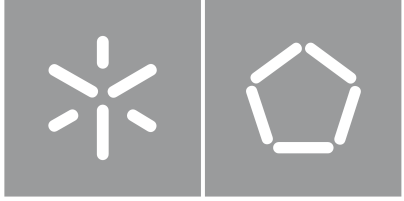
Universidade do Minho
Escola de Engenharia

Waldir Spencer Pimenta Lima

Face recognition using 3D structural geometry of rigid features extracted from 2D images

Novembro de 2010



Universidade do Minho
Escola de Engenharia

Waldir Spencer Pimenta Lima

Face recognition using 3D structural geometry of rigid features extracted from 2D images

Tese de Mestrado
Mestrado em Informática

Trabalho efectuado sob a orientação de
Doutor Luís Paulo Peixoto dos Santos

Novembro de 2010

DECLARAÇÃO

Nome

_____

Endereço electrónico: _____ Telefone: _____ / _____

Número do Bilhete de Identidade: _____

Título dissertação □ / tese □

_____

_____

_____

Orientador(es):

_____

_____ Ano de conclusão: _____

Designação do Mestrado ou do Ramo de Conhecimento do Doutoramento:

_____

Universidade do Minho, ____ / ____ / _____

Assinatura: _____

Acknowledgements

"No man is an island."
John Donne (1572-1631)

This project was developed during an internship at VicomTech, as part of my master's degree program. As I approach its end, I can appreciate how much help I received, and realize that without it I couldn't have done but a fraction of what was achieved. I thus want to express my highest gratitude to the following people:

- My thesis supervisor, Luís Paulo Santos, for the initial proposal and coordination which made this work possible, for the following-up on subsequent developments with keen interest and thoughtful suggestions, and for the help reviewing this thesis in several stages of its development;
- My supervisor during the internship, Luis Unzueta, for the constant support and guidance, many crucial suggestions and advice, timely and relevant discussions, and countless valuable pointers to theoretical and practical approaches, many of which shaped the project's direction, this document's content, or both;
- Several people at Vicomtech, for their instrumental help in many steps of the project: Jon Goenetxea (Qt primer), Javier Barandiran (POSIT, Levenberg-Marquardt, OpenGL and some matrix madness), Marcos Doncel (optimization techniques), John Congote (porting to Linux), Harbil Arregui (review of the Eigenfaces section) and everyone else at the company, for the friendly and helpful environment they provided;
- Others who have generously provided their help, including Jairo Sanchez Tapia, for help understanding AAMs and some basic underlining concepts; Fadi Dornaika, for the insightful theoretical advice and encouraging comments; Pedro Gomes, for help with TeX and formatting this thesis; and members of the IRC channels #qt and #latex at freenode.net, for the valuable help, guidance and suggestions.

I also mustn't leave out the people who, despite not having helped me personally, created the resources upon which I built my work, thus deserving accolades for giving away the products of their work for anyone to use. These include Qt (and its excellent documentation), Yao Wei's AAMlibrary and ASMlibrary, OpenCV, OpenGL, Daniel DeMenthon's POSIT, Joachim Wuttke's Levenberg-Marquardt implementation, Mislav Grgic's SCFace database, the Cohn-Kanade face database, among many others.

Finally, this project would also not have been possible without the support infrastructure provided by Vicomtech, which promptly accepted my internship proposal and provided me with all resources to carry on my project; and the Erasmus program, which financed my internship abroad with a scholarship.

Abstract

The problem of face recognition by computers has been the subject of over 40 years of research in computer vision, and a plethora of increasingly sophisticated techniques have been proposed throughout the years, with varying degrees of success. However, a definitive solution for uncontrolled, high-performance and/or high-precision environments has been elusive, mainly because most of these methods work on 2D images, which are mere projections of the the three-dimensional human face, thus being heavily variable in shape and appearance depending on pose, illumination and expressions of the face. More recently, 3D approaches have emerged, but despite good results, they present technical hurdles that 2D methods can avoid.

This thesis aims to evaluate the hypothesis that rigid facial shape contributes significantly to the recognition problem. While many 2D approaches use shape-free patches for comparison, no approach has attempted partial shape normalization, canceling facial expression-based deformations, but preserving intrinsic proportions of the facial shape. We propose a hybrid 2D+3D recognition method on which we attempt to achieve this *selective* shape normalization by fitting the Candide 3D morphable face model into a 2D mesh generated through Active Shape Models, thus obtaining explicit estimates for pose, facial shape and facial expression. Our results show that normalizing pose and expression, with shape unmodified, increases performance in distinguishing faces.

Keywords: face recognition, pose estimation, parametric models, photogrammetry, anthropometry.

Resumo

O problema do reconhecimento facial por computadores tem-se mantido um dos grandes focos de investigação em visão por computador há mais de quatro décadas, e uma miríade de técnicas cada vez mais sofisticadas, propostas ao longo dos anos, tem resultado em vários níveis de sucesso. No entanto, uma solução definitiva para ambientes não controlados, de alto desempenho e/ou alta precisão tem-se revelada evasiva, principalmente porque a maioria destes métodos trabalha em imagens 2D, que na prática não passam de meras projeções da face humana como objecto tridimensional, tendo portanto uma aparência extremamente variável, dependendo da pose, da iluminação e das expressões faciais. Muita da investigação recente começou a optar por abordagens 3D, mas apesar de bons resultados, estas apresentam dificuldades técnicas que os métodos 2D podem evitar.

Esta tese tem como objetivo avaliar a hipótese de que as estruturas rígidas da face contribuem de forma significativa para o problema do reconhecimento facial. Embora muitas abordagens 2D usem extractos das imagens com formas normalizadas para comparação, nenhuma até agora tentou obter uma normalização parcial, cancelando deformações faciais baseadas em expressões, mas preservando as proporções intrínsecas da estrutura facial. Propõe-se um método de reconhecimento híbrido 2D+3D, no qual se tenta atingir este efeito de normalização *selectiva* através do ajuste do modelo Candide, um modelo tridimensional da face humana, a uma malha de pontos 2D gerada por meio de modelos maleáveis (*Active Shape Models*), obtendo assim estimativas explícitas para a pose, a forma da face e as expressões faciais. Os resultados indicam que a normalização de pose e expressões, mantendo a forma inalterada, aumenta o desempenho do reconhecimento facial.

Palavras-chave: reconhecimento facial, estimação de pose, modelos paramétricos, fotogrametria, antropometria.

Contents

List of Figures

Chapter 1

Introduction

Face recognition has been recently receiving more and more attention in research and industry. This is evident from the increasing number of face recognition conferences such as the International Conference on Audio- and Video-Based Authentication (AVBPA) since 1997 and the International Conference on Automatic Face and Gesture Recognition (AFGR) since 1995. Also reflecting the increase in research production on the subject is the proliferation of systematic empirical evaluations of face recognition techniques, including the FERET tests, FRVT 2000 [Blackburn et al., 2001], FRVT 2002 [Phillips et al., 2003], the Face Recognition Grand Challenge [Phillips et al., 2005] and the XM2VTS [Messer et al., 1999] protocol.

It is clear that robust face detection and recognition in real-time video would provide great value to many industry needs. However, even though there have been many advances in the last few years, the aforementioned events, exhibitions and evaluations show that there is still a long way to go until we achieve accuracy rates that approach the human visual system. Facial recognition technology is still not robust enough for the more demanding applications, such as automated security platforms, as these have an extremely low tolerance for errors due to their high cost in potential material or even human losses. Thus, there are many research projects currently active, aiming to improve the reliability and robustness of these systems. The next paragraphs will detail specific uses of facial detection and recognition.

Figure 1.1: A screenshot of Google Street View, showing people whose faces have been automatically detected and blurred, to protect their privacy. ©2009 Google.

Face detection consists in automatically locating generic human faces in still images or video streams. This ability can be (and has been) used in many contexts. Applications that can already be found in the consumer market include webcams that automatically focus on the user and smart digital cameras that detect smiles for automatic shooting. Companies have been using face detection for longer; for example, to remove/blur faces in public image databases, for privacy reasons (a well-known example is Google's StreetView, see Figure 1.1), or to count people in a room or crowd. Also, much interest is dedicated to the development of intelligent human computer interfaces (see [Toyama, 1998]) such as virtual reality environments, training programs, or video games [Zhao et al., 2003]. Many other uses are possible, but these examples should supply a general overview of current trends.

Face detection can be further enhanced by **face recognition**, that is, the identification of a specific person by comparing the detected face to a database of known faces.

This is very useful in biometrics, since it requires no explicit cooperation from the person being identified, and can operate on several people at once (Figure 1.2), as opposed to other biometric sensors such as fingerprint or eye iris scanners, which depend on the subject's cooperation. A natural area of application is thus access control for places (automated board crossings in airports, restricted areas/rooms) or information (medical records, content under parental control).

Figure 1.2: A scene from the movie Minority Report, depicting a crowd passing through a biometric identification system. ©2002 20th Century Fox.

Two main uses of face recognition can be distinguished: *identification*, which is the attempt to find a match for a new face in the database of known faces, and *authentication* (or verification), which consists in comparing the new image to a specific entry on the database and determining whether they are the same person. Common analogs [ISO/IEC 19794-5:2005, Bowyer et al. 2004] are "one-to-many searching" and "one-to-one matching", respectively.

Blanz and Vetter [2003] state that "for identification, all gallery images are analyzed by the fitting algorithm, and the [...] coefficients are stored. Given a probe image, the fitting algorithm computes coefficients which are then compared with all gallery data in order to find the nearest neighbor." This kind of approach allows applications such as detecting individuals in video surveillance streams, searching for a criminal's face in a large mugshot database, or automatically tagging, cataloging and indexing large photo and video collections.

Authentication, on the other hand, "can be accomplished without a database by computing a representation of the person's face and comparing it to one stored on [a] pass key" [Gordon, 1991]. Access control is the most obvious application, but other possible uses include the detection of false identification cards, or searching one or more recorded video streams for a specific person. Robust implementations could in the future even replace (or complement) authentication mechanisms such as passwords, fingerprints, PINs or credit cards [Zhao et al., 2003].

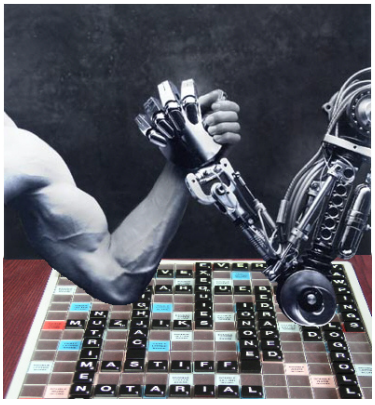


Figure 1.3: Man vs. Machine. © *SingularityHub.com*.

Humans are exceptionally good at detecting and recognizing faces. However, achieving this algorithmically is far from trivial. In fact, despite decades of research, face recognition in an unconstrained environment with changes in illumination and pose is still an unresolved problem. More specifically, current systems are still very far from the human visual perception mechanism in detecting and identifying faces. Zhao et al. [2003] consider that "It is futile to even attempt to develop a system, using existing technology, which will mimic the remarkable face recognition ability of humans. However, the human brain has its limitations in the total number of persons that it can accurately 'remember'". Apart from this large, near-perfect memory, other advantages such as the convenience, accuracy, consistency and "tirelessness" of computers make fully automated facial recognition systems very attractive, and provide a strong motivation for continuing research in the subject.

This document describes work conducted with the aim to enhance the performance of face recognition, especially in uncontrolled environments that present variations in certain factors (such as illumination, pose, or expression), which are detrimental to 2D face recognition. For this effect, the work was based on the premise that rigid facial geometry contains valuable information that should be taken into account when performing facial recognition. Specifically, we implemented a process for shape normalization that preserves the proportions of the face and cancels pose and facial expressions. This concept of selective shape normalization is relatively new in the current literature, especially concerning 2D face recognition. The framework we propose can be applied in a pre-processing step to improve 2D face recognition systems, for both authentication and identification. Although the implementation here presented is based on still images, its performance is fast enough to suggest the applicability of the system to video processing, after proper optimization.

The remainder of this document is organized as follows: In chapter 2 some basic theoretical concepts for this area are introduced. In chapter 3, an overview of the main developments in this field of research is provided. chapter 4 describes in detail the concepts behind the rigid-shape face recognition hypothesis, and some background on historical and current work that tackles the face recognition problem from this perspective. chapter 5 presents the work developed in the scope of this research project. Then, chapter 6 presents some results of the work, by performing a comparison with a standard 2D face recognition method with and without our normalization step. Finally, chapter 7 provides concluding thoughts, summarizing the work outlaid in this document, and setting proposed paths for future enhancements to the approach taken.

Chapter 2

Theoretical background

Performing face recognition through computational methods is a complex task. While there are several different approaches to this problem, some basic concepts are consistently used throughout most of them, due to their practicality and usefulness. This chapter contains a brief, non-exhaustive overview of the most common tools and techniques, mostly mathematical and statistical devices, which are on the basis of many of the approaches presented on chapter 3.

Faces are multidimensional, organic entities, characterized by a high degree of variation across many attributes [Hjelmäs and Low, 2001]. Thus, for detection and recognition purposes, faces are commonly represented as points[1] in a high-dimensional space. In image-based face recognition, specifically, a face is described by the gray-level intensity of every pixel in the image. This means that for an image with 20×20 pixels, the face description will consist of values in 400 coordinates, and therefore will be represented in a space with 400 dimensions. Of course, humans are unable to visualize or intuitively process information in more than 3 dimensions, but higher-dimensional representations are mathematically possible, and operations can be carried on data described in such high-dimensional spaces.

Manipulating data in such a high-dimensional space is a computationally intensive task, especially when we take into consideration the need to compare a new face with every face in the database in order to determine which one (if any) is the most similar, and hopefully declare a match. Thus, dimensionality reduction techniques must be employed to make these

[1] Usually called vectors, in a mathematical context.

calculations less time- and resource-consuming. In other words, this is equivalent to a "lossy compression" of the data describing a face – a technique that allows more efficient storage and faster processing (with real-time performance in current state-of-the-art algorithms) while still achieving good approximations of the actual fully defined face images, since the attributes kept for describing the faces are those determined to be the most significant (either for a generic face, or for each specific face).

The simplest, most straightforward measure of facial similarity in this configuration is the shortest distance between the positions where the faces stand in this space (that is, the basis of a nearest-neighbor classifier). Ideally, the distance between a new face image and a given database entry should be zero for a match, and large otherwise. But the reduced dimensionality of faces, allied to the fact that the same face will have a different appearance depending on illumination, pose, and facial expressions, means that there can only be approximations. It is also usual to store several images per person, to account for these potential variations in appearance.

Therefore, for each new image being compared to the database, there will be a double distribution of distances to database entries. That is, (hopefully) compact clouds of points in the database will be presumed matches, with short distances to the new face, and the remaining points will represent presumed non-matches.

Since these clouds will seldom form a perfectly spherical distribution, it is misleading to use the Euclidean notion of distance between the new point and the mean of the cloud (or center of mass, to use a physical analogy) in this context. A statistical distance measurement, called Mahalanobis distance [Mahalanobis, 1936], which takes into account the shape of the cloud (see Figure 2.1), is thus used.

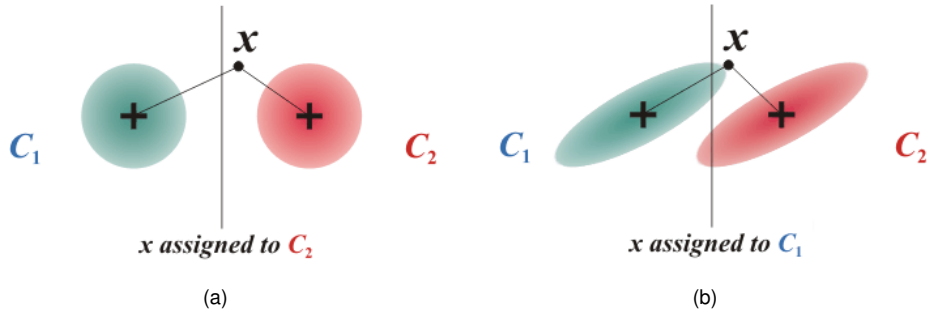


Figure 2.1: The Mahalanobis distance assigns the point x to different clouds according to their shape, despite x in both cases being closer, in Euclidean distance, to the geometrical center of the second cloud. ©*AI Access*.

By setting a sufficiently small threshold criterion for this distance, we can minimize the rate of accepting impostors (false positives) but at the expense of also increasing the rate of rejecting authentic matches (false negatives). The ratio between false positives and false negatives is commonly expressed as a Receiver Operating Characteristic, or ROC curve [Barrett, 1997]. Refer to Figure 2.3. Those ratios can be also expressed independently, in a graph with two curves (see Figure 2.2) representing false positive and false negative rates for each distance threshold value. The point where these two lines intersect is called the Equal-Error Rate (EER), and is the most commonly reported single number from the ROC curve [Bowyer et al., 2006].

