints as the surfaces are fitted to each other using ICP. Figure 3.10 shows the fitting process.

3.2.4 Global similarity-based methods

The techniques mentioned so far rely on local information to register surfaces. Contrary to these, the approaches of Johnson and Hebert [100] and Campbell and Flynn [37] perform registration based on the global surface geometry and can deal potentially better with featureless patches. Johnson and Hebert presented a technique for surface characterization that does not require feature extraction or segmentation. Instead, they use *spin-images* to describe an object. Using a single point basis constructed from an oriented point, they described the position of other points on the surface using two parameters. The accumulation of these parameters for many points on the object's surface results in an image at each oriented point. Since these images describe the relationship between points on the same surface, they are invariant to rigid transformations. Using the correlation between images, they managed to build correspondences between a model and measured data.

3.3 Conclusions

In the remainder of the thesis surface registration techniques will be used to enable comparison between faces. Some of the registration techniques presented will be used in order to align homologous parts of the faces to each other. Since we are trying to detect differences and similarities in the facial features across all faces it is imperative that these features are registered to each other. This way, pose differences between faces are minimized and the residual differences between the facial surfaces are due to differences in the actual facial structure.

Chapter 4

3D Face Recognition Using Surface and Texture Registration

4.1 Introduction

In this chapter a face recognition algorithm is proposed which consists of two steps [151]: The first step involves a 3D rigid registration of the facial surfaces to determine the correspondences between two faces. In the second step, different similarity metrics are introduced and evaluated in order to measure the distance between pairs of closest points on the two faces. A key advantage of the proposed technique is the fact that it can automatically identify faces irrespective of the posture of the subjects. The effect of illumination differences on the recognition rate is measured and discussed. Finally, since this technique is used on subjects with various facial expressions, the effects of non-rigid facial changes in the face are assessed and the strength of the discriminatory power of 3D data is demonstrated.

In this chapter face data from two different databases is used for validating the proposed technique: In both cases it consists of a dense 3D mesh of vertices describing the facial geometry and 2D texture map describing the facial appearance of each subject producing a photo-realistic model of each face.

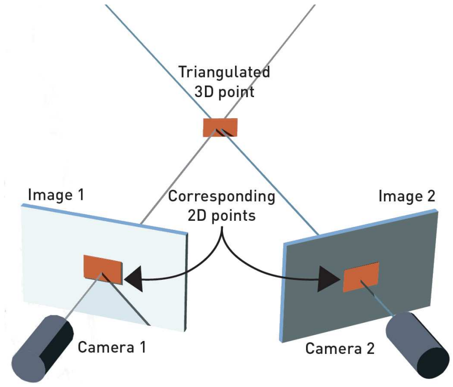


Figure 4.1: A passive sensor. Two cameras with a known geometric relationship capture two images and by establishing the correspondence between their pixels recover 3D shape [123].

4.2 3D face data

The data used in the experiments are 3D face surface scans. Various shots of 57 subjects were taken by us using the VisionRT VRT3D [123] and 300 surface scans were produced by the University of Notre Dame using a Minolta Vivid 900 range scanner [46].

4.2.1 VRT3D face scans

To collect 3D data for the first experiments a commercial stereo camera system was used, which is a hybrid between a passive stereo sensor and a structured light one. In the passive stereo approach two cameras (minimum) with known geometric relationship are used to capture images of the subject. Correspondences between the 2D images are established and the location of 3D points can be computed (Figure 4.1). In the structured light approach one camera is used along with a light projector and the geometric relationship between the two must be known. In the hybrid approach a light pattern is projected on the surface merely as an aid for finding correspondences between the images of the two cameras. The VRT3D camera is made up of three video cameras and a speckled pattern projector. The projector projects a random light pattern of dots on the surface of the face to aid 3D shape recovery. The output is a 3D face model. The third camera uses a filter to eliminate the speckled pattern projected onto the faces to capture greylevel texture in-

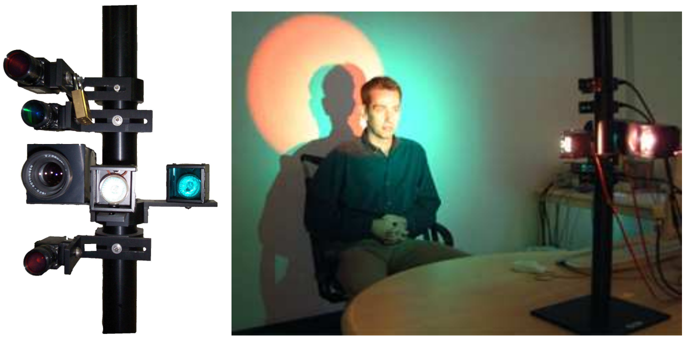


Figure 4.2: The VisionRT VRT3D camera system.

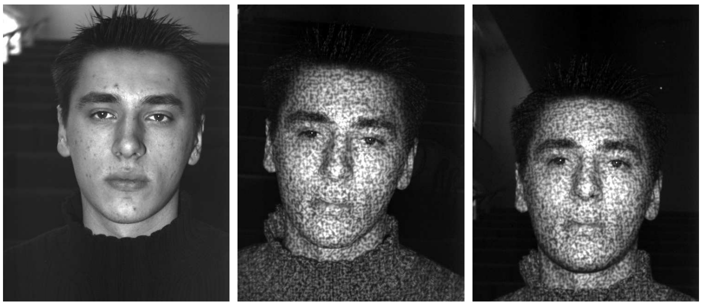


Figure 4.3: The three images used for the reconstruction of the surface.

tensity information. Furthermore, to aid correct texture capture, a halogen light is used to illuminate the surface of the face. Figure 4.2 shows the VRT3D capture system and a typical data capture session. Figure 4.3 shows the images collected by the three cameras which are used for the reconstruction of the textured facial surface. This technology was chosen over laser scanners because of its speed of acquisition (up to 30 frames/sec) and the speed of data reconstruction (< 5 sec. on a 1GHz machine), which allows near real-time processing in realistic scenarios. The speed of the data acquisition also prevents motion artifacts from being introduced in the 3D acquisition process. Finally, the system is built in a cost effective fashion, reducing hardware costs significantly compared to other capturing techniques. The accuracy of the VRT3D is fairly high with an RMS error of under $1mm$ for a typical 3D surface acquisition. The accuracy has been determined in a study where a phantom human torso was tracked using an Optotrak LED tracking system (RMS accuracy of $0.1mm$) as a gold standard [189]. Ten LEDs were positioned around the torso and the VRT3D and Optotrak systems were co-calibrated. The phantom

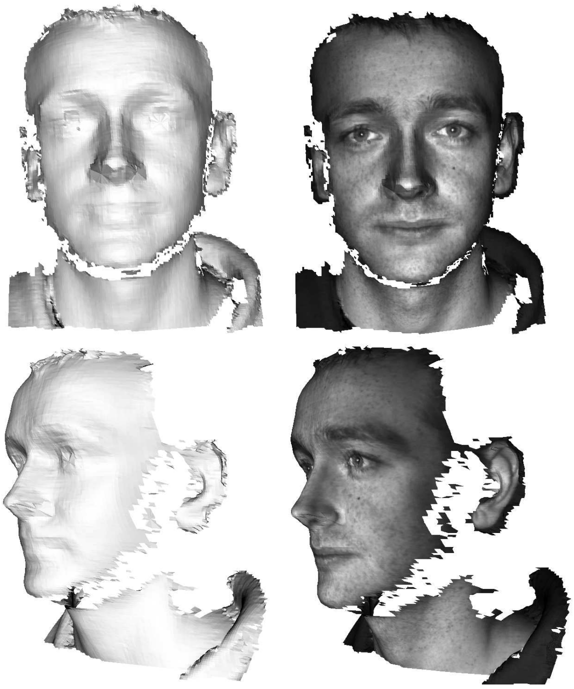


Figure 4.4: An example dataset after reconstruction with and without texture.

was then moved to different positions and the LED coordinates are tracked by both systems. The discrepancy between them was used to establish the accuracy of the VRT3D system. The human torso is admittedly a simpler surface to model than the high curvature regions of the face, but gives an idea about the capturing capabilities of the system. The drawbacks of the acquisition system are its relative bulkiness and since the cameras and projector use lenses, its limited depth of field. A typical stereo camera sensor has a depth of field of about $0.3m$ or less while a structured light sensor has about $1m$. This makes the 3D face acquisition more intrusive as the faces need to be placed within a specific distance from the camera, as with laser scanners. Figure 4.4 shows the output surface from the reconstruction process with and without texture.

4.2.1.1 Preprocessing of data

In order to speed up the processing and reduce the registration errors every subject's face was preprocessed. An ellipse outlining the subject's face is drawn manually on the 2D texture map. The ellipse was drawn manually by placing the mouse cursor on one side of

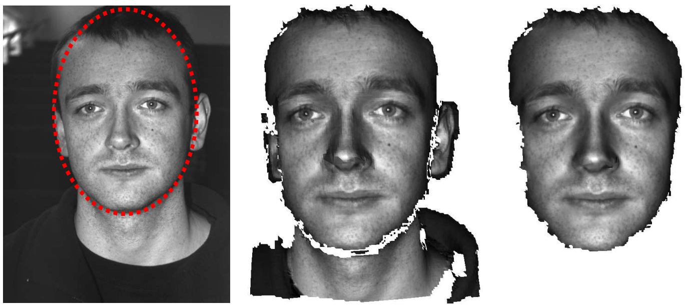


Figure 4.5: A VRT3D dataset example before and after manual cleaning.

the face and dragging it across the 2D facial surface. This process was done manually but it could theoretically be automated by detecting the edges of the facial surface. However, our objective was not face detection and such an endeavor was not pursued. Since the mapping between the surface points and the 2D texture map is known, the vertices which have texture coordinates that lie outside the ellipse can be eliminated. Each point consists of a set of 3D coordinates $\boldsymbol{v}_i = (v_{i_x}, v_{i_y}, v_{i_z})$, which defines its location in space, and a 2D texture coordinate $\boldsymbol{t}_i = (t_{i_x}, t_{i_y})$, which defines the corresponding pixel on the texture image. Let $\boldsymbol{c} = (c_x, c_y)$ be the center of a user-defined ellipse on the face. If the major axis of the ellipse is of size a and the minor of size b then the points which are kept are defined by:

$$\left\{(\boldsymbol{v}_i, \boldsymbol{t}_i) : \left(\frac{(t_{i_x} - c_x)^2}{a^2} + \frac{(t_{i_y} - c_y)^2}{b^2} - 1\right) < 0\right\} \tag{4.1}$$

In this fashion it is possible to discard parts of the mesh that correspond to the neck and hair which confuse the registration process and can introduce errors. An example of this processing is shown in Figure 4.5.

4.2.1.2 Data collection protocol

The data was collected over various sessions with student participants. Each subject was added to the database within the same session with the images of each subject being captured within $20sec$ of each other. Each participant was asked to take three different head positions. These were a 45^o left and right turn of the head, a 0^o position (looking straight into the camera), as well as a 45^o upward tilt. The above data was captured

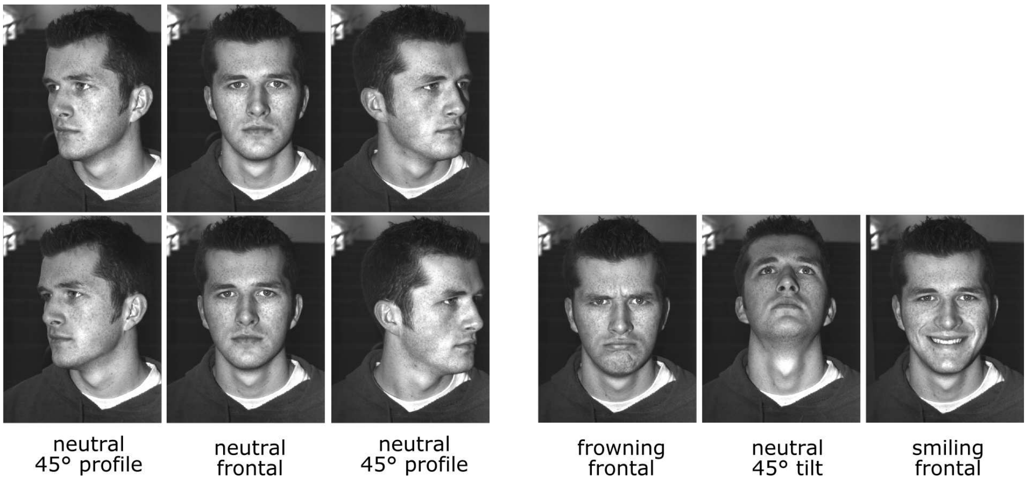


Figure 4.6: Type of data captured with the VRT3D camera.

twice in order to have a plethora of data for training and testing purposes as proposed in [137]. Additionally, three different emotional expressions of each subject at 0^o were acquired; one where the subject is frowning, one with the subject smiling and a neutral one. Figure 4.6 shows examples of the type of data captured with the VRT3D camera. In most cases each reconstructed face has between 8,000 and 12,000 points. The final number of subjects was 57 which was composed of 4 females and 53 males. Furthermore, based on the subjects' ethnic self-classification, the database contained images of eight South Asians, six East Asians, one Black and forty two Caucasians between the ages of 19 and 25.

For a patch on the surface to be captured by the VRT3D sensor there are two basic requirements that need to be satisfied; both stereo cameras must have the patch in their field of view and the speckled pattern needs to be reflected adequately from that area. There are five major sources of error that arise if one of the previous requirements is not satisfied (Figure 4.7): Areas that reflect the speckled pattern poorly such as eyes, eyebrows, beards and teeth sometimes appear as holes in the reconstructed surface (Figure 4.7(a)). Additionally, since the speckled pattern appears as a series of black dots on the face, people with particularly dark skin tend to have more holes than lighter-skinned people (Figure 4.7(b)). Areas that are occluded from at least one camera due to the subject's head posture are also not reconstructed as shown in (Figure 4.7(c)). Occlusion can also

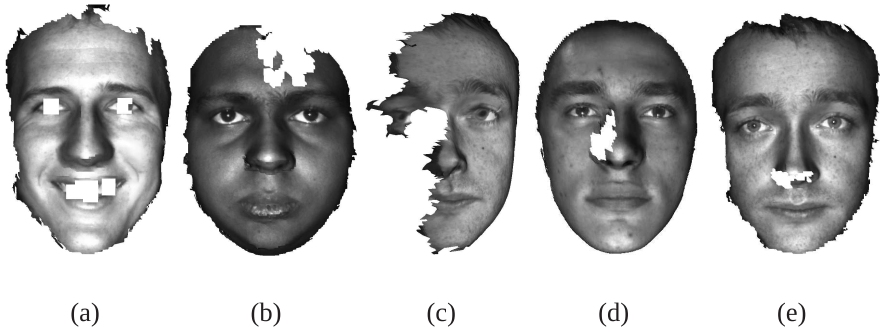


Figure 4.7: Surface reconstruction errors in the VRT3D data.

occur when part of the face (usually the nose) hides another part of the face because the subject was not looking completely straight into the camera (Figure 4.7(d)). Finally, areas of high curvature pose difficulties for the VRT3D camera system. The tip of the nose and its nostrils are sometimes missing from the final mesh (Figure 4.7(e)). A few images (< 10) were deemed to contain too many capture and reconstruction errors due to some of the above reasons and they were excluded from the final database.

4.2.2 University of Notre Dame face scans

In order to validate the experimental results with a known database, a number of datasets were obtained from the University of Notre Dame [46]. These are part of a larger collection of mainly 2D face images. The database has been built using the Minolta VIVID 910 camera which uses a structured light sensor to scan surfaces. Light reflected from the surfaces is captured by a CCD camera and the surface is reconstructed by inferring 3D shape from the distortion of the light pattern. Measurements are usually completed within $2.5sec$. The camera also captures color texture information which it applies on the surfaces. An advantage of this data over the VRT3D datasets is that they are usually of higher resolution. A typical face has about 20,000 points before any preprocessing. The facial features are better defined and there are less holes and surface artifacts. On the other hand, the datasets obtained from the University of Notre Dame database were only frontal images of neutral expression. The 2D texture maps are in color and in order to create similar experimental conditions they were converted to greylevel. Figure 4.8 shows a dataset example from the Notre Dame database. Another reason why tests with this dataset is particularly important is because, contrary to the VRT3D dataset, the data for each subject was collected over two sessions separated by 11 to 13 weeks. This is significant as data collected on different dates can affect the recognition rates [157] with Gross *et al.* [80] noting significant differences even between images taken two weeks apart.

Figure 4.8 shows that there are parts of the face that need to be discarded in order for the recognition to be enhanced as was the case with the VRT3D data. Areas such as shoulders and ears were eliminated in a similar fashion to Section 4.2.1.1 in order to improve the automatic registration. Notice how this dataset also suffers from the same problems as the VRT3D database. Holes and/or spikes are present in almost all datasets as are areas of low reflectance, such as eyebrows, which are difficult to capture by a range sensor. Another drawback of the datasets by Notre Dame University is the lower quality of the texture data compared to the VRT3D data. As demonstrated later, the 2D texture data applied on the 3D datasets is often less than 130 pixels wide from one ear to another and less than 200 pixels from chin to the beginning of the hairline, while the VRT3D textures are 300 and 450 respectively. More importantly however, the contrast in the data is particularly poor with facial images often being over-exposed. The dataset tested on contained two biometric samples, spaced two weeks apart, for each of the 150 subjects and they were assessed by the author to contain: 114 Caucasians, 38 East Asians, 2 South Asians, 3 Blacks and 3 undetermined subjects. The subject pool was made out of 81 males and 69 females. Once again certain images that had a significant amount of the facial surface missing were removed from the face pool and were not used. This was particularly important for the Nore Dame database because these same images were going to be used in the next chapters in landmark-based techniques. What this entails is that if the faces could not be accurately landmarked because the area around the fiducial point is missing, then the dataset was discarded.

The same database has been used in previous work. Chang *et al.* used a larger pool of these subjects with PCA in a single probe study. With an optimal set of eigenvectors in 2D and 3D, they achieved a rank-one recognition rate of 89.0% for 2D, and 94.5% for 3D.

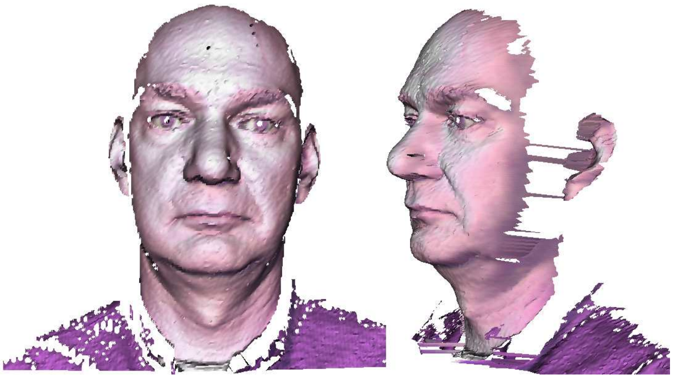


Figure 4.8: Example of a Notre Dame dataset.

In a multi-probe study, where one or more biometric sample for each subject is used for testing, the 3D performance dropped to 92.8% while 2D performance improved, reaching 89.5%. After combining the two modalities in a multiple-probes scenario they were able to obtain significantly better performance, at 98.8%, than for either 2D or 3D alone [39]. Russ *et al.* [175] used 200 subjects from the same database to perform verification tests on 30 probes. In order to simplify the processing they converted the 3D point clouds to 2D range images which they aligned to each other by using the nose of each face as a reference point and then implementing a registration technique which improves the alignment of the images using the Hausdorff distance and the mean square error to establish correspondence between points on the two surfaces. Two similarity metrics, the Hausdorff distance and the mean square error were finally used after the faces were registered. The former reached a verification rate of $P_V = 98\%$ with a false acceptance rate of $FA = 0$ while the latter produced a $P_V = 95\%$ at a $FA = 0$.

4.3 Face matching

4.3.1 Face registration in 3D

If one assumes that the face is a rigid body which can be captured perfectly by sensors, then the mean square error difference between the points on the two surfaces can only be due to difference in identity and 3D pose. Based on these assumptions there is a rigid transformation $\boldsymbol{T}_{rigid}$ that will bring the faces of the same subject in near-perfect alignment and will eliminate errors due to pose differences and isolate differences due to identity. In a more realistic scenario and more formally, let $\overline{\Delta}_{intra}$ be the average difference between all probes p_i and a gallery faces g_i where id(p_i)=id(g_i) such that:

$$\overline{\Delta}_{intra} = \frac{1}{M}\sum_{i=1}^{M}||p_i - g_i|| \tag{4.2}$$

where M is the number of probes. The average difference between all possible gallery/probe combinations is set by:

$$\overline{\Delta}_{inter} = \frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}||p_i - g_j|| \tag{4.3}$$

where N is the number of faces in the gallery. By aligning all probes to all gallery faces we are trying to maximize the ratio ρ of the above differences:

$$\rho = \frac{\overline{\Delta}_{intra}}{\overline{\Delta}_{inter}} \tag{4.4}$$

Registration of surfaces or pose estimation is a key problem in computer vision that has been studied in depth (see Chapter 3). Given two facial surfaces, i.e. a probe (moving) face $A = \{\boldsymbol{a}_i\}$ and a gallery (fixed) face $B = \{\boldsymbol{b}_i\}$, the goal is to estimate the optimal rotation $\boldsymbol{R}$ and translation $\boldsymbol{t}$ that best aligns the faces. To find the optimal rigid transformation $\boldsymbol{T}_{rigid} = (\boldsymbol{R}, \boldsymbol{t})$ the ICP algorithm is used [12]. The function to be minimized is the mean square difference function between the corresponding points on the two faces:

$$f(\boldsymbol{T_{rigid}}) = \frac{1}{|A|}\sum_{i=1}^{|A|}||\boldsymbol{b}_i - \boldsymbol{R}\boldsymbol{a}_i - \boldsymbol{t}||^2. \tag{4.5}$$

where points with the same index correspond to each other. The correspondence is established by looping over each point $\boldsymbol{a}$ on probe face A and finding the closest point, in

Euclidean space, on gallery face B:

$$d(\boldsymbol{a}, B) = \min_{\boldsymbol{b} \in B} ||\boldsymbol{b} - \boldsymbol{a}|| \tag{4.6}$$

The squared difference between two points $||\boldsymbol{a} - \boldsymbol{b}||^2$ is defined as:

$$||\boldsymbol{a} - \boldsymbol{b}||^2 = (a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2 \tag{4.7}$$

Unlike Besl and McKay [12] who first proposed the ICP algorithm, we do not use the quaternion method for calculating $\boldsymbol{R}$. Instead, we use the *singular value decomposition* (SVD) approach [6], based on the cross-covariance matrix of the two point distributions which generalizes well to n dimensions. This calculates a least-squares estimate of the rotation matrix $\boldsymbol{R}$ for a set of points $\boldsymbol{a_i}$ and the rotated points $\boldsymbol{b_i} = \boldsymbol{Ra_i} + \boldsymbol{t} + \boldsymbol{n_i}$ for some translation vector $\boldsymbol{t}$ and noise vectors $\boldsymbol{n_i}$. This is achieved by translating the points by their mean locations

$$\boldsymbol{a'_i} = \boldsymbol{a_i} - \frac{\Sigma_{i=1}^{N} \boldsymbol{a_i}}{N} \qquad \boldsymbol{b'_i} = \boldsymbol{b_i} - \frac{\Sigma_{i=1}^{N} \boldsymbol{b_i}}{N}, \tag{4.8}$$

calculating the matrix $\boldsymbol{H} = \Sigma_{i=1}^{N} \boldsymbol{a'_i} \boldsymbol{b'^{t}_i}$, decomposing it as $\boldsymbol{H} = \boldsymbol{U \Lambda V^t}$ using SVD and calculating $\boldsymbol{X} = \boldsymbol{VU^t}$. This results in the required rotation matrix. The translation $\boldsymbol{t}$ is then calculated by subtracting the centroid of the probe from the gallery face after rotation.

Apart from the basic elements of the ICP a few optimizations have been implemented in order to improve the registration. Before rigid registration is performed on the faces, the center of mass of all surfaces is moved to the origin of the coordinate system. This compensates for large differences in the distance between subjects and the chances of correct pairings between the points of the two surfaces increases. Furthermore, because the probe faces are not only frontals but semi-profile as well, several initial transformations are used in order to further fine-tune the starting position. Profile faces, as seen in Figure 4.7(c), contain only a subset of a frontal face, since occlusion prevented the

VRT3D sensor from capturing the rest of the face. After moving all faces to the origin, profile faces will most likely be misregistered if their starting position is not adjusted. Note that the pose (frontal, profile or upward tilt) is unknown and given that we want to implement an *automatic* face registration, a coarse initial pose estimation is necessary to avoid misregistrations. The upwards tilted faces have been translated to the origin like all other datasets and given that they are frontal and contain most of the face, the registration between them and the frontal gallery faces is not problematic. We assume that each probe will either be a frontal face, a 45^o profile face or a -45^o profile face. In order to ascertain which is the ideal translation component $\boldsymbol{t} = (t_x, t_y, t_z)$ of the initial transformation, three hypothetical starting transformations with different translations are tested:

$$\begin{aligned}\boldsymbol{t}_0 &= (0,0,0)\\ \boldsymbol{t}_1 &= (+30,0,0)\\ \boldsymbol{t}_2 &= (-30,0,0)\end{aligned}$$

Based on our experience the above translations are sufficient to cover all head postures cases during data capture. Assuming that the difference between a probe face A and a gallery face B is $|A - B|$ as defined in eq. 4.9, then the ideal translation to align their respective point sets $\boldsymbol{a}$ and $\boldsymbol{b}$ is:

$$f(j) = \frac{1}{|A|}\sum_{i=1}^{|A|} ||\boldsymbol{b}_i - \boldsymbol{a}_i - \boldsymbol{t}_j||^2. \tag{4.9}$$

In other words, the similarity between the faces is assessed at three different starting positions and the registration starts from the position where the mean square difference between the corresponding points of the faces is the smallest. This way we achieve good registrations even between profiles and frontal datasets without compromising the principle of automation.

Another optimization implemented, as proposed in [204] and [174], was the rejection of “hazardous” pairings during the closest point search in ICP. According to ICP one must

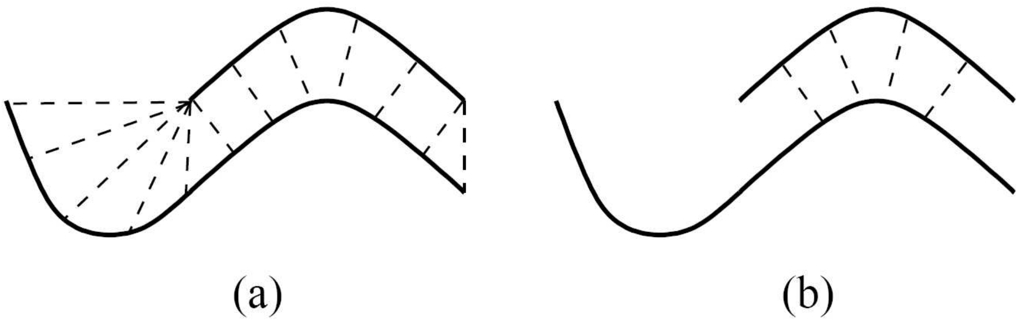


Figure 4.9: Establishing correspondence between two surfaces. Closest-point pairings that contain edge points are used in (a) but ignored in (b) when calculating the optimal transformation. Using these pairings in (a) will cause incorrect registration by forcing the top surface to move to the left (from [174]).

loop over all the points in the probe (source) dataset and find the closest points in gallery (target) datasets. In that case, hazardous pairings are those where the closest point on the gallery face is a boundary point, which is in turn defined as a point that is found on a cell edge that is not shared by another cell. What this means in effect, is that during any one iteration only correspondences between those parts of the two surfaces that overlap are taken into consideration. Furthermore, this optimization also deals with problems in data such as holes, because an area on a face that corresponds to an area on another face that is missing will not be used. Figure 4.9 shows an example of pairings that should be ignored in order for the surfaces to be correctly registered. Finally, in order to speed up the calculation of the closest match we use a k-D tree locator as presented in [10] and [65] where points are organized in such a way that the number of point-to-point comparisons are minimized (see Section 3.2.2). The ICP and its variants implemented (mentioned above) allow us to register two surfaces in about 20 iterations in under < 15sec.

The extentions of ICP that one chooses to implement are very much application specific. In this case it was felt that the above would provide an ideal balance between accuracy and speed. For example Chen and Medioni [43] report that the point-to-plane distance makes the convergence process less susceptible to local minima. Lu *et al.* [127] agree and state that the point-to-plane distance reflects the true distance between the two surfaces better. However, the latter reports that the point-to-plane distance is significantly slower to calculate. In this work speed was one of the main drivers for choosing the registration method and thus significantly slower architectures were ruled out. Furthermore, given that the faces were moved to the originin before any ICP iteration had started and given the nature of the facial surface (highly regular across the population), getting stuck in local minima was uncommon. As a result something better than point-to-point distance was not needed. Extensions for the ICP, such as the ones presented above, were implemented if there was a need based on the specific datasets used.

4.3.2 Face registration in 4D

So far only geometric information has been used to align the faces. However, that is only one source of information, which can be used to drive the registration. Since all acquired faces include geometric information as well as a 2D texture map, a 4D registration has been developed, which incorporates textural information in the ICP algorithm [101, 60]. Here, each point is represented by a 4D vector $\boldsymbol{a} = (a_x, a_y, a_z, a_t)$ where a_t is the texture intensity of the point. The cost function to be minimized is still defined by eq. 4.9 but this time calculating the distance between points $\boldsymbol{a}$ and $\boldsymbol{b}$ in surfaces A and B respectively is an operation on 4D points:

$$||\boldsymbol{a} - \boldsymbol{b}||^2 = (a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2 + w(a_t - b_t)^2 \tag{4.10}$$

where w is a weight variable that normalizes the intensity value differences (since the texture value is usually a unitless number) and thus determines the importance of the texture in the search for the closest point. Since there is no longer a closest form solution for the optimal transformation, we treat the registration as an optimization problem trying to minimize the cost function in eq. 4.9 by using *gradient descent* to find the global minimum.

4.3.3 Measuring facial similarity

After two faces have been registered (using either the 3D or 4D ICP algorithm described above) the residual 3D or 4D distance between the points in the probe face and the closest point correspondences in the gallery face can be used as a similarity metric. Let $A' = \boldsymbol{T}(A)$ be the probe surface after registration and B be the model or gallery surface. The Euclidean distance between the two surfaces is then defined as:

$$\Delta_E = \frac{1}{|A'|}\sum_{i=1}^{|A'|}((\boldsymbol{a'}_{i_x} - \boldsymbol{b}_{i_x})^2 + (\boldsymbol{a'}_{i_y} - \boldsymbol{b}_{i_y})^2 + (\boldsymbol{a'}_{i_z} - \boldsymbol{b}_{i_z})^2 + w(\boldsymbol{a'}_{i_t} - \boldsymbol{b}_{i_t})^2) \quad (4.11)$$

To measure only differences in surface shape the parameter w is set to 0. The similarity metric has been intentionally separated from the registration process in order to increase the modularity of the algorithm, which enables easy switching between different similarity metrics.

As discussed earlier, the purpose of aligning the facial surfaces was to minimize errors due to registration and thus isolate the distance error due to difference in identity. Let face A and B belong to the same subject while C belongs to a different subject. Furthermore, let A' and C' be the faces after rigid registration to face B. Figure 4.10 shows a 2D schematic representation of the facial surfaces before and after registration. The yellow area between each pair of lines denotes the mean square error between them. Despite the rigid registration there are still significant differences between faces B and C' while the difference between faces A' and B belonging to the same subject have been mostly eliminated.

4.4 Experimental protocol

As proposed in the FERET evaluation protocol in [157] and discussed in Chapter 2 we will assess different methodologies for recognizing faces by measuring performance in three tasks: verification, open-set identification and closed-set identification. In order to perform open-set identification subjects are usually split into three groups. The first

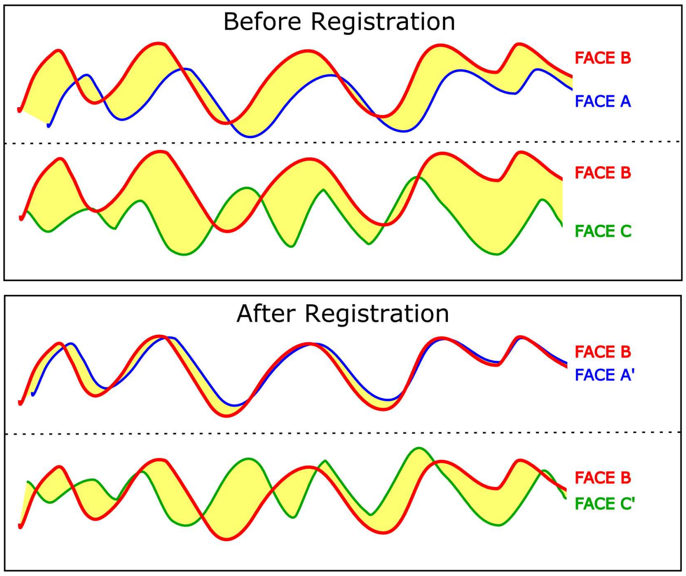


Figure 4.10: Schematic mean square distance before and after registration. Surface A and B are different biometric samples of the same subject while surface C belongs to a different subject. Notice that after rigid registration the difference between A' and B is very small compared to the one between C' and B.

group comprises of faces that are known to the system and are referred to as the *gallery* $\mathcal{G}$. The other two are the probe set $\mathcal{P}_{\mathcal{G}}$, containing different biometric samples of the same subjects as the ones contained in the gallery and the probe set $\mathcal{P}_{\mathcal{N}}$ containing samples of people not in the gallery. Given that our population pool is limited, having only 57 subjects captured by the VRT3D and 150 in the University of Notre Dame database we are not going to use different datasets for open-set identification. Instead we are going to divide the subjects into two pools, the gallery set $\mathcal{G}$ and the probe set $\mathcal{P}$ where $\mathcal{P} = \mathcal{P}_{\mathcal{N}}$. The correct detection and identification rate P_{DI} is still the same as the traditional open-set identification but the false alarm rate P_{FA} is calculated differently. In the open-set identification already presented, P_{FA} is computed by:

$$P_{FA}(\tau) = \frac{|\{p_j : \max_i s_{ij} \geq \tau\}|}{|\mathcal{P}_{\mathcal{N}}|} \tag{4.12}$$

Since there is no $\mathcal{P}_{\mathcal{N}}$ set in this case, the P_{FA} is calculated by:

$$P_{FA}(\tau) = \frac{|\{p_j : \max_i s_{ij} \geq \tau \text{ and } \mathrm{id}(g_i) \neq \mathrm{id}(p_j)\}|}{|\mathcal{P}| - 1} \tag{4.13}$$

This means that for every face in $\mathcal{P}$ we check if there is any face in $\mathcal{G}$ other than the face belonging to the same subject that would cause a false alarm, given a threshold τ. In other words, for every probe p_i we remove the correct match g_i from $\mathcal{G}$ and check if any other gallery face would provide a match above τ. This provides a very close approximation of the original open-set identification measurement. The trade-off relationship between P_{FA} and P_{DI} is normally plotted on an ROC and the area under the curve will be reported as an evaluation measure. In the worst case scenario the area under the ROC curve is 50% while in the best case 100% (Figure 2.3). The second measurement related to P_{FA} and P_{DI} that is reported is the $P_{FA} = P_{DI}$ rate.

The third measure reported is the rank 1 rate. The cumulative count in this case is given by:

$$C(1) = |\{p_j : \text{rank}(p_j) \leq 1\}| \tag{4.14}$$

The closed-set identification for rank 1, $P_I(1)$, is the fraction of probes at rank 1 and is described by:

$$P_I(1) = \frac{|C(1)|}{|\mathcal{P}|} \tag{4.15}$$

It is important to note that the rank 1 or $P_I(1)$ rate will return higher scores than the stricter $P_{FA} = P_{DI}$ rate. This is only natural, since the $P_{FA} = P_{DI}$ measure expects not only the correct gallery face g_i to be matched to each probe p_j, but that the similarity s_{ij} between them will be greater than threshold τ. One, therefore, expects the $P_{FA} = P_{DI}$ and the ROC curve measurements (to which they are related) to be more volatile to changes in the experimental parameters.

For calculating the verification rate we use the round-robin method [156], which is designed for two groups $\mathcal{G}$ and $\mathcal{P}$. The rate reported will be the verification rate P_V with a false acceptance rate $P_{FA} = 1\%$.

$P_I(1)$ **Rates of the VRT3D database**					
w	**frontal** 	**profile** 	**smiling** 	**frowning** 	**tilted**
0	100%	94.7%	43%	96.4%	100%
0.02	100%	94.7%	50.8%	98.2%	98.2%
0.04	100%	80.7%	64.9%	98.2%	98.2%
0.06	100%	70.1%	75.4%	98.2%	98.2%
0.08	100%	52.6%	85.9%	98.2%	94.7%
0.1	100%	45.6%	87.7%	96.4%	92.9%
0.12	100%	42.1%	89.4%	96.4%	89.4%
0.14	100%	33.3%	91.2%	96.4%	84.2%
0.16	100%	28%	92.9%	94.7%	80.7%
0.18	100%	26.3%	94.7%	94.7%	80.7%
0.2	100%	26.3%	94.7%	94.7%	75.4%

Table 4.1: The rank 1 rates ($P_I(1)$) of the different types of queries using various texture weights on the VRT3D database. The texture weights are expressed in factors by which one multiplies the texture value of a specific point, as indicated in eq. 4.10.

4.5 Results

4.5.1 Automatic recognition using the VRT3D database

Table 4.1 and Figure 4.12 show the rank 1 rates of the various expressions and head positions using the VRT3D face database. The rank 1 recognition rates for frontal faces are very high, reaching an impressive $P_I(1) = 100\%$. Profile faces also have high rank 1 rates but the rates are decreasing as more weight is put on the texture. The same is seen for tilted faces, that perform best when no texture is used. For smiling and frowning faces using texture improves the results significantly. A similar trend can be observed in Table 4.2 and Figure 4.13 showing the correlation between false alarm rate and the detection and identification rate ($P_{FA} = P_{DI}$). The profiles and the tilted faces perform best when no texture is used. On the other hand, the frontal faces, the smiling and frowning ones seem to peak in terms of performance in the middle of the texture weight range and then start dropping again. The same behaviour is observed when the area under the ROC curves in table 4.3 and figure 4.14 is measured as well as in the verification rates P_V in Figure 4.15 and Table 4.4.

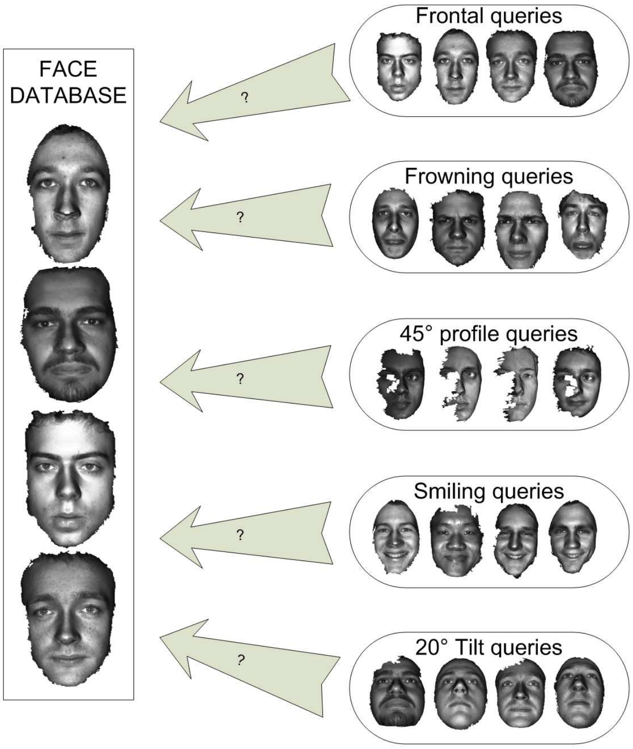


Figure 4.11: Type of queries according to the experimental protocol.

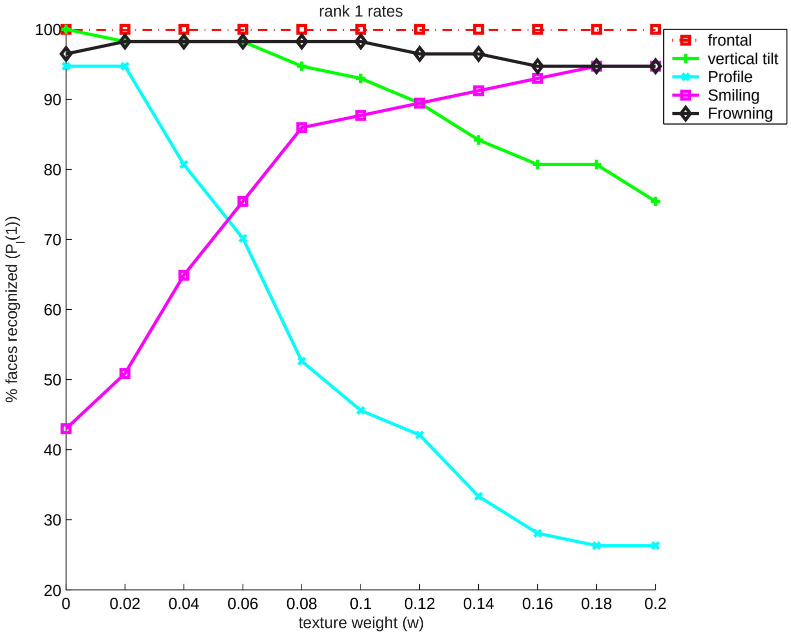


Figure 4.12: All rank 1 rates ($P_I(1)$).

$P_{FA} = P_{DI}$ rates of the VRT3D database					
w	**frontal** 	**profile** 	**smiling** 	**frowning** 	**tilted**
0	98.2%	91.7%	61.4%	90.3%	94.7%
0.02	99.1%	88.5%	66.6%	94.7%	93.8%
0.04	100%	77.1%	69.2%	96.4%	91.2%
0.06	100%	62.2%	71%	94.7%	87.7%
0.08	100%	57.8%	74.5%	93.8%	80.7%
0.1	98.2%	57.1%	75.4%	92.1%	71%
0.12	98.2%	57.8%	76.3%	88.5%	66.6%
0.14	97.3%	56.1%	75.4%	87.7%	64%
0.16	97.3%	55.2%	75.4%	87.7%	62.2%
0.18	97.3%	54.3%	75.4%	87.7%	60.5%
0.2	96.4%	54.3%	75.4%	85.9%	61.4%

Table 4.2: The $P_{FA} = P_{DI}$ rates of the different types of queries using various texture weights on the VRT3D database.

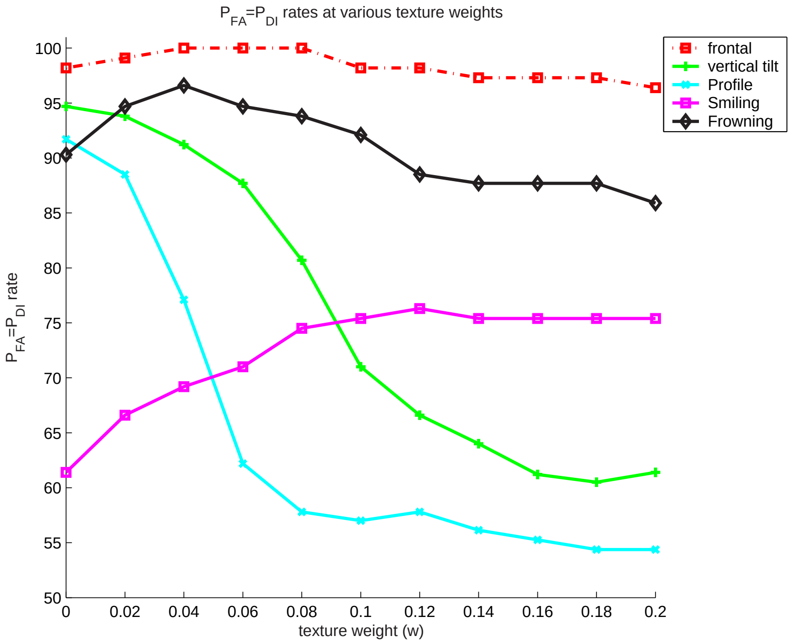


Figure 4.13: $P_{FA} = P_{DI}$ rates at various texture weights on the VRT3D database.

ROC Area percentages of VRT3D database					
w	**frontal** 	**profile** 	**smiling** 	**frowning** 	**tilted**
0	99.7%	94.4%	67.4%	97.4%	96.6%
0.02	99.8%	94.3%	73.2%	98.6%	96.6%
0.04	100%	82.9%	77.8%	99.1%	93.8%
0.06	100%	69.1%	81.1%	99%	90.8%
0.08	100%	61.8%	82.6%	98.4%	85.7%
0.1	99.9%	58.2%	83.7%	97.5%	80.1%
0.12	99.8%	56.3%	83.9%	96.3%	75.4%
0.14	99.9%	54.6%	84.6%	95.1%	72.2%
0.16	99.6%	53.6%	84.7%	93.8%	69.5%
0.18	99.4%	53.1%	84.7%	93.2%	67.5%
0.2	99.3%	52.8%	84.8%	92.5%	65.9%

Table 4.3: The ROC area rates of the different types of queries using various texture weights on the VRT3D database.

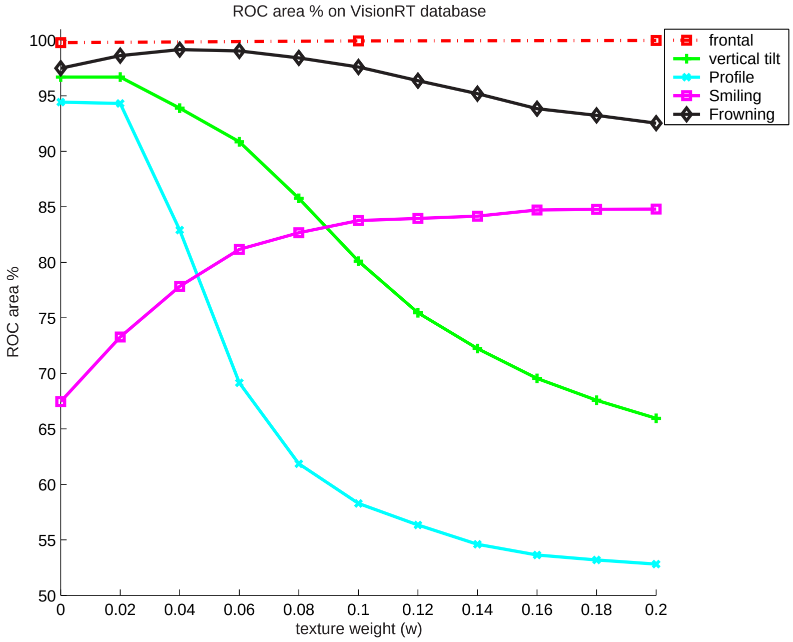


Figure 4.14: Percentage of graph under the ROC curve using the VRT3D database with various texture weights.

Verification rates P_V of the VRT3D database					
w	**frontal** 	**profile** 	**smiling** 	**frowning** 	**tilted**
0	100%	91.2%	33.3%	94.7%	94.7%
0.02	100%	92.9%	49.1%	98.2%	96.4%
0.04	100%	77.1%	61.4%	98.2%	91.2%
0.06	100%	50.8%	70.1%	98.2%	91.2%
0.08	100%	43.8%	73.6%	96.4%	78.9%
0.1	100%	28%	75.4%	96.4%	63.1%
0.12	100%	24.5%	75.4%	94.7%	50.8%
0.14	100%	22.8%	73.6%	92.9%	47.3%
0.16	100%	21%	75.4%	92.9%	43.8%
0.18	100%	21%	75.4%	91.2%	43.8%
0.2	100%	21%	73.6%	87.7%	42.1%

Table 4.4: The verification rates P_V of the different types of queries using various texture weights.

