UNIVERSIDADE FEDERAL DO PARANÁ
DEPARTMENT OF INFORMATICS

MAURÍCIO PAMPLONA SEGUNDO

Real-time 3D face recognition using low-cost acquisition devices

Curitiba
2013

MAURÍCIO PAMPLONA SEGUNDO

Real-time 3D face recognition using low-cost acquisition devices

A thesis submitted to the Department of Informatics, Universidade Federal do Paraná in partial fulfullment of the requirements for the degree of Doctor of Philosophy in Computer Science.

Advisor:	Prof. Dr. Luciano Silva
Co-advisors:	Prof. Dr. Olga R. P. Bellon
	Prof. Dr. Sudeep Sarkar

Curitiba
2013

DEDICATION

This thesis is dedicated to my parents, Maurício Pamplona and Rosângela de Oliveira Heinig Pamplona.

ACKNOWLEDGEMENTS

First, I would like to acknowledge my advisors, Drs. Luciano Silva, Olga Bellon and Sudeep Sarkar, for their guidance and support throughout my doctoral work. Their knowledge and insights were essential to my thesis completion and also to my personal growth. I especially thank Dr. Sarkar for receiving me at University of South Florida (USF) during my doctoral stage, which was a great opportunity that gave me the final push to complete this work. I would like to extend this special thanks to Dr. Dmitry Goldgof for his co-advising in this period at USF. I would also like to acknowledge my committee members, Drs. Paulo Gotardo, Siome Goldenstein and Daniel Weingaertner, for their time, comments and suggestions. Their valuable inputs lead to a considerable improvement in the quality of this thesis. I am very grateful to my labmates, colleagues and friends Rubisley de Paula Lemes, Leonardo Gomes, Rafael Henrique Tibães, Flávio Henrique de Bittencourt Zavan, Rodrigo Alves Nunes, Jong Wan Silva, Caroline Mazetto Mendes, Beatriz Trinchão Andrade, Karl Apaza Agüero, Ronaldo dos Santos Alburnio, Ícaro Oliveira de Oliveira, Leandro Bispo de Oliveira, Maxwell Schner, Henrique Costa, Fillipe Dias Moreira de Souza, Fabio Faria, Matthew Shreve, Timur Luguev, Ravi Panchumarthy, Ravikiran Gurumkonda Krishnan, Rajmadhan Ekambaram, Kester Duncan, Henry Krewer, Benjamin Geiger, Ravi Subramanian, Mona Fathollahi, Baishali Chaudhury, Bingxiong Lin and Imrul Kayes for their support and valuable discussions. Finally, I would like to acknowledge my family for their constant support and encouragement during this graduate journey: my father Maurício Pamplona, my mother Rosângela de Oliveira Heinig Pamplona, my brothers Luciano Pamplona Sobrinho and Gabriel Filipe Pamplona, my sister Carolina Heinig Pamplona, my girlfriend Mirian Angélica da Silva, my uncle Ari José Coelho Filho, my aunt Silvana Pamplona Coelho, my cousins Pedro Pamplona Coelho and João Pamplona Coelho, and my family away from home James Vaughan and Gisele Carion.

ABSTRACT

Biometric identification seeks to distinguish humans by physical or behavioral characteristics, but there is not a perfect biometric feature. Face biometrics show up as a viable option, since they are well accepted by the public, the capturing process requires minimal collaboration or even no collaboration at all, and the cost is fairly low. As a counterpoint, its performance degrades in uncontrolled conditions, which may include variations in pose, illumination, resolution, environment, facial expressions and age. Among different facial properties that could be used for recognition, the geometry stands out for its invariance to pose and illumination. Within this context, this doctoral work aims to propose a solution for the problem of recognizing people using the facial geometry. What differ this work from other works addressing this problem in the literature is that we have added real-time performance and low-cost acquisition as a requirement. To accomplish this, we have designed a novel face detection method, which was thoroughly evaluated and compared to the state-of-the-art, and we have also optimized the normalization, description and matching stages of the recognition process. We have shown the operation of our system in one of the possible applications, which consists in continuously authenticating the identity of a user to provide a more secure session for high security environments. By doing so, we have developed the first continuous authentication system based on the geometry of the face, which is robust to a wide range of facial variations. Finally, we have addressed the compatibility between our system and the current forms of identification (*e.g.* ID cards, passports, driver licenses). To this end, we have designed a 3D face reconstruction method that uses a single or multiple 2D views of a face to retrieve its geometry. Our results show that the method can effectively create realistic 3D face models, which are suitable for person identification.

RESUMO

A identificação biométrica busca distinguir humanos através de caractersticas físicas ou comportamentais, mas não existe uma característica biométrica perfeita. A biometria facial surge como uma opção viável, uma vez que ela é bem aceita pelo público, o processo de captura requer uma colaboração mínima ou nenhuma colaboração, e o custo é relativamente baixo. Em contrapartida, o seu desempenho decai em condições não controladas, como variações na pose, iluminação, resolução, ambiente, expressões faciais e idade. Entre as diferentes propriedades faciais que podem ser utilizadas para o reconhecimento, a geometria se destaca por sua invariâcia à pose e iluminação. Neste contexto, este trabalho de doutorado busca propor uma solução para o problema do reconhecimento de pessoas utilizando a geometria facial. O que diferencia este trabalho de outros trabalhos abordando este problema na literatura é que adicionamos o desempenho em tempo-real e a aquisição de baixo-custo como requisitos. Para isto, projetamos um novo método de detecção facial que foi meticulosamente avaliado e comparado com o estado-da-arte, e também otimizamos os estágios de normalização, descrição e correspondência do processo de reconhecimento. Nós demonstramos o funcionamento do nosso sistema em uma das possíveis aplicações, que consiste em autenticar continuamente a identidade de um usuário para assegurar uma sessão mais segura para ambientes de alta segurança. Com isto, ns desenvolvemos o primeiro sistema de autenticação contínua baseado na geometria da face, que é robusto a uma vasta gama de variações faciais. Por fim, abordamos a compatibilidade entre nosso sistema e as formas de identificação atuais (*e.g.* documentos de identidade, passaportes, carteiras de habilitação). Para isto, projetamos um método de recontrução facial 3D que usa uma ou múltiplas vistas 2D de uma face para recuperar a sua geometria. Os nossos resultados mostram que o método pode criar modelos faciais 3D realísticos efetivamente, que são apropriados para a identificação de pessoas.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

1:1 one-to-one

1:N one-to-many

2D two-dimensional

3D three-dimensional

B3DO Berkeley 3-D Object Dataset

BOSPHORUS The Bosphorus Database

BU-3DFE Binghamton University 3D Facial Expression

CDF Cumulative Distribution Function

EER Equal Error Rate

FAR False Acceptance Rate

FDR False Discovery Rate

fps frames per second

FRGC Face Recognition Grand Challenge

FRR False Rejection Rate

HOG Histogram of Oriented Gradients

IATA International Air Transport Association

ICP Iterative Closest Point

LDA Linear Discriminant Analysis

LM Levenberg-Marquardt

mm millimeters

PCA Principal Component Analysis

RGBDFD RGB-D Face Database

ROC Receiver Operating Characteristic

ROI Region of Interest

SFM Structure From Motion

SIFT Scale-Invariant Feature Transform

SURF Speeded Up Robust Features

TEXAS3D Texas 3D Face Recognition Database

TPS Thin-Plate Splines

CONTENTS

CHAPTER 1

INTRODUCTION

Since the beginning of human history people have intuitively used physical and/or behavioral characteristics as a mean of recognition, which is called biometric recognition. In the last century there was an increasing interest in developing biometric-based systems for person identification due to their numerous applications in security, accessibility and law enforcement, among others [Jain et al., 2004a]. To this end, several human characteristics have been investigated, such as voice [Doddington, 1985], face [Zhao et al., 2003], iris [Daugman, 1993], fingerprint [Jain et al., 1997], retina [Xu et al., 2005], palmprint [Zhang et al., 2003], hand geometry [Sanchez-Reillo et al., 2000], gait [Sarkar et al., 2005], and signature [Plamondon and Srihari, 2000]. Figure 1.1 shows an evaluation of the biometric features most commonly used in commercial systems, similar to the evaluation presented by Jain et al. [2004b], according to five different criteria: 1) accuracy, which relates to the performance of biometric systems and to the singularity of biometric traits; 2) acceptability, which means how well people agree to use such technologies; 3) permanence, which refers to variations in biometric traits over time; 4) cost, which relates to the investment needed to build such systems; and 5) compatibility, which relates to the capability of integration with current identification documents.

As may be seen in Figure 1.1, there is not a perfect biometric feature. Furthermore, practicability (*i.e.* acceptability and cost) seems to be inversely proportional to reliability (*i.e.* accuracy and permanence). For instance, iris recognition is very reliable but it is not well accepted by the public because people hesitate in taking pictures of their eyes. On the other hand, face recognition is quite feasible but its performance heavily depends on the acquisition conditions. However, according to Hietmeyer [2000] the face is the most likely biometric to be used in a global traveler identification system. Some of the reasons are listed below:

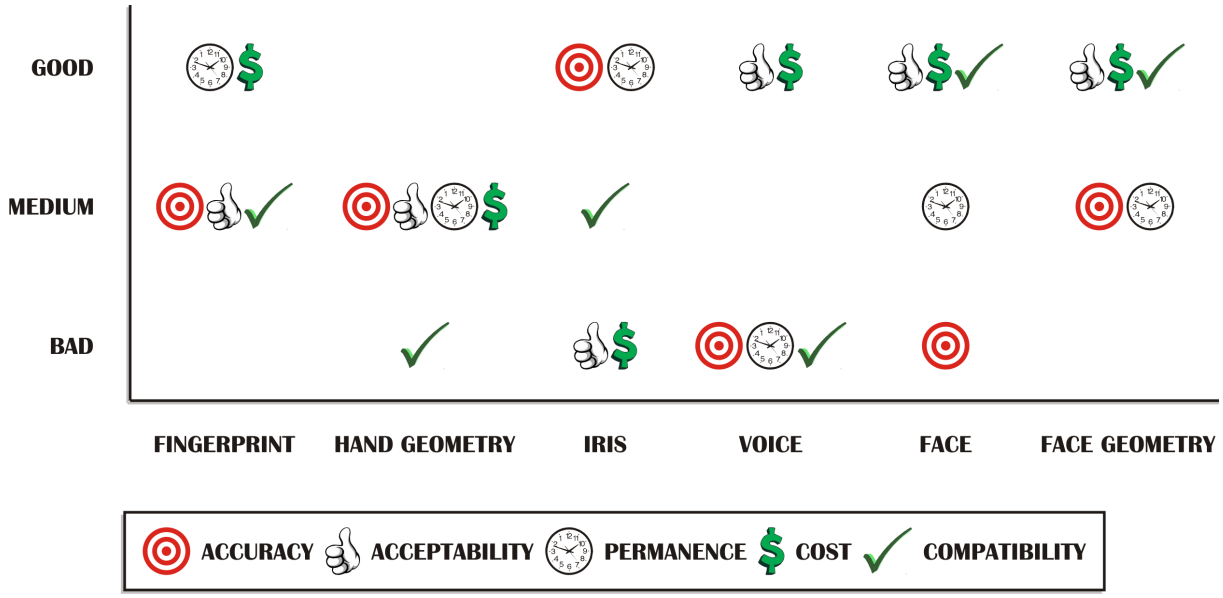


Figure 1.1: Evaluation of the biometric features most commonly used in commercial systems.

Acceptability: human beings intuitively use faces as a mean of recognition [Jain et al., 2004b].

Measurability: faces can be captured in multiple ways, and they have different measurable properties (*e.g.* texture, shape and heat).

Flexibility: faces have a great potential to recognize non-cooperative individuals since they do not require human intervention to be measured, unlike other famous biometrics such as fingerprint and iris.

Applicability: most of the biometric features are only used for access control, but faces are also useful for surveillance, games, animation and entertainment, among others.

Biometric recognition is very useful in highly crowded areas, such as airports, malls and stadiums. It can be used not only for security and law enforcement purposes, but also to provide an efficient access infrastructure to the people. For example, the International Air Transport Association (IATA) proposed a new airport security checkpoint paradigm, called *Checkpoint of the Future*, that uses corridors with state-of-the-art security technology to avoid invasive procedures. According to the IATA, biometrics play a major role in this project, once the identity is used to know how threatening a passenger is based

on a previous risk analysis. There are only a few biometric features that can be used to recognize people walking through a corridor, and the face is probably one of the most viable options [Hietmeyer, 2000].

Face recognition based on texture images, also called two-dimensional (2D) or color images, was the main focus of researches regarding face biometrics for many years [Zhao et al., 2003]. However, recognizing individuals based on 2D images is a challenging problem due to variations in pose, illumination, facial expressions and age, as illustrated in Figure 1.2. Pose changes are caused by out-of-plane rotations, such as looking up and down, left and right. Figure 1.2(b) shows an example of pose variation. Illumination changes are due to uncontrollable lighting conditions in the environment, as shown in Figure 1.2(c). Facial expressions are a very common nonverbal communication that uses facial muscle movements to express emotions. Figure 1.2(d) shows an expression that gives the impression of happiness. Finally, aging brings wrinkles, blemishes and skin discoloration and therefore causes considerable changes in facial appearance, as illustrated in Figure 1.2(e) (*i.e.* this image was created by the app AgingBooth[1]).

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

Figure 1.2: Example of variations in 2D images: (a) an appropriate face image and the effects of changes in (b) pose, (c) illumination, (d) facial expression and (e) age.

In order to avoid some of these problems, recent works consider other facial properties such as shape and heat for recognition. Infrared images were employed to acquire the heat information of faces as a straightforward enhancement of the previous face recognition approaches based on 2D images, since infrared images are not affected by illumination variations at all [Socolinsky et al., 2001]. However, such images are affected by perspi-

[1] `http://www.piviandco.com/apps/agingbooth/`

ration [Buddharaju et al., 2007], which is a common response to stressful situations and physical activities. Figure 1.3 shows an example of the effect of perspiration on infrared images.

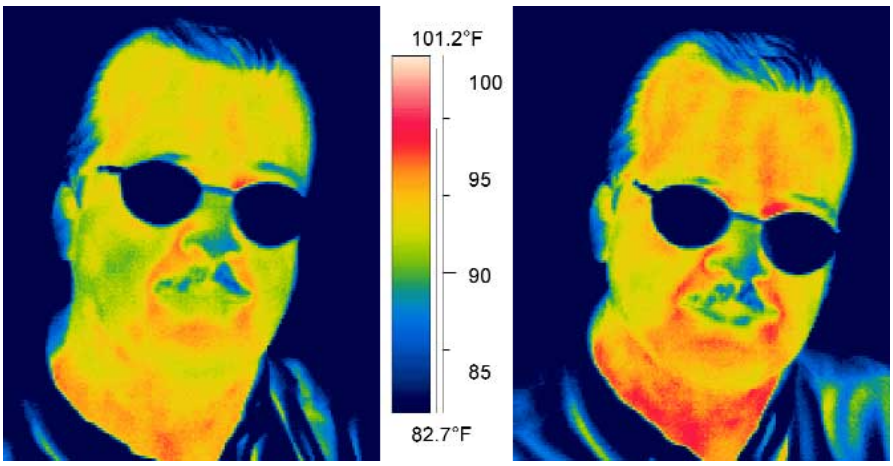


Figure 1.3: Variations in facial heat caused by a stressful situation (image taken from [The Snell Group, 2002]).

After that, there was an increasing interest in using the shape information (*i.e.* three-dimensional (3D) or geometry images) to recognize faces [Bowyer et al., 2006], despite the high cost of 3D data acquisition. Logically, 3D images are not affected by pose variations since orientation can be easily rectified by using data rotation. Also, many 3D acquisition systems based on laser triangulation [Marshall and Stutz, 2011], structured light [Batlle et al., 1998] and other technologies are not affected by illumination variations. For these reasons, the shape information is more accurate than other facial properties, such as texture and heat, as also illustrated in Figure 1.1.

The main challenges for practicable 3D face recognition systems use to be the acquisition (*i.e.* acquisition time, capture area) and the cost (*i.e.* price ranging from US$ 5,000 to US$ 50,000). However, recent 3D sensors have a good trade-off between speed, usability and price. This new generation of sensors, which includes the Microsoft Kinect[2], the ASUS Xtion PRO[3] and the Primesense Carmine[4], is able to capture up to 100 frames per second (fps) of 3D data, the reach extends to up to five meters, and the average cost

[2]http://www.xbox.com/kinect
[3]http://www.asus.com/
[4]http://www.primesense.com/

is about US$ 200. Their 3D images are not as good as the ones acquired by expensive sensors, but the gap is narrowing fast with new sensors being released every year.

These new sensors have opened up a vast array of real-time applications that was not possible with previous acquisition devices. Some examples are user interaction [Ren et al., 2013], action recognition [Mansur et al., 2013] and live object reconstruction [Izadi et al., 2011]. This is also true for 3D face recognition, which now can be performed in real-time for video applications, such as surveillance and continuous authentication [Pamplona Segundo et al., 2013a].

1.1 Objectives

The main objective of this doctoral work is the design of a 3D face recognition framework that is fully automatic, runs in real-time and uses low-cost sensors for data acquisition. By doing so, we allow the development of 3D face recognition systems that are economically viable in industry.

There is also a secondary objective, which is to provide compatibility between a 3D face recognition system and the current forms of identification (*e.g.* ID cards, passports, driver licenses). This is done by reconstructing the geometry of the face using 2D images. This way, as an example, travelers could be matched to their passport photos using facial geometry.

1.2 Contributions

To achieve our objectives, a number of contributions have been made:

- we have designed, to the best of our knowledge, the first real-time 3D face recognition framework that handles multiple fps and uses a low-cost device for image acquisition [Pamplona Segundo et al., 2013a];

- as far as we know, we have created the most accurate 3D face detector in the literature, which works successfully for different acquisition scenarios, including substan-

tial variations in resolution, noise, pose, and facial expressions [Pamplona Segundo et al., 2011, 2013b];

- we have designed a scale-invariant image representation, named orthogonal projection images, that allows using the size of an object to optimize the detection process [Pamplona Segundo et al., 2011, 2013b];

- we have successfully applied our 3D face recognition framework to the continuous authentication problem, making it the first continuous authentication system based on 3D face images [Pamplona Segundo et al., 2013a];

- we have shown a more intuitive way of evaluating continuous authentication systems using well-known biometric terms [Pamplona Segundo et al., 2013a];

- we have developed a new 3D face reconstruction method that uses only a single 2D face image with arbitrary pose as input [Pamplona Segundo et al., 2012];

- we have corroborated neuropsychology works [Hole and Bourne, 2010], showing quantitatively that half-frontal face images have more information about the geometry of the face than frontal and profile ones [Pamplona Segundo et al., 2012].

1.3 Outline

This thesis is organized as follows: Chapter 2 describes our real-time 3D face recognition system and its application for continuous authentication, as well as our 3D face reconstruction approach; the experimental evaluation and the obtained results are presented in Chapter 3; Chapter 4 presents our conclusions, followed by the references.

CHAPTER 2

REAL-TIME 3D FACE RECOGNITION

A 3D face recognition system is divided in the following modules, as illustrated in Figure 2.1: (1) 3D data acquisition module, in which a raw digital representation of the geometry of the face is acquired by a 3D sensor; (2) face detection module, responsible for identifying the location of the face; (3) face normalization module, which standardize the pose and the resolution for further analysis; (4) description module, responsible for creating a biometric template using discriminant features of the normalized face image; (5) matching module, where the obtained template is compared to previously obtained templates; and (6) evaluation module, which analyzes the matching results to decide on the identity.

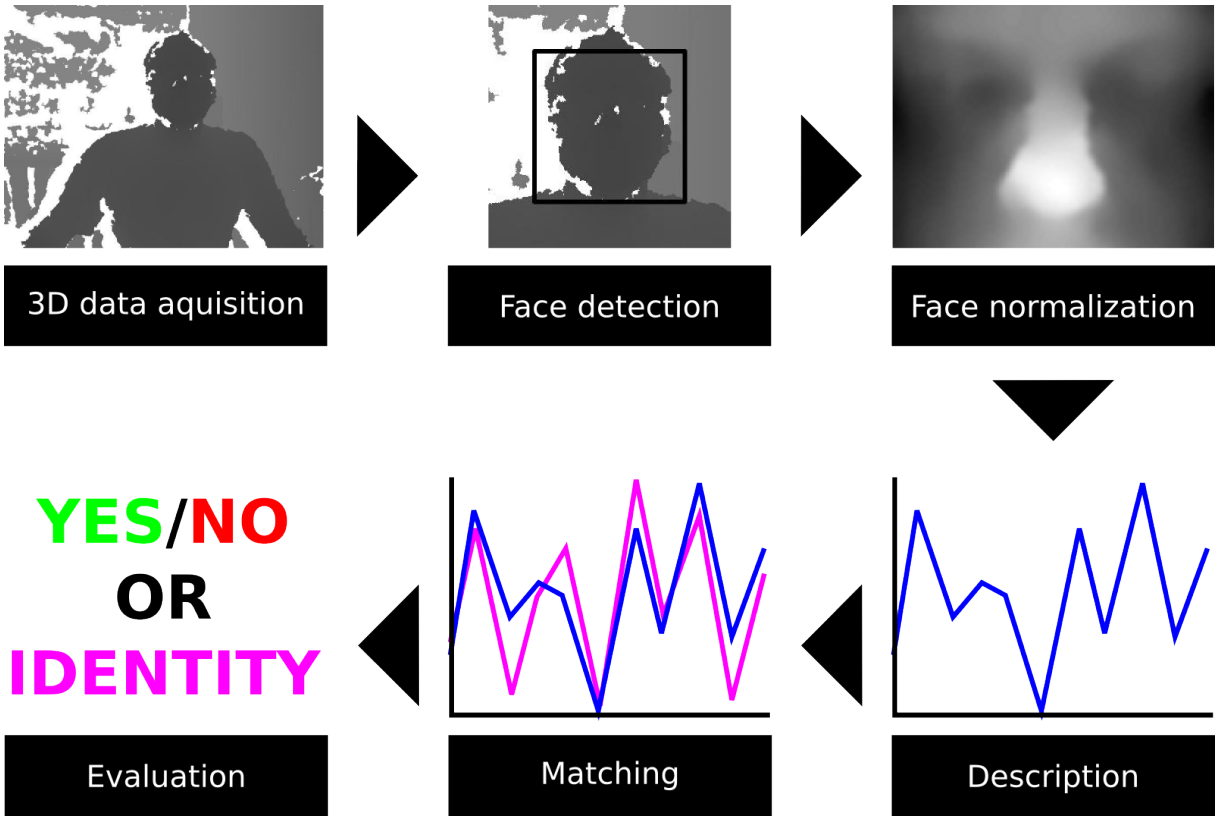


Figure 2.1: Diagram of a 3D face recognition system.

Different systems were proposed in the literature following a similar sequence of mod-

ules [Chang et al., 2006, Kakadiaris et al., 2007, Mian et al., 2007, Queirolo et al., 2010, Wang et al., 2010], but none of them have considered real-time performance or low-cost acquisition devices, which are the target of this work. A detailed explanation about each module is given in Sections 2.1-2.6, and the application of the system for continuous authentication purposes is presented in Section 2.7.

A system as presented in Figure 2.1, however, lacks compatibility to the current forms of identification. To overcome this problem, a 3D reconstruction module can be added to convert 2D face images, which are usually found in identification documents, into 3D information. Our solution to this task is presented in Section 2.8.