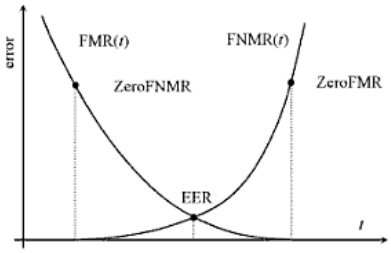


Figure 2.2: False Match Rate (FMR), or false positives, and False Non Match Rates (FNMR), or false negatives, as a function of the threshold t. Reprinted from [Maltoni et al., 2003]

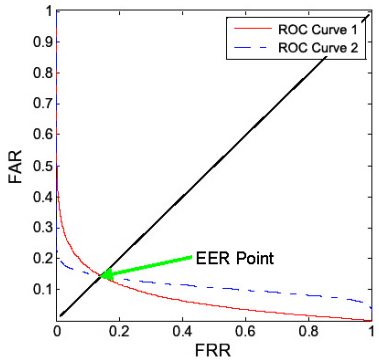


Figure 2.3: Two ROC curves, plotted by matching the False Accept Rate (FAR) with the False Reject Rate (FRR) for each threshold value. The Equal Error Rate (EER) point lies where the curves intersect the diagonal line (in this case, the two curves have the same EER). Reprinted from [Du and Chang, 2007]

The distance threshold is also a critical element distinguishing face detection systems from face recognition ones. In the first case, the threshold can be higher, as there will only be two classes: faces, or non-faces. However, when we want to discriminate between faces in the database, the threshold will then have to be lower than the average inter-class[2] separation, but higher than the average intra-class[2] sample distance.

[2]Intra-class variation occurs between samples of the same class (for example, the skin color of a person across different images may vary due to lighting, tanning, or image sensor quality), while inter-class variation is only present when comparing samples of different classes (for example, the distance between the eyes typically varies between different people, but for a given adult person it remains roughly constant across several measurements).

Principal Component Analysis and eigenvectors

Principal Component Analysis (PCA) is one of the most used methods for dimensionality reduction. It consists on a statistical analysis of a cloud of data points in order to find the vectors which account for most of the variation in the set – that is, the principal components of that set. Figure 2.4 provides an illustration for this concept. This procedure allows all data points to be approximated using a linear combination of these components.

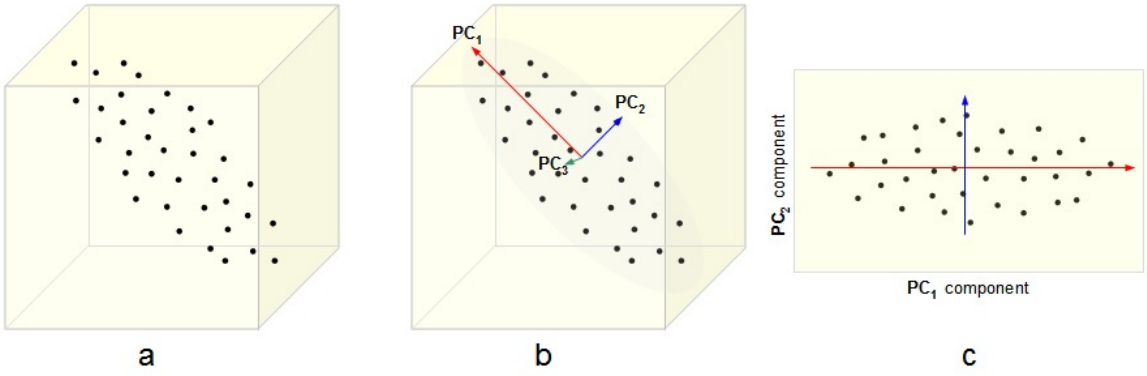


Figure 2.4: In PCA, data are projected into a lower dimensional space (in this case, from 3D to 2D), preserving the directions that are most significant for representing its variation. Note how the directions of the reduced space's axis system are orthogonal to each other, but not necessarily to the axes of the original space. CC-BY Lydia E. Kavraki, cnx.org/content/m11461/

A specific application of dimensionality reduction techniques to face descriptors that has received much attention in face recognition research is Turk and Pentland's "eigenfaces" [1991]. These are developed using the mathematical concepts of eigenvectors and eigenvalues, which are commonly associated with PCA, but can also be obtained with different dimensionality reduction techniques as well. The word "eigenvalue" comes from the German *eigenwert*, which means "characteristic, innate value". Simply stated, eigenvectors are vectors, that when paired to a given transformation, keep their direction. In other words, they are "aligned" with the transformation's direction. For example, when a rubber band is stretched longitudinally, an arrow drawn along its length would only be scaled, while one drawn diagonally would have a different angle after the transformation (see Figure 2.5).

The direction pointed by the first arrow would be an **eigenvector** of the stretching transformation, and the amount by which it was stretched would be its **eigenvalue**. The principal

components of a PCA can be considered eigenvectors, because to obtain the approximations of the original data points, they are linearly combined – that is, they are multiplied by scalar values, or, in this case, eigenvalues, and then summed together.

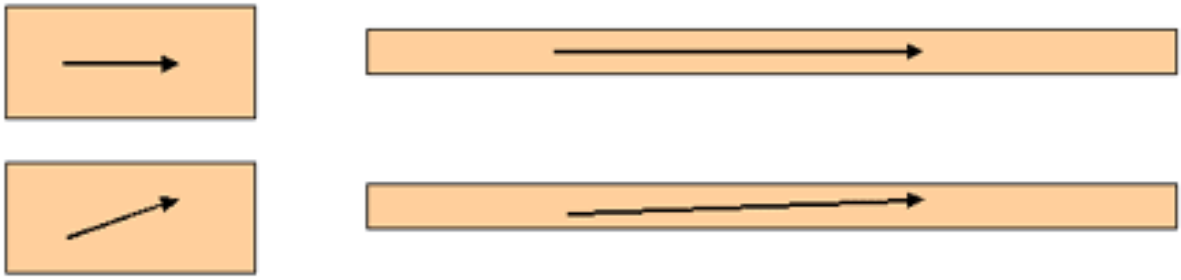


Figure 2.5: A simple demonstration of the concept of eigenvectors: in a rubber band that is expanded (stretched), an arrow drawn in the direction of the expansion keeps its orientation, changing only its value (magnitude). The arrow's direction can thus be considered an eigenvector for that particular stretching transformation, and the amount by which it was scaled is the eigenvalue of the transformation. © *Sally Riordan / PhysLink.com*

Chapter 3

Contextual overview

Automated face detection and recognition by computer vision is a field of research that has garnered much interest, especially in the security industry. As a result, many techniques were developed throughout the years, generating a need for surveys that allow researchers to gain a global perspective of the various approaches taken to the subject.

However, as computing power increases and more powerful applications become possible, the throughput of a growing number of research projects makes the surveys become obsolete or outdated much more quickly. This section thus aims to present an up-to-date, state-of-the-art survey of this research area, aiming both at comprehensiveness and comprehensibleness, and benefitting the latter over the former when necessary.

A historical overview will be first presented, to provide context and background. Also included is an exposition of the classification problem and the scheme adopted for this document. Then, the specific techniques will be presented, according to the taxonomy defined in the classification section.

Naturally, given the hypothesis this project aims to evaluate, special focus will be given in the analysis of 2D and 3D feature-based approaches that attempted specifically the measurement of rigid features of the face, as hints to reveal its underlying static structure, and using these for recognition. While the latter will be introduced in section 3.5, a more detailed analysis will be present in chapter 4.

3.1 Historical summary

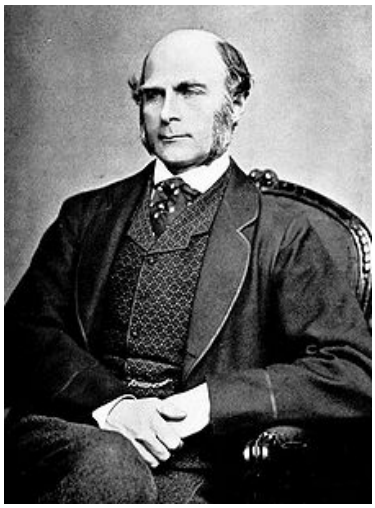


Figure 3.1: Sir Francis Galton conducted research on person identification using profile-based measurements, fingerprints, and other biometric data. Photo in Public Domain.

It is widely assumed that the first formal proposal for automatic face classification was made in [Galton, 1888], who investigated facial profile-based biometrics. The earliest implementation reported on the engineering literature is by Bledsoe [1964], describing work on automatic facial recognition in a mug shot database. But research on automatic machine recognition of faces really started in the 1970s [Kelly, 1970] especially after the seminal work of Kanade [1973]. All these early implementations used typical pattern classification techniques, in order to identify, measure and compare distances and relative positions of facial features. For most of the 1980s, research on face recognition remained essentially dormant. Since the early 1990s, though, research interest in the area has been consistently growing, with many developments, surveys and new applications being presented [Zhao et al., 2003].

The early face detection techniques could only handle images with a single face (or a few well-separated ones) in a frontal pose and with plain backgrounds. Hjelmäs and Low [2001] refers that in those systems, “any change of image conditions would require fine-tuning of the algorithms, or even a complete redesign”. More recent algorithms can already detect faces and their poses in cluttered backgrounds [Gu et al., 2001; Heisele et al., 2001; Schneiderman and Kanade, 2000; Viola and Jones, 2001].

With the recently increasing activity in face recognition research, several survey papers have been presented, aiming to summarize the developments that were being made. One of the earliest was [Samal and Iyengar, 1992], followed by [Chellappa et al., 1995], which presented a comprehensive survey of face recognition techniques at the time. Surveys targeting specific areas were also produced, such as [Hjelmäs and Low, 2001] and [Yang et al., 2002], whose focus was restricted to the face detection problem, or [Brunelli and Poggio, 1993], who evaluated the effectiveness of the two main approaches at the time: feature-based and template-based (refer to the next section for an elaboration on classification taxonomies). Finally, a very thorough general review was presented in [Zhao et al., 2003].

3.2 Classification taxonomy

Zhao et al. [2003] state that the wide range of technical challenges posed by the face recognition problem requires an equally wide range of techniques from several areas, having attracted researchers from very diverse backgrounds: psychology, pattern recognition, neural networks, computer vision, and computer graphics. For such a vast array of solutions presented, a classification system is required to better organize these approaches.

Several attempts have been made to organize and classify the systems developed by researchers working on the face recognition problem. The system most commonly applied in previous surveys consists in describing existing approaches as either **geometric**, or feature-based, or **photometric**, also commonly referred to as template matching, or image-based approaches. Face recognition is a high-level part of the human visual system, and the feature-based approaches put this in practice by using explicit high-level cognitive information to aid in the task, such as knowledge about the position and appearance of characteristic facial features (eyes, mouth, nose, etc.). Image-based approaches, on the other hand, tend to work on an implicit level, taking advantage of recent developments in pattern recognition theory to leave to the computer the task of deciding which parts of the image will be used to describe the faces. This allows them to avoid dependency in manually or heuristically defined features, while allowing measures that might be less intuitive but more accurate. For example, most early feature-based techniques do not work if the eye is closed or if the mouth is open [Zhao et al., 2003], but image-based approaches can be less dependent on these details. However, image-based methods are more sensitive to variations in illumination, camera viewpoint and face orientation.

It is worth noting that, for both these methods, it would be necessary to scan the whole image (aggravating to many frames per second when working over video), applying these techniques to sub-areas in every possible size, position and orientation, in order to find the faces. Faster, more efficient low-level pre-processing is thus applied; it is, of course, much more ambiguous, frequently returning target areas that aren't faces, but processing only these areas, rather than the full image, brings considerable performance improvements nevertheless.

Another classification approach commonly used was inspired in the study of the human visual system. Psychological and neurological research findings have, for instance, described a condition called prosopagnosia, where patients are unable to recognize previously familiar

faces, while having no other significant memory loss or cognitive process deficiency. They recognize people by their voices, hair color, dress, etc. A particularly relevant detail is that prosopagnosia patients do recognize whether a given object is a face or not, but are afterwards unable to identify who the face belongs to [Zhao et al., 2003]. This hints that face detection and face recognition are independent processes, executed separately by different parts of the brain, and indeed many surveys have used this criteria to distinguish existing systems.

We thus have two main axes to define the approaches to automated face processing: geometric/photometric, and detection/recognition.

But on one hand, a closer look reveals that the geometric/photometric division only defines pure techniques, while on practice these are almost always applied in conjunction, which makes it hard to classify existing systems in either one or another category. Hybrid approaches such as Active Appearance Models (see subsection 3.4.5) are naturally even harder to classify. Zhao et al. [2003] do recognize this: "Often, a single system involves techniques motivated by different principles. The usage of a mixture of techniques makes it difficult to classify these systems based purely on what types of techniques they use for feature representation or classification". Nevertheless, they provide a comprehensive listing of techniques used for face recognition [Zhao et al., 2003, Table III].

On the other hand, the detection vs. recognition axis, as shown in chapter 2, depends mainly in the comparison threshold between new faces and stored ones. This classification is therefore subjective and relates to the purpose of each system rather than the potential of the techniques it uses.

While no system is perfect for the classification purpose, as demonstrated above, a compromise solution, aimed at serving the objectives and contents of this document, was devised as follows:

1. low-level face detection/location routines (section 3.3);
2. higher-level, still-image-based face detection/recognition methods (section 3.4);
3. 3D approaches (section 3.5).
4. adaptations to video streams (section 3.6);

3.3 Face location

A low-level face location system is needed to identify areas of the image that should be processed by more advanced (and expensive) face detection techniques (both feature- and image-based). This pre-processing also helps in normalizing the faces' pose (canceling their transformations in position, scale and rotation), but their main advantage is improving performance of the system, since these early techniques are relatively simple and fast to apply. These filters build upon basic routines called morphological operators, which have long been used in image processing, and very efficient techniques have been developed to detect a wide range of image features. For a thorough introduction to morphological operators, see [Gonzalez and Woods, 1978].

Usually, coarse operators are used to detect heads, and then operators meant to detect facial features can be applied to these areas in order to either reinforce the confidence that the candidate areas actually represent faces, or to remove false matches, before stepping into more advanced face detection algorithms. This "early exit" approach helps speeding up the facial detection process, by interrupting the processing (and thus preventing unnecessary analysis) very early in the detection pipeline.

For these filters, multi-resolution search are often used. This not only speeds up the process, but also is more robust against image artifacts that might hinder the functioning of these operators. Also weighting in favor of this approach are several studies [Ginsburg, 1978; Harmon, 1973] which have concluded that information in low spatial frequency bands plays a dominant role in face detection. Sergent [1986] further demonstrated that low-frequency components are useful to the generic face detection process, while high-frequency components contribute to the finer details needed in identification [Zhao et al., 2003].

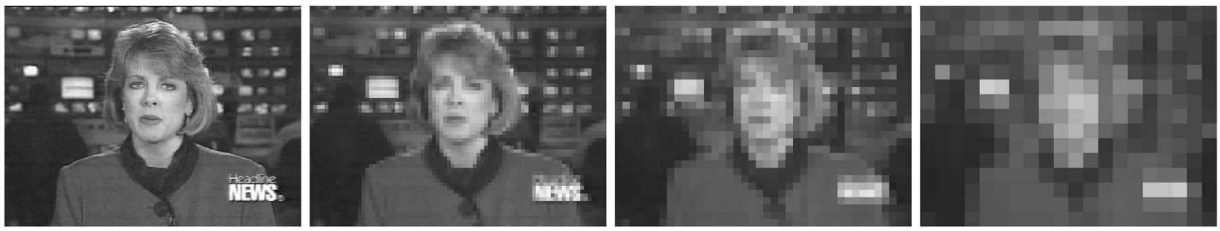


Figure 3.2: A full-resolution image is gradually downsampled to lower resolutions. Search normally starts at the lowest resolution (at right) and is progressively refined in increasingly higher resolutions. Reprinted from [Yang et al., 2002]

Besides location of faces and facial features, morphological operators can also be very useful in the image normalization process. For example, global gradient direction detection can provide a way to reduce lighting variations. Another useful application of these filters is augmenting contrast by adjusting the image's gray-levels histogram. Several normalization/ equalization algorithms exist; see [Gonzalez and Woods, 1978] for a detailed overview.