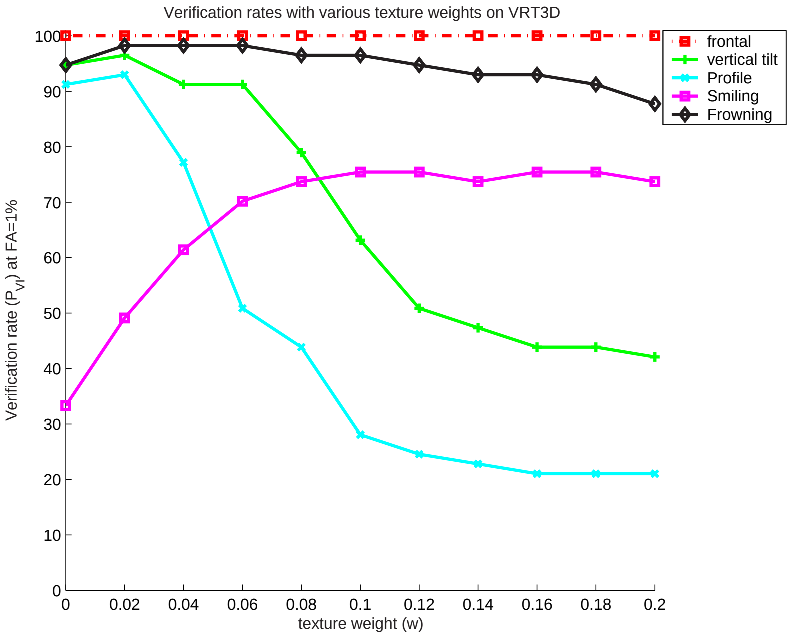


Figure 4.15: Verification rates P_V at $FA = 1\%$ using various texture weights on the VRT3D datasets.

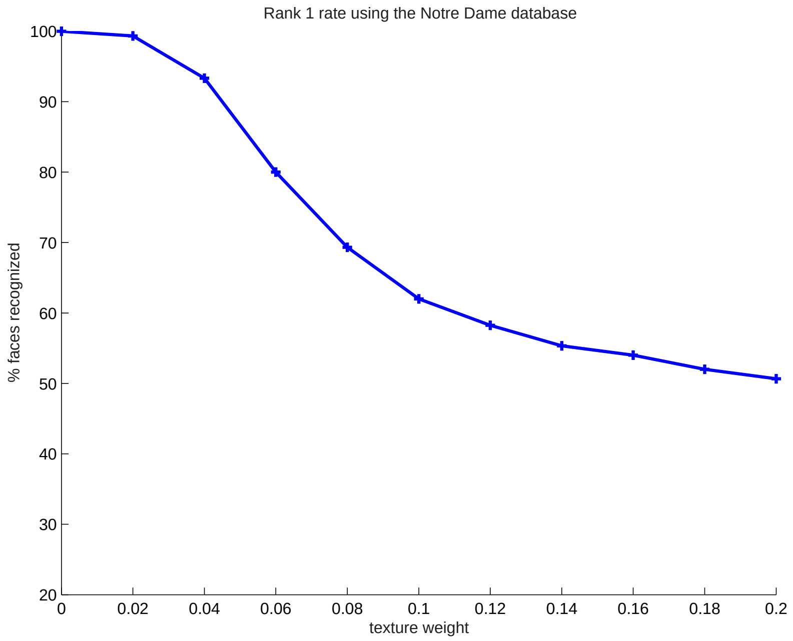


Figure 4.16: All rank 1 rates $P_I(1)$ using the Notre Dame database.

4.5.2 Automatic recognition using the Notre Dame database

Table 4.5 and figures 4.16, 4.17, 4.18 and 4.19 show a different trend for the Notre Dame Database. In this case 150 frontal faces were used as queries and 150 faces as gallery faces and the same tests were run on them as on the VRT3D database. The rank 1 rates when no texture is used are very high again, reaching 100%, but drop sharply as the texture weight increases even though both gallery and probe sets contain frontal faces. The same phenomenon is observed with the $P_{FA} = P_{DI}$ and P_V rates as well as the rate measuring the area under the ROC graph. Notice the difference in contribution of the texture information compared to the VRT3D database. The reasons behind this are discussed in the next section.

4.6 Discussion

The experimental results reported in Section 4.5 above demonstrate the wealth of information that exists in texture 3D surface models and how it can be used for accurate face recognition. The technique is automatic in the sense that it does not require the use of

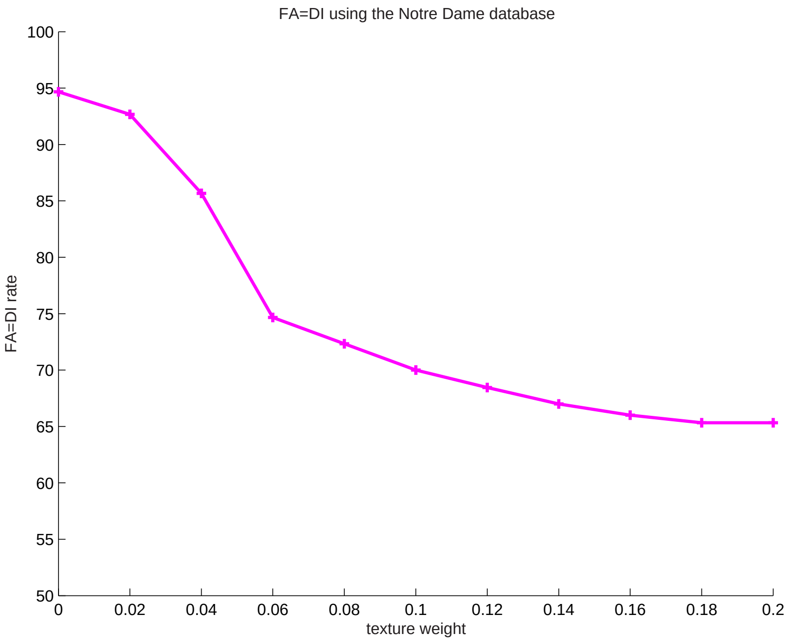


Figure 4.17: $P_{FA} = P_{DI}$ rates at various texture weights using Notre Dame datasets.

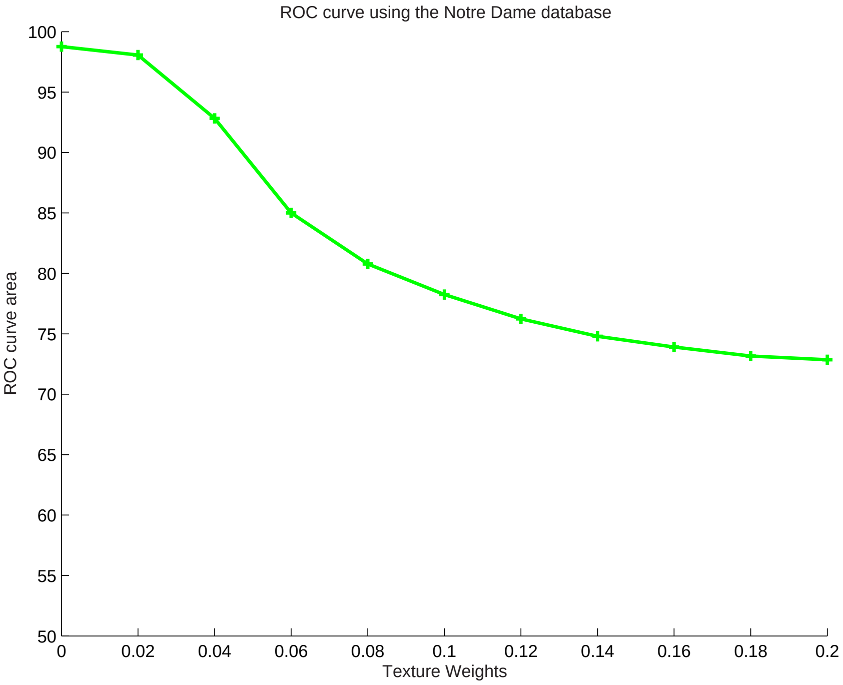


Figure 4.18: ROC curve rates using the Notre Dame database.

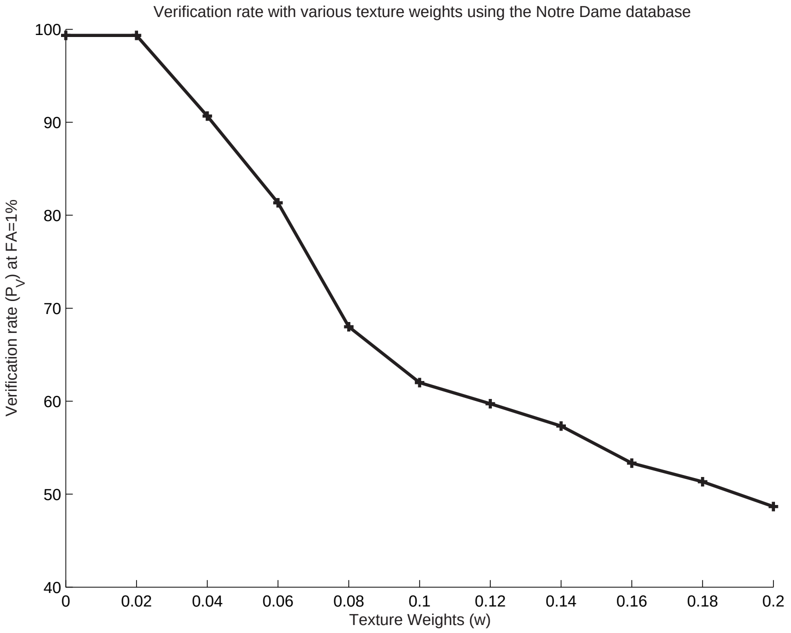


Figure 4.19: Verification rates (P_V at $FA = 1\%$) using the Notre Dame database.

Notre Dame Statistics				
$\boldsymbol{w}$	$P_I(1)$	$P_{FA} = P_{DI}$	ROC area	P_V at $FA = 1\%$
0	100%	94.6%	98.7%	99.3%
0.02	99.3%	92.6%	98%	99.3%
0.04	99.3%	85.6%	92.8%	90.6%
0.06	80%	74.6%	84.9%	81.3%
0.08	69.3%	72.3%	80.7%	68%
0.1	62%	70%	78.2%	62%
0.12	58.2%	68.4%	76.2%	59.7%
0.14	55.3%	67%	74.7%	57.3%
0.16	54%	66%	73.9%	53.3%
0.18	52%	65.3%	73.1%	51.3%
0.2	50.6%	65.3%	72.8%	48.6%

Table 4.5: The $P_I(1)$, $P_{FA} = P_{DI}$, ROC area and P_V rates of the Notre Dame database using various texture weights.

landmarks or any other information about a face prior to registration. However, for the faces to be correctly registered, the facial surfaces need to be relatively free from non-facial points, such as shoulders, which could adversely affect the alignment. In this chapter this is done by manually tracing an ellipsoid over each face's "area of interest" (see Section 4.2.1.1).

4.6.1 VRT3D data

4.6.1.1 Frontal faces

When using no texture the probe set that performs the best out of the VRT3D datasets is the group that has frontal faces with neutral emotional expression. This is expected as a face from the frontal, neutral gallery set is most similar to the frontal, neutral probe set. Figure 4.20 shows a distance map of two frontal faces of the same person where red colors indicate small distances while blue ones indicate large. The 3D distance between the faces is very small and it is only natural to expect high recognition rates. Furthermore, the VRT3D datasets are taken under a controlled environment where the angle and distance between the subject and the lights is constant. There is therefore very little within subject variability in the textures of the 3D data collected (assuming same head posture). As a result the texture information, up to a certain texture weight, can only increase the recognition rate. An example of this is demonstrated by the two VRT3D texture datasets on the left in Figure 4.21. In order to correctly measure the textural differences between the faces one can not simply subtract the 2D texture images from each other. Instead, the difference must be calculated by subtracting the intensity values of *corresponding* points from each other. Once again, to establish correspondence between points we loop over all points on one surface and we build closest-point pairings between them and the points on the other surface. Figure 4.22 shows the textural differences between two sets of corresponding points being computed and projected back onto a 2D image. Since the resulting value might be a negative number all values are normalized so that pixel intensities on the difference image are positive. The closer a pixel is to 128 (grey) the

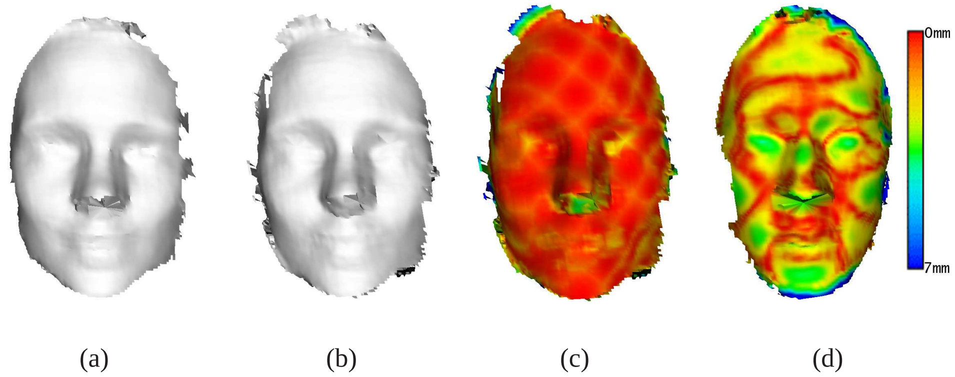


Figure 4.20: (a) and (b) shows two frontal datasets from the same subject, (c) shows a color map of the distances between them after registration while (d) is the color map of distances between (a) and a frontal dataset belonging to another subject (cool colors indicate large distance, warm colors indicate small distance).

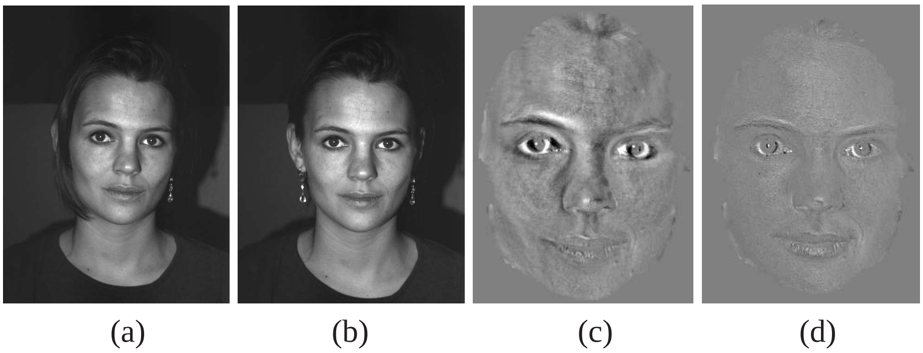


Figure 4.21: Images (a) and (b) are the original images of the subject. (c) shows the differences in illumination between the two images before they have been registered and (d) shows the differences after registration.

smaller the difference in texture between the points that projected it. Darker or lighter than grey values indicate greater differences in those face patches. In Figure 4.21(c) there are pronounced textural differences because the faces are not properly registered and merely translated to the origin. In Figure 4.21(d) of the same figure the textural differences have been reduced after the rigid registration.

As mentioned in Section 2.1.1 the $P_{FA} = P_{DI}$ rate is a stricter measurement and more sensitive to the experimental parameters. This can already be seen from the frontal, neutral datasets where the $P_{FA} = P_{DI}$ rates fall (Table 4.2) as more emphasis is put on the texture.

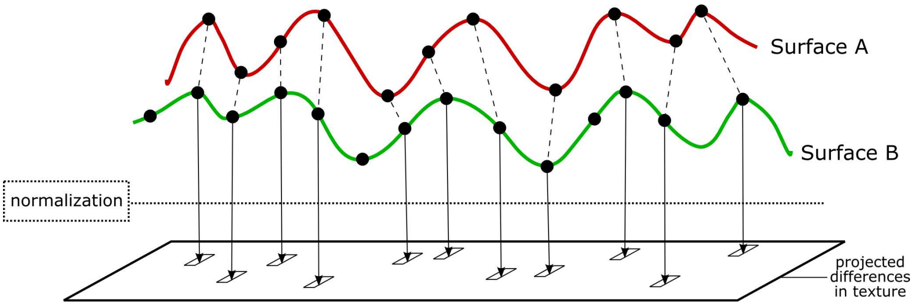
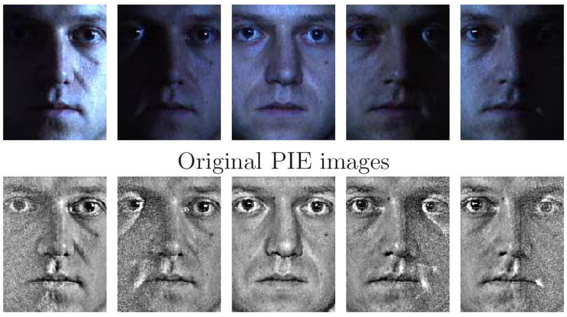


Figure 4.22: Subtracting texture intensities of corresponding surface points. The texture intensities of corresponding points are subtracted from each other and the difference is projected on a 2D bitmap. The normalization allows for a signed representation of the results.

4.6.1.2 Profile faces

When comparing $45\deg$ profile faces the rank 1 rate is very high when no texture is used (94.7%), something which would seem very difficult to reach with 2D information with such dramatic differences in posture. Images in Figure 4.24(a) and Figure 4.24(b) show the frontal and profile dataset of a subject. Despite almost half of the face in Figure 4.24(b) missing, the distance between the surfaces after registration is small. As mentioned in Section 4.3.1, because surfaces with different sizes were compared, parts of the faces that do not overlap are ignored and thus, such different surfaces can be dealt with. When profile datasets belonging to different subjects are registered to each other the differences between them are greater as seen in image Figure 4.24(d). In addition, when texture is used, the rank 1 rate starts dropping (Figure 4.12 and Table 4.1). The same is observed with the $P_{FA} = P_{DI}$ and the P_V rates. This is due to the differences in posture between the datasets. This difference due to posture between surfaces is minimized using a rigid registration, but the illumination differences are not dealt with as the 2D texture maps are not pre-processed. Therefore, the use of texture has detrimental effects on recognition tests using profile faces as probes. Figure 4.25 shows the differences between the aligned textures of the frontal and a profile 2D shot. The textural differences are visibly greater than the images in 4.21. The colored area is the area where the surfaces do not overlap and is ignored as it was not included in the metric.



Original PIE images

Processed PIE images

Figure 4.23: Preprocessing 2D images of the PIE database in order to normalize illumination variation (from [78]).

The textural differences due to illumination can in principle be corrected by performing some normalization on the texture images themselves. A relatively simple preprocessing algorithm is presented in Gross and Brajovic [78] that compensates for illumination variations in images. Taking a single image, the algorithm first estimates the illumination field and then compensates for it in order to recover most of the scene reflectance (Figure 4.23). The technique does not require any training steps or any knowledge of statistical face models. The technique was applied before using several standard 2D face recognition algorithms on many databases and they demonstrated large performance improvements.

In Section 4.3.1 it was reported how an initial estimate of position is performed in order to bootstrap the ICP algorithm. This starting position can be used to infer the angle at which the data was captured. Assuming that the light source does not change location and given a known head posture (profile or frontal), one can compensate for the difference in illumination across poses by using a model as presented in [129]. We did not conduct experiments with such normalized data since the scope of our study was to investigate the effects of posture on shape and texture and not to tackle the illumination variability problem.

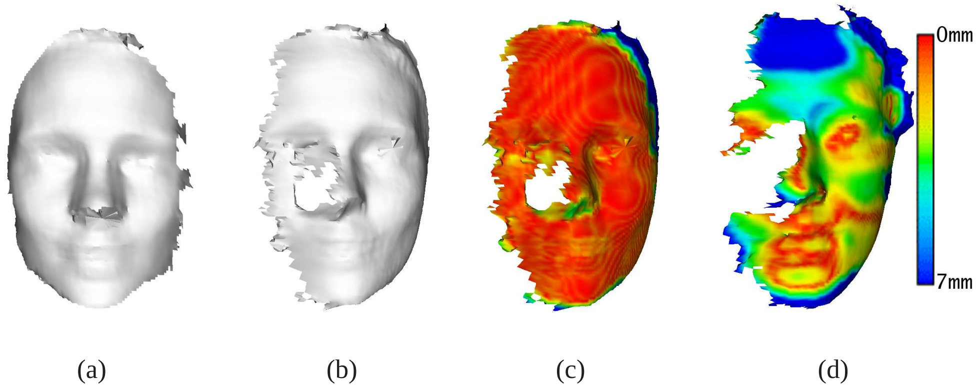


Figure 4.24: (a) shows a frontal dataset, (b) a profile dataset from the same subject, (c) is a color map of the distances between them after registration while (d) is the color map of distances between (a) and a profile dataset belonging to another subject.

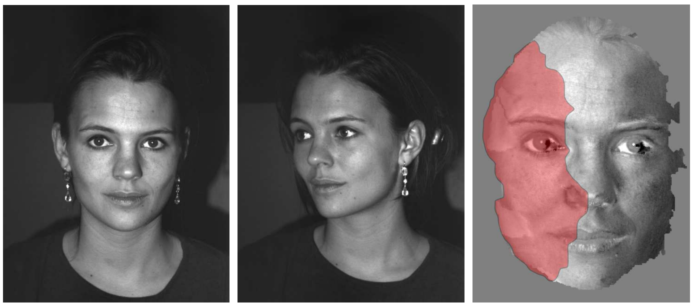
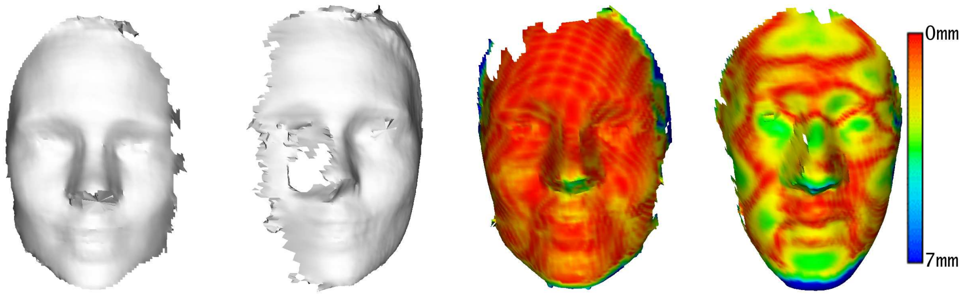


Figure 4.25: The two images on the left are the original images of the subject. The third image shows the differences in illumination between the two images after they have been registered. The red area indicates the part of the gallery face that does not overlap with the profile probe face and is therefore not taken under consideration in the similarity metric.



(a) (b) (c) (d)

Figure 4.26: (a) shows a frontal dataset, (b) a tilted dataset from the same subject, (c) is a color map of the distances between them after registration while (d) is the color map of distances between (a) and a tilted dataset belonging to another subject.

4.6.1.3 Tilted faces

When upward-tilted faces are used as probes the rates are very high when no texture is used (Table 4.5 and Figures 4.16, 4.17, 4.18 and 4.19). As already mentioned the VRT3D system requires both stereo cameras to "see" the same part of the face for the software to be able to infer the 3D shape. An upwards tilted face does not contain as many occlusions as a profile face. Assuming that a face is a cylinder, the height h of the cylinder (from forehead to chin) is larger than the diameter d (from one ear to the other). In other words the face's principal curve is the one between the ears and as a result, occlusion occurs at smaller angles when the face is turned from left to right than if it is tilted upwards or downwards. Consequently, the surfaces of a frontal and a tilted face do not differ significantly. Once the faces are registered they look very similar in 3D. Figure 4.26 shows a frontal and a upwards tilted face and the small residual distance between the two once they have been registered. However, as with profiles, when texture is used the recognition rate falls sharply. This is because as mentioned above changes in posture cause illumination differences between faces. Figure 4.27 shows a frontal and tilted face of the same subject and the difference between them after their textures have been aligned. Once again the illumination differences in the textures are greater than the ones shown in Figure 4.21.

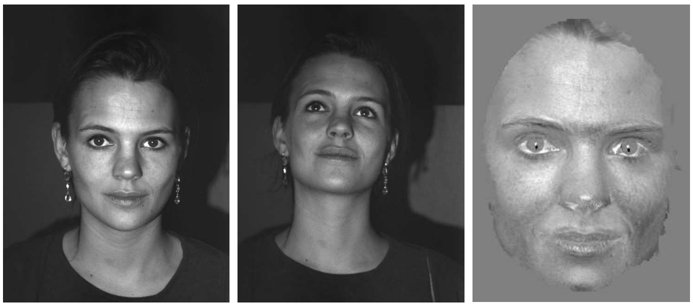


Figure 4.27: The two images on the left are the original images of the subject. The third image shows the differences in illumination between the two images after they have been registered.

4.6.1.4 Faces with expressions

When the faces used as probes have facial expressions then the actual geometry of the surface is different. These non-rigid deformations violate the rigid-body assumption presented earlier. A smiling or frowning face has a different geometry around the parts of the face where muscles are stretching and pulling the soft tissue. This results in reduced recognition rates compared to emotionally neutral frontal faces. Figure 4.29(a) shows a frontal gallery face and image (b) of the same figure shows a smiling probe belonging to the same subject. In Figure 4.29(c) it is clear that the area around the mouth is the area where most differences between probe and gallery face are found while, in Figure 4.29(d) the differences between the neutral gallery face and a smiling probe belonging to a different person are evident. Similarly in Figure 4.30 showing the surface of a frowning face it is clear that most of the differences are located around the eyebrows and around the mouth. The similarity score gap between faces (c) and (d) in Figures 4.29 and 4.30 is smaller than in experiments where identity is the major source of difference between a probe and a gallery. Both different expression and different identity increase the 3D distance between faces, thus reducing the recognition rate. Since, however there are no big differences in posture there are also no dramatic differences in illumination. There are some local differences in the areas where the surfaces are deformed due to a facial expression, which causes the light to be reflected differently, but these differences are not enough to render the texture map useless. As seen in Figures 4.28 and 4.31 there are

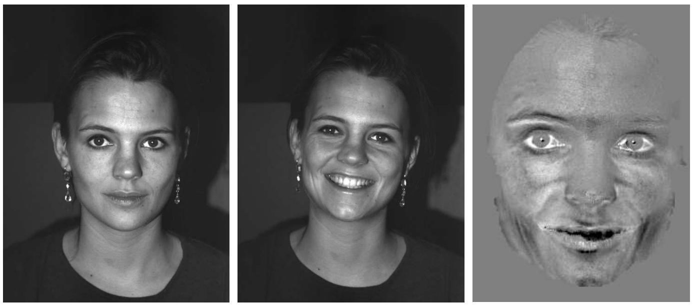


Figure 4.28: The two images on the left are the original images of the subject. The third image shows the differences in illumination between the two images after they have been registered. Notice the difference in the non-rigid parts of the face that are active during a smile.

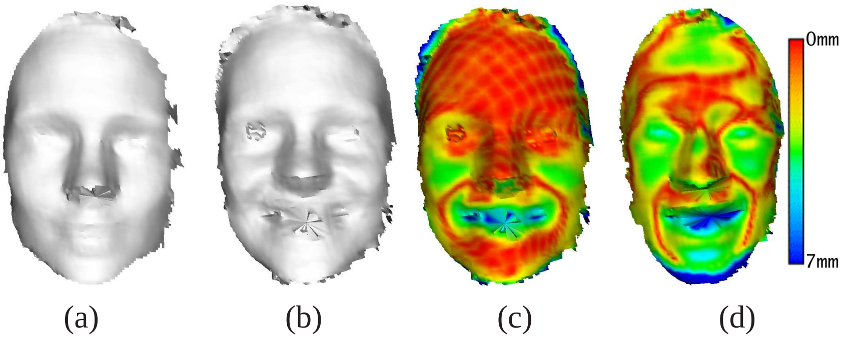


Figure 4.29: (a) shows a frontal (neutral) dataset, (b) a smiling dataset from the same subject, (c) is a color map of the distances between them after registration while (d) is the color map of distances between (a) and a smiling dataset belonging to another subject.

differences in the textures but especially on the frowning example 4.31 they are mostly located around the eyebrows where the surface deformation is greater. Also the local surface deformations of the average smiling face in the VRT3D datasets are more dramatic than the surface deformations on a frowning face (partly because of the reluctance of subjects in a multicultural environment to express sorrow as enthusiastically as joy). This partly explains the significantly higher recognition rates achieved with frowning probes compared to smiling ones.

4.6.2 Notre Dame data

High recognition rates were also achieved with the Notre Dame database using just the surface information. As table 4.5 shows, all 150 faces were matched correctly in rank 1

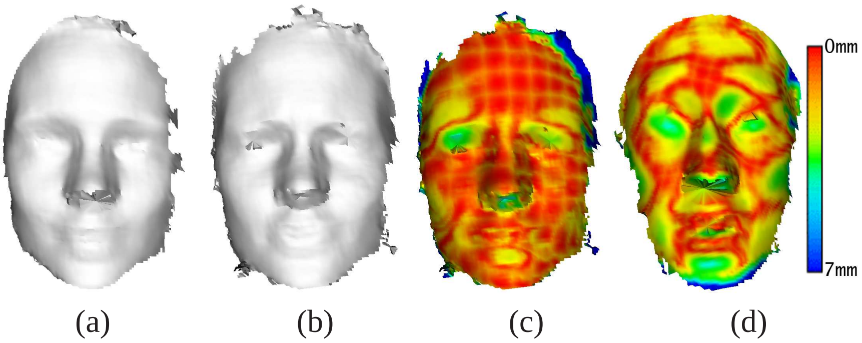


Figure 4.30: (a) shows a frontal (neutral) dataset, (b) a frowning dataset from the same subject, (c) is a color map of the distances between them after registration while (d) is the color map of distances between (a) and a frowning dataset belonging to another subject.

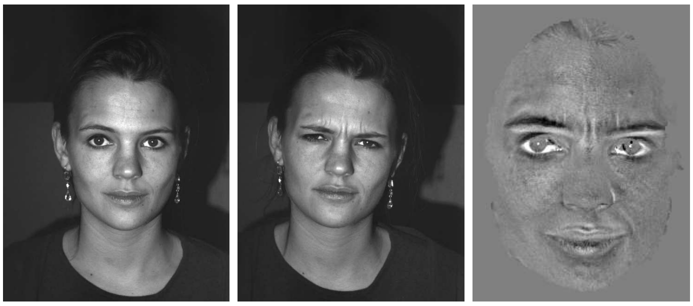


Figure 4.31: The two images on the left are the original images of the subject. The third image shows the differences in illumination between the two images after they have been registered. Notice the non-rigid parts of the face that are active during a frown.

tests. The $P_{DI} = P_{FA}$ rate is also very high, showing that the average surface similarity between two correct matches was significantly smaller than the average surface similarity between the smallest incorrect match. In other words the average client yields a significantly closer match than the average best imposter. When texture is used the rates fall sharply. This is because the datasets of the Notre Dame face database were not captured under the controlled conditions that the VRT3D datasets were captured. One can easily notice that there often are great variations in illumination between images of the same subject. This is either due to different light sources being used across sessions or due to the subject changing position with regards to the light source (or both). Figure 4.33 (top) shows these kind of illumination differences even when the position of the subject is relatively the same across images. The image on the right shows the result when the two images are subtracted from each other. Figure 4.33 (bottom row) shows an example of variation in position with respects to the camera and the consequences it has on the illumination of the face. It is important to note that these dramatic differences in illumination do not exist in frontal scans of the VRT3D set as illustrated in Figure 4.21. Part of the reason for achieving such uniformity in the VRT3D datasets is that contrary to the Notre Dame database, all images of each subject were captured within a couple of minutes from each other, which significantly reduces the within-subject variability. As a result when more and more weight is put on texture with the Notre Dame datasets the average best imposter score is significantly closer to the average score of the clients and at some point even outperforms the latter. Figure 4.32 shows the average client score and the average best imposter score, demonstrating that as the texture weight increases so do the scores of the average client and the average best imposter. Note that when a texture weight greater than 0.12 is used a crossover occurs and the average client performs worse than the average best imposter.

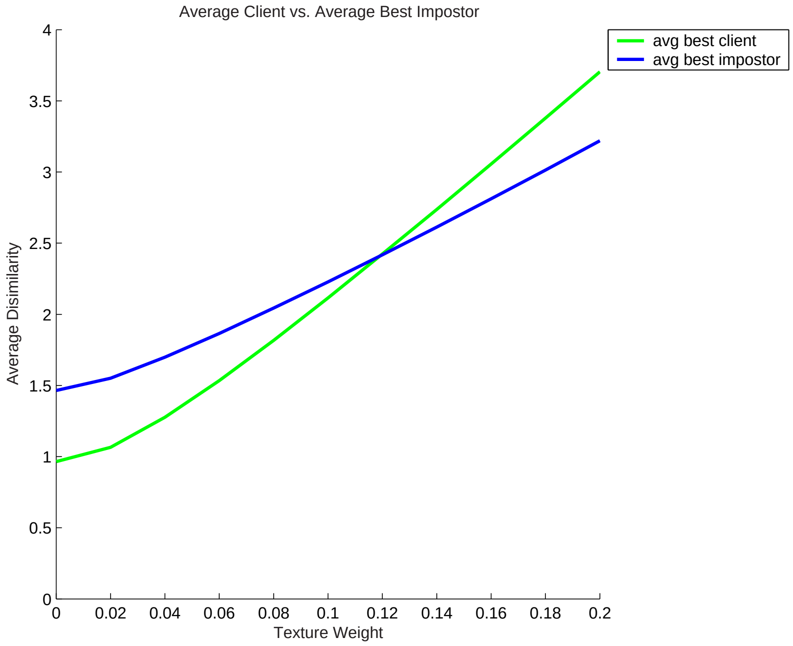


Figure 4.32: Average client score plotted against the average best imposter score. Notice how after a certain texture weight the average best imposter scores better than the average client (0=identical).

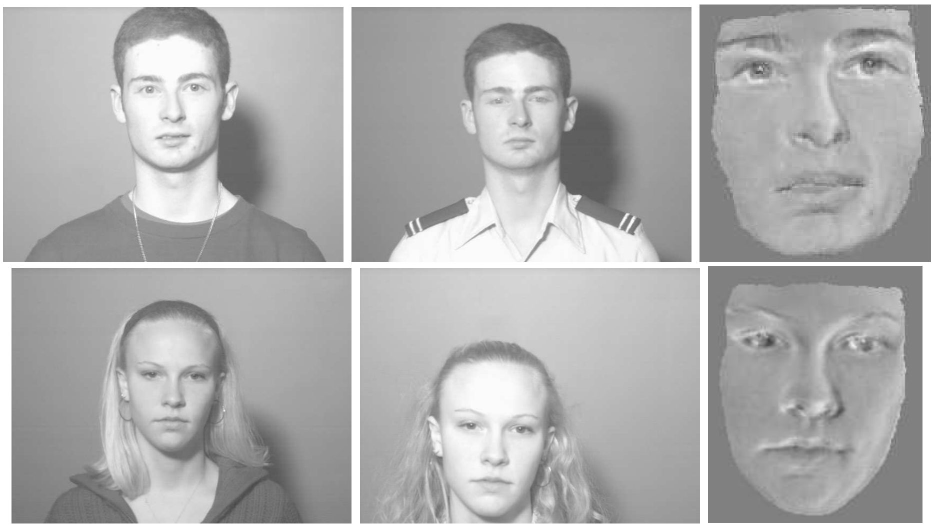


Figure 4.33: The images on the left are the original images of the Notre Dame subject while the third one shows the differences in illumination between the first two images after they have been registered.

4.7 Conclusions

In this chapter the usefulness of three-dimensional information for face recognition is demonstrated. Despite dramatic postural differences between the datasets of each subject very high recognition rates were achieved. Although the surface-only recognition rates are not affected by postural differences, when 2D texture information is used alongside the 3D shape in the similarity metric, the recognition rate is significantly affected. Furthermore lower rates are achieved when the subjects engaged in a facial expression as the surface geometry was affected. Both the issue of posture and the facial expressions can be dealt with to decrease their effect. As mentioned in Section 4.6 the textures of the faces were not pre-processed. It is likely that pre-processing the 2D texture maps with a technique, such as the ones proposed in [78], would improve the recognition scores. In order to deal with facial expressions on the other hand, one could account for differences by landmarking the non-rigid anatomical features and bringing them into some sort of close alignment across all subjects. Finally, using 4D registration does not improve the results. A 4D ICP performed significantly worse in aligning faces, independently of the texture weighting used. The increase in misregistrations of facial surfaces is caused by a relatively great number of local minima in the texture data which causes incorrect point correspondences to be formed. As a result the recognition rates after using a 4D ICP were much lower than using a 3D ICP.

As mentioned in Chapter 2 Medioni *et al.* [135] used a similar ICP-based technique which reached 90% rank 1 rates using 100 subjects. His method however relied only on 3D surface information and not 2D. Maurer *et al.* [133] also used an ICP-based technique but during the facial comparison stage they implemented a weighted sum technique, like the one presented in this chapter, taking advantage of both 2D and 3D information. When using neutral images in the probe and gallery set the verification rate at FAR$= 1\%$ reaches 99.2%. What is more important, however, is that using sets containing facial expressions as well as neutral faces, it reaches a verification rate of 93.5%. Lu *et al.* [127] focused on making the registration faster and more robust. In the evaluation stage, apart from point location and texture they also used the shape index at each point. Using 18 faces some of which were semi-profile and with facial expressions they reached a rank 1 rate of 92%. Finally, Lu and Jain [128] tried to deal with facial expressions by using thin plate splines to model the intra- and inter-subject variation. Using these models they tried to distinguish deformations due to identity from deformations due to facial expressions. On a database which included faces with facial expressions they managed to reach rank 1 rates of 89% using 3D data only and 91% using 3D+2D.

The method introduced in this chapter compares faces by using all facial points. No effort is made to represent the faces in a more compact way by taking advantage of the inherent redundancy that exists in such data. Using such high-dimensional input is computationally costly. Subspace analysis could potentially eliminate the redundancy and isolate the modes of variation that are important for recognition. Furthermore, the technique presented here is used on two relatively small databases and therefore, it is difficult to predict how such a technique would fair with a very large set of data where the likelihood of an imposter with a mean square distance smaller than the correct match increases. An alternative representation of reduced dimensionality could prove more effective. In the next chapter we introduce and compare two model-building techniques for face recognition based on PCA.

Chapter 5

Automatic construction of statistical 3D face models for face recognition

5.1 Introduction

Given the redundancy in facial data, a common approach to face recognition is to use statistical face models. Two statistical model-building techniques are presented in this chapter and their effectiveness in 3D face recognition is evaluated and compared. The motivation behind the use of statistical models is the fact that it can potentially allow a more compact description of faces and thus decrease the vulnerability of face recognition algorithms, such as the one presented in Chapter 4, to noise. Such an increased robustness would also allow the scaling of face recognition techniques to larger databases. Furthermore, even for a small database of 1000 subjects, using the method of the previous chapter would require 1000 rigid registrations before a face can be compared to the rest of the gallery faces. This is computationally very expensive. The techniques presented in this chapter require only one registration to a template face in order to minimize variation due to pose. Then, all comparisons take place in a reduced dimensionality subspace which allows for face-to-face comparisons to happen in a fraction of the time.

In this chapter we propose the use of two registration techniques for building models of 3D face recognition. The first technique is a landmark-based technique in which landmarks, placed manually on fiducial anatomical points, are used for registration. The landmark points are brought into close alignment using rigid and non-rigid registration and the correspondences are established between a base mesh, or template, and the subject's face. The other technique implemented is an automatic rigid surface registration used to establish correspondences between surface points without the use of manually placed landmarks. Once correspondences are established using either techniques, statistical models of 3D faces are created using PCA. The two models were tested using the face recognition protocol described in Chapter 4. Furthermore three important properties of the model are examined: its generalization ability, its specificity and its compactness.

5.2 Principal component analysis revisited

Early attempts in face recognition used local face features in order to describe a face [108, 74], but these techniques have proven largely ineffective as they lack robustness [205]. Given the structural regularity of the faces, one can exploit the redundancy in order to describe a face with less parameters. In intensity images the dimensionality of the space depends on the number of pixels in the input images while in 3D data it depends on the number of points on the surface. Let us assume a set of 3D face surfaces Γ_1, Γ_2, Γ_3,...,Γ_M, each with n surface points. The average 3D face surface is then calculated by:

$$\overline{\Gamma} = \frac{1}{M}\sum_{i=1}^{M}\Gamma_i \tag{5.1}$$

and using the vector difference

$$\boldsymbol{\gamma}_i = \Gamma_i - \overline{\Gamma} \tag{5.2}$$

the covariance matrix $\boldsymbol{C}$ is computed by:

$$\boldsymbol{C} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^T \tag{5.3}$$

The eigenanalysis of $\boldsymbol{C}$ yields the eigenvectors $\boldsymbol{u}_i$ and their associated eigenvalues $\boldsymbol{\lambda}_i$ sorted by decreasing eigenvalue. All surfaces are then projected on the facespace by:

$$\boldsymbol{\beta}_k = \boldsymbol{u}_k^T(\Gamma - \overline{\Gamma}) \tag{5.4}$$

where $k = 1, ..., m$. Every surface can then be described by a vector of weights $\boldsymbol{\beta}^T = [\beta_1, \beta_2, ..., \beta_m]$, which dictates how much each of the principal eigenfaces contributes in describing the input surface. The value of m is application and data-specific, but in general a value is used such that 98% of the population variation can be described. More formally [96]:

$$\frac{\sum_{k=1}^{m} \boldsymbol{\lambda}_k}{\sum_{j=1}^{M} \boldsymbol{\lambda}_j} \geq 0.98 \tag{5.5}$$

The similarity between two faces A and B can be assessed by comparing the weights $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_B$ which are required to parameterize the faces. We will use two measurements for measuring the distance between the shape parameters of the two faces. The first one is the Euclidean distance which is defined as:

$$d_E(\boldsymbol{\beta}_A, \boldsymbol{\beta}_B) = ||\boldsymbol{\beta}_A - \boldsymbol{\beta}_B|| = \sqrt{\sum_{i}^{m}(\beta_{Ai} - \beta_{Bi})^2} \tag{5.6}$$

Turk and Pentland also calculated the distance of a face from the feature-space. There are then four possibilities for an input image: (1) the face is near the feature-space and near a face class (the face is known), (2) near feature-space but not near a face class (face is unknown), (3) distant from feature-space and face class (image not a face) and finally (4) distant from feature-space and near a face class (image not a face). This way images that are not faces can be detected. Typically case (3) returns a false positive in most recognition systems.

By computing the sample variance along each dimension one can use the Mahalanobis distance to calculate the similarity between faces [221]. In the Mahalanobis space, the variance along each dimension is normalized to one. In order to compare the shape parameters of two facial surfaces the difference in shape parameters is divided by the corresponding standard deviation σ:

$$d_M(\boldsymbol{\beta}_A, \boldsymbol{\beta}_B) = \sqrt{\frac{\sum_i^m (\beta_{A_i} - \beta_{B_i})^2}{\sigma_i^2}} \tag{5.7}$$

Another alternative is to use the MahCosine distance metric as reported in Chawla and Bowyer [42]. The MahCosine measure is the cosine of the angle between the biometric samples after they have been transformed to the Mahalanobis space. More formally, the MahCosine measure between images Γ_A and Γ_B with projects $\boldsymbol{\beta_A}$ and $\boldsymbol{\beta_B}$ in the Mahalanobis space is computed as:

$$d_{MahCosine}(\Gamma_A, \Gamma_B) = cos(\theta_{\Gamma_A,\Gamma_B}) = \frac{|\boldsymbol{\beta_A}||\boldsymbol{\beta_B}|cos(\theta_{\Gamma_A,\Gamma_B})}{|\boldsymbol{\beta_A}||\boldsymbol{\beta_B}|} \tag{5.8}$$

Some people report improved results with distance metrics other than the basic Mahalanobis and Euclidean distance [42] but in order to increase the comparability of our results we have opted for the more widely used former two distance metrics.

5.3 Building the face model

One of the key problems in computer vision has been the correspondence problem. As Vetter *et al.* [170] have shown in 2D, the linear combination of pixels does not result in a truly average face (Figure 5.1). The reason for this is that the pixel-wise linear combination does not account for variations of shape. Many face modeling techniques [50, 113, 97] assume that the data, whether 3D or 2D forms a linear vector space. However, the data is not always in that form. The 2D or 3D face data are usually of various geometries and sizes with varying number of pixels or points. The correspondence problem in this case is the problem of finding points on the facial surface that correspond, anatomically speaking, to the same surface points in other faces. In other words, the ith point on every surface should correspond to the same feature across all faces [15]. Early statistical approaches for describing a face did not attempt to match features together [205, 110]. It was assumed that faces were photographed under controlled conditions and no explicit

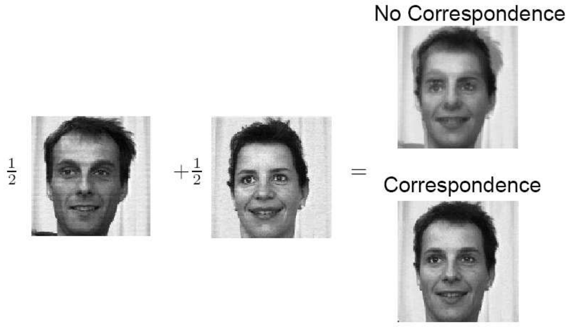


Figure 5.1: A mean calculated with and without correspondences. (from [170])

attempt was made to match corresponding areas. Later research focused on establishing correspondence between features of objects in order to generate more meaningful face models.

One way to establish correspondence is by using manually placed landmarks to mark anatomically distinct points on a surface. As this can be a painstaking and error-prone process some authors have tried to automate it by using a model trained on manually placed landmarks and employing it to find landmarks on other surfaces [48, 96, 64, 51].

Davies [51] developed a framework for establishing correspondence between datasets automatically by treating this task as part of the learning process. In order to achieve this, an objective function was established which measured the utility of a model based on the minimum description length principle. The problem of finding correspondences is then treated as a problem in which correspondences are manipulated in order to optimize the objective function.

Another way of establishing correspondence between points on two surfaces, as discussed in Chapter 3, is by analyzing their shape. Wang *et al.* [214] used curvature information to find similar areas on a surface in order to construct 3D shape models. In Brett *et al.* [26, 25] the surfaces were decimated in a way that eliminates points from areas of low curvature. High curvature areas are then assumed to correspond to each other and are thus aligned.

Contrary to such techniques which use a small number of corresponding feature points, the ICP algorithm aligns shapes by minimizing the sum of squared distances between closest points on the surfaces. Using automatic techniques like these, points on one surface are matched with closest points on the other surface. Some work has been done on combining automatic techniques such as ICP with a semi-automatic statistical technique, such as active shape models, in order to take advantage of the strengths of each [96]. An example of such a technique is disccussed in the concluding section of this chapter.

Finally, as discussed in earlier chapters Vetter *et al.* [209] established correspondence between 3D facial surfaces by using optical flow on 2D textures to match anatomical features to each other. In the next sections the two methods for building the face model are presented followed by a detailed comparison of the two.

5.3.1 Building the model using landmark registration

Initially, the facial surfaces are visualized in 3D using a publicly available visualization library (www.vtk.org). Landmarks are then manually placed on features of the face using the mouse to select points on the 3D surface. The face could be rotated in space which made the selection of the correct feature somewhat easier. The use of texture on the faces helped the selection of the correct landmark further. The images were landmarked by two undergraduate volunteers and each of them landmarked different parts of the database. Since there was no overlap between the datasets they worked on, we did not perform any tests to check the consistency of their efforts. However, all faces were checked in the end of the above process by the author in order to correct some inconsistencies and reduce errors.

Parts of the face such as the cheeks are difficult to landmark because there are no uniquely distinguishable anatomical points across all faces. The landmarks used were placed on anatomically distinct points of the face in order to ensure proper correspondence. Some of the landmarks were placed on the “outer” anatomical features such as chin and eyebrows in order to better capture the overall dimension of a face. It was important to choose landmarks that contain both local feature information (eg. the size of the mouth and nose) as well as the overall size of the face (eg. the location of the eyebrows).

Anatomical points landmarked	
Points	**Landmark Description**
Glabella	Area in the center of the forehead between the eyebrows, above the nose which is slightly protruding (1 point).
Eyes	Both the inner and outer corners of the eyelids are landmarked (4 points).
Nasion	The intersection of the frontal and two nasal bones of the human skull where there is a clearly depressed area directly between the eyes above the bridge of the nose (1 point).
Nose tip	The most protruding part of the nose (1 point).
Subnasal	The middle point at the base of the nose (1 point).
Lips	Both left and right corners of the lips as well as the top point of the upper lip and the lowest point of the lower lip are landmarked (4 points).
Gnathion	The lowest and most protruding point on the chin (1 point).

