FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

3D Face Recognition Under Unconstrained Settings Using Low-Cost Sensors

Tiago Daniel Santos Freitas

U. PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Bioengenharia - Ramo de Engenharia Biomédica

Supervisors: Jaime S. Cardoso and João C. Monteiro

June 2016

3D Face Recognition

Under Unconstrained Settings
Using Low-Cost Sensors

Tiago Daniel Santos Freitas

Mestrado Integrado em Bioengenharia - Ramo de Engenharia Biomédica

June 2016

Abstract

Real-world face recognition systems face major challenges in unconstrained environments, due to variations in pose, illumination, facial expression an disguises. Due to these challenges the performance of the systems is limited, when using only RGB images.

The integration of 3D information as a complement to RGB can improve the robustness of face recognition systems, diluting the performance losses in challenging environments.

In this dissertation, a new face recognition framework is proposed, highly focused in 3D-low-cost images. The developed system uses a combination of FHOG and PHOW for feature extraction, showing promising results in unconstrained environments.

With the appearance of the Intel® RealSense$^{\text{TM}}$ depth sensors, a new dataset was created, the RealFace dataset, acquired with the Intel® RealSense$^{\text{TM}}$ R200 and F200 models, composed by images characterized by variations in pose, illumination facial expression and disguise are included. This dataset is designed for multimodal face-recognition systems performance assessment in unconstrained environments.

The proposed 3D framework was extended to integrate other modalities - IR and RGB - using the features extracted from VGG-Face as descriptors for these modalities. The proposed framework has proven to be competitive with other state-of-the-art methodologies having good results in unconstrained environments.

Finally, a prototype using Intel® RealSense$^{\text{TM}}$ F200 sensor was created for real-time face recognition in challenging scenes.

Acknowledgements

Firstly I would like to express my gratitude to Prof. Jaime S. Cardoso and João Carlos Monteiro, for their patience, guidance and knowledge during the development of this dissertation.

I would also like to show my gratitude to INESC-TEC for providing the Intel® RealSense$^{\text{TM}}$ sensors and to the FEUP group Visual Intelligence Lab (VILAB) for helping in the creation of the RealFace Dataset.

Equally, to all the participants of RealFace dataset, I have to give a special word of appreciation, since without them the dataset could not be created.

To my family and friends for their continuous support. To Beatriz for her unconditional encouragement, specially in the most difficult moments.

Finally, a special thankful note to my parents, for educating me, supporting me, and, although I usually do not admit it, for helping me find Bioengineering, since without them I would not be here. If I found out what I would like to do in the future it was because of them.

“The essence of being human is that one does not seek perfection.”

George Orwell

Contents

List of Figures

List of Tables

Abbreviations and Symbols

AFR	Automatic Face Recognition
CLAHE	Contrast Limited Adaptive Histogram Equalization
CMC	Cumulative Match Characteristic
CNNs	Convolutional Neural Networks
CPD	Coherent Point Drift
DCT	Discrete Cosine Transform
DJBF	Directional Joint Bilateral Filtering
DOF	Degrees-of-Freedom
DoG	Difference of Gaussians
dSIFT	Densely-Sampled Sets of Scale-Invariant Feature Transform
DWT	Discrete Wavelet Transform
EGI	Extended Gaussian Images
EHMM	Embedded hidden Markov model
eLBP	Extended Local Binary Patterns
ELMDP	Enhanced Local Mixed Derivative Pattern
FAR	False Acceptance Rate
FHOG	Felzenszwalb's Histogram of Oriented Gradients
FPLBP	Four-Patch Local Binary Patterns
FRR	False Rejection Rate
GMM	Gaussian Mixture Models
HAOG	Histogram of Averaged Oriented Gradients
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients
ICA	Independent Component Analysis
ICP	Iterative Closest Point
IMU	Inertia Measurement Unit
IR	Infra-Red
JBF	Joint Bilateral Filter
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LDP	Local Derivative Patterns
LGBP	Local Gabor Binary Patterns
LNP	Local Normal Patterns
LPA	Local Polynomial Approximation Filter
MAP	Maximum a Posteriori
mLBP	Modified Local Binary Patterns
PCA	Principal Component Analysis

RANSAC	Random sample consensus
RBF	Radial Basis Function
R1RR	Rank-1 Recognition Rate
RGB	Red-Green-Blue
RGB-D	Red-Green-Blue-Depth
ROC	Receiver Operating Rate
ROI	Region of Interest
SDK	Software Developer Kit
SFR	Spherical Face Representation
SIFT	Scale-invariant feature transform
SNR	Signal-to-Noise Ratio
SSI	Sphere-Spin-Images
SURF	Speeded-Up Robust Features
SVM	Support Vector Machine
TPLB	Three-Patch Local Binary Patterns
TOF	Time-of-Flight
UBM	Universal Background Model
VR	Verification Rate
WSRC	Weighted Sparse Representation-based Classifier

Chapter 1

Introduction

1.1 Motivation

The face is a natural, easily acquirable trait with a high degree of uniqueness, being capable of discriminating a subject's identity.

Real-world face recognition systems, based on 2D information, are still a challenging problem, due to variations in illumination conditions, the presence of occlusions, pose variations, facial expression changes and disguises. In uncontrolled environments, the problem is even bigger as these factors increase due to non-collaborating failures of the identification process (Abate et al., 2007).

Human beings are able to instinctively recognize hundreds of familiar faces using memory as a support (Diamond and Carey, 1986). The face recognition task can be considered as a particular problem of general object recognition (Tsao and Livingstone, 2008), that could be solved automatically by computer vision algorithms. It is known that face recognition capability has earlier physiological development than object recognition and is more affected by orientation (Farah, 1996). Also, the neurological process involved in face recognition is different from the recognition of other non-facial optical stimuli (Farah et al., 1999).

Biometric Recognition is a problem that has been strongly discussed and researched due to its multiple real-world applications. Security biometric systems are the most important application of this task, although homeland security, law enforcement and identity management are some of other possible applications (Sang et al., 2015).

In this kind of systems the most common biometric traits are fingerprint and iris. These two types of systems have a lot of drawbacks, as iris recognition systems are expensive (although highly accurate) and fingerprints are not suitable for non-collaborative individuals. In real-time systems, face recognition seems to be a good solution due to the fact that it can be less costly than iris-based solutions and also work in non-collaborative conditions (Abate et al., 2007).

Variations in illumination, either due to skin reflectance properties in different environment conditions, or due to internal camera controls, can affect the performance of some systems that only perform consistently under moderate illumination variations. Pose variations, that introduce different views of the head and face, generally decrease the system's accuracy, as most systems are prepared for recognition mainly in frontal poses. Also, different facial expressions can generate variations in performance, although only extreme variations can lead to significant errors in performance. Occlusions, especially if they occlude the upper half of the face, remove significant discriminative information that drastically diminish the recognition rate (Abate et al., 2007). These occlusions can occur due to a multiplicity of reasons and can be occasional or deliberate, owing to accessories or facial hair. They may be present in criminal cases, but also can be due to religious matters (burkas). A not least important factor, is the time delay (Bennamoun et al., 2015), which is less explored in the state-of-the-art. Due to aging, the face varies in an unpredictable way, making difficult to recognize an individual. Illness and massive changes in weight can also produce alterations in the person face, posing a hard to solve problem on biometric face systems. All these factors must be taken into account when building systems aimed for automatic face recognition.

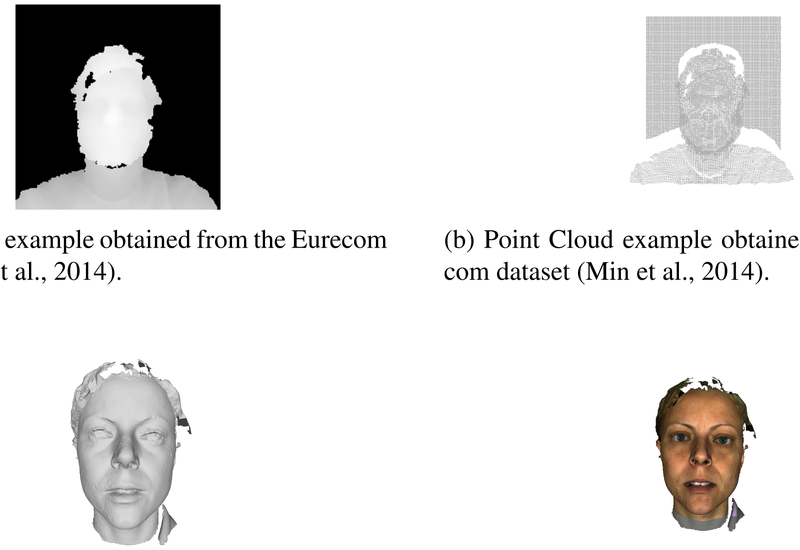
Due to the inherent 3D structure of the face, changes in illumination conditions and non-frontal pose from the individuals can lead to shadows that change some visual facial features, making recognition systems less effective. To overcome the decrease in performance in these situations, 3D face recognition can be used to improve the recognition rate, yielding a more robust facial description less affected by illumination variation, as well as compensating for the changes in pose by making multiple views available to the face recognition system (Bennamoun et al. (2015), Abate et al. (2007)).

In 3D face recognition we can have two main representations of the 3D facial structure: 2.5-Depth images and 3D images (Abate et al., 2007). 3D images have a face and head representation, retaining all the facial geometry information. On the other hand, 2.5D, or range images, are bi-dimensional representations of a set of 3D points in which each pixel in the XY plane stores the depth z value (that corresponds to the distance to the acquisition sensor). The disadvantage of this representation is that it only takes information from one point of view, resulting only in a partial facial model (and not the complete head). Using a group of 2.5D scans from different points of view we can build a 3D model, though. Also, a 3D image depends only on internal anatomical structure, while 2.5D scans are affected by environmental conditions (Abate et al., 2007) and external appearance (it is partially affected by illumination).

The 3D model can be represented either as a Point Cloud or as a mesh image. A Point Cloud can be considered as an unordered collection of the tridimensional coordinates, while a mesh is an aggregate of vertices and edges (generally triangular) that are used for the representation of the face (Bennamoun et al., 2015). The depth map can also be converted in a point cloud or in a 3D mesh. These 3D meshes can be fused with 2D-RGB information to obtain textured meshes. Examples of all these 3D representations are shown on Figure 1.1.

All these representations could lead to an increase in performance of recognition algorithms, especially when combined with RGB image information, as will be discussed in the next chapter, where a review of the most common 3D sensors, public datasets, and state-of-the-art methodologies discussion will be presented.

This dissertation will focus on the development of a real-time system that can perform face



(a) Depth map example obtained from the Eurecom dataset (Min et al., 2014).

(b) Point Cloud example obtained from the Eurecom dataset (Min et al., 2014).

(c) Mesh example obtained from the Florence Superface dataset (Berretti et al., 2012).

(d) Textured Mesh example obtained from the Florence Superface dataset (Berretti et al., 2012).

Figure 1.1: Some examples of the four different types of 3D face image representations.

recognition using 2D and 3D data, acquired with a low-cost sensor. The multimodal framework should be competitive and robust across pose variation, uncontrolled illumination environments, facial expression variations and occlusions.

1.2 Contributions

This dissertation has produced some scientific contributions, namely:

1. Creation of a framework for 3D and multimodal face recognition. The developed system presented overall results superior to some state-of-the-art algorithms and showed to be robust to variations in pose, illumination, facial expression and disguises. The proposed system was also extended for multi-modality, being able to integrate RGB and/or IR data.

2. Creation of a face dataset using the novel sensor Intel® RealSense$^{\text{TM}}$, the RealFace Dataset, designed for multimodal face recognition, providing three types of modalities (IR, RGB and depth + Point Cloud) in controlled and uncontrolled environments. This dataset allowed to evaluate RealSense$^{\text{TM}}$ sensors in face recognition systems, providing an alternative to the more common Kinect-based databases. The presence of images in darkness conditions, pose and facial expression variations, allow to provide a challenging dataset which can be used by the scientific community for algorithm development and testing.

3. Construction of a real-time face recognition prototype that uses Intel® RealSense$^{\text{TM}}$, capable of operating in diverse conditions and environments.

The referred contributions resulted in a publication in U.Porto Journal of Engineering (Monteiro et al., 2016), where a review of 3D-face recognition state-of-the art methods, sensors and datasets is performed. Additionally, the article includes an extension of the framework developed in Monteiro and Cardoso (2015) to integrate 2D and 3D data. The paper is on in the last stage of revision.

The next section will lay out the structure of the rest of the dissertation.

1.3 Dissertation Structure

The rest of this dissertation is divided in six chapters. The motivation and introduction to 3D acquisition has already been carried out in this first chapter.

Chapter 2 will include a brief review of 3D sensors used for face recognition, with special focus on the low-cost alternatives. Additionally, the publicly available datasets for 3D face recognition, that can be use in algorithm testing and refinement, will be detailed while also reviewing some of the state-of-the-art approaches in 3D and multimodal face recognition.

Chapter 3 will describe the creation of a new dataset acquired with Intel® RealSense^{TM} Sensors, the RealFace Dataset.

Chapter 4 will include the testing setups as well as the corresponding results for a variety of pre-processing, feature extractors and classifiers. The results will justify the proposal of a new framework, which is presented with the correspondent results in Eurecom and RealFace datasets.

In Chapter 5 an initial real-time face recognition prototype developed with Intel® RealSense^{TM} will be described.

Finally, Chapter 6 will include the conclusions of this work as well as some suggestions for future work.

Chapter 2

3D Face Recognition: A State-Of-The-Art Review

2.1 Introduction

To assess the identity of a specific unknown query face, Automatic Face Recognition systems (AFR) rely on previous knowledge gained from samples of a database of the subjects that need to be identified. In scientific research, for robustness evaluation, public datasets are generally used, allowing the comparison of performance of different approaches in the same conditions.

Face recognition systems can work in two different modes: identification / recognition and authentication / verification (Abate et al., 2007). The verification problem consists in a comparison between a query face against a specific template face image of the claimed identity. Face identification is more complex, consisting in a multiclass classification problem that compares the input face against all image templates in a database. Therefore in the first category we have a 1:1 problem whereas in the second one we have a 1:N, leading to a higher probability of incorrect classification (Bowyer et al., 2006).

These systems have a typical pipeline design as shown in Figure 2.1 (Bennamoun et al., 2015), generally including the following steps:

1. *Data Acquisition:* in this phase 2D and/or 3D images are captured using a camera, 3D sensor or depth sensor (frames from video can also be acquired). The captured images serve as input to the following blocks of the system.

2. *Preprocessing:* after acquisition, the images or video frames are preprocessed to reduce the influence of noise in the image, improving the signal-to-noise ratio of acquired data. Due to the noisy nature of depth images captured by low-cost sensors, preprocessing plays a major role in the pipeline. This phase includes smoothing, spike and holes removal, but also the detection of the Region of interest (ROI), which will include the face region, removing the influence of surrounding uninformative pixels. Additionally, it can include normalization of data against illumination, scale and orientation variations.

3. *Feature Extraction:* this phase aims at extracting discriminant characteristics from the image, allowing the system to assess the identity of the face image. In the end, the system uses these features and not the whole images for identity assessment. The extracted features can be divided in two categories: holistic features (describing the face) or local features (get information from specific regions).

4. *Classification:* In the final procedure, one or more previously trained classifiers rely on knowledge from previous data for the attribution of some identity to the query input (or rejection of the image in real-time systems).

After classification, performance is assessed by different criteria for verification or identification frameworks.

For verification systems, some of the most used criteria are based on receiver operating characteristic curves (ROC), the equal error rate (EER) and the verification rate (VR) at a certain false acceptance rate (FAR), generally at 0.1 % (VR@0.1%FAR). Using different thresholds, the ROC curve can be plotted as the false rejection rate (FRR) versus the FAR or as the VR versus FAR. The area under this curve is also one of the metrics used, being intrinsically dependent with the system accuracy: a larger area under the ROC curve indicates a more robust system (Bennamoun et al., 2015).

Regarding identification systems, the cumulative characteristic curves (CMC) and the rank–1 recognition rate (R1RR) are the most commonly used. The first one is a representation of the samples correctly classified (in percentage) versus the rank at which the correct match is detected. R1RR is represented as the percentage of all the dataset samples in which the best match corresponds to the correct subject. Additionally, it is very common to use R5RR to verify how common is that the true identity is included in the most 5 likely identities classified by the system (Bennamoun et al., 2015).

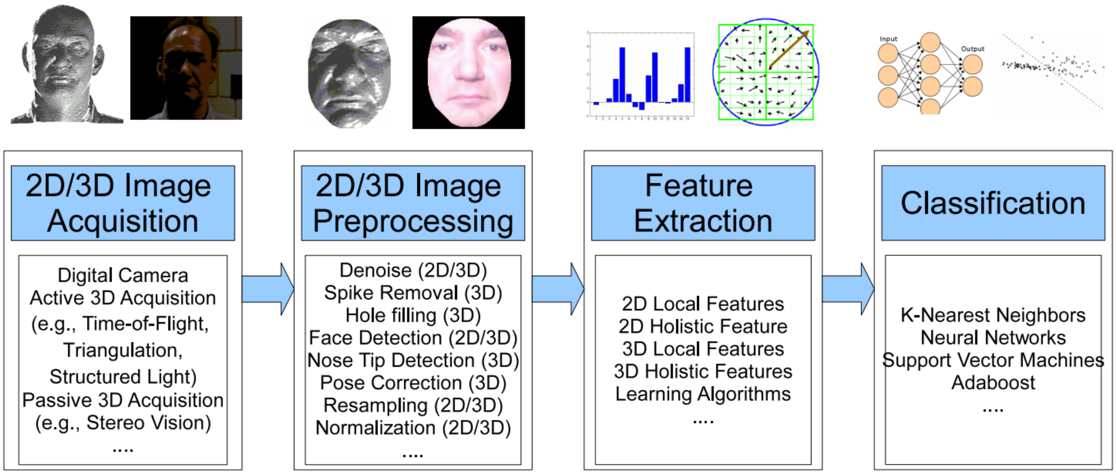


Figure 2.1: Typical work flow of a Facial Recognition System. Obtained from Bennamoun et al. (2015).

Before discussing state-of-the-art approaches, it is important to first evaluate in which sensors this type of 3D facial data can be acquired, with special consideration for the low-cost alternatives.