Low-level processing techniques

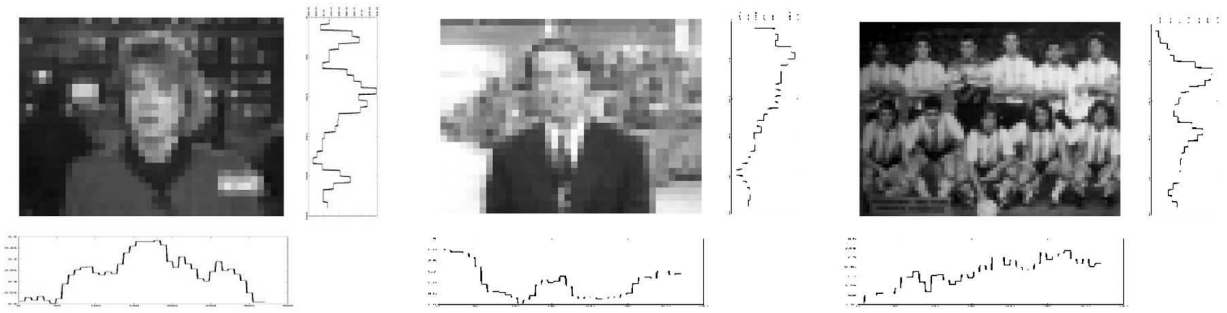


Figure 3.3: Horizontal and vertical histograms of images might contain characteristic signatures that allow the location of faces. However, more complex examples present greater challenges. Reprinted from [Yang et al., 2002]

One of the earliest techniques for automatic location of faces was the conversion of the image into a black-and-white (binarized) representation, and then analyzing the signature of the binarized image's histogram across the horizontal and vertical directions – that is, the sum of black pixels across the rows or columns of pixels (as opposed to the usual image histogram representation, which displays the amount of pixels in the whole image for each gray-level intensity). As depicted in Figure 3.3, the location of the face can be estimated by searching for specific patterns of local maxima and minima (peaks and valleys) in the horizontal and vertical profiles. From the initial location candidates, secondary patterns are applied to look for the expected signatures of features such as eyes and mouth (which will typically be darker areas). Locations that present similar patterns to the expected signatures indicate a good match probability. This technique was used, for example, in [Kanade, 1973] and [Turk and Pentland, 1991].

However, as can be inferred from the second and third images of Figure 3.3, this method has difficulty detecting faces in complex backgrounds or multiple faces. Other approaches

were presented, using edge detection. These edges are then analyzed to find facial features [Sakai et al., 1972] or to match the typical human head outline [Craw et al., 1987]. Brunelli and Poggio [1993] proposed the use of gradients, which convey direction information. Horizontal gradients are useful to detect the left and right boundaries of face and nose, whereas vertical gradients are useful to detect the head top, eyes, nose base, and mouth. An illustration of their technique is presented in Figure 3.4.

Yet another technique that has been used in several implementations is a morphological operator that matches pairs of dark circles, in order to find the eyes – one of the most distinctive facial features, for their typical dark appearance surrounded by lighter areas, and for their symmetry. The eye candidate positions are then used for triangulation aimed at obtaining the expected position of other features such as the mouth or the nose, where further analysis to confirm their existence is performed. The locations of these features relative to each other can then be used to estimate the facial pose and apply the inverse transformation to line up the face into a frontal view.

Figure 3.4: Horizontal and vertical components of a face edge map. Reprinted from [Brunelli and Poggio, 1993]

Figure 3.5: A grayscale image and the result of applying a symmetry detector on it. Reprinted from [Reisfeld and Yeshurun, 1995]

The symmetry of eyes and other facial features has been further exploited with the use of symmetry detectors [Reisfeld and Yeshurun, 1995], having achieved notable accuracy in low-level face detection. The result of one of these detectors is shown in Figure 3.5.

Maio and Maltoni [2000] put forth an implementation using a gradient-type operator over local windows (7×7 pixels) to create a binary image with multi-directional edge information. They then applied a two stage face detection process by first using a Hough transform that detects oval shapes, and then a set of 12 binary templates representing face

features which are matched against the face candidates generated by the Hough transform. Since these are all low level processing techniques, they were able to run the system in real time while achieving very good detection ratio: they report correct face location in 69 out of 70 test images with no false alarms, using test images with faces of varying sizes and complex backgrounds. Figure 3.6 depicts the different steps of this algorithm.

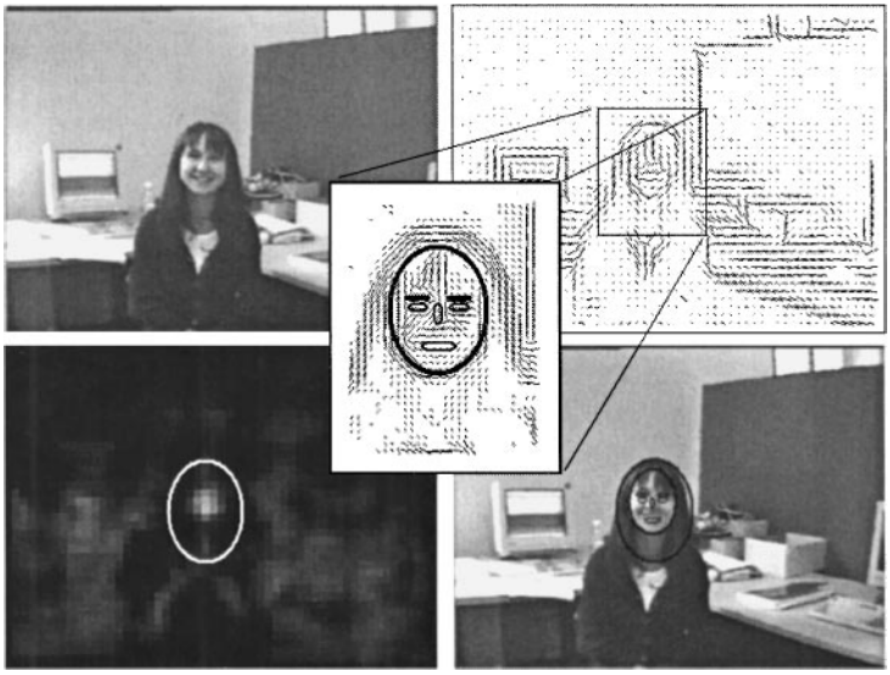


Figure 3.6: A visual summary of Maio and Maltoni's implementation, which reportedly achieved a 99% performance with real time execution. Reprinted from [Maio and Maltoni, 2000]

When using video, the temporal variation also provides raw information that can be used to help in this low-level phase of the facial detection pipeline. A more detailed overview of video-based facial detection is presented in section 3.6.

3.4 Face detection and recognition

With a reasonable face location candidate, more complex methods are necessary to confirm matches and declare a high-confidence result. This is needed because low-level techniques often fail when the appearances of the features change significantly (for example, closed eyes, eyes with glasses, open mouth, etc.).

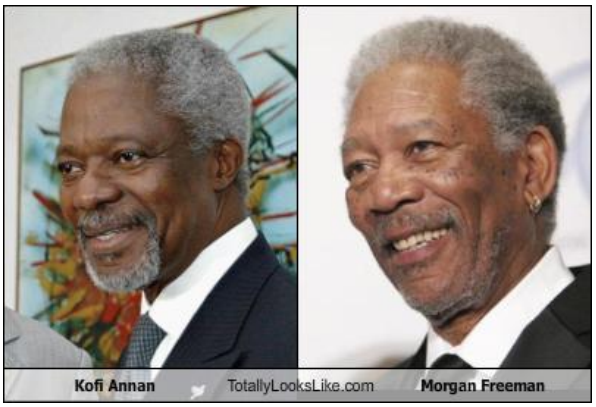


Figure 3.7: It is important that a system be able to distinguish between very similar people, especially in security applications. ©TotallyLooksLike.com

Several systems have been presented since the first research efforts on the area. An overview of the most relevant methods is presented in the following sections. For more detailed listings and descriptions, refer to the surveys mentioned on section 3.1.

3.4.1 Eigenfaces

One of the first applications of PCA for face analysis was made by [Sirovich and Kirby, 1987], who attempted a system for compact face representation. [Turk and Pentland, 1991] extended this principle for face recognition. The general concept lies in using the principal components, produced by a PCA applied on the faces, as eigenvectors for linear combinations which would yield close approximations of the original faces. Since these components are vectors expressed in the same space as the faces (and thus have the same dimension) they can also be displayed in a pixel-based representation, resulting in an image that Sirovich and Kirby called an eigenpicture, and Turk and Pentland called an eigenface.

Eigenvectors represent the most relevant features, which are devised statistically, without human intervention, and thus may not be intuitive features such as eyes, wrinkles, etc. The typical appearance of eigenfaces is depicted on Figure 3.8. It is possible to find a set of eigenvectors that are influent enough to provide an approximation to all the possible vectors representing faces (under approximate imaging conditions), because these should vary in a limited region inside the high-dimensional space they are represented on.

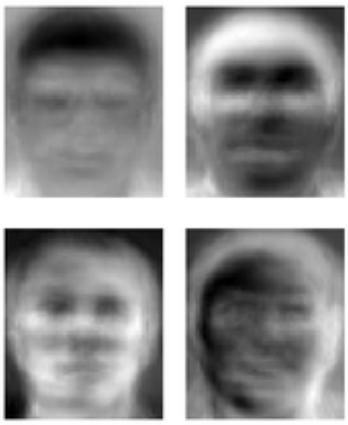


Figure 3.8: Examples of eigenfaces. © AT&T Laboratories

An eigenface deviates from average gray in areas where average variation among the images is more pronounced. Each face in the database from which the eigenfaces are derived can be represented exactly as a combination of all the eigenfaces extracted. Mathematically, this is a linear combination, that is, a sum of all eigenfaces, each multiplied by specific eigenvalues (weights). The faces from the database (and others) can also be *approximated* using a subset composed of only the best eigenfaces (those that have largest eigenvalues). This reduces representation size and matching complexity with minimal loss of descriptive features.

A new face can thus be recognized by checking that the weights for each eigenface which describe a new face image are consistent with the collection of weights stored in the database for a given face. Calculating a weight distribution for a new image is equivalent to projecting it into the n-dimensional "face space" of n eigenvectors (eigenfaces). Faces whose weights are not close enough to any of the recognized faces' values are stored as new faces. If a subset of these new faces clusters in a region of the face space (that is, they don't deviate from their local mean more than a given threshold), it is assumed that they are the same face and a new entry (the average face of that subset) is added to the database of known faces.

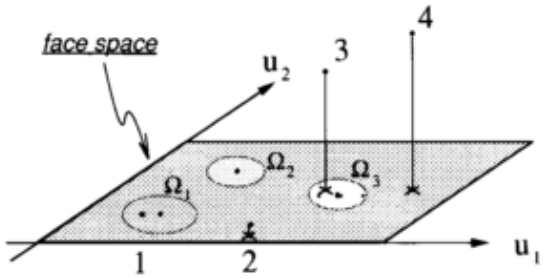


Figure 3.9: A simplified depiction of the "facespace" and the thresholds used for defining a match. Reprinted from [Turk and Pentland, 1991]

The construction of the eigenface set is computationally intensive but shouldn't need to be frequently updated, and can be done offline or in parallel with the execution of the system. For each candidate image, calculating its projection into the reduced eigenface space is marginally intensive (Barrett [1997] reported a few seconds). The matching is highly efficient and according to Turk and Pentland, for a database of up to a few hundred faces, it could be done at the frame-rate of a video camera – that is, in real-time using a video stream.

3.4.2 Fisherfaces

Belhumeur et al. [1997] describe an approach for face recognition that is much less sensitive to large variations in lighting and facial expressions, based on the principle that all of the images in a Lambertian surface[1], taken from a fixed viewpoint but under varying illumination, lie in a compact 3D subspace of the high-dimensional image space [Shashua, 1994].

The eigenface method, which uses PCA for dimensionality reduction, yields projection directions (eigenvectors) that optimize the separation of the face images in classes according to the global variation in the images. This means it also includes variations which are due to lighting and facial expressions, which may result in the clustering of different images of the same face into separate classes. In fact, as stated by Moses et al. [1994], variations between the images of the same face due to pose and lighting are almost always larger than the differences between the faces of two different people taken under similar conditions (see Figure 3.10).

It has been suggested that the variation due to lighting can be reduced by discarding the most significant components (eigenfaces). However, it is unlikely that these correspond only to lighting variations; therefore, removing them might result in the loss of information that could be useful for between-class discrimination.

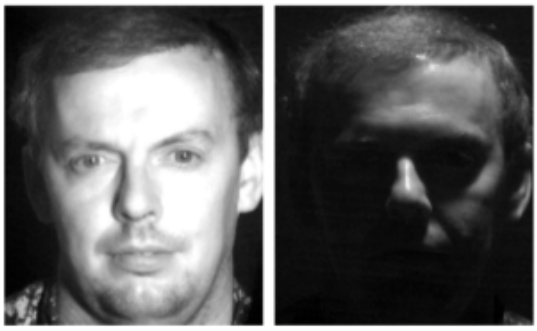


Figure 3.10: The same face may look radically different due to pose and lighting. This affects the performance of image-based face recognition algorithms. Reprinted from [Belhumeur et al., 1997]

Instead of PCA, the authors use Linear Discriminant Analysis (LDA), first developed by Robert Fisher in 1936 for taxonomic classification [Fisher, 1936] (hence the term “fisherface”). In a comparison with Shashua’s correlation measure and Turk and Pentland’s eigenfaces, the authors claim that fisherfaces presented better results, with lower error rates and faster execution. Still, the accuracy of the fisherface method is hindered by variations in images such as facial expressions, different poses and illumination artifacts (self-shadowing, bright spots and subsurface scattering, for

[1] A Lambertian surface reflects light with the same intensity in all directions. These surfaces are only theoretical. In practice, no physical surface is purely Lambertian [Wikipedia::800155]. Human skin, for instance, is close to a Lambertian surface, but also reflects light in other forms, for example through a model called subsurface scattering (SSS), where light exits the surface through a different point than the one it entered, after multiple internal reflections (scattering).

instance). Further work using LDA has been reported by Etemad and Chellappa [1997] and Zhao et al. [2003].

3.4.3 Neural networks

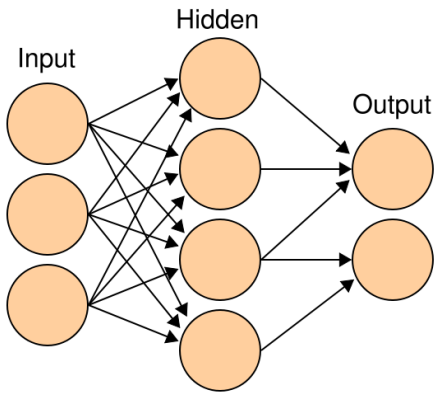


Figure 3.11: A simple neural network. CC-BY-SA Cburnett commons.wikimedia.org

Neural networks (computational systems inspired on the natural brain structure) have been known to yield good approximations to complex problems that defy a deterministic description. With faces described as points in a high-dimensional space and high variation for images of the same face, the problem of face recognition is thus suited to neural network analysis.

Each pixel of the face images is mapped to one input neuron. The intermediate (hidden-layer) neurons are as many as the number of reduced dimensions that are intended. The network "learns" what patterns are faces or not using the backpropagation adjustment method, which consists having the weights for each neuronal connection updated after each iteration, according to the error in the final result and their contribution to it.

This approach has produced promising results, but according to Cottrell and Fleming [1990], they can at best be comparable to an eigenface approach.

3.4.4 Gabor wavelets

First proposed by Dennis Gabor [Gabor and Stroke, 1968] as a tool for signal detection in noise, wavelets can be constructed as continuous waves modulated by a Gaussian envelope – that is, harmonic functions multiplied by a Gaussian function (commonly called "bell curves"). This means that they assume the global shape of a signal whose intensity gradually increases from zero to the maximum, and then decrease in the same way back to zero, while maintaining its local fluctuations (see Figure 3.12).

Applying the same principle that is used in Fourier series, an image is decomposed in a series of wavelets, each with distinct parameters, which when superimposed to each other

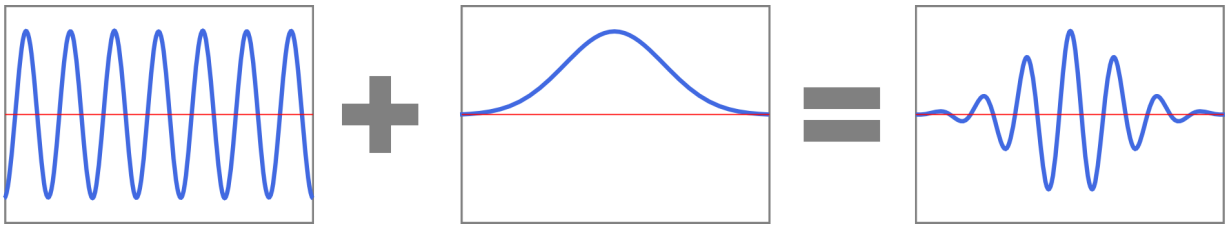


Figure 3.12: Wavelets may be constructed as cyclic waves modulated by a Gaussian envelope. This example was generated using the equations $\sin(x+\frac{\pi}{2})$ *for the continuous waveform,* $e^{-\frac{x}{10}^2}$ *for the Gaussian modulator, and a multiplication of both for the resulting wavelet. The range sampled was* $[-7\pi, 7\pi]$ *(i.e., 7 full cycles).*

reconstruct the original pattern. This allows a very compact representation, since it is only necessary to store information regarding the frequency, length and directions of the wavelets. This decomposition is done locally, in specific points of the image, resulting on several sets of superimposed wavelets (called "jets") that represent their region. The number of wavelets used in each jet determines the precision of the approximation to the original image. The full set of points where wavelets are calculated is called a grid, even though they need not be evenly distributed in the image.