Table 5.1: The 13 manually selected landmarks chosen because of their anatomical distinctiveness.

Previous work on 3D face modeling has shown that there is not much difference between the use of 11 and 59 landmarks [96]. It was therefore decided that 13 landmarks are sufficient for building the model while making the landmarking process less tedious. On average it took about a minute to landmark a single face accurately. Table 5.1 shows the landmarks that were used and Figure 5.2 shows an example of a face that was manually landmarked.

5.3.1.1 Rigid landmark registration

Initially the mean landmarks were calculated by registering all landmark sets to a landmark set belonging to one of the subjects using the least square approach presented in Arun *et al.* [6]. Subsequently, the mean position of each landmark is computed. All datasets are then re-registerered to this mean landmark set. The transformations generated from the registration of each landmark set to the mean landmarks is applied to each facial surface. As a result, all faces are brought into close alignment using only the landmarks to calculate the rigid transformation. Figure 5.3 (top row) shows two faces aligned to the mean landmarks while the bottom row shows a frontal 2D projection of the outer landmarks of the same faces before and after rigid landmark registration.

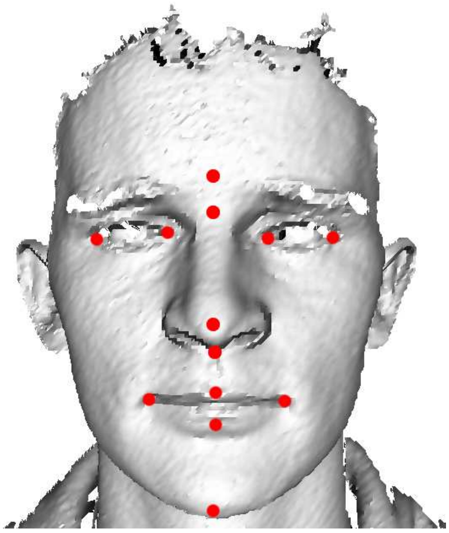


Figure 5.2: The 13 manually selected landmarks chosen because of their anatomical distinctiveness.

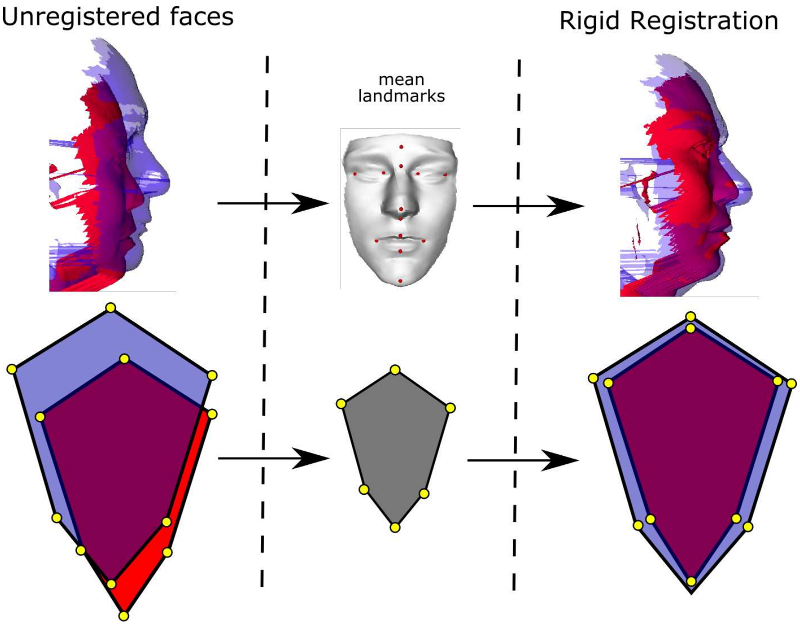


Figure 5.3: Rigid registration of faces using landmarks. The top row shows the two faces aligned to the mean landmarks. The bottom row shows a frontal 2D projection of the outer landmarks of the same faces before and after registration.

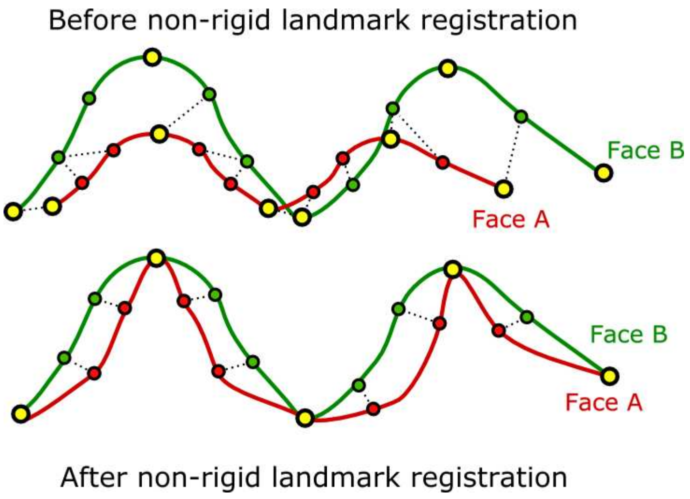

Before non-rigid landmark registration
Face B
Face A
Face B
Face A
After non-rigid landmark registration

Figure 5.4: Before and after non-rigid landmark registration. Yellow points represent landmarks.

5.3.1.2 Non-rigid landmark registration

The next step involves the use of non-rigid registration in order to maximize the correspondence between the faces. This is necessary, because after rigid registration facial features are not always paired-up with the corresponding points on the other surface. Figure 5.4 shows two idealized surfaces before and after non-rigid landmark registration. The yellow points represent the manually placed landmarks that were deemed to correspond to each other. Because the surfaces are not of the same size and shape, even after being closely aligned using a rigid registration (Figure 5.4 (top)), many points are not paired up with the correct points on the other surface. Registering the landmarks non-rigidly (Figure 5.4 (bottom)) allows for corresponding points to be paired up.

Thin plate splines (TPS) can be used in order to warp each set of landmarks to the mean landmarks [169]. For example, Hutton [96] used TPS to accurately non-rigidly register all landmarks and interpolate the alignment in the space between the them. TPS use radial basis functions which have infinite support and therefore each control point has a global effect on the entire transformation. The TPS calculation is therefore complex and inefficient. One of the advantages of using a smaller number of landmarks is the decrease in the computational cost of estimating the TPS warping between the two point sets.

An alternative approach for the non-rigid registration of 3D faces is to use a so-called

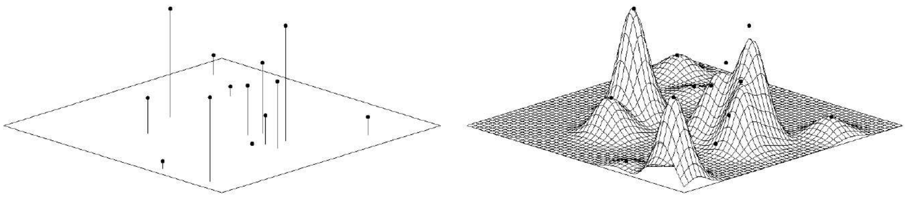


Figure 5.5: Example of B-spline approximation. The image on the left shows the input data and the image on the right the approximation of the control grid [116].

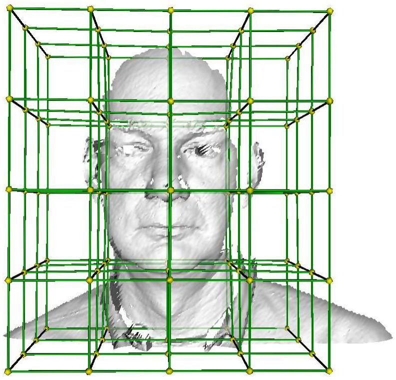


Figure 5.6: A free-form deformation and the corresponding mesh of control points.

free-form deformation (FFD) [183] which can efficiently model local deformations. As discussed in Chapter 3, B-spline transformations, contrary to thin-plate splines, have local support, which means that each control point influences a limited region. Furthermore, the computational complexity of calculating a B-spline is significantly lower than a thin-plate spline. In the following, a non-rigid registration algorithm for landmarks based on multi-resolution B-splines is proposed.

Lee *et al.* [116] described a fast algorithm for interpolating and approximating scattered data using a coarse-to-fine hierarchy of control lattices in order to generate a sequence of bicubic B-spline function whose sum approximates the desired interpolation function. Figure 5.5 shows an example of a B-spline approximation. We extend this method in order to calculate the ideal free-form transformation for 3D landmarks. A rectangular grid of control points is initially defined (Figure 5.6) inside which the set of landmarks are placed. The control points of the FFD move in order to precisely align the facial landmarks placed within the grid to the corresponding landmarks of another face.

The transformation is defined by a $n_x \times n_y \times n_z$ grid $\mathbf{\Phi}$ of control point vectors $\boldsymbol{\phi}_{lmn}$ with uniform spacing δ:

$$\boldsymbol{T}(x, y, z) = \sum_{i=0}^{3}\sum_{j=0}^{3}\sum_{k=0}^{3} B_i(r)B_j(s)B_k(t)\boldsymbol{\phi}_{l+i,m+j,n+k} \tag{5.9}$$

where $l = \lfloor \frac{p_x}{\delta} \rfloor - 1, m = \lfloor \frac{p_y}{\delta} \rfloor - 1, n = \lfloor \frac{p_z}{\delta} \rfloor - 1, r = \frac{p_x}{\delta} - \lfloor \frac{p_x}{\delta} \rfloor, s = \frac{p_y}{\delta} - \lfloor \frac{p_y}{\delta} \rfloor$ and $t = \frac{p_z}{\delta} - \lfloor \frac{p_z}{\delta} \rfloor$ and where B_i, B_j, B_k represent the B-spline basis functions which define the contribution of each control point based on its distance from the landmark [63]:

$$\begin{aligned} B_0(u) &= (1-u)^3/6 \\ B_1(u) &= (3u^3 - 6u^2 + 4)/6 \\ B_2(u) &= (-3u^3 + 3u^2 + 3u + 1)/6 \\ B_3(u) &= u^3/6 \end{aligned}$$

Given a moving point set (source) $\boldsymbol{p} = \{(p_{e_x}, p_{e_y}, p_{e_z})\}$ and a fixed point set $\boldsymbol{q} = \{(q_{e_x}, q_{e_y}, q_{e_z})\}$, the algorithm estimates a set of displacement vectors $\boldsymbol{d} = \boldsymbol{p} - \boldsymbol{q}$ associated with the latter. The output is an array of displacement vectors $\boldsymbol{\phi}_{lmn}$ for the control points which provides a least squares approximation of the displacement vectors.

Since B-splines have local support, each source point $\boldsymbol{p_e}$ is affected only by a limited number of surrounding 64 control points. The displacement vectors of the control points associated with that part of the lattice can be denoted as $\boldsymbol{\phi}_{ijk}$ where $i, j, k = 0, 1, 2, 3$ and are given by:

$$\boldsymbol{\phi}_{ijk} = \frac{w_{ijk}\boldsymbol{d}}{\sum_{a=0}^{3}\sum_{b=0}^{3}\sum_{c=0}^{3} w_{abc}^2} \tag{5.10}$$

where $w_{ijk} = B_i(r)B_j(s)B_k(t)$. In other words, eq. 5.10 entails that the closer a control point is to a source point $\boldsymbol{p}_e$, the larger the former's associated w_{ijk} value and therefore, the larger the influence of the corresponding displacement vector. Listing 2 is a pseudocode representation of the B-spline registration algorithm.

Because B-splines have local support, if the distance between the control points is too

Listing 2 The 3D B-spline registration algorithm

```
Input: a source landmark set $\boldsymbol{p} = \{(p_x, p_y, p_z)\}$ and
corresponding displacement vectors $\boldsymbol{d} = \{(d_x, d_y, d_z)\}$
Output: control lattice $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_{lmn}\}$
for all $l, m, n$ do
    reset $\boldsymbol{\Delta}_{lmn}$ and set $\boldsymbol{\omega}_{lmn} = 0$
end for
for each point $(p_x, p_y, p_z)$ in $\boldsymbol{p}$ do
    let $l = \lfloor \frac{p_x}{\delta} \rfloor - 1, m = \lfloor \frac{p_y}{\delta} \rfloor - 1, n = \lfloor \frac{p_z}{\delta} \rfloor - 1$
    let $r = \frac{p_x}{\delta} - \lfloor \frac{p_x}{\delta} \rfloor, s = \frac{p_y}{\delta} - \lfloor \frac{p_y}{\delta} \rfloor, t = \frac{p_z}{\delta} - \lfloor \frac{p_z}{\delta} \rfloor$
    compute $w_{ijk}$ and $\sum_{a=0}^{3} \sum_{b=0}^{3} \sum_{c=0}^{3} w_{abc}^2$
    for $i, j, k = 0, 1, 2, 3$ do
        compute $\boldsymbol{\phi}_{ijk}$ with eq. 5.10
        add $w_{ijk}^2 \boldsymbol{\phi}_{ijk}$ to $\boldsymbol{\Delta}_{(l+i)(m+j)(n+k)}$
        add $w_{ijk}^2$ to $\boldsymbol{\omega}_{(l+i)(m+j)(n+k)}$
    end for
end for
for all $l, m, n$ do
    if $\boldsymbol{\omega}_{lmn} \neq 0$ then
        compute $\boldsymbol{\phi}_{lmn} = \boldsymbol{\Delta}_{lmn} / \boldsymbol{\omega}_{lmn}$
    else
        $\boldsymbol{\phi}_{lmn} = 0$
    end if
end for
```

small, then the transformation will be affecting only a small area of the surface. If on the other hand, the distance between control points is great (i.e. not many control points are used), then the transformation will be more global and will not allow for a precise alignment of the surface.

In order to avoid these problems, two extensions were implemented: The first is a multilevel version of the B-spline approximation as presented in Lee *et al.* [116] where an initial coarse grid is used initially and then iteratively subdivided to enable closer and closer approximation between two point sets. Before every subdivision of the grid the current transformation $\boldsymbol{T}$ is applied to points $\boldsymbol{p}$ and the displacement vectors $\boldsymbol{d}$ are recomputed. Listing 3 shows the multilevel implementation of the 3D B-spline registration algorithm while Figure 5.7 shows a grid being subdivided.

Listing 3 The multilevel 3D B-spline registration algorithm

```
Input: a source landmark set $\boldsymbol{p} = \{(p_x, p_y, p_z)\}$ and displacement vectors $\boldsymbol{d} = \{(d_x, d_y, d_z)\}$
Output: a control lattice hierarchy $\boldsymbol{\Phi}_0, \boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_n$
let $g = 0$
while $g \leq n$ do
    let $\boldsymbol{p'} = \boldsymbol{T}(\boldsymbol{p})$
    let $\boldsymbol{d'} = \boldsymbol{p} - \boldsymbol{p'}$
    compute $\boldsymbol{\Phi}_g$ from $\boldsymbol{p'}$ and $\boldsymbol{d'}$ as presented in Listing 2
    let $g = g + 1$
end while
```

The second extension prevents control points from deforming an area excessively. If one defines a control point grid with uniform spacing δ, then the most a control point can move is $\frac{\delta}{2}$. In addition, as with the rigid registration, all landmarks sets are non-rigidly registered to the *mean* landmarks. This ensures that the amount of deformation across all landmark sets are minimized. Figure 5.8 (top row) shows two faces being aligned to the mean landmarks while the bottom row shows a 2D outline of the outer landmarks of the same faces before and after non-rigid landmark registration.

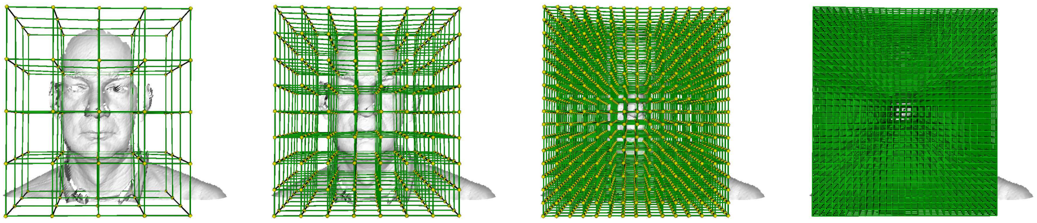


Figure 5.7: Four grid subdivision for landmark registration.

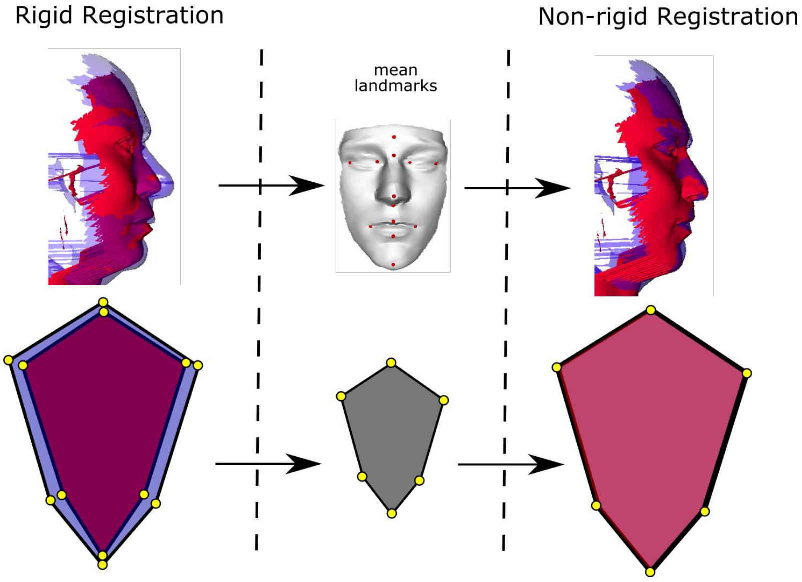


Figure 5.8: Non-rigid registration of faces using landmarks. The top row shows the two faces aligning themselves to the mean landmarks. The bottom row shows a 2D outline of the outer landmarks of the same faces before and after a non-rigid registration.

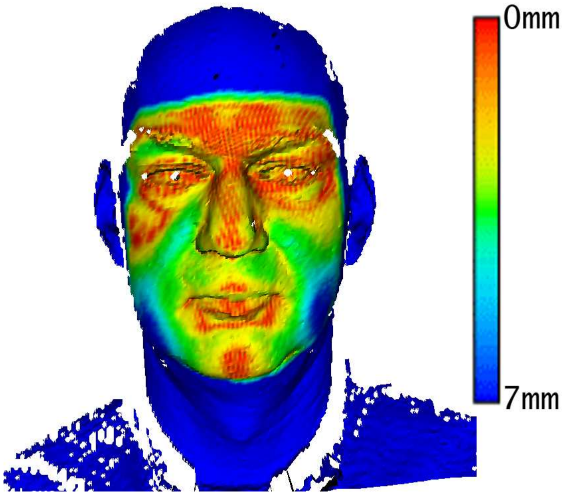


Figure 5.9: Color map of the distances between a face and a template mesh after non-rigid registration of the landmarks.

5.3.1.3 Establishing correspondences

Each rigidly registered face A is transformed using a non-rigid transformation generated from the non-rigid registration of its landmarks to the mean landmarks resulting in A'. A face, which is not part of the population and which is free of artifacts is chosen and further manually pre-processed to make the point distribution on its surface very regular. This face, referred to as the template face, is used later to resample each dataset, the reasons for which are discussed in the next section. The actual template size does not affect the model. Most of the experiments that follow were repeated with a different template and the recognition rates did not change significantly. Just as with all other datasets, the template face B is warped to the mean landmarks by first calculating the B-spline transformation to register the landmarks of the template face to the mean landmark set and applying the resulting transformation to the whole surface of the template face. Face A' and a template face B have now been brought into close alignment particularly near their landmarked areas. Figure 5.9 shows a face where each point is color-coded to show the distance to the closest point on the template mesh. Notice how the area around the landmarked areas is small while the distance in the outer parts of the face, such as the cheeks, is greater. Once the template mesh B is warped to the mean landmarks it is used to resample surface A'. This is done for several reasons: First of all, because each surface has a different number of points, which makes it impossible to use PCA. Thus, the template is used to resample each dataset and represent it with the same number of points. Moreover, non-facial areas (eg. shoulders, hair, ears) are automatically removed. Finally, artifacts such as holes and spikes on each dataset are eliminated. The latter is done automatically since all points on surface B are going to be matched to a point on surface A. What this does in effect is link up points around the edge of a hole into one cell. The holes are not refilled by points but they are covered by a cell. The removal of artifacts reduces the amount of noise that is encoded in the model and thus produces a better model. All of the above are achieved by following a resampling protocol.

For every point $\boldsymbol{b}$ in template mesh B the closest point $\boldsymbol{a'}$ on surface A' is located. Since we would like our statistical shape model to encode the differences in shape between the faces but not the difference in pose, we will use the point $\boldsymbol{a}$ before applying the non-rigid transformation as corresponding point. This results in A'', a surface with the geometry of A but the topology of B. Figure 5.10 gives a schematic illustration of the process of establishing correspondence between the two surfaces. Applying a non-rigid transformation on the two surfaces allows an alignment of surface points that would not have been possible with a mere rigid transformation, as demonstrated earlier in Figure 5.4.

5.3.1.4 PCA of the 3D faces

Figure 5.11 shows the complete pipeline of the aforementioned process. After correspondence has been established and the number of points on each surface is the same, PCA can be used. The following results were constructed using the Notre Dame database made up of 150 subjects with two datasets per subject. We decided to landmark the Notre Dame and not the VRT3D data as the quality of the facial surfaces was superior. As before, both gallery and probe consisted of 150 datasets. Figure 5.12 shows the variance described by each of the significant components of the face data. The first mode describes 51% of the model's variance. Table 5.2 and Table 5.3 show the first four principal components of shape variation between -3, 0 and $+3$ standard deviations using the landmark-based

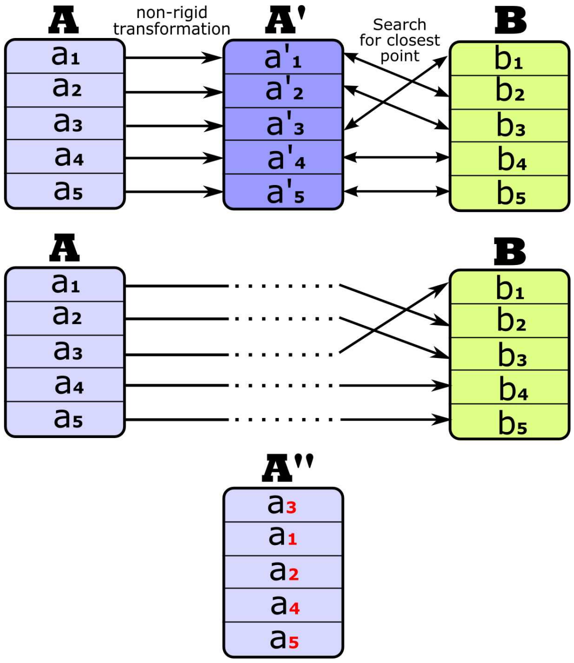


Figure 5.10: The process of establishing correspondence between faces in the landmark-based method. The top row shows face A being non-rigidly transformed into A' and the closest points in B located. In the middle row the coordinates from A are copied over resulting in A'' which has has the connectivity of points in B but the point coordinates of A. All cases of A'' will have the same number of points thus enabling the use of PCA.

registration model. Figure 5.13 shows the population distribution on the first two principal modes of variation. Notice how the population distribution on the two principal axes is unimodal, forming only one peak. This means that there are no sub-clusters in the population distribution and that all faces are spread uniformly in space in the first two dimensions. The arrows show the boundaries of three standard deviations and all faces are represented within these limits.

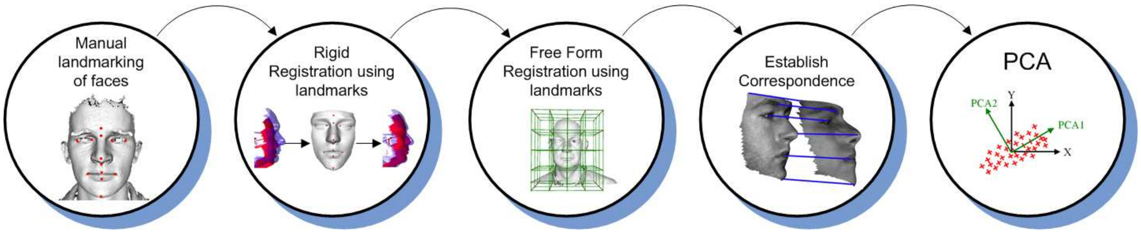

Manual landmarking of faces
Rigid Registration using landmarks
Free Form Registration using landmarks
Establish Correspondence
PCA
PCA2
Y
PCA1
X

Figure 5.11: The process pipeline of the landmark-based registration model.

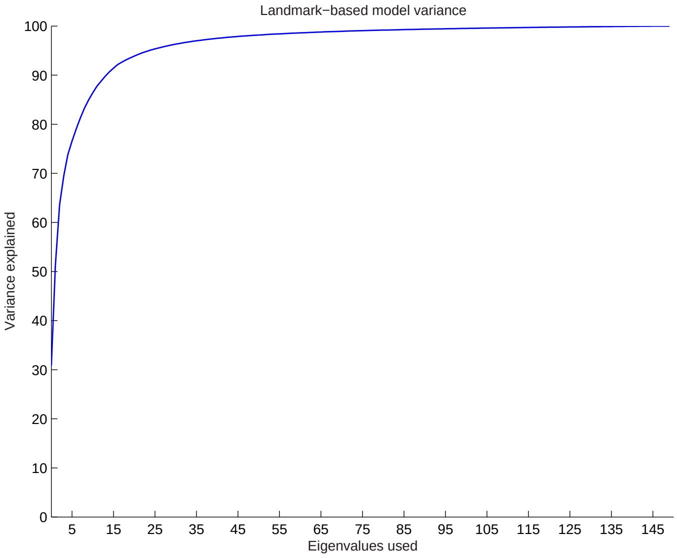


Figure 5.12: The variation described by the modes of the landmark-based statistical face model.

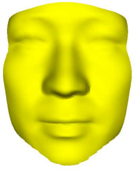
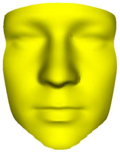
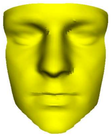
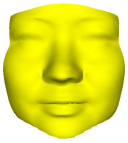
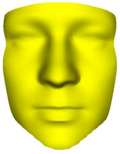
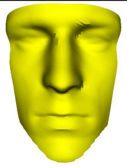
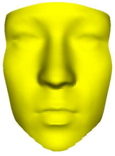
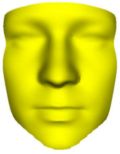
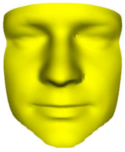
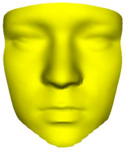
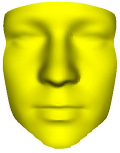
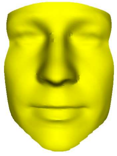
Landmark-Registration-Based Principal Shape Modes			
	$-3\sqrt{\boldsymbol{\lambda}}$	**mean**	$+3\sqrt{\boldsymbol{\lambda}}$
mode 1			
mode 2			
mode 3			
mode 4			

Table 5.2: The first four principal shape modes of the landmark-registration-based model (frontal view).

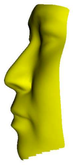
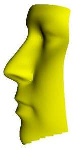
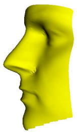
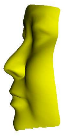
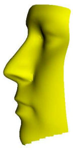
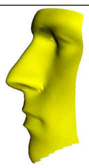
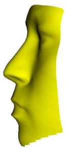
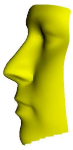
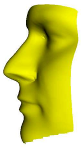
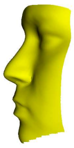
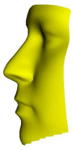
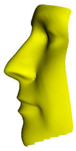
Landmark-Registration-Based Principal Shape Modes			
	$-3\sqrt{\boldsymbol{\lambda}}$	**mean**	$+3\sqrt{\boldsymbol{\lambda}}$
mode 1			
mode 2			
mode 3			
mode 4			

Table 5.3: The first four principal shape modes of the landmark-registration-based model (profile view).

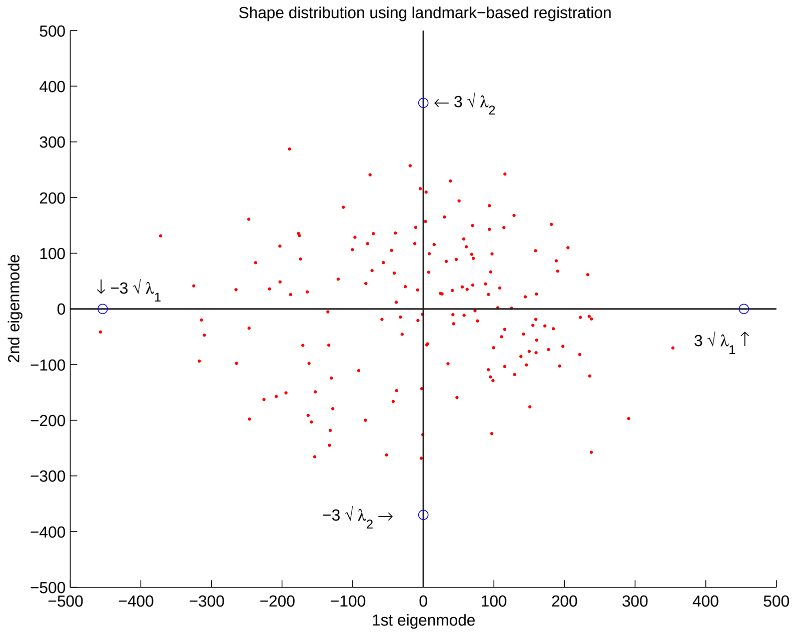


Figure 5.13: The shape distribution of the landmark-based registration model projected on the first two principal modes. The arrows show the boundaries of three standard deviations.

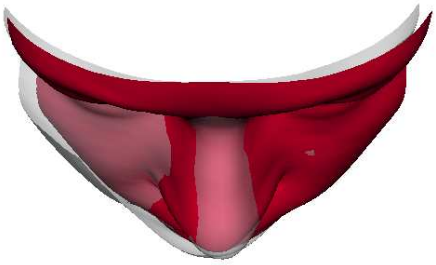


Figure 5.14: Registration errors encoded in the landmark-registration-based model.

5.3.2 Building the model using automatic surface registration

Closer examination of the principal components showed that some were describing positional changes in the face. The statistical face model based on landmarked features was found to encode registration errors generated by the manual selection of features. Figure 5.14 shows the 7th principal component of the model generated using landmarks. It is evident that this mode encodes some rotational component. At the same time the actual facial surface changes insignificantly within that mode. We therefore decided to compare models built by landmark registration with those built by surface registration. For the latter the automatic surface registration algorithm presented in Chapter 4 can be used.

Instead of using landmarks to register the faces to the template the ICP algorithm was employed on the assumption that if all surfaces are rigidly registered to the template mesh, they are also registered to each other. The optimal transformation to align surface A to template surface B is estimated and applied on the former resulting in A'. Subsequently, the template mesh B, with uniform point distribution, is used to resample each face A'. For every point $\boldsymbol{b}$ in template mesh B the closest point $\boldsymbol{a'}$ on A' is located. The point coordinates of $\boldsymbol{a'}$ are then used as corresponding points for point $\boldsymbol{b}$ while maintaining the connectivity between the points in B. This results in A'' which, has the geometry of A' but the topology of B. Figure 5.15 shows a schematic illustration of the process.

As mentioned earlier, the template face has been manually “cleaned” in order to eliminate unwanted facial areas. As a result, all those areas in A representing hair, shoulders and other non-facial parts are automatically removed since these points in A' have no

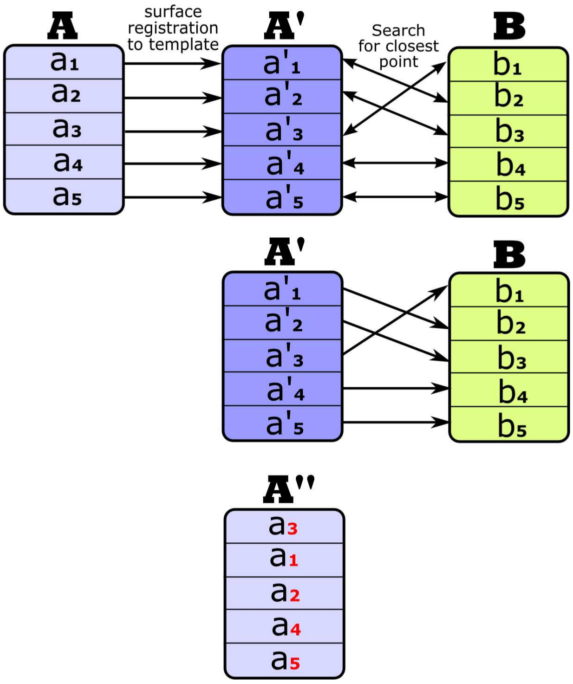


Figure 5.15: The process of establishing correspondence between faces in the ICP-based method. In the top row, a rigid surface registration of A to B results in A'. The correspondence between points in A' and B is then established and the point coordinates are copied over from A' while maintaining the connectivity of point in B. Once more, all cases of A'' will have the same number of points thus enabling the use of PCA.

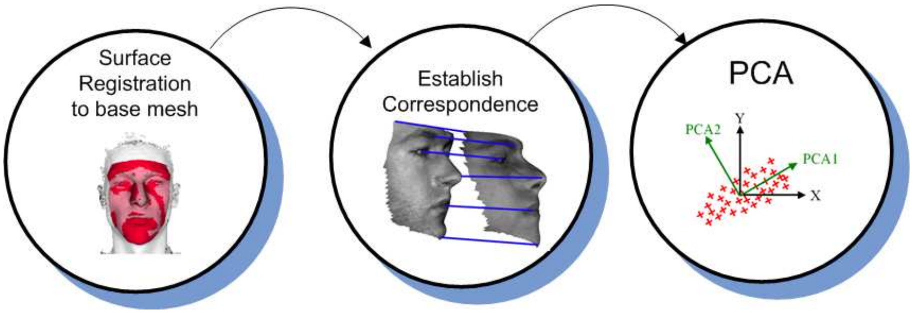


Figure 5.16: The process pipeline of the ICP-based model.

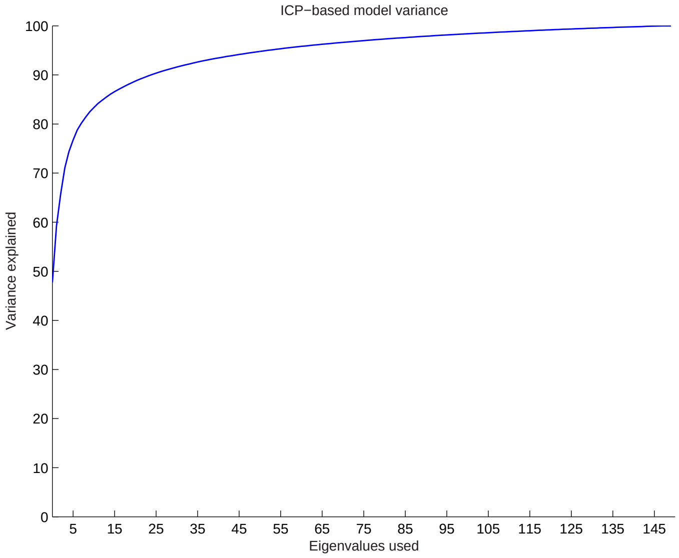


Figure 5.17: The variation described by the modes of the ICP-based statistical face model.

corresponding point in B. Figure 5.16 shows the pipeline of the model building using automatic surface registration.

Figure 5.17 shows the variance described by each of the principal components. The first mode describes 48% of the model's variance. Table 5.4 and Table 5.5 show the first four principal components of shape variation between -3, 0 and $+3$ standard deviations using the ICP-based model. Figure 5.18 shows the population distribution on the first two principal modes of variation. The arrows show the boundaries of three standard deviations. Again, the population distribution in the first two principal axes is unimodal. Furthermore, all the subjects are within three standard deviations of the mean (in the first two dimensions).

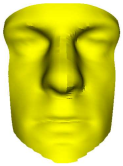
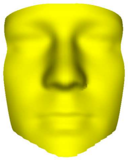
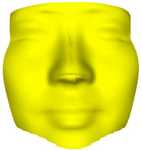
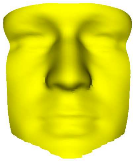
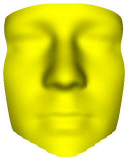
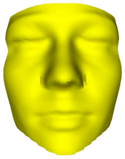
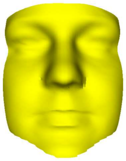
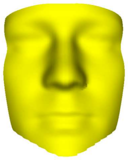
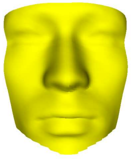
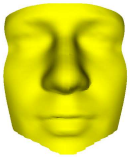
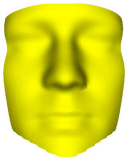
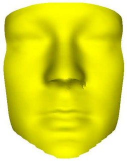
ICP-based Principal Shape Modes			
	$-3\sqrt{\lambda}$	**mean**	$+3\sqrt{\lambda}$
mode 1			
mode 2			
mode 3			
mode 4			

Table 5.4: The first four principal shape modes of the ICP-based registration model (frontal view).

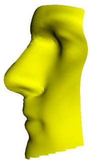
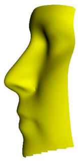
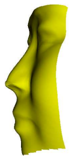
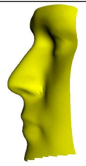
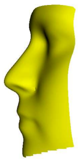
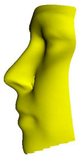
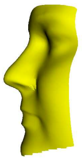
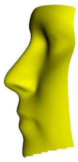
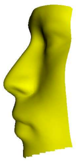
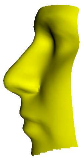
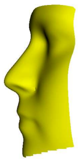
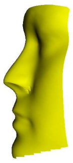
ICP-based Principal Shape Modes			
	$-3\sqrt{\lambda}$	**mean**	$+3\sqrt{\lambda}$
mode 1			
mode 2			
mode 3			
mode 4			

Table 5.5: The first four principal shape modes of the ICP-based registration model (profile view).

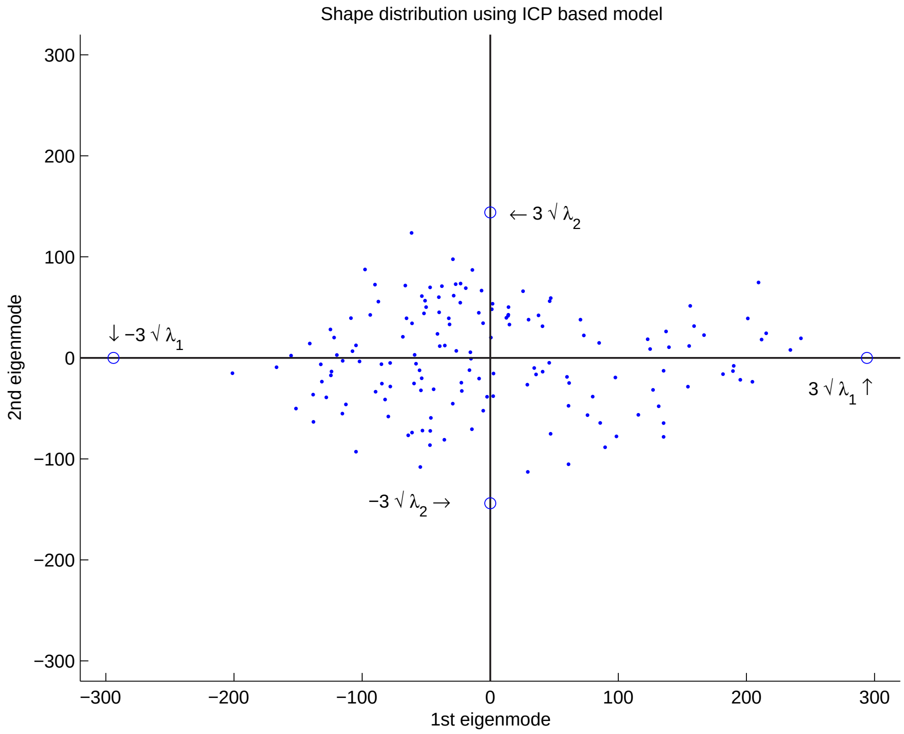


Figure 5.18: The shape distribution of the ICP-based model projected on the first two principal modes. The arrows show the boundaries of three standard deviations.

5.4 Comparing the face models

5.4.1 Qualitative comparison

A visual comparison of the models generated by the two methods already shows some differences between them. Figure 5.19 shows two views of the landmark-based mean (left) and the ICP-based mean (right). The facial features on the landmark-based model are much sharper than the features of the ICP-based one. Given that the features of the surfaces are aligned to each other using non-rigid registration, it is only natural that the resulting mean would be a surface with much more clearly defined features. For example, the lips of every face in the landmark-based model are always aligned to lips and therefore the points representing them would approximately be the same with only their location in space changing. On the other hand the lips in a ICP-based model are not always represented by the same points. The upper lip on one face might match with the lower lip on the template face, which results in an average face model with less pronounced features.

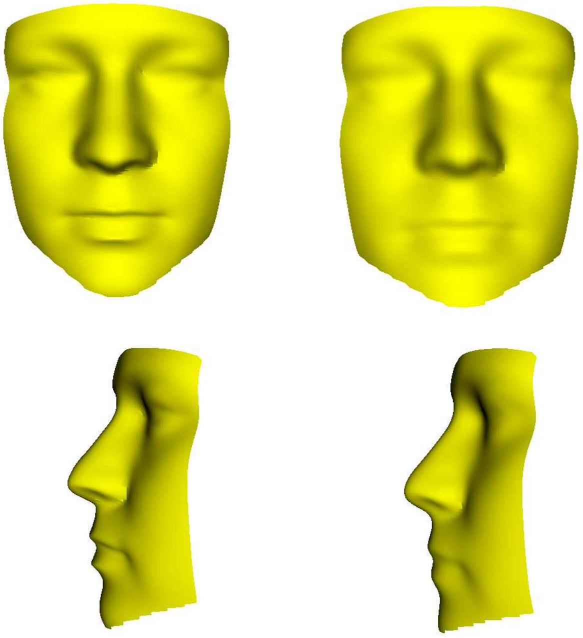
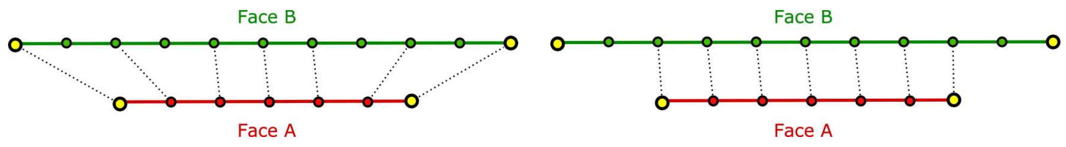


Figure 5.19: Comparison of a landmark-based model mean (left) and a ICP-based model mean (right).

This is expected, as the faces are aligned using a global transformation and there is no effort made to align individual features together. A similar phenomenon is demonstrated earlier in a 2D image in Figure 5.1.

Another visual difference between the two models is the fact that facial size is encoded more explicitly in the landmark-based model. The first mode in Table 5.2 and Table 5.3 clearly encodes face length. On the other hand the ICP-based model in Table 5.4 and Table 5.5 does not describe size. The closest-point correspondence in the ICP-based model is established after rigid surface registration while in the former the faces were actually morphed to fit the landmarks. Figure 5.20 shows schematically the type of correspondences likely to occur in the landmark- and ICP-based model. In the landmark-based model in Figure 5.20(a), face A is scaled due to the non-rigid transformation generated by registering the landmarks. This scaling does not take place in the ICP-based model of Figure 5.20(b). It is this scaling that is explicitly encoded in the former and not in the latter. The ICP-based model encodes just the surface changes over the overlapping areas of face A and face B.



(a) Landmark-based model

(b) ICP-based model

Figure 5.20: (a) shows schematically the correspondences established by the use of landmarks and (b) shows them for the ICP-based model. Notice how the correspondences established in (b) are mostly along the normal direction.

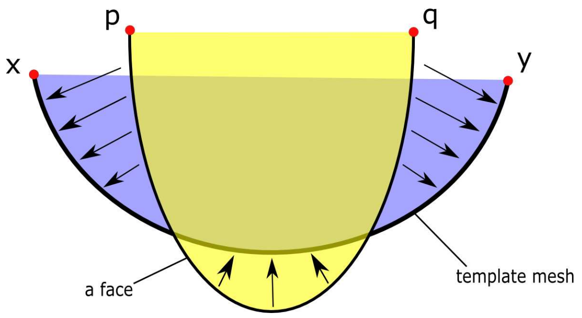


Figure 5.21: Once the faces are registered using ICP the closest points are selected. The geodesic distance between points $\boldsymbol{x}$ and $\boldsymbol{y}$ in the template mesh and $\boldsymbol{p}$ and $\boldsymbol{q}$ in the subject's face remains relatively unchanged.

The first mode of the surface-registration model in Table 5.4 might, at first sight, look like it is describing facial width but in reality the geodesic distance from one side of the face to the other (i.e. left to right) changes very little. Figure 5.21 shows a schematic representation of a template mesh and a face as seen from the top. The geodesic distance between points $\boldsymbol{x}$ and $\boldsymbol{y}$ in the template mesh is the same as the geodesic distance between points $\boldsymbol{p}$ and $\boldsymbol{q}$ in the subject's face. In other words the "height" and the "width" of the template face that is used to resample a facial surface does not change significantly. What does change and is therefore encoded in the first principal component of the ICP-based model is the "depth" (protrusion) of the template face.

The above characteristics, however, merely describe some visual aspects of the face and do not necessarily describe their ability to perform good classification.

Rank 1 rates				
Modes Used	**Euclidean**		**Mahalanobis**	
	LR	ICP	LR	ICP
10	84.67%	98%	79.33%	98.67%
20	89.33%	99.33%	88.67%	99.33%
30	90.67%	99.33%	94%	99.33%
40	92%	99.33%	97.33%	99.33%
50	92.67%	99.33%	97.33%	99.33%
60	92.67%	99.33%	98.67%	99.33%
70	92%	100%	98.67%	98%
80	92%	100%	98%	98%
90	92%	100%	98.67%	98%
100	92%	100%	98.67%	98%
110	92%	100%	98%	98%
120	92%	100%	97.33%	98%
130	92%	100%	98.67%	97.33%
140	92%	100%	98.67%	98%

Table 5.6: The rank 1 rates of the landmark-based (LR) and ICP-based (ICP) models.