2.1 3D data acquisition

The acquisition module is the first stage of any biometric system, and it is responsible for the digitization of the biometric trait. There are different ways of capturing the geometry of the face, and the most common ones are based on stereo vision, laser triangulation, structured light or time-of-flight. A small description for each method is given below:

Stereo vision: this method is based on the process called *stereopsis*, which is the depth perception given by the binocular vision. It captures the scene using two cameras placed side by side, and then measures the displacement between corresponding points in the pair of images to retrieve the depth value. The major challenge in this process is to find reliable point correspondences, which directly affects the accuracy of the resulting image.

Laser triangulation: this method projects a laser point on the scene, which is reflected back to the sensor. Since the distance between laser source and sensor and the projection angle are known, the depth is computed through trigonometric triangulation. Although it can be extended to capture a laser line instead of a single point, it still requires scanning the scene, which is time consuming.

Structured light: this method projects a known light pattern on the scene, and then analyzes the distortion in this pattern to retrieve the depth information. It obtains

more accurate results than laser triangulation, but it may require a specific lighting to work properly.

Time-of-flight: this method estimates the depth by measuring the time taken by a pulse of light to reach the scene and reflect back to the sensor, which is possible because the speed of light is known. Its depth measurements are noisier than the structured light ones, but the process is simple and fast, being able to capture more than 100 fps.

In order to achieve our goal, the acquisition must be performed in real-time and the depth measurements must be accurate enough to perform the recognition. For these reasons, acquisition devices based on laser triangulation or stereo vision are not considered for this work. Since both remaining options satisfy our speed requirements, we have chosen a sensor based on structured light because it is more accurate than time-of-flight.

Among the sensors based on structured light, there are some low-cost options (*i.e.* cost US$200 or less): Microsoft Kinect, ASUS Xtion PRO and Primesense Carmine. The Kinect, which is used in this work, was the first commercially available and is still the cheapest one. More details about this sensor are given in Section 2.1.1. Other sensors with equal or better accuracy than the Kinect can also be used with small or no modifications in the proposed system.

2.1.1 Microsoft Kinect

The Microsoft Kinect is a sensor that was originally designed to enable a hands-free gaming environment in the XBOX 360 console[1]. It captures color, depth and sound data from a scene, and then uses such information to identify user interactions (*i.e.* body movements, voice commands) and map them as game controls. The range of applications for the Kinect is, however, much wider. It has been used for object recognition, scene understanding, sign language recognition and many others.

The Kinect is able to capture up to 30 fps of depth and color data. For each acquisition,

[1] http://www.xbox.com/

the Kinect projects a pattern of points using infrared light and then uses the displacement of these points to recover their depth. This pattern was reverse engineered by Reichinger [2011], and is shown in Figure 2.2. Reichinger [2011] also pointed out that the same pattern is projected nine times in a 3×3 shape, as illustrated in Figure 2.3 (*i.e.* the repetition is also shown in the patent of the sensor [Shpunt and Pesach, 2010]). In total, we have about 30,000 projected points that return real depth measurements. This is much less than the 307,200 depth pixels returned by the Kinect (*i.e.* depth image with size 640×480), so we assume that the remaining pixels are interpolated.

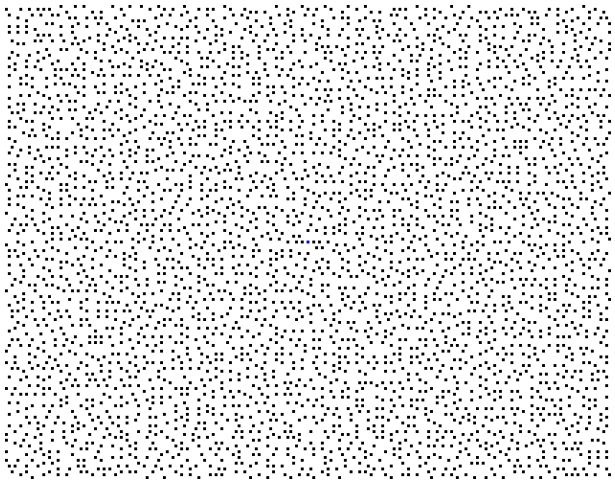


Figure 2.2: Kinect pattern.

Retrieving the exact location of each point is a very difficult task, which results in noisy displacement values, and consequently noisy depth measurements. This effect is even worse in objects that are far from the Kinect [Herrera C. et al., 2012], since the displacement values get smaller due to the perspective distortion. Figure 2.4 shows the real distance and the average error in millimeters (mm) for each disparity value of the Kinect (*i.e.* the disparity value is inversely proportional to the displacement value). As may be seen, the error grows considerably with the increasing distance.

In order to be recognizable, a Kinect image of a human face must be captured no more than one meter away. Images acquired at larger distances will result in too much noise and will strongly affect the recognition performance. This problem limits the applicability of our system in scenarios involving recognition at distance, such as video surveillance and access control of large areas (*e.g.* hallways, as in Figure 2.5(a)). However, the Kinect can

Figure 2.3: Kinect pattern after repetition.

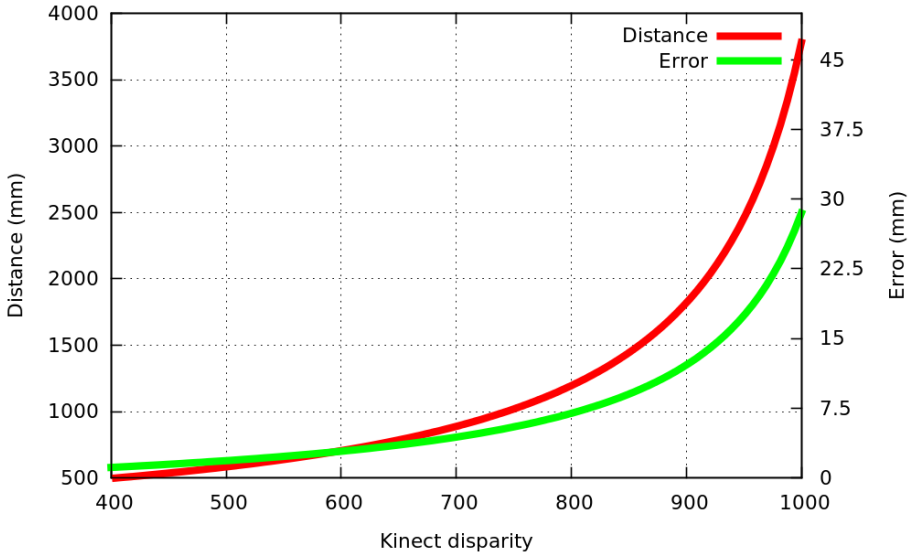


Figure 2.4: Kinect disparity (*i.e.* values returned by the Kinect) versus real distance and average error in millimeters.

be used in combination with the current infrastructure of stadiums, airports and other venues, that use turnstiles (see Figure 2.5(b)) or similar devices to control the access of a large number of people. With the evolution of 3D sensors, our system might then be adapted to cover larger areas. With the Kinect, we can also protect vital/expensive resources, such as computers and medical equipment.

(a) (b)

Figure 2.5: Types of access control: (a) hallways – image from *http://www.openphoto.net*; and (b) turnstiles – image from *http://www.sxc.hu*.

2.2 Face detection

The objective of a face detection method is to determine whether there are faces in an input image and their respective location. This is a key step in any fully automatic face recognition system [Zhao et al., 2003], and several methods were proposed in the literature to solve this problem using both 2D and 3D images or only 3D images.

Some works considered a controlled acquisition environment to correctly locate the face [Chang et al., 2006, Lu and Jain, 2006, Mian et al., 2007, Pamplona Segundo et al., 2010, Tsalakanidou et al., 2005]. Some of the employed constraints are: (1) each image must have only one face; (2) faces must be frontal; (3) faces must be close to the camera; and (4) cluttered background is not allowed. In Figures 2.6(a)-2.6(e) we illustrate methods that take at least one of these constraints into account. Chang et al. [2006] used connected components in depth images and skin color classification to locate and extract

faces (see Figure 2.6(a)). Mian et al. [2007] analyzed peaks in horizontal slices of depth images to locate the nose tip (see Figure 2.6(b)) and then considered the region around it as the face. Pamplona Segundo et al. [2010] and Tsalakanidou et al. [2005] combined depth clustering and shape detection in different ways for facial segmentation purposes (see Figures 2.6(c) and 2.6(d)). Finally, Lu and Jain [2006] used horizontal and vertical histograms to estimate the face position (see Figure 2.6(e)). All these methods obtained very accurate results for images that meet their respective constraints. However, recent 3D sensors (*e.g.* Kinect) have a larger area of capture than the previous ones. For this reason, they can easily capture depth images that contain multiple subjects at different distances or cluttered background, so the constraints listed above are overridden.

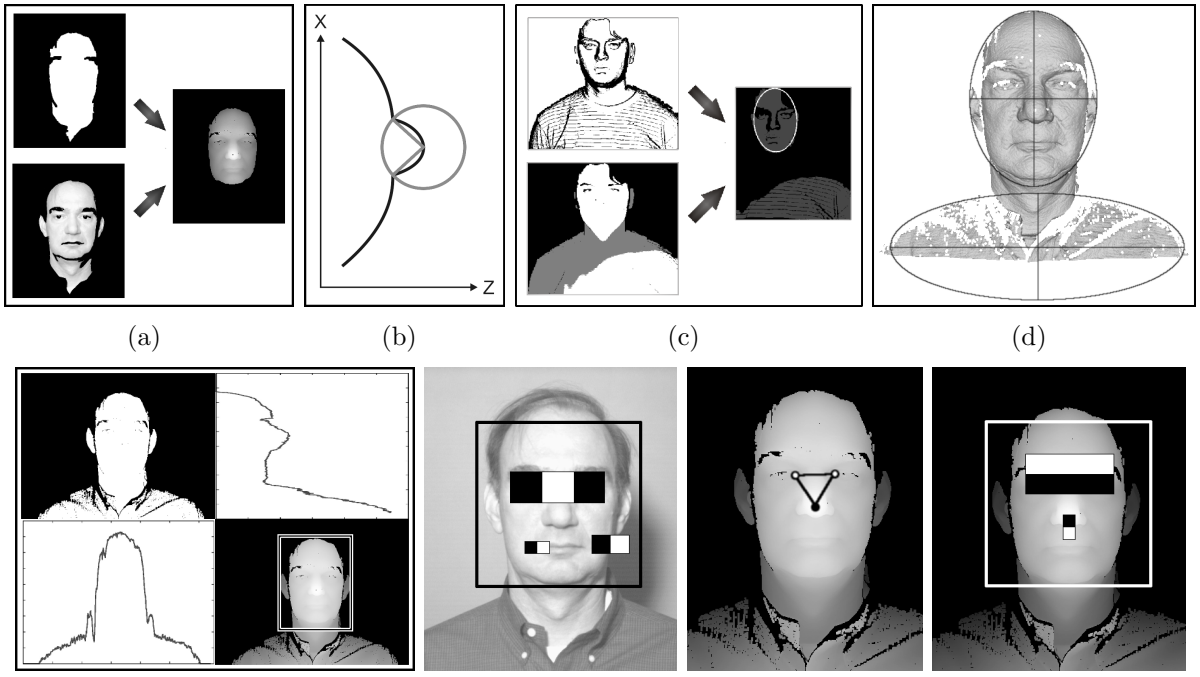


Figure 2.6: Face detection/extraction methods considering (a)-(e) controlled and (f)-(h) uncontrolled acquisition environments.

A more flexible acquisition environment is considered in other works [Böhme et al., 2009, Colombo et al., 2006, Fischer et al., 2010, Lu and Jain, 2005, Wang et al., 2010], illustrated in Figures 2.6(f)-2.6(h), and are consequently more suitable to these new sensors. Lu and Jain [2005] used Viola and Jones' 2D face detector [Viola and Jones, 2004] to find faces in the color image (see Figure 2.6(f)) and then extracted the correspondent

region from the depth image. Wang et al. [2010] extended the previous method by using a pose invariant detector [Huang et al., 2007]. Colombo et al. [2006] employed surface curvature analysis to find possible face regions (see Figure 2.6(g)) and then used Principal Component Analysis to classify these regions as face or non-face. Finally, Böhme et al. [2009] and Fischer et al. [2010] used Viola and Jones's detector on depth images (see Figure 2.6(h)) and combined the results from both color and depth detectors in different ways.

Table 2.1 summarizes all mentioned works according to the following attributes: ability to detect multiple faces; ability to handle cluttered background; robustness to pose variations; independency to lighting conditions; and real-time performance (*i.e.* ability to process multiple images per second). These attributes are not always explicitly defined for all works, so we included our opinion based on our understanding of the method (*e.g.* methods that use color images can be affected by intense lighting variations, so they are not independent of illumination; methods that look for an ellipse cannot detect multiple faces or handle cluttered background). As may be seen, none of the mentioned works received a positive evaluation for all attributes. Although Colombo et al. [2006] dealt with all the problems that can be solved/eased by 3D data, their method cannot process multiple images per second, which is a desirable attribute for many applications, such as face modeling [Hernandez et al., 2012], identification [Min et al., 2012] and continuous authentication [Pamplona Segundo et al., 2013a].

We present a new 3D face detector that handles all variations presented in Table 2.1 in real-time, based on Viola and Jones' detector. To this end, we make use of boosted cascade classifiers [Viola and Jones, 2004] to detect faces using depth data. Böhme et al. [2009] and Fischer et al. [2010] followed a similar path, but unlike these methods, we do not use depth images only as an additional source of texture information. Faces from different subjects have similar size in real world values, so the size information is extremely relevant for face detection [Pamplona Segundo et al., 2011, 2013b]. Since it is possible to take advantage of this fact when using depth images, we introduce the use of orthogonal projection images to represent faces with a fixed size disregarding their distance to the

Table 2.1: Classification of face detection/extraction methods in the literature according to the following attributes: ability to detect multiple faces (MULT); ability to handle cluttered background (BACK); robustness to pose variations (POSE); independency to lighting conditions (LIGH); and real-time performance (TIME).

Method	**MULT**	**BACK**	**POSE**	**LIGH**	**TIME**
Chang et al. [2006]	No	No	Yes	No	No
Mian et al. [2007]	No	No	No	Yes	Yes
Pamplona Segundo et al. [2010]	No	No	No	Yes	No
Tsalakanidou et al. [2005]	No	No	No	No	No
Lu and Jain [2006]	No	No	Yes	Yes	Yes
Lu and Jain [2005]	Yes	Yes	No	No	Yes
Wang et al. [2010]	Yes	Yes	Yes	No	Yes
Colombo et al. [2006]	Yes	Yes	Yes	Yes	No
Böhme et al. [2009]	Yes	Yes	No	Yes	Yes
Fischer et al. [2010]	Yes	Yes	No	No	Yes
Proposed method	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

acquisition sensor. By doing so, we improve the speed and the accuracy because the number of face candidates to be tested is considerably reduced.

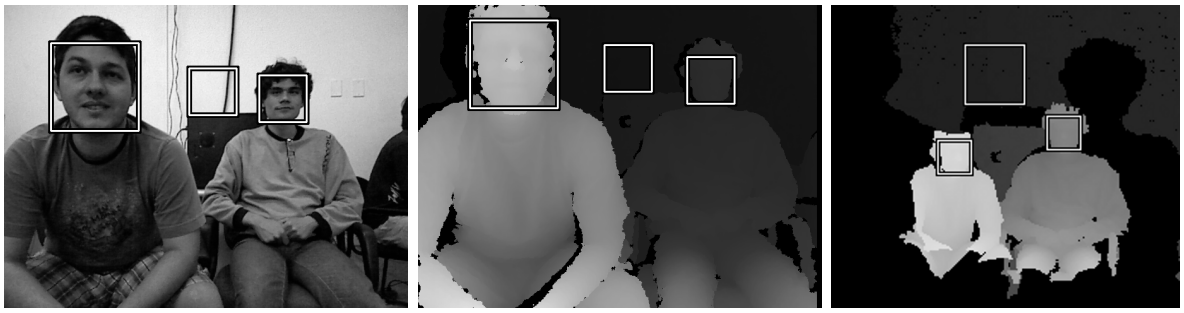
A brief explanation about Viola and Jones' detector is given in Section 2.2.1, since it has inspired our detector. A detailed explanation of the proposed 3D face detector is presented in Section 2.2.2.

2.2.1 Viola and Jones' 2D face detector

The method described by Viola and Jones [2004] is based on Haar features, which are rectangular masks with different size, shape and location. These features return similar values when applied to images with the same pattern, such as faces. However, a single feature is not enough to discriminate complex patterns, so several features are combined in order to obtain a stronger classifier. To select the set of Haar features that will be used in the classifier, face and non-face samples are required for training. All possible features are evaluated using these training samples and the Adaboost algorithm [Freund and Schapire, 1995] is employed to iteratively select the most discriminative features. In the first iteration, all samples are equally considered during the feature selection. After that, the Adaboost gives more weight to samples that were misclassified in the previous

iterations to prioritize the unsolved part of the problem.

The detection process consists of scanning an input image with a sub-window and classifying the sub-window region as face or non-face. Faces may present different sizes due to the perspective distortion, as illustrated in Figure 2.7(a), so the sub-window must be scaled to all possible sizes in order to detect faces with any size. As may also be seen in Figure 2.7(a), a typical input image has much more non-face regions than face regions. For this reason, the set of Haar features is divided into multiple stages that are sequentially applied to each face candidate, and one candidate must be accepted by all stages to be considered a face. By doing this, the overall cost of the detection process is substantially reduced since non-face regions are usually discarded in the initial stages.



(a) (b) (c)

Figure 2.7: Illustration of the detection process using different input images: (a) color images, (b) depth images and (c) orthogonal projection images.

This method can be employed to detect different objects in different types of image, like faces in color [Viola and Jones, 2004], infrared [Li et al., 2007] and depth images [Böhme et al., 2009], pedestrians [Munder and Gavrila, 2006], vehicles [Vural et al., 2012] and so on.

2.2.2 Proposed approach

The size of an object is a very useful information to discard non-object candidates in the detection process, which is also true for faces. However, 2D face detectors cannot use such information due to the perspective distortion. This distortion causes a change in the size of the face according to its distance to the acquisition device, as may be observed in both color and depth images shown in Figures 2.7(a) and 2.7(b) respectively. These figures

also show a non-face region that has the same size of one of the detected faces, thus the size cannot be used to eliminate it.

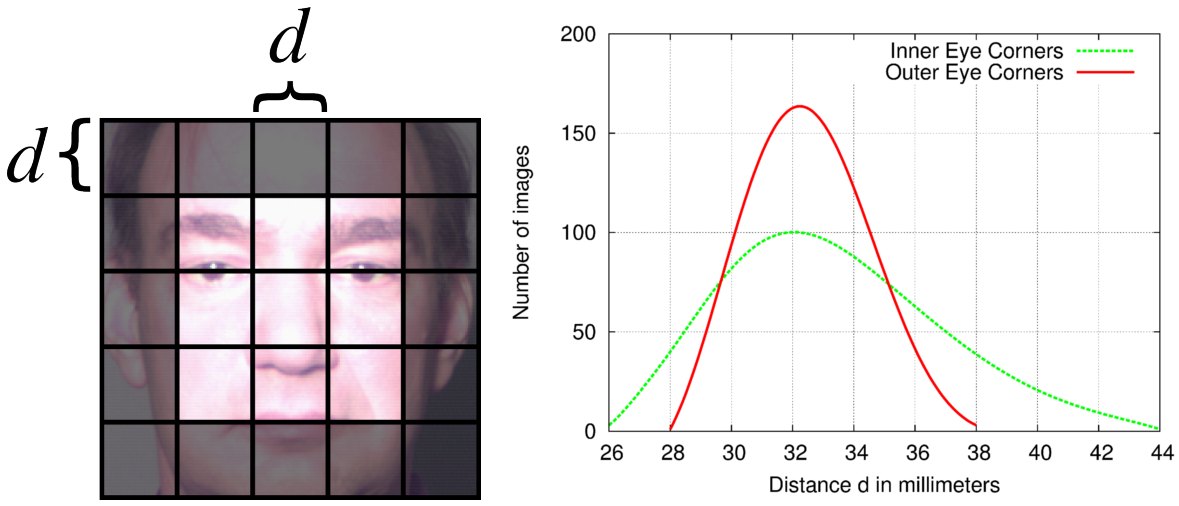
We propose using the face size as a way to eliminate the need to look for faces at different scales, which is done by using orthogonal projection images. Objects with similar size in real world have similar size in such images, as may be seen in Figure 2.7(c), and we no longer need to scan images at multiple scales to detect all faces, we only need to evaluate regions with the face size. By doing this, the same non-face region shown in Figures 2.7(a) and 2.7(b) is not even considered for evaluation in Figure 2.7(c).

To accomplish this, four steps are required: (1) face size estimation; (2) generation of orthogonal projection images; (3) training; and (4) detection. More details about each step are given below.

Face size estimation

According to the neoclassical canon of facial proportions [Prendergast, 2012], one face can be vertically divided in fifths, where one-fifth is the width of an eye or the distance between inner eye corners (see Figure 2.8(a)). Although there are some variations in these proportions, this relation holds quite well for different ethnicities, genders and ages. This work is built on the premise that faces from different subjects have a similar size, so the size of one-fifth cannot vary too much. To show that, we computed the distance d, with d being the size of one-fifth, for all training images of the Face Recognition Grand Challenge (FRGC) database [Phillips et al., 2005] (*i.e.* details about this database are given in Section 3.1.1) in two different ways: (1) the distance between inner eye corners; and (2) one third of the distance between outer eye corners. The ground truth location of these landmarks were used to this end, and the results are shown in Figure 2.8(b). As may be seen, the average d value is about the same in both cases (*i.e.* around 33mm). Also, the variation of the d value is very small, especially when using the distance between outer eye corners, which takes three-fifths into account.

We consider the face as a $5d \times 5d$ region, as illustrated in Figure 2.8(a). On average, $5d$ is equal to 165mm, so this value was used as face size in this work. This value is



(a) (b)

Figure 2.8: (a) Illustration of the face size employed in this work, which is equal to $5d \times 5d$; and (b) the histogram of the distance d, measured from inner and outer eye corners for all FRGC training images.

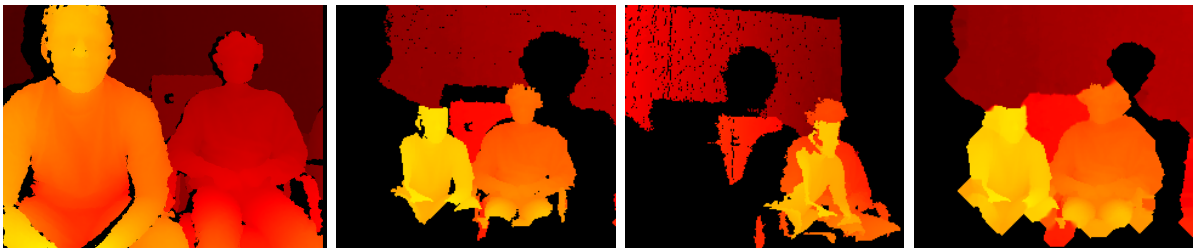
not critical and may be applied to other databases, since it is an attribute of human adult faces and is not peculiar to the FRGC database. However, a smaller size should be considered to detect infant faces. For example, the average d value for newborns is around 22mm [Omotade, 1990], so the face size in this case should be 110mm.