2.2 Low-Cost 3D Sensors

The trend in face recognition is the use of low-cost sensors that still allow the creation of facial recognition frameworks capable of a high recognition performance. Although, in the past, sensors with high precision like Minolta (Minolta, 2006), Inspeck (Savran et al., 2008), CyberWare (Cyberware, nd) and 3dMD (3dMD, nd) were used, their high prices led to the need for cheaper alternatives. Additionally, these systems are usually not fast and it is desirable to have a real-time system that performs identification as fast as possible. These limitations led to the appearance of low-cost 3D sensors that offer a cheap solution, while also being able to work in real-time, most of them being also portable. Low-cost devices, although offering a lower resolution, should present information with enough quality to perform face recognition in adverse conditions.

These 3D sensors can be classified either as stereoscopic camera systems, structured light systems or laser range systems, obtaining both 3D and RGB information. Some of the most common 3D sensors used in face recognition systems are summarized in Table 2.1.

Table 2.1: List of some sensors used in 3D Facial Recognition

Sensor	Type	Resolution (mm)	Working Distance (m)	Price ($)
Minolta Sensors (Minolta, 2006)	*3D* Laser Scanning	0.041-0.22	∼ 2.5	25000
3dMDface (3dMD, nd)	Vision Cameras	<0.2	——	10k - 20k
CyberWare 3030RGB/PS (Cyberware, nd)	Low-Intensity Laser Light Source	0.08 - 0.3	0.35	∼ 72000
Inspeck Mega Capturer II (Savran et al., 2008)	Structured-Light	0.7	1.1	Not Available
Kinect v1 (Microsoft, 2010)	IR laser Emitter	∼ 1.5 - 0.5	0.5 - 4.5	Not Available
Kinect v2 (Microsoft, 2014)	Time-of-Flight	-	∼ 0.5 - 8	149.99
SoftKinetic DS325 (SofKinetic, 2007)	Diffused Laser	1.4 at 1 m distance	∼ 0.15 - 1	259
Structure (Structure, 2013)	IR Structured Light	0.5 - 30	3.5	379
PrimeSense Carmine (I3DU, nd)	IR Laser Emitter	0.1 - 1.2	3.5	Not Available
ASUS Xtion Pro Live (Asus, 2011)	IR Laser Emitter	-	0.8 - 3.5	169.99
Intel RealSense (Intel, 2015a)	Structured Light	<1	∼ 0.2 - >10	99 - 399

While the Minolta and Inspeck sensors are generic 3D sensors, CyberWare and 3dMD were designed specifically to 3D face scanning. All these sensors were used for 3D face recognition, but as referred before, have been replaced through time with low-cost alternatives.

The original Kinect (Microsoft, 2010), Kinect v1, is the most used sensor for depth acquisition. It consists in an infra-red (IR) laser emitter, an IR camera and a RGB camera. The latter captures color images directly, whereas a conjugation of the laser emitter and IR camera capture the depth information, resulting in a final RGB-D map. This depth map is obtained using a triangulation process based on these two sensors. Primarily the IR laser emitter, using a raster, projects a predesigned pattern of spots in the scene, allowing the capture of the reflection of the pattern by the IR camera.

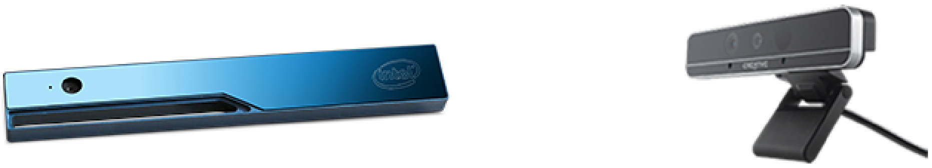
Recently, a new version of this sensor was launched to replace the Kinect v1: the Kinect v2 (Microsoft, 2014) (or Kinect for XBOX One) operates with a different principle, the time-of-flight (TOF). With this methodology the depth images are obtained calculating the time between emitted IR light and its reflections (Dal Mutto et al., 2012). The Kinect v2 offers the possibility of using IR images, which was not possible in the original Kinect. Due to being a recent sensor, the specifications for depth resolution could not be obtained. Recent experiments have resulted in an improved resolution and precision in this new sensor, as discussed in Amon et al. (2014) and Lachat et al. (2015) .

Other low-cost sensors have been developed to compete with Kinect in depth map acquisition. The SoftKinetic DS325 (SofKinetic, 2007), Structure sensor (mobile depth sensor in tablets) (Structure, 2013), Intel® RealSense^{TM} (Intel, 2015a) , ASUS Xtion Pro Live (Asus, 2011) and PrimeSense (I3DU, nd) (recently bought by Apple (Guardian, 2013) and currently not available) have also been used in depth acquisition systems.

From the previously referred sensors, one of the most promising is the Intel® RealSense^{TM} family of depth sensors (Intel, 2015a). Intel® provides two models, the SR300 (previously named F200), for short range applications, and the R200 for long range acquisitions (Figure 2.2). This sensor comes with a Software Developer Kit (SDK), with already implemented modules for facial tracking and detection, and, similarly to Kinect v2, also provides IR images. Additionally, a ZR300 camera, developed for smartphones is also available.

Both these models have the same technology, consisting in 3 cameras that provide RGB images and stereoscopic IR that produce depth maps. With a laser projector, the sensors perform a scene perception and enhanced photography (the depth map can be 3D filtered, allowing re-lighting, re-focusing and background segmentation). More precisely, the R200 camera has two stereo cameras, allowing improved outdoor acquisitions. Using stereoscopy photography, the depth images are computed from the difference (pixel shift) between the two cameras using a triangulation method (Intel, 2015b). The model developed for Android applications (ZR300) has an array of six sensors: the R200 camera, a high-precision accelerometer, a gyroscope, a wide field-of-view camera from motion and feature tracking and, additionally, a rear RGB camera (with 8MP) and a frontal 2MP camera. This is currently the most expensive product, and can be acquired by 399 $. The price of the models has changed along time, but currently the R200 model can be acquired by 99 $, while SR300 had its price increased to 129 $.

The main difference between these camera models is in the operating ranges. The R200 works from 0.5 to 3.5 meters and has an outside range up to 10 meters, while the SR300 model only operates from 0.2 meters to 1.2 meters.



(a) R200 Model.

(b) SR300 Model.

Figure 2.2: Intel® RealSense^{TM} depth camera models.

Intel® provides a powerful SDK with some samples for possible applications of the Intel®

RealSense^{TM}, allowing the development of applications in 5 programming languages: C++, C#, Unity, Java and JavaScript. It provides the user with 78 application samples for the most diverse applications, whether for object recognition, face and hand tracking, eye tracking, facial emotion detection, virtual reality, among others.

All the referred sensors provide powerful tools for face recognition frameworks and can be used to build datasets for algorithm performance assessment.

2.3 Datasets

As discussed in the previous section, depth sensors can be used for the creation of public datasets, which provide important material for algorithm testing and refinement. These databases should try to mimic the main challenges faced in face recognition.

The ideal dataset should have unlimited samples and subjects while also including a large variety of conditions (pose, facial expressions, illumination, occlusions, low resolution).

Publicly available 3D datasets can be included in two main classes: high-resolution scans datasets that are acquired using expensive 3D scanners as Minolta or 3dMDface systems, and low-resolution scans datasets that are obtained using the low-cost sensors mentioned in the previous section.

Naturally, the first datasets created for the 3D facial recognition problem used the high precision sensors. Some of the most important datasets are the Bosphorus (Savran et al., 2008), York (of York, nd), FRGC (Phillips et al., 2005), GavabDB (Moreno and Sanchez, 2004), BinghamtonUniversity(Yin et al. (2006), Yin et al. (2008)), Texas-3D (Gupta et al., 2010), UMB-DB (Colombo et al., 2011), 3D-RMA of Applied Sciences of the Royal Military Academy (nd) and FRAV3D (Kussul et al., 2013)(not available anymore).

Alongside the evolution of sensors towards low-cost, lower resolution and faster acquisitions, more recent databases were also built with this type of sensors. Although the number of 2D+3D datasets is still comparatively low in number to the 2D and high quality 3D datasets, these databases are increasing in number and variety. The specifications of some of these datasets are summarized in Table 2.2.

Table 2.2: Some of the available low-resolution depth maps datasets

Dataset	RGB	*3D* Sensor	Scans	Subjects	Expression	Illumination	Pose	Occlusion	Video
Aalborg University RGB-D Face Database (Hg et al., 2012)	Yes	Kinect v1	1581	31	Yes	No	Yes	No	No
Florence Superface dataset (Berretti et al., 2012)	Yes	Kinect v1	> 14000	20	No	No	Yes	No	Yes
CurtinFaces (Li et al., 2013)	Yes	Kinect v1	>5000	52	Yes	Yes	Yes	Yes	No
UWA Kinect database (Hayat et al., 2015)	Yes	Kinect v1	> 15000	48	Yes	No	Yes	No	No
NASK-StructureFacebase (Gutfeter and Pacut, 2015)	Yes	Structure	330	13	No	No	Yes	No	Yes
BIWI Kinect Head-Pose (Fanelli et al., 2013)	Yes	Kinect v1	> 15000	20	No	No	Yes	No	No
UWA Kinect (Hayat et al., 2015)	Yes	Kinect v1	> 15000	48	Yes	No	Yes	No	No
FaceWareHouse (Cao et al., 2014)	Yes	Kinect v1	3000	150	Yes	No	No	No	Yes
AVL-RGBD Face Database (Hsu et al., 2014)	Yes	Kinect v1	1280	28	No	No	Yes	No	No
Eurecom (Min et al., 2014)	Yes	Kinect v1	> 450	52	Yes	No	Yes	Yes	Yes
IIIT-D face database (Goswami et al., 2014)	Yes	Kinect v1	4605	104	Yes	No	Yes	No	No
Labeled Infrared-Depth Face database (Cao and Lu, 2015)	No	Kinect v2	918	17	Yes	No	Yes	No	No

Some examples of low-resolution databases are the Aalborg University RGB-D Face Database (Hg et al., 2012), Florence Superface dataset (Berretti et al., 2012), CurtinFaces (Li et al., 2013), NASK-StructureFacebase (Gutfeter and Pacut, 2015) , BIWI Kinect Head Pose Dataset (Fanelli et al., 2013), AVL-RGBD Face (Hsu et al., 2014), UWA Kinect dataset (Hayat et al., 2015), Face-WareHouse (Cao et al., 2014), Eurecom dataset (Min et al., 2014) and IIIT-D face database (Goswami et al., 2014). It is important to evaluate the frameworks in these datasets, in order to evaluate the true robustness of the algorithms in challenging environments.

The Aalborg University RGB-D Face Database (Hg et al., 2012) was one of the first available public datasets created. With 1581 samples from 31 different persons, 17 different poses and facial expressions were captured for each subject using Kinect v1.

Florence Superface dataset (Berretti et al., 2012) includes RGB-D video for 20 subjects with large pose variations. Additionally, this dataset also includes 3D high-resolution textured face scans obtained with the 3dMD scanner.

BIWI Kinect Head-Pose dataset (Fanelli et al., 2013) consists in a large database specifically created for head-pose estimation. With more than 15000 scans of 20 individuals, large pose variations are explored. Although not being designed for recognition purposes, it could be an important tool to test algorithms performance across pose variations.

CurtinFaces (Li et al., 2013) contains over 5000 scans of 52 individuals, including variations in pose, illumination, facial expression and occlusions (sunglasses).

The University of Western Australia Kinect Face database was also described by Hayat et al. (2015), with 48 different subjects, each with between 289 to 500 scans. Variations in facial expression and pose were also included.

NASK-StructureFacebase (Gutfeter and Pacut, 2015) presents itself as the only dataset that uses a sensor other than Kinect, using the Structure alternative. Although not being available for public use, it can be considered as an important development in this type of datasets.

FaceWarehouse (Cao et al., 2014) was constructed using facial scans captured with Kinect V1, with 20 different facial expressions, being important to assess the algorithm efficiency against such variations. It is the most complete dataset in terms of variations of facial expression.

AVL-RGBD Face Database (Hsu et al., 2014) also explores pose variations with variation in distances containing 13 different poses (only in one plane) at 5 different distances (with a maximum distance of 2 meters).

Eurecom Kinect Face database seems to be the most balanced database, although the number of scans is limited. It is likely the most used dataset in algorithm testing. It consists in a multimodal RGB-Depth facial images of 52 individuals (38 males and 14 females), with two sessions acquired with a time lapse of 5 to 14 days, captured with Kinect v1. It includes subjects from different ethnicity (21 Caucasians, 11 from Middle East/Mahgreb, 10 East Asian, 4 Indian, 3 African-American and 3 Hispanic) and all images were taken under controlled conditions. Each subject has 9 different facial expressions or partial occlusions, and the dataset also includes video with slow movements in the horizontal and vertical directions.

The IIIT-D face database (Goswami et al., 2014) includes 4605 scans (both RGB and depth) from 106 subjects, captured in two sessions with the Kinect v1 Sensor. The number of images per person varies from a minimum of 11 images and a maximum of 254 images. It includes variations in pose and expression (in some cases, there are variations due to eyeglasses as well).

The Labeled Infrared-Depth Face database (Cao and Lu, 2015)is, to the extent of our knowledge, the most recent dataset and was acquired with Kinect v2. It does not provide RGB data, but it does provide the data from Infrared stream, aligned with depth data. It includes data from 17 individuals, with a total 918 scans. Each subject has scans in 9 different poses with 6 different expressions (with a total of 54 scans per subject). The manually labeled keypoints for each image are also provided.

A clear pattern is recognized in these datasets, as only Kinect sensor is used (except in NASK-StructureFacebase). There is still not enough available relevant public data that uses the more recent sensors like the Kinect v2 or the Intel® RealSense^{TM} models, that would be useful for advances in facial recognition. There's clearly space for the creation of such datasets using more recent sensors, as that would be scientifically relevant for the facial recognition research community.

The use of these sensors and datasets serves as the main basis for creation of frameworks that are capable of automatic facial recognition in unconstrained conditions.

The appearance of low-cost depth sensors and datasets leads to a necessary adaptation of the 2D-based image frameworks already implemented, for being capable of receiving 3D information as input. A review of the state-of-the-art 3D face recognition methods will be performed in the next section.

2.4 State-Of-The-Art Review

Although some authors tried to explore unimodal 3D recognition systems, the more interesting and discriminating ones are the multimodal systems that combine 2D and 3D information.

Concerning the type of input information, 3D-facial recognition approaches can be classified in three main types: 2D-Based, 3D-based and multimodal. The first uses synthetic 3D face models to increase the robustness of 2D images with respect to pose variations as well as changes in illumination and facial expression. 3D-based methodologies do not use RGB or grayscale information, using only 3D or 2.5D data for the development of recognition algorithms. Finally, multimodal approaches take advantage of information from both previous approaches in order to obtain a better classification performance (Bowyer et al., 2006).

After analyzing the typical pipeline of recognition systems, previously shown on Figure 2.1, it is important to define how the two types of information can be fused to perform a final decision in multimodal systems (Bennamoun et al., 2015). If we perform fusion at the *sensor level*, a textured mesh can be obtained, instead of two different inputs. At the *feature level*, modality fusion also can be performed, fusing 2D and 3D features in a single feature representation. *Score-level fusion*, occurs when the scores obtained from individual classifiers are combined to obtain a final global score (sum rule, minimum rule and the product rule are some of these techniques). We can also have a *fusion at the rank level*, where the ranks obtained from different feature classifications are fused using techniques like consensus voting, highest rank fusion or Borda count rank fusion. Finally, *decision-level fusion* can be made using methodologies like majority voting or behavior knowledge (Bennamoun et al., 2015).

After knowing the typical pipeline and how we can fuse different modalities, we are able to discuss some of the most relevant algorithms available in the state-of-the-art. Hereinafter, will be presented a review of some of the most important 3D approaches with high emphasis on low-resolution depth systems.

2.4.1 2D-based Approaches

2D-based approaches were in the genesis of 3D facial recognition and, although they only use a 2D input query face, 3D models are used to improve the robustness of the system. In this category a set of virtual 3D models are generated to simulate the variations in pose and facial expression.

One of the first works using this approach was described by Blanz and Vetter (2003), where a morphable model is fitted to the input image, estimating the tridimensional shape and texture of the face. The morphable model is based on a vectorized representation of the gallery faces in a convex combination of texture and shape. Using manual feature point selection, the system compares the faces using the coefficients obtained from the generated models. Despite the manual selection, this paper presented a major step in the facial recognition field, introducing the potential of 3D models to increase the robustness of 2D systems. The process of adaptation of the generic morphable model to a specific individual is shown in Figure 2.3.

Lu et al. (2004) also used a 3D generic face model in conjugation with 2D face images to generate facial and texture information. Using also manual feature points, a depth model is created from which models with variations in pose, illumination and facial expression are generated to increase the variability of the training dataset. Face images are classified using the minimum Euclidean distance between the two affine spaces defined for the query face and each identity in the database.

In the same year, Hu et al. (2004) proposed a similar approach. Using a frontal face image, 83 key points are automatically detected and aligned, sufficient for 3D face model reconstruction, assisted by a generic tridimensional model. An orthogonal projection of the 2D intensity images on the generated 3D model is then performed. From these models different pose and facial expression images are generated. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were tested for dimension reduction and the Nearest Neighbor method is used for classification. Here, LDA dimension reduction achieves the best recognition rate.

The previous works have proven that 3D could have a major role in facial recognition, introducing significant improvements in performance. These approaches generally generate non-realistic models and additionally the generation of models from a single frame has several limitations. This led to the research of 3D scanning methods for more accurate models (Bowyer et al., 2006).