5.4.2 Quantitative comparison of the models in face recognition

In order to assess the quality of the two model-building methods for face recognition, the faces of the Notre Dame database were divided in two, the gallery set $\mathcal{G}$ and the query set $\mathcal{P}$. $\mathcal{G}$ is used to build the model as described previously and $\mathcal{P}$ is used for identification. Both sets of faces are projected into the facespace and their parameters are used for similarity comparisons as described in Section 5.2. Using the Euclidean and Mahalanobis similarities between the faces, open- and closed-set identification as well as verification was used as in Chapter 4 to describe the task-specific effectiveness of the models. Table 5.6 and Figure 5.22 shows the rank 1 ($P_I(1)$) rates of the two models. The difference between them is clear as the ICP-based model performs significantly better than the landmark-based, achieving rank 1 rates of 100%. It is important to note that the ICP-based model performs significantly better with a low number of eigenmodes, managing to reach 98% and 98.67% using Euclidean and Mahalanobis distance measures respectively, when only 10 eigenmodes are used. The landmark-based model scores 84.67% and 79.33% in the same measurements. Similar trends can be observed in Table 5.7 and Figure 5.23 showing the $P_{FA} = P_{DI}$ rate and in Table 5.9 and Figure 5.25 showing the ROC area rates. The difference is less pronounced in Table 5.8 and Figure 5.24 which shows the verification rates (P_V). Figure 5.26 shows the recognition rates of the landmark-based model building technique which uses non-rigid registration against the results of a rigid-

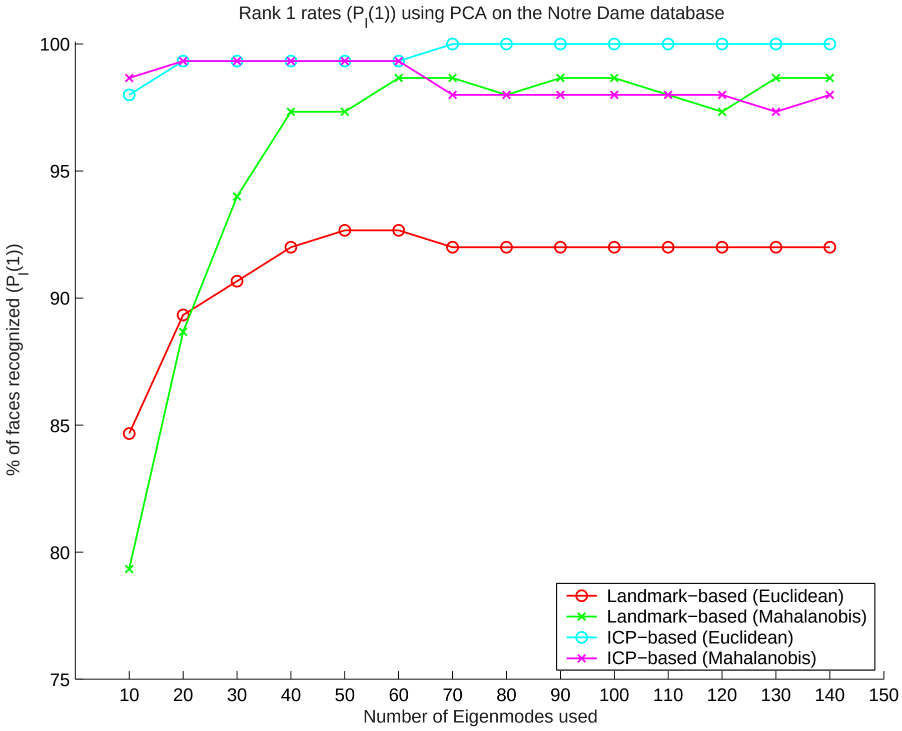


Figure 5.22: The rank 1 rates of the landmark-based and ICP-based models.

FA=DI rates				
Modes Used	**Euclidean**		**Mahalanobis**	
	LR	ICP	LR	ICP
10	79.67%	91.33%	78.33%	93.67%
20	84%	92.67%	80.33%	95.67%
30	84.67%	94%	83.33%	95.33%
40	86%	94%	86.67%	96%
50	87%	94%	89%	95.33%
60	87.67%	94.67%	91%	94.67%
70	87.67%	95%	89.33%	94.67%
80	88%	95%	92.33%	94.67%
90	88%	95%	91.33%	94.33%
100	88.33%	95%	92%	93.67%
110	88.33%	95%	92%	94.67%
120	88.33%	95%	92.67%	94.67%
130	88.67%	95%	94%	96.33%
140	88.67%	95.33%	95%	97.67%

Table 5.7: The FA=DI rates of the landmark-based (LR) and ICP-based (ICP) models.

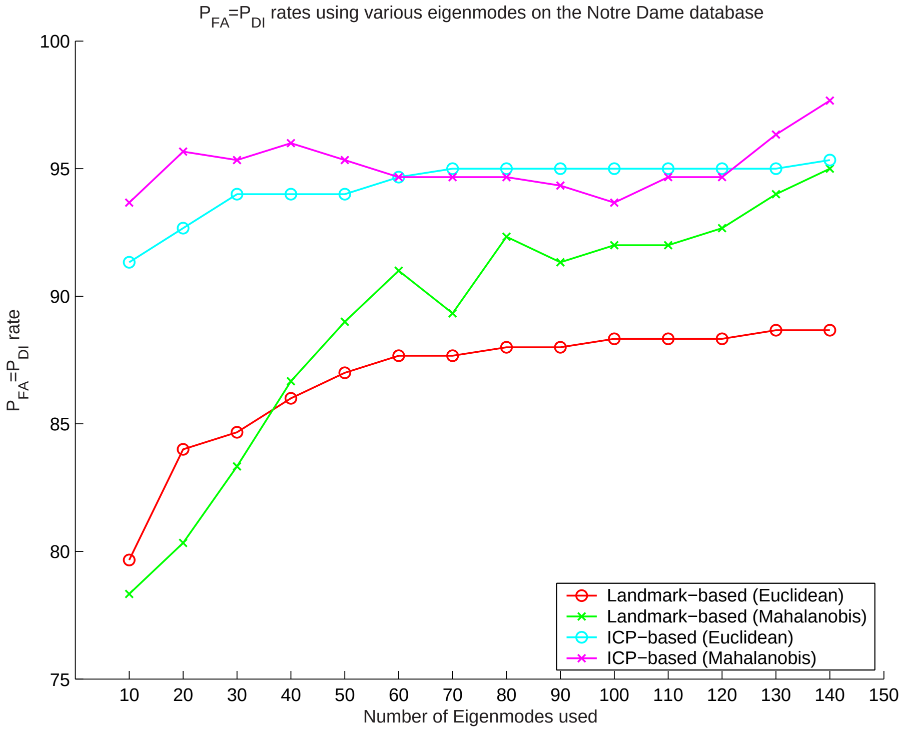


Figure 5.23: The FA=DI rates of the landmark-based and ICP-based models.

	Verification rates			
Modes Used	**Euclidean**		**Mahalanobis**	
	LR	ICP	LR	ICP
10	88.6%	97.3%	82%	99.3%
20	93.3%	98.6%	92%	98%
30	93.3%	99.3%	94.6%	99.3%
40	94%	98.6%	96%	98%
50	94%	98.6%	96.6%	98.6%
60	94%	98.6%	96.6%	98.6%
70	94%	98.6%	97.3%	98.6%
80	94.6%	98.6%	97.3%	97.3%
90	94.6%	98.6%	97.3%	97.3%
100	94.6%	98.6%	96.6%	97.3%
110	94.6%	98.6%	96.6%	98%
120	94.6%	98.6%	97.3%	98%
130	94.6%	98.6%	98%	98.6%
140	94.6%	98.6%	98%	98.6%

Table 5.8: The verification rates (P_V at $FA = 1\%$)

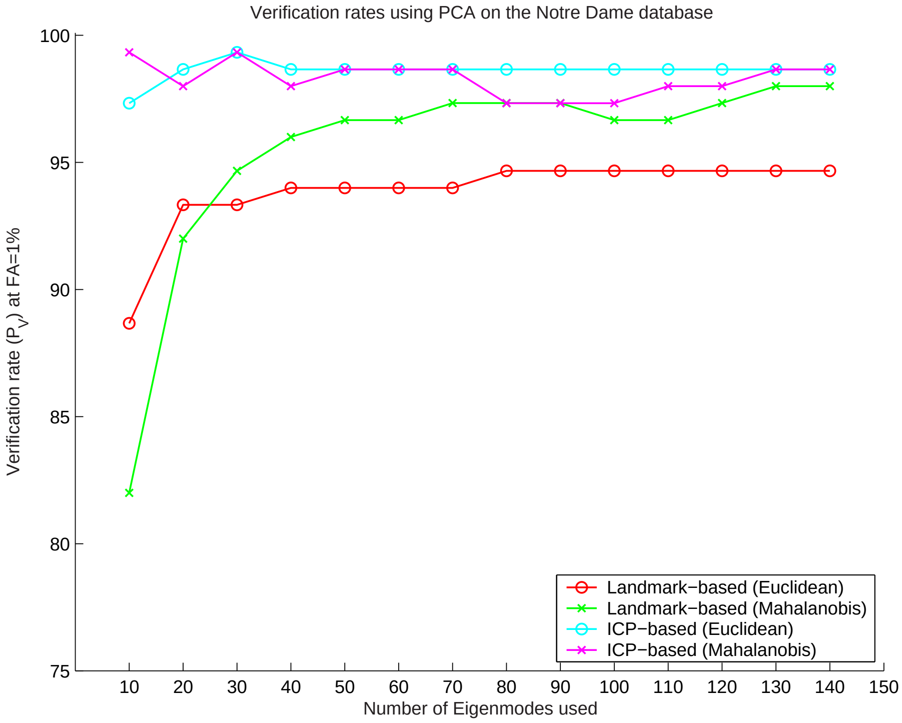


Figure 5.24: The verification rates (P_V at $FA = 1\%$)

ROC curve rates				
Modes Used	**Euclidean**		**Mahalanobis**	
	LR	ICP	LR	ICP
10	88.29%	96.54%	85.73%	98.48%
20	91.98%	97.73%	89.42%	98.41%
30	92.89%	98.08%	92.17%	98.95%
40	93.33%	98.17%	94.32%	98.79%
50	93.44%	98.15%	95.53%	98.58%
60	93.74%	98.21%	96.55%	98.45%
70	93.80%	98.25%	96.61%	98.60%
80	93.85%	98.26%	97.33%	98.23%
90	93.93%	98.32%	97.28%	98.15%
100	94%	98.32%	97.14%	98.28%
110	94.03%	98.39%	97.16%	98.38%
120	94.06%	98.42%	97.75%	98.46%
130	94.09%	98.46%	98.31%	98.78%
140	94.10%	98.59%	98.97%	99.22%

Table 5.9: The ROC curve rates of the landmark-based (LR) and ICP-based (ICP) models.

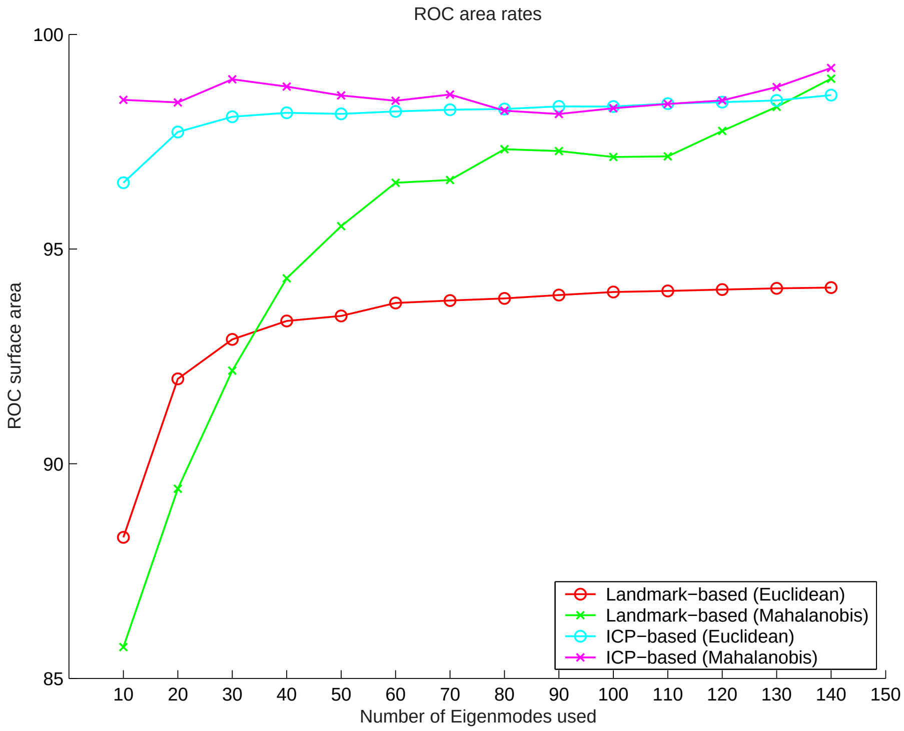


Figure 5.25: The ROC curve rates of the landmark-based and ICP-based models.

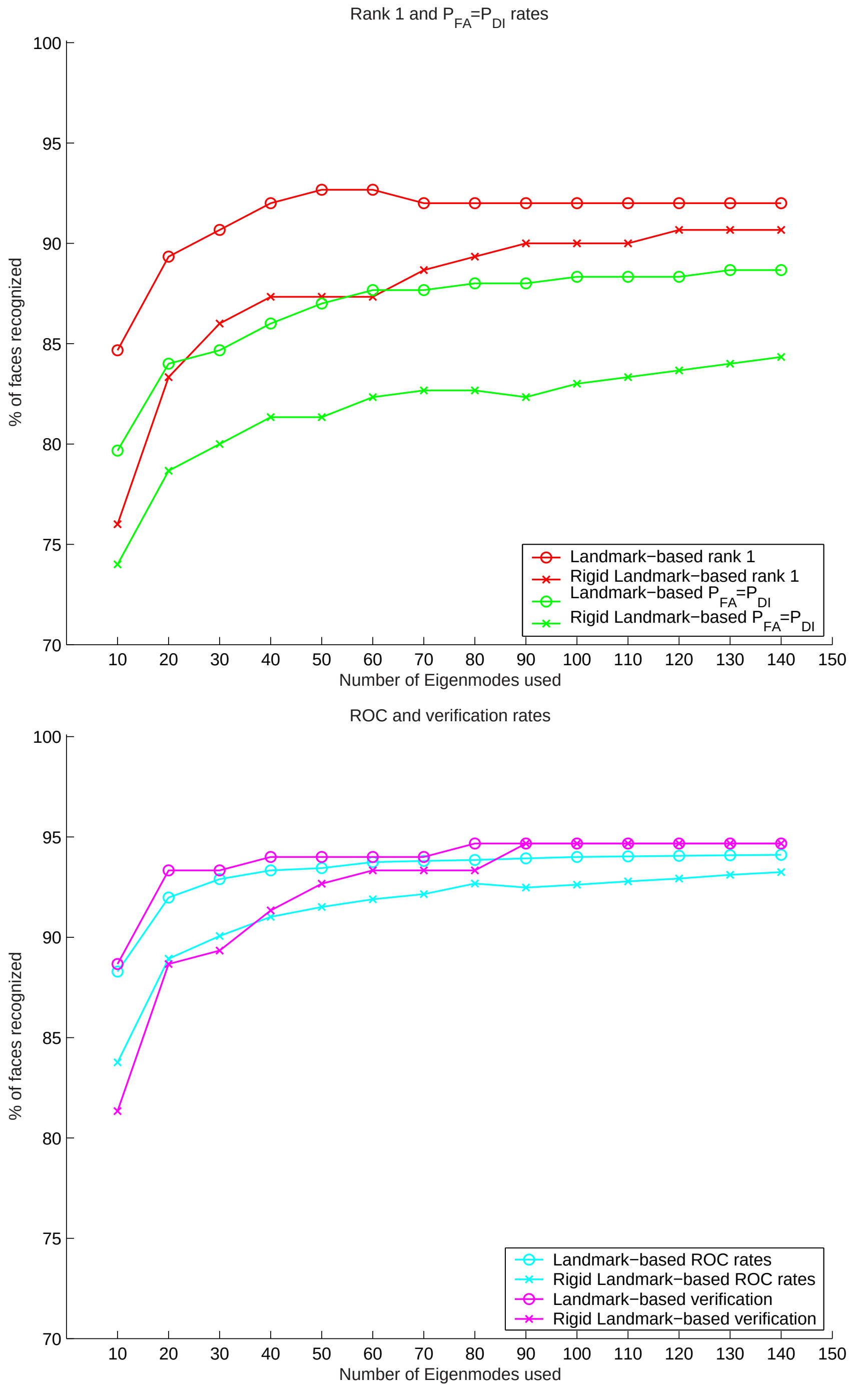
only landmark registration model-building method. Conclusions regarding all of these results are drawn in the concluding section of this chapter.

5.4.3 Comparing the properties of the models

In addition to using task-specific assessments of 3D face models other more generic objective measures can also be used to assess the quality of 3D statistical models.

5.4.3.1 Generalization ability

The generalization ability of a face model is its ability to represent a face that is not part of the training set. This is of importance, as the model needs to be able to generalize to unseen examples and not be overfitting to the training set. Generalization ability can be measured using leave-one-out reconstruction [51, 96]. The face model $\boldsymbol{u}$ is built using datasets $\{\Gamma\}$ and leaving one face Γ_i out. The left-out face is projected into the facespace



Rank 1 and $P_{FA}=P_{DI}$ rates

ROC and verification rates

Figure 5.26: Rigid vs. non-rigid landmark registration.

created using the remaining 149 faces:

$$\boldsymbol{\beta} = \boldsymbol{u}^T(\Gamma_i - \overline{\Gamma}) \tag{5.11}$$

The face Γ_i is then reconstructed using its face parameters $\boldsymbol{\beta}_s$ generating a surface $\Gamma'_i(s)$:

$$\Gamma'_i(s) \approx \overline{\Gamma} + U\boldsymbol{\beta}_s \tag{5.12}$$

where s is the number of shape parameters $\boldsymbol{\beta}$. Then the average square approximation error between the original face Γ_i and the reconstructed Γ'_i is measured:

$$\delta_i(s) = |\Gamma_i - \Gamma'_i(s)|^2 \tag{5.13}$$

This process is repeated for all faces. For a more robust assessment of the model the generalization ability was measured as a function of the number s of shape parameters $\boldsymbol{\beta}$. The mean square approximation error is the generalization ability score and is given by:

$$G(s) = \frac{1}{M}\sum_{i=1}^{M}\delta_i(s) \tag{5.14}$$

Where M is the total number of faces used. For two model building methods X and Y, if $G_X(s) \leq G_Y(s)$ for all s and $G_X(s) < G_Y(s)$ for some s then the generalization ability of method X is better than that of method Y. In this case s is the number of shape parameters $\boldsymbol{\beta}$ that are used to build the left-out face. In order to assess the differences between the models' generalization scores the standard error of each model has to be calculated [191]:

$$\sigma_{G(s)} = \frac{\sigma}{\sqrt{M-1}} \tag{5.15}$$

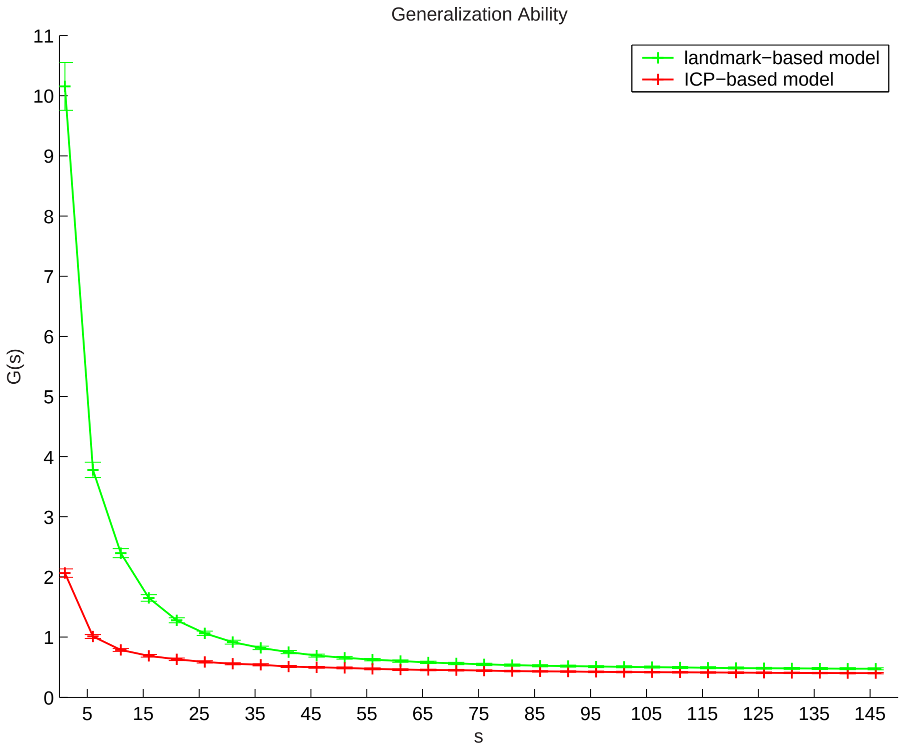


Figure 5.27: The Generalization ability of the landmark-based and ICP-based models. Note that in the graph the better a model generalizes to unseen examples the lower its generalization scores are. The error bars are computed as shown in eq. 5.15 and they show a relatively small standard error in $G(s)$ which allows us to safely conclude that the differences in the generalization scores of the two models are significant.

where M is the total number of faces used to build the model and σ is the sample standard deviation of $G(s)$ defined as:

$$\sigma = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M-1}(x_i - \overline{x})^2} \tag{5.16}$$

As can be seen in Figure 5.27 the ICP-based model has greater capacity to model unseen examples. For all number of parameters s used the ICP-based model performs significantly better. These differences are most obvious when 1 to 30 parameters are used. In literature [51, 96] the generalization ability of the model is plotted with 0 being the most general model and anything greater than that being less general.

5.4.3.2 Specificity

Specificity measures the ability of the face model to generate face instances that are similar to those in the training set. To test the specificity N random faces Γ' were generated as a function of s, the number of face parameters $\boldsymbol{\lambda}$. The generated faces are then compared to the closest faces Γ in the training set:

$$S(s) = \frac{1}{N} \sum_{i=1}^{N} |\Gamma_i - \Gamma'_i(s)|^2 \tag{5.17}$$

For two model-building methods X and Y, if $S_X(s) \leq S_Y(s)$ for all s and $S_X(s) < S_Y(s)$ for some s then method X builds a more specific model than method Y. Once again the standard error of each model has to be calculated in order to be able to assess whether the differences between the two models are significant:

$$\sigma_{S(s)} = \frac{\sigma}{\sqrt{N-1}} \tag{5.18}$$

To calculate the specificity 500 random faces were generated. Figure 5.28 shows that the ICP-based model is also significantly more specific than the landmark-based model.

5.4.3.3 Compactness

Compactness measures the ability of the model to reconstruct an instance with as few parameters as possible. A compact model is also one that has as little variance as possible and it is described as a plot of the cumulative covariance matrix:

$$C(s) = \sum_{i=1}^{s} \boldsymbol{\lambda}_i \tag{5.19}$$

To assess the significance of the differences the standard error in C(s) is calculated once again. The standard deviation in the i^{th} mode is given by [191]:

$$\sigma_{\boldsymbol{\lambda}_i} = \sqrt{\frac{2}{M}} \boldsymbol{\lambda}_i \tag{5.20}$$

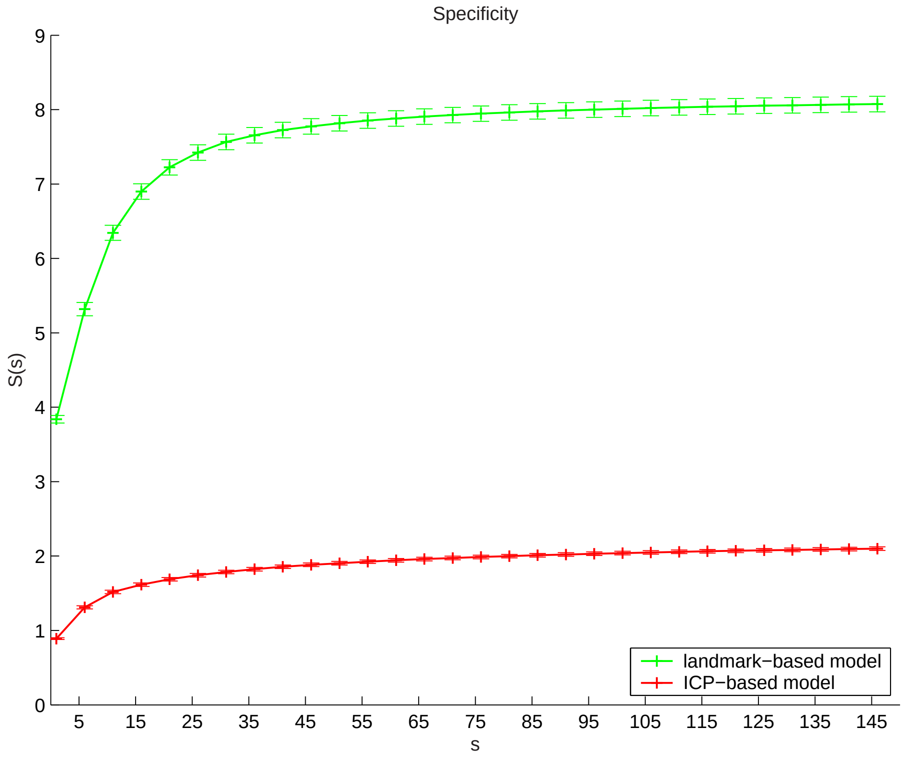


Figure 5.28: The specificity scores of the landmark-based and ICP-based models. Small standard error in $S(s)$ (as shown from the error bars) also allows for us to conclude safely that the difference in specificity scores is indeed significant.

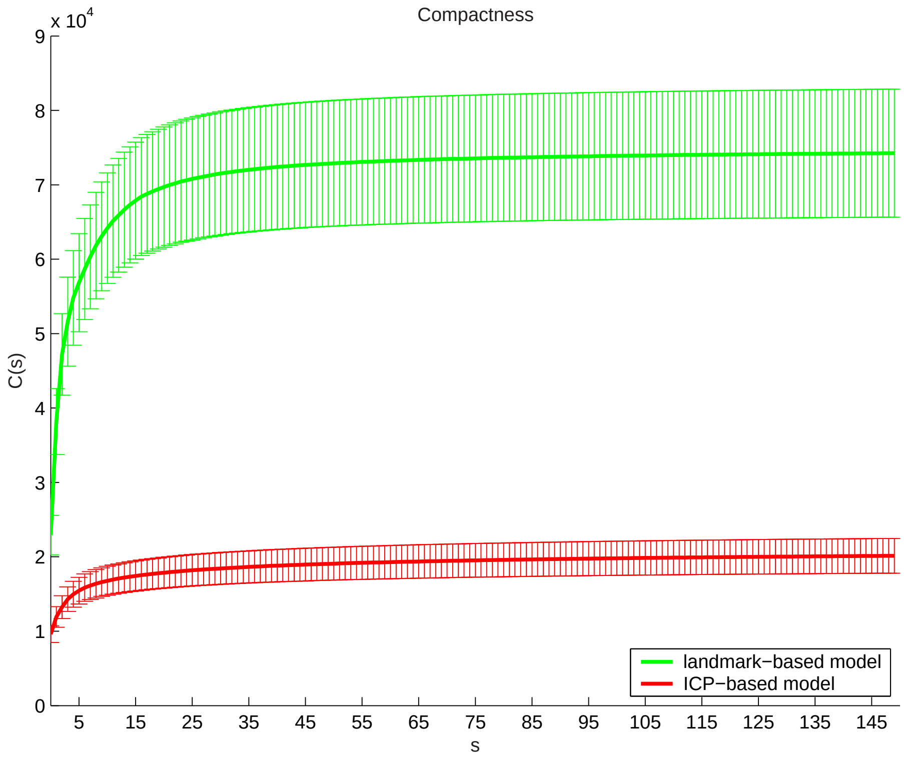


Figure 5.29: The compactness scores of the landmark-based and ICP-based models. The standard error bars indicate the likely error of $C(s)$ allowing one once more to conclude that there is significant difference between the compatness scores of the two models.

where $\boldsymbol{\lambda}_i$ is the i^{th} eigenvalue of the covariance matrix. The standard error is then given by:

$$\sigma_{C(s)} = \sum_{i=1}^{s} \sqrt{\frac{2}{M}}\boldsymbol{\lambda}_i \tag{5.21}$$

Figure 5.29 shows that the ICP-based model is significantly more compact than the landmark-based model.

5.5 Conclusions

In this chapter it was shown that the ICP-based model is superior to the landmark-based one because it allows for the faces to be better aligned to each other. The landmark-based model might produce visually better defined features, but it performs significantly worse in face recognition experiments. It was also shown that the ICP-based model is more specific, generalizes better to unseen examples and is more compact. In the results of this chapter, it was demonstrated that the Mahalanobis distance is a better distance metric to use for the landmark-based model. This is because the landmark-based model contains noise in the form of surface misregistrations since it only uses landmarks to align the facial surfaces to each other. The effect of this within-class variation is minimized when one uses the Mahalanobis distance. Eq. 5.7 describing the latter shows the difference between two shape parameters being divided by the corresponsing standard deviation. This effectively removes within-subject variation. The Euclidean distance on the other hand does not take the standard deviation into account and thus is more sensitive to within-class variation. Another

A landmark-based model, whether it models faces or other objects, might work better on datasets of more varying size. Using an ICP-based model which mainly encodes surface variation of overlapping areas without encoding size would probably not be very appropriate. In the case of the human face, the cranium is covered by a layer of muscle, fat, skin and in certain areas cartilage, all of which introduce a lot of local surface-based variability from person to person. Furthermore, the population of faces used in this study for building the statistical model consisted of adults. If children had been included in the facespace, there would surely have been more head and facial feature size variation. It is not clear how well smaller faces would be reconstructed using the ICP-based technique for such a case. A landmark-based technique might, for example, capture and model the aging trajectory of the population more accurately as it allows for more radical head size differences to be correctly encoded [98].

The importance of establishing good correspondence between surfaces is further demonstrated by another finding: If the landmark-based model-building technique is repeated but instead of performing a non-rigid landmark registration, faces are only rigidly registered using the landmarks, then the recognition rates fall significantly (Figure 5.26). Including a non-rigid registration of landmarks and thus establishing a better correspondence between the surfaces and their anatomical features, leads to a better model.

In principle, texture could have been incorporated into the face models producing two spaces: One for shape and one for appearance. Early experimental results using the textures of the Notre Dame datasets were relatively poor due to the low quality of the texture. We, therefore, decided to focus on shape only and how different registration techniques affect the statistical face model.

Hutton [96] used a similar approach for creating face models. Instead of using multilevel B-splines, Hutton employed thin plate splines to achieve similar goals. The *dense surface model* (DSM) Hutton built was used for a two-class classification problem where subjects were diagnosed as having Noonan syndrome (a genetic disorder subtly affecting the facial structure) or not, but the model was never used for a multi-class classification problem. In two-class classification tests of 674 subjects it managed to correctly detect the ones with Noonan syndrome with 90% accuracy. He later reports similar rates for Velo-cardiofacial syndrome (88%), Smith-Magenis syndrome (93%) and Williams syndrome (94%). He also uses the DSM to distinguish males from females with an overall accuracy of 76.8%. Since the population in [96] contained children and adults the use of landmarks was perhaps more appropriate in order to detect developmental problems. Therefore, the choice of model-building technique might be application-specific. Furthermore, apart from discarding pairs of closest points which are far from each other, there is no other effort made in Hutton to improve the correspondence, such as filling holes or regularizing the surface mesh. In Chapter 6 the classification problem is used as a testbed to explore some techniques that improve surface-to-surface correspondence as well as surface regularization.

Chapter 6

Improving the correspondence using non-rigid surface registration

6.1 Introduction

In the previous chapter we have proposed two registration algorithms for estimating correspondences for the automatic construction of statistical shape models. Both methods suffer from a similar problem: Particularly in areas of high curvature where the faces might differ significantly, such as around the lips or nose, the correspondence established between surface points tends to be incorrect. The top image in Figure 6.1 demonstrates some closest-point correspondences between two surfaces after imperfect registration. The circles show areas where due to differences in curvature the correspondence established is not the ideal one. When using the template face to resample the training set samples the aforementioned errors in establishing correspondence result in an irregular mesh, which is a form of noise in the training set. An example of such mesh irregularity on a training set sample is shown in Figure 6.2. Notice how the nose and eyebrows are particularly prone to these kind of artifacts.

In this chapter two techniques are proposed based on non-rigid surface registration which aim to improve correspondences between surfaces. The bottom image in Figure 6.1 shows schematically how the correspondence between a facial surface and a base mesh

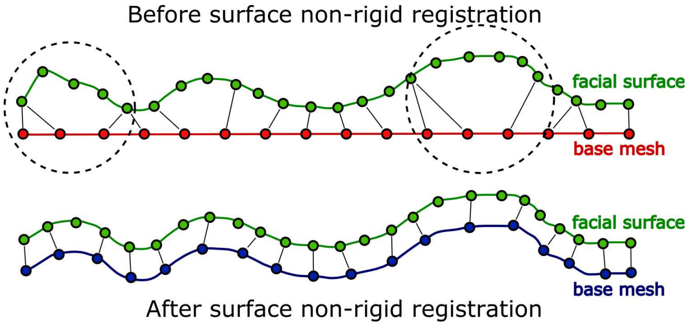

Before surface non-rigid registration
facial surface
base mesh
facial surface
base mesh
After surface non-rigid registration

Figure 6.1: Schematic representation of errors in establishing correspondence. Due to differences in spatial morphology the correspondence between the surfaces is not ideal resulting in non uniform data used for PCA.

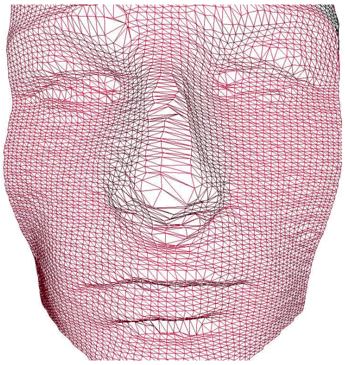


Figure 6.2: In areas of high curvature errors in establishing correspondence occur. This results in an irregular mesh where some areas have higher point density than others. This image shows such artifacts on a facial surface used for training the model. Examples like this introduce noise into the model and our assumption is that reducing such occurences will result in a better model.

can be improved if the base mesh is allowed to deform towards the other mesh via non-rigid registration. Ultimately, however, the point distribution on the original facial surface determines the distribution of the points on the resulting surface. It is for this reason that a second technique is also introduced which uses a synthetically generated uniform surface, such as a sphere, which deforms to approximate the original dataset while maintaining uniform spacing between the resulting surface points. The assumption throughout this chapter is that by improving the correspondence, higher face recognition rates should be expected.

6.2 Non-rigid surface registration using a template face

The method presented in this section is a way of calculating a non-rigid transformation to align two sets of points. Given surfaces A and B, made up of two point sets $\boldsymbol{a}$ and $\boldsymbol{b}$, the similarity function that we want to minimize is:

$$f(\boldsymbol{T}_{nonrigid}) = \frac{1}{|A|}\sum_{i=1}^{|A|} ||\boldsymbol{b}_i - \boldsymbol{T}_{nonrigid}(\boldsymbol{a}_i)||^2. \tag{6.1}$$

where $\boldsymbol{T}_{nonrigid}$ is a non-rigid transformation. In Chapter 5 we proposed a non-rigid registration algorithm for sets of landmarks with known correspondence. In this chapter we assume that the correspondence between surface points is unknown. In order to pair points on two surfaces to each other, just as with ICP, we assume that corresponding points will be closer to each other than non-corresponding ones. A distance metric d is defined between an individual source point $\boldsymbol{a}$ and a target (model) shape B:

$$d(\boldsymbol{a}, B) = \min_{\boldsymbol{b}\in B} ||\boldsymbol{b} - \boldsymbol{a}|| \tag{6.2}$$

Using this distance metric the closest point in B from all points in A is located. Even though the non-rigid registration uses closest point correspondences it is different from simply establishing the pairs of points as it is done in Figure 6.1 (top). The reason for that is that the non-rigid transformation is computed by a type of “voting”. All points on the surface are assigned their corresponding points and a transformation is computed that would minimize the error between these correspondences. Given the number of points that exist on a surface, the surface deformation resulting from the displacement of the control point grid tends to be smooth. This is particularly the case when the initial control point spacing is large, which enables better pairings to be established (see 6.1 (bottom)). Let Y denote the resulting set of closest points and $\mathcal{C}$ the closest point operator:

$$Y = \mathcal{C}(A, B) \tag{6.3}$$

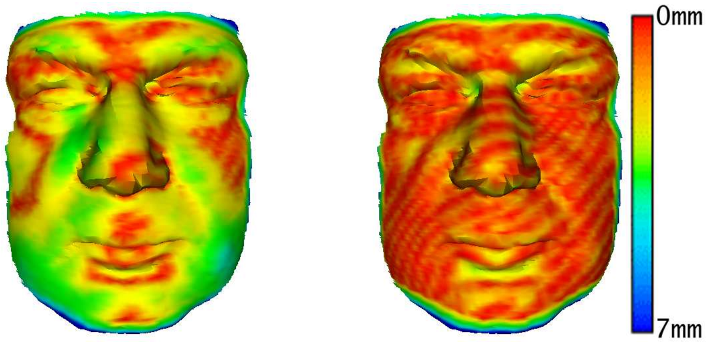
After closest-point correspondence is established, the point-based non-rigid registration algorithm developed in Chapter 5 can be used to calculate the optimal non-rigid transformation $\boldsymbol{T}_{nonrigid}$. This is represented here by the operator $\mathcal{M}$. In order for the deformation of the surfaces to be smooth, a multiresolution approach similar to the one in Chapter 5 was adopted, where the control point grid of the transformation is subdivided iteratively to provide increasing levels of accuracy. The non-rigid surface registration algorithm is displayed in Listing 4. Figure 6.3 shows two color maps of the distances between a face and a base mesh of the landmark-based model. Image (a) shows the distance between a face and a base mesh after non-rigid landmark registration. Notice that the areas near landmarks are the closest to the base mesh (eyes, mouth, nose, chin). Image (b) shows the color map after non-rigid surface registration has been performed and as a result the distance between the face and the base mesh has been reduced globally.

Listing 4 The non-rigid surface registration

```
Start with surfaces $\boldsymbol{A}$ and a target point set $\boldsymbol{B}$.
Set subdivision counter $k = 0$, $A^{(0)} = A$ and reset $\boldsymbol{T}_{nonrigid}$.
repeat
    Find the closest points between $A$ and $B$ by: $Y^{(k)} = \mathcal{C}(A^{(k)}, B)$
    Compute the ideal non rigid transformation to align $Y^{(k)}$ and $A^{(0)}$ by:
    $\boldsymbol{T}^{(k)}_{nonrigid} = \mathcal{M}(A^{(0)}, Y^{(k)})$.
    Apply the transformation: $A^{(k+1)} = \boldsymbol{T}^{(k)}_{nonrigid}(A^{(0)})$
until $k$ equals user-defined maximum subdivisions limit
```

Similarly for the ICP-based model, Figure 6.4(a) shows a color map of the distance between a face and a base mesh after rigid surface registration. Notice that some areas of the face are close to the base mesh while others are further away. Notice also how the color map pattern of the ICP-based model differs from the landmark-based one of Figure 6.3(a). Figure 6.4(b) shows the color map after non-rigid surface registration has been performed.

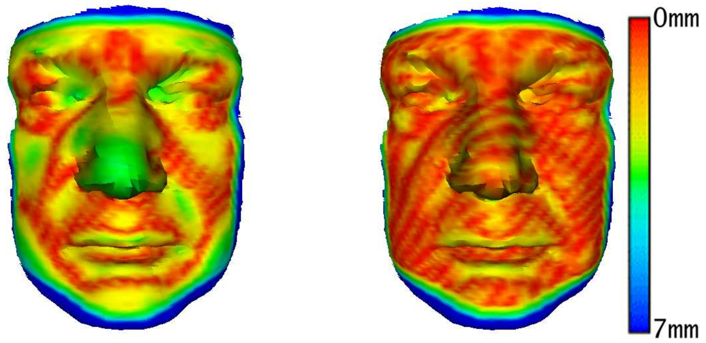
This processing step is added to the model construction pipelines of the statistical models discussed in Chapter 5. The landmark-based statistical model involves the rigid registration of the landmarks followed by a non-rigid registration of the latter. Correspondence is then established between the landmark-registered surfaces. It is right before the



(a) After point-based non-rigid registration.

(b) After non-rigid surface registration.

Figure 6.3: The effects of non-rigid surface registration on non-rigidly registered data using landmarks. The closest points were originally the ones surrounding the areas that contain landmarks (a). After the non-rigid surface registration the two surfaces have been brought into much closer global alignment (b). The stripes on the facial surface on image (b) are the result of the non-rigid surface registration. Under certain control point grid resolutions a faint array of stripes can be seen on the surface which is caused by a slightly greater deformation to surface points closer to the grid's control points. This has no effect on the actual correspondence established after the non-rigid registration.



(a) Before non-rigid surface registration.

(b) After rigid surface registration.

Figure 6.4: The effects of non-rigid surface registration on surfaces that have been rigidly registered using ICP. The surfaces in (a) were in closer alignment compared to the landmark-based registration but a non-rigid surface registration (NRSR) brought them into better alignment (b).

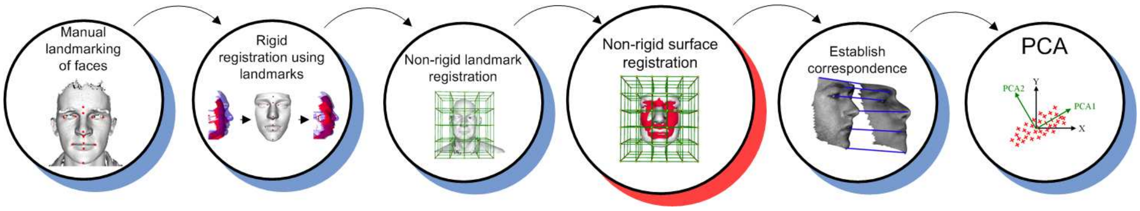

Manual landmarking of faces
Rigid registration using landmarks
Non-rigid landmark registration
Non-rigid surface registration
Establish correspondence
PCA
PCA2
Y
PCA1
X

Figure 6.5: The process pipeline of the landmark-based statistical model which contains a non-rigid surface registration (NRSR).

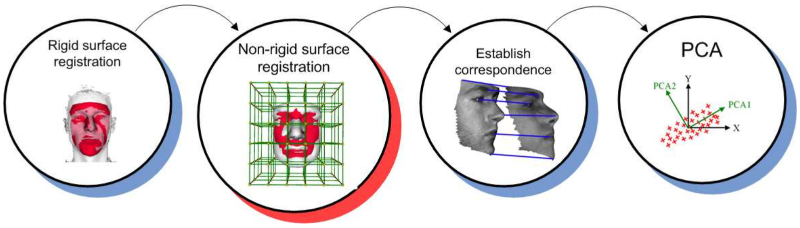
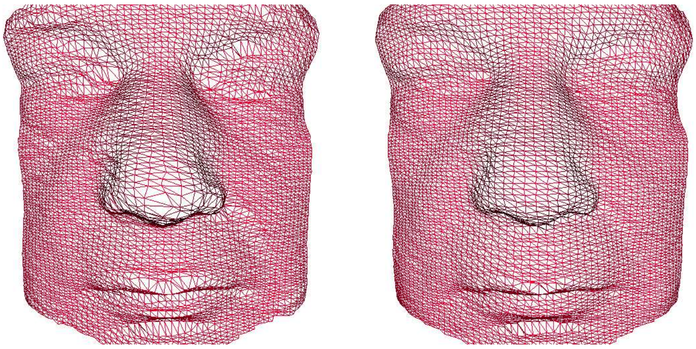


Figure 6.6: The process pipeline of the ICP-based statistical model which contains a non-rigid surface registration (NRSR).

establishment of correspondence that the non-rigid surface registration is added in order to make the resulting correspondence smoother, which in turn results in smoother meshes. Just as with landmark-based non-rigid registration (Figure 5.10) the non-rigid surface registration (NRSR) is used in order to align the facial points with their corresponding ones on the template mesh. Once the pairings have been established the point coordinates of the surface *before* the non-rigid registration are copied over to the template mesh while maintaining the point connectivity of the latter (see Section 5.3.1.3). Figure 6.5 shows the new processing pipeline for generating a non-rigid surface registration-based statistical model. The step in red is the newly added step that performs the non-rigid surface registration before the correspondence is established. Similarly for the ICP-based statistical model in Chapter 5, after the faces have been rigidly registered a non-rigid surface registration is performed to bring the facial points into closer alignment before the dense surface correspondence is established. Figure 6.6 shows the new processing pipeline for generating an ICP-based statistical model. Once more, the step in red performs the non-rigid surface registration before the correspondence is established. The non-rigid surface registration can register two surfaces in under $4secs$ on a $2Ghz$ machine.



(a) Mesh generated before non-rigid surface registration.

(b) Mesh generated after non-rigid surface registration.

Figure 6.7: Differences in surface uniformity before and after non-rigid surface registration.

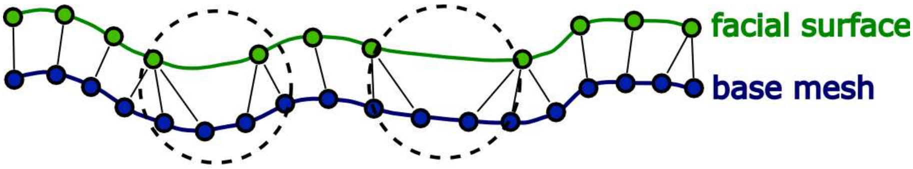


Figure 6.8: Further problems with non-rigid surface registration.

Figure 6.7(a) shows a facial surface that has been created without non-rigid surface registration while Figure 6.7(b) shows the same surface when non-rigid surface registration has been used. Notice how the surface mesh in image (b) is significantly smoother and more uniform than the one in image (a).

6.3 Non-rigid surface registration with a uniform surface

As mentioned before, ultimately the uniformity of the point distribution used for building the statistical face models depends on the uniform spacing of the sensed data which the base mesh is sampling. If there are holes on the face or even very large cells then the resulting base mesh will not be uniform. Figure 6.8 shows how such artifacts in the original surface cause irregularities in the resulting dataset even though non-rigid surface registration is used and the surfaces are closely aligned.

One way to solve this problem is by searching for a closest point anywhere on the surface.

In this section an alternative technique is presented for generating surfaces with uniformly spaced points that can be used for building the statistical model. To achieve this we propose to use a synthetically generated uniform grid of points on a sphere. The deformation process is, furthermore, restricted along the normal direction of the sphere, ensuring that point uniformity is maintained throughout the morphing process.

6.3.1 Preparing for the non-rigid registration

Initially all faces are rigidly registered to a template face using ICP. The choice of template face is discussed in Section 5.3.1.2. As mentioned in Chapter 4 we treat the alignment problem as an optimization problem whose goal is to minimize the Euclidean distance between the points $\boldsymbol{b}$ on template face B and the points $\boldsymbol{a}$ on a dataset A in order to find the optimal rotation $\boldsymbol{R}$ and translation $\boldsymbol{t}$:

$$f(\boldsymbol{T}_{rigid}) = \frac{1}{|A|}\sum_{i=1}^{|A|} ||\boldsymbol{b}_i - \boldsymbol{R}\boldsymbol{a}_i - \boldsymbol{t}||^2. \tag{6.4}$$

After all faces have been registered to the template they are translated by a transformation $\boldsymbol{T}_{init}$ to an optimal position near the surface of the sphere, as shown in Figure 6.9(b) and 6.9(c). All faces are moved to the same location on the sphere by applying the same $\boldsymbol{T}_{init}$ to all of them:

$$A'' = \boldsymbol{T}_{init}(\boldsymbol{T}_{rigid}(A)) \tag{6.5}$$

6.3.2 Spherical non-rigid registration

In order to deform the sphere to match a face, a new type of registration is developed which we term *spherical B-spline registration* (SBR). For this, a sphere of radius r is generated. Any point on the surface can be defined by two angles ϕ and θ (see Figure 6.9(a)). We fit a *spherical non-rigid transformation* (SNRT) (Figure 6.10) to approx-

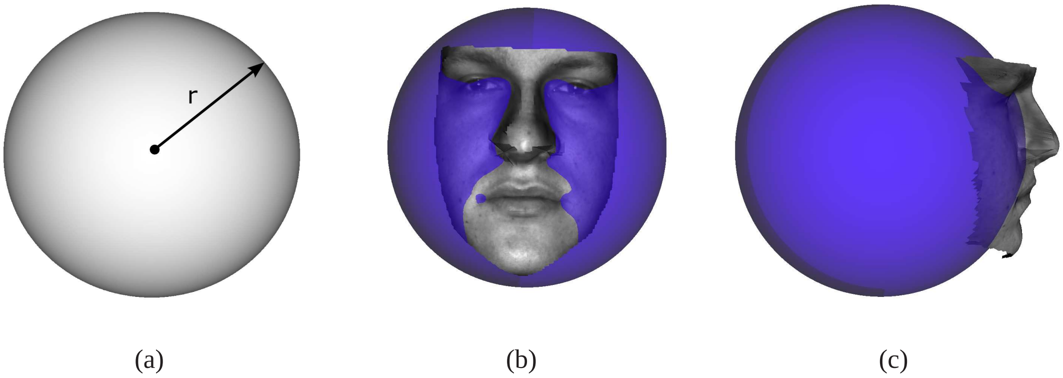


Figure 6.9: (a) The initial sphere and (b) and (c) with a face placed on the optimal position on it.