Orthogonal projection images

An orthogonal projection image g is a 2D representation of a 3D image f (*i.e.* f can be a depth image, a surface mesh or a point cloud as long as 3D point coordinates, which are in relation to the sensor coordinate system, are in real world scale). It can be interpreted as a depth image taken by a camera located infinitely away from the scene with an infinite focal length. So, although g looks like a depth image, objects on it are not affected by the perspective distortion (see Figures 2.9(a) and 2.9(b)).

It is obtained by transforming two axes of a 3D point $p = [X_p, Y_p, Z_p]^T$ into row and column and the remaining axis into the pixel value, for all points in f, according to Equations 2.1-2.4:

$$p' = R(p - \bar{p}) \tag{2.1}$$



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

Figure 2.9: (a) Depth image; (b)-(c) orthogonal projection images from different viewpoints; and (d) hole filling result for (b).

$$row_p = \frac{H_g}{2} - \frac{Y_{p'}}{r} \tag{2.2}$$

$$col_p = \frac{W_g}{2} + \frac{X_{p'}}{r} \tag{2.3}$$

$$g(col_p, row_p) = \frac{Z_{p'}}{r} \tag{2.4}$$

where $\bar{p}$ is the average of all points in f, R is a 3D rotation matrix, H_g and W_g are respectively the height and the width of g in pixels, and r is the resolution of g. Equations 2.2 and 2.3 give the coordinates of p in g, and Equation 2.4 gives the pixel value in this location. Equation 2.1 centralizes f and may also rotate it to obtain orthogonal projection images from different viewpoints. The values of H_g, W_g and $\bar{p}$ can be estimated based on f or can be predetermined if the area of capture of the sensor is known. The resolution r is the only parameter to create orthogonal projection images, and it defines the size of the pixel in mm. The r value is obtained by the following equation:

$$r = \frac{n}{n'} \tag{2.5}$$

where n is the face size in the real world (*i.e.* 165mm, as defined during the estimation of the face size) and n' is the desired face size in g. For example, to have faces with 33×33 pixels in g, r must be equal to 5.

Since our orthogonal projection images are created by a forward mapping, more than

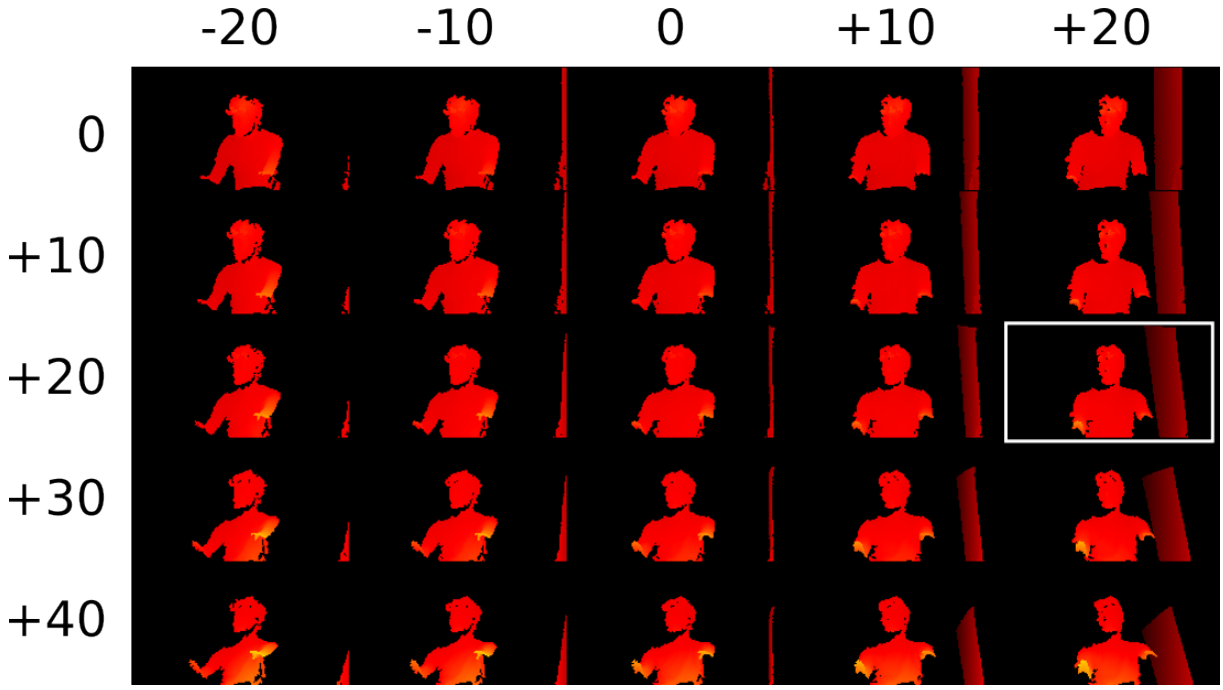
one point in f may be mapped to the same pixel in g, and some pixels in g may not have a value. To solve the first problem, we take the closest point to the sensor among all points mapping the same pixel. To fill blank pixels, we assign the nearest neighbor pixel value to them if they are close enough to valid pixels (*i.e.* distance smaller than the face size), as shown in Figure 2.9(d). Both problems could be solved at the same time by using an inverse mapping, which consists in finding the corresponding point in f for each pixel in g. However, the inverse mapping cannot be done as fast as the forward mapping because it requires searching correspondences. Nevertheless, inverse mapping would be the best option for a parallel implementation, since it solves both problems in a single step and also replaces write collisions (*i.e.* mapping points to the same pixel in g) with read collisions (*i.e.* finding correspondences in f simultaneously).

Figures 2.10 and 2.11 show multiple orthogonal projection images from different viewpoints of a same input image. This artifice can be used to detect faces under pose variation without having a classifier for rotated faces. As may be seen, in Figure 2.10(a) the subject is looking to his right. His face looks frontal when we rotate the scene to the left, as may be seen in the right side of Figure 2.10(b). Figure 2.11(a) shows another example where the subject is now looking down. In this case, his face looks frontal when we rotate the scene up, which is shown in the center of Figure 2.11(b). One problem that may occur is nearer objects occluding farther ones (*e.g.* the closest subject occludes the other one in Figure 2.9(c)), which can make some faces undetectable without further processing.

Training step

Efficient and effective classifiers were successfully obtained for face images with sizes ranging from 18×18 to 24×24 pixels in the literature [Lienhart and Maydt, 2002, Viola and Jones, 2004]. So, in this work we used face images with 21×21 pixels, with $r \approx 7.9$ according to Equation 2.5. Orthogonal projection images were created for all 943 images of the FRGC training set, and the ground truth location of the outer eye corners was employed to extract 21×21 face samples. Other parts of the orthogonal projection images were used as non-face samples together with a set of 3,019 intensity

(a)



(b)

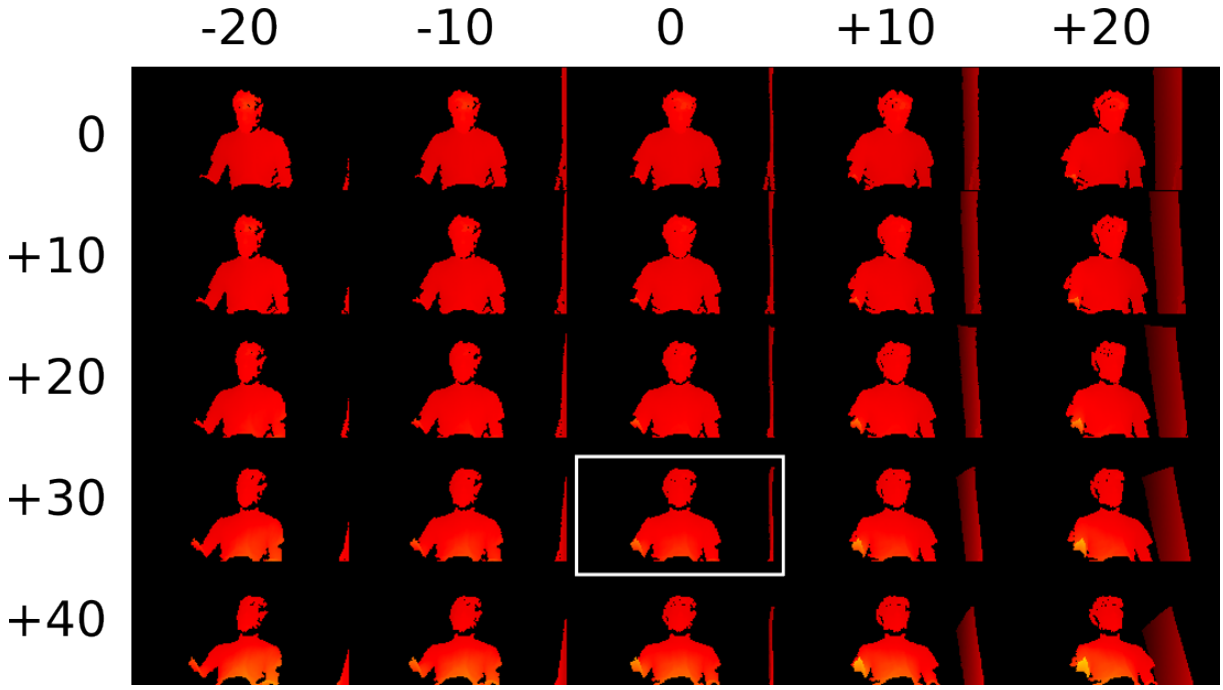
Figure 2.10: (a) Texture image of the original pose, and (b) resulting orthogonal projections from 25 different viewpoints. In this example, the subject is looking to his right and his face looks frontal when we rotate the scene to the left, as indicated by a white square.

non-face images[2]. These intensity images were included as non-face samples to make the classifier more robust against unknown patterns.

The OpenCV library[3] was employed to create our cascade classifier. The extended set of Haar features was used [Lienhart and Maydt, 2002], and target detection and false positive rates were respectively set to 0.999 and 10^{-6}. The target was reached after

[2] `http://tutorial-haartraining.googlecode.com/svn/trunk/data/negatives/`
[3] `http://opencv.org/`

(a)



(b)

Figure 2.11: (a) Texture image of the original pose, and (b) resulting orthogonal projections from 25 different viewpoints. In this example, the subject is looking down and his face looks frontal when we rotate the scene up, as indicated by a white square.

selecting 32 features divided among six stages. Our classifier is far smaller than the state-of-the-art one for 2D face detection available in OpenCV[4], which has 22 stages and more than $2,000$ features. This is achieved because 3D faces are more invariant than 2D faces, considering both intersubject and intrasubject variability (*i.e.* skin color, facial hair and illumination are more troublesome in 2D face images). The reduced size of the classifier also leads to a gain in speed.

[4]The following classifier was used: haarcascade_frontalface_alt.xml (OpenCV-2.4.1)

Detection step

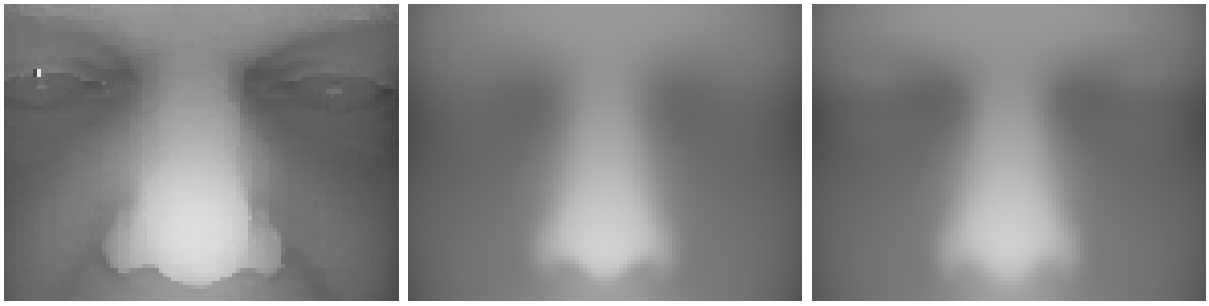
A 21×21 sub-window is employed in the detection stage to scan orthogonal projection images looking for faces. There is no need to scan the image using sub-windows with other sizes. To deal with multiple detections, we perform two steps: first, all detections are divided into disjoint groups of detections by analyzing the overlapping area between detections (*i.e.* at least 70% of overlap is required to group two detections); then, we compute the median location for each group and use it as the final location of this group. As in Viola and Jones [2004], a minimum number of detections in a disjoint group can be used as a threshold to distinguish between face and non-face groups, which may help to eliminate false positives. In this work, this threshold varies based on the number of viewpoints used to create orthogonal projections images. We have empirically found that the threshold value is 1 for 1–10 viewpoints, 2 for 11–50 viewpoints and 3 for 50+ viewpoints.

2.3 Face normalization

The goal of a face normalization method is to standardize the pose and resolution of a face image to help further analysis of it. There are different ways of performing this task, and three of them were more prevalent in the literature: (1) facial landmarks detection [Pamplona Segundo et al., 2007, 2010], which is used to pre-align facial surfaces [Lu et al., 2006] or to move a facial surface to a standard location [Mian et al., 2007]; (2) registration to a reference model [Chang et al., 2006], which consists in aligning an input facial surface to a generic face model; and (3) fitting a deformable model [Kakadiaris et al., 2007], in which a model is transformed to reflect the shape of an input facial surface. All of them have advantages and disadvantages, but the second method is computationally cheaper and is also straightforward applicable to video sequences (*i.e.* it does not require any processing other than its own to be optimized for video). For these reasons, we use a method based on the registration to a reference model as the normalization method, as presented in Section 2.3.1.

2.3.1 Normalization through registration to a reference model

The registration to a reference model consists in aligning an input facial surface to a generic face model, and the first step is to obtain a generic face to be used as a reference. A noise-free average face $\mathbf{\Psi}$ is computed using the images of the FRGC training set, which contains 943 images from 275 different subjects. To this end, first all training images $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2, \ldots, \mathbf{\Gamma}_M$ are aligned to an initial estimation $\mathbf{\Psi}'$ using the Iterative Closest Point (ICP) algorithm. $\mathbf{\Psi}'$ is the first training image resampled on a uniform grid with resolution of 1mm, as shown in Figure 2.12(a). The grid is centered in the nose and eyes area and has size of 96×72mm, totaling 97×73 points. This face region was chosen because it is not as affected by facial expressions as other parts of the face [Chang et al., 2006, Queirolo et al., 2010], making the normalization process robust to such variations.



(a) (b) (c)

Figure 2.12: Iterative computation of the average face: (a) initial estimation, (b) result after the first iteration and (c) result after convergence.

After the alignment, we have the images $\mathbf{\Gamma}'_1$, $\mathbf{\Gamma}'_2$, ..., $\mathbf{\Gamma}'_M$ in the same coordinate system of $\mathbf{\Psi}'$. With these images we compute the residual vectors $\mathbf{\Phi}'_i = \mathbf{\Gamma}'_i - \mathbf{\Psi}'$, where each value in $\mathbf{\Phi}'_i$ is the distance in the Z-axis between one point in $\mathbf{\Psi}'$ and its closest point in $\mathbf{\Gamma}'_i$. Then, we recompute $\mathbf{\Psi}'$ using the Equation 2.6 and repeat the entire process until convergence (*i.e.* convergence is achieved when the standard deviation of the aligned training images stops decreasing), which took 12 iterations. Finally, the last $\mathbf{\Psi}'$ is assigned to $\mathbf{\Psi}$, and the resulting average face is presented in Figure 2.12(c).

$$\mathbf{\Psi}' = \mathbf{\Psi}' + \frac{1}{M}\sum_{i=1}^{M} \mathbf{\Phi}'_i \tag{2.6}$$

Once we have a reference face, any detected face is aligned to it using the ICP algorithm [Besl and Mckay, 1992] to standardize the pose, and a uniform grid sampling with resolution of 1mm is employed to standardize the resolution. ICP correspondences are obtained by the project-and-walk strategy [Rusinkiewicz and Levoy, 2001], and the transformation between corresponding points is computed using orthonormal matrices [Horn et al., 1988]. The transformation obtained for one frame can be used as an initial estimation for the next frame, making the normalization process much faster in video sequences. Figure 2.13 shows some examples of the normalization process for faces with pose variation. As may be seen, the results are very consistent even though the rotation angles are very different.

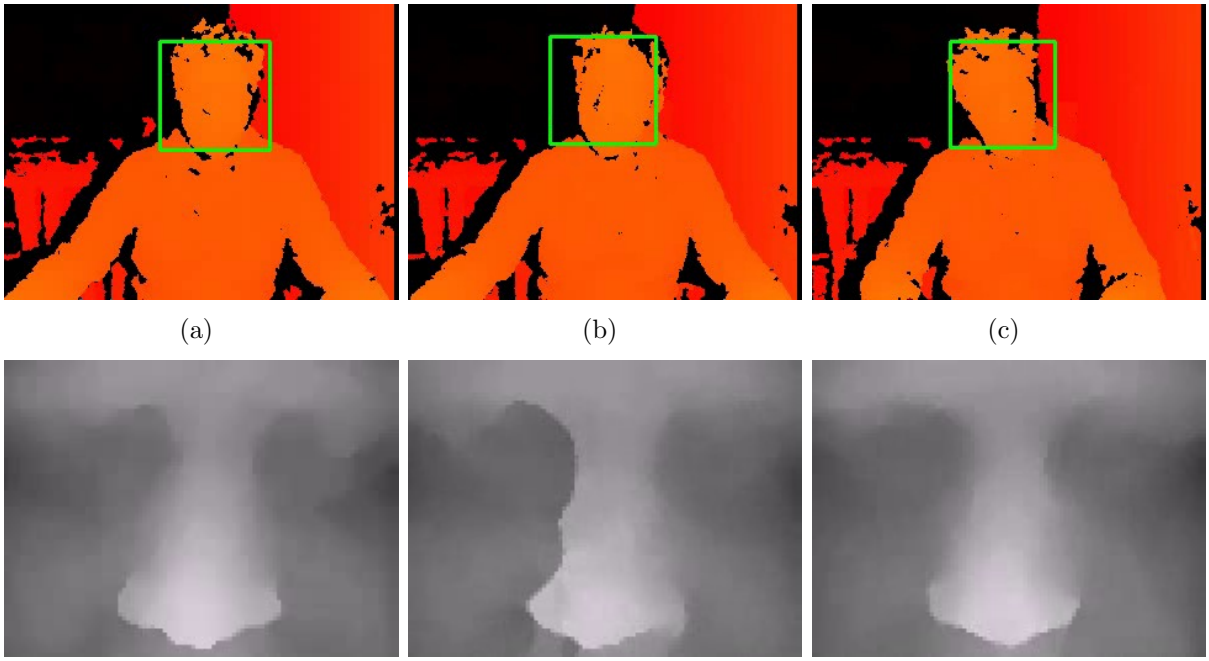


Figure 2.13: Normalization results: (a)-(c) detected faces and their respective (d)-(f) normalized images.

2.4 Face description

The goal of the description module is to create a biometric template using discriminant features of the normalized face image. The normalized image by itself can be used as biometric template, but its dimensionality is high (*i.e.* the template would have 7081

dimensions for a 97×73 image). Turk and Pentland [1991] presented one of the most successful description methods in the literature, where the Principal Component Analysis (PCA) [Pearson, 1901] is used to compress a normalized face image with minimal data loss in order to minimize the matching time. This method, however, does not differ between discriminative and non-discriminative image information. To address this problem and also keep the dimensionality low, Belhumeur et al. [1997] used the Linear Discriminant Analysis (LDA) [Fisher, 1936]. In this method, the focus is not only in compressing the image, but also in maximizing the discriminativeness of the template. On the other hand, it requires a much larger training set to work properly. Both methods were then applied to 3D images [Chang et al., 2005, Hiremath and Hiremath, 2013].

PCA and LDA can also be used after other descriptors in order increase their compression and eventually increase their discriminative power [Déniz et al., 2011, Ocegueda et al., 2011]. Among them, the Histogram of Oriented Gradients (HOG) [Dalal and Triggs, 2005] has shown to be a very efficient and effective descriptor for face recognition in color images [Déniz et al., 2011]. HOG descriptors are not indicated when there are orientation variations, but this is not the case for our normalized images. Also, HOG descriptors showed themselves more invariant than the normalized image [Pamplona Segundo et al., 2013a], which supports our choice of this method for face description.

2.4.1 Histogram of Oriented Gradients for face description

HOG descriptors follow the idea that images can be described by the distribution of its gradients. Other descriptors like Scale-Invariant Feature Transform (SIFT) [Lowe, 1999] and Speeded Up Robust Features (SURF) [Bay et al., 2006] have a similar idea, but they are only used to describe local points. Unlike SIFT and SURF, HOG describes the appearance and shape of an entire object by dividing the image into cells and computing a histogram of gradients for each cell, and then scanning the image with a block and normalizing the histogram of the cells inside it. The normalized histograms are concatenated to form the HOG descriptor. In this work, the input image is scaled to 64×64 pixels, and the HOG is computed using cells of size 8×8 pixels with histograms of 9 gradient bins.

The block has size 2×2 cells, and it scans the image with a step size of 1 cell. In the end, we have a descriptor with 1764 elements to be used as template, which is 75% smaller than the size of the normalized image.

2.5 Face matching

The matching stage consists in comparing the biometric template from a probe image to one or more biometric templates previously obtained. The result of the matching must reflect how close two templates are, and it should return a high similarity value for templates from the same subject and a low similarity value for templates from different subjects. Some works have performed this task by registering two facial surfaces and then measuring how well they fit each other [Chang et al., 2006, Lu et al., 2006, Queirolo et al., 2010]. Although this process is very effective in one-to-one (1:1) recognition mode, it is time consuming and does not allow real-time responses, specially in one-to-many (1:N) recognition mode. To achieve an acceptable performance in 1:N mode, other works represent faces through histograms, vectors or normalized images in order to make the matching process easier [Chang et al., 2005, Kakadiaris et al., 2007, Mian et al., 2007, Pamplona Segundo et al., 2013a]. In this case, it is possible to use simple distance measures to compute the similarity between biometric templates. Among many possible distances [Deza and Deza, 2009], some of them are more common in face recognition works, such as Euclidean distance, Manhattan distance, and Mahalanobis distance [Moon and Phillips, 2001]. In this work, we use the Manhattan distance, also known as L_1 distance, for two reasons: first, it does not require previous knowledge about the distribution of the biometric templates, unlike the Mahalanobis distance; second, it is not so influenced by noise, unlike the Euclidean distance which penalizes large residuals much more than small ones. Figure 2.14 shows a comparison in terms of recognition rate when using Euclidean, Manhattan and Mahalanobis distances to match images from the FRGC training set. As may be seen, the Manhattan distance presents slightly better results in comparison to Euclidean and Mahalanobis distances, and its computational cost is also lower.