2.4.2 3D-based Approaches

The use of 3D unimodal methodologies has shown to be a good alternative their RGB counterparts in conditions of varying illumination, facial expression and pose. The main advantages of this type of methodologies is the preservation of geometrical information about the face even with variation

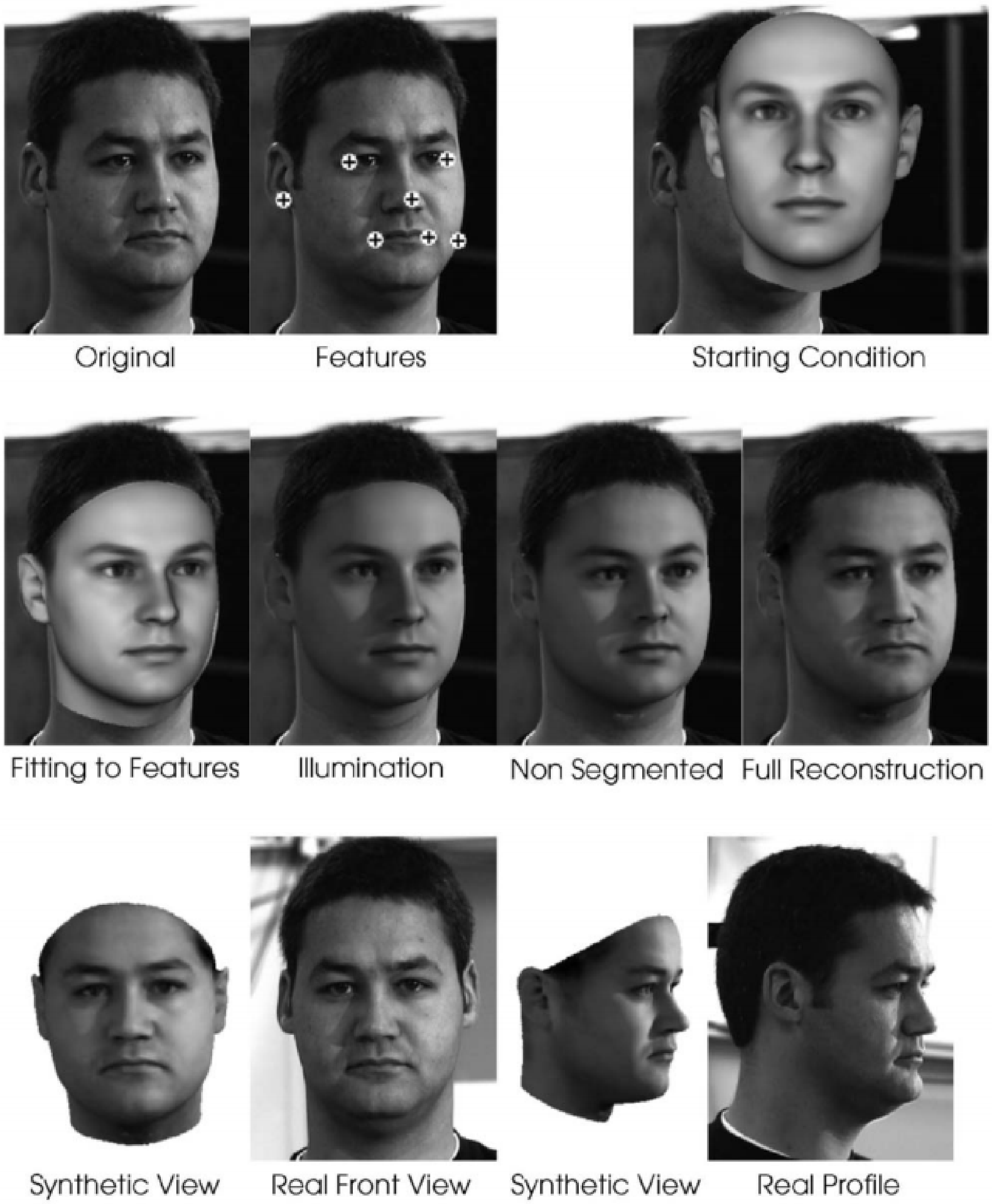


Figure 2.3: Adaptation of a generic 3D morphable model to a specific subject. Obtained from Blanz and Vetter (2003).

in illumination conditions.

Despite that, 3D scanning is not totally invariant to illumination (Bowyer et al., 2006). The use of stereo or structured light sensors involves the capture of one or more RGB images, that can be influenced by illumination, conditioning the quality of the obtained models. Although sensors are less influenced by light than 2D images, this factor can still have a non-negligible effect.

The main problem of using 3D data is the need of a correct alignment of tridimensional data between face surfaces. To help to solve this problem, in 1992, Besl and McKay (1992) introduced the Iterative Closest Point (ICP) to perform alignment of facial 3D models. The ICP algorithm can be used in two different manners: for reducing the volume difference between two point clouds, and for calculating the volume difference between two models. This technique fails for a misalignment superior to approximately 15°, leading to a non-convergence in the algorithm. Despite that, this is still one of the most used algorithms for solving the alignment problem.

One of the first works with 3D facial recognition was introduced by Gordon (1991), in which the author stated that some facial descriptors (like shape of forehead, jaw line, eye corner cavities and cheeks) remain similar in different facial expressions. Here, principal curvatures are calculated in the depth images, and are then used to detect feature points. Euclidean distance is used as the similarity metric.

A few years later, Tanaka et al. (1998) proposed a different curvature-based approach. By extracting the principal curvatures and their orientations in a facial model, some features were extracted and mapped on two unit spheres, named Extended Gaussian Images (EGI). Finally, the similarity match was performed using a Fisher spherical approximation on the obtained EGIs.

Similarly to Gordon (1991), Chua et al. (2000) defended that some rigid facial regions (nose, eye socket and forehead) deform much less in case of facial expression. These ‘rigid regions’ are found by a point signature two-by-two comparison among different facial expressions of the same person, and similarity is measured and used by a rank vote process with a training indexed table.

The use of PCA in 3D models was introduced by Hesher et al. (2002), applying it directly on depth maps. Then, Euclidean distance is used to match the resulting vectors. In two different works, PCA was used on Eigenfaces (Heseltine et al., 2004b) and in Fisherfaces (Heseltine et al., 2004a) representations of depth images. Euclidean and cosine distance measures were used for classification, respectively.

Moreno et al. (2003) introduced a different methodology, segmenting regions according to their median signs and their Gaussian curves. Regions with significant curvatures are then isolated. From here, local features are extracted (areas, distances, angles, area ratios, mean of areas, mean curvatures, variances, etc.). Different combinations of features were tested, resulting in a reduction of the feature space from 85 to 35 features (the group of features with highest performance). Finally, a classification based on Euclidean distance was performed.

Ansari et al. (2003) proposed a new acquisition method from a single frontal and a single profile scan. A generic morphable 3D model is deformed to be adapted to the input query 3D model, with a two-phase deformation. First there’s a scale adjustment and alignment of the model to the real facial surface, and then a second local deformation tries to approximate the vertices of the model, using feature points. Using the Euclidean distance between 29 different point coordinates, an identity for the query face is chosen.

A new free-form representation, the Sphere-Spin-Images (SSI), was introduced by Wang et al. (2004). Here, an SSI is associated with a point in a face surface, consisting in a 2D histogram constructed using the neighborhood surface of a point using position information (local shape). Therefore, SSIs locally describe the shape of facial surface. Using a correlation coefficient, different SSIs of specific keypoints were compared. Recognition is achieved by a SSI-comparison-based voting method for each of the SSIs in each face.

In Cook et al. (2004), to avoid the problem of having a misalignment superior to 15° when applying ICP, nose tip, nasal bridge and brow are first detected. From here, some depth and curvature-based features were extracted from scans of the 3D_RMA dataset. A PCA feature space reduction is performed, and a Gaussian Mixture Model (GMM) is used for statistical modeling of the error distribution in facial surfaces, allowing to differentiate between intra- and inter-personal comparisons of range images.

In 2012, Huang et al. (2012) proposed a hybrid system (both holistic and local feature-based), based solely on depth information. Here, extended Local Binary Patterns (eLBP) are applied on depth images, resulting in multiscale extended LBP Depth faces that contain all the 2D information of range images. The SIFT methodology is also applied, using a classification based on three similarity measurements. The main advantage of this work was not needing an alignment process, and, although it has only good results in nearly frontal pose, it appears to be robust to facial expression variations and partial occlusions.

In more recent works, Naveen and Moni (2015) tested their framework on the FRAV3D dataset, using 2D-DWT (Discrete Wavelet Transform) and 2D-DCT (Discrete Cosine Transform) for spectral representation of high resolution depth images. Feature dimension reduction is performed by PCA, obtaining the corresponding weight vectors. These weight vectors are fed to a classifier which uses Euclidean distance for classification. The score fusion of the two techniques was performed at the score-level.

Tang et al. (2015) performed landmark detection using the three main principal curvatures, obtained from the construction of asymptotic cones which describe the local geometry of the mesh model. In the three main curvature faces, the Local Normal Patterns (LNP) is applied and then classified by a Weighted sparse representation-based classifier (WSRC).

2.4.2.1 3D-based Approaches Using Low-Cost Sensors

3D-based methodologies that use low-cost scans have also been explored in the past. Min et al. (2012) used low-cost sensor depth scans from PrimeSense. Using the depth discontinuity and an empirical threshold, the head region is segmented and then subsampled. In the training phase, M faces are randomly selected to form a set of canonical faces. Using a modified ICP, the EM-ICP (faster), each face is aligned with the M canonical faces. The main problem with this approach is that the nose tip detection is manual in order to segment a region of interest. The facial region obtained is divided on nose, eye region, cheeks and the remaining parts (each region is associated with a respective weight). A feature vector is formed containing the $L2$ distances between each facial region and their corresponding areas. The decision phase is made using the Euclidean distance between feature vectors. This system uses frontal poses and does not use RGB information, although it introduces the use of a new sensor in this type of biometric systems, while also working in real-time.

Cardia Neto and Marana (2015) proposed a new feature extraction method: the 3DLBP, designed specifically for Kinect depth maps, based on the LBP (Local Binary Patterns). In this method each pixel is described by 4 LBP values. Also a variant of HOG (Histogram of Oriented Gradients) feature extractor was proposed: the Histogram of Averaged Oriented Gradients (HAOG). First, the nose tip is detected automatically and a sphere is cropped around that point. After this, to increase the quality of the depth model, symmetric filling is applied, improving the robustness of the depth map mainly in high pose variations (profile views for example). 3DLBP is applied on the cropped ROI and this region is divided in 8×8 sub-regions, obtaining 4 histograms (one for each LBP value) that are concatenated in one global feature vector. The same histogram procedure is applied in parallel with HAOG (where a single histogram per region is obtained). An SVM (Support Vector Machine) classifier is applied independently to 3DLBP and HAOG feature vectors, and a weighed score is used for the final decision. A summary of this approach can be seen in Figure 2.4.

Bondi et al. (2015) also used real-time Kinect v1 video sequences to generate high resolution models every time someone passes through the sensor. Iteratively, 3D low resolution frames are aligned with a reference frame using a Coherent Point Drift (CPD) algorithm, filtering the 3D data with a variant of the Lowess method. A combination of SIFT (Scale-Invariant Feature Transform) and spatial clustering was used to detect stable keypoints on depth data. Facial curves were used to model variations in depth between the pairs of keypoints detected. A random sample consensus (RANSAC) algorithm is used for outlier removal in the keypoint matching phase. A match between keypoints from different face models is performed using a new distance metric that takes into account the saliency of the curvatures between keypoints.

Table 2.3 summarizes the most relevant information regarding the works described in this section, focusing on a series of parameters (feature extraction, classifiers, datasets, etc.) used on these 3D unimodal approaches.

Although these approaches are a good solution to the problems faced by 2D images, they do not take fully advantage of all the information available, as RGB information is not used. The next section will focus on multimodal approaches that attempt to fuse both sources of data in a single classification.

2.4.3 Multimodal Approaches

The inclusion of two modalities has shown to be promising for real-world systems and uncontrolled environments, especially when high pose variations and low illumination environments are a possibility. The fusion of 2D and 3D scans have always improved the performance of the systems (Abate et al., 2007), when compared to the use of one unimodality.

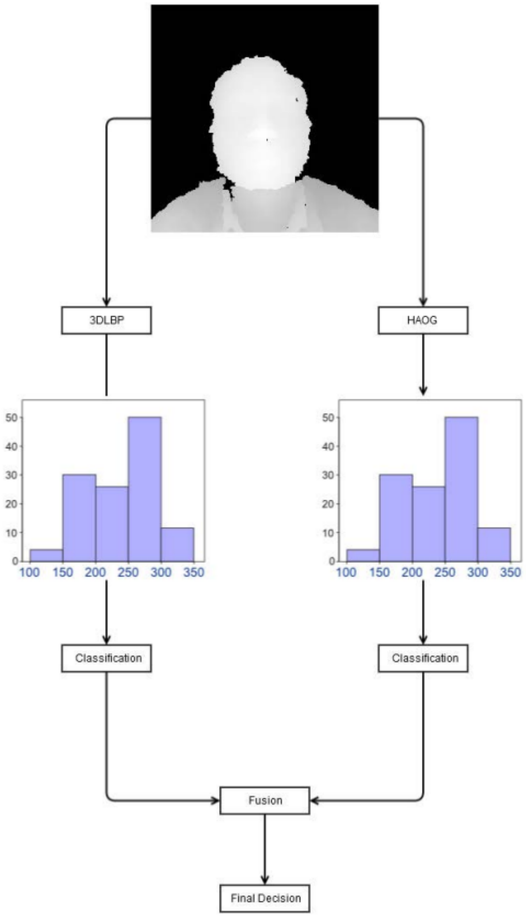


Figure 2.4: Process for Facial Recognition used by Cardia Neto and Marana (2015).

Table 2.3: Summary of the most relevant works concerning unimodal 3D face recognition

Author	Feature Extraction	Classifier	Dataset	r_1
Gordon (1991)	Distance Measures	Euclidean Distance	8 subjects (23 scans)	97.00
Tanaka et al. (1998)	Curvature features	Fisher Spherical Approximation	37 subjects	100
Hesher et al. (2002)	PCA	Euclidean Distance	37 subjects (222 scans)	100
Heseltine et al. (2004b)	PCA on Eigenfaces	Euclidean Distance	York 3D Face	87.3
Heseltine et al. (2004a)	PCA on Fisherfaces	Cosine Distance	York 3D Face	88.7
Chua et al. (2000)	Point Signature Comparison	Ranked vote	6 subjects (24 scans)	100
Moreno et al. (2003)	Geometric statistics	Euclidean Distance	GavabDB	78.00
Ansari et al. (2003)	3D Coordinates	Euclidean Distance	26 subjects (52 scans)	96.2
Wang et al. (2004)	Sphere-Spin-Images	SSI-Comparison-based Voting Method	6 subjects (31 scans)	91.68
Cook et al. (2004)	Depth and Curvature	Gaussian Mixture Model	3D_RMA	97.33
Huang et al. (2012)	eLBP + SIFT	3 Different Similarity Measurements	Bosphorus, Gavab DB, FGRC v2	97 / 95.49 / 97.6
Naveen and Moni (2015)	2D-DCT and 2D-DWT	Euclidean Distance	FRAV3D	96
Tang et al. (2015)	Principal Curvatures + LNP	WSRC	FRGC v2	93.33
Min et al. (2012)	*L2* Distances	Euclidean Distance	20 subjects	100
Cardia Neto and Marana (2015)	3D-LBP + HAOG	SVM	Eurecom	98
Bondi et al. (2015)	SIFT and Curvatures	RANSAC + Distance and Salience Metric	Florence Superface Dataset	75

Two of the first works with multimodal facial recognition, (Chang et al. (2003), Chang et al. (2005)) investigated the benefits of integrating 3D data (using a Minolta Vivid 900 sensor) with 2D images, using PCA separately on both modalities. The authors state that 2D and 3D individually get similar performances, but when combined (with a simple weighing system), a significant increase in the performance is observed using the Mahalanobis cosine distance for the decision.

Tsalakanidou also investigated the robustness of multimodal approaches in two of his works. In his first work (Tsalakanidou et al., 2003), which is very similar to Chang et al. (2003), an Eigenfaces extension to depth scans is used. The Eigenfaces are applied on both 2.5D and 2D scans and Euclidean distance are computed separately. The final score is obtained through multiplication of both individual results and assigning the query face to the smallest product template image. Here the multimodal approach has shown significance improvements over the independent 2.5D and 2D recognition. In his second work (Tsalakanidou et al., 2005), an embedded hidden Markov model (EHMM) was used to combine 2.5D and intensity images. Two EHMM classifiers trained with 2D-DCT coefficients are used for classification (one for each modality). To increase the performance of the system and to augment the training dataset samples, scans with different poses and variations are generated for each subject.

Papatheodorou and Rueckert (2004) proposed a simple 4D Euclidean Distance to measure the facial similarity (calculating also the textural differences). The results were overall promising, although with pose and facial expression variations, performance decreased.

In 2007, a new approach was proposed by Mian et al. (2007), where, first a pose correction is performed using the Hotelling transform. Using a combination between 3D Spherical Face Representation (SFR) and 2D Scale Invariant Feature Transform (SIFT), a large percentage of the candidate faces is removed (SFR-SIFT-based rejection classifier). Then the eyes-forehead and the nose regions are automatically segmented and matched using a modified ICP algorithm.

One year later the same author (Mian et al., 2008) proposed a new a 3D keypoint detection using a PCA-based method, achieving results in terms of keypoint repeatability similar to SIFT. Once a 3D point is identified, a tensor representation locally describes the keypoint. In parallel, in 2D images, a SIFT approach was implemented for keypoint detection. A feature-level fusion with vector concatenation as well as a score level fusion (with 4 different similarity criteria) assesses the best performing alternative. The higher performance was achieved with score level fusion.

Using high resolution face scans, Hiremath and Manjunatha (2013), used Radon transform on both texture and depth images in order to obtain binary maps to crop the facial region. Gabor features are extracted from both types of scans. PCA is applied to reduce dimensionality, and feature vectors are then inputted in an AdaBoost classifier that selects the most discriminant features. Finally, Nearest Neighbor scheme decides the identity of query face.

Elaiwat et al. (2015) proposed a multimodal approach using 3D textured high-resolution face models, in which Curvelet coefficients are used to represent facial geometrical features. Primarily, curvelet transform is applied on textured faces in order to identify the keypoints in the curvelet domain. Only repeatable keypoints (those appearing nearly in the same location) are saved. Using both depth and textural information, each face is decomposed into multi-scale and multi-angle decompositions. A local surface descriptor is applied around the keypoints considering all the sub-bands of the scale in which the keypoints are detected. To compensate variance to rotation, circular shift is applied to the keypoint orientation. Finally, 2D and 3D feature vectors are created and a cosine distance metric is used for facial matching. The final results are obtained using a confidence weighted sum rule.