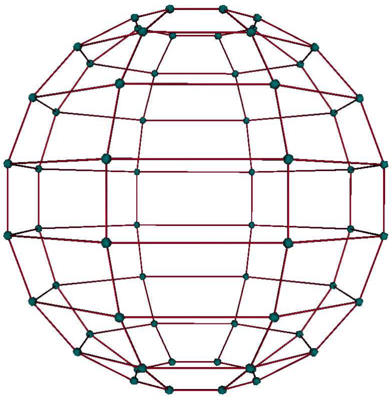


Figure 6.10: The control point grid of the spherical non-rigid transformation.

imate the facial surface. The spherical B-spline registration is similar to the B-spline registration presented in Section 6.2. The difference is that the displacements of the transformation control points are calculated in spherical rather than Cartesian coordinates. In Listing 2 of Chapter 5 the 3D B-spline registration algorithm takes in a source landmark set $\boldsymbol{p} = \{(p_{e_x}, p_{e_y}, p_{e_z})\}$ and the displacement vectors $\boldsymbol{d} = \{(d_{e_x}, d_{e_y}, d_{e_z})\}$ associated with these landmarks. We start off by calculating the distance Δ of each facial point $\boldsymbol{a} = (a_x, a_y, a_z)$ from the center of the sphere (the origin of the coordinate system) as:

$$\Delta = \sqrt{a_x^2 + a_y^2 + a_z^2} \tag{6.6}$$

If every point on the face is a projection of a ray starting from the center of the sphere, the angle θ and ϕ of each ray is calculated. θ is calculated as:

$$\theta = arcos(\frac{a_z}{\Delta}) \times \frac{180}{\pi} \tag{6.7}$$

ϕ is calculated as:

$$\phi = arctan(\frac{a_y}{a_x}) \times \frac{180}{\pi} + 180 \times n \tag{6.8}$$

where n specifies the sphere quadrant given by a_x and a_y. Given a spherical grid with uniform angular spacing δ, the SNRT can then be formulated as the 2D tensor product of B-splines:

$$\mathbf{T}_{SNRT}(\phi, \theta) = \sum_{l=0}^{3} \sum_{m=0}^{3} B_l(u) B_m(v) \phi_{i+l,j+m} \tag{6.9}$$

where $i = \lfloor \frac{\theta}{\delta} \rfloor - 1, j = \lfloor \frac{\phi}{\delta} \rfloor - 1, u = \frac{\theta}{\delta} - \lfloor \frac{\theta}{\delta} \rfloor, v = \frac{\phi}{\delta} - \lfloor \frac{\phi}{\delta} \rfloor$ and where B_l and B_m represent B-spline basis functions [63]. By using a sphere to reconstruct the surface we succeed in keeping the grid of the surface very uniform. The approximating of the original data points takes place only along the normal direction on the sphere's surface and therefore the spacing between the sphere points is uniform. Listing 5 provides the pseudo-code of the SBR algorithm. As with previous cases of non-rigid registration, in order to fit our model to the data we use a multi-resolution approach. After approximating the data using an initial coarse spherical control point grid, we subdivide it and we approximate again achieving greater accuracy (Figure 6.11). This process is repeated until the face is approximated accurately. Figure 6.12 shows the sphere being iteratively transformed to approximate a face places near its surface. Figure 6.13 shows an original dataset (left), a

Listing 5 The sphere B-spline registration algorithm

```
Start with sphere $\boldsymbol{A}$ and a target point set $\boldsymbol{B}$.
Set subdivision counter $k = 0$, $A^{(0)} = A$ and reset $\boldsymbol{T_{SNRT}}$.
repeat
    Represent every point in $B$ by two sphere angles $\phi$ and $\theta$ and calculate distance $\Delta$
    from the sphere's surface.
    Compute the ideal spherical non-rigid transformation using arrays $\boldsymbol{\phi, \theta, \Delta}$
    Apply the transformation: $A^{(k+1)} = \boldsymbol{T_{SNRT}}^{(k)}(A^{(0)})$
until $k$ equals user-defined maximum subdivisions limit
```

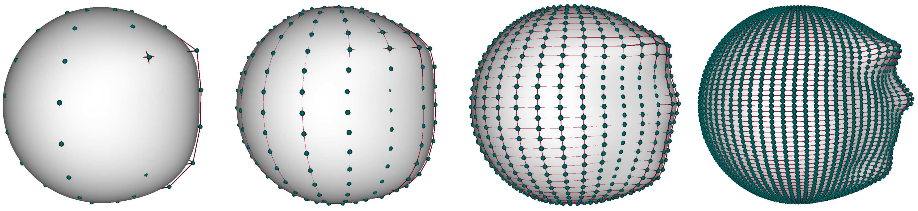


Figure 6.11: The iterative subdivisions of the spherical control point grid.

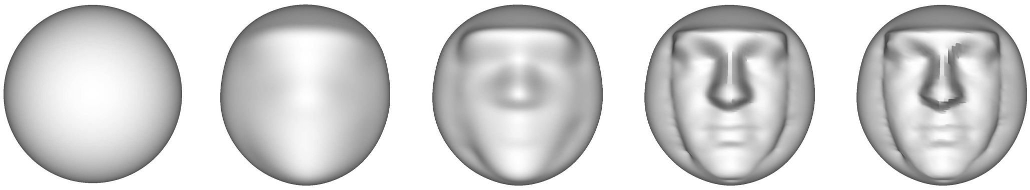
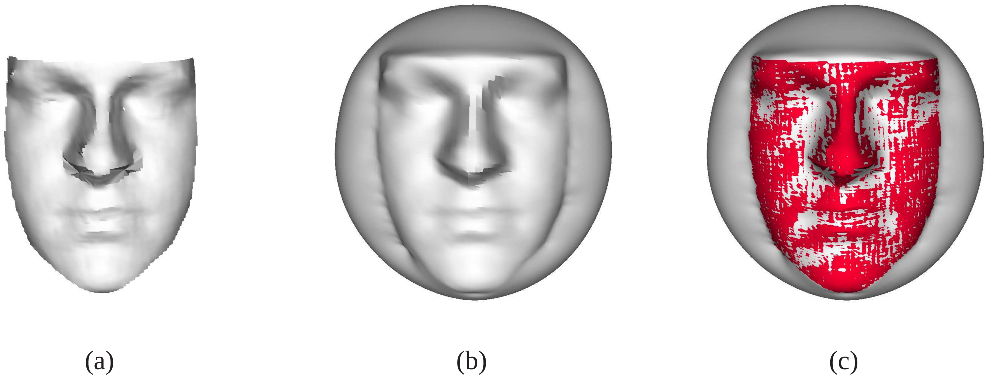


Figure 6.12: The iterative registration of the facial surface.

deformed sphere after registration (middle). The right image of the same figure shows the two aforementioned surfaces on top of each other. It is evident that the registration is of high definition and accuracy.

Apart from yielding faces with the same number of points (ideal for PCA), the B-spline registration of the faces allows for the correction of artifacts of the surface. Holes in the face that arise from the limitations of the data capturing hardware are corrected automatically as the sphere is smoothly deformed by fitting to the hole's surrounding points. The dataset in image (a) and (b) in Figure 6.14 has artifacts in areas where the structured light is not reflected well. After deforming the sphere these artifacts disappear as Fig-



(a) (b) (c)

Figure 6.13: (a) The original face, (b) the approximated face and (c) the overlap between the two.

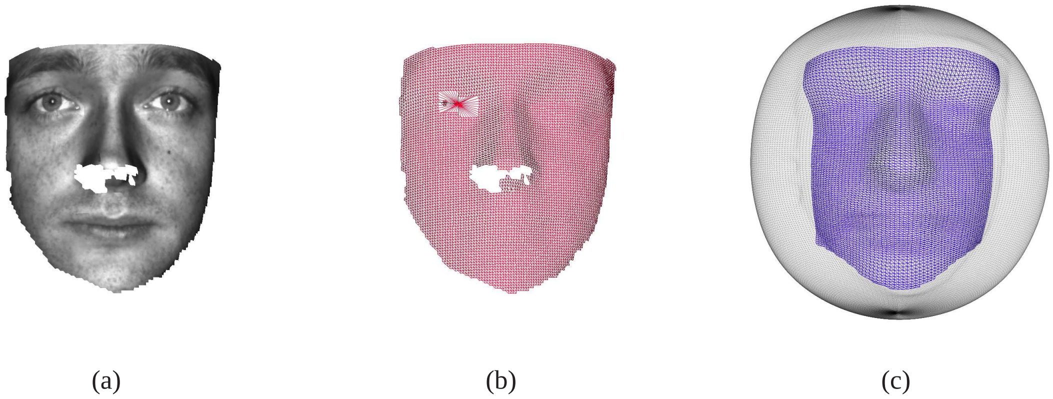

(a) (b) (c)

Figure 6.14: Holes are covered and grid is regularized after *SBR*.

ure 6.14(c) shows. The landmark- and ICP-based model building methods, presented in the previous chapter, were also eliminating holes by linking up points around the edge of a hole into one cell as discussed in section 5.3.1.3. In this case, however, actual points are generated to fill up the hole. The SBR algorithm can approximate a face in under $6secs$ on a $2GHz$ machine.

6.3.3 Face extraction

The SBR technique is applied to all the faces and an approximation of their original surface is generated. The surfaces have now the same number of points and a PCA-based model can be readily created. Figure 6.15 shows the mean face generated using this data. The spheres we used have a total of 40,000 points, but most of the points on them are not part of the original facial surface and it would therefore be desirable to eliminate them. More importantly, the border area on the sphere where the sphere points start approximating the facial points has the potential to negatively affect the recognition process. The two images on the left in Figure 6.16 show two approximated faces of the same subject. It is clear that in the border there is random variation that depends on how the data was preprocessed, which could affect the statistical methods employed. The image on the right of the same figure shows a top view of faces belonging to the same subject. It is evident that one surface is more protruded than the other in the borders of the faces. Eliminating “non-facial” points allows one to build the PCA model without variation that is not di-

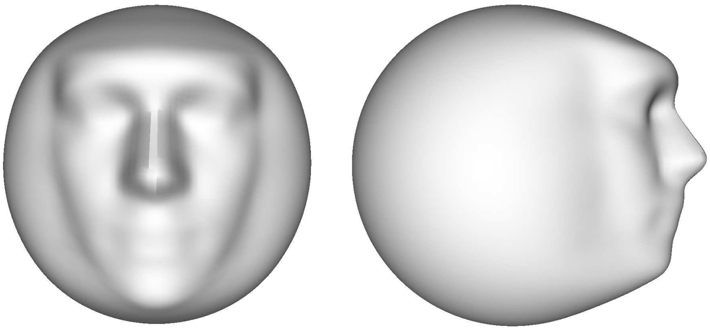


Figure 6.15: Mean face when using SBR.

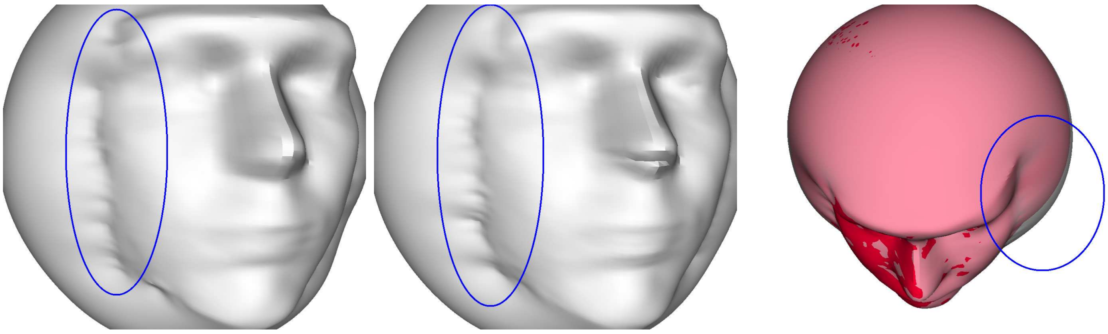


Figure 6.16: Variation in sphered data of the same subject. The left and centre images show two different biometric samples of the same subject. Notice how there is a difference between the highlighted areas of the two surfaces. The image on the right shows the two biometric samples (red and white to differentiate them) registered to each other. Notice how, despite belonging to the same subject the spheres exhibit significant variation where the face meets the sphere. These variations introduce noise into the model which can have a negative effect on the recognition rates.

rectly linked to the facial structure of the subjects. As already mentioned the initial sphere is deformed using a coarse grid which is iteratively subdivided to achieve greater degree of detail. During the first coarse iterations a significant part of the sphere points are deformed to approximate the data. This "global" movement will inevitably be contained in the data and needs to be eliminated. Figure 6.17 shows the variation encoded in the first principal component of the full sphere. The circled areas show examples of variation that does not explicitly relate to the subject's face. In order to eliminate this variation we chose to select a smaller area on the sphere, which represents the area deformed to approximate points on the face. The original faces are placed on top of their corresponding versions after SBR as seen in Figure 6.13(c) and the area on the sphere that is under the

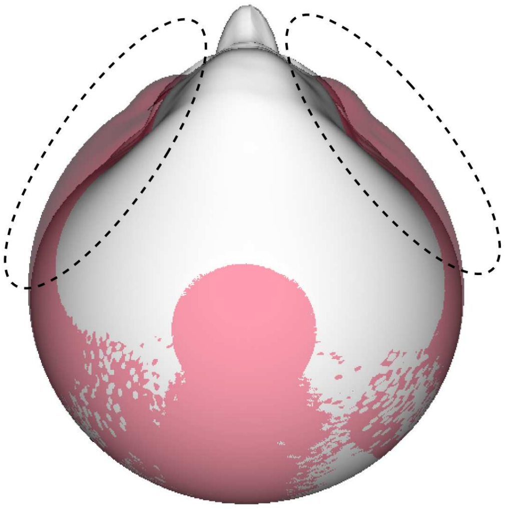


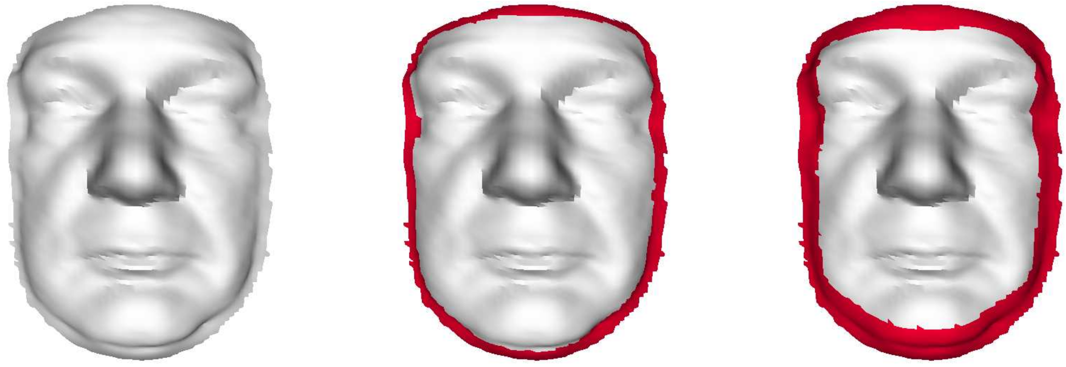
Figure 6.17: 1st principal component of sphere space. Notice how the noise introduced into the model by datasets such as the one in figure 6.16 is encoded explicitly into the model's 1st principal component.

original face is extracted. We define a sphere point $s = (s_x, s_y, s_z)$ as being exactly under the original facial surface A by finding the closest point on A and determining whether the point is an edge (boundary) point or not. An edge point is defined as a point that is part of a side that is used by only one polygon. Points that are in the boundary of each surface are identified and pairings that include them are ignored during the calculation of the transformation.

We, thus, define a function $edge(p)$ that takes a point and returns true if it is an edge point and false if it is not an edge point. Given a set of points $\boldsymbol{s}_i$ on the sphere and the set of corresponding closest points $\boldsymbol{a}_i$ on the original facial surface, the points which are then retained are given by:

$$\{\boldsymbol{s}_i : ||\boldsymbol{s}_i - \boldsymbol{a}_i|| < \tau, !edge(\boldsymbol{a}_i)\} \tag{6.10}$$

where τ is a user defined distance threshold, the value of which is heuristically chosen. The symbol "!" reverts the returned value of function $edge(p)$. Based on the above criteria the resulting surfaces will have different number of points, given that the faces are of different size and therefore occupy different area of the sphere. The points which were kept are the points that are present on $n\%$ of subjects. In other words after the faces are extracted from the sphere we discard a further set of points that does not exist on a user-



(a) $n = 1\%$ (b) $n = 50\%$ (c) $n = 100\%$

Figure 6.18: Choosing which points on the sphere to keep. (a) contains the points that are present in at least 1% of the subjects ($n = 1\%$). (b) contains the points that are present in at least half of the subjects ($n = 50\%$) and (c) contains the points that are present on all subjects ($n = 100\%$). Notice how the number of points decreases as the value of n increases.

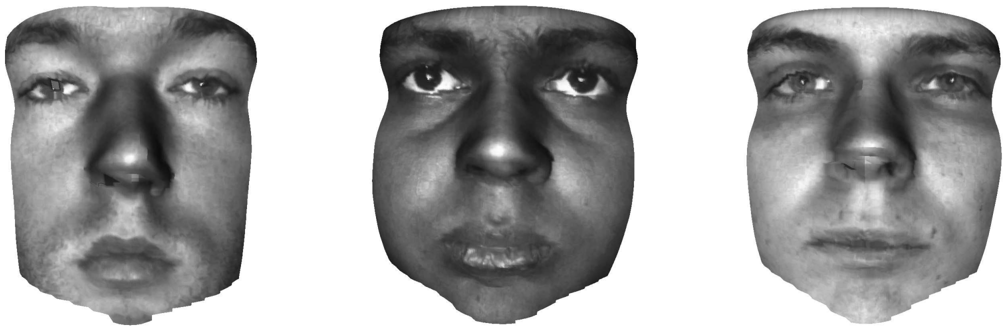


Figure 6.19: Examples of extracted faces.

defined percentage of faces. Figure 6.18 shows how changing this percentage changes the size of the face. Notice how setting n to higher values decreases the number of points kept. The red area around faces (b) and (c) shows the area that was clipped from the original image (a) by modifying the value of n. The datasets in Figure 6.19 are examples of the datasets (7000 points) after they have been fully processed. All datasets contain the same number of points and have no surface artifacts or irregularities that would have adverse effects on the recognition effort. Figure 6.20 shows the processing pipeline when SBR is used.

Figure 6.21(a) shows the facial surface that has been created without non-rigid surface registration while Figure 6.7(b) shows the surface when non-rigid surface registration

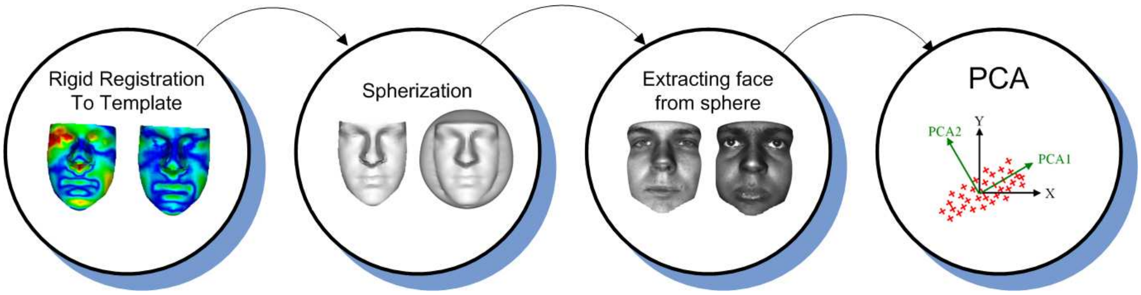
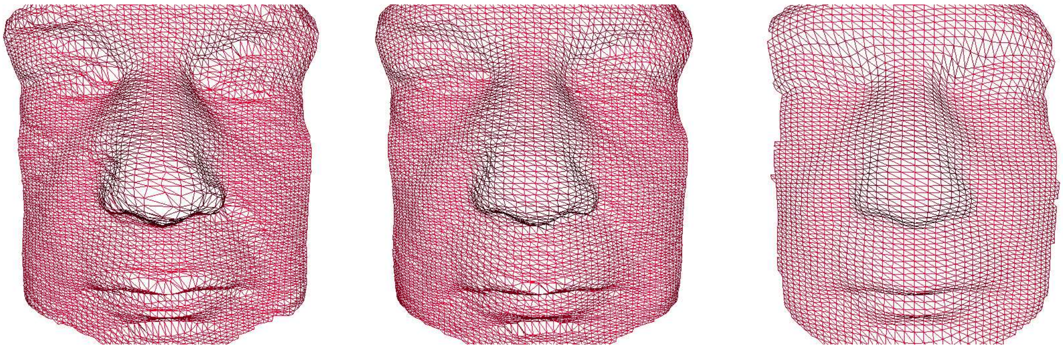


Figure 6.20: The processing pipeline when SBR is used.



(a) Before non-rigid surface registration.

(b) After non-rigid surface registration.

(c) After SBR

Figure 6.21: Effects of non-rigid surface registration and SBR on the uniformity of face points.

(NRSR) has been used. Notice how Figure 6.21(c) which has been created using the SBR, has uniform point distribution with no dense or sparse point clusters.

The SBR itself can not be added as a step in the pipeline of the landmark-based model, as it was done with the ICP-based one. As mentioned earlier, the point ids are used in the landmark-based model to temporarily store the established correspondences. When using SBR the surface points are regenerated and therefore the point ids that are stored become irrelevant. Nevertheless, to demonstrate that uniform surfaces allow better correspondences to be established, we use SBR to regenerate the surfaces. These regenerated surfaces are then used in the landmark-based technique as presented in Chapter 5. This model is referred to as the SBR landmark-based model. The type of registrations involved in each of the three landmark-based techniques and their order is contrasted in Table 6.1.

The type of registrations involved in each of the three ICP-based techniques and their

Landmark-based model techniques				
Technique	SBR	Rigid landmark reg.	Non-rigid landmark reg.	Non-rigid surface reg.
Original landmark-based	-	step 1	step 2	-
SBR landmark-based	step 1	step 2	step 3	-
NRSR landmark-based	-	step 1	step 2	step 3

Table 6.1: The three landmark-based techniques and the order of the registrations involved in the creation of the model.

ICP-based model techniques			
Technique	Rigid surface reg.	SBR	Non-rigid surface reg.
Original ICP-based	step 1	-	-
SBR ICP-based	step 1	step 2	-
NRSR ICP-based	step 1	-	step 2

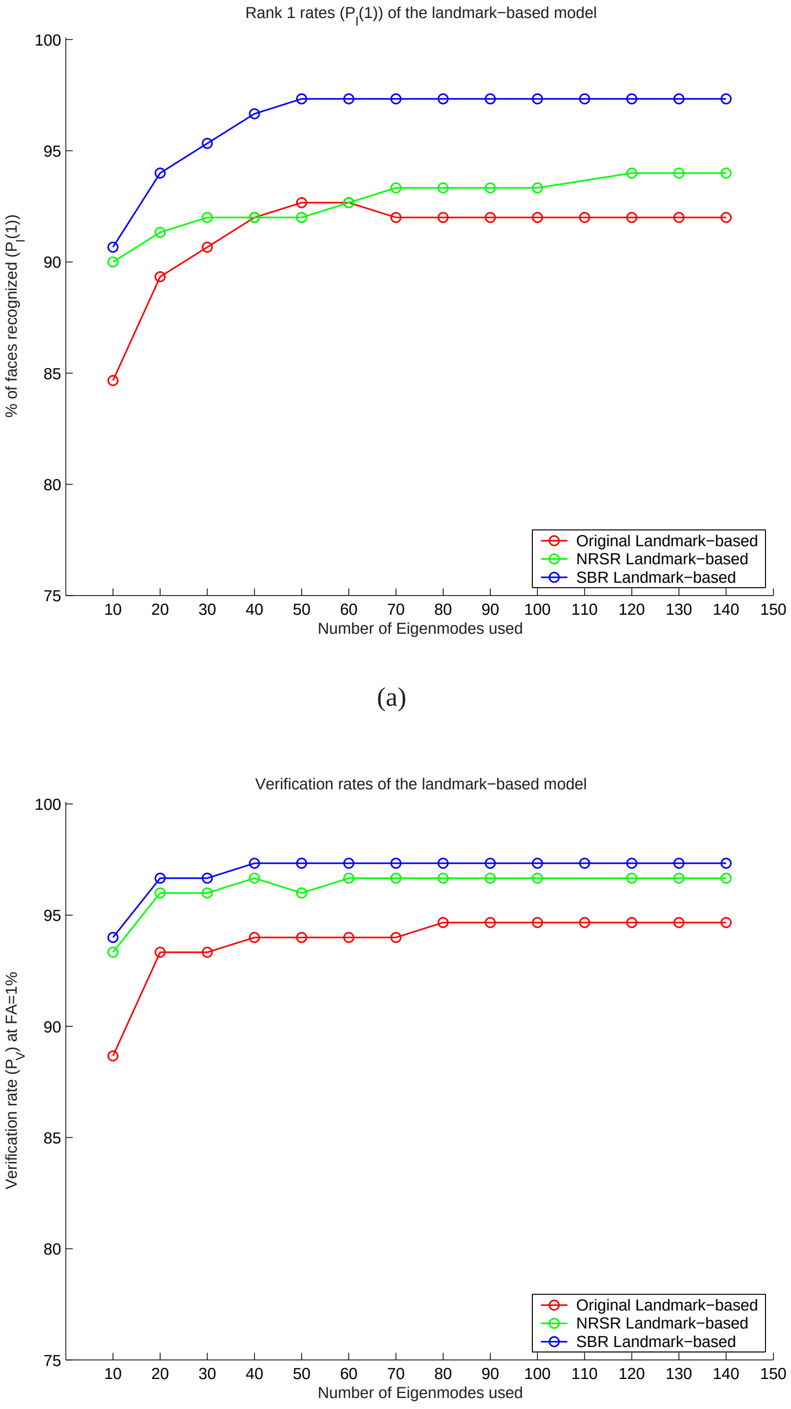
Table 6.2: The three ICP-based techniques and the order of the registrations involved in the creation of the model.

order is contrasted in Table 6.2.

6.4 Results

In order to assess the strength of the model-building techniques we perform the same face recognition tests as we did in the previous chapter. First, we will assess the effect of the non-rigid surface registration (NRSR) step in both the landmark-based and ICP-based statistical model. We will also evaluate the effectiveness of the spherical non-rigid registration (SBR) in generating datasets to be used with PCA. As demonstrated, the model that includes an SBR, already contains a rigid registration before the surfaces are transformed into spheres and extracted from them. In that way it is very similar to the ICP-based model. For this reason we are going to refer to those results as SBR ICP-based.

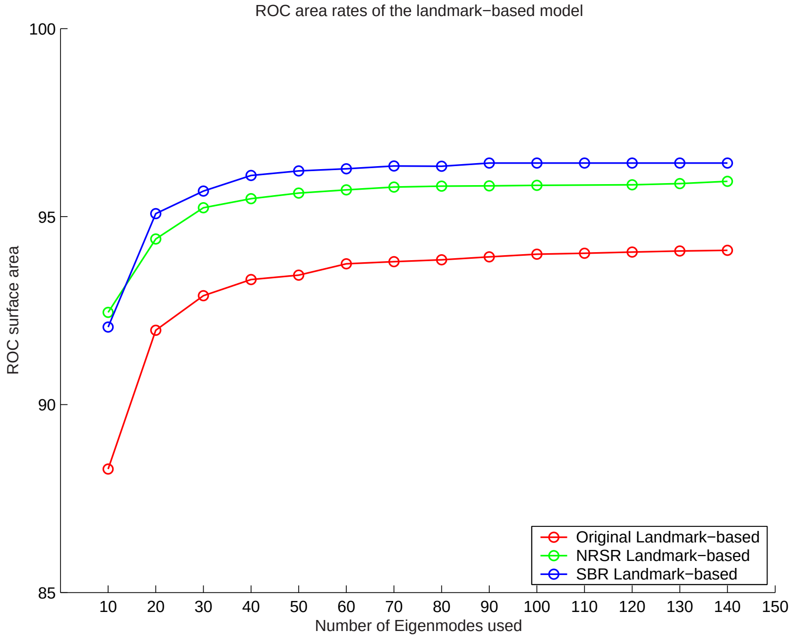
Figure 6.22(a) shows the rank 1 rates of the original landmark-based model, as well as the landmark-based models that contain a non-rigid surface registration (NRSR) and spherical non-rigid registration (SBR). Figure 6.22(b) shows the verification rates for the aforementioned landmark-based models. Figure 6.23(a) and 6.23(b) show the ROC and $P_{FA} = P_{DI}$ rates respectively. According to all the measurements the models that include an NRSR and SBR step, improve the rates compared to the original version of the landmark-based statistical model-building technique.



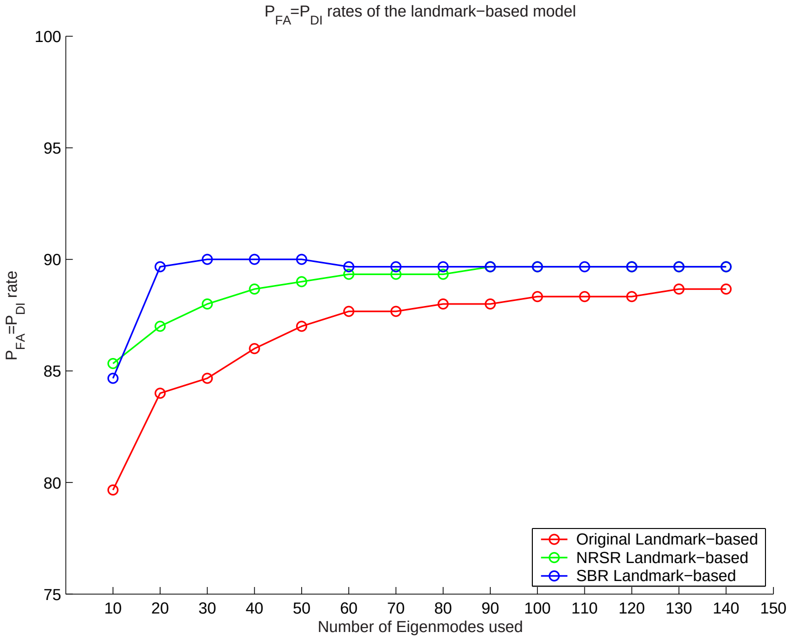
(a)

(b)

Figure 6.22: The rank 1 and verification rates of the landmark-based models.



(a)



(b)

Figure 6.23: The ROC and $P_{FA} = P_{DI}$ rates of the landmark-based models.

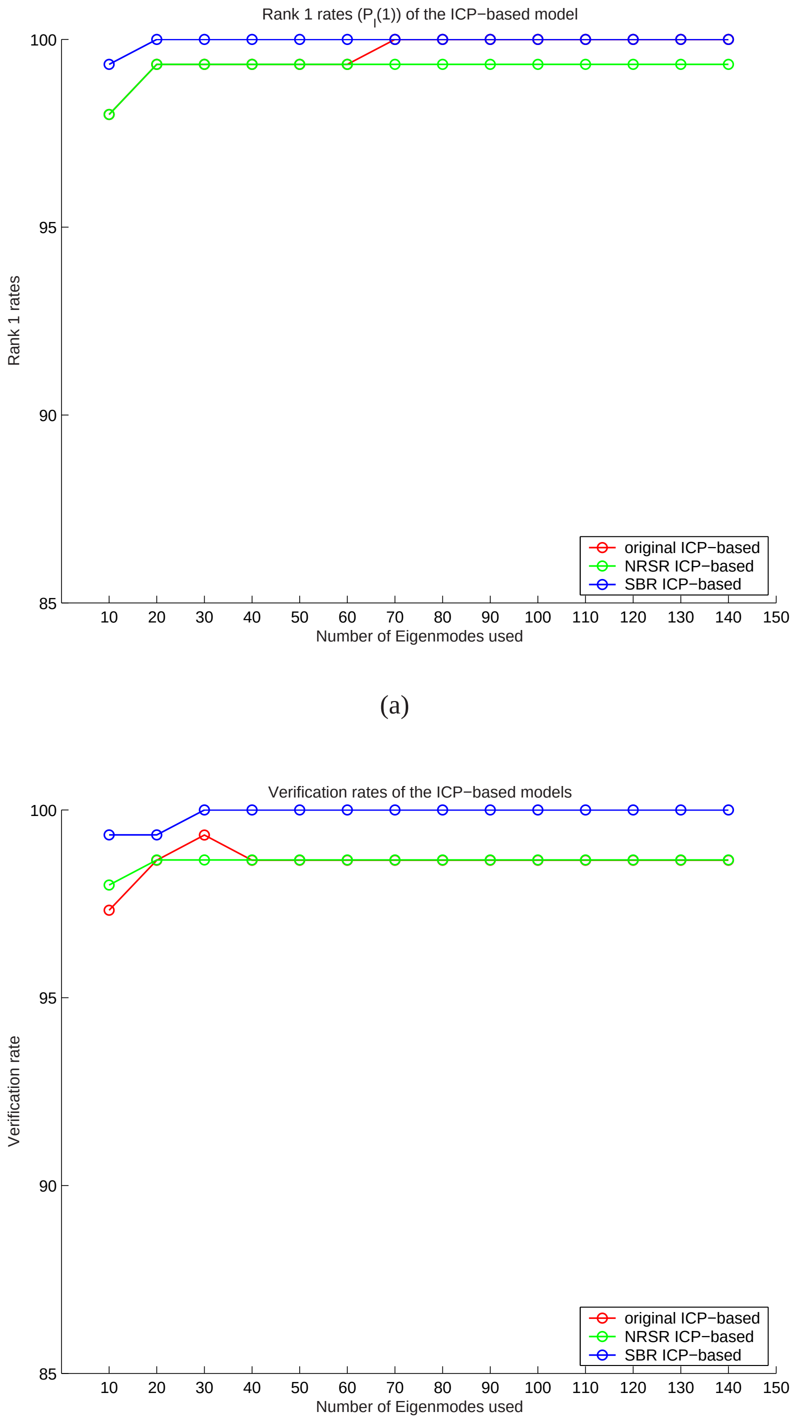
Figure 6.24(a) shows the rank 1 rates of the original, NRSR and sphered ICP-based model. Figure 6.24(b) shows the verification rates for the aforementioned ICP-based models. Figure 6.25(a) and 6.25(b) show the ROC and $P_{FA} = P_{DI}$ rates respectively. In this case an improvement is visible in the rates when using the sphered data but no significant difference between the original ICP-based technique and the one that involves a non-rigid surface registration in its pipeline.

Figure 6.26 and Figure 6.27 show the statistical evaluation tests, generalization ability, specificity and compactness for the six types of model generation.

6.5 Discussion

The results clearly demonstrate the effectiveness of the non-rigid surface registration in providing a better, more uniform correspondence between point sets and thus improving the model. In the case of the landmark-based model, where the correspondence is problematic because it is established using a manually selected array of points the effect is even greater. The landmark-based model, that has an NRSR step included, performs significantly better in classification tests than the baseline landmark-based technique. Further improvement is achieved when the surfaces are resampled with the SBR to create a uniform grid with no artifacts which allows for a relatively noise-free correspondence to be established. More concretely, the basic landmark-based technique reaches a maximum of 92% (rank 1) while the technique which includes the NRSR step reaches 94%. On the other hand using the SBR technique allows the rank 1 rake to 97.5%.

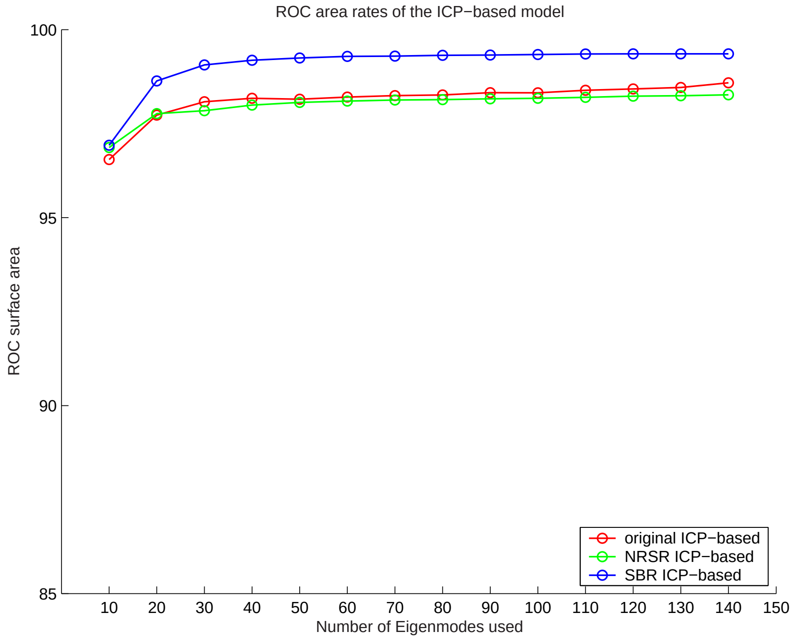
The effects on the ICP-based model are less pronounced. The NRSR step does not produce the effects seen with the landmark-based model. The correspondence established with the ICP-based is already quite good and more uniform than the landmark-based one and this result is expected. Nevertheless, the SBR improves the results of the ICP-based model further bringing the rank 1 and verification rates up to 100% even when using only 20 and 30 parameters respectively. The rank 1 rate of the ICP-based technique reaches 100% recognition but only after 70 eigenmodes are used. In contrast the technique which



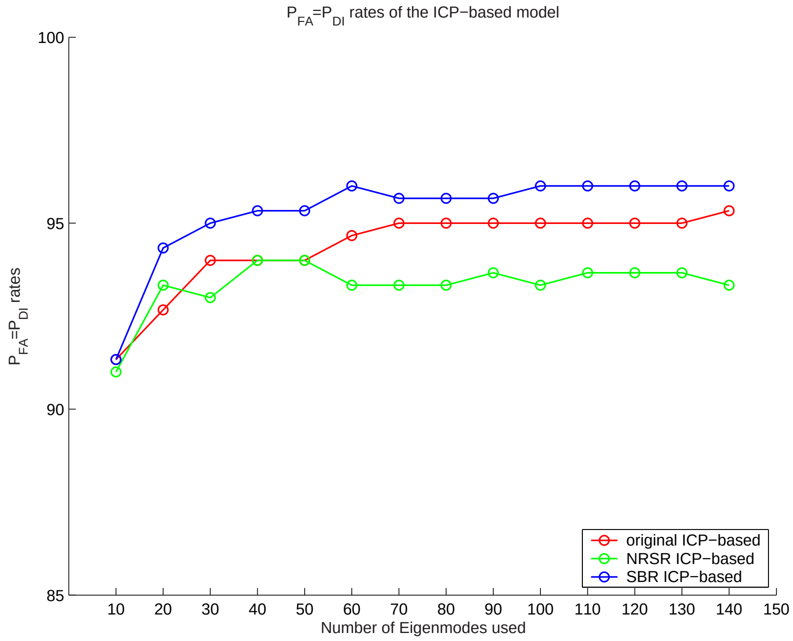
(a)

(b)

Figure 6.24: The rank 1 and verification rates of the ICP-based model.

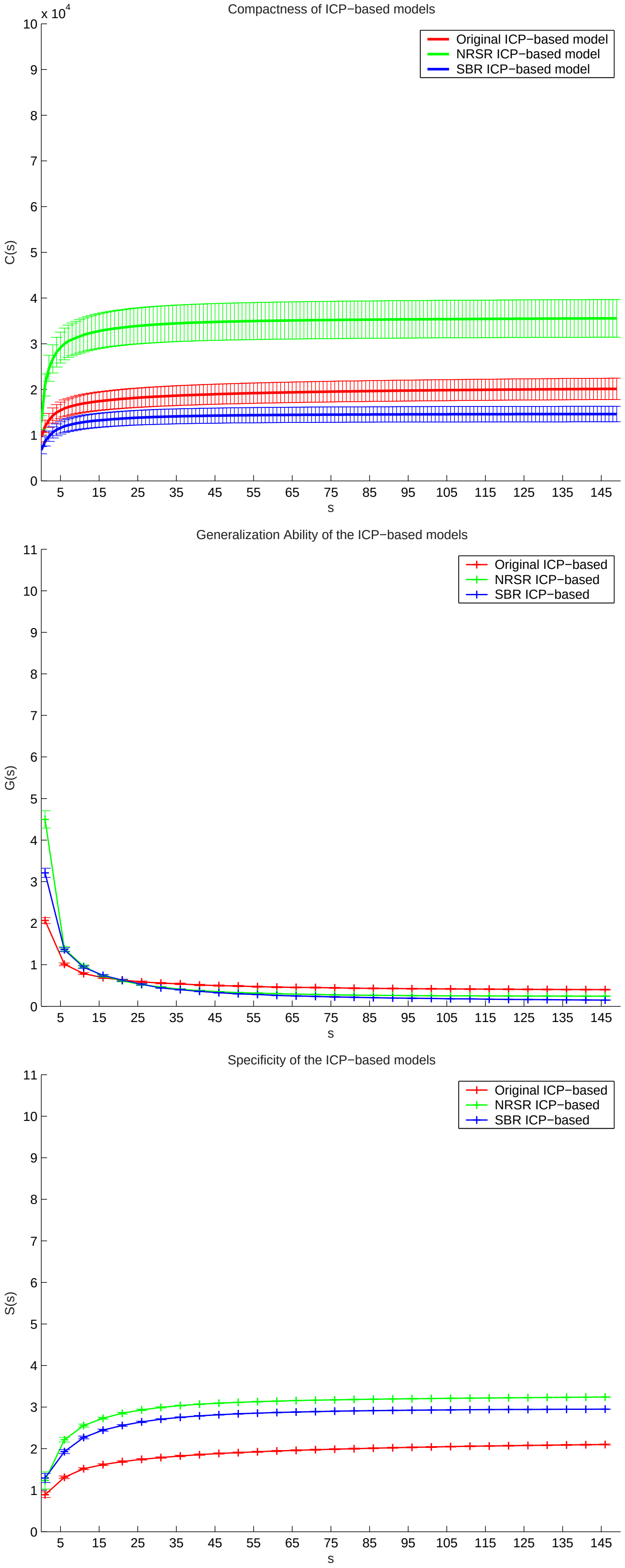


(a)



(b)

Figure 6.25: The rank 1 and verification rates of the ICP-based model.


Compactness of ICP−based models

Generalization Ability of the ICP−based models

Specificity of the ICP−based models

Figure 6.26: Statistical tests on the ICP-based models.

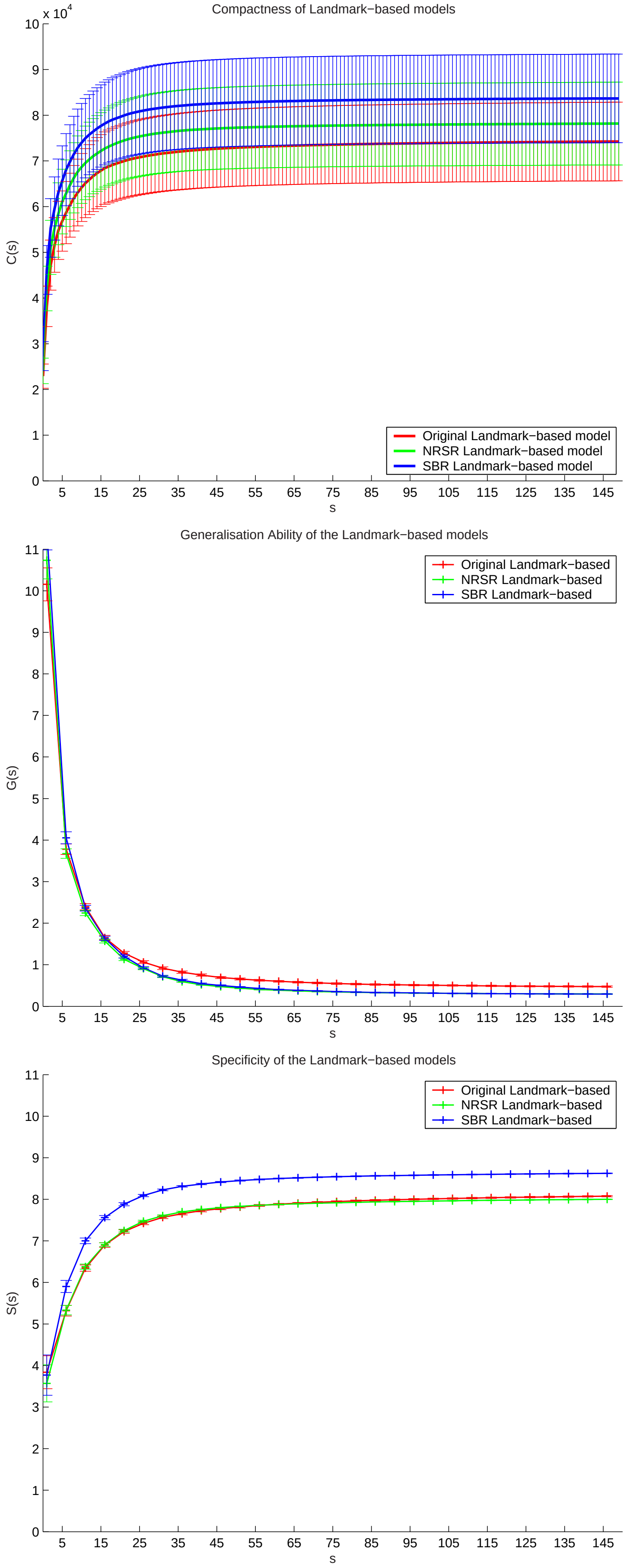


Figure 6.27: Statistical tests on the landmark-based models.

include an SBR step reaches 100% even when 20 eigenmodes are used.

Performing generic tests on the models we observed that they are quite similar. The results demonstrated no significant differences in the models in terms of generalization ability, compactness and specificity.

Chapter 7

Facial feature analysis and automatic model optimization

7.1 Introduction

In Chapter 5 we proposed two general model-building techniques for the construction of statistical face models: a landmark- and an ICP-based method. Chapter 6 expanded on these by introducing a non-rigid surface registration step in the model-building process. Both chapters dealt with building a face model that uses the entire face. In this chapter we investigate which parts of the 3D facial shape are most useful for face recognition. Two methods are presented: the first one involves the manual segmentation of the facial data while the second one involves the exclusion of facial parts based on their stability across facial expressions.

Early on in face recognition, the use of certain regions of the human face was examined as a possible alternative to using the entire face. Pentland *et al.* [154] automatically detected facial features and using these eigenfeatures by themselves a 95% rate was achieved in rank 1 tests. This shows that in lower dimensions the eigenfeatures outperform the eigenface recognition. Emidio *et al.* [56] used just the eye to perform classification. When only a small number of model parameters were used, the whole face was a more powerful classifier than the eye. However, when more than 15 model parameters were used, the eyes were performing better in classification tests.

As discussed in Chapter 1, face recognition can be affected by a number of factors such as facial expressions, occlusion, viewing angle as well as aging. The human brain uses both global and local features to recognize faces, possibly as a way of making person identification more robust. Similarly, in machine face recognition, techniques might become more robust by using parts of the face that are less affected by the aforementioned extraneous variables. For example, when trying to identify a person from a smiling image, the mouth and the eyes are probably not the best characteristics to use because their shape changes significantly. However, the upper part of the nose changes very little due to facial expressions and might be more appropriate to use. Some researchers report that the eyes are less prone to the effects of time [93] and thus using just the eyes might improve the classifiers' performance. Especially in cases when the training set is not particularly large, adding facial features that are noisy or highly correlated to each other can have detrimental effects on the recognition rates.

In the sections that follow, the two approaches for using subsets of the facial surface for facial recognition are presented. These subsets are extracted from the ICP- and landmark-based statistical face models presented in Chapter 5. The extensions proposed in Chapter 6 were not employed in order to evaluate the techniques presented in this chapter using the baseline methods. The techniques of Chapter 6 increase the recognition rates significantly and given that they are already relatively high, it would have been more difficult to detect differences in performance between experimental conditions.

7.2 Anatomical face segmentation

The first method to identify the parts of the face that are most useful for recognition is to segment the face into anatomical regions and test each of them individually. All model-building techniques presented so far, whether surface- or landmark-registration based, have a common goal: establishing point correspondence across all faces in the population. To this end, it is sufficient to segment the average face. The resulting segmentation can

Region size		
Region	Number of points	Percentage of whole face
cheeks	2241	40%
eyes	881	15%
nose	814	14%
mouth	733	13%
forehead	578	10%
chin	461	8%

Table 7.1: The size of the segmented facial parts.

then be transferred to all other faces using the established point correspondences.