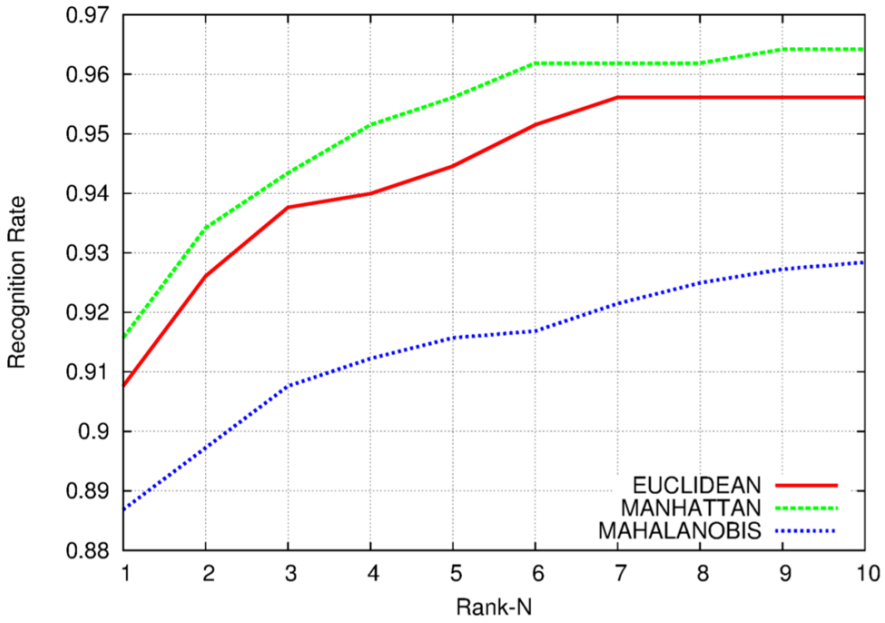


Figure 2.14: Recognition results using Euclidean, Manhattan and Mahalanobis distances.

2.5.1 Manhattan distance for matching

Given a HOG descriptor $H_p = \{h_1^p, h_2^p, \ldots, h_N^p\}$ from a probe template p, where N is the size of the descriptor, the Manhattan distance between p and a gallery template g is given by Equation 2.7:

$$L_1(H_p, H_g) = \sum_{i=1}^{N} |h_i^p - h_i^g| \tag{2.7}$$

This distance can then be used to discover if both templates are from the same subject or not, since templates from the same subject usually obtain lower L_1 distances than templates from different subjects.

2.6 Evaluation

In the evaluation stage, the matching results are analyzed and a decision on the identity is made. Different procedures are adopted in 1:1 and 1:N scenarios. In a 1:N recognition mode, what is usually done is to rank the gallery templates according to their similarity to the probe template and then use 1:1 evaluation methods for the top candidates. In a 1:1 recognition mode, the distribution of matching distances between biometric tem-

plates from the same subject and from different subjects of a training set, illustrated in Figure 2.15, can be used to define a classification rule that distinguishes genuine and impostor matchings.

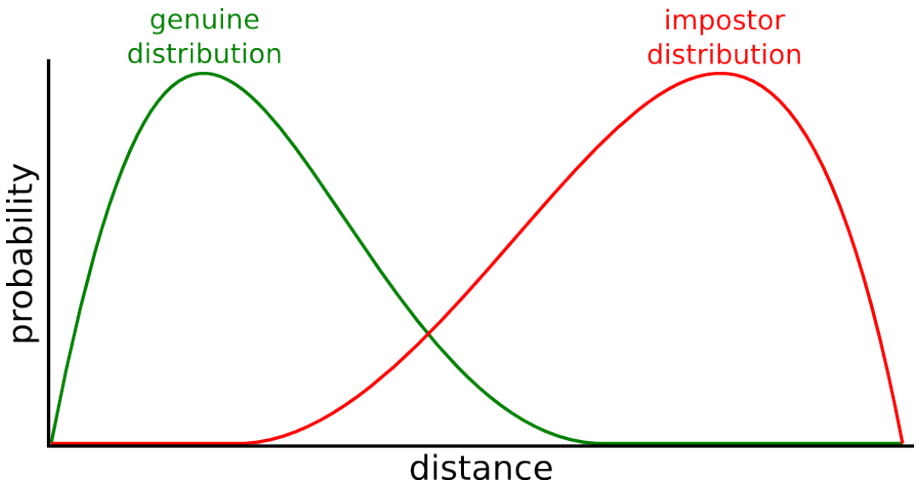


Figure 2.15: Illustration of the distribution of the matching distances between biometric templates from the same subject (genuine) and from different subjects (impostor).

The simplest classification rule that can be applied is a thresholding operation, as illustrated in Figure 2.16. In this case, a threshold t is defined and matching distances below t are classified as genuine and equal or above t are classified as impostor. Genuine matchings misclassified as impostor are evaluated through the False Rejection Rate (FRR), and impostor matchings misclassified as genuine are evaluated through the False Acceptance Rate (FAR), as also illustrated in Figure 2.16. In an ideal biometric system, both FRR and FAR would be equal to zero, but in practice they are inversely proportional. That means that one increases when the other decreases, and if one of them is equal to zero, the other will probably be very high. The threshold is used to set the relation between FRR and FAR. A low threshold may lead to a highly secure system, in which an impostor matching is hardly going to be classified as genuine at the cost of misclassifying many genuine matchings. On the other hand, a high threshold may lead to a permissive system that misclassifies many impostor matchings. When we have the same value for FRR and FAR, this error is called Equal Error Rate (EER). The lower the EER, the higher is the performance of the system.

Another classification rule consists in converting every distribution of matching dis-

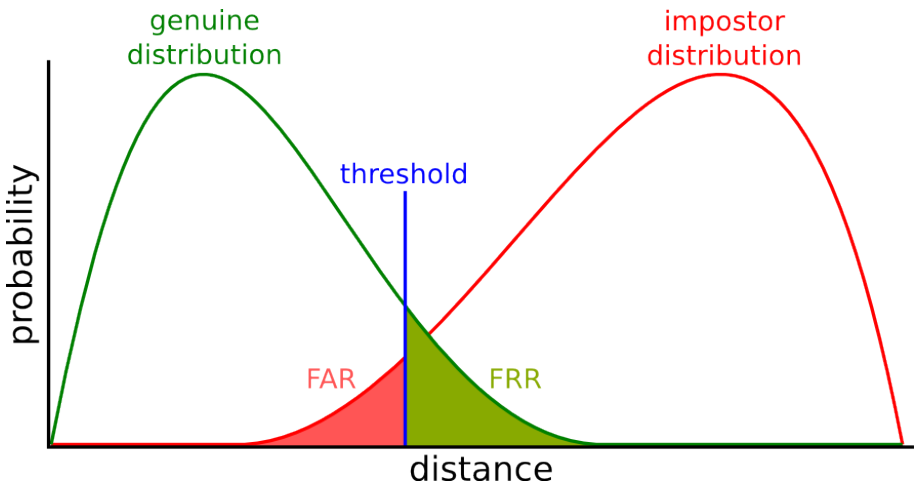


Figure 2.16: Illustration of the classification of genuine and impostor matchings through a threshold.

tances (see Figure 2.15) into a Cumulative Distribution Function (CDF), as illustrated in Figure 2.17, and then using the obtained CDFs to compute the probability of a genuine matching and the probability of an impostor matching, respectively called $P(d \mid genuine)$ and $P(d \mid impostor)$, with d being the matching distance. If $P(d \mid genuine)$ is greater than $P(d \mid impostor)$ we consider the matching as genuine, otherwise, it is considered an impostor matching. In this classification rule, we always have the same value for FRR and FAR.

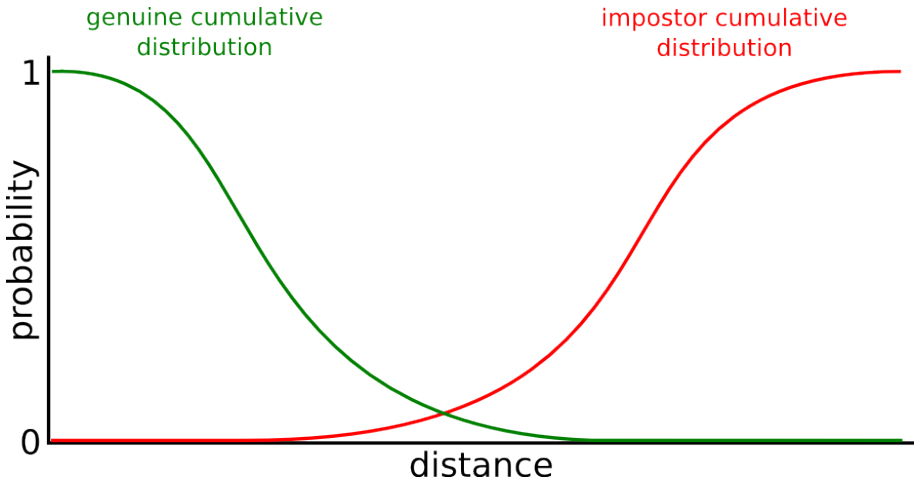


Figure 2.17: Illustration of the classification of genuine and impostor matchings through a threshold.

We use CDFs as the evaluation method because converting matching distances into probabilities is very useful when dealing with multiple probes and CDFs allow weighting

results according to their discriminativeness. Since our system capture up to 30 instances of the same face every second, it is very important to combine the results in a way that the best ones are prioritized.

2.6.1 Cumulative Distribution Functions for matching evaluation

A matching distance d can be classified as genuine or impostor, which respectively represent scores resulting from matching biometric templates from the same subject and from different subjects, by using a CDF for each possibility. To this end, each CDF can be represented as a vector with all possible distances and their respective probabilities, or as a mathematical function, which is a more concise representation. In this work, $P(d \mid genuine)$ and $P(d \mid impostor)$ are respectively given by Equations 2.8 and 2.9:

$$P(d \mid genuine) \propto 1 - \frac{1}{2}\left[1 + erf\left(\frac{d - \mu_{genuine}}{\sigma_{genuine}\sqrt{2}}\right)\right] \tag{2.8}$$

$$P(d \mid impostor) \propto \frac{1}{2}\left[1 + erf\left(\frac{d - \mu_{impostor}}{\sigma_{impostor}\sqrt{2}}\right)\right] \tag{2.9}$$

where $\mu_{genuine}$ and $\mu_{impostor}$ are the average matching distances for genuine and impostor matchings, with $\sigma_{genuine}$ and $\sigma_{impostor}$ being their respective standard deviations.

The values for $\mu_{genuine}$, $\mu_{impostor}$, $\sigma_{genuine}$ and $\sigma_{impostor}$ can be discovered using a training set, or through a parameter search that maximizes the recognition results when there is no training set available.

2.7 Application in Continuous Authentication

For many years biometrics have been proposed as a substitute for common authentication methods, such as passwords and tokens [Bolle et al., 2003]. However, in most authentication systems, once someone gets access to the desired resource no further verification is performed. Although these systems stop an unauthorized individual from getting access,

they cannot ensure that the accessing user is the allowed one, which is not acceptable in high security environments. The continuous authentication addresses this issue by constantly monitoring accessing users to make sure no unauthorized access occurs after the initial verification, as illustrated in Figure 2.18. As may be observed, after capturing some initial samples as biometric template, the identity is continuously verified through subsequent samples. Its major advantage is to provide a more secure session, which may be used in computer access control [Monaco et al., 2012], online examinations [Flior and Kowalski, 2010] or to protect health information[5], and it only requires biometric samples to be captured continuously.

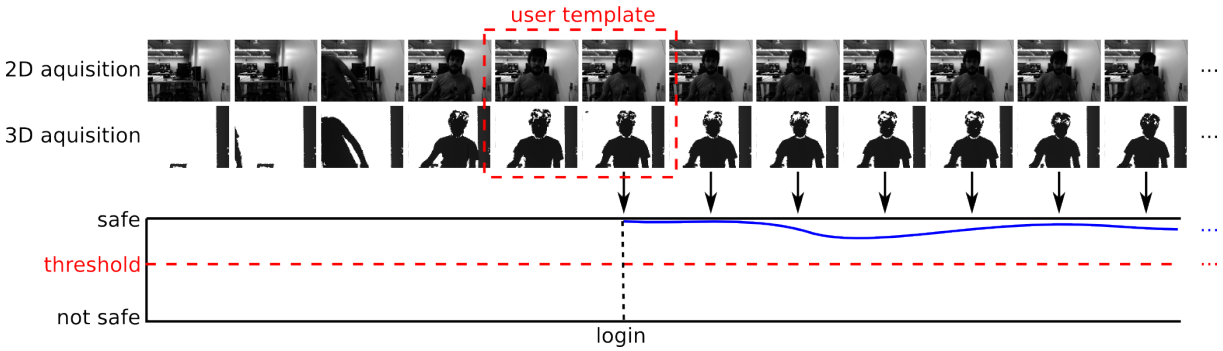


Figure 2.18: Illustration of the operation of a continuous face authentication system based on face images.

In this context, keystrokes appeared as the most straightforward feature for continuous authentication and were the first biometric feature used for this purpose [Leggett et al., 1991, Gunetti and Picardi, 2005, Monaco et al., 2012]. Although the use of keystrokes for continuous authentication does not require additional hardware in a traditional computer configuration, according to Monaco et al. [2012] it requires more than 200 keystrokes to identify an impostor (*i.e.* at least one minute considering an average computer user). However, as pointed out by Sim et al. [2007], impostors can damage a protected system with much less effort (*e.g.* the command line **"rm -rf *"** in a Linux console can be typed in a few seconds). To overcome this problem, different biometric features with a higher discriminant power were employed, such as electrocardiograms (ECG) [Agrafioti and Hatzinakos, 2009], faces [Janakiraman et al., 2005, Niinuma et al., 2010] and finger-

[5] `http://www.imprivata.com/products-and-solutions/authentication-management/onesign-secure-walk-away`

prints [Sim et al., 2007], as well as multimodal systems [Altinok and Turk, 2003, Damousis et al., 2008, Sim et al., 2007]. Despite the advantages in accuracy, fingerprint-based systems cannot obtain samples continuously without user cooperation making the continuous authentication too inconvenient for the user, and ECG biometrics require users to wear body sensors and can reveal other information than the identity (*e.g.* health conditions such as arrhythmia [Agrafioti and Hatzinakos, 2009] and stress).

Facial images can be captured without any user cooperation by low-cost cameras, which are built-in in most of today's computers. However, face recognition based on 2D images is substantially affected by pose, illumination and facial expression variations [Zhao et al., 2003]. To avoid these variations Niinuma et al. [2010] introduced the concept of soft biometrics, which are color distributions of faces and clothes. This type of description is, however, less discriminant and easier to mimic. We propose using a depth sensor to perform 3D face authentication continuously, since 3D outperforms 2D face recognition in many aspects [Bowyer et al., 2006]. First, pose robustness is better achieved when 3D data is available. Second, the Kinect is able to capture 3D images in a wide range of lighting conditions. Finally, the 3D data allows a better classification of foreground and background objects, which facilitates tasks like object detection and tracking.

2.7.1 Proposed approach

The recognition system described in Sections 2.1-2.6 is employed to the 3D continuous authentication process, as shown in Figure 2.19. The Kinect is used for acquisition, but other depth sensors with equal or better accuracy than the Kinect can also be used with small or no modifications. As shown in Figure 2.4, the accuracy of the Kinect depends on the distance between object and sensor. Due to this problem, we only use faces up to 1500mm away from the acquisition device for recognition purposes.

In the detection stage, we create multiple projection images from different viewpoints to detect rotated faces. We only considered viewpoint changes around x- and y-axes because pitch (see Figure 2.20(a)) and yaw (see Figure 2.20(b)) rotations are the most common pose variations of a regular computer user. Although not considering roll rota-

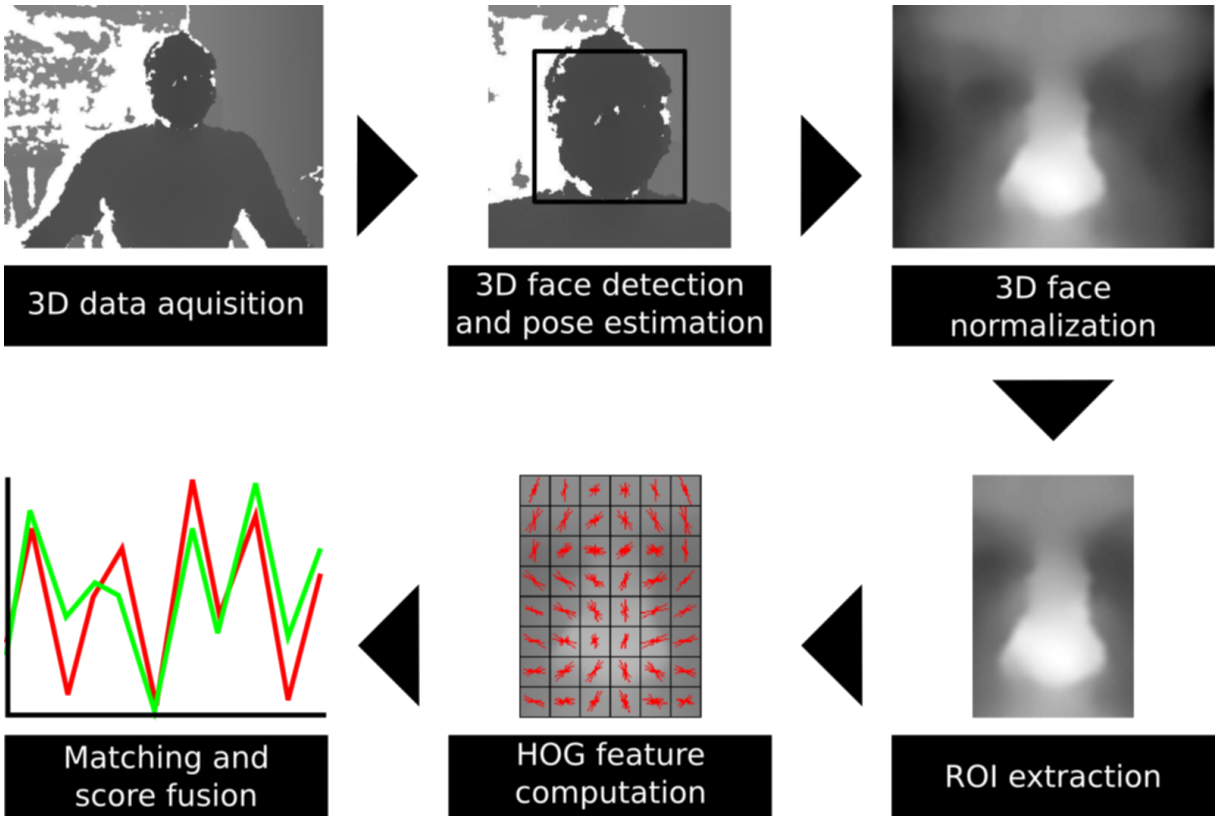
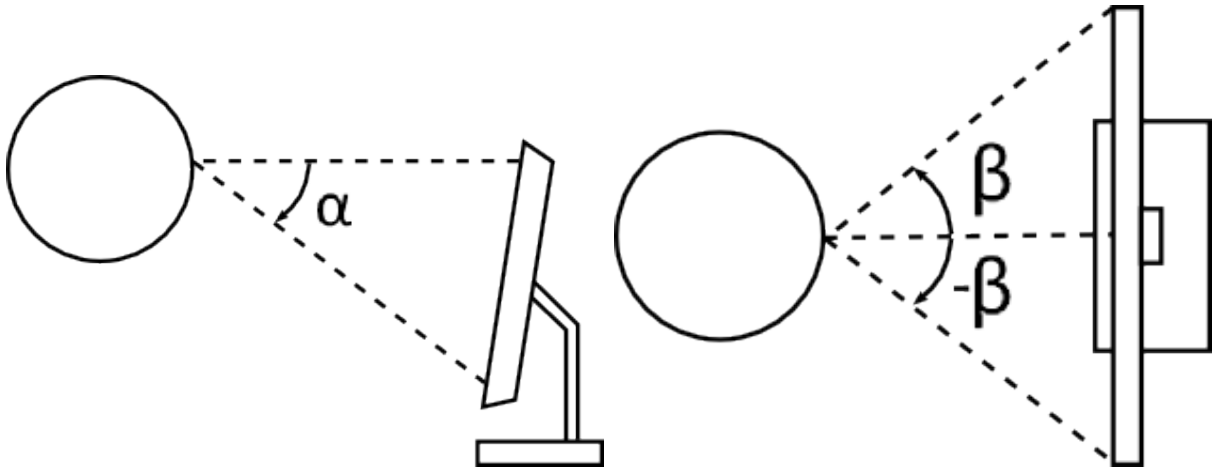


Figure 2.19: Diagram of the process applied to each frame in the continuous authentication system.

tions (*i.e.* changes around the z-axis) speeds up the detection process, but it may result in a few false negatives. In Figure 2.20, the parameters α and β are the maximum values for pitch and yaw rotations, respectively (*i.e.* $\alpha = 40$ and $\beta = 20$ to cover the rotations of the face in front of the display). Projection images were created for all viewpoints within the range specified by α and β at 10 degrees steps, and the detection result is also used to obtain a rough estimation of the head pose. This estimation is given by the rotation values of the viewpoint in which the face was detected.

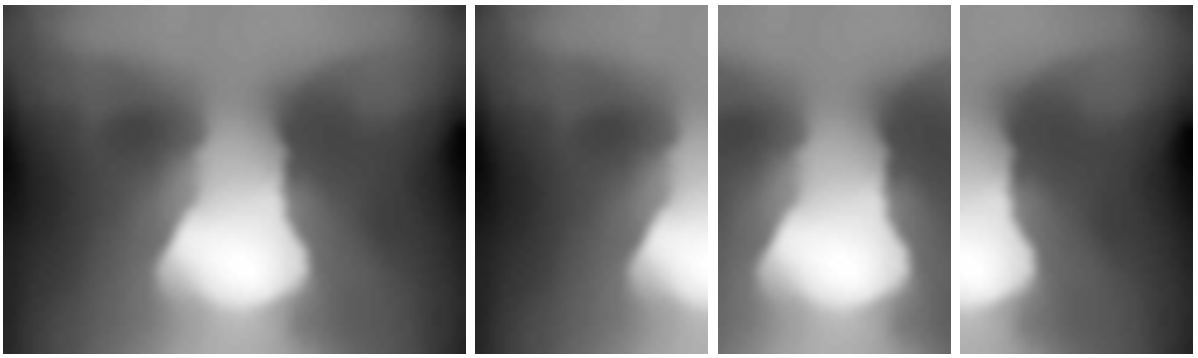
After detection, the face is normalized as in Figure 2.21(a), but it is not possible to use the entire face image every frame because pose variations can substantially affect one side of the face. When this happens, the affected side may present holes and excessive noise due to self-occlusions in the face, as may be observed in Figure 2.13(e). To solve this problem, we divide each image in three different Regions of Interest (ROIs): the left half of the face, the nose region and the right half of the face, respectively shown in Figures 2.21(b), 2.21(c) and 2.21(d). We only use one of these regions for each frame



(a) Side view (b) Top view

Figure 2.20: Common pose variations of a regular computer user: (a) pitch and (b) yaw.

according to its pose, which is obtained in the detection stage. The nose ROI is used for frontal faces (*i.e.* a face is considered frontal if the yaw rotation is smaller than 5 degrees), while we use the left ROI when the user is looking to the right and the right ROI when the user is looking to the left. This way we avoid using too noisy image parts and also use the most invariant facial region when frontal faces are available [Chang et al., 2006].



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

Figure 2.21: (a) Example of resulting face image after normalization, and its different ROI: (b) left region, (c) nose region and (d) right region.

A HOG descriptor is then computed for the chosen ROI, and its matching distance to the user template is used to evaluate the current safety status of the system. More details about enrollment and continuous authentication processes are given below.

User enrollment

At the login moment, when the user identifies itself to get access to the computer, the system is assumed to be safe. So we take N biometric samples at this point to be used as the user template (*i.e.* $N = 3$), as illustrated in Figure 2.18. Since all three ROIs must be acquired at the same time, the access is locked until the user turns his face frontally, which is done because frontal faces do not have any side damaged by self-occlusions. In the end, we have N descriptors for each ROI that form the biometric template.