Naveen et al. (2015), used a Local Polynomial Approximation Filter (LPA) to obtain directional faces to each modality. These faces are optimized using the Intersection of Confidence Interval Rule (ICI). For feature extraction a modified LBP (mLBP) is computed and concatenated in a histogram, to which the Discrete Fourier Transform (DFT) is applied. Finally, Euclidean distance measure is applied on a PCA reduced feature vector, using score-level fusion to obtain the final decision.

2.4.3.1 Multimodal Approaches Using Low-Cost Sensors

Similar to 3D approaches, some multimodal methods using low-cost sensor have also been proposed. In 2013, Li et al. (2013) used Kinect v1 to develop a facial recognition system invariant to pose, expression, illuminations and disguise. A query face is registered using a reference 3D model (obtained from high resolution scans). In order to compensate the missing data from pose variations, a symmetric filling step is carried out (an example of this process is shown on Figure 2.5). Although the human face is not entirely symmetrical, this approach proves to increase the recognition rate of the systems when presented with high pose variations. The RGB scans are transformed to the Discriminant Color Space and a Sparse Representation Classifier (SRC) is applied in parallel to both types of scans. Then two sets of similarity scores are obtained based on individual class reconstruction error for both depth and texture images. These two scores are normalized using the z-score technique and summed for final decision.

Goswami et al. (2013), tested a new multimodal system in Eurecom and in III-TD datasets. Here, the saliency and entropy maps for RGB is computed. In parallel, the entropy maps for depth images is also computed. To the resulting images HOG is applied, concatenating the resulting feature vectors to get the final descriptors. In this work a Random Forest Classifier is used for the identity assessment.

Recently Ajmera et al. (2014) proposed the use of Speeded-Up Robust Features-based (SURF) descriptors in Kinect scans (tested on EURECOM database and CurtinFaces dataset). Using a Graph Based Iterative Hole filling interpolation, images with variation in pose are generated using the depth model. In parallel, the RGB image is processed with an Adaptive Histogram Equalization (low contrast enhancement), a non-local means filter (for pepper noise removal) and a steerable filter. The SURF algorithm is computed, detecting keypoints that are then matched using a nearest neighbor approach. A weighed score fusion for the three methods is applied for final decision making. This methodology has the disadvantage of relying in manual face cropping and not being robust to pose variations, occlusions and illumination variations. Depth images data are not considered for feature extraction, only being used to generate images in different conditions.

Mracek et al. (2014) developed a work on low-cost sensors using Kinect v1 and SoftKinetic DS325. A feature-preserving mesh denoising algorithm is applied to the depth images to deal with noise and peak presence. All models were aligned using the Iterative Closest Point (ICP) and were converted to six representations of depth texture and curvature. Gabor and Gauss-Laguerre filters are then applied to the mesh, and an individual feature vector is obtained using z-score normalized PCA projections. A correlation metric is used for making the final decision.

Hsu et al. (2014) proposed a 3D face reconstruction using RGB-D images, performing the generation of multiple 2D faces, to increase the gallery size. Additionally, a landmark detection to perform face alignment is used. Sparse Representation was used for classification.

A multimodal approach has also been proposed by Dai et al. (2015), where a new local descriptor is used for feature extraction: Enhanced Local Mixed Derivative Pattern (ELMDP). Gabor features are extracted from RGB images. After this, ELMDP is applied independently on depth and texture images leading to two histogram representations. For the matching phase a Nearest Neighbor classifier is applied on both histograms, and the two modalities are combined with a weighed score fusion methodology.

Hayat et al. (2015) proposed a new raw depth pose estimation and automatic crop of facial region. The images are clustered based on their poses, resulting in a model for each cluster. Using Riemannian manifold, a Block-based covariance matrix is applied and an SVM classifier is used for each modality. The final decision is made by fusion of classification results for each of the image subsets. This approach shows clear improvements of multimodal approaches over their unimodal counterparts.

Krishnan and Naveen (2015) introduced a new multimodal framework. Entropy maps of depth and RGB images were obtained and a saliency map is also constructed from RGB images. HOG is

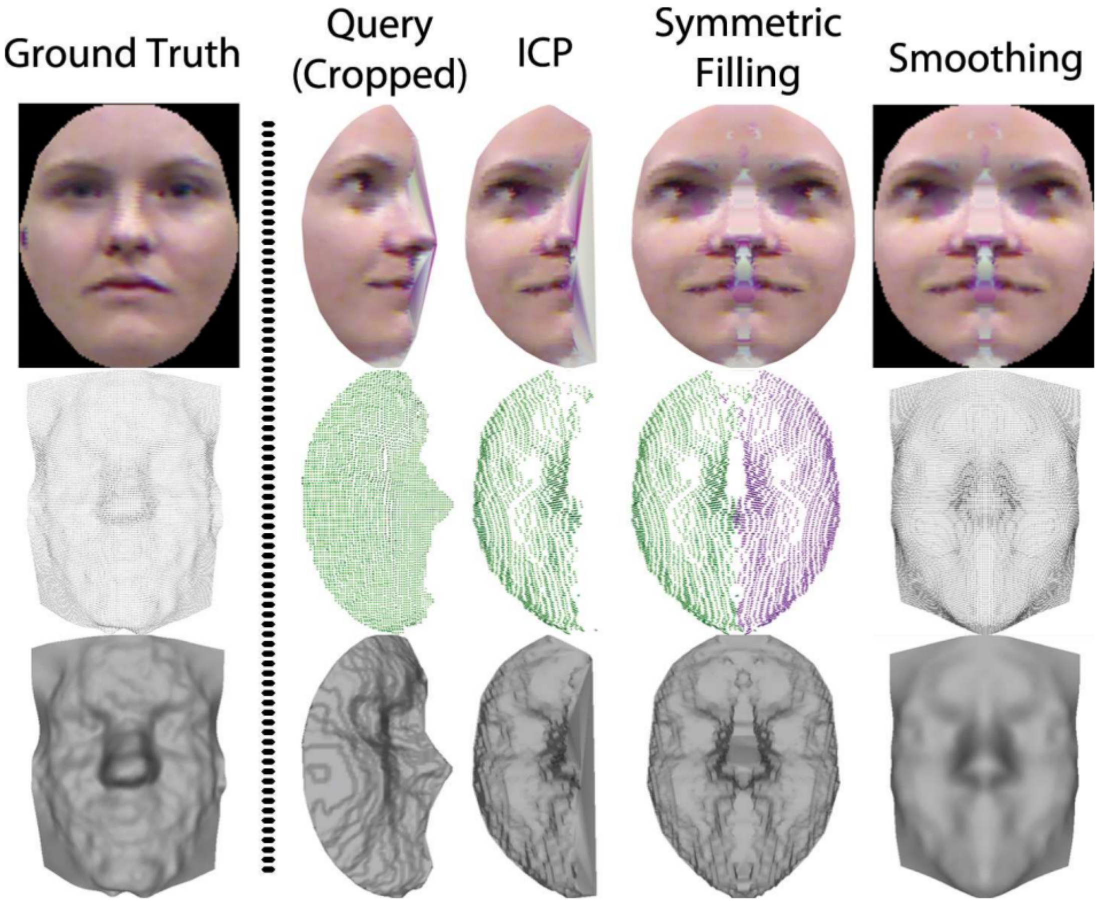


Figure 2.5: Symmetric Filling Process. Obtained from Li et al. (2013).

applied on the three resulting maps and concatenated in three resulting histograms. A Tree Bagger classifier is used for identity assessment.

Sang et al. (2015) proposed a new multimodal system, which included pose correction. ICP is used for face alignment, also allowing the alignment of RGB images. To each modality HOG is applied, and a Joint Bayesian Classifier is used for classification.

Table 2.4 summarizes the most relevant information extracted from the works described above, regarding a series of parameters (feature extraction, classifiers, datasets, etc.) used for the development of these multimodal approaches.

Table 2.4: Summary of the most relevant works concerning multimodal face recognition

Author	Feature Extraction	Classifier	Dataset	r_1
Chang et al. (2003)	PCA	Mahalanobis Cosine Distance	275 subjects	98.8
Tsalakanidou et al. (2003)	PCA	Euclidean Distance	295 subjects (3540 scans)	98.75
Tsalakanidou et al. (2005)	2D-DCT	EHMM	50 subjects	79.2
Papatheodorou and Rueckert (2004)	-	4D Euclidean Distance	62 subjects	66 - 100
Mian et al. (2007)	SIFT and SFR	modified ICP	FRGC v2	98.31 - 99.7
Mian et al. (2008)	Tensor Representation+ SIFT	4 Different Similarity Measurements	FGRC v2	96.6 - 99.9
Hiremath and Manjunatha (2013)	Gabor	Nearest Neighbor	Texas 3D + Bosphorus + CASIA 3D	99.5
Elaiwat et al. (2015)	Curvelet Coefficients	Cosine Distance	FRGC, BU-3DFE, Bosphorus	99.2 / 95.1 / 91
Naveen et al. (2015)	LPA + DFT	Euclidean Distance	FRAV3D	91.68
Li et al. (2013)	-	SRC	CurtinFaces	96.7
Goswami et al. (2013)	Saliency + Entropy HOG	Random Decision Forest	III-TD and Eurecom	80.0 / 88.0
Ajmera et al. (2014)	SURF	Nearest Neighbor	Eurecom and CurtinFaces	89.28 / 98.07
Mracek et al. (2014)	Gabor and Gauss-Laguerre	Correlation Metric	Kinect (9 subjects), Kinectic (26 subjects), FRGC v2	<89
Hsu et al. (2014)	-	SRC	Curtin Faces, Eurecom, AVL-RGBD Face	97.2 / ∼89 / ∼76
Dai et al. (2015)	ELMDP + Gabor	Nearest Neighbor	CurtinFaces	∼ 95
Hayat et al. (2015)	Riemannian manifold	SVM	BIWI, CurtinFaces, UWA	96.56 / 96.42 / 96.56
Sang et al. (2015)	HOG	Joint Bayesian	Bosphorus, BIWI, Eurecom, CurtinFaces	∼97 /∼84 / ∼93 / ∼98
Krishnan and Naveen (2015)	Saliency + Entropy + HOG	Tree Bagger	CurtinFaces, FRAV3D	∼ 65 / ∼ 70

As a preliminary study regarding this dissertation, some tests have been carried in EURECOM dataset, using an extension of a 2D framework developed by Monteiro and Cardoso (2015). This work has been published in U.Porto Journal of Engineering (Monteiro et al., 2016) and, as this framework will be one of the base-lines for this thesis development and results comparison, it will be further detailed in the next section.

2.4.3.2 A Cognitively-Motivated Recognition Method

The framework was developed for 2D facial recognition using a UBM-based (Universal Background Model) hierarchical system modeled by GMM (Monteiro and Cardoso, 2015). Recognition is done in a hierarchical way, such that global models take precedence over more detailed ones. A set of partial models is built and organized into levels, each one containing equal size non-superimposing subregions (that are ordered in an arbitrary way), I_l. The algorithm starts with trying to assess an identity using the global image, only advancing to the next level if a decision with a certain level of confidence cannot be made. Each decision made for each region in the same level is independent and only the most relevant one is kept.

The UBM approach, originally proposed for voice recognition was introduced by Reynolds et al. (2000). It can be interpreted as simple hypothesis test: given a query face image $Y$and a claimed ID, S, the two hypothesis can be defined as:

$$H_0\text{: } Y \text{ belongs to } S$$
$$H_1\text{: } Y \text{ does not belong to } S$$

being H_0 the null hypothesis and H_1 the alternative. A likelihood-ratio test can be used to achieve the optimal decision:

$$\frac{p(Y|H_0)}{p(Y|H_1)} \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{accept } H_1 \end{cases} \tag{2.1}$$

in which θ is the threshold of acceptance for a certain identity. For this decision to be made $p(Y|H_0)$ and $p(Y|H_1)$ need to be computed. H_0 will, therefore, represent a model λ_{hyp} that characterizes the hypothesized identity, whereas H_1 will describe the model of all the alternatives to the hypothesized identity, $\lambda_{\overline{hyp}}$. For an unknown input sample, the most likely identity, Id_{max}, will correspond to the identity with highest likelihood-ratio value for all possible identities.

GMMs are used to model both $\lambda_{\overline{hyp}}$ and λ_{hyp}, as they are capable of generating smooth parametric densities for these models. GMMs are trained using a variant of the original SIFT, the densely-sampled set of scale-invariant feature transform (dSIFT). Here, the original SIFT is applied on dense grids of locations at a fixed scales and orientations, allowing a decrease of detection in non-interesting points. To simplify the complexity of the training, while not hurting the performance, GMMs are simplified, by being trained using diagonal covariance matrices. Contrarily to the original SIFT, PCA reduction is applied to the 128 resulting dimensions of the keypoints, reducing the dimensionality to 32, leading to a reduced computation time in training phase, and improving the distinctiveness and robustness of the extracted feature vectors. The dSIFT is used to train the GMMs using all the keypoint descriptors obtained from all individuals (UBM), $\lambda_{\overline{hyp}}$, and specific subject data for modeling the individual-specific models, λ_{hyp}.

In UBM training, a set of extracted data from all the subjects, named the ”impostor data”, is modeled by a k-mixture GMM. Using this UBM, the ”genuine” specific model for each enrolled user can be computed by adaptation of the global UBM parameters, using specific subject data.

The UBM adaptation for each specific model is done by tuning of the parameters in a maximum *a posteriori* (MAP) sense. For each component of the UBM, a set of statistics is obtained from M individual-specific feature vectors, $X = \{\mathbf{x}_1, ..., \mathbf{x}_M\}$:

$$n_i = \sum_{m=1}^{M} p(i|\mathbf{x}_m) \tag{2.2}$$

$$E_i(\mathbf{x}) = \frac{1}{n_i} \sum_{m=1}^{M} p(i|\mathbf{x}_m)\mathbf{x}_m \tag{2.3}$$

$$E_i(\mathbf{x}\mathbf{x}^t) = \frac{1}{n_i} \sum_{m=1}^{M} p(i|\mathbf{x}_m)\mathbf{x}_m\mathbf{x}_m^t \tag{2.4}$$

where $p(i|\mathbf{x}_m)$ represents the probabilistic alignment of $\mathbf{x}_m$ into each UBM component i. This allows each component to be adapted using the diagonal of the covariance matrices.

Knowing the UBM parameters, $\{w_i, \boldsymbol{\mu_i}, \boldsymbol{\sigma_i}\}$, the adaptations for each specific user, $\{\hat{w}_i, \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i\}$ can be computed as:

$$\hat{w}_i = [\alpha_i n_i/M + (1-\alpha_i)w_i]\,\xi \tag{2.5}$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i E_i(\mathbf{x}) + (1-\alpha_i)\boldsymbol{\mu_i} \tag{2.6}$$

$$\hat{\Sigma}_i = \alpha_i E_i(\mathbf{x}\mathbf{x}^t) + (1-\alpha_i)(\boldsymbol{\sigma_i}\boldsymbol{\sigma}_i^t + \boldsymbol{\mu_i}\boldsymbol{\mu}_i^t) - \hat{\boldsymbol{\mu}}_i\hat{\boldsymbol{\mu}}_i^t \tag{2.7}$$

$$\boldsymbol{\sigma_i} = \mathrm{diag}(\Sigma_i) \tag{2.8}$$

To assure that $\sum_i w_i = 1$, a weighting parameter ξ is introduced. The α parameter is an adaptation coefficient, which depends on a relevance factor r, which measures the relative weight of the original values. Formally, it can be defined as:

$$\alpha_i = \frac{n_i}{r+n_i} \tag{2.9}$$

The relevance factor, r, has been defined as $r = 16$ (Monteiro and Cardoso, 2015). This training is repeated 14 times, for each of the subregions shown in Figure 2.6. For testing a query face, the dSIFT method is applied and PCA is applied to each of the keypoints detected. A vector of features is obtained, $X_{test} = \{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,N}\}$, where $\mathbf{x}_{t,i}$ is the i-th PCA-reduced SIFT vector extracted from a given subregion m. X_{test} is projected into the UBM and the individual specific models. The final score, $s_{t,m}$, is calculated as the mean likelihood-ratio of all keypoint descriptors obtained, $\mathbf{x}_{t,i}$, $s_{t,m} = \frac{1}{N}\sum_{i=1}^{N} s_{t,m}^{(i)}$. The assigned subject's identity is computed by maximum likelihood-ratio.