Using publicly available software (www.blender.org), the average face of the landmark- and ICP-based techniques was segmented into separate anatomical regions. The six regions are: the forehead, the two eyes, the nose, the two cheeks, the mouth and finally the chin as shown in Figure 7.1. Areas of the same color were regarded as a single feature even if they are not connected to each other. It is worth noting that the so-called "forehead" in the following sections is not what is colloquially referred to as forehead. It has been substantially trimmed during the model-building stage and does not include more than a couple of centimeters above the eyebrows. Additionally, the two eyes, just like the two cheeks, are grouped together. The assumption made is that the two regions around the eyes of the face are equal in terms of information and discriminatory power. In contrast, to perform modular face recognition Pentland *et al.* [154] used the left and right eye (separately), the nose and the mouth as features. Part of the reason for that is the fact that they detected the features automatically and these are the most easily identifiable features on a 2D face. On the 3D data that used in this work, where correspondence has been established across all faces, the template face is manually segmented and the segmentation is propagated across the population. This way, one could select any features of the face without considering issues such as feature detection. The regions we selected were based on the commonly used regions (eyes, nose, mouth) plus a segmentation of the remaining features (cheeks, chin, forehead). Table 7.1 lists the facial regions, the number of points that make up each of them and the percentage that each part is of the whole.

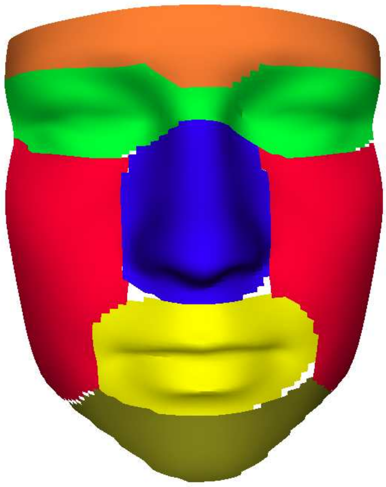


Figure 7.1: The manually segmented features of the average face generated by the landmark-based technique. Areas of the same color are considered as one even if they are not connected to each other.

7.2.1 Eigenfeatures

Once all faces are automatically segmented using the manually generated segmentation of the average face from each of the two models, a separate eigenspace can be build for each feature using PCA. This results in six independent featurespaces referred to as *eigeneyes*, *eigenforeheads*, *eigennoses*, *eigenmouths*, *eigenchins* and *eigencheeks*. Tables 7.2 and 7.3 show the first two principal modes of variation for the nose and the cheeks of the ICP-based statistical model. A comparison of the features' first modes of variation and of the whole face from Tables 5.5 and 5.4 in Chapter 5 readily demonstrates that the modes of variation of the features are not the same when the features are analyzed individually as when they are part of the face. Each feature varies differently when it is part of a collection of other facial features than when it is by itself. This observation is at the heart of the assumption that individual features might perform well in face recognition tests even though they might only be a subsection of the whole face.

Given that the correspondence has been established between the surfaces in 3D and given that the correspondences are also known between the 3D surfaces and the 2D texture data, this technique could also be used to segment the bitmaps of the face for calculating separate facespaces for 2D information. If this were to be combined with a feature

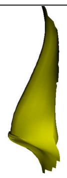
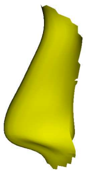
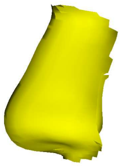
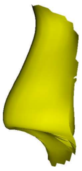
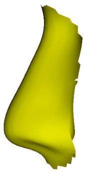
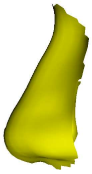
Eigennose of the ICP-based model			
	$-3\sqrt{\boldsymbol{\lambda}}$	**mean**	$+3\sqrt{\boldsymbol{\lambda}}$
mode 1			
mode 2			

Table 7.2: The first two principal components of the nose-space.

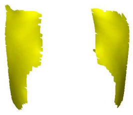
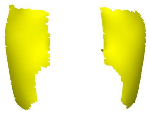
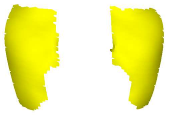
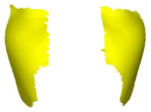
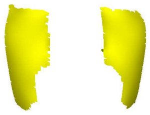
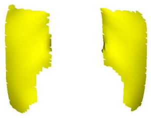
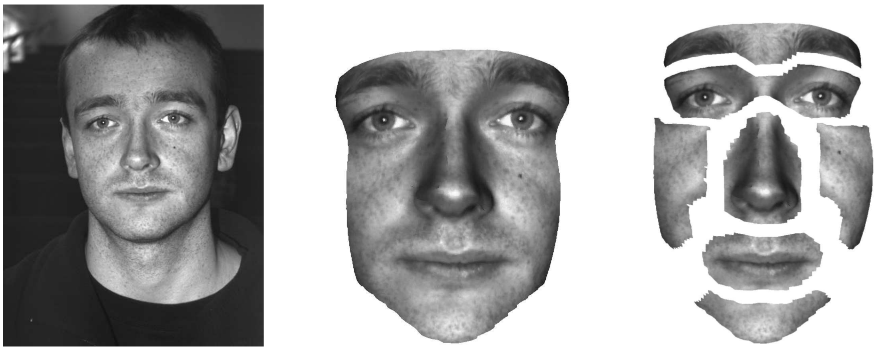
Eigencheeks of the ICP-based model			
	$-3\sqrt{\boldsymbol{\lambda}}$	**mean**	$+3\sqrt{\boldsymbol{\lambda}}$
mode 1			
mode 2			

Table 7.3: The first two principal components of the cheek-space.



(a) original texture

(b) original 3D model

(c) segmented texture

Figure 7.2: Segmented facial parts with texture.

detection method, it would provide a powerful 2D segmentation technique. Figure 7.2 shows the 2D information of a face being segmented into regions automatically by using the correspondence propagated to it from the manually segmented model mean. For demonstration purposes, the face segmented in this case belonged to the VRT3D database because of the poor quality of textures in the Notre Dame datasets.

7.2.2 Face recognition using eigenfeatures

7.2.2.1 Comparison of eigenfeatures

As in previous chapters these feature models can be used for classification by calculating the distance between subjects in the featurespace. The population used to build each featurespace is the 150-strong Notre Dame database. Each feature was split into a gallery set $\mathcal{G}$ and a probe set $\mathcal{P}$. The results that are reported in the following were obtained using the Euclidean distance metric to measure similarity. Similar results were obtained using the Mahalanobis distance, but those graphs are not displayed in order to reduce the size of result section.

The first task is to compare the classification ability of each of the eigenfeatures individually. Figures 7.3 and 7.4 show the recognition rates for each anatomical region. For the landmark-based model, the strongest feature for classification is the nose followed by the cheeks, eyes, mouth, forehead and chin. The dashed line shows the recognition rates when using the entire face. The results demonstrate the great discriminatory power of some of the features. The nose in particular, with only 814 points (14% of whole face), manages to reach a rank 1 rate of 90% when more than 90 parameters are used and a verification rate of 95%.

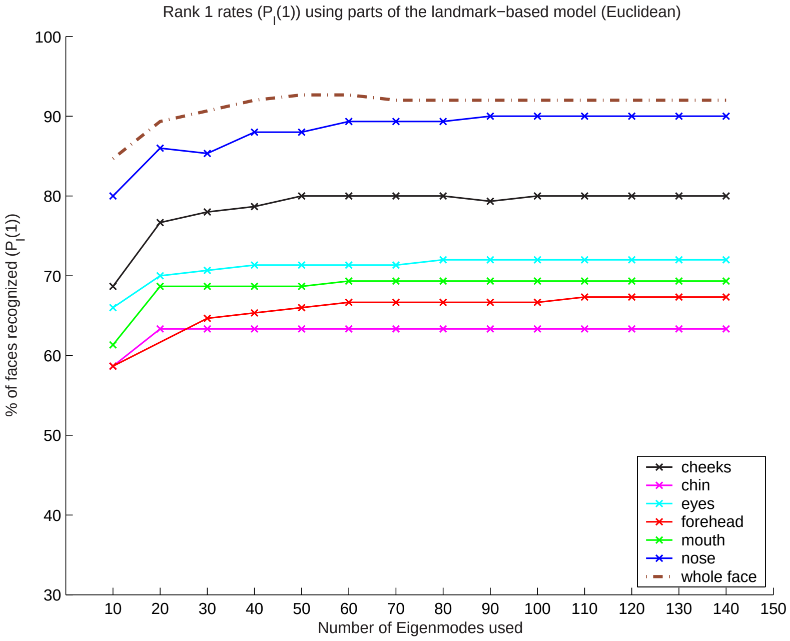
The same experiments were conducted using the ICP-based model and Figures 7.5 and 7.6 show the classification rates for the facial regions. In the case of the ICP-based model, the strongest feature for classification is, once again, the nose (rank 1 rate of 98.6%), followed by the eyes, cheeks, mouth, chin and forehead.

7.2.2.2 Combining eigenfeatures

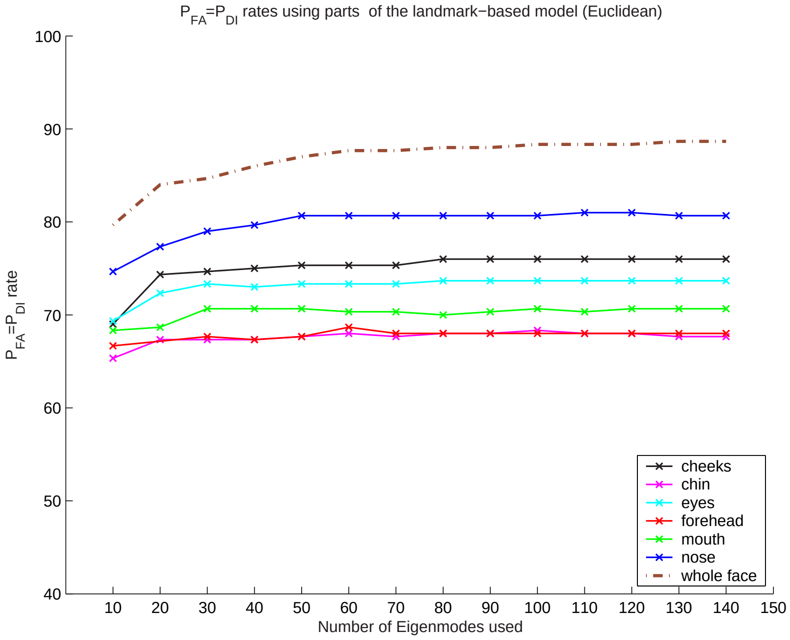
Another way to take advantage of the classification ability of the individual features is to use them in a combined fashion. An experiment conducted using a combination of facial regions involved the incremental addition of the most powerful regions to the statistical model and testing each of these models individually.

The regions are incrementally concatenated to the model, in the case of the landmark-based model starting with the most powerful discriminant (nose) to the weakest (chin), as demonstrated in Figures 7.3 and 7.4. Figure 7.7 shows the different models of the landmark-based type that were tested. For the ICP-based model the order was slightly different, given that the classification power rank of the different regions was slightly different itself.

Figures 7.8 and 7.9 show the recognition rates as the segments of the landmark-based type are concatenated to the model. The results are surprising: Adding the two regions which perform best when used individually into one dataset does not improve the results. In fact, when used by itself, the nose performs better than the nose and the cheeks together. As more regions are added to the nose-cheek model the closer the rates get to the recognition rates of the whole face. Since Figures 7.8 and 7.9 showed that the recognition rates for the landmark-based model dropped significantly when cheeks were used and subsequent additions improved the results, a different order of features was used. The cheeks were added last while the rest of the order remained the same (nose, eyes, mouth,

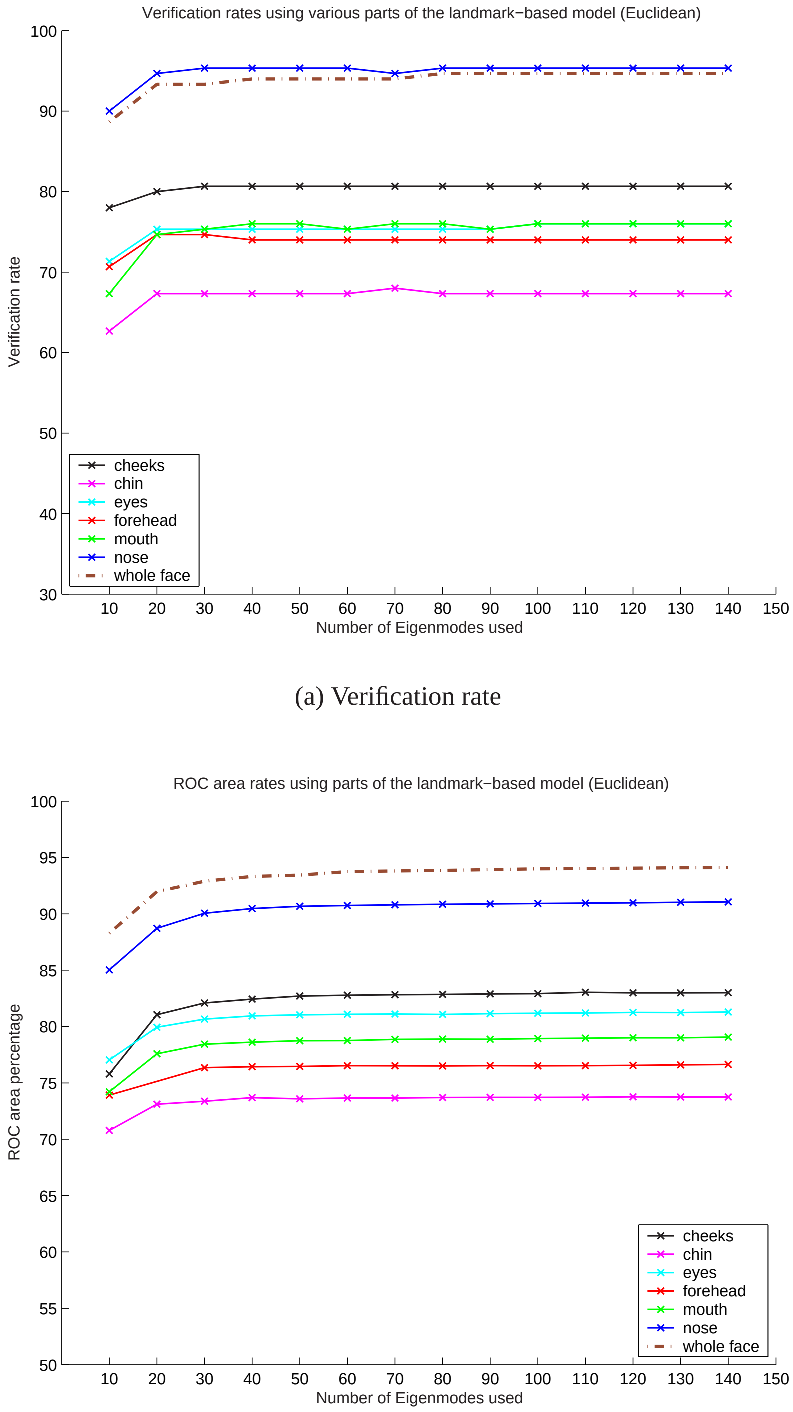


(a) Rank 1 rate



(b) $P_{FA} = P_{DI}$ rate

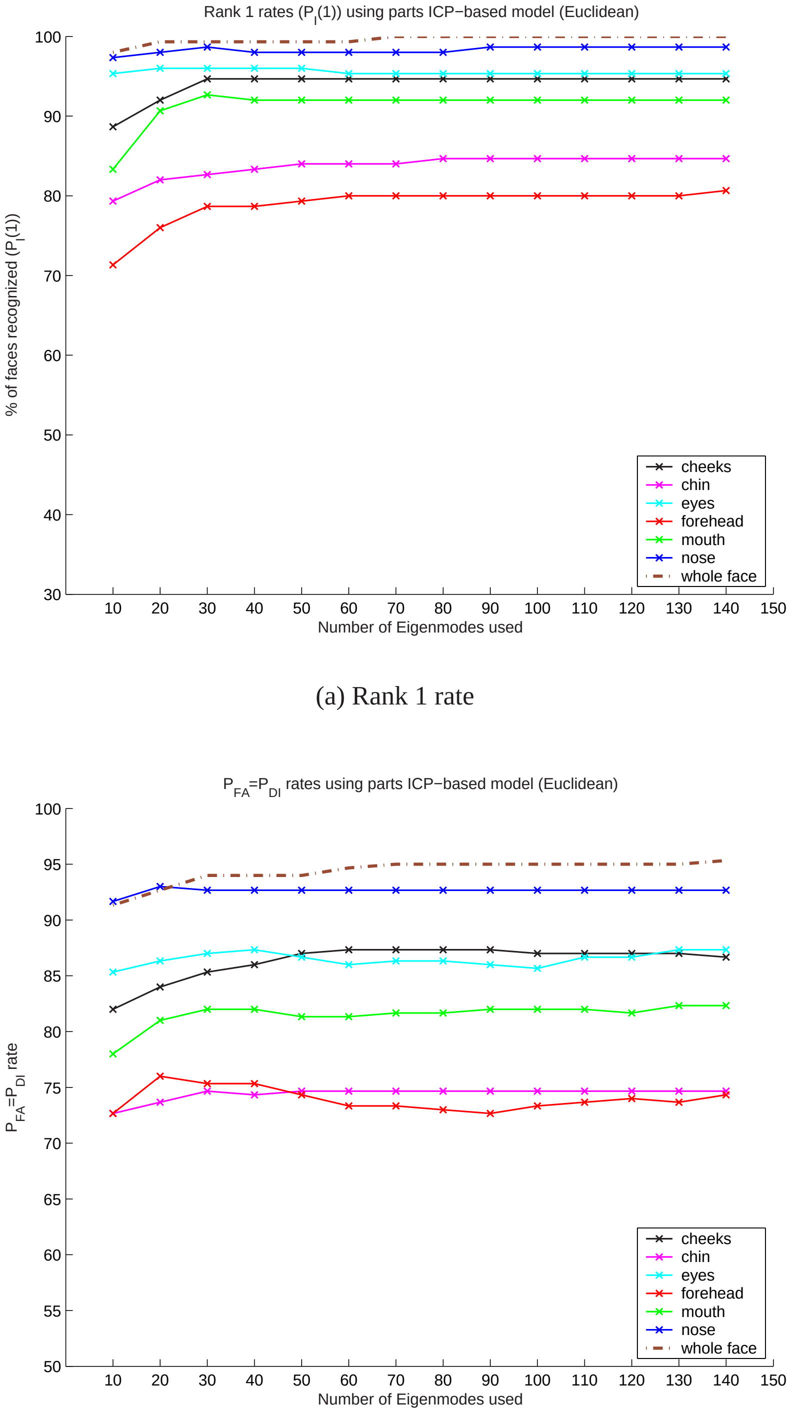
Figure 7.3: The classification rates for the segments of the landmark-based model.



(a) Verification rate

(b) ROC rate

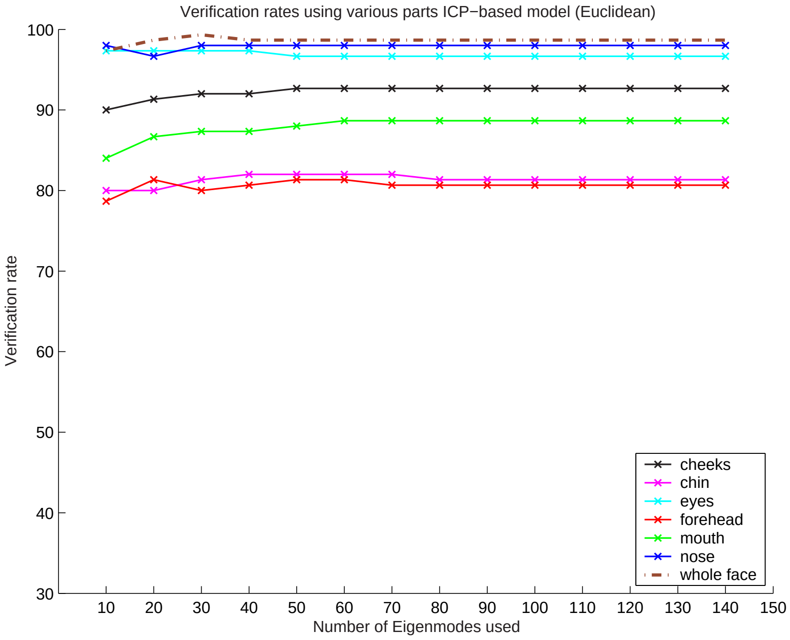
Figure 7.4: The classification rates for the segments of the landmark-based model.



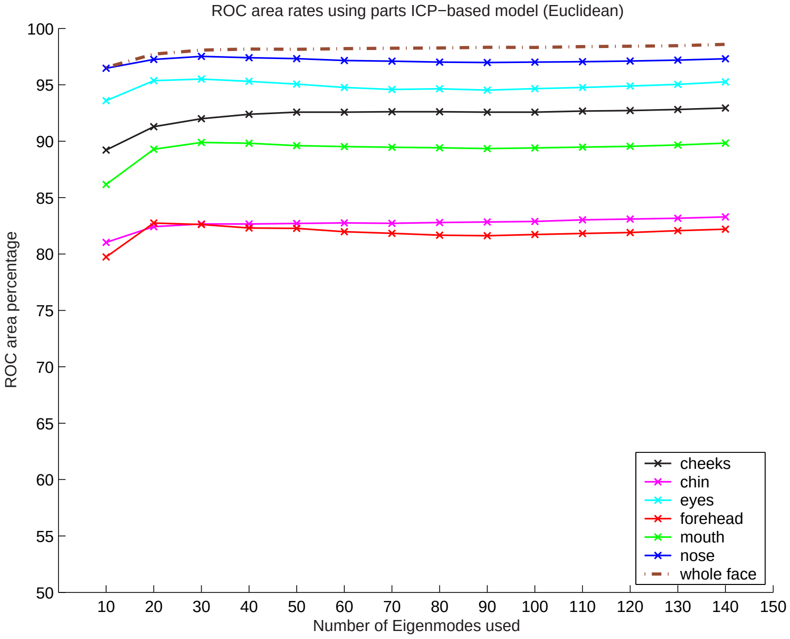
(a) Rank 1 rate

(b) $P_{FA} = P_{DI}$ rate

Figure 7.5: The classification rates for the segments of the ICP-based model.



(a) Verification rate



(b) ROC rate

Figure 7.6: The classification rates for the segments of the ICP-based model.

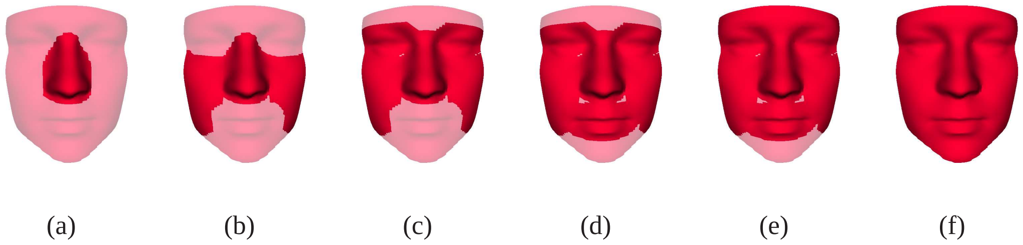


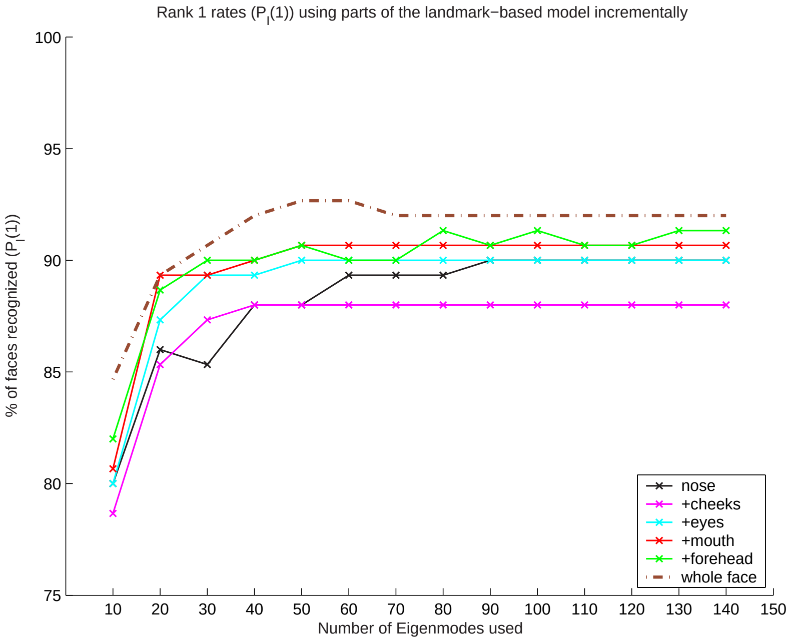
Figure 7.7: Incremental addition of the facial parts of the landmark-based model. The bold areas are the parts of the face that make up the model. The opaque ones are not used in the facespace. Model (a) is a nose-space while (f) in the end uses the whole face to create a model.

forehead, chin, cheeks). Figures 7.10 and 7.11 show the results for this order of region additions. Also in this case, the two first regions (nose and eyes) perform worse when used in conjunction than when the nose is used individually. Adding the third component (mouth) improves the rank 1 rates. According to all four measurements, using all but the cheeks either improves or has no effect on the recognition rates compared to when the whole face is used. This might be explained by the fact that most points on the cheeks are far away from any landmarks and thus the correspondence between the points across all subjects might not be as good as for other features.

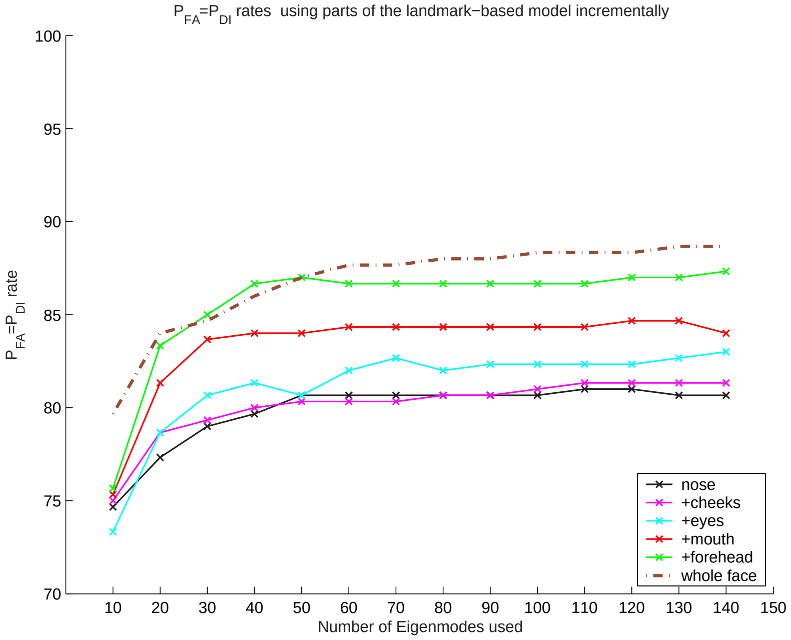
The same incremental approach is adopted for the ICP-based model, based on the order of classification power of the specific model (nose, eyes, cheeks, mouth, chin, forehead). Figures 7.12 and 7.13 show the recognition rates as the individual segments are concatenated to the model. Once again, it can be seen that certain combinations of features perform better than the whole face. In Figure 7.12(a), the nose and the eyes combined perform almost as good as the whole face with only 29% of the points of the original surface. In Figure 7.12(b) and Figure 7.13(b) of the same figure, the nose-eyes model manages to perform consistently better than the whole face combined.

7.2.2.3 Decision fusion using eigenfeatures

Another way of combining information from the different regions is not by concatenating their points into one statistical model but by fusing the scores of each of the feature classifiers into one score for classification (ie. adding distances in the feature-space). This

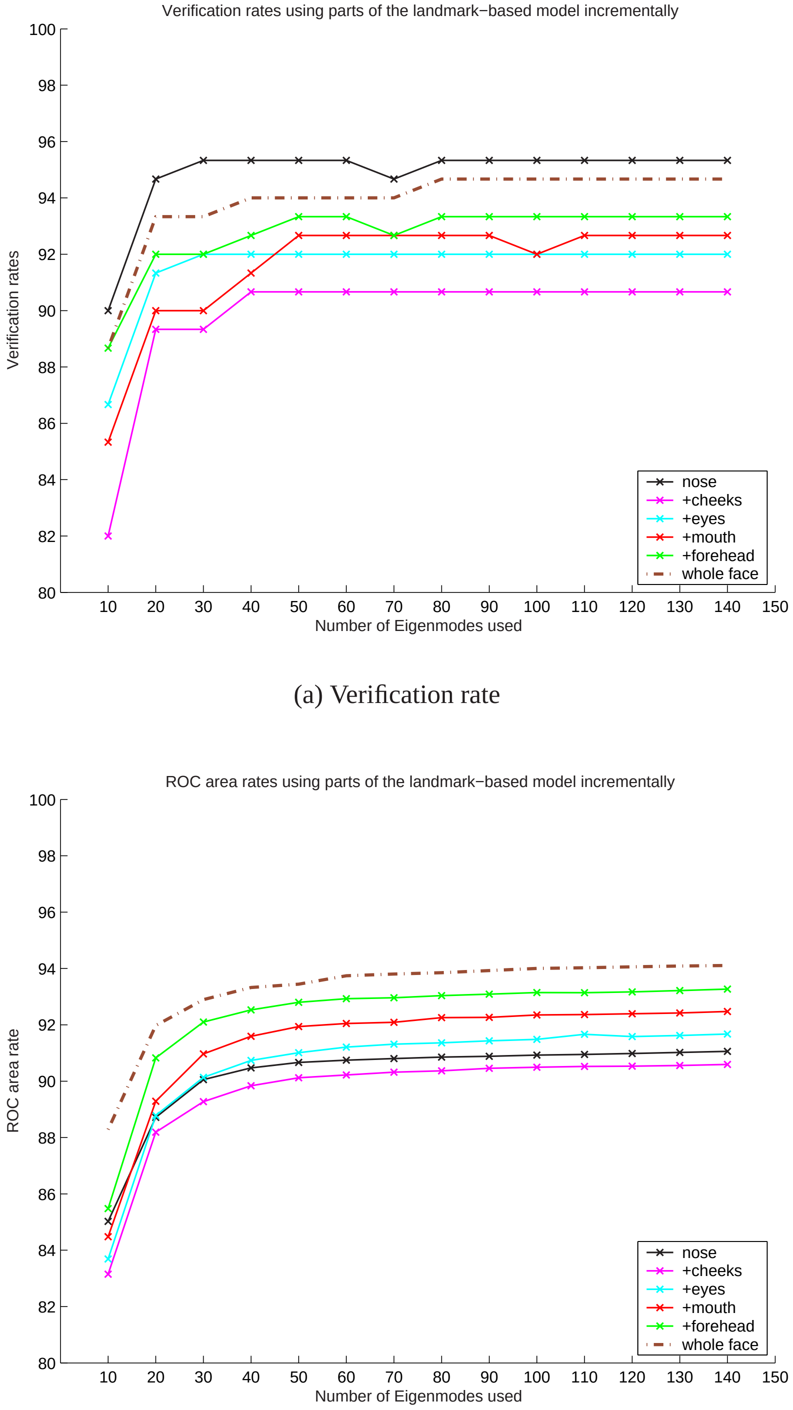


(a) Rank 1 rate



(b) $P_{FA} = P_{DI}$ rate

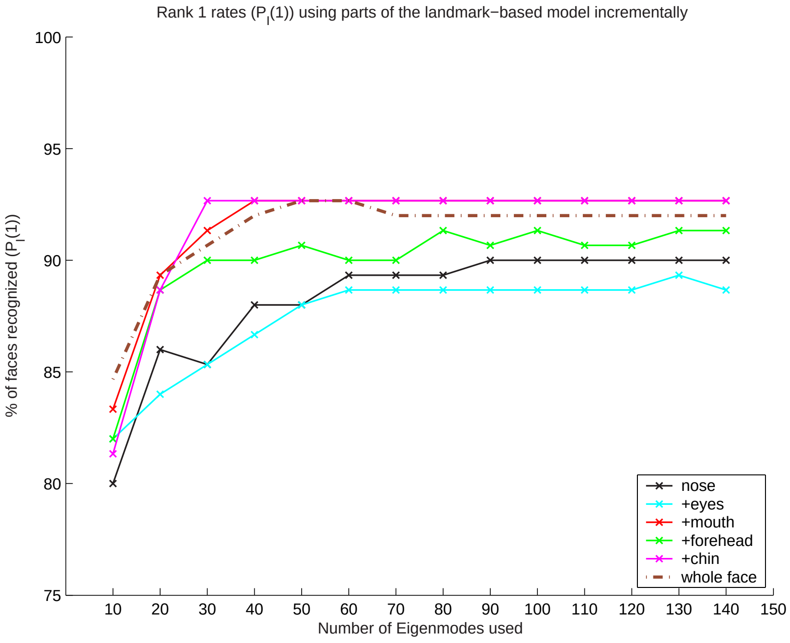
Figure 7.8: Rates of the incremental addition of the segments of the landmark-based model.



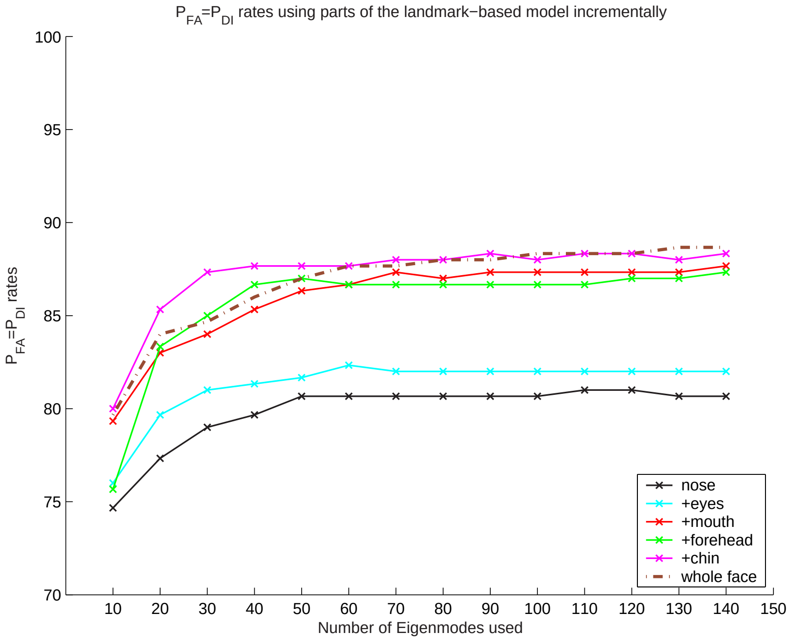
(a) Verification rate

(b) ROC rate

Figure 7.9: Rates of the incremental addition of the segments of the landmark-based model.

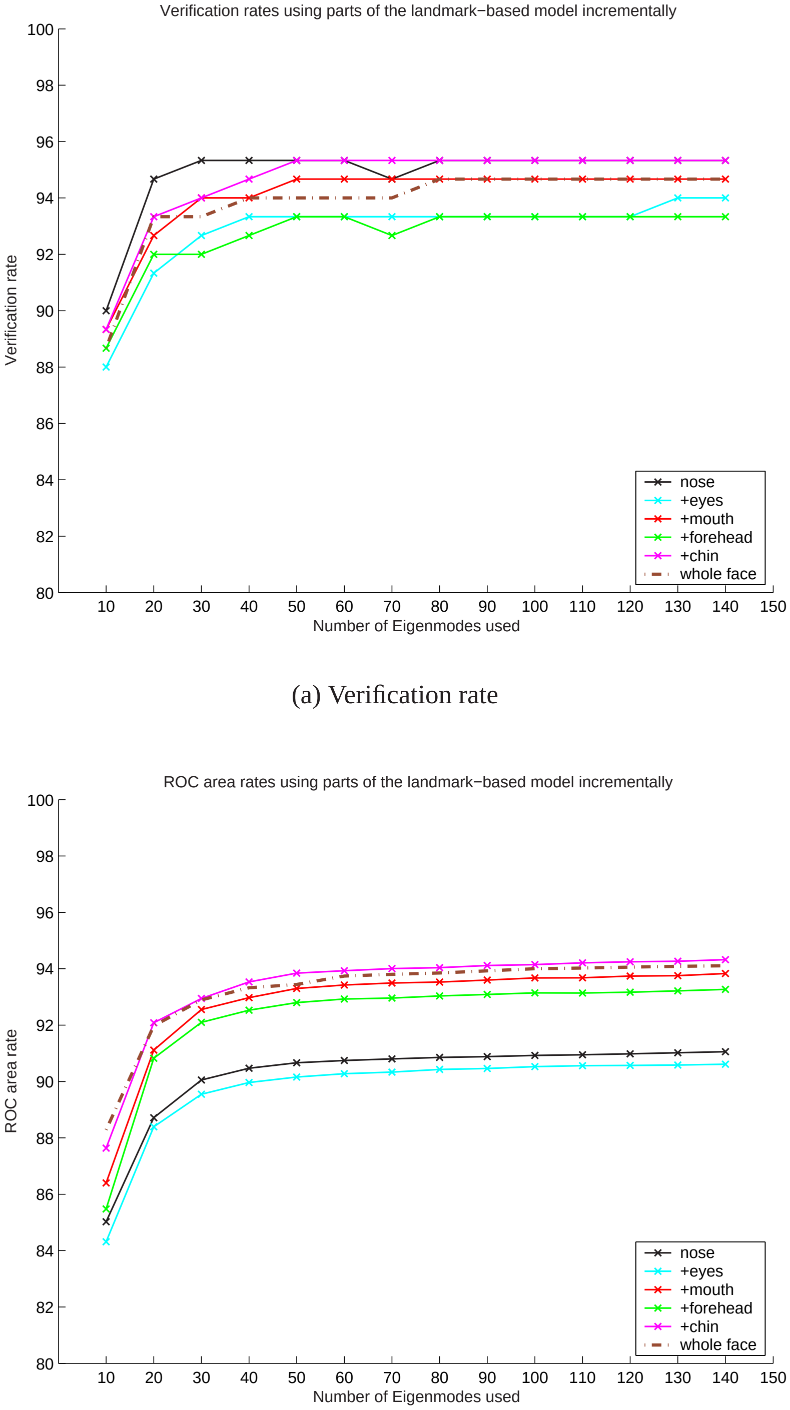


(a) Rank 1 rate



(b) $P_{FA} = P_{DI}$ rate

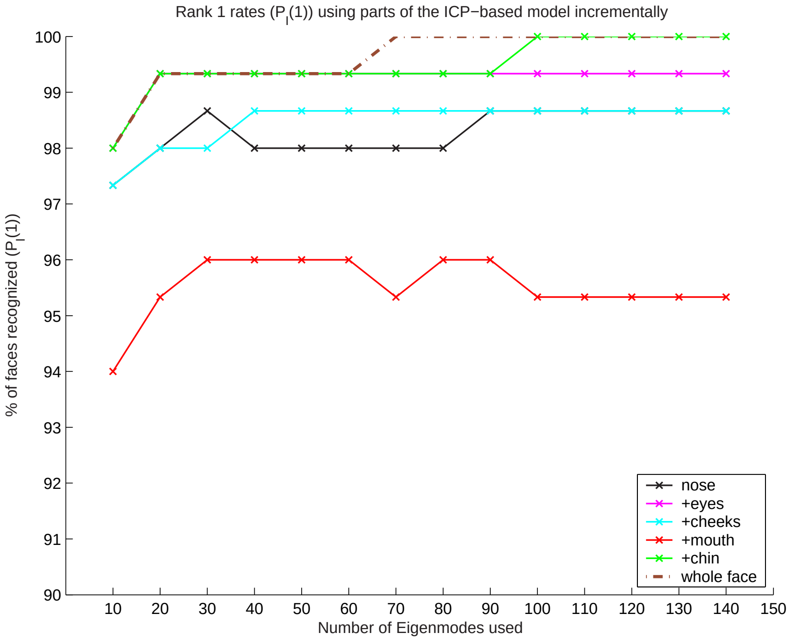
Figure 7.10: Rates of the incremental addition of the segments of the landmark-based model with the cheeks added last.



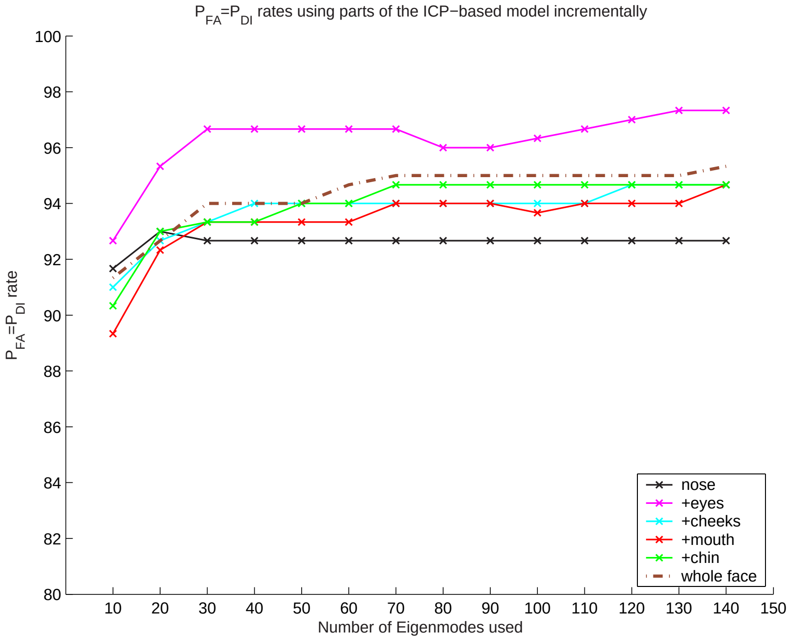
(a) Verification rate

(b) ROC rate

Figure 7.11: Rates of the incremental addition of the segments of the landmark-based model with the cheeks added last.

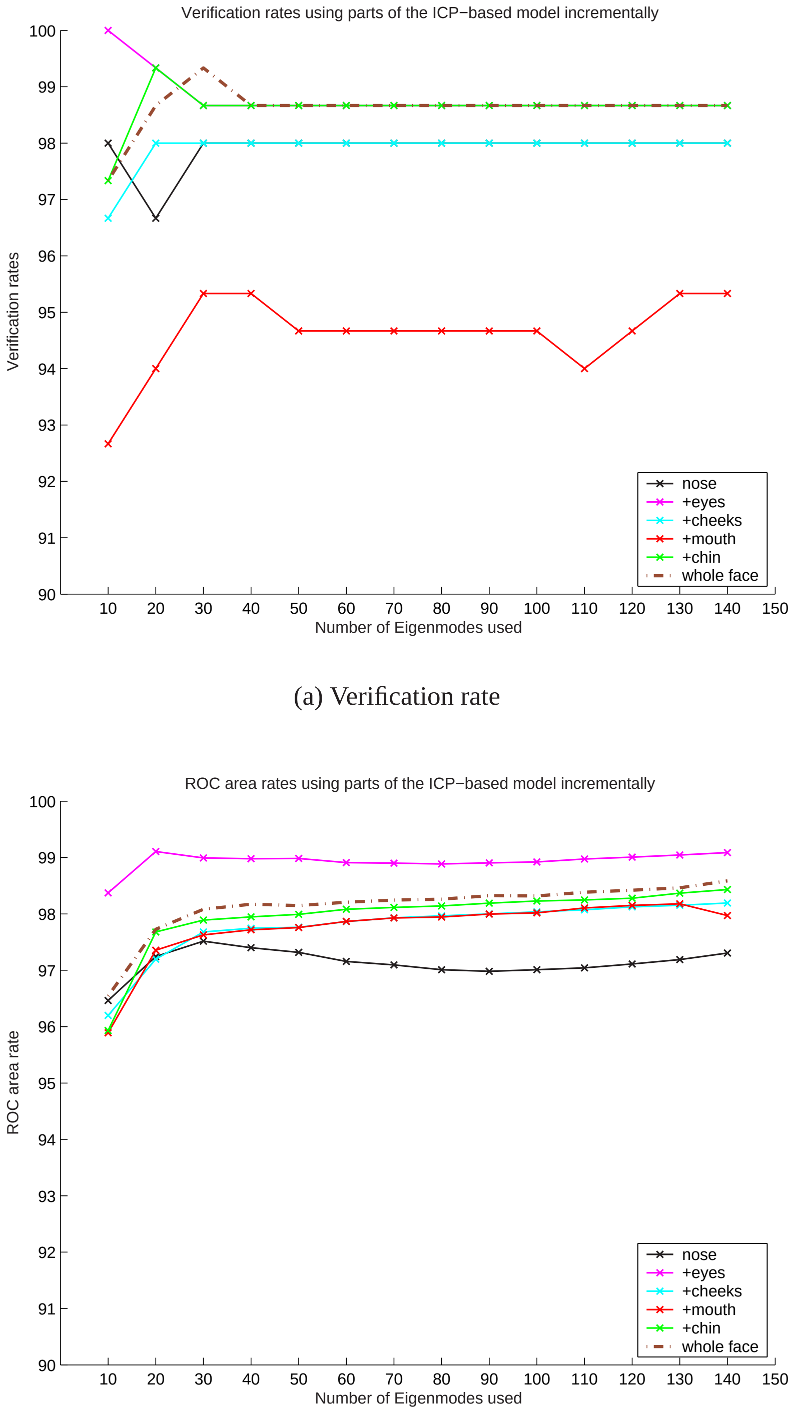


(a) Rank 1 rate



(b) $P_{FA} = P_{DI}$ rate

Figure 7.12: Rates of the incremental addition of the segments of the ICP-based model.



(a) Verification rate

(b) ROC rate

Figure 7.13: Rates of the incremental addition of the segments of the ICP-based model.

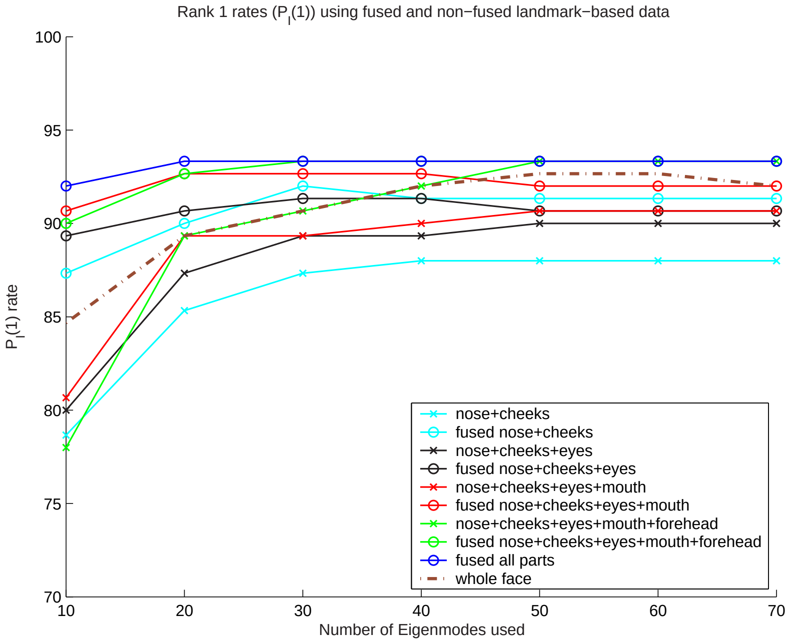
means that rather than using the score of a combined model, the score of each model is computed separately after nearest neighbor search is performed within each space. A simple approach is presented in Ross and Jain [171] where the similarity scores from different classifiers are combined based on the sum rule. The sum rule is a weighted average of the scores from the multiple modalities.