The wavelet approach has been further extended [Lades et al., 1993] to models with flexible grids, that can be slightly transformed in order to match faces with different poses from the original, thus allowing a better performance in detecting faces that are not facing the camera (but only to a certain extent — a problem common to all 2D approaches). This method is commonly called the elastic grid matching approach. It is worth noting that while fixed grid wavelet matching is comparable in performance to the eigenface approach, elastic grid matching will be relatively slower.

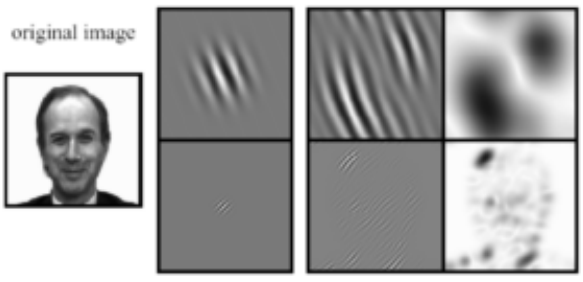


Figure 3.13: An image and a subset of the wavelets it has been decomposed into. Reprinted from [Wiskott et al., 1997]

3.4.5 Active Shape Models and Active Appearance Models

The first approaches to face location based on active shape contours were presented by Kass et al. [1987], who called them "snakes" – deformable curves that attempt to minimize "energy" in pixel intensity local gradients. [Yuille et al., 1987] extended the approach to deformable templates (sets of geometrically related facial features) achieving better performance.

Active Shape Models (ASM) use the same principle of matching geometric templates to boundaries (edges) in the target image. The difference is that statistical analysis is used to model and restrict the variation of the points that define the template. PCA is applied in order to extract the eigenvectors that describe variation of the models for each face from the average template. Each face (or generically, shape) is represented as a vector with as many components as the number of points it contains.

The elasticity of the model is an important component of this approach. Balance must be achieved between a general approach – that is, being able to generate, from the stored models, any plausible example of the class they represent (in this case, faces), in order to recognize new faces – and a specific approach that avoids generation of illegal examples [Cootes et al., 1995].

The average model for a given class of objects (such as a specific face) is built by marking control points in each image of that class in the database and then warping the images so that their control points match their average positions across all images. From the shape-normalized images, grayscale intensity over the region around each shape point is sampled. To minimize the effect of global lighting variation, the samples for each point are normalized in intensity, so that they spread across the full range of possible intensity values even if that specific area of the image is darker than the areas around other points. Shape and appearance of any image in the database can then be rebuilt using only the average image and the variation parameters (eigenvectors + eigenvalues).

The model fitting process starts with the mean model shape, but on each iteration the points of the model are adjusted within their valid range in order to attempt a better fitting into the image being processed. Eventually, the parameters of the model converge to the values that make it adjust to the new image, even though potentially none of the samples in the database had that specific configuration. This process is what allows applications such as face recognition and tracking.

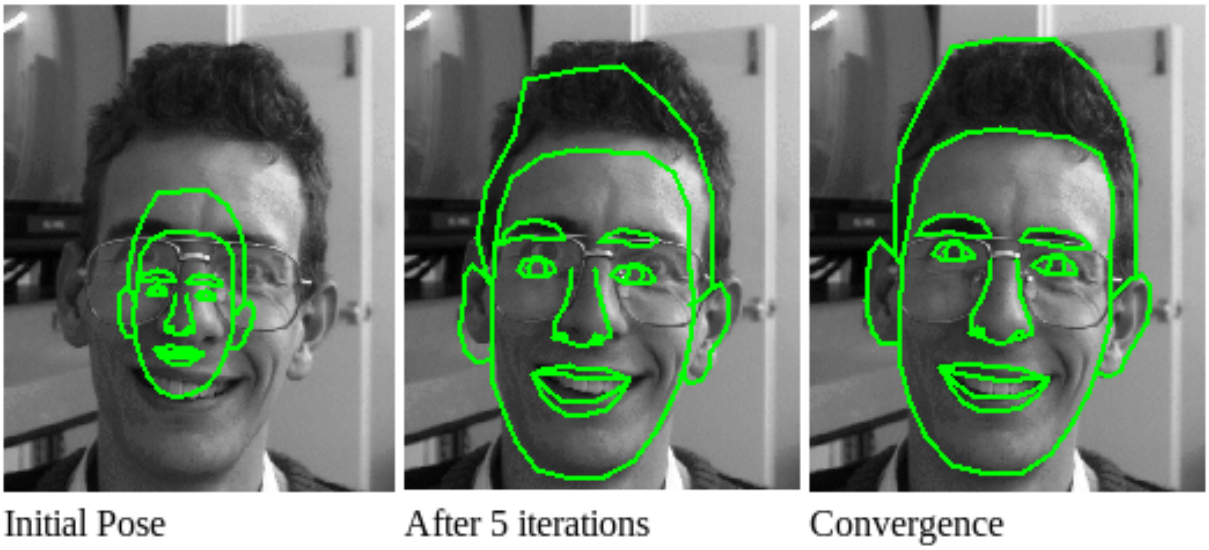


Figure 3.14: An example of a model being iteratively adjusted to fit a new face. Reprinted from [Cootes et al., 1995]

To avoid the application of 2D pattern matching for each point's neighborhood, a simplified approach that has been proved much more efficient while still effective is comparing the intensity variation across a line normal (perpendicular) to the model's shape in that point. The intensity values are normalized to reduce the influence of global intensity changes. Comparison is made by sampling such an "intensity profile curves" from the target image using a longer line (that is, fetching a set with more values than those the stored samples contain), then "sliding" the model point's average curve across the fetched values' curve, and recording the position that best fits (that is, the one where the sum of differences for each pixel value on the curve is the smallest). This is repeated for every point in the model, giving a suggested new position for each point in every iteration. The variation parameters are then updated, with constraints that guarantee values within the expected range.

One weakness of this approach is that it depends on a relatively good initial positioning of the model, since adaptations are only performed locally for each point. The process could go wrong if the model is placed too far from the actual feature it is attempting to match. To improve the efficiency and robustness of the algorithm, a multi-resolution approach is applied, starting with low resolution images and progressively advancing to higher resolution ones. This allows faster convergence and reduces chances of the model getting trapped in wrong parts of the image.

Per the description above, it is clear that the ASM method uses characteristics of both

feature-based and image-based approaches. Furthermore, the ASM paradigm has been expanded to statistical appearance models, such as the Flexible Appearance Model (FAM) [Lanitis et al., 1995] and the Active Appearance Model (AAM) [Cootes et al., 2001]. An AAM is essentially an ASM with texture superimposed. The system can generate a synthetic image by manipulating the parameters of shape and appearance variance, and then match this synthetic image with the target image.

3.5 3D approaches

2D approaches have serious drawbacks in coping with pose variations. Several techniques can overcome in-plane rotation (head tilt). But when there is out-of-plane rotation, also called "rotation in depth" (head turning, or nodding), even though the transformations can still be calculated and reversed in the less acute cases, most 2D algorithms begin to fail as the rotation angles increase. A solution to this problem that has demonstrated some potential is the use of several poses for each face, with interpolation occurring to match intermediate poses. However, this increases the amount of storage needed for the image database, as well as the complexity of the system – and yet the results are not robust enough for application in more demanding systems.

Thus, several approaches have recently surfaced that use 3D estimation from the 2D images. As stated by Bronstein et al. [2005],

> "Three-dimensional face recognition is a relatively recent trend that in some sense breaks the long-term tradition of mimicking the human visual recognition system, like the 2D methods attempt to do. As evaluations such as the Face Recognition Vendor Test (FRVT) demonstrate in an unarguable manner that current state of the art in 2D face recognition is insufficient for high-demanding biometric applications [Phillips et al., 2003], trying to use 3D information has become an emerging research direction in hope to make face recognition more accurate and robust."

These proposals present greater robustness to rotated faces (for example, profile views). Common techniques include the use of stereo vision [Lao et al., 2000] or motion analysis (the video-based approach is described in section 3.6). Still, Gordon [1991] states that

many features of the face are difficult to detect or measure because of variability of lighting conditions or low dynamic range in the input image. However, areas of the face which are difficult to describe with standard intensity based methods can be extracted with specialized data acquisition systems, which are described below.

Early work on 3D recognition using 3D sensors was performed by Cartoux et al. [1989]. Later on Gordon [1991] reported on research using range data (depth maps). These maps are obtained using active sensors called range cameras. The sensors are called "active" because as opposed to a typical (passive) camera that merely captures incoming light, they actually emit radiation whose reflection they capture. A typical implementation of these systems consists of a scanner that emits laser rays and computes the distance to each point using the delay of the "echo" and the known speed of light. As it is evident, this approach suffers from the higher cost and lower availability of the specific equipment (3D scanners) they require, compared to the ubiquity of photographic and video cameras.

Figure 3.15: Example of a 3D model of a face, built using a range scanner. Reprinted from [Bowyer et al., 2006]

Another common approach is the "structured light", which consists in projecting a grid on the surface that will be scanned, and then acquiring one or more images in order to extract, from the distorted grid, the topology of the surface. This hybrid (active+passive) approach may reduce the cost of 3D data acquisition, but still shares the disadvantages of intrusion and lack of scalability with the remaining systems described above. They might also not be as robust to illumination variance as other 3D approaches. Bowyer et al. [2006] describes a popular misconception in the literature about 3D face recognition, which he calls *"The myth of 'illumination variance'"*:

> "It is often asserted that 3D is, or should be, inherently better than 2D for purposes of face recognition. One reason often asserted for the superiority of 3D is that it is 'illumination independent' whereas 2D appearance can be affected by illumination in various ways. It is true that 3D shape *per se* is illumination independent, in the sense that a given 3D shape exists the same independent of how it is illuminated. However, the sensing of 3D shape is generally not illumination independent—changes in the illumination of a 3D shape can greatly

> affect the shape description that is acquired by a 3D sensor. The acquisition of 3D shape by either stereo or structured-light involves taking one or more standard 2D intensity images[, which] are typically taken with commercially available digital cameras."
>
> – Bowyer et al., 2006

Despite all these different methods used to obtain the 3D data, the most common approach for recognition has been to transform (and some cases deform) the 3D model, in order to render the result in 2D for matching with new images either by the location of facial features or by image-based comparison. The latter naturally requires that the model be texturized, which consists in "mapping" a 2D image of the face into the 3D model [Blanz and Vetter, 2003].

Even though most approaches have followed a hybrid model (true 3D for data acquisition, 2D projection for comparison with input images), some research has indeed focused in comparisons directly on 3D data using geometric measurement. chapter 4 features a closer look at these methods, as they provide specific background for the current project.

For more detailed surveys of recent developments in 3D face recognition, the interested reader can consult [Bowyer et al., 2006, Table 1], which provides a comprehensive survey of algorithms developed and their performance. Another detailed paper summarizing developments in 3D face recognition was presented by Scheenstra et al. [2005].

3.6 Video

Video-based face recognition techniques typically have to face lower quality images, since the frames must be compressed for transmission or storage. They do, however, have the advantage of temporal information. This can (and should) be used to compensate for the loss of spatial information. For instance, a high-resolution image can be reconstructed from a sequence of lower-resolution video frames and used for recognition.

Another method made possible by the temporal variation is the calculation of frame difference (simply subtracting differences between each frame at pixel-level) over which temporal and

spatial analysis is applied. A simple example is searching for oval shapes in the frame-difference image, to detect the head blob, which allows cropping and scaling the face from the image, for posterior processing. This approach was used in [Turk and Pentland, 1991]. More advanced techniques include analysis of the moving image contour, or the optical flow [Lucas and Kanade, 1981].

Going even further, there are methods that track the face in the video after locating it, and use the relative motion of its points to reconstruct the 3D shape of the head. This approach is called Structure-from-Motion (SfM) and, as described in section 3.5, can greatly increase face recognition performance in uncontrolled environments. Recently, many results in face tracking have been presented, which use a flexible 3D model to track faces in video including large variations in pose and expression [Dornaika and Davoine, 2004; Lefèvre and Odobez, 2009].

A serious drawback of SfM is the lack of accuracy in the recovered 3D shape, due to the typical low resolution of video frames, especially in surveillance applications. The lack of accuracy may not hamper the face detection task, but it is quite harmful for face recognition, which must differentiate the 3D shapes of very similar objects. This can be countered by combining approaches. For example, a possible solution proposed by Zhao et al. [2003] is the use of Shape-from-Shading (SfS), that is, using illumination information to recover 3D shapes. An early implementation was in fact presented in [Zhao and Chellappa, 2000]. In chapter 4 this problem (when applied to still images or single video frames) will be addressed in more detail.

Chapter 4

Methodology

The main hypothesis this research project aimed to investigate was the informational value of rigid facial geometry for the problem of facial recognition, with the assumption that the various ratios, dimensions and angles between the rigid features of the face are sufficiently unique to differentiate a sizeable portion of the population.

This notion is in fact quite old. The first scientific document to propose such an approach to face recognition was authored by Matthews [1888], and is quoted below:

> "It is generally held that men have attained their natural stature about the age of twenty-four. [...] During subsequent life—excepting from the loss of teeth, which would deduct proportionately from the depth of the chin—no appreciable change in the osseous fabric can be theoretically assumed.
>
> Externally, appearances may differ much; but in that case the issues raised are mostly those of beard or no beard, fat cheeks or lean cheeks, blotches, wrinkles, and crows' feet; or no blotches, wrinkles, and crows' feet. Now, without doubt, fatness in lieu of leanness produces a very perceptible difference to the eye, and with superficial observers might invalidate the admeasurement. But [...] [w]ithin the boundaries of the area measured, the same identical proportions and distances subsist between the eyes, the lips, and the chin."

In fact, Matthews even claims [Matthews, 1884] his efforts predate Galton's method of facial image superposition (see section 3.1), having devised an instrument which he called "identiscope" (based on a *camera lucida*[1]) for performing the measurements he proposed.

[1] A *camera lucida* (Latin for "light room", as opposed to the common photographic *dark room*) is an optical device usually employed

This early example embraces concepts that resonate with basic intuitions regarding the automation of face recognition. As such, when computers started to become widely available in academic and research centers, in the 1960s, this was also the method used by the first approaches toward computer-based face recognition [Brunelli and Poggio, 1993], such as [Bledsoe 1964; Kelly 1970 and Kanade 1973].

However, all these approaches took into account both rigid and non-rigid points of the face (including Matthews' own implementation!, see [Matthews, 1884]), therefore invalidating the very premise they were set in. It is, of course, understandable why they chose to do so, since (citing Gordon [1991]) "even the highest contrast features of the face, such as the eyes, are a challenge to identify and describe reliably. Low contrast features such as shape of jaw boundary, cheeks, and forehead are currently impossible to describe from general intensity images". Too few rigid key (i.e., fiducial) points can be reliably detected from 2D images: essentially the eyes' corners and pupils, and some features of the nose. By including more easily detectable (but non-rigid) points of the face, such as mouth corners, and the apparent face boundary (rather than the true jaw boundary), the information stored for each face was richer, but eventually only useful for cases where the faces didn't feature variations that impact these points' locations. The position and orientation are part of these variation factors, and were accounted for with sophisticated cancellation methods (see aside) whose usage traces back to [Bledsoe, 1964]. But other factors

---

An aside: Gordon [1991] mentions a study by Harmon [1973], which at the time of Gordon's writing represented "the sole example of a system which explicitly [incorporated] shape information into the task of face recognition". Harmon's system was not fully automatic: the tagging was performed by a group of human subjects, who identified 21 qualitative facial features from photographs, including forehead, cheeks, and chin. The faces were encoded by assigning a value from 1 to 5 to each feature. For example, for the forehead and chin, these values corresponded to the range from "receding" to "bulging", and for the cheeks to the range from "sunken" to "full". After inserting the data into a computer and calculating the similarity between the faces, this technique showed a relatively good success rate, thus reinforcing the idea of comparing facial features for the recognition task. However, despite this and other attempts using features marked by humans (e.g., [Cox et al., 1996]), Brunelli and Poggio [1993] state that "features are only as good as they can be computed."

---

as a drawing aid by artists. It performs an optical superimposition of the subject being viewed upon the surface where the artist is drawing. This allows the artist to duplicate key points of the scene on the drawing surface, thus aiding in the accurate rendering of the subject [Wikipedia::339562].

which wouldn't be canceled, referring to local shape variation, include:

- occlusions caused by hair style or significant facial hair changes (e.g. long/thick beards) or accessories (e.g. glasses);
- weight gain/loss;
- facial expressions.

This meant that overall, however well-founded, this strategy kept the recognition potential of these systems from reaching consistent reliability in non-controlled environments. As such, they were quickly outnumbered by the image-based methods, which make use of a richer[4] representation of faces, based on the intensity of pixel values. This change of focus became especially evident after the groundbreaking work by Turk and Pentland [1991]. The few rigid points reliably detectable in images were enough to perform normalization of position and rotation, while the comparison itself was done in the much larger (see chapter 2) space of image intensity values.