The hierarchical scheme is depicted in Figure 2.6c, in which holistic representations precede to more specific ones, only advancing to the next level if necessary. After obtaining the recognition scores for each possible identity, using the full-face image, a certainty index, c_m, is computed for measuring the certainty of the match obtained with the likelihood ratios vector, $\boldsymbol{s}_{t,m}$. This index is calculated using the following expression:

$$c_m = \boldsymbol{s}_{t^*,m} - \frac{1}{T-1}\sum_{t=1,t\neq t^*}^{T} \boldsymbol{s}_{t,m} \tag{2.10}$$

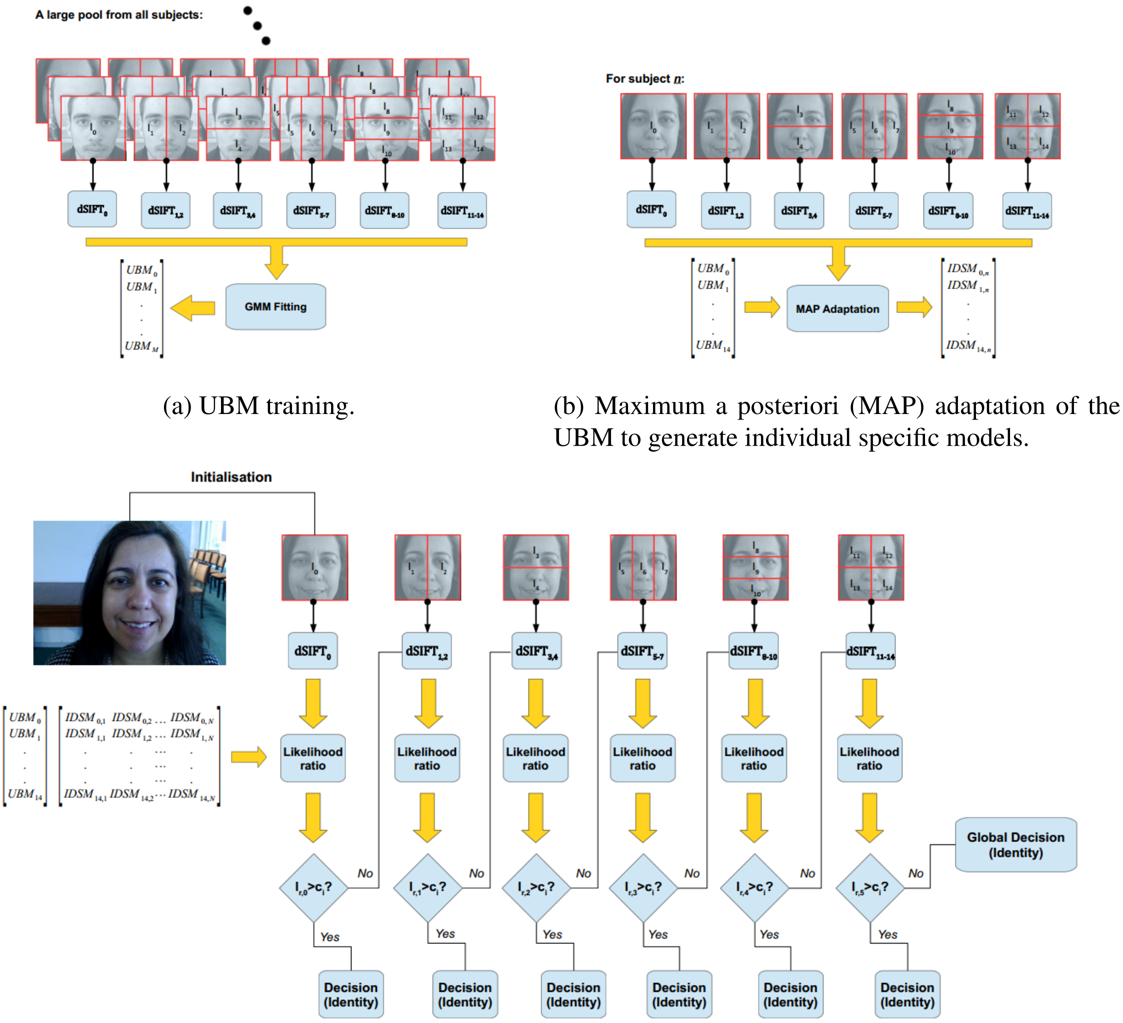
where $\boldsymbol{s}_{t,m} = \{s_{1,m}, \ldots, s_{T,m}\}$ corresponds to the scores of all the T enrolled users. With no false positive corruption in the vector $\boldsymbol{s}_{t,m}$, there should be a significant difference between the highest value (true identity), $\boldsymbol{s}_{t^*,m}$, and the average impostor score, $\frac{1}{T-1}\sum_{t=1,t\neq t^*}^{T} \boldsymbol{s}_{t,m}$.

If the c_m value is greater than a previously defined threshold, θ_l, the decision is accepted and the algorithm will not advance to the next level. Otherwise, the system advances to the next level of the hierarchy, meaning that a more detailed analysis needs to be carried out for identity assessment to be performed with a higher degree of confidence. The algorithm advances to the subregions I_{1-2} (corresponding to the second level), and the process is repeated. In levels with multiple regions only the maximum c_m among these regions is considered for the decision criterion. If, in the end, no level achieved a decision with significant confidence, the decision will correspond to the highest c_m score among all levels.

Another alternative for real-time applications, if no level is capable of making a decision, is that images can be considered doubtful and no decision made.

2.4.3.3 Extension of the Framework to 3D and EURECOM dataset testing

The framework described before was extended to deal with depth map image inputs (Monteiro et al., 2016). The extension was made by applying the pipeline in parallel to depth images. For



(a) UBM training.

(b) Maximum a posteriori (MAP) adaptation of the UBM to generate individual specific models.

(c) Testing block for input query face.

Figure 2.6: Representation of the main blocks of the UBM-based framework (Monteiro and Cardoso, 2015).

this, two feature extractors were tested: the original dSIFT approach (similar to the used in the original 2D) and LBP description (dividing the image in 4 × 4 subregions). The LBP results in a histogram for each region, that are concatenated in one single vector. This results in two hierarchical models that operate in parallel for each modality, 2D and 3D.

The fusion of modalities was performed at the score level, using the likelihood-ratio values from the two hierarchical pipelines. The final score is obtained using a simple weighted mean, averaging the two scores. The optimal values for the weights were found by grid search, under the constraint that the sum of the weights is unitary.

The framework extension was tested on the Eurecom dataset (Min et al., 2014). For training, the neutral images (with no facial expression and pose variation) were used, while the other conditions were used for testing. The profile images were not used, and the faces were manually cropped in order to only analyze the facial region.

The main results obtained with the framework are presented in Table 2.5. For each tested scenario the individual performance observed for each condition present in the EURECOM dataset is presented: light on (LO), occluded eyes (OE), occluded mouth (OM), occluded paper (OP), open mouth (OPM) and smile (S). The obtained results were compared with the baseline results in Eurecom dataset presented in Min et al. (2014).

Table 2.5: Accuracy Results Obtained in the Eurecom dataset.

Modality: Methodology	LO	OE	OM	OP	OPM	S
RGB: SIFT (Monteiro and Cardoso, 2015)	0,990	0.962	0.952	0,625	0,913	0,990
RGB: SIFT (Min et al., 2014)	0.837	0.712	0.885	0.375	0.913	0.990
RGB: LGBP (Min et al., 2014)	0.990	0.904	0.990	0.817	0.952	1.000
3D: SIFT (Monteiro and Cardoso, 2015)	0.721	0.615	0.308	0.048	0.490	0.731
3D: LBP (Monteiro and Cardoso, 2015)	0.798	0.635	0.433	0.106	0.538	0.788
3D: SIFT (Min et al., 2014)	0.049	0.020	0.020	0.029	0.029	0.249
3D: LBP (Min et al., 2014)	0.952	0.789	0.519	0.125	0.817	0.837
MM: 2D-SIFT + 3D-LBP (Monteiro and Cardoso, 2015)	1.000	0.981	0.952	0.625	0.933	0.990
MM: LGBP (Min et al., 2014)	1.000	0.894	0.981	0.846	0.981	1.000
MM: LBP (Min et al., 2014)	0.990	0.934	0.981	0.817	0.962	1.000

In RGB images, the results obtained had similar or higher performance in all tested conditions, even though a fair comparison can only be performed between the framework results and the SIFT approach presented in (Min et al., 2014). The LGBP, Local Gabor Binary Patterns, was the approach with the best results. The hierarchical model by Monteiro and Cardoso (2015) outperformed the SIFT by Min et al. (2014) in all conditions except for occlusion with paper.

In range images, the SIFT approach presented by Min et al. (2014), achieved considerably worse performance than the hierarchical framework. Concerning the results from the extension of the hierarchical framework with LBP in depth images, all the results are considerably better than the ones obtained by SIFT, corroborating the conclusions presented in Min et al. (2014). For that reason, for multimodal results assessment of the UBM, the original 2D formulation with dSIFT was used in parallel with the LBP extension for depth images.

When analyzing the multimodal results, significant improvements were obtaining when comparing to their unimodal counterparts. The occlusion with paper (OP) was an exception, where the results were not improved. When comparing with the state-of-the-art results in Min et al. (2014), the obtained performance was either in the same range or slightly better than the ones reported in literature (except again for OP).

It can be concluded that for the extended version of the Monteiro and Cardoso (2015) results follow the trend observed in previous works, where multimodal fusion with multiple sources of information leads to an improvement over all individual performances. An improvement in depth results would lead to an improved multimodal performance. For this reason, the focus of the current dissertation will be, mainly, on improving depth description and classification, to benefit the joint classification with RGB data.

As we verified, there is clearly space and need for a creation of a new dataset, that uses a different sensor than Kinect. To help in the development of the framework, a new dataset was built, with its detailed analysis being the focus of the next chapter.

Chapter 3

A New Dataset for Multimodal Face Recognition: The RealFace Dataset

After the analysis of the publicly available datasets for 3D facial recognition, and with the appearance of the Intel® RealSense$^{\text{TM}}$ depth sensors, it was clear that the creation a new dataset was needed. This contribution would be important, not only for evaluating these sensors in face recognition systems, but also to offer alternatives to available databases based on Microsoft Kinect.

Although there are no databases for facial recognition using sensors like Asus Xtion or Intel® RealSense$^{\text{TM}}$, a recent publication revealed a RGB-D Scene Understanding Benchmark Suite (Song et al., 2015) that has depth data from Kinect v1, Kinect v2, Intel Realsense R200 and Asus Xtion. When compared to the remaining sensors, the authors refer that the quality of Intel® RealSense$^{\text{TM}}$ R200 model is too low for accurate object recognition. The raw depth images appear to be worse than the other sensors. Nevertheless, they refer that the RealSense$^{\text{TM}}$ R200 sensor is the most portable sensor and also the one that consumes less power, which can be both advantages for most applications (Song et al., 2015). In Figure 3.1 we can see a comparison between these sensors, where clearly the R200 model presents the noisier depth image. Although the improved depth image (using information from multiple consecutive frames) corrects some of the noise, the depth image quality never comes closer to the remaining sensors.

It important to realize that RealSense$^{\text{TM}}$ sensors are still in the development phase, changing very quickly. The new versions of the sensors seem to be improving the overall quality of the images. During the development of this dissertation the models of the sensors have changed many times, as has the SDK.

Although the quality of R200 sensor seems to be indicative of poor performance in recognition, a face recognition acquisition protocol was built using both R200 and F200 models, as will be outlined in the following sections.

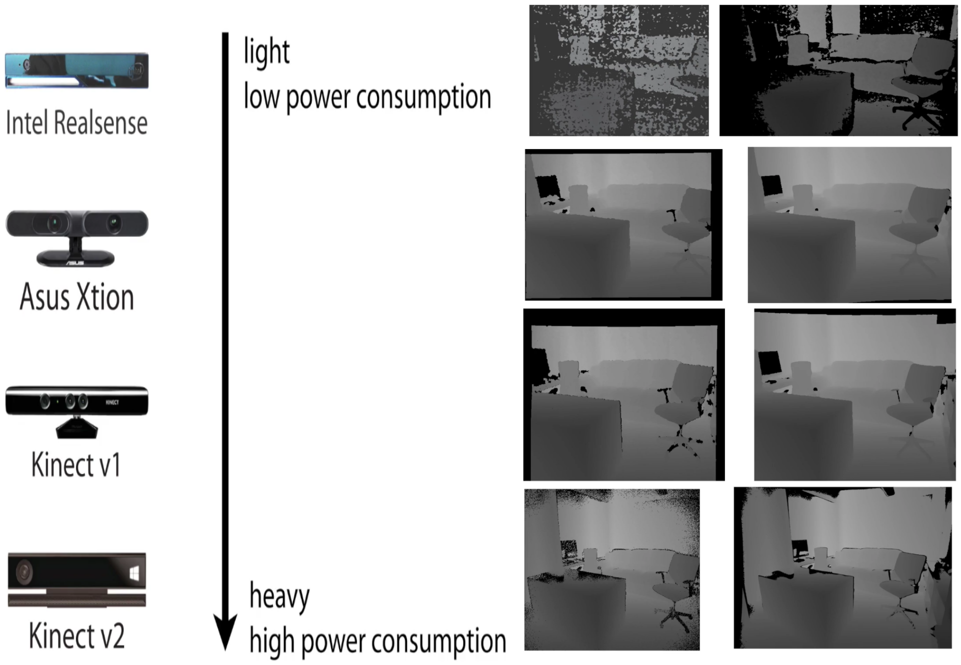


Figure 3.1: Comparison of image quality, weight and power consumption in different sensors. Also the raw depth images and the corresponding improved depth images by using multiple frames are displayed. Images acquired in Song et al. (2015).

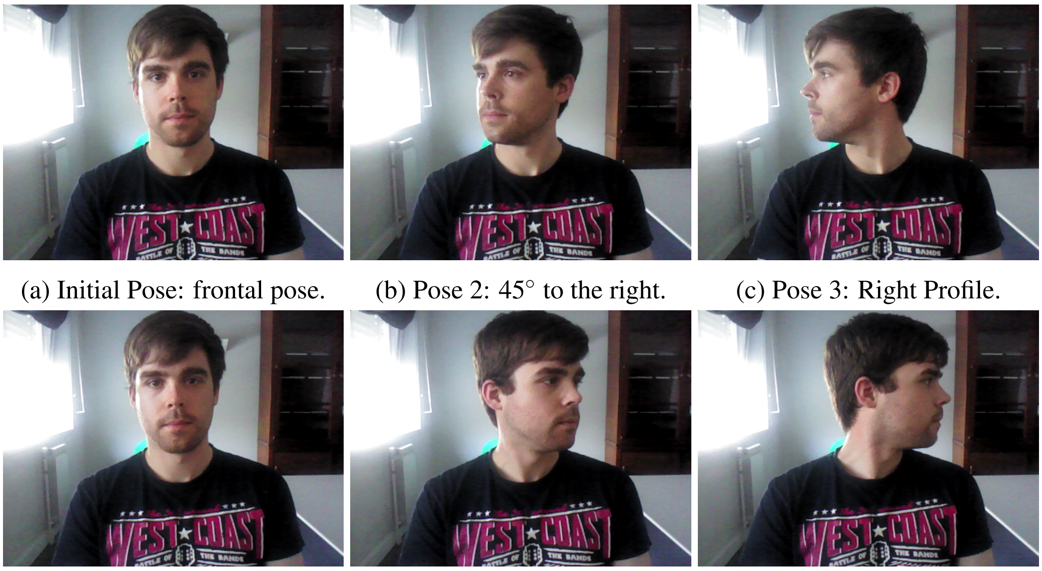
3.1 Data Acquisition Protocol

Several variations in pose, facial expression, disguises and different illumination conditions were considered to simulate, as much as possible, the conditions that characterize unconstrained acquisition environments. It was also important to ensure that the acquisition time would not be too long for the volunteers. Finally it was important to create a dataset that would differ from the datasets presented in Section 2.3.

To evaluate the performance of algorithms in different illumination conditions, images were captured in three illumination modes: natural, artificial and no illumination. Obviously, it is impossible to maintain constant natural illumination, but the focus was on testing various situations and not constant ones. Additionally, the three modes of lighting allow to understand if image quality of the depth sensor is, in fact, independent of the illumination. The darkness condition, was achieved closing the blinds of the room and using a black cloth that prevented the presence of most illumination in the room during the acquisition.

Concerning pose variations, only 5 poses were tested: frontal pose, profile poses (right and left) and $\pm 45^\circ$ pose in the lateral plane. For each pose variation, and in each illumination condition, 4 additional variations were tested concerning facial expression/disguise: neutral expression, open mouth, occlusion by a handkerchief and presence of glasses.

In each condition the user goes sequentially from frontal pose to 45° to the right and then profile. Then another acquisition is done with frontal pose and the subject repeats the movement to the left. In Figure 3.2 we can see an example of the visited poses in each condition in natural illumination with neutral facial expression.



(a) Initial Pose: frontal pose. (b) Pose 2: 45° to the right. (c) Pose 3: Right Profile.

(d) Pose 4: frontal pose. (e) Pose 5: 45° to the left. (f) Pose 6: Left Profile.

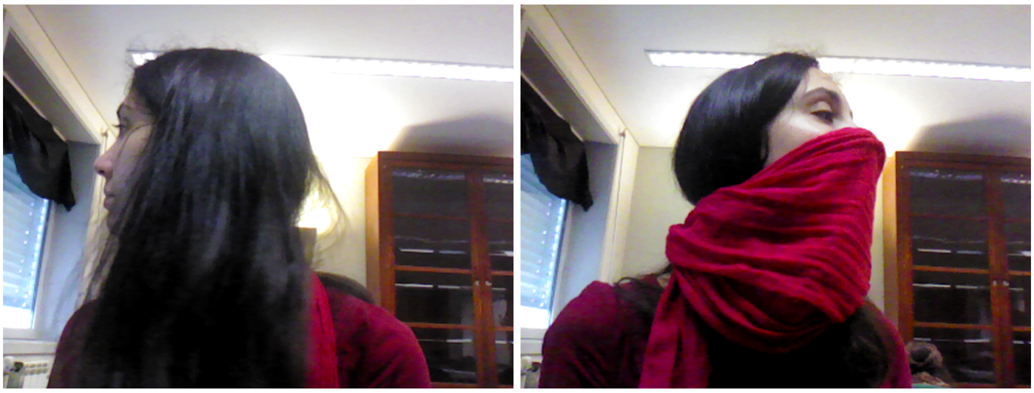
Figure 3.2: Representative scheme of the poses visited by each subject in each of the conditions during enrollment (in this case neutral expression with natural illumination is represented).

The whole process resulted in a total of 72 different conditions for each volunteer. Ideally we would have tested a lot more conditions, but it was important not making the acquisition too tedious for the subjects.

The selected conditions are thought to be enough to create a complete dataset adequate to evaluate the sensor performance, while also creating challenging conditions to face recognition systems. The occlusion by scarf, which in some cases occludes most of the subject's face, is clearly the most challenging condition, even for the human eye. Also, the presence of natural hair occlusions in profile views, mainly in female subjects, also makes it difficult to identify subjects in some images. Two examples of such conditions are presented in Figure 3.3, where, most probably, the algorithms will have difficulties to assess the identity of some volunteers.

Due to the different optimal operating ranges, the acquisition could not be done with the two sensors simultaneously, since the F200 model works well for closer ranges (optimal range of approximately 0.2 m) and the R200 model does not work optimally at close range (no depth measurements for distances smaller than 0.5 meters). Therefore, it was decided that the acquisition process would be repeated for each sensor (for logistic reasons the illumination conditions order was inverted when passing from the F200 model to the R200 model). The distance of acquisition was approximately 0.5 meters for the F200 model and 1.3 meters for R200.

To take advantage of all the modalities of Intel® RealSense™ it was decided to acquire all streams of each sensor. Therefore, for the R200 model, the two IR images provided by the two IR sensors were captured as well as the depth stream, the RGB stream and the respective Point Cloud. As for the F200 model the same streams were captured, with the difference that this sensor



(a) Profile case where we can see a natural occlusion by hair.