We found that fusing the scores of the different parts, rather than joining their vertices into a single model, significantly improved the recognition rates. To produce these fused scores, equal weights were used for all classifiers, as finding the ideal balance between the feature scores was not the objective. Figures 7.14 to 7.17 show some experiments in which classification scores were fused. The scores are not displayed after 70 model parameters because they do not change significantly. Using the fused scores of the eigenfeatures returns significantly higher classification rates than joining the vertices of the segments into a model. It is worth noting that the recognition rates increase when scores of more eigenfeatures are fused together. This results in the fused scores of all segments performing significantly better than when they are part of the same statistical face model (whole face).

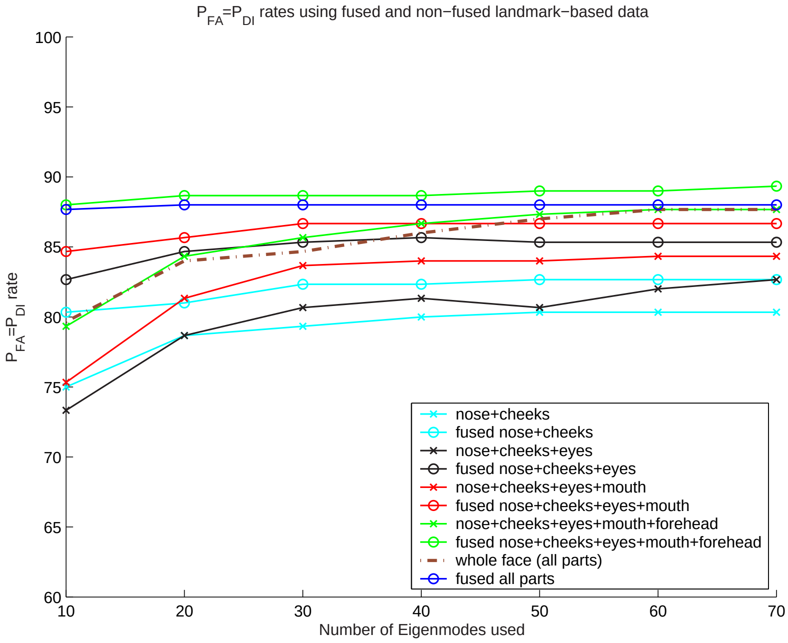
A similar pattern is also observed in Figures 7.16 and 7.17 showing the recognition rates of the fused segments of the ICP-based model. Once again using the fused scores of the individual segments yields significantly higher rates managing to reach 100% rank 1 rates across all parameters, including the case when only ten parameters are used.

7.3 Automatic model optimization

In the previous section, a manual segmentation of the face was used and the individual regions were examined in order to identify the one that contributes most to high recognition rates. This, however, involves an arbitrary choice of regions which a human observer considers as distinct and separate from other regions. One problem with this approach is that all segmented regions of the face contain some points that assist the correct identification of a person and other points that impede it.

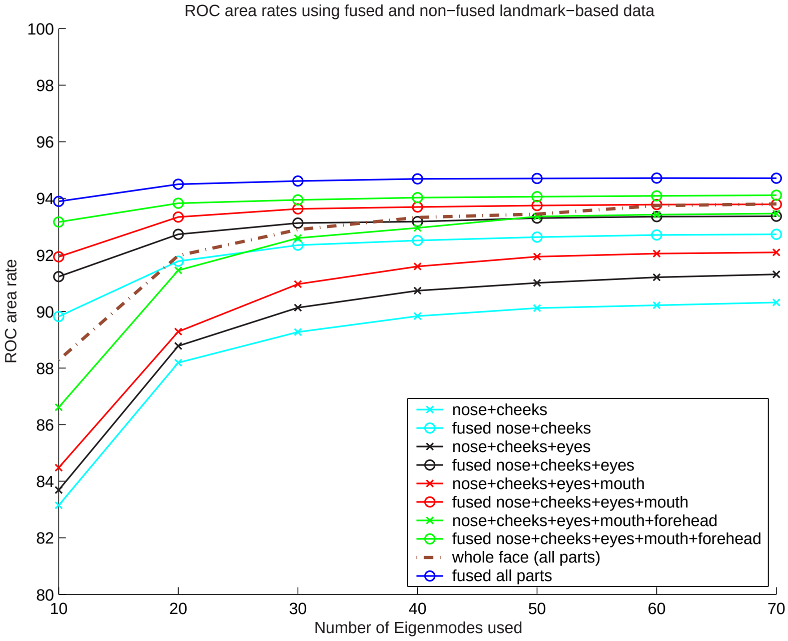


(a)

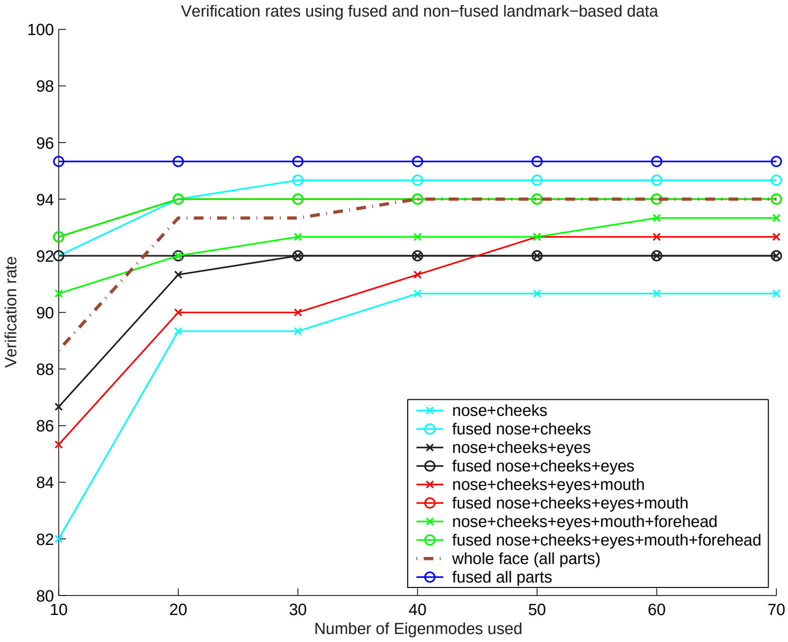


(b)

Figure 7.14: Comparing the recognition rates of features generated with the landmark-based approach and their vertices joined together in the model-building stage against the rates of features which have their individual scores fused after classification.

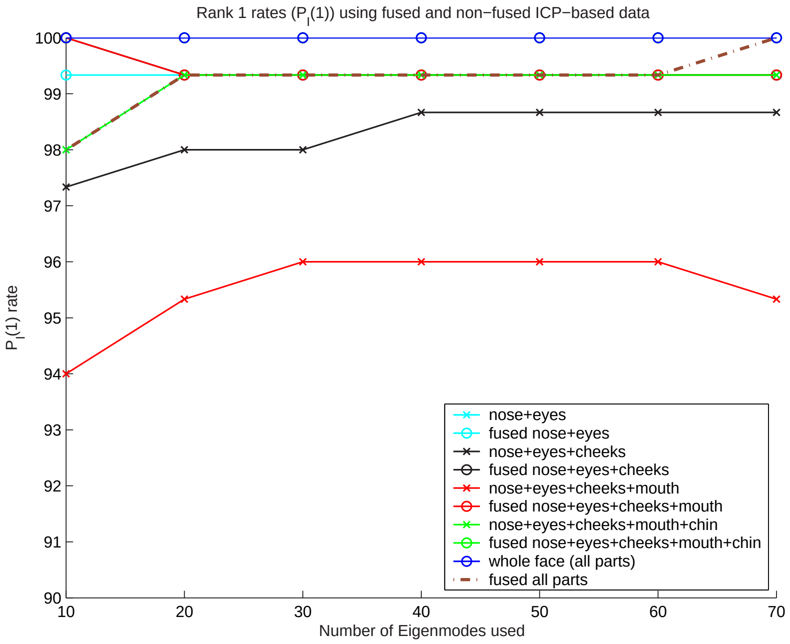


(a)

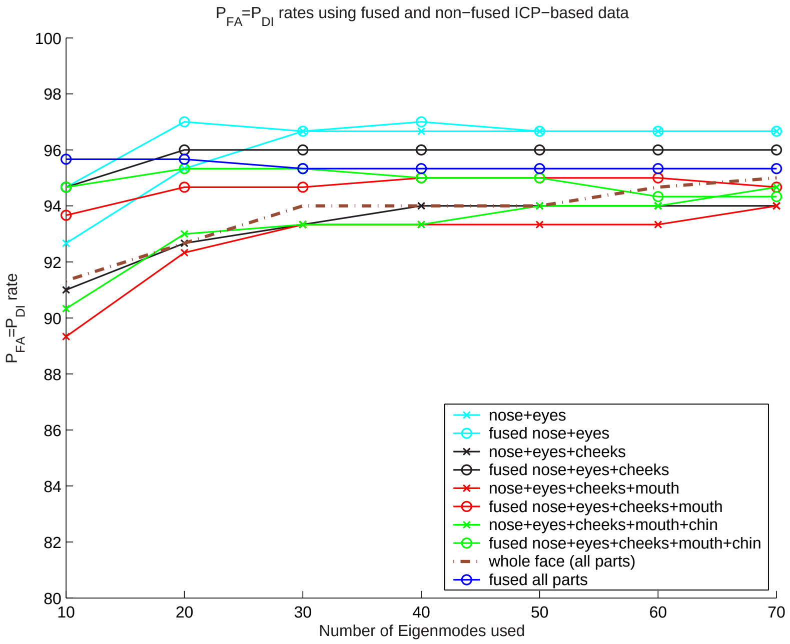


(b)

Figure 7.15: Comparing the recognition rates of features generated with the landmark-based approach and their vertices joined together in the model-building stage against the rates of features which have their individual scores fused after classification.

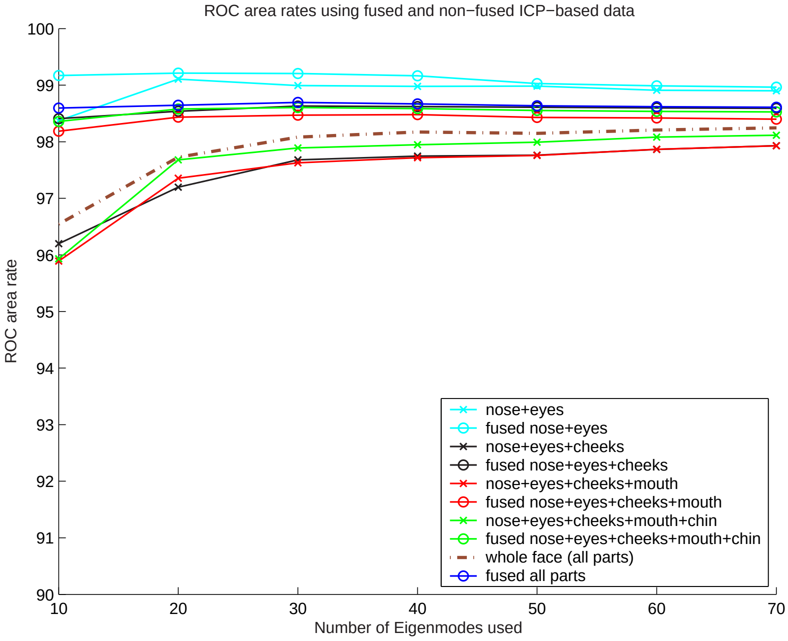


(a)

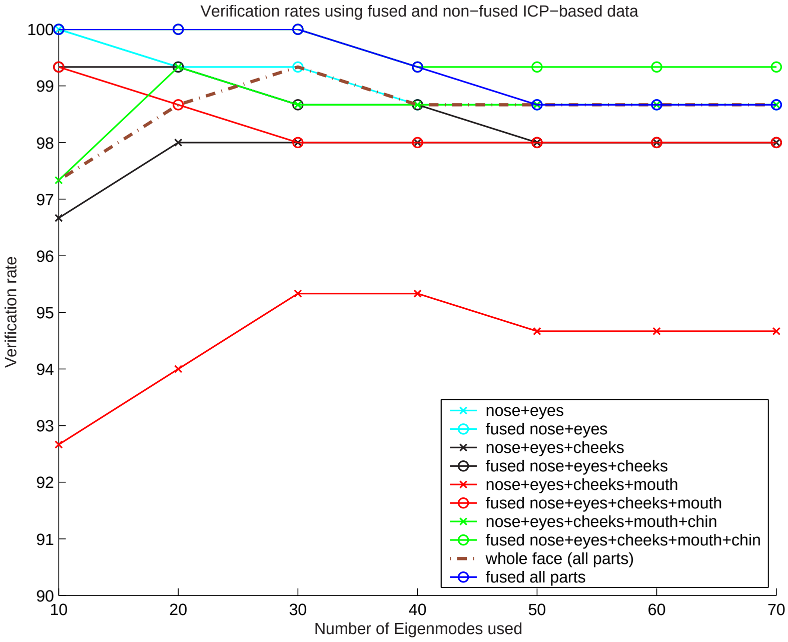


(b)

Figure 7.16: Comparing the recognition rates of features generated with the ICP-based approach and their vertices joined together in the model-building stage against the rates of features which have their individual scores fused after classification.



(a)



(b)

Figure 7.17: Comparing the recognition rates of features generated with the ICP-based approach and their vertices joined together in the model-building stage against the rates of features which have their individual scores fused after classification.

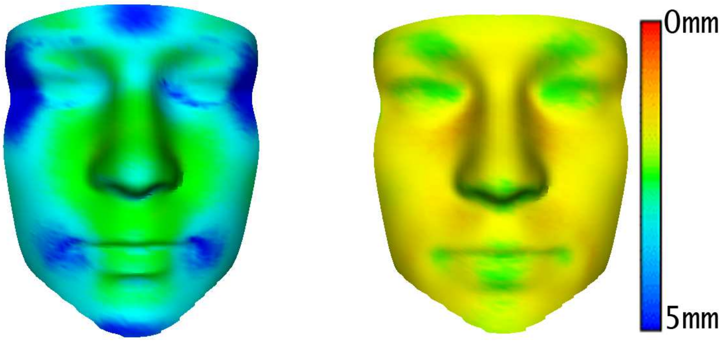
7.3.1 Statistical variability

Chapter 1 briefly discussed that a face that has a prominent characteristics (large nose, big lips) is more easily remembered by humans than an average face. Theorists postulate that the more distant a face is from a hypothetical “mental” mean the more easily it can be distinguished and identified.

Similarly, when it comes to machine face recognition, the greater the ratio of the within subjects variability (due to pose, facial expression etc.) and the between subjects variability (due to identity), the harder face recognition becomes. In other words, one can hypothesize that as the ratio of the within and between class variability decreases, the recognition rates improve.

7.3.1.1 Within-subject variability

With the correspondence between the points of the surfaces established, one can compute those points that vary the most within the subject class. To calculate this variability, 150 gallery faces of the Notre Dame database were paired up with their corresponding 150 probe faces. Notice how in this part the probe faces are actually used in the model building stage to optimize the model. This is in generally not the right thing to do because you train your system to the specific population. Ideally, you want a completely separate probe set from a gallery set. However, in this case, given that we only have two biometric samples per subject, it was impossible to implement the technique that follows without violating the aforementioned rule. In earlier chapters, the reason we did not use LDA was precisely because we did not want to use both probe and gallery set into the same model. If we had done so, the results might have been impressive but all biometric samples would have been seen by the model and any results would have been misleading. The point variability was calculated by finding the average distance between corresponding points of datasets belonging to the same subject. Let $\boldsymbol{g}_i$ be a gallery point set while $\boldsymbol{p}_i$ be a probe point set belonging to the same subject with index value i. If a single point on the gallery point set



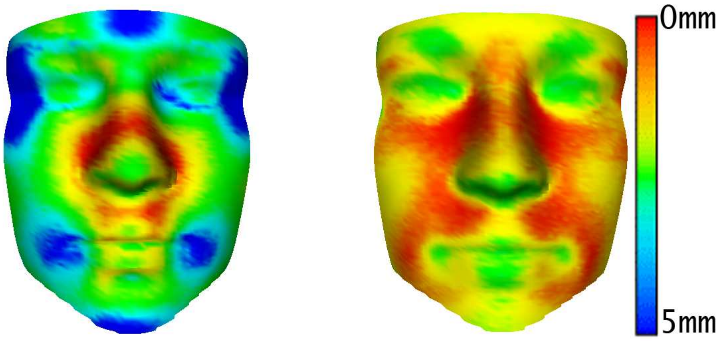
(a) landmark-based (b) ICP-based

Figure 7.18: Within-class variability of the (a) landmark-based and (b) ICP-based model

is defined as $\boldsymbol{g}_{ij}$ then the within-class variability $\boldsymbol{w}_j$ for each point is calculated by:

$$\boldsymbol{w}_j = \sqrt{\frac{\sum_{i=1}^{N}(\boldsymbol{g}_{ij} - \boldsymbol{p}'_{ij})^2}{N}} \tag{7.1}$$

where N is the number of datasets in the gallery set that were used to calculate the within subject variability. Figure 7.18 shows the within-class variability of the two types of statistical face models. Colors toward blue indicate that a certain point varies a lot within subjects while colors closer to red indicate that the point is relatively invariable within pairs of datasets from the same subject. The color scale is the same for both models. Figure 7.19 shows the same variability but this time the scalar values on each face have been normalized in order to maximize the spread of the variability and to highlight the areas on either side of the extreme. Notice how the areas around the landmarks vary the most (Figure 7.18(a) and 7.19(a)). This is an expected outcome, given the way the model is build according to the landmark-based technique. The latter uses 13 fiducial points to rigidly register the surfaces. The step after this rigid alignment is a non-rigid landmark registration which further aligns anatomical areas to each other in order to establish point correspondence. The multilevel B-spline registration that is used to achieve that (see Chapter 5) deforms the face in order to register these 13 fiducial points perfectly. The areas that deform most, given the local emphasis of the multilevel B-splines, are the areas surrounding these landmarks. However, since these landmarks are manually selected,



(a) landmark-based normalized

(b) ICP-based normalized

Figure 7.19: Normalized within-class variability of the (a) landmark-based and (b) ICP-based model.

they are prone to error. Naturally, this error is greatest in the areas of maximum non-rigid distortion. For this reason points around the landmarks in particular, vary significantly within-subjects.

To test this hypothesis the within-subject variability of the landmarks after rigid alignment was calculated. The first column in Table 7.4 shows the variability in mm between the pairs of landmark sets belonging to the same subject. If the surfaces have no artifacts, are of very high resolution and perfectly landmarked, then the within-class variability of the landmarks would be close to zero. This, however, is not the case as some landmarks vary significantly within-subjects. The second column in the same table shows the ranking order of the most variable landmark point (rank 1) to the least variable one (rank 13). Notice the correlation between the color mapping of the within-subjects variability of the whole face and the ranking order of the landmarks (Figure 7.20). In general, the greater the average error of a landmark, the greater the within class variability of the surface points surrounding that landmark.

The eye landmarks, the ones on the left and on the right of the mouth, the chin and the glabella ones vary the most and that explains the variability distribution in Figure 7.18(a) and 7.19(a). Compared to the rest of the landmarks, these areas are more problematic to landmark for a number of reasons. First of all, the glabella and the chin do not have very clearly identifiable characteristics. Locating the landmark can be quite arbitrary. The left and right of the mouth is also challenging to mark because there is not such a sharp surface difference between lip and non-lip tissue. Furthermore, the poor quality of the texture data in the Notre Dame datasets makes it particularly difficult to landmark. Finally, given that there are often holes around the areas of low-reflectance and high curvature such as near the eyes, it is often impossible to place a landmark in the “correct” place and a landmark is instead placed in the nearest possible surface point. Contrary to the aforementioned, placing landmarks on the tip of the nose, nasion or subnasal is easier as these areas have distinctive geometrical characteristics.

However, the error distribution in the landmark sets is not just due to errors in locating the “correct” fiducial point. Fitzpatrick and West [62] showed that after the rigid registration the registration error will tend to be smaller, on average, at landmarks close to the centroid of the landmark configuration and its value increases as the distance of the point from the principal axes of the landmark configuration increases. It is for this reason that the “outer” landmarks vary relatively more than the more central ones (nose, nasion, subnasal, etc.) and that the error is greater towards the edges of the face and smaller towards the center (Figure 7.20). The non-rigid registration that follows the rigid one aligns the landmark sets almost perfectly to each other. It is therefore expected that the areas that will be distorted the most are the ones that have greater registration error after rigid registration (i.e. the points that are far from the centroid).

Figure 7.18(b) shows the variability for the ICP-based datasets. Once again, the extremes of the distribution are highlighted with the variability distribution normalized in Figure 7.19(b). In the case of the ICP-based model, the faces are rigidly registered to the template face using the whole surface and not just a manually selected group of landmark points. For this reason, the blue areas have disappeared and the areas with most within-class variability are areas such as the eyes, mouth, eyebrows that tend to contain errors during the data capture as well as the subnasal area which tends to be occluded by the tip of the nose. The within-class variability is insignificant in this case compared to the variability of the landmark-based model.

Landmark variability (mm)		
Landmark	**within-class**	
	variability	rank
left of left eye	2.49	1
right of left eye	2.14	8
glabella	2.21	6
left of right eye	2.39	2
right of right eye	2.31	4
nasion	1.75	9
nose tip	1.63	10
subnasal	1.61	11
left of mouth	2.30	5
top of mouth	1.57	13
right of mouth	2.32	3
bottom of mouth	1.58	12
chin	2.19	7

Table 7.4: Within-class variability of the landmark points.

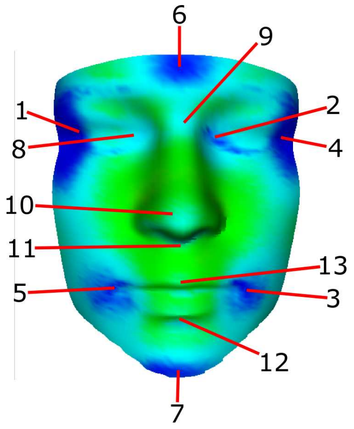


Figure 7.20: The rank of the most within-class variable landmark points.

7.3.1.2 Between-subject variability

In order to calculate the between-subject variability, each of the 150 faces of the facespace was compared to the mean face and the average digression of each point from the mean point position was computed. Let $\boldsymbol{g}_i$ be a gallery point set while $\overline{\boldsymbol{g}}$ be the mean gallery point set. If a single point on the gallery point set is defined as $\boldsymbol{g}_{ij}$ then the within-class variability $\boldsymbol{b}_j$ for each point is calculated by:

$$\boldsymbol{b}_j = \sqrt{\frac{\sum_{i=1}^{N}(\boldsymbol{g}_{ij} - \overline{\boldsymbol{g}}_j)^2}{N}} \tag{7.2}$$

where N is the number of datasets in the gallery set.

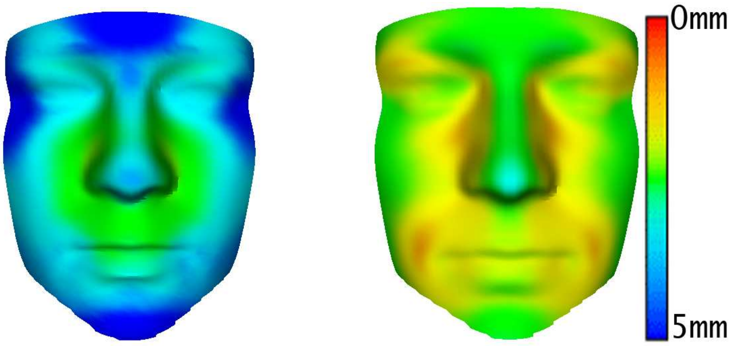
Figure 7.21 shows the between class variability of the point sets in absolute values while 7.22 shows the normalized values to utilize the full color range and to emphasize the differences between the areas. The color values in the between-class variability image can not be directly compared to the values in the within-class variability (Figure 7.18), because the range of the lookup table of the color mapping has been adjusted to produce more meaningful rendering of the values. It is evident that the landmark-based model has more between-class variability than the ICP-based model. The latter provides a closer overall fit for all the point sets, which corroborates the findings in Chapter 5.

Just as with the within-subject variability, the between-subject is also affected by the nature of error distribution on the face after rigid registration as presented in [62]. In this case it is even more evident how points further from the landmark centroid have greater registration error than landmarks near the centroid. Table 7.5 and Figure 7.23 clearly show that the outer landmarks vary the most while the ones closer to the center of the face like the subnasal and the mouth fiducial points vary very little, comparatively speaking. In other words, the outer landmarks (further from the landmark centroid) become more explicit encoders of head size. Furthermore, the surface points surrounding the landmarks far from the landmark centroid, like the glabella, left of the left eye, right of the right eye and the chin have maximum error and are thus distorted the most when the non-rigid registration is performed (Figure 7.21(a)).

This kind of variability distribution is not seen in the between-class variability of the ICP-based model because landmarks are not used (Figure 7.21(b)). It can be seen that, with the ICP-based model, there are areas toward the center of the face with small variation while areas near the borders, which are more susceptible to face size differences, vary the most. In both models, in other words, there is little between-subject variability toward the center of the face and significantly more in the edges. A sole exception to this is the nose, which is a part of the face that differs significantly from face to face even after registration. Despite being a central feature, it is evident from the between-class variability distributions of both models that the nose tip varies significantly from subject to subject. This might explain why the nose was a particularly good discriminant factor in the experiments of Section 7.2.1.

7.3.1.3 Face recognition using an optimized facespace

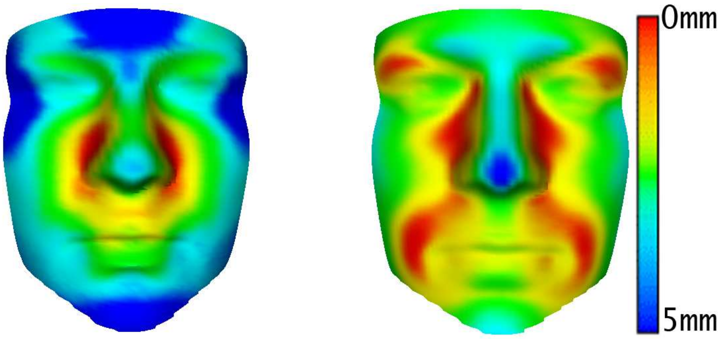
By using analysis of the point variability discussed above, one can decide which facial points to include in the statistical model. The points with the greatest within-subject variability are removed and the recognition rates for the new model with the remainder of the points are recalculated. In order to test this hypothesis, various increments of the most varying points were removed and the model was reconstructed. Figure 7.24



(a) landmark-based

(b) ICP-based

Figure 7.21: Between-class variability of the (a) landmark-based, (b) ICP-based statistical face models.



(a) landmark-based normalized

(b) ICP-based

Figure 7.22: Normalized between-class variability of the (a) landmark-based and (b) ICP-based model.

Landmark variability (mm)		
Landmark	**between-class**	
	variability	rank
left of left eye	4.24	3
right of left eye	3.73	7
glabella	4.72	2
left of right eye	3.76	6
right of right eye	4.03	4
nasion	3.83	5
nose tip	3.70	8
subnasal	2.94	12
left of mouth	3.55	10
top of mouth	2.63	13
right of mouth	3.66	9
bottom of mouth	3.22	11
chin	5.07	1

Table 7.5: Between-class variability of the landmark points.

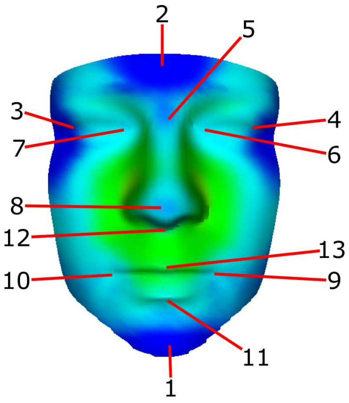
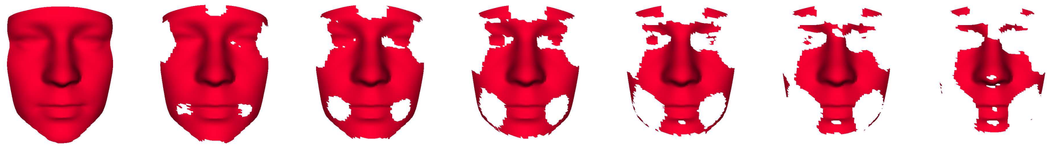


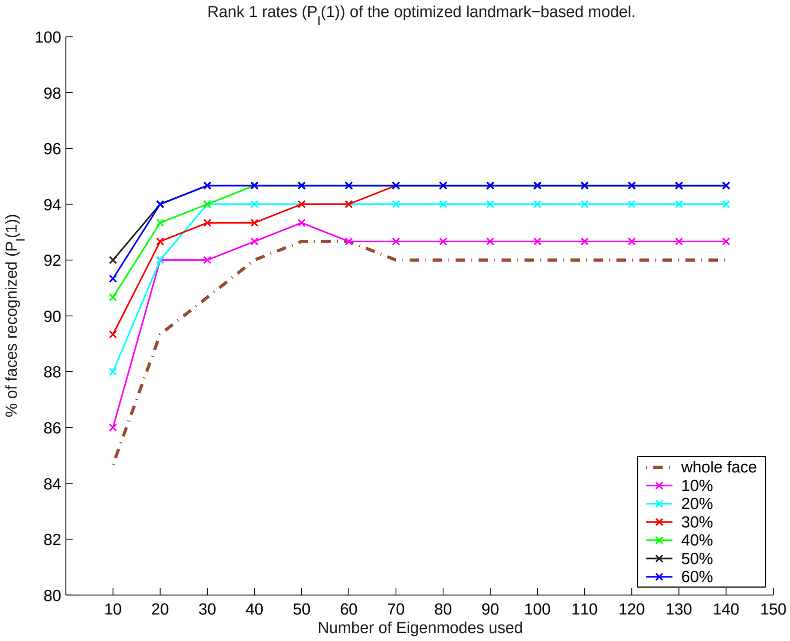
Figure 7.23: The rank of the most between-class variable landmark points.



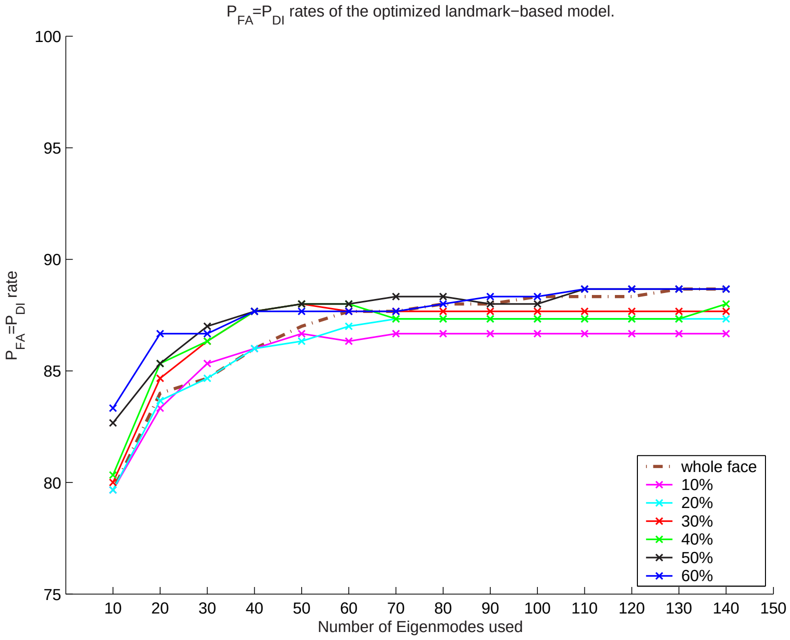
(a) 0% (b) 10% (c) 20% (d) 30% (e) 40% (f) 50% (g) 60%

Figure 7.24: Removing the most within-class varying points.

shows the models that result from the removal of the $10\%, 20\%, 30\%, 40\%, 50\%$ and 60% most within-class varying points. The same experiments were conducted for both model-building methods. Figures 7.25 and 7.26 show the evaluation scores of the optimized landmark-based models. Notice how in most cases, the gradual removal of the most varying points significantly improves the rates. Even when 60% of the facial points are removed leaving the model with just over 2000 points, its classification ability is still better than all other point removal steps. As predicted, even a conservative removal of points at a level of 10% manages to improve the baseline scores where no points are removed. Finally figure 7.27 shows the same results but the x-axis shows the percentage of points removed in order to emphasize the general behaviour of the model as less of the within-variable points remain. Notice also how, in contrast to the other figure, this figure shows the results when removing 70% and 80% of the most variable points. Figures 7.28 and 7.29 show the evaluation scores of the various optimized ICP-based models. In this case, removing the most within-class variable points does not significantly improve the recognition rates. Within a range of model parameters removing 10% of the most variable points improves the $P_{FA} = P_{DI}$ and verification rates, but at other times, using the whole face without removing points performs better. This is somewhat understandable if one considers two characteristics of the ICP-based model. Firstly, the recognition rates where already quite high and it can get increasingly difficult to improve on these. Secondly, and perhaps more importantly, the within-subject variability of the ICP-based model as displayed in Figure 7.21 is not great and there are no extrema as in the landmark-based model where a removal of the points of those areas would clearly improve the results. As a result, there is no clear advantage of removing the most within-class variable points of

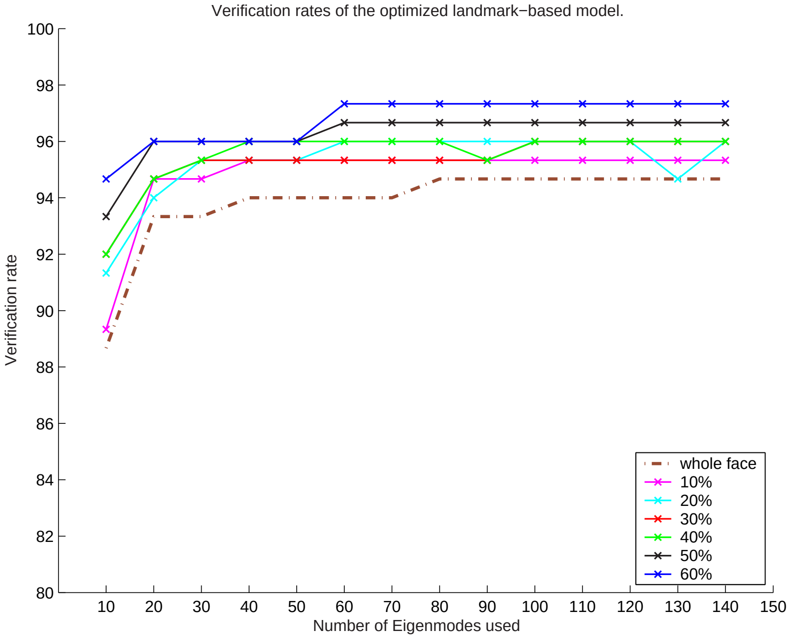


(a) Rank 1 rate

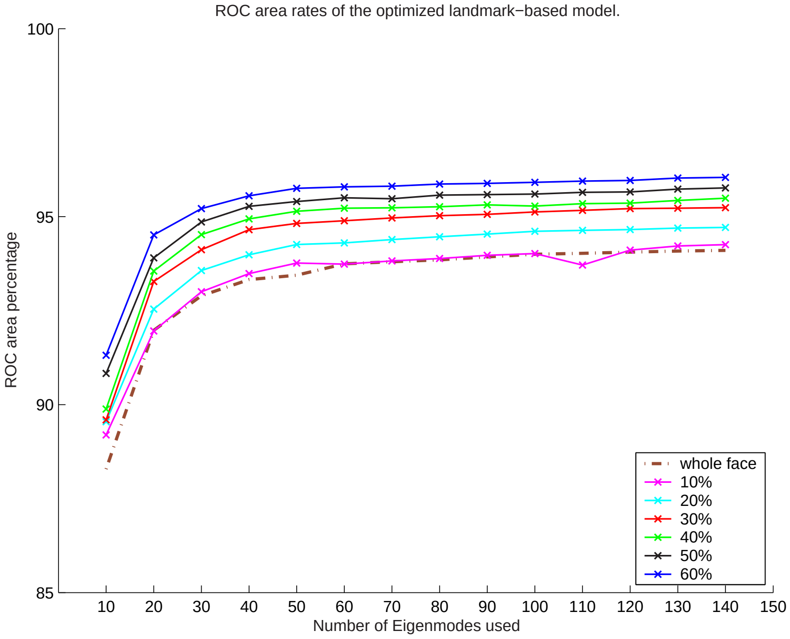


(b) $P_{FA} = P_{DI}$ rate

Figure 7.25: The recognition rates resulting from the gradual removal of the most within-class varying points from the landmark-based model.

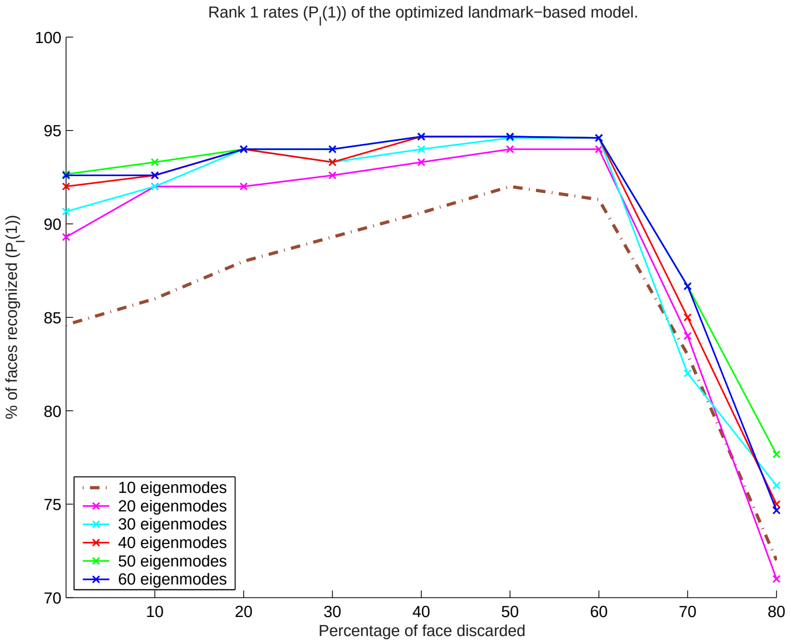


(a) Verification rate

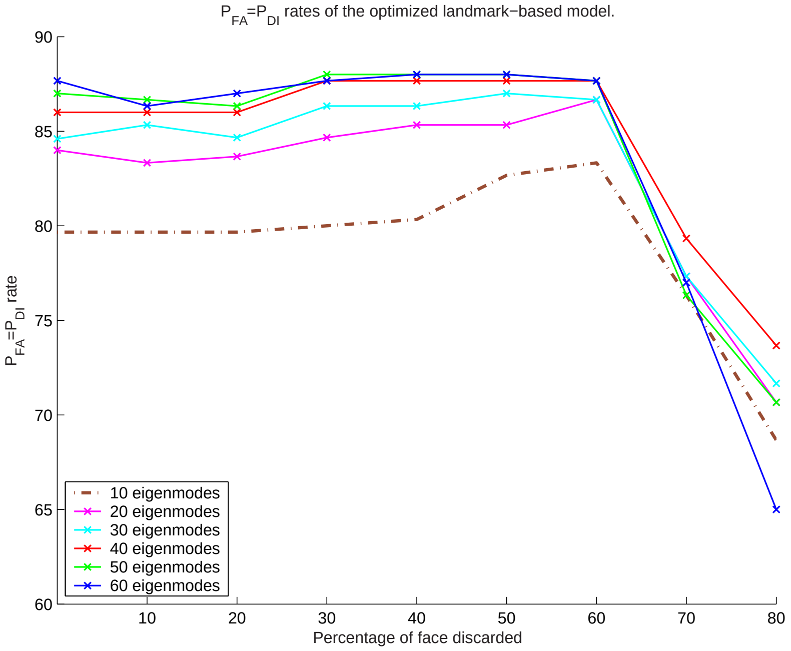


(b) ROC rate

Figure 7.26: The recognition rates resulting from the gradual removal of the most within-class varying points from the landmark-based model.



(a) Rank 1 rate



(b) $P_{FA} = P_{DI}$ rate

Figure 7.27: The recognition rates resulting from the gradual removal of the most within-class varying points from the landmark-based model. Notice how the x-axis has the percentages of points removed in order to demonstrate more clearly the general pattern of behaviour of the model.

the ICP-based model.

7.3.1.4 Discussion

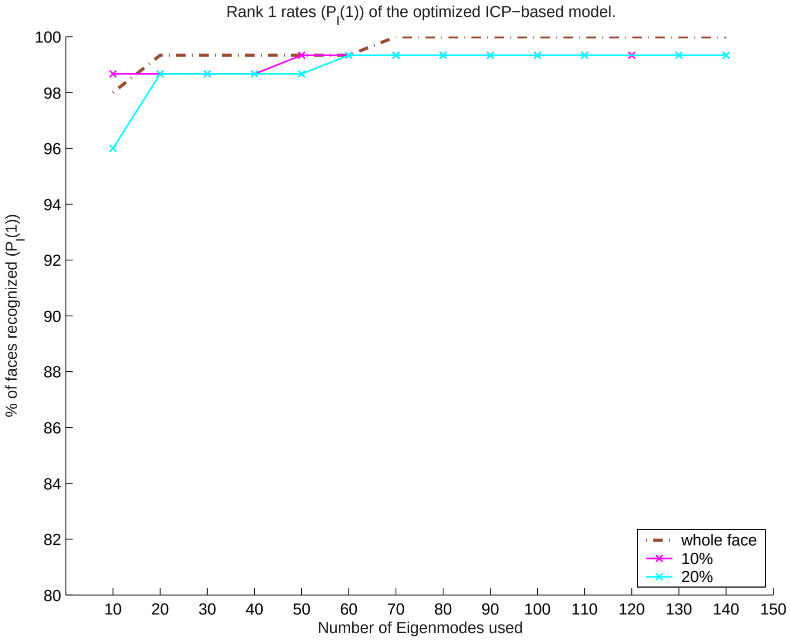
The discriminatory power of individual features is surprisingly high, opening up possibilities for the use of 3D eigenfeatures as classifiers. As it is demonstrated, the relationship between individual features is different from technique to technique and this strongly suggests that the findings are model-specific and in no way do they reflect the behavior of any statistical face model. Specific conclusions drawn from one model do not necessarily apply to another. Nevertheless, some observations, like the discriminatory power of the nose, might be true for other models too.

Another observation in this chapter is that different combinations of features improved the recognition results compared to using the entire face. Only a couple of combinations were presented in this chapter. If one segments the face in n regions, there are $2^n - 1$ combinations of facial features. In this case, since $n = 6$, there are 63 different combinations of features to assess, something which is impractical.

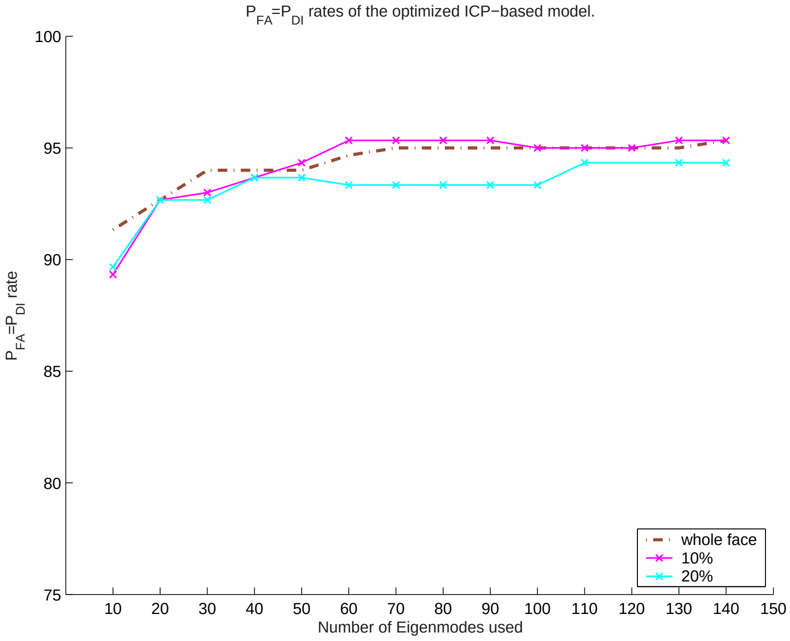
The score fusion results of the various feature-based classifiers could be further improved if a more sophisticated decision fusion algorithm than a simple sum rule is used. In this chapter, an unweighted sum was employed. An alternative would be to use a separate training set and find a weight combination between the facial features that minimizes the classification errors. A good start would be to use the within-class variability that was calculated as an indication of which areas would be better classifiers. For example, facial segments of high within-class variability would have a lower weight than segments with low within-class variability. At the same time segments with high between-class variability would be associated with a greater weight value in the fusion function than areas with low between-class variability.

Furthermore, alternative combinations of the scores of the models could be performed. For example, one could fuse the classifier scores of the whole face with the scores from the strongest features etc. Future work in this area will be discussed in Chapter 8.

Generally speaking, different facial regions produce different recognition rates be-

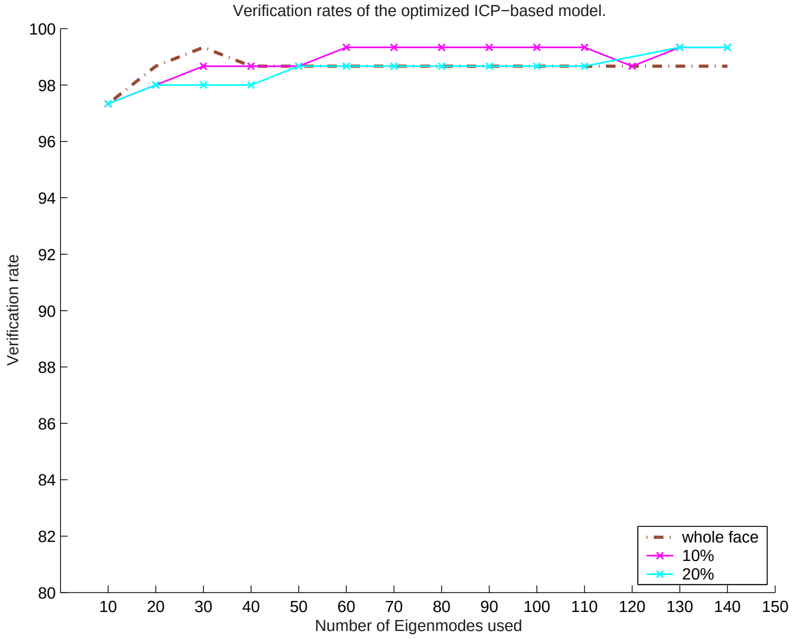


(a)

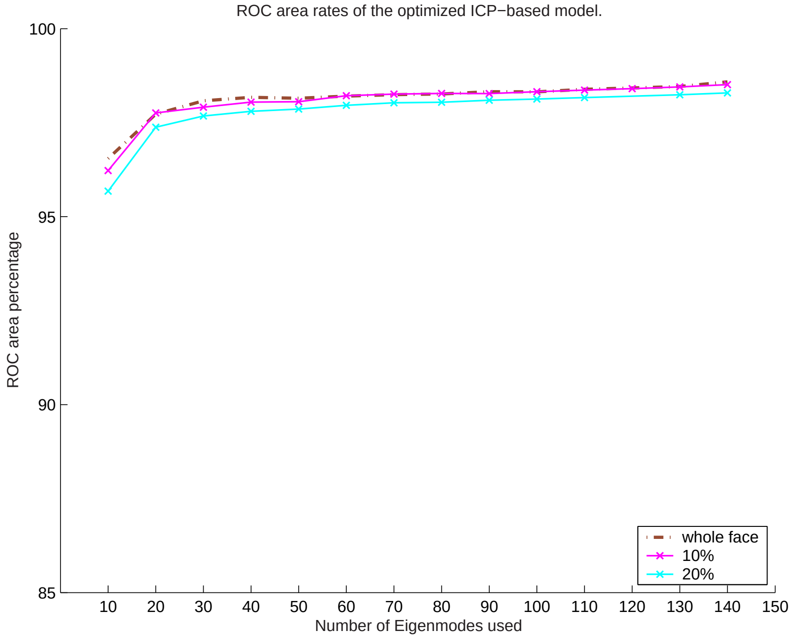


(b)

Figure 7.28: The recognition rates resulting from the gradual removal of the most within-class varying points from the ICP-based model.