Despite achieving generally better results than the previous distance/angle-based approach (Brunelli and Poggio's well-known comparison from 1993 reported 90% recognition accuracy with geometric methods and perfect performance with image-based template matching), this general approach was not without faults. Two main factors are at the root of its weakness.

As the field of photography matured, some approaches for geometrical analysis of photographies based on feature points were developed, such as photogrammetry, which consists in determining the geometric properties of objects from photographic images. Such techniques, whose principles were used in some of these approaches, to normalize the different faces for comparison, were eventually expanded into computer implementations that embody the pose cancellation method used in virtually all recent 3D face recognition; namely, the definition of the correspondence problem[2], algorithms for bundle adjustment[3], and the generic specification of the 3D pose estimation problem.

The first one is that faces are three-dimensional objects, while a 2D image of a face is merely its projection into a lower dimensional space. Like the dimensionality reduction techniques

[2]The correspondence problem consists in to identifying which parts of an image correspond to which parts of another image, after the camera has moved, time has elapsed, and/or the objects have moved around [Wikipedia::6498435].

[3]Given a set of images depicting a number of 3D points from different viewpoints, bundle adjustment can be defined as the problem of simultaneously refining the 3D coordinates describing the scene geometry as well as the parameters of the relative motion and the optical characteristics of the camera(s) employed to acquire the images [Wikipedia::13754920].

[4]For instance, visual appearance might also contain quite unique clues that can't be represented by numeric parameters, unless a classification system and methods for automatically detecting them are devised. These include scars and skin marks such as spots, naevi (moles), etc.

described in chapter 2, this is a "lossy" transformation, which means that recovery of the original shape is only partially possible, as an approximation. The reason why the vast majority of face photos are taken in a frontal pose is that such a projection is the one that minimizes the distortions to the principal components of the face shape (assumed, for simplicity's sake, as the x and y axes), ignoring the least influential component (the z axis, which contains depth variations); this an intuitive application of the PCA concept: minimizing the inevitable loss information due to the projection.

Nevertheless, even though pose cancellation methods can be used by employing high-level previous knowledge of face geometry (for instance, the assumed horizontality of the eye-to-eye axis), with larger rotations more serious distortions, or even occlusions, might occur, therefore yielding an incorrect normalized image for comparison. Some proposals to counter this effect include the use of images for several poses, encoding techniques that minimize illumination differences (see subsection 3.4.2) and exploiting the vertical symmetry of the face to correct the half of the face that deviates more from the image plane. It should be noted that the latter can to some degree counter out-of-plane rotation in the horizontal direction (i.e. head turning) but not on the vertical (i.e. nodding). Another effect of this conversion is that illumination direction might cause shadows in the face. Research on how to overcome this problem has been carried with good results [Zhao and Chellappa, 2000; Belhumeur et al., 1997; Georghiades et al., 2001], but edge cases are still problematic.

The second factor that influences image-based facial recognition is the set of parameters that might change the appearance of the same person under different configurations (even all above conditions being equal). Illumination intensity and color affect the appearance of the skin, and so does skin tone variation (e.g. from tanning). These issues can be reduced to some degree with intensity normalization and image grayscaling. But other modifiers, which are harder to cancel, include:

- light facial hair (e.g. mustache, short beard)
- makeup
- aging
- disguises

These problems are actually more prevalent in image-based recognition techniques than in geometric ones, since for the latter only the fiducial points' location is used for comparison, rather than the texture of the areas between them. Adding to this, the same weaknesses

attributed the geometric methods described above (namely, the need to cancel local shape variations) arise in image-based approaches from allowing flexible deformations to the templates. These apparent drawbacks are outweighed successfully by the reliance in the visual representation of the face, which is richer in information, but as was seen above, the accuracy of such an approach has an intrinsic upper bound in uncontrolled environments.

Not only the computer's weaknesses in using a 2D representation of faces were evident in image-based approaches, but even human recognition revealed flaws derived from the near-2D nature of human vision[5]. According to Gordon [1991], "it is unlikely that humans base their representation or comparison of shape on the accurate perception of depth". Bronstein et al. [2005] goes further, to suggest that even though simple texture mapping creates natural-looking faces, the individuality of the subject concealed in the 3D geometry of his face is completely lost. As a demonstration, they presented the result (see Figure 4.1) of different textures mapped into a single 3D face shape, yielding 2D appearances that could, according to them, deceive any 2D face recognition method.

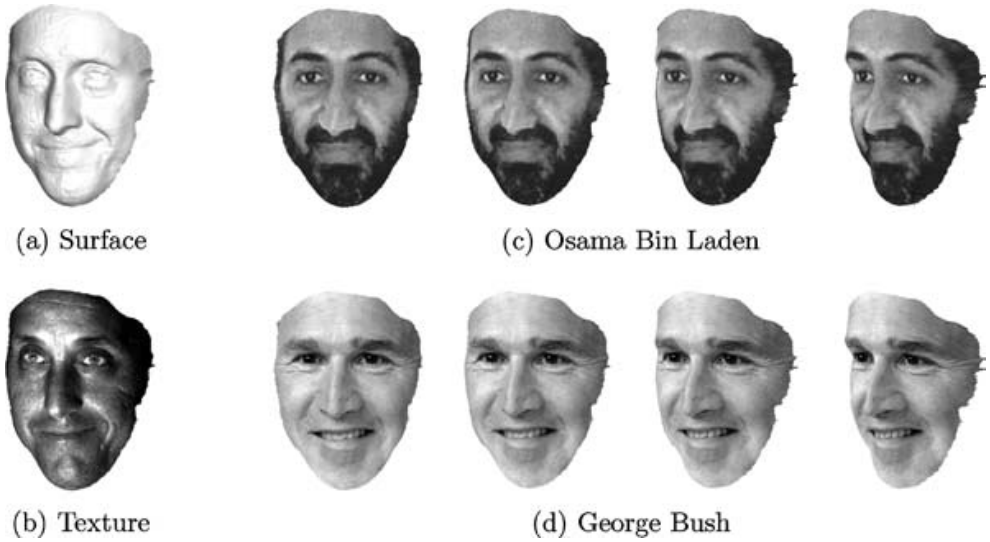


Figure 4.1: Bronstein et al. [2005]'s demonstration of the importance of 3D shape which is concealed in 2D images. Through simple texture mapping, a given face shape (a), whose original texture is displayed in (b), can be made to look like Osama bin Laden (c), or George W. Bush (d). Reprinted from [Bronstein et al., 2005].

With these premises, and thanks to the recent technological developments such as the availability of specialized equipment and processing power to tackle 3D face data, face recognition systems using true 3D (depth) data started to emerge in the research landscape, using

[5] "the human visual system, [...] uses mainly the 2D information of the face to perform recognition." Bronstein et al. [2005].

the devices described in section 3.5. These techniques have now developed their own niche in facial recognition research, and amount to a sizeable number, thus begetting their own classification system. Gordon [1991] proposes that, just as the geometric vs. photometric classification of 2D facial recognition (see section 3.2) has proved useful for cataloging the different 2D approaches presented through the years, one can also apply an analog of this taxonomy to 3D recognition, in that analysis and matching of the full surface curvature corresponds to the photometric approach, and the comparison of feature points, detectable through the characteristic shape of the regions surrounding them, matches the geometric approach. He further states that, accordingly, the overlap currently observable in 2D approaches would also translate to the 3D environment, in that to automatically align the scanned surfaces for curvature comparison, some features would have to be detected and combined with high-level knowledge of face anatomy.

Regarding the usage of rigid geometry, Gordon [1991] states:

> "One obvious shortcoming of [comparing aligned 3D range data] is that it doesn't allow for elastic changes in the surface such as varying facial expression. To reduce these effects we can define regions of the face over which comparisons are most meaningful. For instance the mouth and chin region are most highly affected by expression change, so we may not wish to consider differences in those regions."

However, it wasn't until Chua et al. [2000] that such an approach was reported. They explain their reasoning as follows: "While the face-shape of the same person may change, sometimes greatly, due to different facial expressions, there will still be regions, such as nose, eye socket region and forehead, which will keep their shape and position or be subjected to much less deformation between any different expression. If these regions can be identified, the 3D non-rigid face recognition problem can be reduced to the rigid case. Based on this observation, our approach is to extract the rigid parts of the face and utilize them to realize the task of recognition". Interestingly enough, they define the rigid parts as those one standard deviation from the mean location of the data points on the models, and compute the resulting regions based on this specification. This adaptive threshold is flexible and automated, as it doesn't need domain-specific previous knowledge, but might be fooled if most or all entries for a face happen to be in a neutral expression.

By using 3D scanners, these approaches can thus obtain a sufficient number of rigid points and surfaces (for example, in the forehead and eyebrow areas, which are not reliably de-

tectable using image-based methods, due to lack of distinctive intensity variation patterns, and the great flexibility that eyebrows display, respectively), which can then be used for recognition under the principles outlined in the beginning of this chapter. Overall, the excellent results obtained by recent works [Chua et al., 2000; Bronstein et al., 2004; Lee et al., 2005] indicate that indeed **face recognition can be performed reliably through geometric measurement of rigid facial features** even in the presence of large local shape variations, differing poses, and appearance-changing parameters such as illumination changes, skin color, makeup, etc. As such, **the original conjecture of this thesis is validated**.

The case for a hybrid approach

Despite the recent successes, 3D methods do have some disadvantages. Citing Bronstein et al. [2005]: "while in 2D face recognition a conventional camera is used, 3D face recognition requires a more sophisticated sensor [...]. This is one of the main disadvantages of 3D methods compared to 2D ones. Particularly, it prohibits the use of legacy photo databases, like those maintained by police and special agencies". In fact, backwards compatibility is hindered not only with existing data, but also with existing equipment: most environments where application of face recognition technology is a recognized need (surveillance, video conferencing, etc.) are widely equipped with 2D cameras, while 3D scanners, though gradually becoming more available, are still quite rare.

The high computational cost of these algorithms is also a problem [Xu et al., 2004], namely for real-time recognition in video streams, or for implementations in embedded devices with limited resources, such as "intelligent" door locks. Brunelli and Poggio [1993] mention the potential higher recognition speed of geometric methods compared to image-based approaches (because comparisons can be made over numerical coordinates or parameter values, rather than over the pixels of the images, either raw or compressed with dimensionality reduction techniques) and smaller memory requirements (in their experiment, they stored information at one byte per feature, requiring only 35 bytes per person). As stated in Gordon's classification, described above, 3D recognition can be performed using specific key points of the rigid areas of the face, rather than their whole surfaces [Zhou et al., 2004; Bronstein et al., 2005]. This alleviates the second problem, but not the first.

The approach presented on this work aims to cover a middle ground between the 2D and 3D methods.

As mentioned above, techniques for reducing the influence of most parameters that cause variation in 2D images have been presented throughout the years, with varied degrees of success, but arguably approaching the theoretical maximum that a lesser-dimensional (2D) representation of faces can achieve. These techniques attempt to cancel intra-class variation while preserving inter-class variation; for example, removing the effects of different illumination directions (such as shadows) but allowing the natural shading of the face to be preserved (e.g., the fact that the eyes and mouth are darker areas than the surrounding skin).

However, despite the long history, described above, of investigations on the contribution of rigid face geometry for face recognition confirming its importance, very few 2D methods have tackled the problem of preserving the rigid facial shape during shape deformation removal; that is, separating intrinsic rigid proportions from flexible deformations caused by facial expressions, and from the projective deformations caused by pose variation. In other words, pose and expression should be canceled while preserving the general shape of the face. Instead, current literature describes usage of shape-free patches [Cootes et al., 2001, Dornaika and Davoine, 2004, and others] which normalize the faces to cancel all shape deformations, thus ignoring the rigid shape information.

Some research projects [Lanitis et al., 1995; Belhumeur et al., 1997; Edwards et al., 1998] experimented with separating the variation modes obtained from the dimensionality reduction process, attempting to set apart those that affect intra-class variation (i.e., within instances of the same face) from those that influence inter-class variation (i.e., between instances of different faces). This is illustrated in Figure 4.2. However, since the dimensionality reduction is automatic and thus agnostic from high-level knowledge about facial anatomy, and due to the strong nonlinear coupling between pose, face shape and expression [Bascle and Blake,

---

Of course, shape alone is insufficient for face recognition. Gordon [1991] states: "Even if we could extract reliably and accurately the position and descriptions of the standard facial features (eyes, nose, mouth, outline of face area), there is good deal of evidence in the psychology literature to suggest that this is not sufficient for humans to perform individuation among many faces". Since reconstruction of 3D shape from 2D input is only approximate, humans' neural centers for facial recognition had to rely on extra clues, namely the texture. Computers followed the same path, as described above. But there's no reason to discard the extra layer of information that the 3D shape approximation can provide us with, and intuition suggests that the human mechanisms for face recognition do indeed make use of this extra knowledge.

1998], these modes likely won't optimally separate these different phenomena [Cootes and Taylor, 2004, Figure 4.8].

Figure 4.2: The effect of changing parameters that affect mostly inter-class variation (top) and intra-class variation (bottom). Reprinted from [Edwards et al., 1998]

A few attempts were made to explicitly separate pose from expression [Bascle and Blake, 1998; Wang and James Lien, 2009]. However, they were aimed at achieving expression identification, rather than face recognition, and could thus to some degree ignore the coupling between the local shape deformations caused by face proportions and those caused by expressions.

The most straightforward way to achieve separation between **pose**, **shape** and **expression**, is to use a standardized, parameterizable 3D model developed using high-level anatomical knowledge about the face. Such a model will solve the vulnerabilities of deformable models [Zhao and Chellappa, 2000, Blanz and Vetter, 2003, and others] whose parameters are defined and limited statistically from existing data, in that it would allow more variation than that present on the database (by theoretically allowing all possible shape deformations of face geometry), and simultaneously reduce the need for allowing large variations in its parameters (which are otherwise needed to account for combinations of the three factors of shape deformation).

The obvious choice for such a model is the Candide face [Ahlberg, 2001]. Refer to Figure 4.3 for an illustration. It was developed specifically as a research tool – a parametrized model based on the concept of Action Units (AU), which were first described by Hjortsjö [1969] and later extended by Ekman and Friesen [1977] as the Facial Action Coding System (FACS). Since these AUs are mostly intended as descriptors for facial movements as components of facial expressions, the Candide model adds Shape Units (SU), which allow the description of the rigid shape of the face, and therefore the aforementioned separation. Terzopoulos and Waters [1993] report the use of another model (Figure 4.4), which was able to

accurately reproduce face expressions through simulating anatomically-correct movements of the underlying facial muscles. But this fidelity meant the model was quite complex, and since their focus was facial expression capture and face synthesis for animation, they didn't implement parameters for control of rigid geometry, which rendered it inappropriate for this project. Moreover, the relative simplicity of Candide model granted it the status of a *de facto* standard. The 3rd revision to the model aimed to make it compatible with the MPEG-4 standard [MPEG-4–Part 2], specifically the chapter on Face and Body Animation. ISO and IEC, the international standardization bodies which have been responsible for the initial version of the MPEG-4 standard, have also recently presented another standard aimed at biometric data interchange formats [ISO/IEC 19794-5:2005], with a specific section describing a three-dimensional face image data interchange format [ISO/IEC 19794-5:2005/Amd 2:2009]. Candide, or whatever model succeeds it, will likely conform to this standard as well.

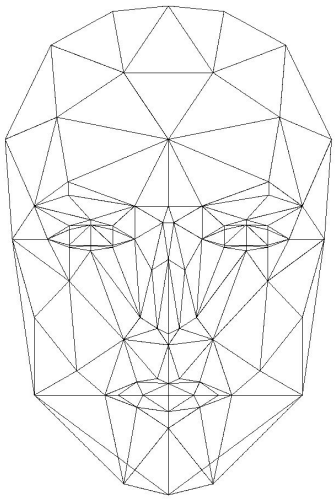


Figure 4.3: The Candide model, version 3. Source: www.icg.isy.liu.se/candide/

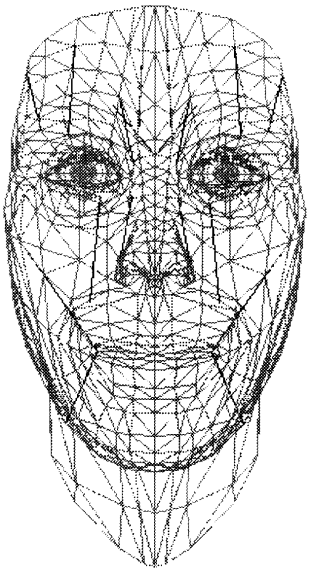


Figure 4.4: Terzopoulos and Waters's anatomical face model. Reprinted from [Terzopoulos and Waters, 1993].

Some research using the Candide model has surfaced recently [Dornaika and Davoine, 2004; Chen and Davoine, 2006; Lefèvre and Odobez, 2009], but most have focused on face tracking, rather than face recognition. Specifically, none has aimed for shape preservation. The current work focused on that front, and the next chapter will describe the specific implementation for validating this theoretical proposal.