(b) Situation where scarf occlusion and pose variation are combined.

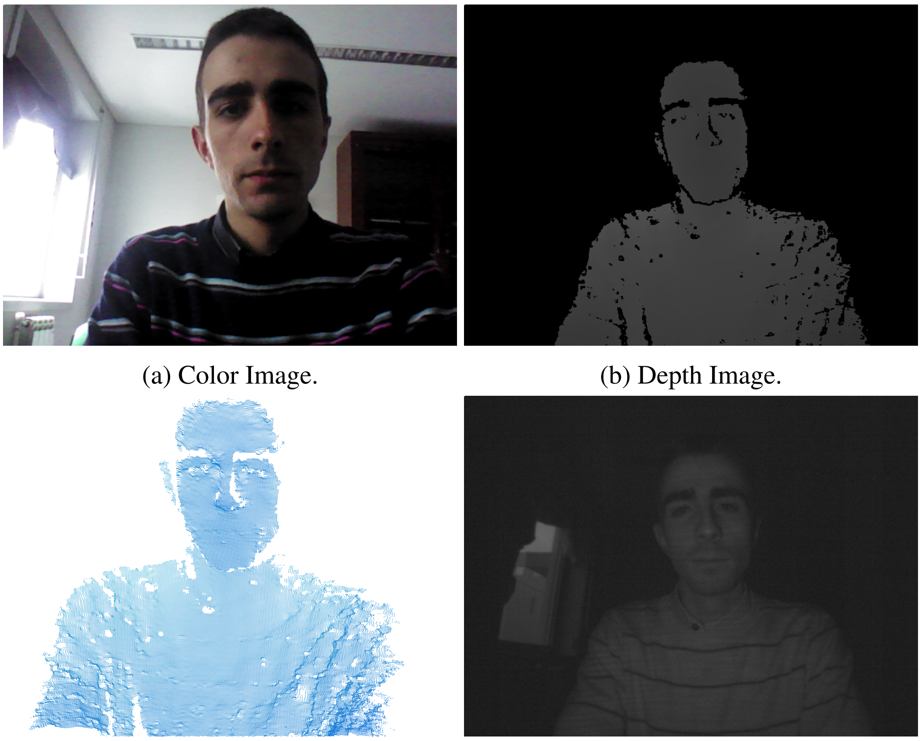
Figure 3.3: Two examples of difficult cases, where, additionally to pose variation, natural occlusion by hair in (a) and occlusion by handkerchief in (b) are also present.

only provides one IR stream. In Figures 3.4 and 3.5, examples of all the modalities captured with the F200 and R200 sensors are depicted, respectively.

Since the Intel® RealSense$^{\text{TM}}$ SDK does not provide a tool to efficiently record and save frames from the sensor, a simple program for the dataset acquisition was built using C++ scripts with the Librealsense database (Intel®, 2015). Although not being an official product by Intel® (it is maintained by developers), Librealsense appears as an alternative for the SDK, for all operating systems, allowing the users to have direct access to all the camera streams, access to calibration information, having some additional features as multi-camera simultaneous use. This allowed the creation of a C++ script to capture all the streams referred previously.

The depth and IR images are captured in 16 bits (in F200 model the depth is provided as a 8 bit image due to different range image scale), the color image is provided in three channels (RGB) and the point cloud is presented in real-world coordinates in meters, where the sensor serves as the origin for the coordinate system.

Clearly, it is noted that the R200 sensor, as referred in (Song et al., 2015), is a noisy sensor, which will probably result in a poor performance in face recognition. Additionally, the presence of the IR laser pattern in the infra-red images may lead to poor performance in IR facial recognition. One could disable the IR emitter, but this would lead to no depth image formation in indoor environments. Contrarily, the F200 model seems to have a cleaner depth image and could be very useful for face recognition, as all modalities seem to provide relevant information. Additionally, the R200 sensor has shown to be a lot slower in the acquisition process, although fast enough for real-time acquisitions. The resulting mean time of acquisition per subject was approximately 12 minutes.



(a) Color Image.

(b) Depth Image.

(c) Point Cloud.

(d) Infra-red Image.

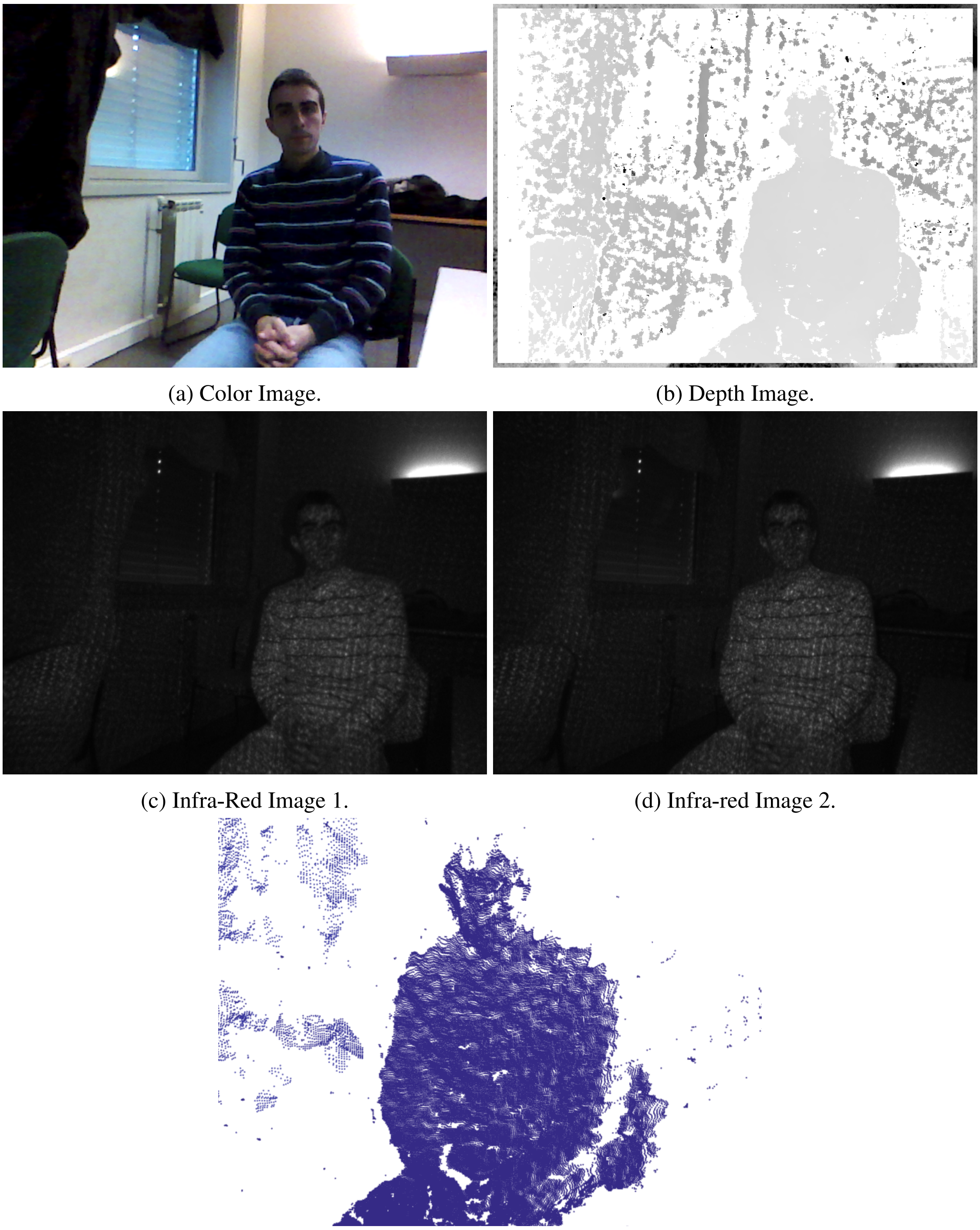
Figure 3.4: Example of the 4 types of streams acquired with the sensor F200.

3.2 Dataset Composition

The dataset includes data from 42 individuals, with ages ranging from 18 to 40 years. The genre distribution was 20 females and 22 males, while the nationalities included 41 Portuguese subjects and one Venezuelan.

Additionally to the raw streams, manually cropped images are also provided for each modality, as well as manually selected keypoints, that can be used in the assessment of keypoint detection algorithms. For profile and 45° images 5 keypoints are marked, namely the eye center, nose tip, mouth corner, chin and ear lobe. For frontal images, the 6 keypoints provided are left and right eyes centers, the nose tip, left and right mouth corners and the chin. For the F200 model images, the keypoints were selected in color images and IR images (since the coordinate system is the same between IR and depth, the keypoints can be used in both streams). As for the R200 model, the keypoints were manually annotated in both IR images and color, since the two IR cameras offer different views. In the case of darkness, no points were marked in RGB images.

For each subject, 72 different conditions were captured in each of the cameras, which resulted in 648 different scans per individual (72 in each of the modalities). Table 3.1 summarizes the number of images captured in each of the considered conditions.



(a) Color Image.

(b) Depth Image.

(c) Infra-Red Image 1.

(d) Infra-red Image 2.

(e) Point Cloud.

Figure 3.5: Example of the 5 types of streams acquired with the sensor R200.

Table 3.1: Number of images for each of the modalities captured in each condition. In the table N stands for Neutral Expression, MO for Mouth Open, S for Scarf and G for Eye Glasses.

Pose	Illumination Conditions											
	Natural				Artificial				Darkness			
	N	MO	S	G	N	MO	S	G	N	MO	S	G
-90°	1	1	1	1	1	1	1	1	1	1	1	1
-45°	1	1	1	1	1	1	1	1	1	1	1	1
-0°	2	2	2	2	2	2	2	2	2	2	2	2
+45°	1	1	1	1	1	1	1	1	1	1	1	1
+45°	1	1	1	1	1	1	1	1	1	1	1	1
Total	72											

3.3 Conclusions

This face recognition dataset offers a positive contribution for the scientific community, as an alternative to the available traditional Kinect datasets.

The created database offers difficult challenges for multimodal facial recognition systems, promoting the creation of systems that are intelligent enough to distinguish from different illumination conditions, and can adapt to the environment. This is crucial in the darkness illumination case, where RGB images provide almost none relevant information, which can be the case in night time face recognition systems, where the 3D and IR images should present more relevant information.

It is important, in the future, to increase the number of subjects of the dataset in order to increase the variability of the dataset and the number of identities. The dataset should be made available for the scientific community in the future.

This dataset, together with the ones presented in the Section 2.3, provides important foundations for algorithm performance testing. In the next chapter, these databases will be used for the development and assessment of the framework used for a new 3D/multimodal facial recognition system.

Chapter 4

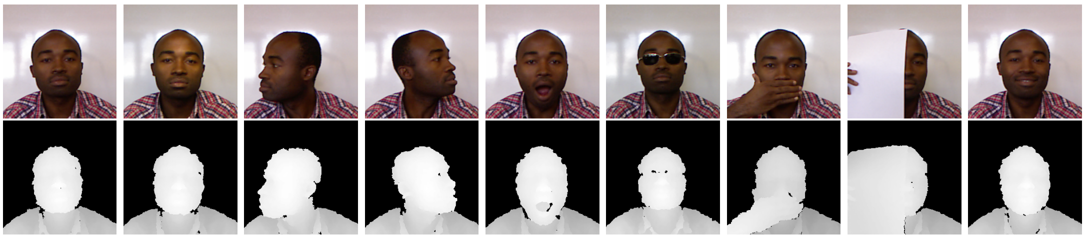
A New Multimodal Face Recognition Framework For Unconstrained Scenarios

The dataset presented in Chapter 3 and the ones discussed in Section 2.3 offer proper testing conditions for performance assessment of face recognition systems. This chapter will focus on experiments on the aforementioned datasets, where some different approaches will be tested, for the creation of a new framework capable of multimodal face recognition. Most of this framework will focus on 3D, since that was the main topic of this dissertation.

As depicted in Figure 2.1, face recognition systems follow a standard work-flow. The data acquisition phase has already been performed, in the acquisition of each tested databases, leaving only preprocessing, feature extraction and classification. In the next sections these topics will be analyzed, describing the setups done to verify the best approaches.

First, some experiments will be carried out to find which method for preprocessing depth data presents the biggest performance boost, using a fixed combination of feature extraction and classification. Then a similar setup will try to assess the best feature extraction methods, where, using no preprocessing, the classifier is not varied. Combining the knowledge gained from the previous experiments, the classifier will be chosen to verify which method of classification has a more positive impact in the overall performance of the system. Finally, an extension of the 3D unimodal system is created, including other modalities (RGB+IR), testing the proposed framework for diverse conditions.

During the chapter the evaluation of the framework will be performed using some of the metrics referred in Chapter 2, namely the global accuracy and the CMC curve, to compare recognition rates at different ranks.



(a) NE. (b) LO. (c) LP. (d) RP. (e) MO. (f) OE. (g) OM. (h) OP. (i) SM.

Figure 4.1: An example of the pairs RGB-Depth in all the conditions presented in Eurecom dataset (Min et al., 2014).

4.1 Datasets

To assess the performance of different approaches using different sensors, two main datasets were chosen to get the results: Eurecom (Min et al., 2014) Kinect Dataset and the RealFace Dataset described in the previous chapter.

The Eurecom dataset was chosen due to its variability of test conditions and also due to being one of the most cited and tested datasets in the state-of-the-art.

This dataset offers the following different conditions: neutral expression (NE), light on (LO), occluded Eyes with sunglasses (OE), occluded Mouth with hand (OM), half-occluded face with paper (OP), mouth open (MO), smile (SM), left profile (LP) and Right Profile (RP). These conditions have been captured in two sessions, having one pair RGB-Depth for each condition in each session (the Point Cloud is also provided, as well as manually annotated face keypoints). An example of each one of these conditions, for one of the subjects, is shown in Figure 4.1. In this dataset, profile poses were not taken into account.

The RealFace Dataset, discussed in the previous Chapter, was also assessed, using only the frontal poses. Pose variations introduce a new challenge for the recognition systems, and before trying to deal with these variations it is important to have a good recognition in frontal poses. The pose variations will be discussed later in this chapter. Due to the fact that the sensor model R200 is still not of enough quality to be used in face recognition applications, only the images from the F200 model were used.

The next section will focus on the analysis of the performance in response to different types of preprocessing.

4.2 Preprocessing

Preprocessing of depth images, due to its intrinsic noise, assumes a major importance in recognition systems built on this type of data.

Some of the most common methods for this purpose are the noise/spike removal and hole filling algorithms (Bennamoun et al., 2015). The noise/spike removal methods should preserve the relevant and discriminative depth information, while reducing the influence of noise. The hole filling algorithms try to compensate the holes created in the middle of depth maps, generally in points where the sensor does not measure depth due to noise or motion.

A few of the most common smoothing methods are Wiener filtering, bilateral filtering, and bilateral mesh denoising. Spike removal can be performed using a median filtering or using a simple thresholding technique. Hole filling techniques can be compensated with interpolation techniques, morphable models, symmetric filling (Bennamoun et al., 2015) and closing operations (Huynh et al., 2013).

Other approaches have been explored. Tepper and Sapiro (2012) described the use of L1 splines for enhancement of depth maps, increasing the Signal-to-Noise ratio (SNR) of the images. Le et al. (2014) proposed a more complex approach. Here a Directional Joint Bilateral Filter (DJBF) has been proposed, where different filters are used for hole and non-hole regions. Li et al. (2013) used a gridfit smoothing method (non interpolant) with a modified ridge estimator to generate the surface. Mracek et al. (2014), for depth model improvement, used a fast and effective feature preserving mesh denoising algorithm. Vijayanagar et al. (2014) implemented a real time multi-resolution anisotropic diffusion-based filtering scheme that only filters hole regions and object edges. Hsia (2015) proposed a different methodology for depth enhancement, depending on the region of the depth map. Different methodologies were applied to non-hole regions, small hole regions and big hole regions.

To determine the best preprocessing approach, some preprocessing methods were tested in Eurecom dataset: average filter, morphological close, Gaussian filtering, Median Filtering, Joint Bilateral Filtering (JBF), L1 Splines Filtering and Wiener Filtering. The baseline for this setups is the setting with no preprocessing.

L1 Splines Filtering (Tepper and Sapiro, 2012) involves minimizing a function that is a linear combination of regularizing and fitting terms, combining DCT to allow the use on grid data. The classical splines usually use L2 weighed norm, while this implementation uses L1 norm, which is shown to respond better to outliers. The DCT (which controls the complexity of the problem) is then combined with a split-Bregman iterative method to solve the numerical problem. The tested parameters used for L1 Splines Filtering were a main smoothing parameter of s = 0.02, $\lambda = 1$ and 100 split-Bregman iterations.

Although being also used in Le et al. (2014), JBF was initially proposed by Tomasi and Manduchi (1998), where this local non-iterative method is presented as a smoothing filter that preserves edges, using a non-linear combination of the local pixel values. It is controlled by two parameters: the geometric spread, σ_d, which controls the blur, and the photometric spread, σ_r, that controls the range actuation of the filter. In the experiments, σ_r was chosen to be 9 and σ_d as 0.2.

Wiener filtering in depth images was proposed by Mohammadzade and Hatzinakos (2013), and involves filtering the image with a low-pass filter. The chosen neighborhood in these experiments was 3×3.

Gaussian, median and average filtering were used with a 3×3 window. Morphological close, was used with a disk structuring element of radius 5.

To evaluate the effects in performance, 3DLBP (Huang et al., 2006) and HAOG (Galoogahi and Sim, 2012) were used as feature extractors, using an SVM (with RBF - Radial Basis Function - kernel) as a classifier. These two descriptors have shown to have good accuracy for recognition using depth maps (Cardia Neto and Marana, 2015), and will be further discussed later in this chapter. The SVM was optimized using grid search, with C and γ being optimized in the range of $[2^{-15};2^{15}]$ and $[2^{-15};2^{5}]$, respectively.

To have a fair comparison of all conditions, neutral images from both sessions were used as training, while the remaining ones were used as test, which means 2 images/subject were used for training and 12 images/subject served for testing. The images were manually cropped to include the face region, without any type of alignment.

Table 4.1: Accuracy Results (%) under different preprocessing approaches in the Eurecom dataset.