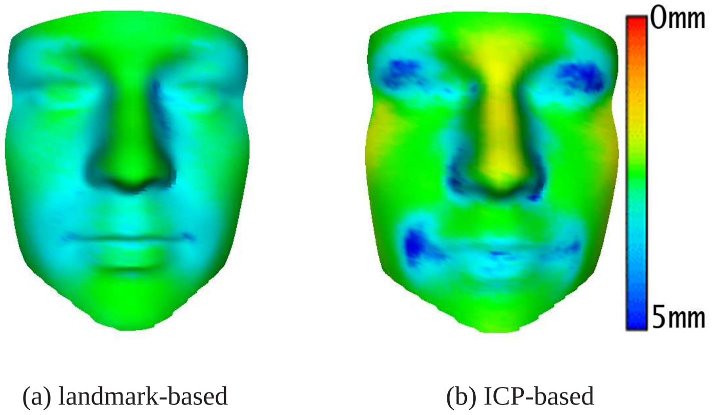


(a)



(b)

Figure 7.29: The recognition rates resulting from the gradual removal of the most within-class varying points from the ICP-based model.



(a) landmark-based (b) ICP-based

Figure 7.30: Ratio of within- and between-class variability.

cause of the variability within these areas. Some tend to vary a lot from person to person allowing for better discrimination while others vary very little. As a result, points on one facial region have different statistical properties from points on other regions. A more optimized solution to the classification problem was explored, where points on the face were selected, which would perform well at classification regardless which anatomical region they belong to. Eliminating points from a surface in order to maximize the discriminating factors has produced promising results. In this chapter, points on the face were eliminated if they were deemed to vary significantly between subjects of the same class. The variability of points between classes, however, was not explicitly used apart from providing explanations for some of the scores. Another approach would be to take both the within- and the between-class variability into account and try to keep points that vary most between class while eliminating the ones that vary most within class. A simple approach was tested where the between-class variability of each point was divided by the within-class. The points where the between- and within-class ratio is small are referred to as *suboptimal* and are removed. Figure 7.30 shows the scalars of the joined variability mapped on the face. Preliminary experiments using division to fuse the within- and between-class variability have not yielded significant improvements in the recognition rates. More specifically removing 10% of the suboptimal points did not allow the rank 1 rate to get over 91% which is lower than using the baseline method which uses the whole face. In contrast removing 10% of the most within-class variable points allows the rank 1 rate to reach 93%. When 20% of the suboptimal points are removed the rank 1 rate drops even further reaching only 89%. Results like these led us to abandon this route. However, other approaches to combine the within- and between-class variability might prove more successful. Generally speaking, an objective function like in Davies [51], enables one to assess which combinations of eigenfeatures to use or which points to eliminate. The objective function allows for the parameterization of the combinations and their evaluation in the search for the optimal statistical model.

Studying the within- and between-class variability of the population could also be used in other types of classification. For example, in the method presented in Chapter 4 where the squared 3D distance between corresponding points is used as a similarity metric. In that case, one could assign weights to each point in the population based on the within-class variability. When calculating the squared 3D distance, points of low within-class variability would contribute more than points with high within-class variability.

Finally, the variabilities calculated in the last section were calculated using the probe and gallery sets that were subsequently used for classification. Ideally, an unseen set of images should be used to calculate the variability before applying the findings on the set of data. Given the small number of datasets available for this study, however, the variability was calculated using the same sets that were used for classification. Nevertheless, the principle behind the optimization of the statistical face model still applies.

Chapter 8

Conclusions

In this work we developed various 3D face recognition techniques based on rigid and non-rigid registration. Moreover, we showed which types of registration can lead to better point correspondence across faces and thus to better models. Using these models, we also explored an eigenfeatures implementation as well as a subspace optimization. We have used various metrics in face recognition to perform a task-specific evaluation of the 3D face models. We have also used metrics such as specificity, generalization ability and compactness to characterize the models. The techniques developed here could have application to other biometric modalities which yield 3D data, such as ears, noses, hands etc. Furthermore, these techniques could assist in general object recognition problems and not just biometrics.

8.1 Summary of contributions

In Chapter 4 we demonstrated the wealth of information that exists in 3D facial data by introducing a fully automatic face recognition technique in which facial surfaces are registered to each other using the ICP algorithm. The remaining 3D square difference between the registered faces was used as a similarity metric, yielding rank 1 rates of up to 100% using frontal images. The proposed method produced encouraging results for faces that were substantially different from frontal neutral faces, such as faces with expression or an extreme pose. Using automatic surface registration compensates for these differences and yields good recognition rates. In the same chapter we investigated the effects of pose, facial expression and illumination on a similarity metric that fuses the 3D and 2D information. The recognition rate using this metric on non-frontal images falls while on faces with expression it increases. Similar surface-based approaches managed to reach 84% using a smaller database of only 10 subjects with a variety of pose and expression [127]. Medioni *et al.*[135] used a database of 100 subjects to test a approach similar to ours achieving similar results but only tested on neutral images and profiles of up to 20°, which is smaller than our profiles of 45°. Furthermore, they did not make an attempt to use texture intensity during registration and classification. In general the technique in Chapter 4 performed better than all surface-based techniques we reviewed. The few techniques that managed to reach as high recognition rates as our reported results were obtained using smaller databases.

In Chapter 5, we presented two model-building techniques for subspace analysis. Based on the assumption that better correspondences between facial points lead to better recognition rates, we proposed a semi-automatic landmark-based method and a fully automatic surface registration-based (ICP-based) one. Both approaches reduced surface artifacts and produced surfaces that have the same number of points, which can then be used in PCA-based techniques.

Subsequently, we demonstrated that the ICP-based model provides a better correspondence than a landmark-based technique and as a result, yields higher recognition rates across all measures. Furthermore, we showed that the ICP-based model is more compact, generalizes better on unseen instances and is more specific than the landmark-based one. The recognition rates with both techniques were better than all PCA-based techniques we reviewed in Chapter 5. The sole exception is a technique which uses PCA both on the shape and the texture of the faces such as the one by Bronstein *et al.* [27] which was also tested on faces with expressions. However, the latter work was tried on a small database and thus the differences might be insignificant.

In Chapter 6 the shortcomings of the models were discussed and addressed with the inclusion of an automatic non-rigid surface registration to improve the point correspondence. We demonstrated that this non-rigid registration improves the recognition rates significantly. Finally, we used a synthetic uniform surface (a sphere) to regularize each surface before it is used for building the model, something which further improved the classification rates.

In Chapter 7 we introduced a novel 3D eigenfeatures technique. We performed PCA on segmented facial features creating a separate featurespace for each anatomical region. Using these eigenfeatures we established which anatomical regions perform better than others in classification tasks and we discussed why that would be the case. Combining these regions in various combinations into a unified model allowed us to achieve better recognition rates when the entire (unsegmented) face was used. An alternative solution was proposed which involved the fusing together of the classification scores of each featurespace, managing to reach significantly higher rates than combining the eigenfeatures into a single model. Moreover, by manipulating the ratio of the within- and between-class variability lead to more powerful models. Using variability information we showed how removing points that vary significantly within-subjects can help improve the classification. Even when the model is reduced by 60% in size it still yields higher rates than when using the entire surface.

8.2 Limitations and future work

8.2.1 Using a larger and improved database

The experimental results reported in this work were generated using 150 subjects at most. As discussed in Chapter 2 the results can vary significantly depending on the size of the database used. In order to increase the validity of our results it would be desirable to replicate the experiments on a larger dataset. The XM2VTS database [144] has about 300 3D datasets, which is still fairly limited. Researchers at Arizona State University are in the process of acquiring about 1500 face scans to be used for 3D face authentication [207].

The scans look to be of the same or higher resolution as the ones used in this study.

Great demographic variety would also be a desirable trait for a database. People from various ethnic backgrounds and of both sexes should be represented in adequate proportions and ideally, the 3D images should be taken at repeated intervals of time. The latter would also allow one to model the growth trajectories of the human face and try to develop models that are more insensitive to facial changes as a result of growth or aging.

A bigger database which contains more than one image per subject would also allow for increased robustness to within-subject variation as the model would encode some of the within-subject variability of the faces. Furthermore, it would allow one to apply other statistical techniques such as linear discriminant analysis which requires the facespace to contain more than one sample per class.

8.2.2 Improved testing protocol

One consequence of using a relatively small database is the fact that some testing protocols used in this thesis had to be adjusted. Many researchers build the model with a group of faces and then test the recognition rates using two probe sets: A gallery probe set $\mathcal{P}_{\mathcal{G}}$, which contains different instances of seen examples and a probe set $\mathcal{P}_{\mathcal{N}}$, which contains unseen examples. This enables one to perform open-set identification and verification tests with “true” imposters. Instead, we divided the subjects into two pools, the gallery set $\mathcal{G}$ and the probe set $\mathcal{P}$. This has a number of ramifications. The false alarm rate P_{FA} is traditionally computed by using the probe set $\mathcal{P}_{\mathcal{N}}$ to find the “best imposter” and check if the similarity score between him and the gallery dataset it matched to has a similarity score s_{ij} smaller than threshold τ. More formally:

$$P_{FA}(\tau) = \frac{|\{p_j : \max_i s_{ij} \geq \tau\}|}{|\mathcal{P}_{\mathcal{N}}|} \tag{8.1}$$

Since there is no $\mathcal{P}_{\mathcal{N}}$ set in this case, the P_{FA} is calculated by:

$$P_{FA}(\tau) = \frac{|\{p_j : \max_i s_{ij} \geq \tau \text{ and } \mathrm{id}(g_i) \neq \mathrm{id}(p_j)\}|}{|\mathcal{P}| - 1} \tag{8.2}$$

In other words, for every face in $\mathcal{P}$ we check if there is any face in $\mathcal{G}$ other than the face belonging to the same subject that would cause a false alarm based on threshold τ. Creating separate probe sets would increase the validity of the findings presented here.

In an ideal environment, where one has access to a large data pool, one would build a PCA-based model on a training set of images. One would then decide on the number of eigenvectors and type of distance metric to use based on tests on a validation set and finally, use a test set to measure the performance of the technique [24]. This would also make the algorithm more compatible to the testing procedures of the FERET and FRVT 2002 evaluations where teams are expected to use a completely new set of images during the test phase (see Chapter 2).

8.2.3 Create a texture and shape model

One of the potential advantages of new 3D face databases is that the texture may be of higher quality than in the dataset currently used. All the experiments performed in this work and the models created can easily be extended to use texture information. One could build a separate texture facespace and use this information in conjunction with 3D shape to perform classification. Alternatively the shape and texture information could be combined into one model by using a weighting to normalize the intensity values with respect to the geometry creating a texture-shape-facespace. Chapter 2 discussed various findings that demonstrate that the recognition rates can improve if the texture and shape modalities are combined into one. Other similarity metrics such as mutual information [131, 212] could also be used to assess the similarity between 2D datasets, which can then be combined with shape similarity scores.

8.2.4 3D to 2D registration

The bulk of face recognition is taking place using 2D data. 2D acquisition remains a very easy way of collecting data and in some scenarios the only possible way. The advantages offered by 3D data, however, such as the insensitivity to small posture changes is particularly attractive. One could use 3D data even when the probes are in 2D by registering the projection of the 3D model to 2D data and using the parameters of the 3D model to perform face recognition. In order to fit the 3D data one would also have to build a separate texture space on the 3D data. This has already been demonstrated with a morphable appearance model in Blanz *et al.* [18]. Additionally, using modular subspaces, as the ones presented in Chapter 7, breaks down the problem in various subtasks and allows for a more flexibility and robustness.

8.2.5 Modeling facial expressions

Facial expressions are one of the greatest sources of variability in facial data. The model building techniques presented in previous chapters could be used to explicitly model facial expressions in order to build an expression-space.

Including faces of various expressions in the facespace allows for the classification of the input faces based on their facial expression. Combining this technique with an automatic landmark technique one can go from using sets of landmarks for expression recognition to using full facial data. Assuming that the database contains various expressions for each subject, one could detect the expression of the input image and only search through the subset of images in the database containing that expression.

Modeling facial expressions could also be useful in cases when the database contains only neutral images. Since the correspondence between facial points can be established with the modeling techniques we have developed, one could generate a free-form transformation that turns a neutral face of a subject to a face with expression of the same subject (and vice versa). The transformations across subjects could then be analyzed with PCA in order to build and expression-space. This effectively creates an eigensmile, eigenfrown, eigendisgust, eigensurprise etc. These models can then be used to animate neutral faces, but more importantly these models could allow one to “neutralize” a facial expression and turn the face into a more suitable form before being submitted for identification. Figure 8.1 shows a face being manipulated by a facial expression model and turned from a

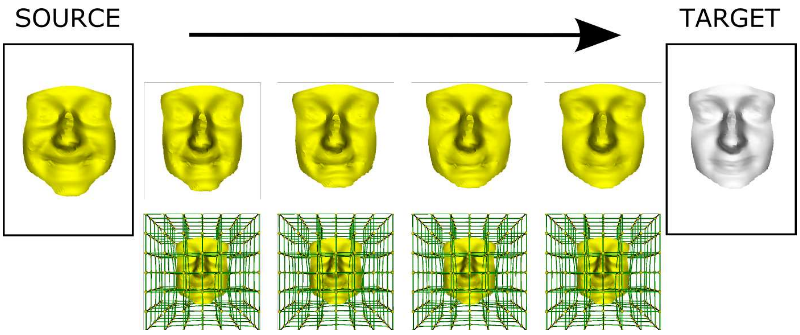


Figure 8.1: Modeling facial expressions. An expression model encodes the non rigid transformation that are needed in order to transform the source surface (smiling) to the target surface (neutral).

smiling face (source) to a neutral one (approximating the target), which would be more easily identified by a face recognition algorithm. The expression model could also allow the modelling of all the expressions between the smiling and neutral state. Yin *et al.* [222] are preparing to make a 3D facial expression database available for the research community, which would be a ideal data source to conduct this kind of work.

The VRT3D system used to capture data for some of the experiments conducted in this work is able to capture up to 30 frames/sec. This is particularly useful because one could capture not just a finalized facial expression but many intermediate steps building a much more “complete” expression-space. Given the capture speed of the VRT3D system, apart from facial expressions one could potentially model the facial movements generated when a subject is talking. Knowing the correspondence between phonemes and facial expressions, one could create models for the speech animation as well as analyze the dynamics of facial expressions.

8.2.6 Creating better correspondence

Throughout the thesis it was often stated that improving the correspondence between faces improves the recognition rates. One of the ways this can be done is by using a larger number of landmarks. Automatic landmark placement is also an option for reducing the tediousness of the process. Areas such as the cheeks and forehead, however, are difficult to landmark and furthermore, the ICP-based model uses no landmarks and outperforms the landmark-based models. A way to create better correspondence is by improving the search for the corresponding point from one surface to the other. Currently, the way this is done is by searching for the closest point in 3D space. An extension of that would be to add various attributes to each point, making the search for the closest point more robust and insensitive to data capture errors. Instead of its xyz coordinates each point could be represented by a feature vector which includes curvature information, texture intensity at that point as well as a signature calculated from the relationship between that point and its neighboring points as presented in [45].

Finally, the point correspondence across faces could be optimized as a function of the recognition rates. Using a classification task as an objective function one could manipulate the correspondence between the points to find point pairings that would maximize the classification rates.

8.2.7 Advanced score fusion

In Chapter 7, simple fusion techniques were used to demonstrate alternative ways of combining eigenfeatures of a face. More advanced methods are available for combining biometric scores. For example, the scores from each feature classifier could be concatenated into a feature vector which itself would be subjected to a second-level classifier that can form a decision boundary in the score space [99]. Alternatively, a larger database, with more than one biometric entry per person, would allow the sequential fusion of scores. This means that a score is collected from each subject's biometric entry and the scores within each subject could be summed or averaged to form a more robust score that would naturally be more immune to within-class variance.

Bibliography

[1] B. Achermann and H. Bunke. Classifying range images of human faces with hausdorff distance. In *15th International Conference on Pattern Recognition*, pages 809–813, 2000.

[2] A. Amini and J. Duncan. Differential geometry for characterizing 3D shape change. In *Proceedings of SPIE Mathematical Methods in Medical Imaging*, number 1768, pages 170–181, 1992.

[3] Wikipedia anonymous user contribution. K-D tree article, 2001-2006.

[4] A. Ansari and M. Abdel-Mottaleb. 3D face modelling using two orthogonal views and a generic face model. In *International Conference on Multimedia and Expo*, 3, pages 289–292, 2003.

[5] A. Ansari and M. Abdel-Mottaleb. 3D face modelling using two views and a generic face model with application to 3D face recognition. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 37–44, 2003.

[6] K.S. Arun, T.S. Huang, and Blostein S.D. Least squares fitting of two 3D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):698–700, 1987.

[7] R. Bartels, J. Beatty, and B. Barsky. *An Introduction to Splines for use in Computer Graphics and Geometric Modelling*. Morgan Kaufmann, Los Altos, CA, 1987.

[8] M. Bartlett, H. Lades, and T. Sejnowski. Independent component representation

for face recognition. In *Proceedings, SPIE Symposium on Electronic Imaging: Science and Technology*, pages 528–539, 1998.

[9] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[10] J.L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[11] P. Besl and R. Jain. Invariant surface characteristics for 3D object recognition in range images. In *Computer Vision, Graphics and Image Processing*, number 33, pages 33–80, 1986.

[12] P.J. Besl and N.D. McKay. A method for registration of 3D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[13] C. Beumier and M. Acheroy. Automatic 3D face authentication. *Image and Vision Computing*, 18(4):315–321, 2000.

[14] C. Beumier and M. Acheroy. Face verification from 3D and grey level clues. *Pattern Recognition Letters*, 22:1321–1329, 2001.

[15] D. Beymer and T. Poggio. Image representations for visual learning. *Science*, 272:1905–1909, 1996.

[16] I. Biederman and P. Kalocsai. *Face Recognition: From Theory to Applications*. Springer-Verlag, 1998.

[17] V. Blanz, P. Grother, J. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In *IEEE Conference on Computer Vision and Pattern recognition*, 2005.

[18] V. Blanz, S. Romdhani, and T. Vetter. Face identification across different poses and

illuminations with a 3D morphable model. In *Proceedings of the 5th Int. Conference on Automatic Face and Gesture Recognition*, pages 202–207, 2002.

[19] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999.

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

[21] W. Bledsoe. The model method in facial recognition. *Technical Report PRI:15, Panoramic Research Inc. Palo Alto CA.*, 1964.

[22] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[23] F. L. Bookstein. Thin-plate splines and the atlas problem for biomedical images. In *12th International conference in Information Processing in Medical Imaging*, Lecture Notes in Computer Science, pages 326–342, 1991.

[24] K. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. *Computer Vision and Image Understanding*, 101:1–15, 2006.

[25] A. Brett, A. Hill, and C. Taylor. A method of automatic landmark generation for automated 3D pdm construction. *Image and Vision Computing*, 18:739–748, 2000.

[26] A. Brett and C. Taylor. A method of automatic landmark generation for automated 3D pdm construction. In *9th British Machine Vision Conference Proceedings*, pages 914–923. Springer, 1998.

[27] A. Bronstein, M. Bronstein, and R. Kimmel. Expression-invariant 3D face recognition. In *International Conference on Audio- and Video-Based Person Authentication*, pages 62–70, 2003.

[28] A. Bronstein, M. Bronstein, and R. Kimmel. Three-dimensional face recognition. *International Journal of Computer Vision*, 64:5–30, 2005.

[29] V. Bruce. Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, 73:105–116, 1982.

[30] V. Bruce. *Recognizing faces*. Lawrence Erlbaum Associates, 1988.

[31] V. Bruce. Stability from variation: the case of face recognition. *Quarterly Journal of Experimental Psychology*, 47(1):5–28, 1994.

[32] V. Bruce, P. Hancock, and A. Burton. *Face Recognition: From Theory to Applications*. Springer-Verlag, 1998.

[33] V. Bruce and S. Langton. The use of pigmentation and shading information in recognising the sex and identities of faces. *Perception*, 23:803–822, 1994.

[34] L. Brunie, S. Lavallée, and R. Szeliski. Using forces fields derived from 3D distance maps for inferring the attitude of a 3D rigid object. In *Second European Conference on Computer Vision*, pages 670–675, 1992.

[35] J. Buhmann, M. Lades, and C. Malsburg. Size and distortion invariant object recognition by hierarchical graph matching. In *Proceedings of the International Joint Conference on Neural Networks*, pages 411–416, 1990.

[36] A. Burton and S. Wilson. Face recognition in poor-quality video. *Psychological Science*, 10:243–248, 1999.

[37] R. Campbell and P. Flynn. Eigenshapes for 3D object recognition in range data. In *Proceedings of Computer Vision and Pattern Recognition*, 1999.

[38] J. Cartoux, J. LaPreste, and M. Richetin. Face authentication or recognition by profile extraction from range images. In *Workshop on Interpretation of 3D Scenes*, pages 194–199, 1989.

[39] K. Chang, K. Bowyer, and P. Flynn. Face recognition using 2D and 3D facial data. In *Multimodal User Authentication Workshop*, pages 25–32, 2003.

[40] K. Chang, K. Bowyer, and P. Flynn. Adaptive rigid multi-region selection for handling expression variation in 3D face recognition. In *IEEE Workshop for Face Recognition Grand Challenge Experiments*, 2005.

[41] K. Chang, K. Bowyer, and P. Flynn. An evaluation of multi-modal 2D+3D face biometrics. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 27(4):619–624, 2005.

[42] N. Chawla and K. Bowyer. Random subspaces and subsampling for 2-d face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 582– 589, 2005.

[43] Y. Chen and G. Medioni. Object modelling by registration of multiple range images. In *IEEE Proceedings of the Conference Robotics and Automation*, pages 2724–2729, 1991.

[44] C. Chua, F. Han, and Y. Ho. 3D human face recognition using point signature. In *International Conference on Face and Gesture Recognition*, pages 233–238, 2000.

[45] C. Chua and R. Jarvis. Point signatures - a new representation for 3D object recognition. *International Journal of Computer Vision*, 25(1):63–85, 1997.

[46] University of Notre Dame Computer Vision Research Laboratory. University of notre dame biometrics database distribution. http://www.nd.edu/ cvrl/UNDBiometricsDatabase.html, 2002-2004.

[47] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *European Conference of Computer Vision*, number 2, pages 484–498, 1998.

[48] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:18–23, 1995.

[49] I. Cox, J. Ghosin, and P. Yianilos. Feature-based face recognition using mixture-distance. In *Computer Vision and Pattern Recognition*, pages 209–216, 1996.

[50] I. Craw and P. Cameron. Parameterizing images for recognition and reconstruction. In *British Machine Vision Conference Proceedings*, pages 367–370. Springer, 1991.

[51] R. Davies. *Learning Shape: Optimal Models for Analysing Natural Variability*. PhD thesis, University of Manchester, 2002.

[52] J. Duchon. Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *R.A.I.R.O. Analyse Numrique*, 10:5–12, 1976.

[53] G. Edwards, T. Cootes, and C. Taylor. Face recognition using active appearance models. In *Proceedings of the European Conference on Computer Vision*, pages 581–695, 1998.

[54] A. Elad and R. Kimmel. Bending invariant representations for surfaces. In *Computer Vision and Pattern Recognition*, pages 168–174, 2001.

[55] H. Ellis. *Aspects of Face Processing*. Nijhoff, Dordrecht, 1986.

[56] T. Emidio, R. Schmidt, and R. Marcondes. Eigenfaces versus eigeneyes: First steps toward performance assessment of representations for face recognition. *Lecture Notes in Artificial Intelligence*, 1783:197–206, 2000.

[57] K. Etemad and R. Chellapa. Discriminant analysis for recognition of human face images. *Journal of the Optical Society of America*, 14:711–720, 1997.

[58] J. Feldmar and N. Ayache. Locally affine registration of free-form surface. In *Proceedings of the Conference of Computer Vision and Pattern Recognition*, pages 496–501, 1994.

[59] J. Feldmar and N. Ayache. Rigid and affine registration of free-form surfaces using

differential properties. In *Proceedings of the European Conference of Computer Vision*, pages 397–406, 1994.

[60] J. Feldmar, J. Declerck, G. Malandain, and N. Ayache. Extension of the icp algorithm to nonrigid intensity-based registration of 3D volumes. *Computer Vision and Image Understanding*, 66(2):193–206, 1997.

[61] R. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376–386, 1938.

[62] J. Fitzpatrick and J. West. The distribution of target registration error in rigid-body. *IEEE Transactions in Medical Imaging*, 20(9):917–927, 2001.

[63] J. Foley, A. van Dam, S. Feiner, and J. Hughes. Computer graphics. *ASME Journal of Biomechanical Engineering*, 122(4):354–363, 2000.

[64] A. Frangi, D. Rueckert, J. Schnabel, and W. Niessen. Automatic construction of multiple-object three-dimensional statistical shape models:application to cardiac modeling. *IEEE Transactions on Medical Imaging*, 21(9):1151–1166, 2002.

[65] J. Friedman, J. Bentley, and R. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3):209–226, 1977.

[66] I. Gauthier, M. Behrmann, and M. Tarr. Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience*, 11:349–370, 1999.

[67] I. Gauthier and N. Logothetis. Is face recognition so unique after all? *Journal of Cognitive Neuropsychology*, 17:125–142, 2000.

[68] I. Gauthier, M. Tarr, A. Anderson, P. Skudlarski, and J. Gore. Activation of the middle fusiform face area increases with expertise recognizing novel objects. *Nature Neuroscience*, 4(2):258–273, 1999.

[69] A. Godil, S. Ressler, and P. Grother. Face recognition using 3D facial shape and colour map information:comparison and combination. In *SPIE*, number 5404, pages 351–361, 2004.

[70] G. Godin, M. Rioux, and R. Baribeau. Three-dimensional registration using range and intensity information. In *Proceedings SPIE Videometrics III*, number 2350, 1994.

[71] B. Gökberk, A. Salah, and L. Akarun. Rank-based decision fusion for 3D shape-based face recognition. In *International Conference on Audio- and Video-based Biometric Person Authentication*, pages 1019–1028, 2005.

[72] D. Goldberg. Time. http://www.zonezero.com/magazine/essays/diegotime/time.html, 1976-2005.

[73] D. Goldof, H. Lee, and T. Huang. Feature extraction and terrain matching. In *Proceedings of Computer Vision and Pattern Recognition*, pages 899–904, 1998.

[74] A. Goldstein, L. Harmon, and Lesk A. Identification of human faces. In *Proceedings of the IEEE*, number 5, pages 748–760, 1971.

[75] G. Gordon. Face recognition based on depth and curvature features. In *Computer Vision and Pattern Recognition*, pages 808–810, 1992.

[76] A. Goshtasby. Registration of images with geometric distortions. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):60–64, 1988.

[77] H. Gray. *Anatomy of the human body*. Lea and Fabiger, 1918.

[78] R. Gross and V. Brajovic. An image preprocessing algorithm for illumination invariant face recognition. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pages 10–18, 2003.

[79] R. Gross, I. Matthews, and S. Baker. Eigen light-fields and face recognition across

pose. In *International Conference on Automatic Face and Gesture Recognition*, 2002.

[80] R. Gross, J. Shi, and J. Cohn. Quo vadis face recognition? In *Third Workshop on Empirical Evaluation Methods in Computer Vision*, pages 145–152, 2001.

[81] R. Grossmann, N. Kiryati, and R. Kimmel. Computational surface flattening: a voxel-based approach grossmann. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:433–441, 2002.

[82] P. Grother, R.J. Michaels, and P.J. Phillips. Face recognition vendor test 2002 performance metrics. Number 2688 in Lecture Notes in Computer Science, pages 937 – 945. Springer, 2003.

[83] MIT Media Lab: VisMod Group. Photobook/eigenfaces demo. http://vismod.media.mit.edu/vismod/demos/facerec/basic.html, 2002.

[84] P. Halliman, G. Gordon, A. Yuille, P. Giblin, and D. Mumford. *Two and three-dimensional patterns of the face*. A.K. Peters, 1999.

[85] P. Hammond, T. Hutton, J. Allanson, A. Shaw, and M. Patton. 3D digital stereophotogrammetric analysis of noonan syndromed. In *British Human Genetics Conference*. Blackwell Synergy, 2002.

[86] L. Harmon and B. Julesz. Masking in visual recognition: Effects of two-dimensional noise. *Science*, 180:1194–1197, 1973.

[87] J. Haxby, E. Hoffman, and M. Gobbini. The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 20(6):223–233, 2000.

[88] B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 657–662, 2001.

[89] C. Hesher, A. Srivastava, and G. Erlebacher. A novel technique for face recognition using range imaging. In *Seventh International Symposium on Signal Processing and Its Applications*, pages 201–204, 2003.

[90] R. Hietmeyer. Biometric identification promises fast and secure processing of airline passengers. *The International Civil Aviation Organization Journal*, 55(9):10–11, 2000.

[91] H. Hill and V. Bruce. Effects of lighting on matching facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, 22:986–1004, 1996.

[92] H. Hill, P. Schyns, and S. Akamatsu. Information and viewpoint dependence in face recognition. *Cognition*, 62:201–202, 1997.

[93] E. Hjelmas and J. Wroldsen. Recognizing faces from the eyes only. page Pattern Recognition, 1999.

[94] J. Huang, B. Heisele, and V. Blanz. Component-based face recognition with 3D morphable models. In *International Conference on Audio- and Video-Based Person Authentication*, 2003.

[95] M. Husken, M. Brauchmann, S. Gehlen, and C. von der Malsburg. Strategies and benefits of fusion of 2D and 3D face recognition. In *IEEE Workshop on Face Recognition Grand Challenge Experiments*, 2005.

[96] T. Hutton. *Dense Surface Models of the Human Face*. PhD thesis, University College London, 2004.

[97] T. Hutton, B. Buxton, and P. Hammond. Dense surface point distribution models of the human face. In *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 153–160, 2001.

[98] T. Hutton, B. Buxton, P. Hammond, and H. Potts. Estimating average growth trajectories in shape-space using kernel smoothing. *IEEE Transactions on Medical Imaging*, 22(6):747–753, 2003.

[99] C. Jacek, M. Sadeghi, J. Kittler, and L. Vandendorpe. Decision fusion for face authentication. In *International Conference on Biometric Authentication*, number 2072, pages 686–693, 2004.

[100] A. Johnson and M. Hebert. Surface matching for object recognition in complex three-dimensional scenes. *Image and Vision Computing*, 16:635–651, 1998.

[101] A.E. Johnson and S.B. Kang. Registration and integration of textured 3D data. *Image and Vision Computing*, 17:135–147, 1999.

[102] M. Johnson, S. Dziurawiec, H. Ellis, and J. Morton. Newborns preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40:1–19, 1991.

[103] A. Johnston, H. Hill, and N. Carman. Recognizing faces: effects of lighting direction, inversion and brightness reversal. *Perception*, 21:365–375, 1992.

[104] M. Jones and T. Poggio. Multidimensional morphable models: A framework for representing and matching object classes. *International Journal of Computer Vision*, 29(2):107–131, 1998.

[105] I. Kakadiaris, G. Passalis, T. Theoharis, G. Toderici, I. Konstantinidis, and N. Murtuza. Multimodal face recognition: combination of geometry with physiological information. Number 2, pages 1022 – 1029, 2005.

[106] T. Kanade. *Computer recognition of human faces*. Birkhauser, Basel, Switzerland, 1973.

[107] N. Kehtarnavaz and S. Mohan. A framework for estimation of emotion parameters from range images. In *Computer Vision, Graphics and Image Processing*, number 45, pages 88–105, 1989.

[108] M. Kelly. Visual identification of people by computer. *Technical Report AI-130, Stanford AI Project*, 1970.

[109] T. Kim and J. Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):318–327, 2005.

[110] M. Kirby and L. Sirovich. Application of the karhunen-loéve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[111] J. Kittler, Y. Li, and J. Matas. On matching scores for lda-based face verification. In *British Machine Vision Conference Proceedings*. Springer, 2000.

[112] Y. Lamdan and H. Wolfson. Geometric hashing: a general and efficient model-based recognition scheme. In *International Conference of Computer Vision*, pages 238–249, 1988.

[113] A. Lanitis, C. Taylor, and T. Cootes. Automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393–401, 1995.

[114] S. Lavallée, R. Szeliski, and L. Brunie. Matching 3D smooth surfaces with their 2D projections using 3D distance maps. In *SPIE Geometric Methods in Computer Vision*, pages 322–336, 1991.

[115] J. Lee and E. Milios. Matching range images of human faces. In *International Conference on Computer Vision*, pages 722–726, 1990.

[116] S. Lee, G. Wolberd, and S. Shin. Scattered data interpolation with multilevel b-splines. *IEEE Transactions on Visualization and Computer Graphics*, 3(3):228–244, 1997.

[117] Y. Lee and J. Shim. Curvature-based human face recognition using depth-weighted hausdorff distance. In *International Conference on Image Processing*, pages 1429–1432, 2004.

[118] Y. Lee, H. Song, U. Yang, H. Shin, and K. Sohn. Local feature based 3D face recognition. In *International Conference on Audio- and Video-based Biometric Person Authentication*, pages 909–918. Springer, 2005.

[119] S. Li and A. Jain. *Handbook of Face Recognition*. Springer-Verlag, 2004.

[120] S. Li and J. Lu. Face recognition using the nearest feature line method. *IEEE Transactions on Neural Networks*, 10(2):439–443, 1999.

[121] Y. Li, S. Gong, and H. Lidell. Support vector regression and classification based multi-view face detection and recognition. In *International Conference on Face and Gesture Recognition*, pages 300–305, 2000.

[122] L. Light, F. Kayra-Stuart, and S. Hollander. Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, 5(3):212–228, 1979.

[123] Vision RT Limited. http://www.visionrt.com, 2001-2006.

[124] S. Lin, S. Hung, and L. Lin. Face recognition/detection by probabilistic desision-based neural network. *IEEE Transactions on Neural Networks*, 8:114–132, 1997.

[125] C. Liu, C. Collin, A. Burton, and A. Chaurdhuri. Lighting direction affects recognition of untextured faces in photographic positive and negative. *Vision Research*, 39:4003–4009, 1999.

[126] C. Liu and H. Wechsler. Evolutionary pursuit and its application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:570–582, 2000.

[127] X. Lu, D. Colbry, and A. Jain. Matching 2.5D scans for face recognition. In *International Conference on Pattern Recognition*, pages 362–366, 2004.

[128] X. Lu and A. Jain. Deformation analysis for 3D face matching. In *7th IEEE Workshop on Applications of Computer Vision*, pages 362–366, 2005.

[129] X. Lu and A. Jain. Intergrating range and texture information for 3D face recognition. In *7th IEEE Workshop on Applications of Computer Vision*, pages 155–163, 2005.

[130] D. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA, 2nd edition, 1984.

[131] F. Maes, A. Collignon, D. Vandermeulen, G. Marechal, and R. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16:187–198, 1997.

[132] T. Masuda, K. Sakaue, and N. Yokoya. Registration and integration of multiple range images for 3D model construction. In *Computer Vision and Pattern Recognition*, 1996.

[133] T. Maurer, D. Guigonis, I. Maslov, B. Pesenti, A. Tsaregorodtsev, D. West, and G. Medioni. Performance of geometrix activeidtm 3D face recognition engine on the frgc data. In *IEEE Workshop on Face Recognition Grand Challenge Experiments*, 2005.

[134] N. Mavridis, F. Tsalakanidou, D. Pantazis, S. Malasiotis, and M. Strintzis. The hiscore face recognition application: Affordable desktop face recognition based on a novel 3D camera. In *International Conference on Augmented Virtual Environments and 3D Images*, pages 157–160, 2001.

[135] G. Medioni and R. Waupotitsch. Face recognition and modeling in 3D. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 232–233, 2003.

[136] J. Meinguet. Multivariate interpolation at arbitrary points made simple. *Zeitschrift fur Angewandte Mathematik and Physik*, 30:292–304, 1979.

[137] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb the extended

m2vts database. In *Proceedings of the Second International Conference on audio and Video-based biometric Person Authentication*, 1999.

[138] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX*, pages 41–48, 1999.

[139] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[140] O. Monga and S. Benayoun. Using partial derivatives of 3D images to extract typical surface features. *Computer Vision and Image Understanding*, 61(2):171–189, 1995.

[141] A. Moreno, A. Sanchez, J. Velez, and F. Diaz. Face recognition using 3D surface-extracted descriptors. In *Irish Machine Vision and Image Processing Conference*, 2003.

[142] T. Nagamine, T. Uemura, and I. Masuda. 3D facial image analysis for human identification. In *International Conference on Pattern Recognition*, pages 324–327, 1992.

[143] A. Nefian and M. Hayes III. Hidden markov models for face recognition. Number 5, pages 2721–2724, 1998.

[144] University of Surrey. The extended m2vts database. http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/, 2006.

[145] K. Okada, J. Steffans, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. Malsburg. *Face Recognition: From Theory to Applications*. Springer, Berlin, Germany, 1998.

[146] A. Oliva and P. Schyns. Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, 34:72–107, 1997.

[147] A. O'Toole, T. Price, T. Vetter, J. Bartlett, and V. Blanz. Three-dimensional shape and two-dimensional surface textures of human faces: The role of ”averages” in attractiveness and age. *Image and Vision Computing Journal*, 18(1):9–19, 1999.

[148] A. O'Toole, T. Vetter, H. Volz, and E. Salter. Three-dimensional caricatures of human heads: distinctiveness and perception of facial age. *Perception*, 26(6):719–732, 1997.

[149] G. Pan, S. Han, Z. Wu, and Y. Wang. 3D face recognition using mapped depth images. pages 175–175, 2005.

[150] G. Pan, Z. Wu, and Y. Pan. Automatic 3D face verification from range data. pages III: 193–196, 2003.

[151] T. Papatheodorou and D. Rueckert. Evaluation of automatic 3D face recognition using surface and texture registration. In *Sixth International Conference on Automated Face and Gesture Recognition*, pages 321–326, 2004.

[152] G. Passalis, I. Kakadiaris, T. Theoharis, G. Toderici, and N. Murtuza. Evaluation of 3D face recognition in the presence of facial expressions: an annotated deformable model approach. pages 1022 – 1029, 2005.

[153] P. Penev and J. Atick. Local feature analysis: A general statistical theory for object recognition. *Computational Neural Systems*, 7:477–500, 1996.

[154] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Computer Vision and Pattern Recognition*, pages 84–91, 1994.

[155] A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modelling and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):715–729, 1993.

[156] J. Phillips, P. Grother, and R. Michaels. *Handbook of Face Recognition*. Springer-Verlag, 2004.

[157] J.P. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[158] P. Phillips. Support vector machines applied to face recognition. *Advanced Neural Information Processing Systems*, 11:803–809, 1998.

[159] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and J. Bone. Face recognition vendor test 2002: Evaluation report. *National Institute of Standards and Technology*, 2003.

[160] P. Phillips, A. Martin, C. Wilson, and M. Przybocki. An introduction to evaluating biometric systems. *Computer*, 63:56–63, 2000.

[161] T. Poggio and S. Edelman. A network that learns to recognize 3D objects. *Nature*, 343:263–266, 1991.

[162] S. Prince and J. Elder. Creating invariance to ’nuisance parameters’ in face recognition. *IEEE Computer Vision and Pattern Recognition*, pages 439–446, 2005.

[163] S. Prince and J. Elder. Tied factor analysis for face recognition across large pose changes. In *British Machine Vision Conference*, 2006.

[164] K. Pulli. Multiview registration for large data sets. In *3DIM*, 1999.

[165] A. Rangarajan, E. Mjolsness, S. Pappu, L. Davachi, P. Goldman-Rakic, and J. Duncan. A robust point matching algorithm for autoradiograph alignment. In *Visualization in Biomedical Computing*, pages 277–286, 1996.

[166] Riya. Riya photo search. http://www.riya.com/, 2005.

[167] S. Rizvi, P. Phillips, and H. Moon. A verification protocol and statistical performance analysis for face recognition algorithms. *IEEE Conference on Computer Vision and Pattern recognition*, pages 833–838, 1998.

[168] R. Robins and E. McKone. Can holistic processing be learned for inverted faces? *Cognition*, 88:79–107, 2003.

[169] K. Rohr, H. Stiehl, R. Sprengel, T. Buzug, J. Weese, and M. Kuhn. Landmark-based elastic registration using approximating thin-plate splines. *Transcations on Medican Imaging*, 20(6):526–534, 2001.

[170] S. Romdhani, V. Blanz, C. Basso, and T. Vetter. *Handbook of Face Recognition*. Springer-Verlag, 2004.

[171] A. Ross and Jain A. Information fusion in biometrics. *Information fusion in biometrics*, 24:2115–2125, 2003.

[172] W. Rucklidge. Efficient visual recognition using the hausdorff distance. *Lecture Notes in Computer Science*, 1173(12):2307–2322, 1996.

[173] D. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5:517–548, 1994.

[174] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *Proceedings of the Third Intl. Conf. on 3D Digital Imaging and Modeling*, pages 145–152, 2001.

[175] T. Russ, M. Koch, and C. Little. 3D facial recognition: a quantitative analysis. In *38th Annual 2004 International Carnahan Conference on Security Technology*, pages 338–344, 2004.

[176] T. Russ, M. Koch, and C. Little. A 2D range hausdorff approach for 3D face recognition. In *Computer Vision and Pattern Recognition*, pages 1429–1432, 2005.

[177] O. Sacks. *The man who mistook his wife for a hat.* Picador, 1985.

[178] J. Sadr, I. Jarido, and P. Sinha. The role of eyebrows in face recognition. *Perception*, 32:285–293, 2003.

[179] F. Samaria. *Face recognition using hidden markov models*. PhD thesis, University of Cambridge, 1994.

[180] P. Sander and S. Zucker. Inferring surface trace and differential structure from 3D images. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 12(9):833–854, 1990.

[181] B. Scholkopf, A. Smola, and K. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1999.

[182] E. Schwartz, A. Shaw, and E. Wolfson. A numerical solution to the generalized mapmaker's problem: flattening nonconvex polyhedral surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(9):1005–1008, 1989.

[183] T. Sederberg and S. Parry. Free-form deformation of solid geometric models. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques*, pages 151–160, 1986.

[184] J. Shepherd, G. Davies, and H. Ellis. *Perceiving and Remembering Faces*, chapter Studies of cue saliency. Academic Press, London, 1981.

[185] F. Simion, V. Cassia, C. Turati, and E. Valenza. The origins of face perception: Specific versus non-specific mechanisms. *Infant and Child Development*, 10:59–65, 2001.

[186] D. Simon. *Fast and Accurate Shape-Based Registration*. PhD thesis, Carnegie Mellon University, 1996.

[187] P. Sinha and T. Poggio. I think i know that face. *Nature*, 384:404, 1996.

[188] L. Sirovich and M. Kirby. A low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 4(3):519–524, 1987.

[189] N. Smith, I. Meir, G. Hale, R. Howe, L. Johnson, P. Edwards, D. Hawkes, M. Bidmead, and D. Landau. Real-time 3D surface imaging for patient positioning

in radiotherapy. *International Journal of Radiation Oncology Biology Physics*, 57(2):187, 2003.

[190] R. Solar and P. Navarrete. Eigenspace-based face recognition: a comparative study of different approaches. *IEEE Transactions in Systems and Cybernetics*, 35:315–325, 2005.

[191] M. Spiegel and L. Stephens. *Schaum's Outline of Statistics*. Schaum, 1998.

[192] A. Srivastava, X. Liu, and C. Hesher. Face recognition using optimal linear components of face images. *Journal of Image and Vision Computing*, 24(3):291–299, 2003.

[193] G. Stockman. Object recognition and localization via pose clustering. In *Computer Vision, Graphics and Image Processing*, number 40, pages 361–387, 1987.

[194] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836, 1996.

[195] R. Szeliski and S. Lavallee. Matching 3D anatomical surfaces with non-rigid deformations using octree-splines. In *IEEE Workshop on Biomedical Image Analysis*, pages 144–153, 1994.

[196] H. Tanaka, M. Ikeda, and H. Chiaki. Curvature-based face surface recognition using spherical correlation principal directions for curved object recognition. In *Third International Conference on Automated Face and Gesture Recognition*, pages 372–377, 1998.

[197] M. Tarr and H. Bulthoff. Is human object recognition better described by geon structural descriptions or by multiple views. *Journal of Experimental Psychology*, 21:71–86, 1995.

[198] D. Terzopoulos and D. Metaxas. Dynamic 3D models with local and global deformations: deformable superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):703–714, 1991.

[199] J. Thirion. Extremal points: definition and application for 3D image registration. In *Proceedings of Computer Vision and Pattern Recognition*, pages 587–592, 1994.

[200] P. Thompson. Margaret Thatcher: A new illusion. *Perception*, 9(4):483–484, 1980.

[201] A. Torralba and A. Oliva. The statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, 2003.

[202] F. Tsalakanidou, S. Malassiotis, and M. Strintzis. Integration of 2D and 3D images for enhanced face authentication. In *Sixth International Conference on Automated Face and Gesture Recognition*, pages 266–271, 2004.

[203] F. Tsalakanidou, D. Tzocaras, and M. Strintzis. Use of depth and colour eigenfaces for face recognition. *Pattern Recognition Letters*, 24:1427–1435, 2003.

[204] G. Turk and M. Levoy. Zippered polygon meshes from range images. In *Proceedings of SIGGRAPH94*, pages 311–318. A. Glassner, 1994.

[205] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal for Cognitive Neuroscience*, 3(1):71–86, 1991.

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

[207] Arizona State University. 3D face authentication database. http://prism.asu.edu/3Dface/default.asp, 2006.

[208] T. Valentine. A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quarterly Journal of Experimental Psychology*, 43(2):161–204, 1991.

[209] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):733–742, 1995.

[210] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 19:733–742, 1997.

[211] T. Vetter and N. Troje. Separation of texture and shape in images of faces for image coding and synthesis. *Journal of the Optical Society of America*, 14:2152–2161, 1997.

[212] P. Viola. *Alignment by maximination of mutual information*. PhD thesis, Massachusetts Institute of Technology, 1995.

[213] Y. Wang, C. Chua, and Y. Ho. Facial feature detection and face recognition from 2D and 3D images. *Pattern Recognition Letters*, 23:1191–1202, 2002.

[214] Y. Wang, B. Peterson, and L Staib. Shape-based 3D surface correspondence using geodesics and local geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 644–651, 2000.

[215] S. Weik. Registration and integration of multiple range images for 3D model construction. In *Proceedings 3DIM*, 1997.

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

[217] H. Wong, K. Chueng, and H. Ip. 3D head model classification by evolutionary optimization of the extended gaussian image representation. *Pattern Recognition*, 37(12):2307–2322, 2004.

[218] Y. Wu, G. Pan, and Z. Wu. Face authentication based on multiple profiles extracted from range data. In *Audio- and Video-Based Biometric Person Authentication*, pages 515–522, 2003.

[219] C. Xu, Y. Wang, T. Tan, and L. Quan. Automatic 3D face recognition combining global geometric features with local shape variation information. In *Sixth International Conference on Automated Face and Gesture Recognition*, pages 308–313, 2004.

[220] Y. Xu, J. Liu, and N. Kanwisher. The m170 is selective for faces, not for expertise. *Neuropsychologia*, 43:588–597, 2005.

[221] W. Yambor, B. Draper, and J. Beveridge. *Empirical Evaluation Techniques in Computer Vision*. Wiley, 2000.

[222] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3D facial expression database for facial behavior research. In *Proceedings of the 7th Int. Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006.

[223] L. Yin and M. Yourst. 3D face recognition based on high-resolution 3D face modeling from frontal and profile views. In *ACM SIGMM workshop on Biometric methods and applications*, pages 1–8. ACM Press, 2003.

[224] R. Yin. Looking at upside-down faces. *Journal of Experimental Psychology*, 81:141–145, 1969.

[225] A. Young, D. Hellawell, and D. Hay. Configurational information in face perception. *Perception*, 16:747–759, 1987.

[226] W. Zhao, R. Chellapa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *International Conference on Automatic Face and Gesture*, pages 336–341, 1998.

[227] W. Zhao, R. Chellapa, and P. Phillips. Subspace linear discriminant analysis for face recognition. *Technical Report CAR-TR-914, Center for Automation Research, University of Maryland, College Park, MD*, 1999.

[228] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: A literature survey. *ACM Computing Survey*, 35(4):399–458, 2003.

[229] G. Zigelman, R. Kimmel, and N. Kiryati. Texture mapping using surface flattening via multi-dimensional scaling. *IEEE Transactions on Visualization and Computer Graphics*, 8:198–207, 2002.