User continuous authentication

In the continuous authentication, we must be able to determine the safety status of the system at any time. This is done by computing the probability of the system being safe at time t, called P_{safe}, from the history of observations $\mathcal{Z}_t$. Each observation $z_i \in \mathcal{Z}_t$ corresponds to the matching distance between a probe image and the user template at time i. The matching distance corresponds to the Manhattan distance between the chosen ROI at time i and its respective ROI in the template. The fusion of continuous scores is based on the Temporal-First integration proposed by Sim et al. [2007], which keeps track of P_{safe} over time with a weighted sum of $\mathcal{Z}_t$. In this fusion scheme, older observations are "forgotten" to ensure the current user is still the allowed one and the probability of the system being safe can be computed at any time, even when there is no observation. Equation 2.10 is used to compute P_{safe}:

$$P_{safe} = P(safe \mid \mathcal{Z}_t)e^{\frac{-\Delta t \ln 2}{k}} \tag{2.10}$$

where k is the decay rate that defines how fast the system "forgets" older observations (*i.e.* P_{safe} drop to half every k seconds without observations, $k = 15$), and Δt is the elapsed time since the last observation z_t. $P(safe \mid \mathcal{Z}_t)$ is obtained by Equations 2.11-2.13:

$$P(safe \mid \mathcal{Z}_t) = \frac{P'(safe \mid \mathcal{Z}_t)}{P'(safe \mid \mathcal{Z}_t) + P'(\neg safe \mid \mathcal{Z}_t)} \tag{2.11}$$

$$P'(safe \mid \mathcal{Z}_t) = P(z_t \mid genuine) + P'(safe \mid Z_u)e^{\frac{(u-t)\ln 2}{k}} \tag{2.12}$$

$$P'(\neg safe \mid \mathcal{Z}_t) = P(z_t \mid impostor) + P'(\neg safe \mid Z_u)e^{\frac{(u-t)\ln 2}{k}} \tag{2.13}$$

where $P(z_t \mid genuine)$ and $P(z_t \mid impostor)$ are given by Equations 2.8 and 2.9, respectively, and u is the time of the last observation before t, z_u. Since the system is assumed to be safe at the login time, $P'(safe \mid \mathcal{Z}_0) = 1$ and $P'(\neg safe \mid \mathcal{Z}_0) = 0$.

The parameters $\mu_{genuine}$, $\sigma_{genuine}$, $\mu_{impostor}$ and $\sigma_{impostor}$) in Equations 2.8 and 2.9 were obtained for each ROI, and the resulting CDFs are shown in Figure 2.22. An exhaustive search was performed to obtain the set of parameters that minimizes the EER in our experiments. The respective values of $(\mu_{genuine}, \sigma_{genuine}, \mu_{impostor}, \sigma_{impostor})$ for the left ROI, the nose ROI and the right ROI are $(89.0, 14.5, 128.3, 17.8)$, $(82.5, 13.2, 122.4, 16.2)$, and $(88.8, 12.9, 129.2, 17.4)$. Since only one ROI is used per frame, only its respective CDFs are used in Equations 2.12 and 2.13

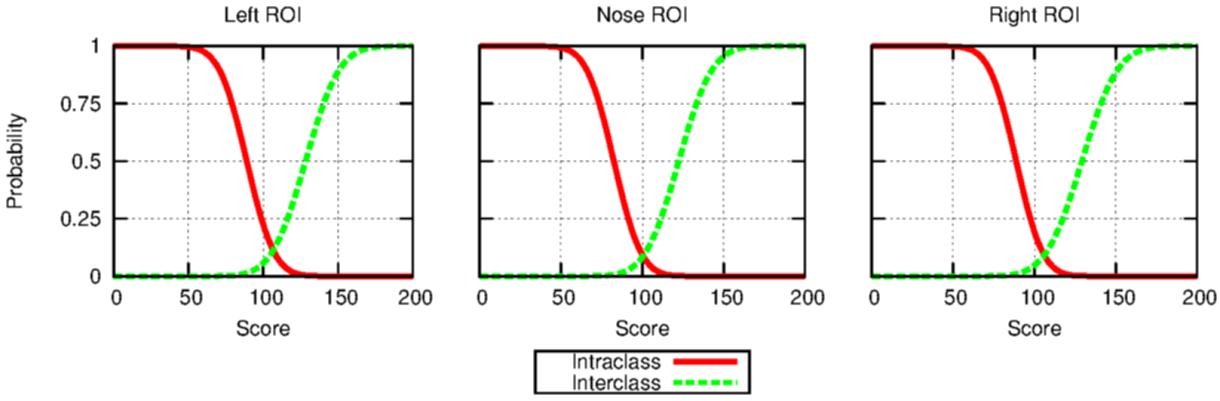


Figure 2.22: CDFs for the three ROIs employed in this work: left ROI, nose ROI and right ROI. Intraclass and interclass curves represent CDFs for genuine matchings and impostor matchings, respectively.

Since our fusion method can be updated for every new observation, we no longer need to keep a history of observations, as it is done by Sim et al. [2007]. Also, our changes in the formulation avoid a continuous decrease in the P_{safe} value in the first k seconds after the login.

2.8 3D Face Reconstruction

One problem of 3D face recognition systems is the lack of compatibility to the current forms of identification that usually have a 2D face image as a biometric record (*e.g.* ID cards, passports and driver licenses). To overcome this problem, 3D reconstruction methods may be used to retrieve the 3D information from 2D images [Blanz and Vetter, 2003, Choi et al., 2010, Jiang et al., 2005, Kemelmacher-Shlizerman and Basri, 2011, Levine and Yu, 2009, Medioni et al., 2009, Wang and Lai, 2011].

Some 3D face reconstruction methods use multiple 2D images to recover the geometry of a face. Choi et al. [2010] used the sparse bundle adjustment algorithm over a set of facial landmarks in five images with specific facial pose to compute a sparse 3D face model. A dense 3D face model is obtained by Medioni et al. [2009] by applying a structure from motion technique to a high resolution video sequence. However, in some cases, there is only a single 2D image available to recover the geometry of a face.

Nevertheless, it is still possible to three-dimensionally reconstruct a face using a single 2D image as input. Jiang et al. [2005] fitted a 3D deformable face model to a set of landmarks in a frontal face image to obtain a 3D face model. Blanz and Vetter [2003] also used a 3D deformable face model, but the fitting process was guided by the texture information and there was no restrictions on the face pose. Finally, Kemelmacher-Shlizerman and Basri [2011] used a shape from shading technique to deform a reference face model and obtain the 3D model of a face. All these face reconstruction methods can be performed fully automatically. However, in case of failures or unexpected scenarios (*e.g.* high illumination variation or painted faces), only landmark-based methods can be manually assisted in a practical way.

For this reason, we have developed a landmark-based face reconstruction method to recover the geometry of a face using a single 2D image as input. It relies on a previously defined set of facial landmarks, which may be automatically located or manually obtained depending on the application. For example, if the reconstruction process is going to be used in conjunction with an access control system, it should be fully automatic since the use is going to be intense. But if the reconstruction process is used to add faces to a

watchlist, the use is very limited, so manual annotations may be used to ensure more accurate results.

2.8.1 Proposed approach

In our 3D face reconstruction method, the Levenberg-Marquardt (LM) iterative minimization technique [Marquardt, 1963] is applied to obtain camera and deformation parameters that fit a sparse 3D deformable face model in a set of facial landmarks located in the input image. These landmarks may be automatically located [Cootes et al., 2001] or manually obtained. After that, the Thin-Plate Splines (TPS) technique [Bookstein, 1989] is used to deform a generic face model in order to obtain a dense 3D model of the input face, as done by Choi et al. [2010] and Park and Jain [2006]. Finally, the texture of the input image is warped into the 3D model. The reconstruction process is summarized in Figure 2.23.

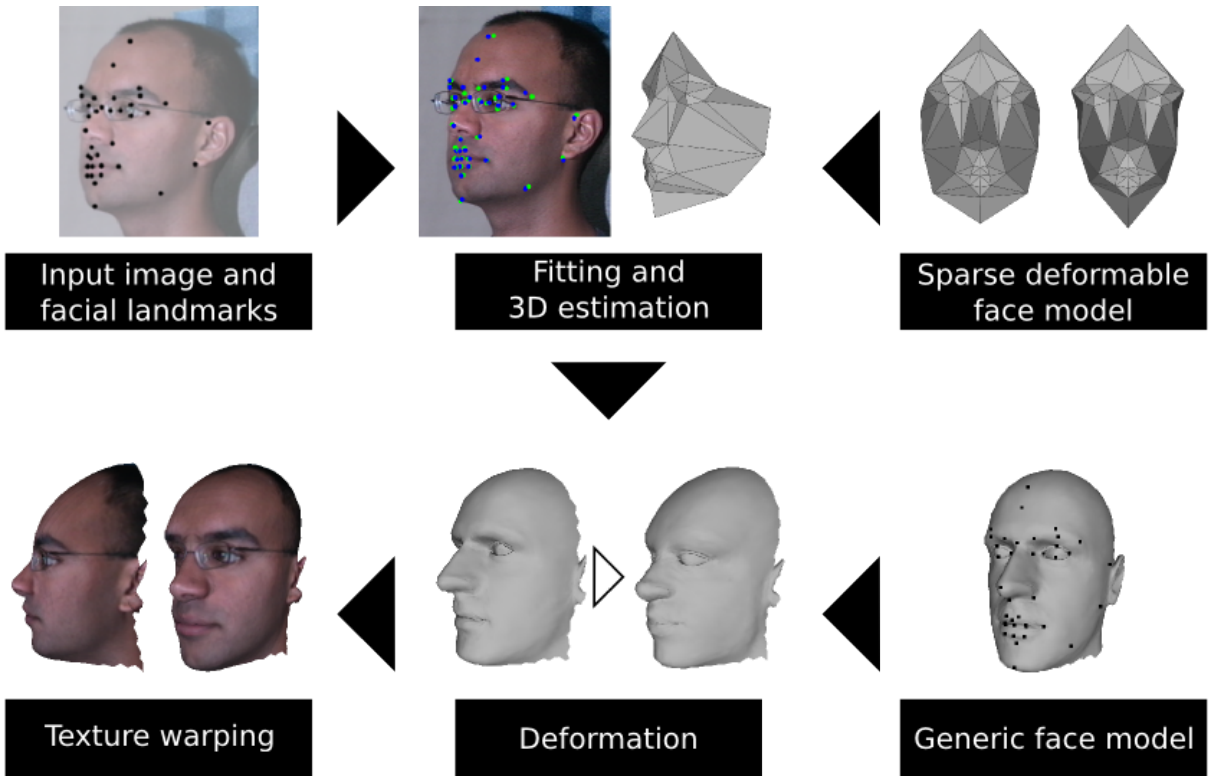
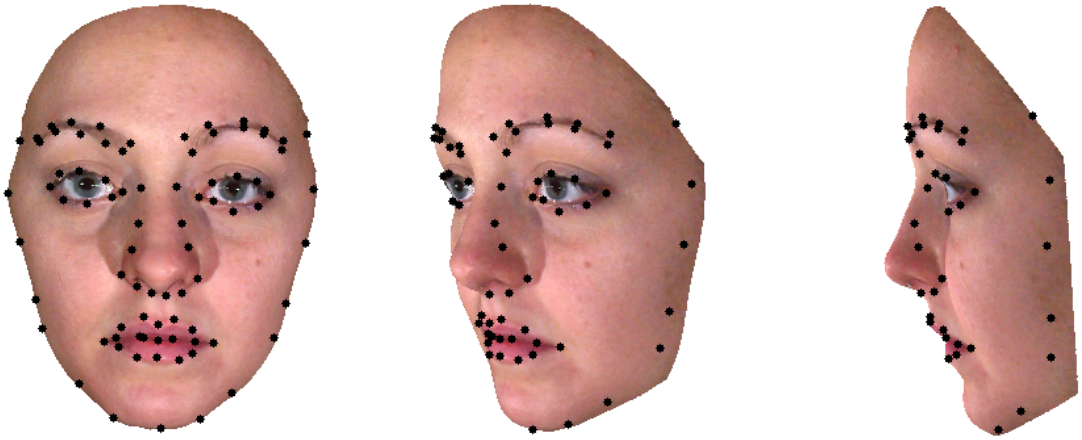


Figure 2.23: Diagram of our 3D face reconstruction method.

Sparse deformable face model creation

To create a sparse deformable face model, we use 50 neutral 3D face images of the Binghamton University 3D Facial Expression (BU-3DFE) database [Yin et al., 2006] (*i.e.* details about this database are given in Section 3.1.2). Each of these images is composed of a textured surface mesh and the ground truth location of 83 facial landmarks. Figure 2.24 shows some renderings of a BU-3DFE image and its facial landmarks.



(a) (b) (c)

Figure 2.24: Renderings of a BU-3DFE subject in (a) frontal, (b) half-frontal and (c) profile poses.

Each training image is then represented as a set of 3D coordinates $Q_i = \{q_1^i, q_2^i, \ldots, q_N^i\}$, where N is the number of landmarks and $q_j^i = \{X_j^i, Y_j^i, Z_j^i\}$, and an average set $\bar{Q}$ is computed after aligning all training images to the same coordinate system. The PCA [Pearson, 1901] is applied to the training images to learn the deformation model, and any set of 3D landmarks Q can be represented by a vector of weights $[w_1, w_2, \ldots, w_K]$, where w_i is the weight of the i-th eigenvector u_i returned by PCA. The original representation of Q is recovered by Equation 2.14:

$$Q = \bar{Q} + \sum_{i=1}^{K} w_i u_i \tag{2.14}$$

We keep 99% of the overall data variance using PCA. Since K is much smaller than the size of Q (*i.e.* $3N$), the vector of weights is much easier to be retrieved during the

fitting process.

Fitting and 3D estimation

The fitting process can be described as follows. Given a set of 2D landmarks $P = \{p_1, p_2, \ldots, p_N\}$ in an input image, with $p_i = \{x_i, y_i\}$, the objective is to find the set $Q = \{q_1, q_2, \ldots, q_N\}$ with 3D coordinates $q_i = \{X_i, Y_i, Z_i\}$ and the transformation T that minimize Equation 2.15:

$$\frac{1}{N}\sum_{i=1}^{N} ||p_i - p'_i|| \tag{2.15}$$

where $p'_i = Tq_i$. The transformation T is given by a camera model with seven parameters: translation in all axes t, rotation in all axes R and focal length f. A 3D landmark q_i is transformed into 2D coordinates on the input image p'_i by applying Equations 2.16-2.18:

$$q'_i = Rq_i + t \tag{2.16}$$

$$x'_i = \frac{W}{2} - \frac{X'_i}{Z'_i}f \tag{2.17}$$

$$y'_i = \frac{H}{2} + \frac{Y'_i}{Z'_i}f \tag{2.18}$$

where W and H are respectively the width and height of the input image.

The LM iterative minimization technique [Marquardt, 1963] is then performed to find the camera parameters and the vector of weights that minimize Equation 2.15 in order to obtain the final set of 3D landmarks through Equation 2.14. The LM method is a combination of the gradient descent and the Gauss-Newton minimization methods, and is more robust than when these methods are applied separately. With the presented method, there is no need to make strict assumptions about the pose Jiang et al. [2005] or use multiple images Choi et al. [2010] to obtain a realistic 3D model.

Enhanced reconstruction through face symmetry

As the rotation of the face increases, its visible area decreases due to the self-occlusion problem. The number of visible landmarks represented in the sparse model also decreases for this reason, as may be seen in Figure 2.24. Figure 2.25 shows the average number of visible landmarks for different rotation angles of the face. As may be seen, the number of visible landmarks decreases with increasing rotation.

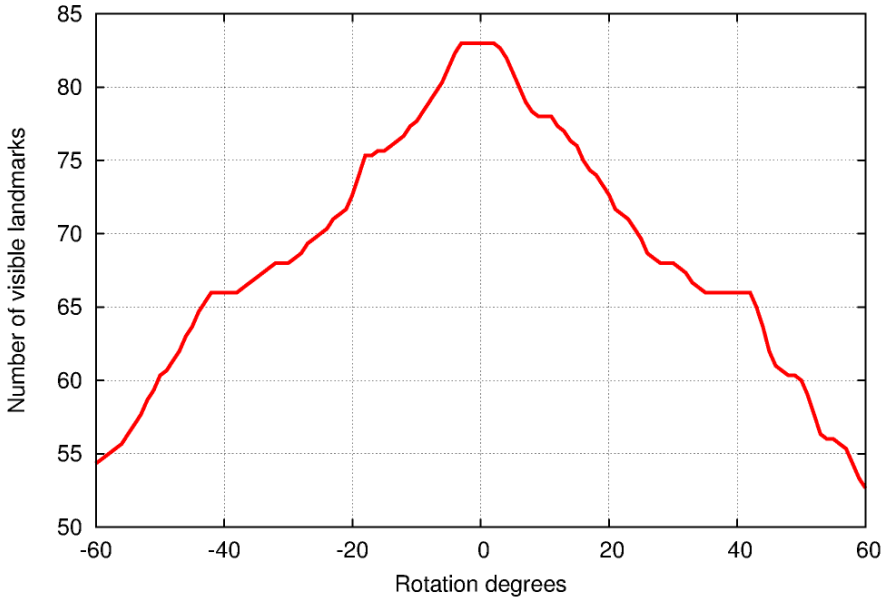


Figure 2.25: Number of visible landmarks in different face poses.

However, most human faces are quite symmetric, allowing us to use the visible side of a face to estimate the occluded side, as previously done in the literature [Shimshoni et al., 1999]. To this end, given two symmetric 3D landmarks q_i and q_j, q_j is mirrored into q_i to create a redundant 3D landmark $q_i^s = \{-X_j, Y_j, Z_j\}$. After that, the minimization equation in the parameter estimation stage is replaced by Equation 2.19:

$$\frac{1}{2N}\sum_{i=1}^{N} ||p_i - p_i'|| + ||p_i - p_i^s|| \tag{2.19}$$

where $p_i^s = Tq_i^s$. Besides providing valid information for occluded landmarks, the use of symmetry also reduces the influence of noise by providing redundant information for symmetric landmarks when both are visible.

Deformation and texture warping

The deformation between landmarks of a generic face model and the reconstructed landmarks of an input face is mapped through the TPS technique [Bookstein, 1989]. After obtaining a deformation map, it is applied to all points of the generic face model to obtain a dense reconstructed face. Then, the camera parameters can be used to project the reconstructed face into the input image to retrieve the texture information. Problems with self-occlusions are handled through the use of symmetry.

CHAPTER 3

RESULTS

3.1 Databases

The following databases were used in our experiments: FRGC database [Phillips et al., 2005]; BU-3DFE database [Yin et al., 2006]; The Bosphorus Database (BOSPHORUS) [Savran et al., 2008]; Texas 3D Face Recognition Database (TEXAS3D) [Gupta et al., 2010]; RGB-D Face Database (RGBDFD) [Hg et al., 2012]; and Berkeley 3-D Object Dataset (B3DO) [Janoch et al., 2011]. A brief description of each one is presented in Sections 3.1.1-3.1.6.

3.1.1 Face Recognition Grand Challenge database

The FRGC database [Phillips et al., 2005] is divided into training (FRGC v1) and testing (FRGC v2) sets. The training set contains 943 images from 275 different subjects and the testing set is composed of 4,007 images from 466 different subjects. There are 184 subjects in both sets and the number of images per subject ranges from 1 to 30. Images are 640×480 and were acquired by a laser scanner. The average number of valid points per image is about 97,000. Faces are frontally posed and present different artifacts that may affect the detection performance: facial expressions (*e.g.* disgust, happy, puffy cheek, sad and surprise), distorted images, noisy points and surface holes. Some examples of these artifacts are shown in Figures 3.1(a)-3.1(e).

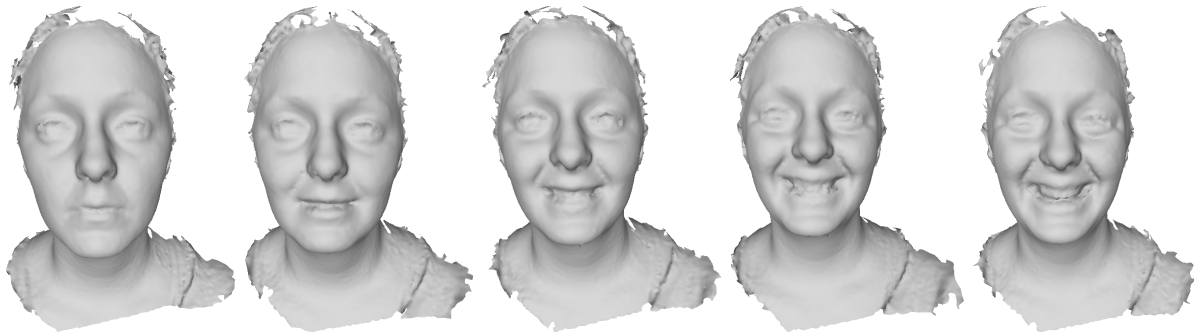
3.1.2 Binghamton University 3D Facial Expression database

The BU-3DFE database [Yin et al., 2006] has 2,500 images from 100 different subjects. The data acquisition was performed using a hybrid sensor based on stereo photogrammetry and structured light. Images have 14,000 valid points on average, and faces are

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

Figure 3.1: FRGC artifacts: (a) facial expressions; (b) spikes; holes caused by (c) limited focal distance or (d) insufficient laser reflectance; and (e) distortions caused by movements at the acquisition time.

not exactly frontal. Each subject has 25 images, one neutral and four for each of six facial expressions (*i.e.* angry, disgust, fear, happy, sad and surprise) at different levels of intensity. Figures 3.2(a)-3.2(e) show the intensity effect on facial expressions.

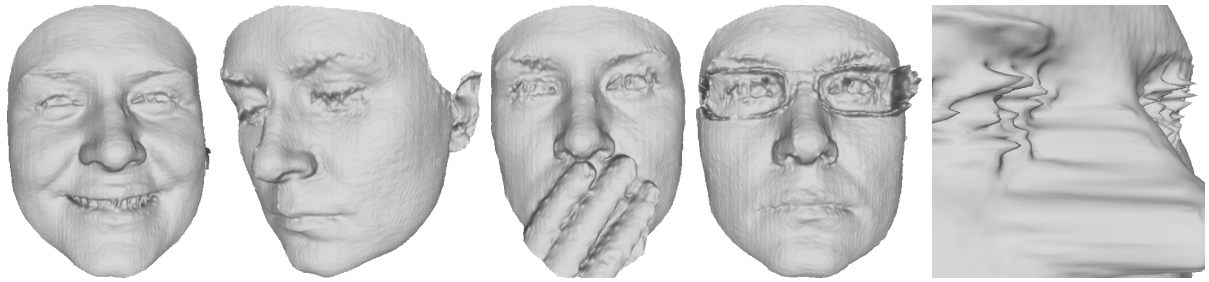


(a) (b) (c) (d) (e)

Figure 3.2: BU-3DFE expression intensities: (a) neutral, (b) mild, (c) moderate, (d) intense and (e) very intense expressions.

3.1.3 The Bosphorus Database

The BOSPHORUS database [Savran et al., 2008] contains 4,666 images from 105 different subjects, and the number of images per subject ranges from 29 to 54. Images have 36,000 valid points on average and were acquired by a structured light-based sensor. Faces present different artifacts, such as pose variation, occlusion, facial expressions and noise, as shown in Figures 3.4(a)-3.4(e).