Preprocessing Method	Feature Extractor	LO	OE	OM	OP	MO	S	Global
No Preprocessing	3DLBP	35.6	7.7	8.7	**18.3**	0.3	19.4	18.6
	HAOG	51.9	36.5	17.3	8.7	28.9	40.4	28.4
Average	3DLBP	46.2	29.8	15.4	6.7	22.1	24	23.7
	HAOG	52.9	26.9	14.4	8.7	26.9	38.5	26.1
Morphological Closing	3DLBP	40.4	37.5	5.8	10.6	20.2	23.1	22.0
	HAOG	**58.7**	**51.0**	**20.2**	9.6	**38.5**	42.3	**35.6**
Gaussian	3DLBP	42.3	30.8	14.4	8.7	20.2	21.2	22.6
	HAOG	50.0	31.7	11.5	12.5	27.9	35.6	26.9
JBF	3DLBP	37.5	35.6	7.7	8.7	18.3	25.0	21.8
	HAOG	51.9	36.5	17.3	8.7	28.9	40.4	28.4
L1 Splines	3DLBP	37.5	35.6	7.7	8.7	18.3	25	21.8
	HAOG	51.9	36.5	17.3	8.7	28.9	40.4	28.4
Median	3DLBP	42.3	35.6	9.6	10.6	24.0	23.1	23.9
	HAOG	50.0	29.8	16.4	10.6	31.7	43.3	29.0
Wiener	3DLBP	48.1	33.7	12.5	3.7	22.1	24.0	24.4
	HAOG	55.8	35.6	17.3	12.5	29.8	**44.2**	30.6

The accuracy results in the Eurecom dataset are shown in Table 4.1, where are displayed the best results obtained for each subset using each type of preprocessing. The preliminary results were clearly not promising, since the maximum global performance achieved was 35.6% using HAOG and with Morphological Closing, followed by a 30.6% performance by Wiener Filtering. This clearly is not good enough for a face recognition system. Morphological closing offers a big boost in performance, due to the clear intrinsic hole presence in the depth maps.

These results led to the possibility that non-normalization in size and non-centering of the depth image could possibly result in a break in performance using these descriptors.

Therefore, additional tests were performed: using the keypoints provided by the Eurecom Dataset, the depth image was cropped and aligned, being also resized to 96 ×96. Additionally, this was also done using the Point Cloud provided by the dataset (instead of the depth map), applying the Morphological Closing, Symmetric Filling (Li et al., 2013) and Gridfit smoothing (Li et al., 2013). The use of Point Clouds in this case was tested because the depth images were provided in 8 bit images, and therefore the Point Clouds should present a more detailed representation of the face. The alignment of the face in the Point Clouds was performed using the provided keypoints and maintaining only the Cloud points in the neighborhood of the nose. The Point Cloud is then converted to a depth image, for feature extraction.

Described in Li et al. (2013), Symmetric Filling is a process which demands an almost perfect alignment and a previous pose correction method. The correctly aligned face Point Cloud is centered in the position (0, 0, 0), and a symmetric Point Cloud is created where the x coordinates are replaced by their negative values. Trying to fill only the missing data, namely the holes, the Euclidean distance to the closest point of the original Point Cloud is calculated. All the points with a distance smaller than a threshold δ are removed.

Also in Li et al. (2013) Gridfit algorithm is used to smooth the resulting Point Cloud and to convert it to a depth image. Gridfit is a non-interpolant algorithm that tries to find the approximate surface that fits the supplied data as closely as possible. It deals well with noise and replicated data.

The symmetric filling process demands a correct alignment and normalization, being that the reason that why it was not tested before under the same conditions. With the provided keypoints the Point Cloud was aligned, allowing the use of this algorithm for the assessments presented below.

Table 4.2 summarizes the accuracy results for these preprocessing methodologies, where are displayed the best results obtained for each subset using each type of preprocessing. The alignment of the face and size normalization, led to a boost in performance of approximately 20 % in global accuracy. Since the used descriptors are based on local features, this enhancement is justified. This increase of performance is, however, not observed in some occluded scenarios. Despite that, performance is obviously upgraded in almost symmetric situation, namely in the Light On, Mouth Open, Smile and Occluded Eyes. Additionally, it is important to notice that the use of Point Cloud values instead of depth map 8 bit images significantly improves the results.

Wiener filtering and Symmetric Fill are the processing strategies that achieved better results. Gridfit, although when applied by itself does not cause a significant improvement in performance, when combined with Symmetric Filling and Wiener Filtering leads to good performances for HAOG feature extractor, leading to accuracies bigger than 90% in LO, SM and OE. Despite that, the global results are still worse than with Winner filtering and no preprocessing with Point Cloud representation.

Due to the fact that Symmetric Filling demands an almost perfect pose correction and alignment, and that it can result in unrealistic Point Clouds in the presence of occlusions, Wiener Filtering seems to be the best approach. It presents a fair performance boost, while not being dependent on pose correction. It is important to notice that the Occlusion with paper is definitely the most difficult circumstance in this dataset.

To confirm these conclusions, some additional tests were done in the RealFace dataset, using solely frontal poses. Therefore, 24 images per subject were used. The manually cropped images provided in the dataset were used and resized to 96×96. Contrarily to the Eurecom dataset, the images were not centered due to the fact that some images had no manually annotated keypoints in depth. This, though, will allow us to conclude the performance of the algorithms in non-aligned

Table 4.2: Accuracy Results (%) for different preprocessing approaches in the Eurecom dataset, with normalization for a size of 96 × 96 and alignment using keypoints provided in the dataset.

Preprocessing Method	Feature Extractor	LO	OE	OM	OP	MO	S	Global
No Preprocessing	3DLBP	78.9	55.8	21.2	12.5	46.2	84.6	49.0
	HAOG	77.9	51.0	**23.1**	**20.2**	27.9	58.7	41.4
Point Cloud No Preprocessing	3DLBP	87.5	83.7	19.2	3.85	83.7	95.2	62.2
	HAOG	85.6	79.9	13.5	4.8	72.1	96.2	57.5
Point Cloud Closing	3DLBP	86.5	80.8	10.6	9.6	38.5	82.7	50.3
	HAOG	80.8	65.4	17.3	6.7	31.7	79.8	45.8
Point Cloud Wiener	3DLBP	86.54	88.46	17.31	1.92	**85.58**	96.15	**62.7**
	HAOG	84.6	84.6	13.5	2.88	67.3	**99.0**	58.7
Point Cloud Symmetric Filling	3DLBP	85.6	78.9	10.6	4.8	47.1	91.4	52.4
	HAOG	87.5	73.1	20.2	8.7	48.1	81.7	50.5
Point Cloud Gridfit	3DLBP	76.0	60.6	7.7	8.7	35.58	78.9	43.3
	HAOG	81.7	65.4	16.4	6.7	31.7	79.8	46.0
Point Cloud Symmetric Filling + Gridfit	3DLBP	69.2	63.6	9.6	5.8	51.0	79.8	44.6
	HAOG	**93.3**	**94.2**	19.2	6.7	59.6	93.3	59.6
Point Cloud Wiener + Gridfit	3DLBP	69.2	68.3	13.5	4.8	39.4	81.7	44.1
	HAOG	92.3	92.3	**23.1**	5.8	55.8	94.2	59.4

conditions. In this case the Point Cloud was not used since the 16 bit depth images already give a good representation of the depth values.

To maintain the conditions tested in the Eurecom database, two images with neutral expression, using artificial illumination, were used as training. Despite that, the images are not perfectly aligned in the image as were those used in the Eurecom database. Neutral images in artificial illumination were used as training because it was the only illumination condition that was controlled in the acquisition. For testing without pose variations, 11 different conditions were tested for each subject: Natural illumination with neutral expression (NN), Natural illumination with Mouth Open (NMO), Natural illumination with scarf occlusion (NS), Natural illumination with eyeglasses (NG), Artificial Illumination with Mouth Open (AMO), Artificial Illumination with Scarf Occlusion (AS), Artificial Illumination with eyeglasses (AG), Darkness with Neutral Expression (DN), Darkness with Mouth Open (DMO), Darkness with Scarf Occlusion (DS) and finally Darkness with eyeglasses (DG). Morphological Closing and Wiener filtering were tested with HAOG and 3DLBP using an SVM with RBF kernel as a classifier, with grid search for optimization using the ranges referred before. Table 4.3 summarizes the results in these conditions.

Table 4.3: Accuracy Results (%) for different preprocessing approaches in the RealFace Dataset, with manual cropping a normalization to size of 96 × 96

Preprocessing Method	Feature Extractor	NN	NMO	NS	NG	AMO	AS	AG	DN	DMO	DS	DG	Global
No Preprocessing	3DLBP	50.0	45.2	**29.8**	64.3	75.0	25.0	59.5	73.8	57.1	23.8	44.1	49.78
	HAOG	64.3	46.4	23.8	52.4	67.9	27.4	54.8	73.8	50.0	26.2	40.5	47.94
Closing	3DLBP	50.0	36.9	20.2	**73**.8	77.38	22.6	67.1	69.1	**59.5**	11.9	**57.1**	49.68
	HAOG	**69.1**	**52.4**	27.4	72.6	**82.1**	23.8	**72.6**	**83.3**	57.1	20.2	54.8	**55.95**
Wiener	3DLBP	48.8	44.1	**29.8**	63.1	75.0	26.2	61.9	75.0	57.1	25.0	44.1	50.00
	HAOG	59.5	50.0	23.8	53.6	70.2	**29.8**	54.8	75.0	51.2	**27.4**	41.7	48.81

Both preprocessing methods seem to have good influence in the accuracy, although Morphological Closing led to a improvement of approximately 8%, using the HAOG feature extractor, being the only method that significantly improves the results in global performance, as 3DLBP seems to keep it constant, although suffering a small improvement with Wiener filtering.

Wiener filtering still has a small improvement in both extractors, and seems to be a good fit in this dataset, despite the good performance of morphological closing with HAOG.

After this analysis of some preprocessing methods, an even more extensive analysis of some feature extraction methods is necessary, to evaluate which approaches achieve better representation and results for depth images.

4.3 Feature Extraction

Feature Extraction plays a major role in biometric systems. A good extractor should be able to get discriminant features that can be used to classify the identity of each subject. As seen on Chapter 2, different approaches can be used for feature extraction. To find which presents better performance, different approaches were tested in both datasets.

To evaluate solely the effect of feature extraction, no preprocessing was applied on depth images and, similarly, a SVM was used as classifier, using a RBF kernel, which was optimized using grid search with C and γ being optimized in the ranges used in the previous section.

In Eurecom, the Point Clouds were used to align the image and crop the face region, similarly to the process described in Section 4.2. As for RealFace Dataset, the manually cropped images were used. Both dataset images were resized for 96×96.

Despite the high number of tested feature extractors, only the 6 with best performance will be presented. Many methods, similar to the ones presented in Chapter 2, were tested, but the ones with better performance were: HOG, HAOG, 3DLBP, PHOW (Pyramid Histogram Of visual Words), FHOG (Felzenszwalb's HOG) and LDP (Local Derivative Patterns). Before presenting the results, a brief description of each of these methods is in order.

Originally being described for Human Detection (Dalal and Triggs, 2005), HOG is a descriptor that has been used in many applications, such as face detection (Cerna et al., 2013) and face recognition (Albiol et al., 2008). This approach divides the image in small regions, and for each of these regions, a histogram of the gradient orientations is computed. The bins correspond to each of the orientations and the intensity of the bins is increased by its magnitude and not its occurrence.

In Galoogahi and Sim (2012), a variation of HOG has been presented, which has been used in depth images by Cardia Neto and Marana (2015), named Histogram of Averaged Oriented Gradients (HAOG). This variation uses the same principle as HOG, but it averages the gradients and the orientation of the gradients.

Another variation of HOG, the Felzenszwalb's HOG (FHOG) has been described in Felzenszwalb et al. (2010) for object detection. Here a feature pyramid is calculated for a finite number of scales, using repeated smoothing and sub-sampling. A parameter λ is used for defining the number of levels in each octave.

Presented in Huang et al. (2006), 3DLBP was described as a variation of traditional LBP, for depth images. The authors state that, statistically, 93 % of the depth differences are smaller than

7. For each pixel, a neighborhood of 8 is presented, and, for each neighbor pixel, the difference between that neighbor and the central pixel is calculated. If any of this differences is greater than 7 or smaller than -7, they are set to 7 and -7, respectively. From these values, 4 codes are created. The first is equal to the traditional LBP, where a 8 bit binary number is created, in which subtractions equal or greater than 0 being set to 1, or to 0 if the opposite happens. The resultant 8-bit binary number is then converted to a decimal number, getting the first LBP code. The difference values in each neighbor pixel is then converted into a 3 bit binary number. The bits are concatenated for each of the neighbors, resulting in 3 8-bit numbers that are converted to decimal to form the remaining 3 LBP codes. An example of this method is presented in Figure 4.2, where a case in which traditional LBP would have poor description of the depth image is improved by 3DLBP. Using this method, 4 LBP values are calculated to each pixel, generating 4 different LBP images. For each image a 14 bin histogram is obtained for each 8 × 8 regions. The histograms are then concatenated for the final descriptor.

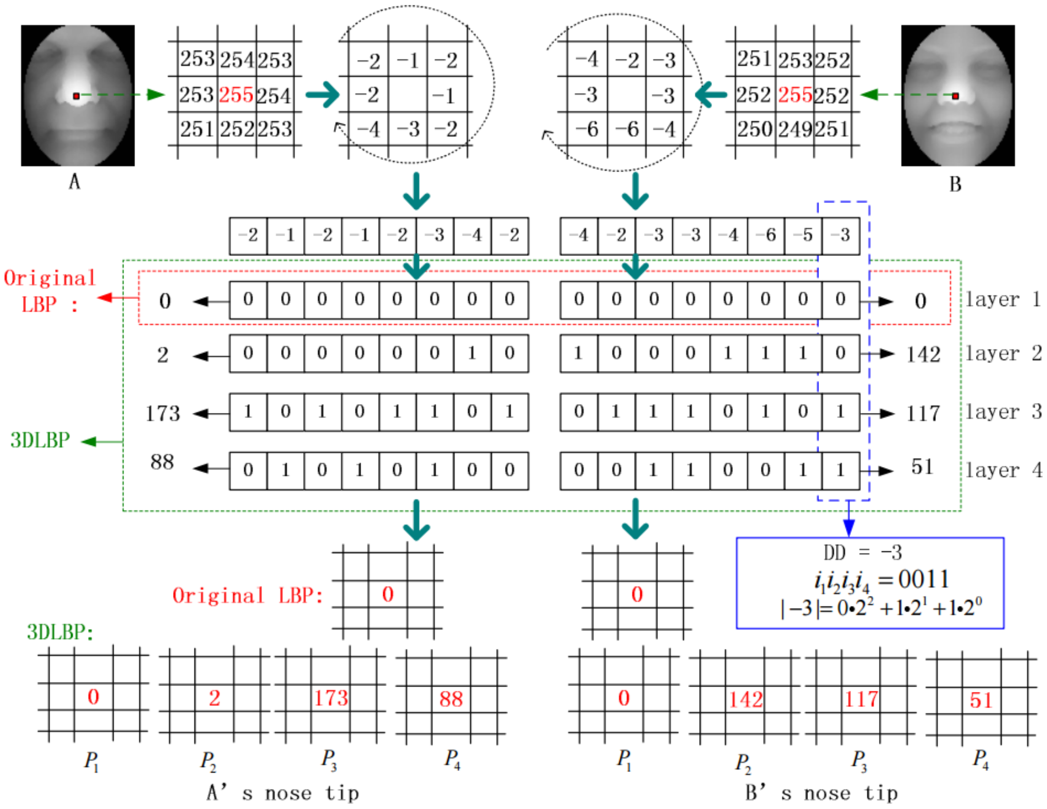


Figure 4.2: Comparison between the traditional LBP and 3DLBP. The image was obtained from Huang et al. (2006).

Another variation of LBP has been described for face recognition: the LDP - Local Derivative Patterns (Zhang et al., 2010). Since the traditional LBP is incapable of describing more detailed information than the first order derivative among local neighbors, the LDP tries to describe the $(n-1)^{th}$ order derivative in various directions, namely 0°, 45°, 90° and 135°. For each direction a binary coding function is defined for encoding the co-occurrence of two derivative directions at different neighboring pixels. This encoding method results in a 32-bit binary sequence by concatenating the 8-bit binary numbers resultant from the 4 directions. The method is extended to the nth-derivative. The final descriptor consists in a histogram of all the generated codes.

Finally, PHOW (Bosch et al., 2007) consists in a variation of dense-SIFT, which has been briefly described in Section 2.4.3.2. Here, dense-SIFT is extended at multiple scales. SIFT is applied on dense grids at multiple scales and orientations for extraction of the interesting keypoints. SIFT keypoints represent local extrema on different DoG (Difference of Gaussians) spaces. From these extrema, candidates with low contrast and edge responses are eliminated, remaining only dominant orientation keypoints.

All these feature extractors were applied on both datasets, for performance assessment. HOG and HAOG were applied in 12 × 12 blocks of the image, resulting in 64 histograms of 7 bins. FHOG was computed for 9 orientation bins, 8 for spatial bin size. Additionally, the value at which to clip histogram bins was set to 0.2. 3DLBP was also applied on 12 × 12 regions, with 14 bins in each histogram. LDP was used with order 2 (it had better performance than other orders), being also calculated for 12 × 12 blocks. Finally, PHOW was used at scales 4, 6, 8 and 10 and the step of the grid was set to 5. After the keypoints are extracted, a k-means clustering is applied to the training keypoints, using 300 clusters. After that a Vector of Linearly Aggregated Descriptor, VLAD (Jégou et al., 2010), is applied and used as a final descriptor. The descriptors are then classified using an SVM with RBF kernel, as previously described.

A summary of the obtained results is shown on Tables 4.4 and 4.5. We can see that PHOW and FHOG are the two methods that present better performance. In the Eurecom dataset, PHOW outperforms all the other approaches in 5 of 6 subsets, getting, by far, the best results in occlusion conditions. FHOG results show that this variant of HOG has potential for recognition in depth images, outperforming HOG and HAOG. These two algorithms also are the two best methods performing in the RealFace dataset, presenting the top results in all subsets. Similarly to Eurecom, the occlusion with scarf (the most difficult condition) shows that PHOW consistently presents the best results in the presence of occlusions.