Chapter 5

Implementation

The process implemented in this project allows new images to be routed through a fully automated pipeline that detects key points in the image, finds the best 3D face configuration that matches those points, and returns a normalized image, with pose and expression cancelled, but –and this is the key difference from other normalization methods– keeping rigid face shape intact. This normalized image can then be used in a typical 2D face recognition system.

Specifically, the workflow consists of roughly three steps, all automatic:

1. Warping of a neutral 2D point-based model to fit a face image (section 5.1). This step, though automatic, requires a previous training phase.
2. Adjustment of a neutral 3D model to the warped 2D model (section 5.2). The 3D model can be manipulated in pose (translation and rotation), facial expressions, and shape (face proportions), so that its projection best matches the 2D model. The matching is completed by automatic texture mapping of the face image into the 3D model.
3. Neutralization of the 3D model's pose and expression parameters, to yield normalized images, which can then be used as input to a standard image-based face recognition system (section 5.3).

Below, each part of the processing flow is explained in detail.

5.1 Fitting the 2D model

The first step is to automatically process an image to detect the location of feature points required by a standard 2D model. By adjusting the coordinates of the model's points to the features' locations in the image, the model is thus fitted to the face.

To achieve automation on this step, a previous training stage is needed. For this effect, a small set of face photos is manually marked in the key points that define the 2D mesh. Then, a statistical model is constructed from the combined appearance of the areas around these points in the images. With this model, new, unmarked images can then have the 2D mesh automatically placed on them by searching for areas in the image that match the appearance of each model point. This is implemented through the AAM and ASM methods, described in subsection 3.4.5.

In order to have means to effect this training process, an application was built to manage face image databases. These consist merely in system folders with image files, each with corresponding metadata files. Specifically, a C++/Qt application was created to load, display and navigate an image collection. For the landmark tagging, support for two 2D meshes commonly used in AAM and ASM frameworks (a 68-point model, and a 58-point model, respectively) was implemented. Using Qt graphics and event manipulation routines, the model is displayed as an overlay to the current image, with circles on each landmark which can be dragged to the correct positions. Lines are drawn connecting the points in a way that resembles the facial shape contours (face outline, nose, mouth, eyes), in order to ease the model manipulation.

The model can also be transformed globally, with rotation, scale and translation commands implemented with modifier keys (Ctrl, Alt) that alter the effects of dragging with the mouse. Moreover, common image manipulation controls were implemented, such as zooming, automatic fitting into the visualization area, and moving it if it exceeds the display region. All these controls allow the accurate tagging of a wide range of images, regardless of size and orientation of the heads. For illustration purposes, Figure 5.1 shows the application developed, with the navigation and statistical model a 2D mesh already adjusted to a face photo.

With some images tagged, a statistical model can be built using either the AAM or the ASM approach. Our experiments reveal that AAM works well for databases where the appearance of the images is roughly similar (such as the surveillance images of the SCFace database

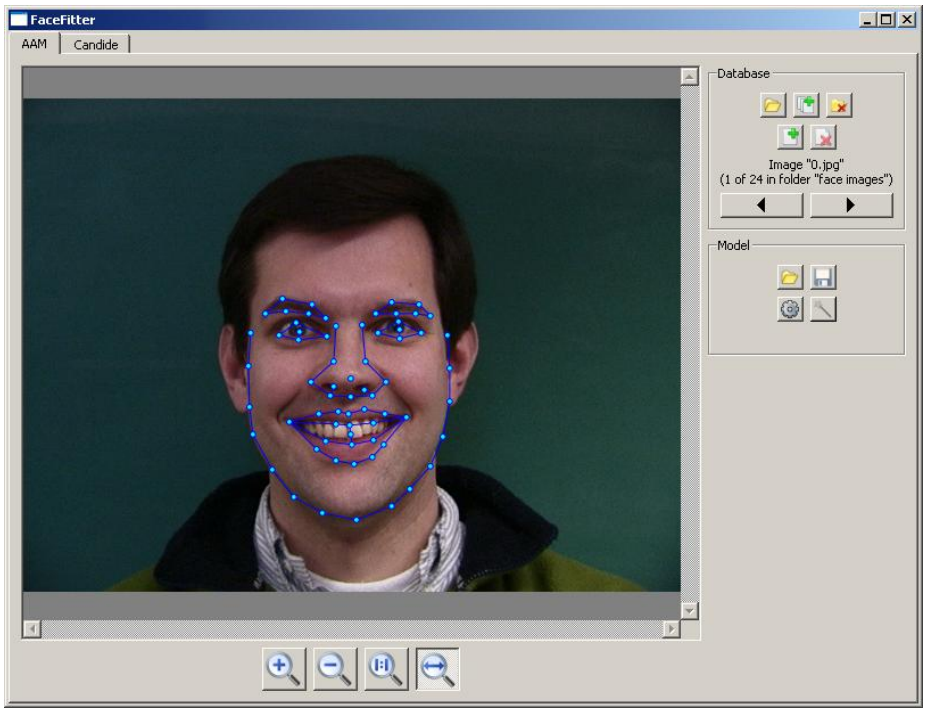


Figure 5.1: A screenshot of the application developed for this project, displaying the tab where visualization and manipulation of the 2D model is carried, as well as the loading, saving and generation of the statistical models. A "magic wand" button allows the loaded (or generated) statistical model to be used to automatically fit the model to a new image.

[Grgic et al., 2009]), but when images vary more widely, the ASM tends to perform better, since it is less dependent on the appearance of the whole face, using instead only the areas around the landmarks. This was the case with the Cohn-Kanade image database [Kanade et al., 2000] and its extended version [Lucey et al., 2010] which we used for the experiments described in chapter 6.

After building the statistical model with a subset of the database, we can use the model to automatically fit the landmarks to the remaining images in the training database, considerably speeding up the training process. The manual manipulation commands allow adjustments to be performed whenever necessary, but the automatic results are often close enough to require few to no manual tweaking. For this automated process, AAMlibrary and ASMlibrary, both by [Wei, 2009], were used.

With a training database of only a few dozen images tagged, the statistical model can be regenerated to provide a refined template which can then be used for the automatic recognition process with new images.

5.2 Fitting the 3D model

With the 2D mesh adjusted to the image, the 3D model (the Candide model, version 3.1.6) can now be configured to fit the photo. Rather than having the 3D mesh matched directly to the photo, it is adjusted to the 2D model instead, through different configurations of pose, shape and expression, until the best adjustment is found. This provides a very fast optimization process, since the error estimation for each configuration of the 3D model consists essentially in a projection of the 3D coordinates into 2D space, and a simple Euclidean distance calculation between the projected coordinates of the 3D model and those of the 2D model.

Before the implementation of this stage as an automatic process, an interface for visualization of the 3D model was built, using the OpenGL bindings of Qt. Controls for manual adjustment were implemented, to allow manipulating the model in pose (rotation and translation), shape and expression. These can be seen in Figure 5.2.

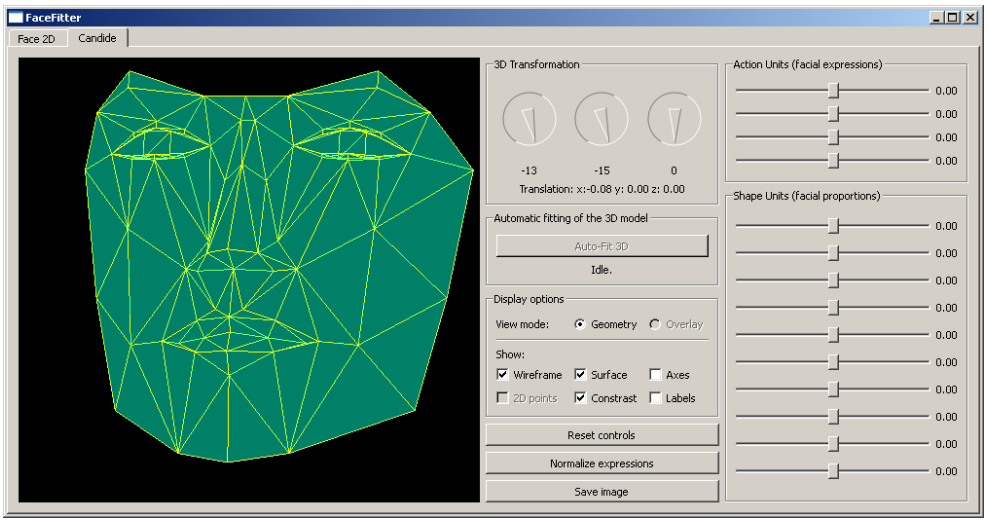


Figure 5.2: A screenshot of the application developed for this project, displaying the tab where visualization and manipulation of the Candide model were implemented.

The visualization and manual configuration controls were provided as an interface to the procedures that implement the adjustment of the 3D Candide model according to its various parameters. The rotation and translation parameters are naturally implemented as 3D transformations to the 3D face object. The shape (SU) and expression (AU) parameters, on the other hand, are defined in the Candide model as sets of displacement vectors: each unit is

specified as a set of vertices and the corresponding translation vectors, and the combined motion of all vertices in the set produces the desired effect of the shape or action unit. The values applied to these vectors serve as modulators for the amount of change they effect. The sliders therefore control this adjustment value, with the neutral position, 0, representing no displacement from the neutral face.

Of course, the primary purpose of these adjustment routines was not manual adjustment, but instead to allow the automatic optimization algorithm to progressively tune and fit the 3D model to the 2D model previously applied to the face image[1]. Given the optimization strategy outlined at the start of this section, another pre-requisite needed to be fulfilled: a direct mapping of the points of the 3D and the 2D models had to be provided. Naturally, the vertices of the Candide 3 model don't match perfectly the points of the 2D model used for AAM/ASM. However, there's a rough correspondence that can be used to generate a comprehensive enough mapping. Figure 5.3 illustrates the correspondence mapping used, and Appendix A presents the complete index of 2D-3D vertex pairs.

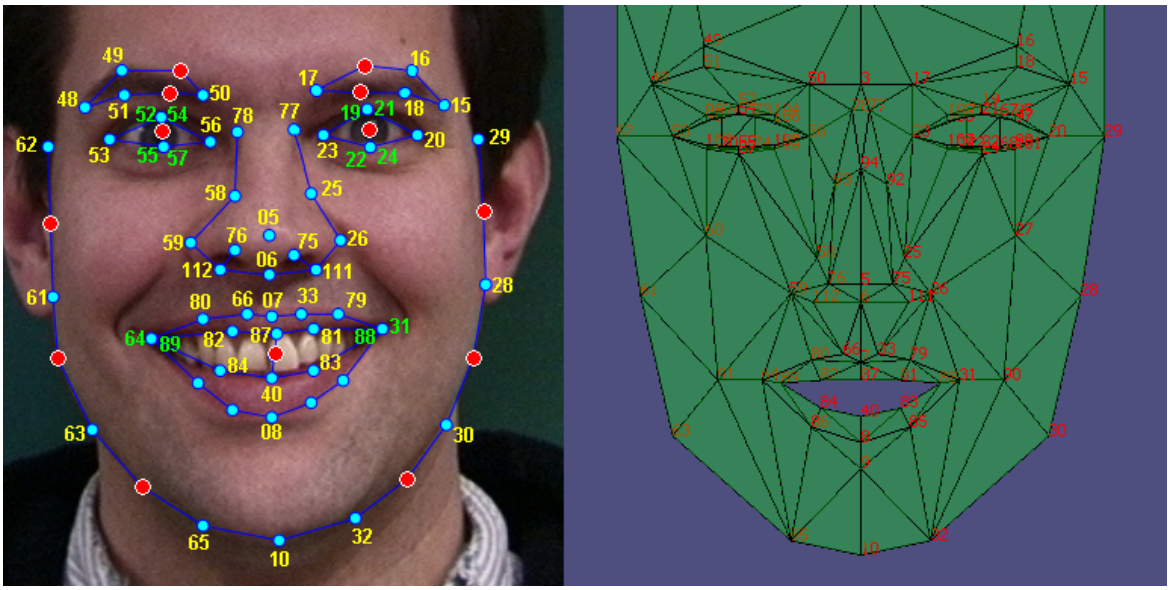


Figure 5.3: The correspondence between points of the 2D model and vertices of the Candide model. The numbers on the left refer to the vertices in the Candide model. Red points do not exist in the Candide model. Points with yellow labels have a direct correspondence with a Candide vertex. Points with green labels correspond to two vertices in the Candide model, whose coordinates are merged together for the comparison.

[1] The manual controls are nevertheless useful in debugging, evaluating the effectiveness of the automatic adjustment (e.g., determining whether a better configuration can be manually obtained) and may even be used for other applications, such as synthesis of new poses or expressions to increase variation in the training database, or animation of a face extracted from a single photograph.

With this set of common points, we build a reduced copy of both the Candide and the 2D models, consisting of only their common vertices. These models are used for the comparison and adjustment process; the complete Candide model can be then configured with the same parameters as the reduced one, after the optimization is completed.

For the initial pose estimation, the POSIT algorithm [DeMenthon and Davis, 1995] was used, since it provides a very fast initial guess assuming the Candide model as a rigid object. POSIT estimates the rotation and translation parameters so that the error between the 2D projection of the 3D vertices and the actual 2D coordinates of the 2D model points is minimized. The reduced Candide model, despite being a subset of the full model, is comprehensive enough to provide a near-perfect approximation for the pose fitting. We used the OpenCV implementation of the algorithm.

This initial configuration is then used as the input for a more sophisticated optimization routine; in this case, the well-known Levenberg-Marquardt algorithm [Levenberg, 1944]. Here, we allow variation of the shape and expression parameters of the Candide model, as well as further adjustment of the rotation and translation vectors, to account for changes in facial shape (both rigid and flexible) that may allow closer matching to the 2D coordinates with an adjusted pose. We only employed a subset of the SUs and AUs for manipulation and optimization, since the strong coupling between some pairs (e.g., the “mouth width” SU and the “lip stretcher” AU) made the optimization last longer and yield less accurate results.

The implementation of the Levenberg-Marquardt algorithm that we used is modular, in that it receives a list of parameters for optimization, and a callback error function which it invokes repeatedly passing different values for the parameters. The return value of this error function is then used by the algorithm to tune the parameters while it searches for the minimum error configuration. The parameters used are the subset of AUs and SUs we selected (as described in the above paragraph), along with the translation and rotation vectors.

Some adjustments were made to the parameters to ensure their validity in this context:

- The AU and SU vectors produce realistic results in the -1 to +1 range. In the manual manipulation mode, these limits were implemented in the interface sliders, which can be seen in Figure 5.2. However, the Levenberg-Marquardt implementation we used does not provide a direct way to limit the values of the parameters of the function to optimize. To obtain values which produce valid configurations of the Candide model, we modulate the unbounded values, returned by the Levenberg-Marquardt algorithm

for these parameters, with a sigmoid function –the hyperbolic tangent– thus ensuring that they get a smooth, continuous mapping into the]-1,1[interval.

- The rotation returned by POSIT is provided in matrix format, and therefore contains interdependent components, which cannot be freely manipulated by a blind algorithm like Levenberg-Marquardt (which treats them as independent parameters) without compromising the integrity of the result as a valid rotation descriptor. Because of this, the matrix is converted to the compact rotation vector format[2], so that each component can safely be modified in an independent manner. This also has the advantage of reducing the number of parameters for optimization, speeding up the process: instead of the 9 components of the rotation matrix, we have only 3 elements of the rotation vector.

After applying the global transformations (rotation and translation) and the local deformations (AU and SU vectors) to the reduced Candide model, we project the 3D coordinates of its vertices to 2D space. The error function then returns the combined difference (that is, the sum of the Euclidean distances) between the projections of the 3D points and their corresponding points from the 2D model (this is the same approach used by the POSIT function). By repeatedly testing parameter configurations and manipulating them to minimize this error, the Levenberg-Marquardt algorithm eventually reaches the best configuration of the 3D Candide face for the 2D image, thus estimating pose, shape and expression.

When this process is complete, we have the configuration of Candide that makes it most closely fit the image. We can then use the projection of the 3D points into the 2D image space, to locate the patches of the image that correspond to each triangle in the Candide model. This allows automatic texturization of the 3D model, completing the generation of a full-3D representation of the face originally presented to the system as a 2D image.

[2] A rotation vector is a compact representation for a rotation that stems from Euler's rotation theorem, which implies that any rotation or sequence of rotations of a rigid body in a three-dimensional space is equivalent to a pure rotation about a single fixed axis. As such, a non-normalized (i.e, non-unit-sized) vector provides the direction of the angle of rotation, and its length (magnitude) describes the angle of rotation.