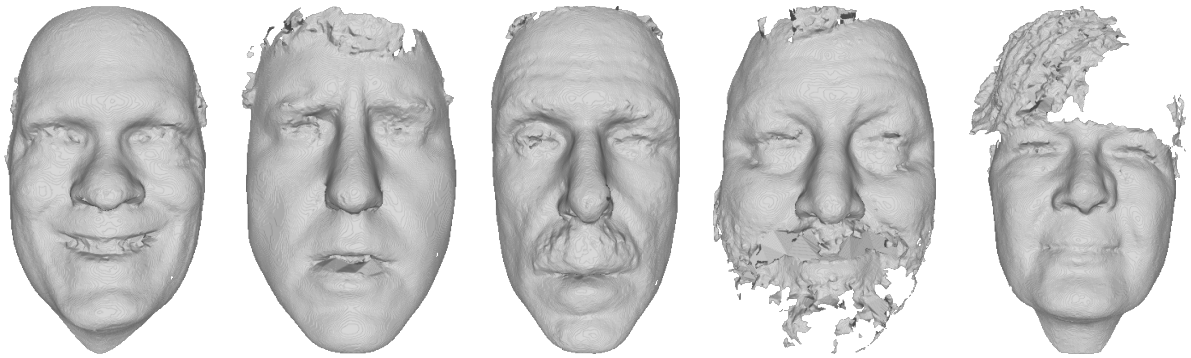


(a) (b) (c) (d) (e)

Figure 3.3: BOSPHORUS artifacts: (a) facial expressions; (b) pose; (c)-(d) different types of occlusion; and (e) noise in eyes and border regions.

3.1.4 Texas 3D Face Recognition Database

The TEXAS3D database [Gupta et al., 2010] is composed of 1,149 images from 116 different subjects acquired by stereo-based sensor. Each image has 242,000 valid points on average and there are four or less images for most of the subjects. Faces are frontal and can present expression variations. Although the TEXAS3D database presents the highest resolution among all tested databases, images are relatively smooth due to the acquisition technology, as shown in Figure 3.4.



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

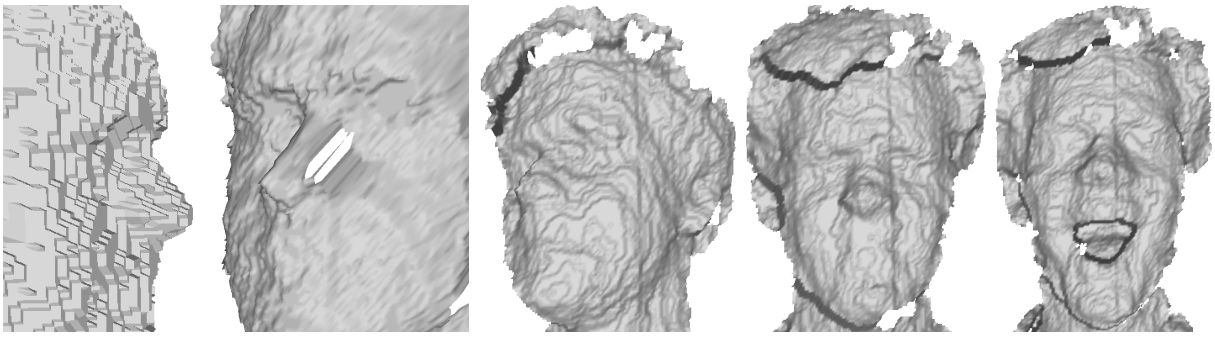
Figure 3.4: TEXAS3D artifacts: (a)-(b) facial expressions; (c)-(d) facial hair; and (e) hair parts.

3.1.5 RGB-D Face Database

The RGBDFACE database [Hg et al., 2012] has a total of 1,581 images from 31 different subjects acquired by a Microsoft Kinect sensor[1]. Images are 640 × 480 and have 82,000

[1]http://www.xbox.com/kinect

valid points on average, although most of these points do not belong to the face region. Each subject has 51 images presenting 13 different poses and four facial expressions (*i.e.* happy, sad, angry and yawn). A "staircase effect" may also be observed in these images due to the low precision of the Kinect. Figures 3.5(a)-3.5(e) show some examples of these artifacts.



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

Figure 3.5: RGBDFACE artifacts: (a) "staircase effect"; (b) holes and noise; (c) pose and (d)-(e) facial expressions.

3.1.6 Berkeley 3-D Object Dataset

The B3DO database [Janoch et al., 2011] provides the raw Kinect data for 645 images from different objects and different scenarios. This database does not contain any faces, and is used to evaluate the performance of the proposed approach on images with complex background in terms of number of false alarms.

3.2 Face detection results

The goal of our experimental evaluation was to analyze the robustness of the proposed face detector against several types of variations, artifacts and acquisition scenarios, and by doing so, demonstrate its advantages over other methods. To accomplish that, six databases were used in our experiments (*i.e.* FRGC, BU-3DFE, BOSPHORUS, TEXAS3D, RGBDFD and B3DO), most of them well-known in the literature and extensively used for 3D face analysis. We also compare our results to an implementation of the Viola and Jones' detector [Viola and Jones, 2004] available at the OpenCV library and to Hg et al.'s im-

plementation [Hg et al., 2012] of the 3D detector proposed by Colombo et al. [2006], since both are state-of-the-art face detectors for 2D and 3D images, respectively.

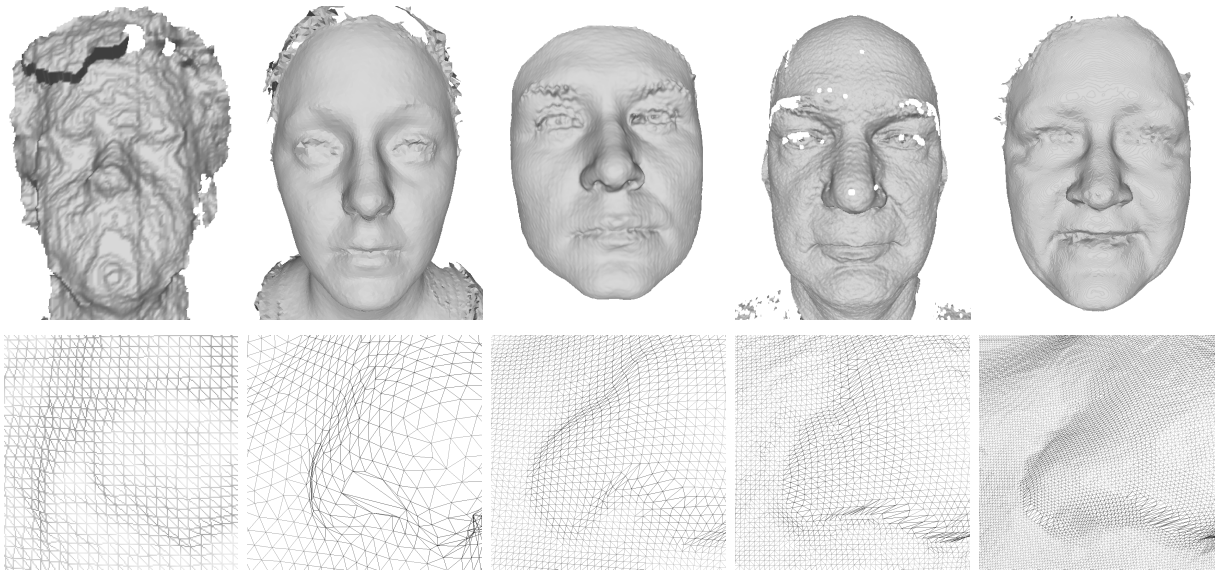
The same cascade classifier obtained from the FRGC training set was employed to the experiments with all databases. One detected face is considered correct if the central part of the face (see the light region in Figure 2.8(a)) is completely inside the detected square. This evaluation was automatically performed for the FRGC, BU-3DFE, BOSPHORUS and TEXAS3D databases based on the ground truth location of the outer eye corners. The RGBDFD database was evaluated through a visual inspection because there is no ground truth data for the images. Results were presented in terms of FRR and False Discovery Rate (FDR), where FRR shows the percentage of face regions that were misclassified as non-face and FDR shows the percentage of false detections among all detections. Our detector is implemented in C without any parallelism, and the experiments were performed on a laptop with a 2.40GHz Intel Core i3-3110M processor.

3.2.1 Database comparison

Table 3.1 summarizes the main aspects of the FRGC, BU-3DFE, BOSPHORUS, TEXAS3D and RGBDFD databases. A visual comparison is also shown in Figure 3.6. As may be seen, these databases were acquired in different scenarios and present different challenges. Obtaining high detection rates and a small number of false detections in all of them is a difficult task, especially when using the same approach (*i.e.* same training and parameters for all databases). TEXAS3D and BOSPHORUS are composed of segmented face images (see Figures 3.6(c) and 3.6(e)), so they are not valuable to evaluate the occurrence of false detections because there are a few non-face regions to be evaluated (*e.g.* face boundaries). However, they are very useful to evaluate the robustness of the detection process against unknown subjects, occlusions, pose, facial expressions and resolution in terms of false rejections. The RGBDFD database is the most challenging one due to the low quality acquisition system, which produces highly noisy, undetailed face images, as may be seen in Figure 3.6(a).

Table 3.1: Classification of the databases used in this work according to the following aspects: pose variations (PV); lighting variations (LV); facial expressions (FE); occlusions (OC); segmented faces (SF); resolution (RS); and noise (NS).

Database	**PV**	**LV**	**FE**	**OC**	**SF**	**RS**	**NS**
FRGC	No	Low	Yes	No	No	High	Low
BU-3DFE	Low	No	Yes	No	No	Medium	Low
TEXAS3D	No	No	Yes	No	Yes	High	Low
BOSPHORUS	High	No	Yes	Yes	Yes	High	Medium
RGBDFD	High	No	Yes	No	No	Low	High



(a) RGBDFACE (b) BU-3DFE (c) BOSPHORUS (d) FRGC (e) TEXAS3D

Figure 3.6: Image examples from all databases used in this work. Facial regions are shown in the top row, and close-up views of the nose corner are shown in the bottom row.

3.2.2 Experimental results

FRGC, TEXAS3D, BU-3DFE and BOSPHORUS evaluation

In this experiment, we employed three different detection scenarios:

1. using a single orthogonal projection image of the original viewpoint (*i.e.* the sensor viewpoint);

2. using five orthogonal projection images to cover small pose variations (*i.e.* the original viewpoint and four slightly different viewpoints with -5 and $+5$ degrees of rotation in X and Y axes);

3. using 45 orthogonal projection images to cover large pose variations (*i.e.* viewpoint rotation from -20 to $+20$ degrees in the X axis and from -40 to $+40$ in the Y axis, with a 10 degrees step).

For each detection scenario, we show in Table 3.2 the obtained results in terms of FRR, FDR and average time in seconds using FRGC, TEXAS3D, BU-3DFE and BOSPHORUS. Face images from TEXAS3D and BOSPHORUS are segmented and false detections can only be caused by inaccurate detections. For this reason, FDR values for these databases are not shown in Table 3.2 since they are not relevant to the experiment (*i.e.* FDR $<$ 0.05% in all cases).

Table 3.2: Detection results (FRR and FDR) and average time (seconds) for the FRGC, BU-3DFE, BOSPHORUS and TEXAS3D databases using 1, 5 or 45 orthogonal projection images. The values presented in bold characters show the best results for each experiment.

# Viewpoints	**1**			**5**			**45**		
FRGC	**FRR**	**FDR**	**Time**	**FRR**	**FDR**	**Time**	**FRR**	**FDR**	**Time**
Neutral (2288)	2.0%	0.2%	0.009	**0.3%**	**0.0%**	**0.026**	0.2%	0.2%	0.198
Expression (1719)	2.6%	0.1%	0.009	**0.3%**	**0.0%**	**0.026**	0.3%	0.1%	0.197
Total (4007)	2.2%	0.2%	0.009	**0.3%**	**0.0%**	**0.026**	0.2%	0.1%	0.197
TEXAS3D	**FRR**	**FDR**	**Time**	**FRR**	**FDR**	**Time**	**FRR**	**FDR**	**Time**
Neutral (813)	1.0%	***	0.018	**0.0%**	***	**0.058**	0.0%	***	0.456
Expression (336)	1.5%	***	0.018	**0.0%**	***	**0.058**	0.0%	***	0.460
Total (1149)	1.1%	***	0.018	**0.0%**	***	**0.058**	0.0%	***	0.457
BU-3DFE	**FRR**	**FDR**	**Time**	**FRR**	**FDR**	**Time**	**FRR**	**FDR**	**Time**
Neutral (100)	56.0%	2.2%	0.002	21.0%	3.7%	0.007	**4.0%**	**4.0%**	**0.048**
Expression (2400)	58.7%	4.3%	0.002	20.5%	2.9%	0.007	**0.8%**	**1.3%**	**0.046**
Total (2500)	58.6%	4.3%	0.002	20.5%	3.0%	0.007	**0.9%**	**1.4%**	**0.047**
BOSPHORUS	**FRR**	**FDR**	**Time**	**FRR**	**FDR**	**Time**	**FRR**	**FDR**	**Time**
Neutral (299)	1.0%	***	0.003	**0.0%**	***	**0.009**	0.0%	***	0.075
Expression (2621)	3.9%	***	0.003	**0.3%**	***	**0.009**	0.2%	***	0.076
Pose (806)	54.6%	***	0.003	35.0%	***	0.009	**1.6%**	***	**0.076**
Occlusion (294)	25.5%	***	0.003	**17.3%**	***	**0.010**	16.7%	***	0.077
Total (4020)	15.4%	***	0.003	8.5%	***	0.009	**1.7%**	***	**0.076**

With this experiment, we show the advantages and the cost of using multiple orthogonal projections. As may be seen in Table 3.2, using only the original viewpoint for detection does not give the best performance in any database. When we add slightly different viewpoints, we get the best cost-benefit in terms of FRR, FDR and time for most testing sets. The exceptions are the testing sets that are affected by large pose variations, in which more viewpoints are required to obtain a low FRR value. The drawback of using

a large number of viewpoints is that the detection time grows linearly with the number of viewpoints.

The results for all databases in Table 3.2 show the robustness of the proposed approach against facial expressions (*i.e.* 60% of the tested images present facial expressions) and unknown subjects (*i.e.* about 600 tested subjects are not in the training set). BU-3DFE and BOSPHORUS also show the robustness against large pose variation (*i.e.* almost 30% of the tested images present pose variation). The worst performance was obtained for occluded face images from BOSPHORUS. Since occlusions were not considered in our training stage, we expected a low detection rate. However, we still obtained a detection rate above 80%. To achieve robustness against occlusions, further processing must be done [Alyuz et al., 2012, Colombo et al., 2009].

In the end, faces were correctly detected in 99.8% of the frontal non-occluded images using only 5 slightly different viewpoints, and in 98.9% of the images presenting large pose variations by using 45 viewpoints. Overall, faces were correctly detected by the proposed approach in more than 99% of the tested images, and the average time shows that it can be performed multiple times per second.

FRGC comparison

The FRGC database was also used to compare our detection approach against the state-of-the-art face detection approach based on a multiscale search proposed by Viola and Jones [2004], which is available in OpenCV (*i.e.* the default parameters were used). This detector has been employed or suggested by 2D and 3D face recognition works in the literature [Böhme et al., 2009, Fischer et al., 2010, Lu and Jain, 2005, Mian et al., 2007, Sim et al., 2007, Zhao et al., 2003], and presents one of the best cost-benefit regarding detection accuracy, false detections and computation time. The FRGC database is used because it provides both color and depth images, so we used the OpenCV cascade classifier for the color images, and our own cascade classifier for depth images.

Table 3.3 summarizes the results of this comparison. For color images, we obtained 0.1% FRR, 3.9% FDR and an average time of 0.135 seconds. The OpenCV detector using

color images obtained a slightly better detection rate, but the number of false detections and the computation time are much worse than those obtained using the proposed approach. The large difference in computation time is due to two reasons. The first one is the optimization of the scanning stage. While Viola and Jones' detector has to look for faces with different sizes, the size of the faces is already known in the proposed approach, as said in Section 2.2.2. Second, our cascade classifier is considerably smaller than OpenCV's one for color images, as presented in Section 2.2.2.

Table 3.3: Detection results (FRR and FDR) and average time (seconds) for the FRGC database using the proposed approach with 5 viewpoints and Viola and Jones' approach for color and depth images.

Method	**FRR**	**FDR**	**Time**
Proposed	0.3%	**0.0%**	**0.026**
OpenCV - color	**0.1%**	3.9%	0.135
OpenCV - depth	0.3%	53.6%	0.060

For depth images, FRR, FDR and average time were 0.3%, 53.6% and 0.060 seconds, respectively. Although the difference in time is reduced when using our depth classifier and Viola and Jones' detector, the amount of false detections is definitely unacceptable (*i.e.* about one of every two detections is a false detection). These results show that our classifier is far less robust than OpenCV's color classifier against non-face patterns, but it also show the potential of knowing the face size by using orthogonal projection images to eliminate false detections, as presented in Section 2.2.2. We observed recurring false detections on the nose tip and chin regions when using Viola and Jones' detector on depth images. Not surprisingly, these regions are not even tested by the proposed approach because they are too small to be considered a face. Thus, even with a simple classifier we are able to achieve both high detection rates and low FDR.

RGBDFD evaluation and comparison

In this experiment, we used the RGBDFD database to show the performance of the proposed approach for low quality images. Table 3.4 shows the obtained results for different

subsets of the database. Subset 07 contains frontal neutral face images. Subsets 01, 02, 05, 06, 08, 09, 12 and 13 have images with pose rotation in a single axis. Images from subsets 03, 04, 10 and 11 present cross rotations and images from subsets 14–17 present facial expression variations. To handle the pose variations presented in this database, we used 53 orthogonal projection images of viewpoints in the range of [-30,30] degrees in the X axis and [-50,50] degrees in the Y axis. As may be observed, faces were detected by the proposed approach in 95.38% of the images, with a 1.31% FDR.

Table 3.4: Comparison between three detectors using the RGBDFD database. The values presented in bold characters show the best results for each subset.

Subset	**Hg et al. [2012]**		**Viola&Jones (color)**		**Proposed**	
	FRR	**FDR**	**FRR**	**FDR**	**FRR**	**FDR**
01	32.81%	**0.00%**	11.83%	7.53%	**0.00%**	**0.00%**
02	75.48%	7.69%	6.45%	9.68%	**1.08%**	**6.12%**
03	18.18%	2.70%	**0.00%**	5.38%	**0.00%**	**1.06%**
04	34.02%	9.09%	2.15%	9.68%	**0.00%**	**0.00%**
05	100.0%	**0.00%**	43.01%	9.68%	**20.43%**	**0.00%**
06	75.54%	42.11%	**3.23%**	12.90%	**3.23%**	**1.10%**
07	6.38%	2.22%	**0.00%**	7.53%	**0.00%**	**0.00%**
08	28.89%	**0.00%**	3.23%	5.38%	**0.00%**	**0.00%**
09	100.0%	**0.00%**	7.53%	10.75%	**6.45%**	**0.00%**
10	89.52%	42.86%	12.90%	9.68%	**11.83%**	**0.00%**
11	8.89%	4.65%	**3.23%**	7.53%	**3.23%**	**1.10%**
12	75.27%	**0.00%**	**3.23%**	8.60%	4.30%	**0.00%**
13	71.74%	**0.00%**	**9.68%**	13.98%	25.81%	2.81%
14	13.33%	2.50%	**0.00%**	8.60%	3.23%	**0.00%**
15	4.26%	2.17%	**0.00%**	4.30%	1.08%	**1.08%**
16	23.40%	10.00%	4.30%	8.60%	**1.08%**	**5.15%**
17	6.67%	10.64%	2.15%	15.05%	**0.00%**	**0.00%**
Total	44.74%	6.54%	6.64%	8.89%	**4.62%**	**1.31%**
Time	N.A.		0.126		0.119	

We also compared our RGBDFD results to Viola and Jones' detector using color images and to an implementation of the 3D face detector [Hg et al., 2012] based on Colombo et al. [2006]. The results for these two detectors are also shown in Table 3.4, and, overall, our detector considerably outperforms their results. As may be seen, both FRR and FDR are much higher for Hg et al.'s detector, which misses the face about ten times more often and has about five times more false detections. As in our previous comparison using FRGC,

Viola and Jones' detector and the proposed approach have comparable detection results, but the amount of false detections is considerably reduced by our approach. Our detection rates for frontal faces (subsets 07, 14-17) and frontal neutral faces (subset 07) are 98.9% and 100%, respectively. Viola and Jones's detector performed similarly, achieving 98.7% and 100%, while Hg et al.'s detector only achieves 89.2% and 93.6%. Under pose (subsets 01-06, 08-13), we correctly detect 93.6% of the faces, Viola and Jones's is slightly worse, with a 91.2% detection rate, and Hg et al.'s detector is much less effective, with a 40.8% detection rate.

Although the proposed approach and Viola and Jones's detector have similar detection rates, our approach gets not only the face location, but the face pose as well. To do this, we just need to save the viewpoint in which the face was detected. This information can be employed in further processing, such as face normalization for recognition [Pamplona Segundo et al., 2013a].

B3DO evaluation and comparison

The B3DO database was used to evaluate the occurrence of false detections on images with complex background. Since both B3DO and RGBDFD were acquired by a Kinect sensor, we used the same configuration of the previous experiment in this one. For comparison, we performed this experiment using the proposed detector and using Viola and Jones' detector for color and depth images. The obtained results are presented in Table 3.5. In total, our detector got 30 false detections in 25 images, while Viola and Jones' detector got 72 false detections in 65 images when using color (OpenCV classifier) and 93 false detections in 79 images when using depth (our classifier). As may be seen, even when images with complex background are considered, the proposed approach is very effective in reducing the amount of false detections.

3.2.3 Discussion

We presented a 3D face detector and demonstrated its efficacy in six different databases. We improved the original detection stage proposed by Viola and Jones [2004] using scale-

Table 3.5: Number of false detections and number of images presenting false detections for the B3DO database using the proposed approach with 53 viewpoints and Viola and Jones' approach for color and depth images.

Method	**# false detections**	**# images**
Proposed	30	25
OpenCV - color	72	65
OpenCV - depth	93	79

invariant orthogonal projection images, eliminating the need for multiple scans in a same image due to the perspective distortion. We also detect faces across large pose variation with a frontal face detector by using multiple orthogonal projection images from different viewpoints of the same scene.

A frontal cascade classifier was trained using a subset of the FRGC database, and we evaluated the performance of the proposed approach for FRGC, BU-3DFE, BOSPHORUS, TEXAS3D, RGBDFD and B3DO databases. More than 13,000 face images and 630 unknown subjects were evaluated, many of them presenting artifacts not present in the training set (*e.g.* some facial expressions, noise, pose, occlusions), and the proposed approach was able to detect 99% of the faces, with less than 1% FDR (considering only 8,000 non-segmented images). The RGBDFD database was the most challenging one due to presence of multiple artifacts, mainly noise, in a same image. In this database the face was correctly detected in more than 95% of the images, reaching 100% for frontal neutral face images.

The proposed detector presents some advantages over similar works in the literature:

1. probe images are no longer scanned in multiple scales, so the computation time is significantly reduced;

2. depth images are more invariant than color images and only regions with a prespecified size are tested, what reduces the number of non-face candidates and make the detector much more reliable;

3. robustness against pose variation can be achieved by increasing the number of orthogonal projection images, and our detector returns not only the face location but

the face pose as well.