5.3 Normalized face recognition

After obtaining the 3D model of the face, we can then manipulate it into novel configurations. At this point, several applications are possible, including facial recognition, face animation, or expression recognition. Our focus, however, lies in pose and expression cancellation with shape preservation, so we are specifically interested in removing rotation, translation and facial expressions of the face. For this, the parameters that control these variations are set to the neutral value, while the remaining ones (SUs, mostly) remain untouched. It should however be noted that, in our implementation, some of Candide's shape units were treated as expressions and cancelled as well, since they represent variations in regions of the face that are particularly flexible, such as the mouth and the eyebrows.

After performing this normalization step, further processing is needed for getting the images into a usable state for image-based comparison. Namely, grayscaling is performed for canceling dependence on chromatic and temperature components of the illumination, or skin color variation. Also, histogram equalization guarantees that the range and distribution of intensities is roughly the same on each image.

The result of this pipeline are normalized face images on which elastic shape deformations have been removed, but rigid deviations from the average face are preserved. This format is suitable for usage in the typical 2D face recognition systems using image-based comparison. For this work, we used the ubiquitous eigenfaces method as the testing algorithm, since it's one of the most well-known and studied techniques, has several implementations available, and embodies the basic principles behind many other related recognition techniques. The results of this test are presented on the next chapter.

Chapter 6

Results

The normalization procedure described in the previous chapter was implemented in a way that, save the training phase, enables fully automatic processing, the whole process running in a few seconds per image. The automatic 2D fitting takes less than a second, the pose estimation even less, and the 3D adjustment was tuned (through selective usage of relevant parameters and discarding unhelpful ones) to last around 3 seconds. Given that these measures refer to a basic, unoptimized implementation, and considering the good results obtained by similar approaches [Dornaika and Davoine, 2004; Lefèvre and Odobez, 2009], it is safe to assume that the full adjustment process could be brought to a performance compatible with real-time video processing.

For evaluation of the impact of this novel normalization routine, a brief test was executed using the Cohn-Kanade face databases, version 1 [Kanade et al., 2000] and 2 [Lucey et al., 2010]. After building a database with five subjects, 23 new images of the enrolled individuals were tested with the procedure described in the previous chapter. Both the training images and the new images used for validation contained variation in pose, expression, illumination, color, image size and face location. Photographs with both plain and cluttered backgrounds were used. Figure 6.1 provides a sample of the images used for training and testing.

The training process used one image of each subject. Manual tagging of a subset of the images was performed, and the remaining were tagged semi-automatically with a transitory statistical model, with manual adjustments whenever needed. A definitive statistical model was generated afterwards.

Figure 6.1: Sample of the unprocessed images used in the experiments. © Jeffrey F. Cohn

The 23 test images (amounting to an average of 4.6 images for each subject) were then processed in a totally automated manner, by fitting the 2D model to the images using the model devised in the training phase, and fitting the 3D model to the 2D points with the optimization algorithms described in section 5.2. Two different recognition experiments were carried, to provide a basis for comparison, using a standard implementation of an image-based recognition system – namely, an eigenface-based application [Hewitt, 2007].

In the first experiment, the images were preprocessed with grayscaling and histogram equalization, and only the face was extracted from the images, with size and location (that is, scale and translation) normalized. No compensation for rotation was attempted. This is consistent with the practice of training the classifier with different head orientations to provide robustness to pose variation, due to the inability to reliably estimate and cancel pose in a pure 2D approach. More importantly, facial expression wasn't taken into account, and its effects were transferred into the eigenfaces. The same preprocessing was applied to the testing images. A sample of the images after this processing is presented in Figure 6.2.

In this experiment, we achieved an eigenfaces recognition accuracy in 17 out of 23 test images, resulting in a 74% recognition ratio. While this is below the current reported ratios for similar approaches, we must stress that not only the database used was rather small, thus making this experiment more of a proof of concept than a sound statistical assessment, but also that it contained variations in many attributes that commonly hinder 2D recognition approaches, as mentioned in the beginning of this chapter.

The second experiment was carried in the same manner, but this time using our inferred high-level 3D knowledge of the faces to also cancel pose and expression, while retaining the

Figure 6.2: Sample of the preprocessed images, without pose cancellation or expression normalization.

rigid proportions of the faces, such as eye separation distance or the nose vertical position. The result of this process is presented in Figure 6.3.

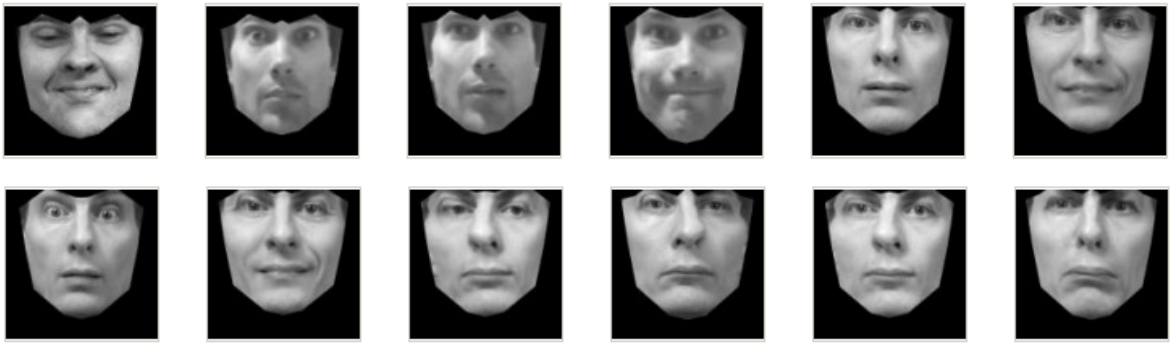


Figure 6.3: Sample of the preprocessed images, with pose cancellation and expression normalization.

By training the eigenface classifier with images treated this way, and also presenting testing images similarly preprocessed, we achieved a recognition rate of 23 out of 23 images – that is, perfect recognition. While we again must stress the small size of the database and the nominal nature of these results, **it is clear that applying this further normalization step –that is, canceling rotation in 3D and removing facial expressions, while preserving the rigid shape of the face– does indeed enhance 2D recognition**, thus validating the proposal made in section 4.

Chapter 7

Conclusion

The initial hypothesis put forth at the beginning of this project was that after achieving an estimation of pose, expression and facial shape of a face (for instance, by obtaining a good fit of the anatomically correct Candide model by tweaking the corresponding parameters), preserving the latter would yield measurements that should remain stable across different images of the same person, thus providing dense statistical distributions, and therefore models capable of distinguishing different people through their characteristic facial proportions.

However, as the implementation grew closer to the fitting objective devised theoretically, it became clear that many non-rigid facial motions do change the location of image-based fiducial points used to infer the bone structure of the face, thus severely reducing the accuracy of the statistical model built from the training images. [Bronstein et al., 2005] acknowledge this problem by stating: “It appears (...) that very few reliable fiducial points can be extracted from a 2D facial image in the presence of pose, illumination, and facial expression variability. As the result, feature-based algorithms are forced to use a limited set of points, which provide low discrimination ability between faces.”

On the other hand, a review of research trends and state-of-the-art techniques for face recognition, as performed in chapter 3, and explored in further detail concerning geometry-based methods in chapter 4, reveals that the usage of such rigid measurements is not only a sound theoretical proposal, but also validated empirically, specifically with the usage of 3D data input systems. These are not entirely ready for widespread adoption, however, due to several factors, including the cost, increased processing power required, and them being more invasive than visible-spectrum image-based acquisition.

There are ways to achieve reliable measurements using relatively non-invasive sensors, such as backscatter X-ray, Terahertz radiation (T-rays) or Millimeter wave scanners, all of which are already deployed and currently in use in many security or medical applications. Approaches based on these technologies would work through the detection of fiducial keypoints directly from data representing the bone structure itself, rather than indirectly from salient features of the skin layer covering the structure. Kakadiaris et al. [2002] provides a good introduction to this kind of approach. Still, the same problems present in the 3D methods plague this approach: lack of available equipment, and especially, inability to provide backwards compatibility with legacy data and capturing devices (that is, cameras).

Aiming to a workable and achievable middle-ground solution, we presented a hybrid 2D+3D approach that, while only providing an approximation of the true rigid proportions of the face (as opposed to flexible deformations caused by expressions), and therefore not being self-sufficient for recognition, carries sufficient value to significantly affect the performance of 2D face recognition methods. This premise is validated by preliminary experiments that produced 100% recognition rate in a database with variations in pose and expression, while a standard 2D approach only reached 74% in the same image database.

When taking into account that the best approaches have used a hybrid model of integrating several techniques, it is quite reasonable to assume that accuracy would be enhanced by integrating these measures into facial recognition systems. Brunelli and Poggio [1993], while extolling the advantages of image-based approaches over the feature-based ones, added that “it is indeed possible that successful object recognition architectures need to combine aspects of feature-based approaches with template matching techniques.”

Indeed, recent research [Pato and Millett, 2010] suggests that despite popular belief, biometric systems are inherently fallible, going as far to state that “[N]o biometric characteristic, including DNA, is known to be capable of reliably correct individualization over the size of the world’s population”, pointing that at the scale often devised for biometric systems, even a small number of false positives or false negatives will cause difficult problems: “[F]alse alarms may consume large amounts of resources in situations where very few impostors exist in the system’s target population.”

These two factors reinforce the value of this approach, suggesting that this extra layer might not only enhance recognition accuracy, as was demonstrated above, but also prove to be a valuable aid in disambiguating hard cases of false positives (twins, look-alikes...) and false negatives (disguise, facial hair, aging, etc.).

In conclusion, having tested the effect of canceling facial expressions, while preserving rigid proportions, for facial recognition systems, we determined that this technique does indeed enrich the recognition data. This assumption has been overlooked in recent research and thus we propose that making use of it would further the state of the art algorithms for face recognition.

For future development, we propose further testing of this hypothesis, to establish its statistical significance, as well as work in optimizing the workflow. For example, real-time Candide fitting was achieved in a tracking context [Dornaika and Davoine, 2004; Lefèvre and Odobez, 2009]. The changes we effected to the Candide-3 model also suggest a revision of its shape and action units to further reduce interdependency between the parameters.

An interesting conjecture to investigate is the extent to which the fact that human faces are non-symmetrical influences recognition performance. Specifically, this asymmetry is mostly ignored by the Candide model, whose shape parameters generally affect both halves of the face equally. If not for recognition, possibly the accuracy of the fitting could be improved by taking this variation into account.

Finally, due to the lack of rigid points detectable in 2D images, we propose that besides the eye and nose key points, ears could also be used as rigid parts of the face, since they are quite invariable both in shape/length/protrusion and position. A 3D model becomes almost unavoidable in this case, since the common simplification of rough coplanarity between the face features cannot be applied when the ears are considered. Lefèvre and Odobez [2009] did extend the Candide to include the side of the head, achieving good results with wide pose variation. This and other such approaches demonstrate the feasibility of this procedure.

We have presented a case for closer inspection of the geometric approach to face recognition, and specifically presented a novel approach which we hope will help further interest and research development in this area in the near future.

Bibliography

Ahlberg, Jörgen. “Candide-3”. Technical report, Linköping University, Linköping, Sweden, 2001. 39

Barrett, William A. “A survey of face recognition algorithms and testing results”. In *Asilomar Conference on Signals, Systems and Computers*, pages 301–305, 1997. 7, 20

Bascle, Benedicte and Blake, Andrew. “Separability of pose and expression in facial tracking and animation”. Technical report, University of Oxford, 1998. 38, 39

Belhumeur, Peter N.; Hespanha, João P.; and Kriegman, David J. “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 21, 34, 38

Blackburn, Duane M.; Bone, Mike; and Phillips, P. Jonathon. “Face recognition vendor test 2000 evaluation report”. Technical report, DoD Counterdrug, DARPA, and NAVSEA, 2001. Available at http://www.frvt.org/FVRT2000/documents.htm. 1

Blanz, Volker and Vetter, Thomas. “Face recognition based on fitting a 3D morphable model”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1063–1074, 2003. 3, 28, 39

Bledsoe, Woodrow W. “The model method in facial recognition”. Technical Report Technical Report PRI 15, Panoramic Research Inc., Palo Alto, California, USA, 1964. 12, 32

Bowyer, Kevin W.; Chang, Kyong; and Flynn, Patrick J. “A survey 3D and multi-modal 3D+2D face recognition”. Technical report, University of Notre Dame – Department of Computer Science and Engineering, Notre Dame, Indiana, USA, January 2004. A short version appears in International Conference on Pattern Recognition (ICPR) 2004. 3

Bowyer, Kevin W.; Chang, Kyong; and Flynn, Patrick J. “A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition”. *Computer Vision and Image Understanding*, 101(1):1–15, 2006. 7, 27, 28

Bronstein, Alexander M.; Bronstein, Michael M.; Spira, Alon; and Kimmel, Ron. “Face recognition from facial surface metric”. Technical report, Israel Institute of Technology – Department of Electrical Engineering / Department of Computer Science, Haifa, Israel, 2004. 37

Bronstein, Alexander M.; Bronstein, Michael M.; and Kimmel, Ron. “Three-dimensional face recognition”. *International Journal of Computer Vision*, 64(1):5–30, April 2005. 26, 35, 37, 53

Brunelli, Roberto and Poggio, Tomaso. “Face recognition: Features versus templates”. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1993. 12, 17, 32, 33, 37, 54

Cartoux, Jean-Yves; LaPresté, Jean-Thierry; and Richetin, Marc. “Face authentication or recognition by profile extraction from range images”. In *Proceedings of the Workshop on Interpretation of 3D Scenes*, pages 194–199, 1989. 27

Chellappa, Rama; Wilson, Charles L.; and Sirohey, Saad. “Human and machine recognition of faces: A survey”. *Proceedings of the IEEE*, 83(5):705–740, 1995. 12

Chen, Yisong and Davoine, Franck. “Simultaneous tracking of rigid head motion and non-rigid facial animation by analyzing local features statistically”. Technical report, HEUDIASYC Mixed Research Unit, Centre national de la recherche scientifique (CNRS) / Compiègne University of Technology, Compiègne, France, 2006. 40

Chua, Chin-Seng; Han, Feng; and Ho, Yeong-Khing. “3d human face recognition using point signature”. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 233–238, 2000. 36, 37

Cootes, Timothy F. and Taylor, Cristopher J. “Statistical models of appearance for computer vision”. Technical report, Manchester University, Manchester, UK, 2004. 39

Cootes, Timothy F.; Taylor, Christopher J.; Cooper, David H.; and Graham, Jim. “Active shape models – their training and application”. *Computer Vision and Image Understanding*, 61:18–23, 1995. Available at: http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/Papers/cviu95.pdf. 24, 25

Cootes, Timothy F.; Edwards, Gareth J.; and Taylor, Christopher J. "Active appearance models". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:681–685, 2001. 26, 38

Cottrell, Garrison W. and Fleming, Michael. "Face recognition using unsupervised feature extraction". In *Proceedings of the International Neural Network Conference*, pages 322–325, 1990. 22

Cox, Ingemar J.; Ghosn, Joumana; and Yianilos, Peter N. "Feature-based face recognition using mixture-distance". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 209–216, 1996. 32

Craw, Ian; Ellis, Hadyn D.; and Lishman, J. Rowland. "Automatic extraction of face-features". *Pattern Recognition Letters*, 5:183–187, 1987. 17

DeMenthon, Daniel and Davis, Larry S. "Model-based object pose in 25 lines of code". *International Journal of Computer Vision*, 15:123–141, June 1995. 46

Dornaika, Fadi and Davoine, Franck. "Online appearance-based face and facial feature tracking". In *Proceedings of the 17th International Conference on Pattern Recognition*, pages 814–817, 2004. 29, 38, 40, 49, 55

Du, Yingzi and Chang, Chein-I. "Discussion the problems of using the roc curve as the sole criteria in positive biometrics identification". In *Proceedings of the SPIE*, volume 6579, pages 65790K1–9, 2007. 7

Edwards, Gareth J.; Taylor, Christopher J.; and Cootes, Timothy F. "Learning to identify and track faces in image sequences". In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 1998. 38, 39

Ekman, Paul and Friesen, Wallace V. *Facial Action Coding System*. Consulting Psychologist Press, 1977. 39

Etemad, Kamran and Chellappa, Rama. "Discriminant analysis for recognition of human face images". In *Journal of the Optical Society of America*, volume 14 of *A*, pages 1724–1733, 1997. 22

Fisher, Ronald A. "The use of multiple measures in taxonomic problems". *Annals of Eugenics*, 7:179–188, 1936. 21

Gabor, Dennis and Stroke, George W. “The theory of deep holograms”. In *Proceedings of the Royal Society of London*, volume 304 of *A*, pages 275–289, 1968. 22

Galton, Francis. “Personal identification and description”. *Nature*, pages 173–177, 201–202, 1888. Available at: http://galton.org/essays/1880-1889/galton-1888-nature-personal-id.pdf. 12, 31

Georghiades, Athinodoros S.; Belhumeur, Peter N.; and Kriegman, David J. “From few to many: illumination cone models for face recognition under variable lighting and pose”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, June 2001. 34