We compared our detector against two state-of-the-art works to verify these advantages. First, we used the FRGC database to compare our detector with the well-know multiscale face detector proposed by Viola and Jones [2004], which was performed on color and depth images. We obtained an equivalent detection rate and substantially reduced the number of false detections and the computation time. In this comparison we show that our classifier is very simple, but is discriminant enough when the size of the object is known, which is possible through our single scan method based on orthogonal projection images. Then, we used the RGBDFD database to compare our detector with a state-of-the-art 3D face detector [Hg et al., 2012] and also with Viola and Jones' detector for color images. Our detector clearly outperformed Hg et al.'s detector, and, as in our previous experiment, reduced the amount of false detections in comparison to Viola and Jones' detector.

To show the robustness to images with complex background, we compared the number of false detections obtained by our detector and Viola and Jones' detector on images of the B3DO database, which contains images from different scenes and objects but no faces. Again, false detections were much less frequent when our detector was performed.

The computation time of the proposed detector allows its use in real-time applications, and we believe it can be easily extended to detect other objects whose size presents a small intraclass variation. The idea of using multiple orthogonal projection images is highly parallelizable and can be easily executed in graphics processing units, as well as the detection process [Ghorayeb et al., 2006]. Figure 3.7 illustrates how parallelism could be exploited in this work. We also believe that orthogonal projection images would be beneficial to other works that extend Viola and Jones' approach [Jain and Learned-Miller, 2011] and to other detection approaches [Yang et al., 2002], since they would also become scale-invariant and the number of non-face candidates would be considerably reduced.

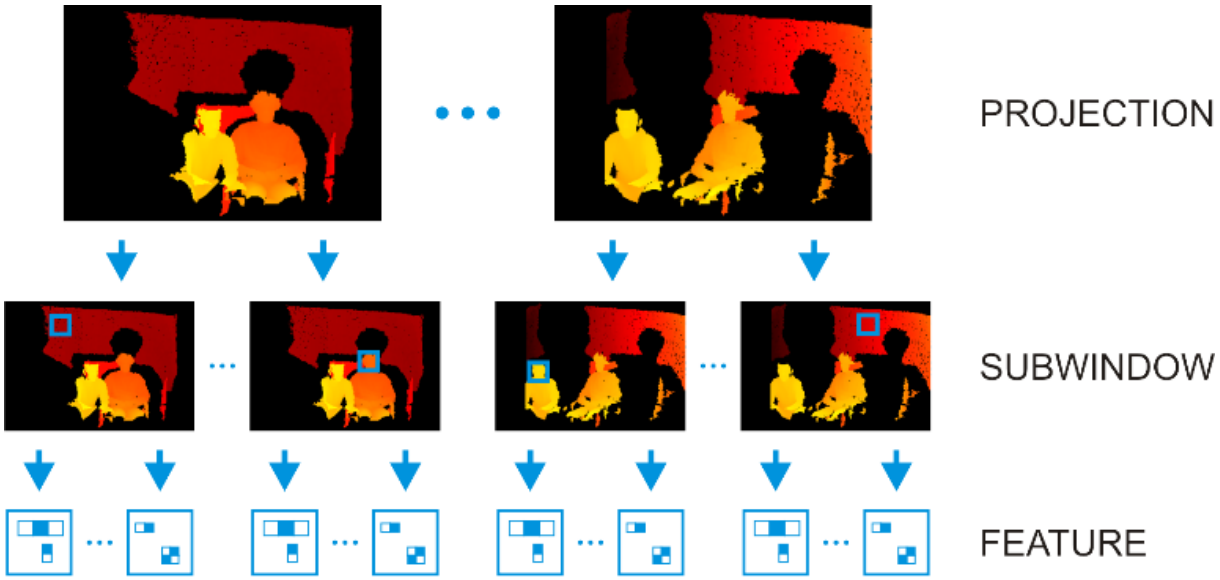


Figure 3.7: In our detector, parallelism could be used in the projection creation, detection and feature computation levels.

3.3 Continuous authentication results

For our experiments, we use four 40 minutes long videos from different subjects acquired by a Kinect sensor. In these videos, the user appears in the scene, logs in the system, uses the computer for approximately 40 minutes and then leaves the scene. The videos were cut so that the first frame shows the user entering the scene and the last picture shows the user leaving the scene. No restrictions were imposed on how the user should use the computer and how the user should behave in front of the computer, but users were not allowed to leave the computer before 40 minutes have passed. Each video sequence has more than 70,000 frames and contains faces with different artifacts that may affect the authentication performance: facial expressions, occlusions, pose and noise. Some examples of these artifacts are shown in Figure 3.8.

Although we use only four subjects in our experiments, the fact that almost 70,000 uncontrolled images were captured for each subject gives us a substantial intraclass and interclass variability.

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

Figure 3.8: Examples of artifacts present in Kinect videos: (a)-(b) facial expressions, (c) occlusion and (d) pose.

3.3.1 Experimental results

Each video was used as input for the proposed continuous authentication system, and the results are shown as solid lines in Figure 3.9. About 2 hours and 40 minutes of genuine access were analyzed, and the system was able to keep the users with high P_{safe} values (*i.e.* above 0.8 in 95% of the frames). After that, we concatenated each video to the end of the remaining videos to simulate impostor accesses and make sure the proposed system is able to detect a user change right after the genuine user leaves the scene. The 12 simulations were then performed (*i.e.* three for each video), and the results are also shown in Figure 3.9 as dashed lines. A total of 8 hours of impostor accesses were considered, and, as may be observed, the P_{safe} value for the authorized user is constantly higher than the P_{safe} value for intruders. This result is corroborated by the Receiver Operating Characteristic (ROC) curve of the P_{safe} values shown in Figure 3.10, in which a 0.8% EER is achieved.

Finally, we present an intuitive way to analyze the potential of the system to detect intruders. We consider the initial frame of each video that was concatenated to another video as the beginning of the impostor access. Then, for a given threshold value, we can see how long the system takes to identify the threat (*i.e.* how many seconds P_{safe} takes to go below the threshold) as presented in Figure 3.11. The solid line was obtained using the EER threshold, which is equal to 0.715. In this experiment, 75% of the impostor accesses are detected in the first second. However, in one case the system takes 19 seconds to detect the intruder. This time can be reduced by increasing the threshold, at the cost

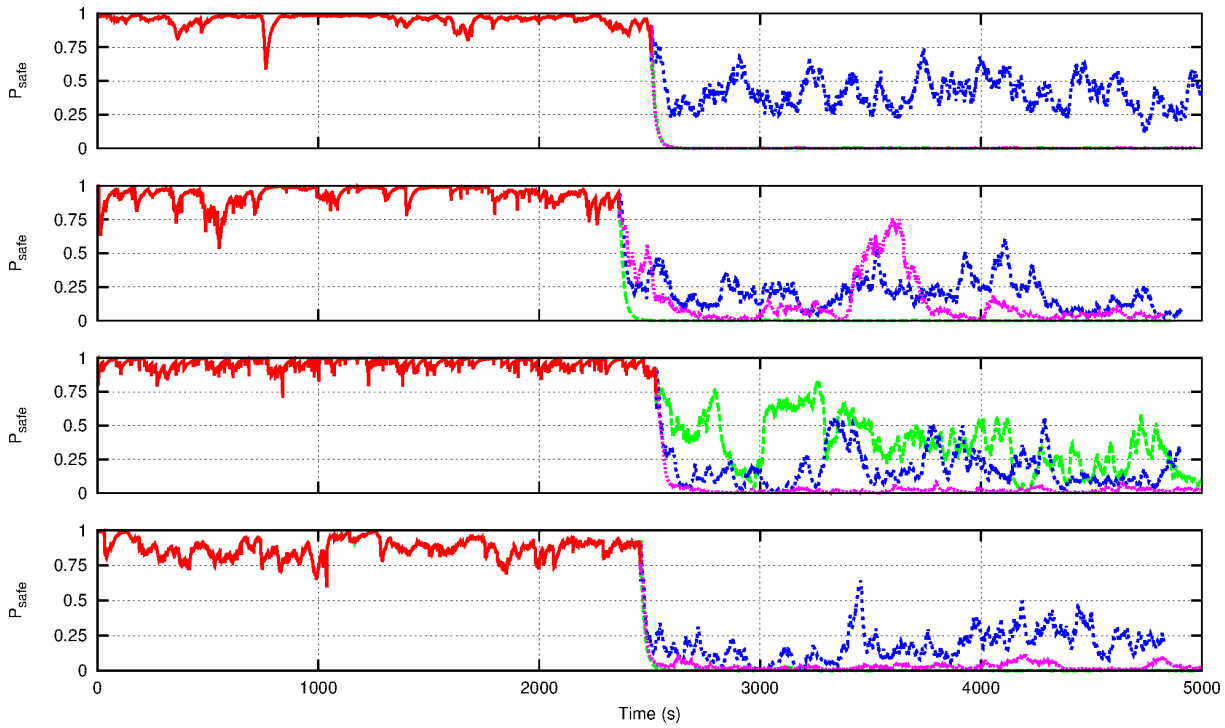


Figure 3.9: Each plot presents the results for the proposed continuous authentication system for a different subject. The solid line represents the authorized user accessing the computer in the initial 40 minutes, and the dashed lines represent the attacks by other subjects starting around 2500s time interval.

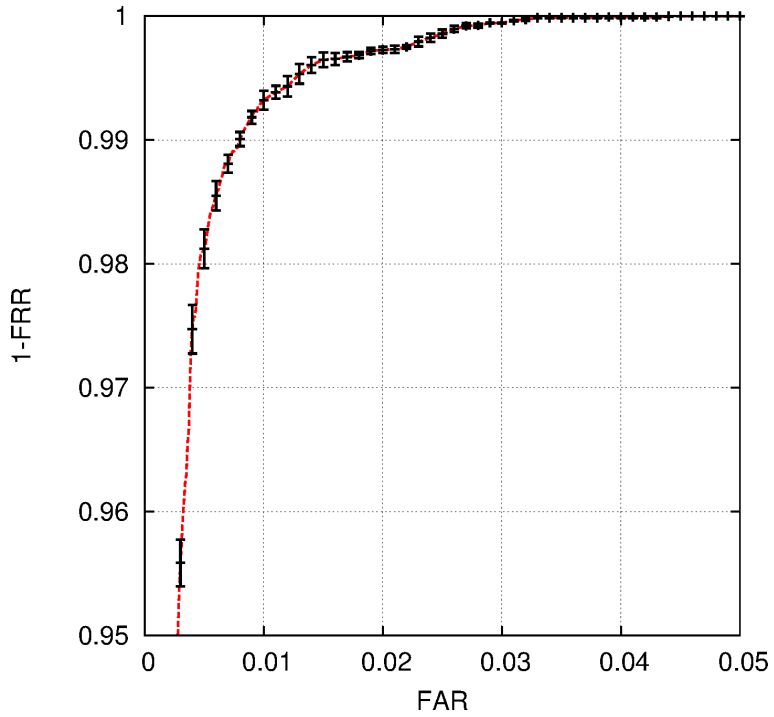


Figure 3.10: ROC curve of the P_{safe} values obtained by our continuous authentication system (see Figure 3.9).

of increasing the FRR. Figure 3.11 shows in dashed lines an example of the results for a higher threshold (*i.e.* 0.758). Although 91.7% of the attacks are detected in the first second and the worst case is reduced to 8 seconds, the FRR grows from 0.8% to 2%.

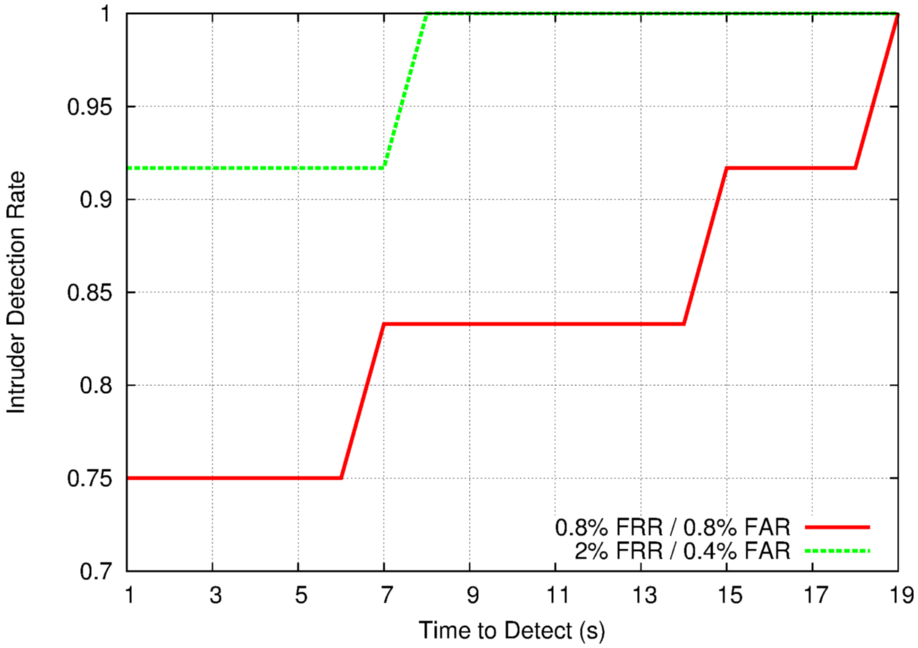


Figure 3.11: Intruder detection rate versus time to detect an intruder: as the time to detect increases, so does the intruder detection rate.

Our experiments were performed at a frame rate of 1 fps in an Intel Core i3 processor, and the remaining frames were discarded by the system. The system is able to process much more fps, but it was not necessary in this specific problem. No parallelism was employed to achieve real-time performance.

3.3.2 Discussion

At the best of our knowledge, this is the first continuous authentication system that uses 3D face images to monitor and ensure that the accessing user is the allowed one. The acquisition was performed by a Kinect sensor, but the system can be used with other depth cameras. The proposed approach automatically detects, normalizes, describes and matches depth images in real-time. Although depth images are invariant to pose, such variations may cause holes and noise due to facial self-occlusions. To solve this problem, in this work we match different regions of the face depending on which facial parts are

clearly visible. In the fusion stage, we present an improved version of the Temporal-First integration approach [Sim et al., 2007] that does not require to keep a history of observations and better controls P_{safe} in the initial part of the continuous authentication process.

More than 2 hours and 40 minutes of genuine accesses and over 8 hours of impostors trying to get access to the system were evaluated in our experiments. The proposed approach obtained a 0.8% EER and was able to detect most of the intruders within a one-second window. We also present a more intuitive way to evaluate the security of the system (see Figure 3.11) by plotting the intruder detection rate along time for different FRR/FAR values.

3.4 3D face reconstruction results

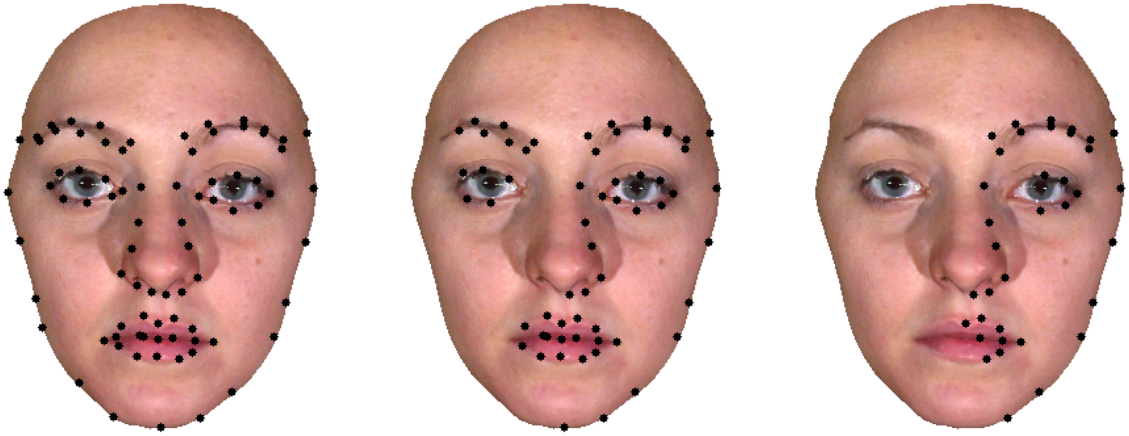
Our reconstruction experiments were designed to evaluate the accuracy of the presented reconstruction method across large pose variations. To this end, we used the neutral images of the BU-3DFE database [Yin et al., 2006] that were not used to create the sparse deformable face model in Section 2.8.1 as a testing set, totaling 50 images.

We have used the test images to create synthetic views with different poses, as illustrated in Fig. 2.24. For each testing image, 121 renderings were created with yaw rotation ranging from -60 to $+60$ degrees to simulate large pose changes. Moreover, each rendering was reconstructed multiple times after adding a random noise with average magnitude ranging from 0 to 5 pixels to the facial landmarks' location in order to show the robustness of the method against noisy data. In our experiments, the reconstruction error is the average Euclidean distance between the ground truth and the reconstructed set of 3D landmarks in millimeters.

3.4.1 Experimental results

In our first reconstruction experiment, only frontal synthetic images were used. First, these images were reconstructed with the original number of landmarks. Then, we sim-

ulated self-occlusion by removing visible landmarks that would be occluded if the face was rotated, as illustrated in Figure 3.12. Figure 2.25 shows the average number of landmarks for different rotation angles. With this experiment, we were able to evaluate the robustness of the reconstruction method against missing information.



(a) (b) (c)

Figure 3.12: Self-occlusion simulation in a frontal face image: (a) original image, and occlusion from (b) half-frontal and (c) profile images.

The obtained results are shown in Figure 3.13, where the curve with label "3D" presents the average reconstruction error of the ground truth, and other curves with label "2D / X px" present the average reconstruction error of frontal synthetic images with self-occlusion simulation, where X is the average noise magnitude in pixels. As may be seen, our sparse 3D deformable model is not able to exactly represent the images of the testing set since there are no subjects in both training and testing sets. For this reason, we obtained an average reconstruction error of about 2 mm for the ground truth. Also, in all cases, the average reconstruction error increases with increasing self-occlusion and noise, as expected.

All synthetic images with different poses were used in our second experiment in order to evaluate the influence of pose variation on the reconstruction results. In the obtained results, shown in Figure 3.14, the reconstruction error remains approximately constant in a wide range of rotation (*i.e.* from -45 to $+45$ degrees), which suggests that the rotation adds information to the image. Also, a slight advantage for half-frontal images can be

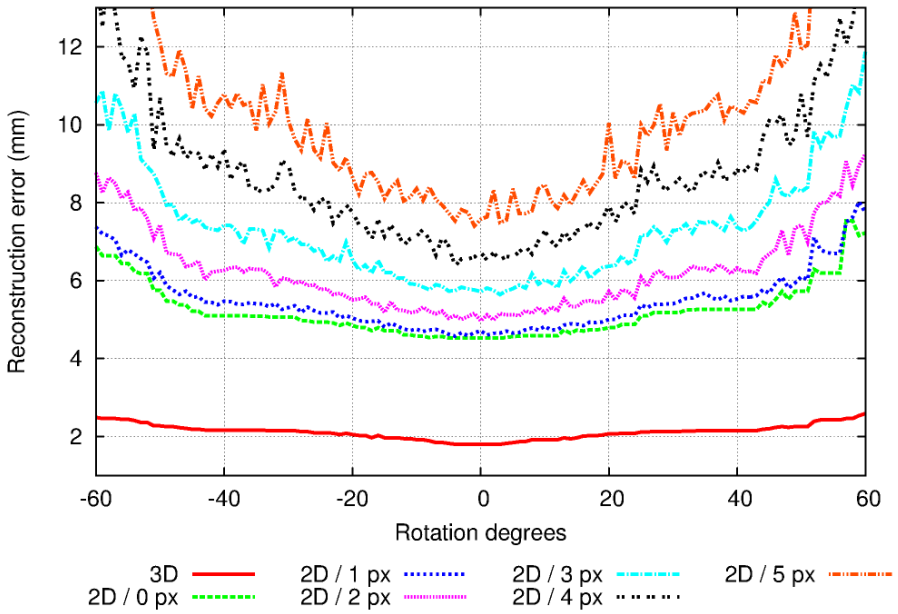


Figure 3.13: Average reconstruction error of frontal synthetic images with self-occlusion simulation.

observed for low noise experiments. However, the results are still being affected by noise and large pose variations.

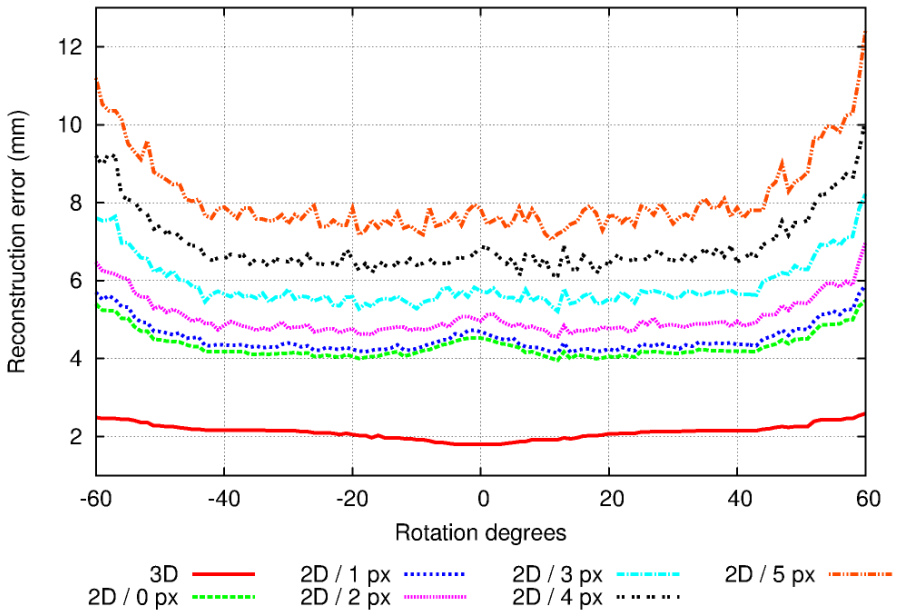
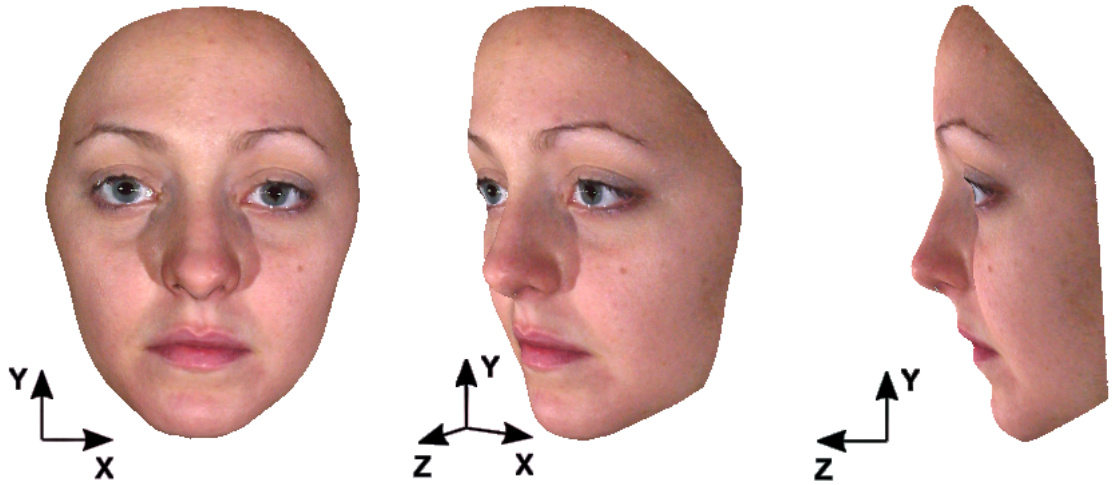


Figure 3.14: Average reconstruction error of synthetic images with pose variation.

The results in Figure 3.14 clearly outperformed the results shown in Figure 3.13, showing that half-frontal faces are better for 3D reconstruction than frontal faces. Half-frontal face images have not been applied to face recognition as much as frontal [Turk

and Pentland, 1991, Zhao et al., 2003] and profile face images [Bhanu and Zhou, 2004, Kakadiaris et al., 2008] due to the difficulty of standardizing acquisition and appearance of such images without any 3D information. However, we have confirmed an observation made by neuropsychology works [Hole and Bourne, 2010], which concluded that half-frontal images allow perceiving all three axes in a single image, as shown in Figure 3.15, which is an interesting property for 3D face reconstruction.



(a) (b) (c)

Figure 3.15: Illustration of visible axes in face images with different pose: (a) frontal, (b) half-frontal, and (c) profile images.

In our final experiment, we repeated the previous experiment using the symmetry of the face as an additional information, and the obtained results are shown in Figure 3.16. We have achieved high reconstruction accuracy and robustness against large pose variations and noise. The superiority of half-frontal images is even more evident in this experiment, in which the reconstruction results are much closer to the ones obtained when using the ground truth information.

Visual comparison

Figure 3.17 shows a comparison between our reconstruction method and the method proposed by Choi et al. [2010] that uses a single or multiple images. If the method uses a single image, only the half-frontal face in Figure 3.17(a) is considered, and the half-frontal

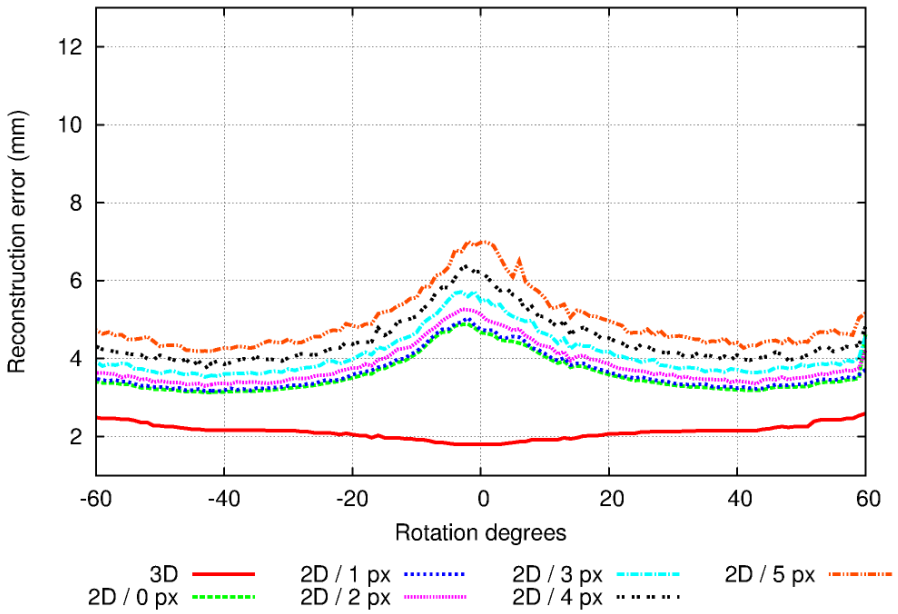


Figure 3.16: Average reconstruction error of synthetic images with pose variation using the symmetry of the face as an additional information.

face was also used in the texture warping stage for all methods.

All these approaches were employed to reconstruct the face of one subject from the Multi-PIE database [Gross et al., 2008], shown in Figure 3.17(a). Figures 3.17(b) and 3.17(c) show Choi et al.'s [2010] results for multiple images and for a single image, respectively. As may be seen, the result for multiple images is very realistic, while the result for a single image is a mixture of the generic face model and the subject's face. This happens because the generic face is not deformed in Choi et al.'s [2010] method when only a single image is available. When using our method for a single image we obtain a realistic result as well, as shown in Figure 3.17(d), showing that we do not need multiple images to achieve a comparable reconstruction accuracy. Figure 3.18 shows another example of the reconstruction result using a single half-frontal image.

3.4.2 Discussion

We have presented a new 3D face reconstruction method that uses only a single face image with arbitrary pose as input. It combines a sparse 3D deformable model and a simple camera model to estimate the 3D coordinates of 2D facial landmarks using an

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

Figure 3.17: Reconstruction results for (a) a subject of the Multi-PIE database using Choi *et al.'s* method for (b) multiple images and (c) a single image, and using (d) the proposed method.

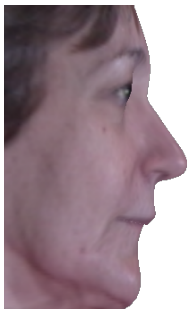
iterative minimization of the reprojection error. The BU-3DFE database was used in our experiments to evaluate the reconstruction accuracy across pose variation and noise.

Corroborating previous neuropsychology works [Hole and Bourne, 2010], our experiments have shown that half-frontal images present advantages over frontal images. Also, assuming that faces are symmetric improved the reconstruction accuracy and the robustness against large pose variations and noise. The average reconstruction error of half-frontal images ranged from 3.3 to 4.5, depending on the noise, using only unknown

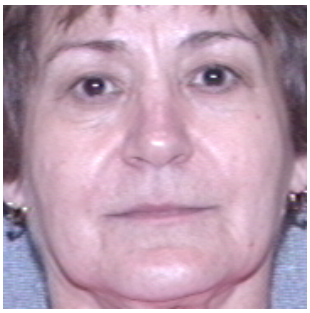
(a)

(b)

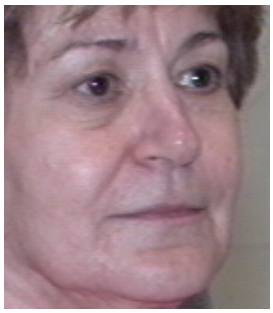
(c)



(d)



(e)



(f)

(g)

Figure 3.18: Reconstructed model using a single (a) half-frontal image rendered from (b)-(d) different viewpoints, similar to (e)-(g) other images from the same subject.

subjects for testing.

CHAPTER 4

CONCLUSION

The objectives of this work were both (1) the development of a fully automatic, real-time 3D face recognition system using low-cost sensors and (2) providing a compatibility mechanism between the developed system and the current forms of identification. To attend our performance requirements, we designed a completely new 3D face detector, optimize all other stages of the recognition framework, and put everything together in an efficient way. To address the compatibility issues, we have developed a new 3D face reconstruction method that requires a single 2D image as input and works across large pose variations.

The developed system is, to the best of our knowledge, the first fully automatic 3D face recognition system using low-cost acquisition devices that handles multiple fps. The Microsoft Kinect is used to capture a 3D video stream. The face detection module, based on boosted cascade classifiers and orthogonal projection images, is one of the most successful in the literature, being robust to the most common problems in 3D images. To standardize the pose and the resolution of the detected faces, they are aligned to a reference model using ICP and then a uniform grid sampling is applied. HOG descriptors are extracted from normalized images and then used as biometric template for matching and evaluation.

This system was successfully applied to the continuous authentication problem, which consists in monitoring the identity of users during the whole access and not only at login. The fusion of the recognition results over time used an improved version of Sim et al.'s [2007] Temporal-First integration, which reduced the memory consumption and stabilized the safety measurements in the initial seconds of the process.

Our 3D face reconstruction method fits a sparse 3D deformable face model to facial landmarks in a 2D image to estimate their 3D coordinates, since it is not possible to

obtain it directly from the 2D image without a prior knowledge about the shape of faces. Our experimental results have corroborated previous neuropsychology works [Hole and Bourne, 2010], which pointed out the advantages of using half-frontal face images over frontal and profile images.

4.1 Achieved results

Our 3D face detector was tested in more than 13,000 face images from 630 unknown subjects and six different databases, and most of these images presented at least one artifact (*e.g.* facial expressions, noise, pose, occlusions). Nevertheless, and it was able to detect 99% of the faces, with less than 1% FDR. These results represent the state-of-the-art in 3D face detection.

Our experiments regarding the continuous authentication system have used more than 10 hours of genuine and impostor accesses, and it obtained a 0.8% EER, which is the lowest EER obtained so far by a continuous authentication system in the literature. Also, the system was able to detect most of the impostors within a one-second window.

The average reconstruction error of our 3D face reconstruction method ranged from 3.3 to 7mm, depending on the pose and the amount of noise, and all testing subjects were completely unknown to the system. As a comparison, these results are similar to the error in Kinect measurements for objects up to 1500mm away from the sensor (see Figure 2.4), which is the maximum distance for a face to be used in our recognition system.

4.2 Future directions

With the rapid advances in 3D imaging, we believe that in the near future 3D sensors will be more accurate and will cover larger areas, and our system could be applied to video surveillance and access control of crowded areas. In these cases, parallelism could be used to handle several faces simultaneously. Also, texture and/or infrared images could be combined to the depth information to investigate if it is possible to obtain more accurate results with a multimodal recognition system.

We would like to apply our 3D detector to other objects, and also try different detection methods on the orthogonal projection images. We could also use orthogonal projection images for other purposes, such as segmentation and recognition.

Finally, we would like to extend our 3D face reconstruction method to include a dense deformable face model. This way, the results are not only going to be realistic, but they are going to be more accurate for recognition purposes as well.

BIBLIOGRAPHY

F. Agrafioti and D. Hatzinakos. ECG biometric analysis in cardiac irregularity conditions. *Signal, Image and Video Processing*, 3(4):329–343, 2009.

A. Altinok and M. Turk. Temporal integration for continuous multimodal biometrics. In *Proceedings of the Workshop on Multimodal User Authentication*, pages 131–137, 2003.

N. Alyuz, B. Gokberk, and L. Akarun. Adaptive registration for occlusion robust 3d face recognition. In *Computer Vision – ECCV 2012. Workshops and Demonstrations*, volume 7585 of *Lecture Notes in Computer Science*, pages 557–566. Springer, 2012.

J. Batlle, E. Mouaddib, and J. Salvi. Recent progress in coded structured light as a technique to solve the correspondence problem: A survey. *Pattern Recognition*, 31(7): 963–982, 1998.

H. Bay, T. Tuytelaars, and L. Gool. Surf: Speeded up robust features. In *Computer Vision ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer, 2006.

P. N. Belhumeur, J. P. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

P. J. Besl and H. D. Mckay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

B. Bhanu and X. Zhou. Face recognition from face profile using dynamic time warping. In *Proceedings of the 17th International Conference on Pattern Recognition*, pages 499–502, 2004.

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

M. Böhme, M. Haker, K. Riemer, T. Martinetz, and E. Barth. Face detection using a time-of-flight camera. In *Dynamic 3D Imaging*, volume 5742 of *Lecture Notes in Computer Science*, pages 167–176. Springer, 2009.

R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. *Guide to Biometrics*. Springer, 2003.

F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:567–585, 1989.

K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006.

P. Buddharaju, I. T. Pavlidis, P. Tsiamyrtzis, and M. Bazakos. Physiology-based face recognition in the thermal infrared spectrum. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):613–626, 2007.

K. I. Chang, K. W. Bowyer, and P. J. Flynn. Adaptive rigid multi-region selection for handling expression variation in 3d face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2005.

K. I. Chang, K. W. Bowyer, and P. J. Flynn. Multiple nose region matching for 3d face recognition under varying facial expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1695–1700, 2006.

J. Choi, G. Medioni, Y. Lin, L. Silva, O. Bellon, M. Pamplona Segundo, and T. C. Faltemier. 3D face reconstruction using a single or multiple views. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 3959–3962, 2010.

A. Colombo, C. Cusano, and R. Schettini. 3d face detection using curvature analysis. *Pattern Recognition*, 39(3):444–455, 2006.

A. Colombo, C. Cusano, and R. Schettini. Gappy pca classification for occlusion tolerant 3d face detection. *Journal of Mathematical Imaging and Vision*, 35(3):193–207, 2009.

Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

I. G. Damousis, D. Tzovaras, and E. Bekiaris. Unobtrusive multimodal biometric authentication: the HUMABIO project concept. *EURASIP Journal on Advances in Signal Processing*, 2008(110):1–11, 2008.

J.G. Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15 (11):1148–1161, 1993.

O. Déniz, G. Bueno, J. Salido, and F. De la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598–1603, 2011.

M. M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, 2009.

G. R. Doddington. Speaker recognition – identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664, 1985.

J. Fischer, D. Seitz, and A. Verl. Face detection using 3-d time-of-flight and colour cameras. In *Proceedings of the 41st International Symposium on Robotics and 6th German Conference on Robotics*, pages 1–5, 2010.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

E. Flior and K. Kowalski. Continuous biometric user authentication in online examinations. In *Proceedings of the 7th International Conference on Information Technology: New Generations*, pages 488–492, 2010.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, pages 23–37, 1995.

H. Ghorayeb, B. Steux, and C. Laurgeau. Boosted algorithms for visual object detection on graphics processing units. In *Asian Conference on Computer Vision*, volume 3852 of *Lecture Notes in Computer Science*, pages 254–263. Springer, 2006.

R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-PIE. In *Proceedings of the Eighth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8, 2008.

D. Gunetti and C. Picardi. Keystroke analysis of free text. *ACM Transactions on Information and System Security*, 8(3):312–347, 2005.

S. Gupta, K. R. Castleman, M. K. Markey, and A. C. Bovik. Texas 3d face recognition database. In *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 97–100, 2010.

M. Hernandez, J. Choi, and G. Medioni. Laser scan quality 3-d face modeling using a low-cost depth camera. In *Proceedings of the 20th European Signal Processing Conference*, pages 1995–1999, 2012.

D. Herrera C., J. Kannala, and J. Heikkilä. Joint depth and color camera calibration with distortion correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):2058–2064, 2012.

R.I. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T.B. Moeslund, and G. Tranchet. An rgb-d database using microsoft's kinect for windows for face detection. In *Proceedings of*

the Eighth International Conference on Signal Image Technology and Internet Based Systems, pages 42–46, 2012.

R. Hietmeyer. Biometric identification promises fast and secure processings of airline passengers. *The Internationl Civil Aviation Organization Journal*, 55(9):10–11, 2000.

P.S. Hiremath and Manjunath Hiremath. Linear discriminant analysis for 3d face recognition using radon transform. In *Multimedia Processing, Communication and Computing Applications*, volume 213 of *Lecture Notes in Electrical Engineering*, pages 103–113. Springer, 2013.

G. Hole and V. Bourne. *Face Processing: Psychological, Neuropsychological, and Applied Perspectives*. Oxford, 1 edition, 2010.

B. K. P. Horn, H. M. Hilden, and S. Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society America*, 5(7): 1127–1135, 1988.

C. Huang, H. Ai, Y. Li, and S. Lao. High-performance rotation invariant multiview face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4): 671–686, 2007.

S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 559–568, 2011.

A. K. Jain, L. Hong, and R. Bolle. On-line fingerprint verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):302–314, 1997.

A. K. Jain, S. Pankanti, S. Prabhakar, L. Hong, A. Ross, and J. L. Wayman. Biometrics: A grand challenge. In *Proceedings of the International Conference on Pattern Recognition*, volume 2, pages 935–942, 2004a.

A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004b.

V. Jain and E. Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 577–584, 2011.

R. Janakiraman, S. Kumar, S. Zhang, and T. Sim. Using continuous face verification to improve desktop security. In *Proceedings of the 7th IEEE Workshop on the Applications of Computer Vision*, pages 501–507, 2005.

A. Janoch, S. Karayev, Y. Jia, J.T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1168–1174, 2011.

D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, and W. Gao. Efficient 3D reconstruction for face recognition. *Journal of Pattern Recognition*, 38(6):787–798, 2005.

I. A. Kakadiaris, G. Passalis, G. Toderici, M. N. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):640–649, 2007.

I. A. Kakadiaris, H. Abdelmunim, W. Yang, and T. Theoharis. Profile-based face recognition. In *Proceedings of the Eighth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8, 2008.

I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):394–405, 2011.

J. Leggett, G. Williams, M. Usnick, and M. Longnecker. Dynamic identity verification

via keystroke characteristics. *International Journal of Man-Machine Studies*, 35(6): 859–870, 1991.

M. D. Levine and Y. Yu. State-of-the-art of 3d facial reconstruction methods for face recognition based on a single 2d training image per person. *Pattern Recognition Letters*, 30(10):908–913, 2009.

S. Z. Li, R. Chu, S. Liao, and L Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (4):627–639, 2007.

R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings of the International Conference on Image Processing*, volume 1, pages 900–903, 2002.

D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157, 1999.

X. Lu and A. K. Jain. Multimodal facial feature extraction for automatic 3d face recognition. Technical report, Department of Computer Science, Michigan State University, 2005.

X. Lu and Anil K. Jain. Automatic feature extraction for multiview 3d face recognition. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 585–590, 2006.

X. Lu, A.K. Jain, and D. Colbry. Matching 2.5d face scans to 3d models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):31–43, 2006.

A. Mansur, Y. Makihara, and Y. Yagi. Inverse dynamics for action recognition. *IEEE Transactions on Cybernetics*, 43(4):1226–1236, 2013.

D. W. Marquardt. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics*, 11(2):431–441, 1963.

G. F. Marshall and G. E. Stutz. *Handbook of Optical and Laser Scanning*. Optical Science and Engineering. Taylor & Francis Group, 2011.

G. Medioni, J. Choi, C.-H. Kuo, and D. Fidaleo. Identifying noncooperative subjects at a distance using face images and inferred three-dimensional face models. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 39(1): 12–24, 2009.

A. Mian, M. Bennamoun, and R. Owens. An efficient multimodal 2d-3d hybrid approach to automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1927–1943, 2007.

R. Min, J. Choi, G. Medioni, and J. Dugelay. Real-time 3d face identification from a depth camera. In *Proceedings of the 21st International Conference on Pattern Recognition*, pages 1739–1742, 2012.

J.V. Monaco, N. Bakelman, S.-H. Cha, and C.C. Tappert. Developing a keystroke biometric system for continual authentication of computer users. In *Proceedings of the European Intelligence and Security Informatics Conference*, pages 210–216, 2012.

H. Moon and P. J. Phillips. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception*, 30(3):303–321, 2001.

S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006.

K. Niinuma, U. Park, and A. K. Jain. Soft biometric traits for continuous user authentication. *IEEE Transactions on Information Forensics and Security*, 5(4):771–780, 2010.

O. Ocegueda, G. Passalis, T. Theoharis, S. K. Shah, and I. A. Kakadiaris. Ur3d-c: Linear dimensionality reduction for efficient 3d face recognition. In *Proceedings of the International Joint Conference on Biometrics*, pages 1–6, 2011.

O.O. Omotade. Facial measurements in the newborn (towards syndrome delineation). *Journal of Medical Genetics*, 27(6):358–362, 1990.

M. Pamplona Segundo, C.C. Queirolo, L. Silva, and O.R.P. Bellon. Automatic 3d facial segmentation and landmark detection. In *Proceedings of the 14th International Conference on Image Analysis and Processing*, pages 431–436, 2007.

M. Pamplona Segundo, L. Silva, O. R. P. Bellon, and C. C. Queirolo. Automatic face segmentation and facial landmark detection in range images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(5):1319–1330, 2010.

M. Pamplona Segundo, L. Silva, and O.R.P. Bellon. Real-time scale-invariant face detection on range images. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 914–919, 2011.

M. Pamplona Segundo, L. Silva, and O.R.P. Bellon. Improving 3d face reconstruction from a single image using half-frontal face poses. In *Proceedings of the 19th IEEE International Conference on Image Processing*, pages 1797–1800, 2012.

M. Pamplona Segundo, S. Sarkar, D. Goldgof, L. Silva, and O.R.P. Bellon. Continuous 3d face authentication using rgb-d cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 64–69, 2013a.

M. Pamplona Segundo, L. Silva, O. R. P. Bellon, and S. Sarkar. Orthogonal projection images for 3d face detection. *Pattern Recognition Letters*, 2013b.

U. Park and A.K. Jain. 3d face reconstruction from stereo video. In *Proceedings of the 3rd Canadian Conference on Computer and Robot Vision*, page 41, 2006.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazin*, 2(11):559–572, 1901.

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

R. Plamondon and S. N. Srihari. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (1):63–84, 2000.

P.M. Prendergast. Facial proportions. In *Advanced Surgical Facial Rejuvenation*, pages 15–22. Springer, 2012.

C. C. Queirolo, L. Silva, O. R. P. Bellon, and M. Pamplona Segundo. 3d face recognition using simulated annealing and the surface interpenetration measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):206–219, 2010.

A. Reichinger. Kinect pattern uncovered. azt.tm's Blog, 2011.

Z. Ren, J. Yuan, J. Meng, and Z. Zhang. Robust part-based hand gesture recognition using kinect sensor. *IEEE Transactions on Multimedia*, 15(5):1110–1120, 2013.

Szymon Rusinkiewicz and Marc Levoy. Efficient Variants of the ICP Algorithm. In *Proceedings of the Third International Conference on 3D Digital Imaging and Modeling*, pages 145–152, 2001.

R. Sanchez-Reillo, C. Sanchez-Avila, and A. Gonzalez-Marcos. Biometric identification through hand geometry measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1168–1171, 2000.

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

A. Savran, N. Alyuz, H. Dibeklioglu, O. Celiktutan, B. Gokberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *Proceedings of the 1st Workshop on Biometrics and Identity Management*, pages 47–56, 2008.

I. Shimshoni, Y. Moses, and M. Lindenbaumlpr. Shape reconstruction of 3d bilaterally symmetric surfaces. In *Proceedings of the International Conference on Image Analysis and Processing*, pages 76–81, 1999.

A. Shpunt and B. Pesach. Optical pattern projection. US Patent Application, Pub. No. US 2010/0284082 A1, 2010.

T. Sim, S. Zhang, R. Janakiraman, and S. Kumar. Continuous verification using multimodal biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):687–700, 2007.

D. A. Socolinsky, L. B. Wolff, J. D. Neuheisel, and C. K. Eveland. Illumination invariant face recognition using thermal infrared imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 527–534, 2001.

The Snell Group. A little white lie. Think Thermally, May 2002.

F. Tsalakanidou, S. Malassiotis, and M. G. Strintzis. Face localization and authentication using color and depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):152–168, 2005.

M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

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

S. Vural, Y. Mae, H. Uvet, and T. Arai. Multi-view fast object detection by using extended haar filters in uncontrolled environments. *Pattern Recognition Letters*, 33(2):126–133, 2012.

S.-F. Wang and S.-H. Lai. Reconstructing 3d face model with associated expression deformation from a single face image via constructing a low-dimensional expression deformation manifold. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2115–2121, 2011.

Y. Wang, J. Liu, and X. Tang. Robust 3d face recognition by local shape difference boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10): 1858–1870, 2010.

Z. Xu, X. Guo, X. Hu, X. Chen, and Z. Wang. The identification and recognition based on point for blood vessel of ocular fundus. In *Advances in Biometrics*, volume 3832 of *Lecture Notes in Computer Science*, pages 770–776. Springer, 2005.

M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

L. Yin, X. Wei, Y. Sun, J. Wang, and M.J. Rosato. A 3D facial expression database for facial behavior research. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006.

D. Zhang, W.-K. Kong, J. You, and M. Wong. Online palmprint identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1041–1050, 2003.

W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.