Ginsburg, Arthur P. “Visual information processing based on spatial filters constrained by biological data”. Technical Report AMRL-TR-78-129, Aerospace Medical Research Laboratory (AMRL), Wright-Patterson Air Force Base, Ohio, USA, 1978. 15

Gonzalez, Rafael C. and Woods, Richard E. *Digital Image Processing*. Prentice Hall, 1978. 15, 16

Gordon, Gaile G. “Face recognition based on depth maps and surface curvature”. In *Geometric Methods in Computer Vision*, volume 1570 of *SPIE Proceedings*, pages 234–247, Bellingham, Washington, USA, 1991. SPIE Press. Available at: http://www.vincent-net.com/spie_sandiego.pdf. 3, 26, 27, 32, 35, 36, 37, 38

Grgic, Mislav; Delac, Kresimir; and Grgic, Sonja. “SCface – surveillance cameras face database”. *Multimedia Tools and Applications*, pages 1–17, 2009. ISSN 1380-7501. URL http://www.scface.org. 43

Gu, Lie; Li, Stan Z; and Zhang, Hong-Jiang. “Learning probabilistic distribution model for multi-view face detection”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 116, Los Alamitos, California, USA, 2001. IEEE Computer Society. 12

Harmon, Leon D. “The recognition of faces”. *Scientific American*, 229(5):71–82, November 1973. ISSN 0036-8733. 15, 32

Heisele, Bernd; Serre, Thomas; Pontil, Massimiliano; and Poggio, Tomaso. “Component-based face detection”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 657–662, Kauai, Hawaii, 2001. 12

Hewitt, Robin. “Seeing with opencv, part 5: Implementing eigenface”. *SERVO Magazine*, May 2007. 50

Hjelmäs, Erik and Low, Boon Kee. “Face detection: A survey”. *Computer Vision and Image Understanding*, 83(3):236–274, 2001. 5, 12

Hjortsjö, Carl-Herman. “Människans ansikte och det mimiska språket (man’s face and the mimic language)”. *Studentlitertur*, 1969. in Swedish. 39

ISO/IEC. “Part 5: Face image data”. In *IS 19794:2005 – Information technology Biometric data interchange formats*. International Organization for Standardization / International Electrotechnical Commission, Geneva, Switzerland, 2005. 3, 40

ISO/IEC. “IS 19794-5:2005/Amd 2:2009 – three-dimensional face image data interchange format”. International Organization for Standardization / International Electrotechnical Commission, 2009. Geneva, Switzerland. 40

Kakadiaris, Ioannis A.; Passalis, Georgios; Theoharis, Theoharis; Toderici, George; Konstantinidis, Ioannis; and Murtuza, Mohammed N. “Multimodal face recognition: Combination of geometry with physiological information”. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR’05)*, pages 1022–1029, 2002. Available at http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1467555. 54

Kanade, Takeo. *Computer Recognition of Human Faces*. Birkhauser, Basel, Switzerland, and Stuggart, Germany, 1973. 12, 16, 32

Kanade, Takeo; Cohn, Jeffrey F.; and li Tian, Ying. “Comprehensive database for facial expression analysis”. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, Grenoble, France, 2000. 43, 49

Kass, Michael; Witkin, Andrew P.; and Terzopoulos, Demetri. “Snakes: active contour models”. In *Proceedings of the First International Conference on Computer Vision*, pages 259–268, London, England, 1987. IEEE Press. 24

Kelly, Michael D. *Visual identification of people by computer*. PhD thesis, Stanford University, Stanford, California, July 1970. Abstract published in the Stanford A.I. Project Memo # AIM-130 (AI-130). Reference STAN-CS-168, AD 713252, available at ftp://reports.stanford.edu/www/pub/public_html/cstr/reports/cs/tr/74/409/CS-TR-74-409.pdf, page 96. 12, 32

Lades, Martin; Vorbrüggen, Jan C.; Buhmann, Joachim; Lange, Jörg; von der Malsburg, Cristoph; Würtz, Rolf P.; and Konen, Wolfgang. “Distortion invariant object recognition in the dynamic link architecture”. *IEEE Transactions on Computers*, 42(3):300–311, March 1993. 23

Lanitis, Andreas; Taylor, Christopher J.; and Cootes, Timothy F. “Automatic face identification system using flexible appearance models”. *Image and Vision Computing*, 13:393–401, 1995. 26, 38

Lao, Shihong; Sumi, Yasushi; Kawade, Masato; and Tomita, Fumiaki. “3D template matching for pose invariant face recognition using 3D facial model built with iso-luminance line based stereo vision”. In *International Conference on Pattern Recognition (ICPR)*, pages II:911–916, 2000. 26

Lee, Yonguk; Song, Hwanjong; Yang, Ukil; Shin, Hyungchul; and Sohn, Kwanghoon. “Local feature based 3D face recognition”. In *Proceedings of the International Conference on Audio- and Video-based Person Authentication (AVBPA)*, pages 909–918, 2005. 37

Lefèvre, Stéphanie and Odobez, Jean-Marc. “View-based appearance model online learning for 3D deformable face tracking”. Technical report, Idiap Research Institute / École Polytechnique Fédérale de Lausanne, Martigny, Switzerland, 2009. 29, 40, 49, 55

Levenberg, Kenneth. “A method for the solution of certain non-linear problems in least squares”. *The Quarterly of Applied Mathematics*, 2:164–168, 1944. 46

Lucas, Bruce D. and Kanade, Takeo. “An iterative image registration technique with an application to stereo vision”. In *Proceedings of Imaging Understanding Workshop*, pages 121–130, 1981. 29

Lucey, Patrick; Cohn, Jeffrey F.; Kanade, Takeo; Saragih, Jason; Ambadar, Zara; and Matthews, Iain. “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression”. Technical report, Carnegie Mellon University / University of Pittsburgh, Pittsburgh, Pennsylvania, USA, 2010. 43, 49

Mahalanobis, Prasanta Chandra. “On the generalised distance in statistics”. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936. 6

Maio, Dario and Maltoni, Davide. “Real-time face location on gray-scale static images”. *Pattern Recognition*, 33:1525–1539, 2000. 17, 18

Maltoni, Davide; Maio, Dario; Jain, Anil K.; and Prabhakar, Salil. *Handbook of Fingerprint Recognition*. Springer, 2003. 7

Matthews, William. "The claims of the identiscope". *Photographic News*, 28 (1365):701, October 1884. Available at: http://www.archive.org/stream/photographicnew01unkngoog#page/n722/mode/1up. 31, 32

Matthews, William. "Personal identification". *Scientific American Supplement*, 26 (659):10533, August 1888. Available at: http://galton.org/essays/1880-1889/galton-1888-nature-personal-id.pdf. 31, 32

Messer, Kieron; Matas, Jiri; Kittler, Josef; Luettin, Juergen; and Maitre, Gilbert. "XM2VTSDB: The extended M2VTS database". In *Proceedings of the International Conference on Audio- and Video-based Person Authentication (AVBPA)*, pages 72–77, 1999. 1

Moses, Yael; Adini, Yael; and Ullman, Shimon. "Face recognition: the problem of compensating for changes in illumination direction". In *European Conference on Computer Vision*, pages 286–296, 1994. 21

MPEG Working Group. "Part 2: Visual". In *International Standard on Coding of Audio-Visual Objects*. 1999. Formerly known as ISO-14496-2. 40

Pato, Joseph N. and Millett, Lynette I., editors. *Biometric Recognition: Challenges and Opportunities*. Whither Biometrics Committee; National Research Council, 2010. 54

Phillips, P. Jonathon; Grother, Patrick; Micheals, Ross J.; Blackburn, Duane M.; Tabassi, Elham; and Bone, Mike. "Face recognition vendor test 2002: Evaluation report". Technical Report NISTIR 6965, National Institute of Standards and Technology (NIST), 2003. Available at: http://www.frvt.org/FVRT2002/documents.htm. 1, 26

Phillips, P. Jonathon; Flynn, Patrick J.; Scruggs, W. Todd; Bowyer, Kevin W.; Chang, Jin; Hoffman, Kevin; Marques, Joe; Min, Jaesik; and Worek, William. "Overview of the face recognition grand challenge". *Computer Vision and Pattern Recognition (CVPR)*, pages I:947–954, 2005. 1

Reisfeld, Daniel and Yeshurun, Yehezkel. "Robust detection of facial features by generalized symmetry". In *Proceedings of the 11th International Conference on Pattern Recognition*, volume 24, pages A:117–120, The Hague, The Netherlands, 1995. 17

Sakai, Toshiyuki; Nagao, Makoto; and Kanade, Takeo. “Computer analysis and classification of photographs of human faces”. In *Proceedings of the First USA–Japan Computer Conference*, pages 55–62, 1972. 17

Samal, Ashok and Iyengar, Prasana A. “Automatic recognition and analysis of human faces and facial expressions: A survey”. *Pattern Recognition*, 25:746–751, 1992. 12

Scheenstra, Alize; Ruifrok, Arnout; and Veltkamp, Remco C. “A survey of 3D face recognition methods”. In *International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA)*, volume 3546, pages 891–899. LNCS, 2005. 28

Schneiderman, Henry and Kanade, Takeo. “Probabilistic modeling of local appearance and spatial relationships for object recognition”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 746–751, 2000. 12

Sergent, Justine. “Microgenesis of face perception”. In Ellis, Hadyn D.; Jeeves, Malcolm A.; Newcombe, Freda; and Young, Andy, editors, *Aspects of Face Processing*, pages 17–33. Martinus Nijhff Publishers, Dordrecht, The Netherlands, 1986. Available at http://books.google.com/books?id=h1YoDHfkVK8C. 15

Shashua, Amnon. *Geometry and photometry in 3D visual recognition*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 1994. 21

Sirovich, Lawrence and Kirby, Michael. “Low-dimensional procedure for the characterization of human faces”. *Journal of the Optical Society of America*, 3(4):519–524, 1987. 19

Terzopoulos, Demetri and Waters, Keith. “Analysis and synthesis of facial image sequences using physical and anatomical models”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, Jun 1993. ISSN 0162-8828. 39, 40

Toyama, Kentaro. “Look, ma—no hands! hands-free cursor control with real-time 3D face tracking”. In *Proceedings of the Workshop on Perceptual User Interfaces (PUI’98)*, pages 49–54, 1998. 2

Turk, Matthew A. and Pentland, Alex P. “Eigenfaces for recognition”. *Journal of cognitive neuroscience*, 3(1):72–86, 1991. 8, 16, 19, 20, 21, 29, 33

Viola, Paul and Jones, Michael. “Rapid object detection using a boosted cascade of simple features”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–518, 2001. Available at http://research.microsoft.com/en-us/um/people/viola/pubs/detect/violajones_cvpr2001.pdf. 12

Wang, Te-Hsun and James Lien, Jenn-Jier. “Facial expression recognition system based on rigid and non-rigid motion separation and 3D pose estimation”. *Pattern Recognition*, 42(5): 962–977, May 2009. 39

Wei, Yao. “Research on facial expression recognition and synthesis”. Master’s thesis, Nanjing University, Nanjing, China, February 2009. http://code.google.com/p/asmlibrary. 43

Wikipedia. “Bundle adjustment”, 2010a. URL http://en.wikipedia.org/w/index.php?title=Bundle_adjustment&oldid=386109821. Unique article ID: 13754920. Online; accessed 25 October 2010. 33

Wikipedia. “Camera lucida”, 2010b. URL http://en.wikipedia.org/w/index.php?title=Camera_lucida&oldid=388535676. Unique article ID: 339562. Online; accessed 24 October 2010. 32

Wikipedia. “Correspondence problem”, 2010c. URL http://en.wikipedia.org/w/index.php?title=Correspondence_problem&oldid=392396612. Unique article ID: 6498435. Online; accessed 25 October 2010. 33

Wikipedia. “Lambertian reflectance”, 2010d. URL http://en.wikipedia.org/w/index.php?title=Lambertian_reflectance&oldid=383333956. Unique article ID: 800155. Online; accessed 24 October 2010. 21

Wiskott, Laurenz; Fellous, Jean-Marc; Krüger, Norbert; and von der Malsburg, Cristoph. “Face recognition by elastic bunch graph matching”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997. 23

Xu, Chenghua; Wang, Yunhong; Tan, Tieniu; and Quan, Long. “Automatic 3D face recognition combining global geometric features with local shape variation information”. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 308–313, 2004. 37

Yang, Ming-Hsuan; Kriegman, David J.; and Ahuja, Narendra. “Detecting faces in images: A survey”. *IEEE Transactions on Pattern analysis and Machine intelligence*, pages 34–58, 2002. 12, 15, 16

Yuille, Alan L.; Cohen, David S.; and Hallinan, Peter W. “Feature extraction from faces using deformable templates”. *International Journal of Computer Vision*, 8:99–112, 1987. 24

Zhao, Wen Yi and Chellappa, Rama. “SFS based view synthesis for robust face recognition”. In *Preceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 285–292, 2000. 29, 34, 39

Zhao, Wen Yi; Chellappa, Rama; Phillips, P. Jonathon; and Rosenfeld, Azriel. “Face recognition: A literature survey”. *ACM Computing Surveys (CSUR)*, 35(4):399–458, 2003. 2, 3, 12, 13, 14, 15, 22, 29

Zhou, Mingquan; Liu, Xiaoning; and Geng, Guohua. “3D face recognition based on geometrical measurement”. In *Sinobiometrics*, pages 244–249, 2004. 37

Appendix A

Correspondence between the 3D and 2D models

Below we present the complete index of the correspondence between the vertices of the Candide 3D model, and the nodes of the 68-point 2D model. This correspondence was used to produce the "unified" models: reduced versions of the 3D and 2D models, which can be compared directly to each other during the optimization (3D model fitting, see section 5.2). Afterwards, the parameters of the reduced Candide model can be passed to the full version, for the normalization step to proceed with greater detail.

Some points in the Candide model that don't have direct correspondence to a 2D point are represented by a set of values in the table; the average location of these vertices is used, as they provide a good approximation to the corresponding 2D points.

The vertex labels are largely based on Candide3 vertex names; those that were altered are displayed in *italics*.

The first row of the table corresponds to the virtual center point, which doesn't exist in either model, but is necessary for running POSIT. Due to the nature of Candide coordinates, this point is the origin of the coordinate space (0,0,0). For the AAM model, four real points were used to calculate its estimated location.

Table A.1: Full Candide-AAM_68 correspondence

Candide #	AAM_68 #	Unified #	Name
N/A	29+34+38+44	0	Center of the model
62	0	1	Upper contact point between right ear and face
61	2	2	Lower contact point between right ear and face
63	4	3	Right corner of jaw bone
65	6	4	Chin right corner
10	7	5	Bottom of the chin
32	8	6	Chin left corner
30	10	7	Left corner of jaw bone
28	12	8	*Lower contact point between left ear and face*
29	14	9	Upper contact point between left ear and face
15	15	10	Outer corner of left eyebrow
16	16	11	*Top center point of left eyebrow*
17	18	12	Inner corner of left eyebrow
18	20	13	*Bottom center point of left eyebrow*
48	21	14	Outer corner of right eyebrow
49	22	15	*Top center point of right eyebrow*
50	24	16	Inner corner of right eyebrow
51	26	17	*Bottom center point of right eyebrow*
53	27	18	Outer corner of right eye
52+54	28	19	Top center point of left eye
56	29	20	Inner corner of right eye
55+57	30	21	Bottom center point of right eye
69+70+73+74	31	22	Right pupil
20	32	23	Outer corner of left eye
19+21	33	24	Top center point of left eye
23	34	25	Inner corner of left eye
22+24	35	26	Bottom center point of left eye
67+68+71+72	36	27	Left pupil
78	37	28	*Right edge of nose bridge*
58	38	29	*Right nose crease*
59	39	30	*Right nostril outer border*

Candide #	AAM_68 #	Unified #	Name
112	40	31	Bottom right edge of nose
6	41	32	*Bottom middle point of nose*
111	42	33	Bottom left edge of nose
26	43	34	Left nostril outer border
25	44	35	*Left nose crease*
77	45	36	*Left edge of nose bridge*
76	46	37	Right side of nose tip
75	47	38	Left side of nose tip
64+89	48	39	Right mouth corner
80	49	40	*Middle point of right outer edge of upper lip*
66	50	41	*Uppermost point of right outer edge of upper lip*
7	51	42	*Center point of outer edge of upper lip*
33	52	43	*Uppermost point of left outer edge of upper lip*
79	53	44	*Middle point of left outer edge of upper lip*
31+88	54	45	Left mouth corner
85	55+56	46	*Middle point of left outer edge of lower lip*
8	57	47	*Center point of outer edge of lower lip*
86	58+59	48	*Middle point of right outer edge of lower lip*
84	60	49	*Middle point of right inner edge of lower lip*
40	61	50	*Center point of inner edge of lower lip*
83	62	51	*Middle point of left inner edge of lower lip*
81	63	52	*Middle point of left inner edge of upper lip*
87	64	53	*Center point of inner edge of upper lip*
82	65	54	*Middle point of right inner edge of upper lip*
5	67	55	Nose tip