3DLBP also showed good results in Eurecom, although not outperforming other approaches in RealSense^{TM} dataset. This may happen due to the fact that these images are not aligned leading to a decrease in performance (which was already discussed in previous section).

Table 4.4: Accuracy Results (%) using different feature extractors in the Eurecom dataset, with Point Cloud representation and normalization for a size of 96 × 96 and aligned using keypoints provided in the dataset.

Feature Extractor	LO	OE	OM	OP	MO	S	Global
HOG	85.58	76.92	13.46	4.81	72.12	96.15	58.17
HAOG	83.65	82.69	12.50	3.85	83.70	97.12	57.86
FHOG	88.46	86.54	14.42	3.85	79.81	**100**	62.18
3DLBP	87.50	83.70	19.20	3.85	83.7	95.2	62.18
LDP	86.54	78.85	15.38	4.81	71.15	91.35	58.01
PHOW	**96.15**	**92.31**	**58.65**	**18.27**	**90.38**	99.04	**75.80**

Table 4.5: Accuracy Results (%) for different feature extractors in the RealSense^{TM} Dataset, with manual cropping a normalized to a size of 96 × 96.

Feature Extractor	NN	NMO	NS	NG	AMO	AS	AG	DN	DMO	DS	DG	Global
HOG	57.14	48.81	35.71	70.24	75.00	29.76	54.76	79.76	54.76	25.00	53.57	53.14
HAOG	64.29	46.43	23.81	53.38	67.86	27.38	54.76	73.81	50.00	26.19	40.48	47.94
FHOG	64.29	46.43	34.52	**82.14**	79.76	32.14	**73.81**	86.90	66.67	29.76	**64.29**	60.06
3DLBP	50.00	45.24	29.76	64.29	75.00	25.00	59.52	73.81	57.14	23.81	44.05	49.78
LDP	53.57	40.48	28.57	72.62	77.38	22.62	64.29	70.24	63.10	15.48	48.81	50.65
PHOW	**75.00**	**66.67**	**48.81**	66.67	**84.52**	**41.67**	70.24	**89.29**	**72.62**	**36.9**	51.19	**63.96**

Although PHOW and FHOG were not designed for depth images, they still can extract relevant information in this type of modality. These two methods and 3DLBP, were then selected for the assessment of different classification methods, which will be discussed in the next section.

4.4 Classification

Different classifiers can be used in a supervised learning problem like face recognition. For evaluating the robustness of different classifiers, similar conditions to the used before were considered with no preprocessing, and only varying the classifier with FHOG, PHOW and 3DLBP.

Tests were performed for SVM with RBF kernel, Linear-SVM, Logistic Regression Classifier, K-Nearest Neighbor Classifier (in this case with k = 1), Naive-Bayes Classifier, GMM-UBM (Monteiro and Cardoso, 2015), Sparse Representation Classifier (Wright et al., 2009), Weighed-Sparse-Representation Classifier (Lu et al., 2013), and Random-Forest Classifier. The three classifiers with best results were Linear-SVM, RBF-SVM and Logistic Regression, whose results are shown in Tables 4.6 and 4.7.

Table 4.6: Accuracy Results (%) under different classifiers using FHOG, PHOW and 3DLBP in the Eurecom dataset, with normalization for a size of 96 × 96 and centering using keypoints provided in the dataset.

Classifier	Feature Extractor	LO	OE	OM	OP	MO	S	Global
SVM-RBF	PHOW	96.15	92.31	58.65	18.27	90.38	99.04	**75.80**
	FHOG	88.46	86.54	14.42	3.85	79.81	100.00	62.18
	3DLBP	87.50	83.70	19.20	3.85	83.70	95.20	62.18
Linear-SVM	PHOW	97.12	89.42	49.04	19.23	83.65	98.08	72.76
	FHOG	88.46	86.54	13.46	3.85	79.81	100	62.02
	3DLBP	88.46	88.46	12.5	2.88	83.65	97.12	62.18
Logistic Regression	PHOW	98.08	89.42	52.88	26.92	88.46	99.04	**75.80**
	FHOG	89.42	90.38	19.23	5.77	87.50	100	**65.38**
	3DLBP	90.38	93.27	15.38	4.81	89.48	100	**65.54**

With only two samples/subject for training, logistic regression classifier outperforms the remaining alternatives. Therefore, this classifier has been the chosen for the final framework.

PHOW has a performance significantly superior to the other feature extractors, but it is important to study if the fusion of these three features can increase further the recognition rates.

Table 4.7: Accuracy Results (%) under different classifiers, using FHOG, PHOW and 3DLBP in the RealSense^{TM} Dataset, with manual cropping a resize size of 96 × 96.

Classifier	Feature Extractor	NN	NMO	NS	NG	AMO	AS	AG	DN	DMO	DS	DG	Global
SVM-RBF	PHOW	75.00	66.67	48.81	66.67	84.52	41.67	70.24	89.29	72.62	36.90	51.19	63.96
	FHOG	64.29	46.43	34.52	82.14	79.76	32.14	73.81	86.9	66.67	29.76	64.29	60.06
	3DLBP	50.00	45.24	29.76	64.29	75.00	25.00	59.52	73.81	57.14	23.81	44.05	49.78
Linear-SVM	PHOW	61.9	41.67	27.38	77.38	77.38	30.95	72.62	85.71	60.71	29.76	61.90	61.58
	FHOG	61.90	41.67	27.38	77.38	77.38	30.95	72.62	85.71	60.71	29.76	61.90	57.03
	3DLBP	50.00	42.86	28.57	65.48	73.81	23.81	58.33	72.62	57.14	21.43	45.24	49.03
Logistic Regression	PHOW	77.38	73.81	48.81	69.05	88.10	42.86	72.62	92.86	75.00	39.29	57.14	**66.99**
	FHOG	67.86	47.62	33.33	78.57	78.57	32.14	76.19	91.67	65.48	30.95	70.24	**61.15**
	3DLBP	54.76	45.24	26.19	66.67	76.19	19.05	63.10	83.33	61.90	20.24	54.76	**51.95**

Therefore, the three built logistic regression output probabilities were combined with different weights, yielding a final global probability, as described below:

$$Global_{prob} = w_1 \times prob_{3DLBP} + w_2 \times prob_{FHOG} + w_3 \times prob_{PHOW} \tag{4.1}$$

Here we force the sum of the weights to be unitary. Since PHOW always outperforms the other methods, a minimum weight of 0.4 was considered, and Wiener filtering was used for preprocessing. The final weights were chosen according to the global accuracy. In both datasets 3DLBP did not contribute positively for the highest classification, leading the final descriptor to a simple combination of PHOW and FHOG. A representative scheme of this approach is shown in Figure 4.3.

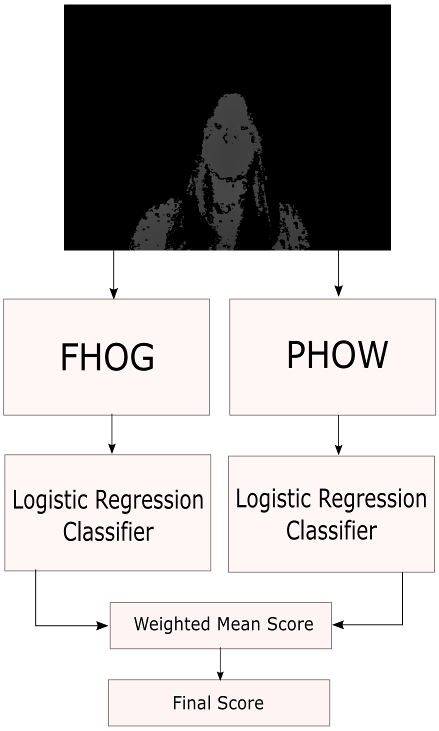

FHOG
PHOW
Logistic Regression Classifier
Logistic Regression Classifier
Weighted Mean Score
Final Score

Figure 4.3: Scheme that resumes the pipeline of the proposed framework for 3D face recognition.

The CMC curves for the two datasets are displayed in Figures 4.4 and 4.5, respectively. Although in Eurecom the fusion combined with preprocessing did not significantly increase the results, in RealFace dataset, on the other hand, fusion improved the results despite high dependence of fusion performance on the PHOW descriptor alone.

Knowing that both Kinect and RealSense^{TM} sensors have different modalities, the next intuitive step is to use those different data sources provided by these sensors, to increase the robustness of the face recognition framework. In the next section this framework will be extended to different modalities.

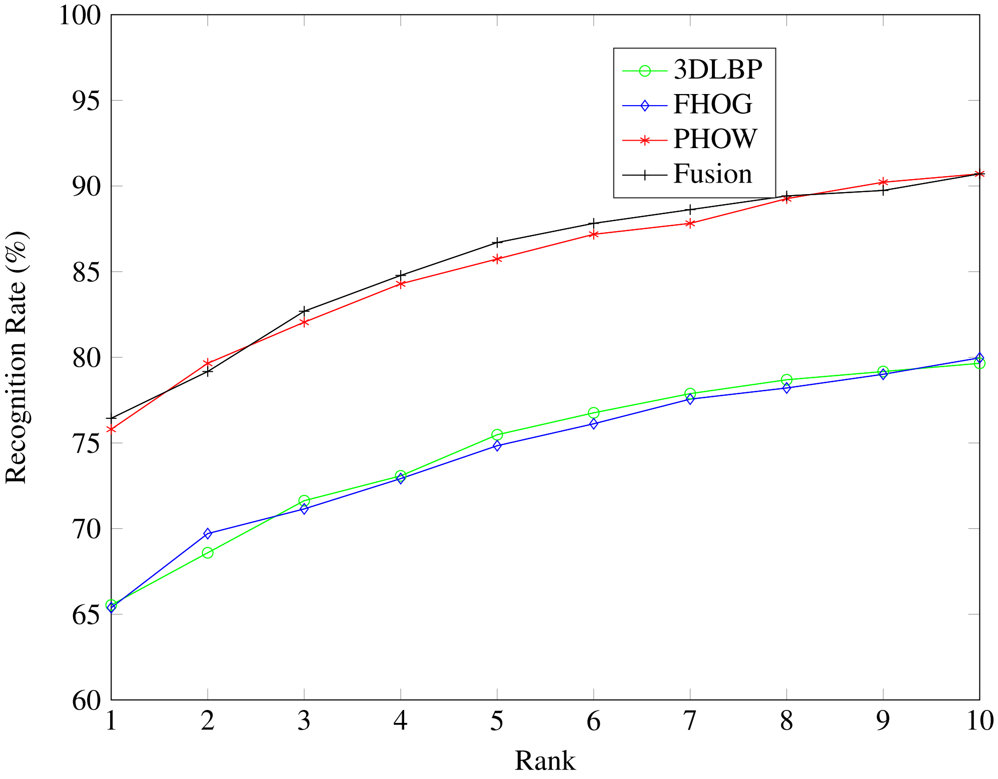


Figure 4.4: CMC curve for different descriptors in Eurecom dataset: 3DLBP (green), FHOG (blue), PHOW (red) and Fusion between FHOG + PHOW (black).

4.5 Multimodal Face Recognition Framework

Both Kinect and RealSense^{TM} sensors provide color and infra-red streams that can also provide important information for face recognition. Therefore, the developed 3D framework will be expanded to include these modalities, to provide an increase in performance. As we saw in Chapter 2, multi-modality proves to improve the robustness of face recognition systems, and should be important in cases like complete darkness, where one of the modalities (RGB) does not provide relevant information.

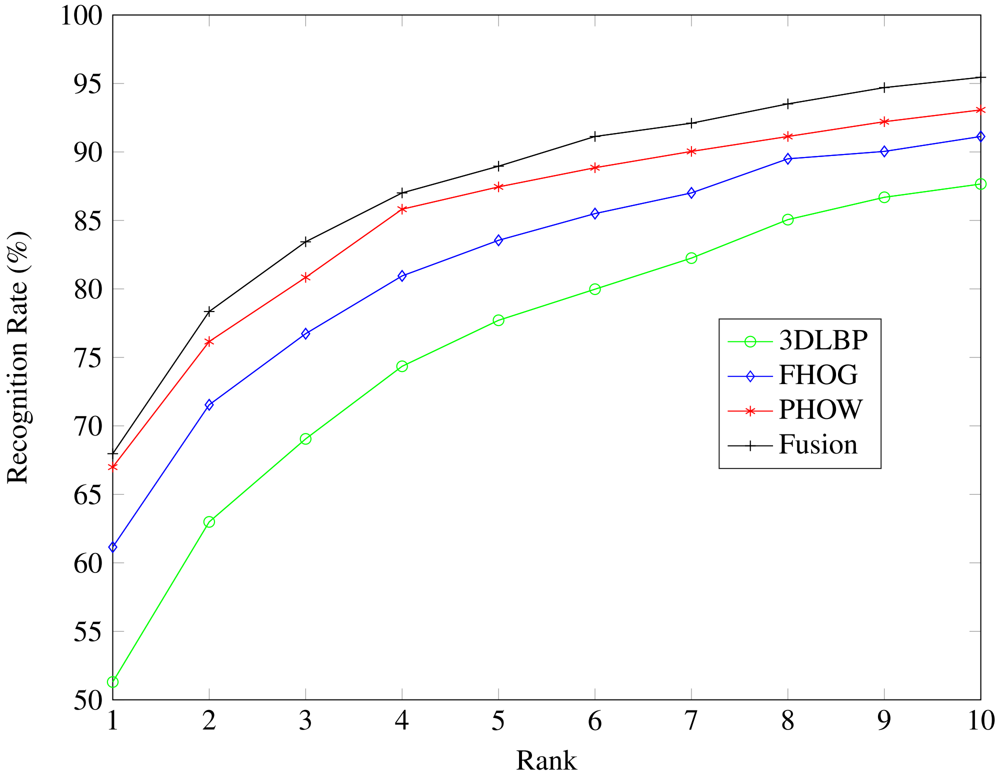


Figure 4.5: CMC curve for different descriptors in RealFace dataset: 3DLBP (green), FHOG (blue), PHOW (red) and Fusion between FHOG + PHOW (black).

The Eurecom Dataset only provides RGB information, while RealFace dataset provides RGB and IR images. It is important, then, to define which strategy to use in order to integrate IR and RGB modalities in the developed framework.

Some of the most promising works in Computer Vision have explored Convolutional Neural Networks (CNNs) in diverse applications, thanks to the appearance of large datasets and more powerful and faster GPUs (Krizhevsky et al. (2012), Simonyan and Zisserman (2014)). More recent works extended CNNs to deep face recognition, elevating the performance of face recognition systems. Some of the most popular approaches are DeepFace (Taigman et al., 2014), VGG-Face (Parkhi et al., 2015) and DeepID3 (Sun et al., 2015). These CNNs were trained in millions of images, to allow the creation of powerful classifiers for face recognition. The constructed deep classifiers are able to extract robust features for face recognition. Using pre-trained CNNs, one could use a CNN for feature extraction instead of classification (Razavian et al., 2014), and use such robust features to train new models for different datasets.

Using the pre-trained model provided by Parkhi et al. (2015), we have tested the robustness of the VGG-Face CNN for both RGB and IR modalities. The resulting feature vectors of length 4096 are L2-normalized and trained using a logistic regression classifier, with similar conditions to the ones used for 3D images. The RGB images of Eurecom dataset were cropped accordingly to the keypoints provided by the dataset, whereas in RealFace, images were manually cropped.

Since RGB images and IR-images are visually similar, one could think of a direct adaptation of the VGG-Face CNN to IR images. Since VGG-Face is expecting RGB-images, the IR image is replicated for the three-channels. One could also think that this approach would lead to similar redundant features, but since the CNN subtracts the mean intensity of the images used in training of the CNN, the features are different, since the inputs in each channel are also different.

The results, shown on Tables 4.8 and 4.9, were promising and showed that the features extracted by VGG-Face, were robust enough to deal with occlusion, facial expression and disguises. In the RealFace dataset, the results for darkness images in the RGB modality were not considered due to their low performance in this conditions. Even being designed for RGB, when applied on IR-images, the VGG-Face CNN is able to extract relevant features that allow a good performance in such images.

Table 4.8: Accuracy Results (%) on 3D (using PHOW+FHOG) and RGB (using VGG-Face features) in the Eurecom dataset.

Modality	LO	OE	OM	OP	MO	S	Global
3D	97.12	93.27	50.96	24.04	93.27	100	76.44
RGB	100	98.08	95.19	96.16	100	100	98.24

Table 4.9: Accuracy Results (%) on 3D (using PHOW+FHOG), IR (using VGG-Face features) and RGB (using VGG-Face features) in the RealFace Dataset.

Modality	NN	NMO	NS	NG	AMO	AS	AG	DN	DMO	DS	DG	Global
3D	78.57	69.05	47.62	71.43	89.29	44.05	73.81	91.67	73.81	48.81	59.52	67.97
IR	100	96.43	76.19	78.57	96.43	75.00	73.81	98.81	95.24	61.90	71.43	83.98
RGB	98.81	96.43	80.95	89.29	100	88.10	95.24	-	-	-	-	92.69

To evaluate the performance of the system with multiple modalities, the individual logistic regression probabilities were combined using an analogous formula to Equation 4.1:

$$Global_{prob} = w_1 \times prob_{3D} + w_2 \times prob_{IR} + w_3 \times prob_{RGB} \tag{4.2}$$

The sum of the weights is forced to be unitary. In Figure 4.6 is shown the proposed multimodal framework. IR images are only used in sensors which this modality is provided.

To overcome the loss of performance in the case of RGB in darkness conditions, particularly in the RealFace, a new method is suggested to deal with illumination conditions: for all test images the mean intensity of gray-scale converted RGB image is calculated, and, depending on this value, a weight for RGB-modality is calculated, using a logistic function:

$$Weight_{RGB} = \frac{1}{1+e^{(-0.5(-20+mean_{intensity})}} \tag{4.3}$$

Figure 4.7 represents this function. The value of 20 was set empirically accordingly with the RealFace dataset images, and was set as the mean transition intensity between fair and poor illumination conditions. This adaptation allows the algorithm to self-adapt its performance by adjusting the RGB-weight to be higher in higher illumination, and lower in less ideal low illumination conditions.

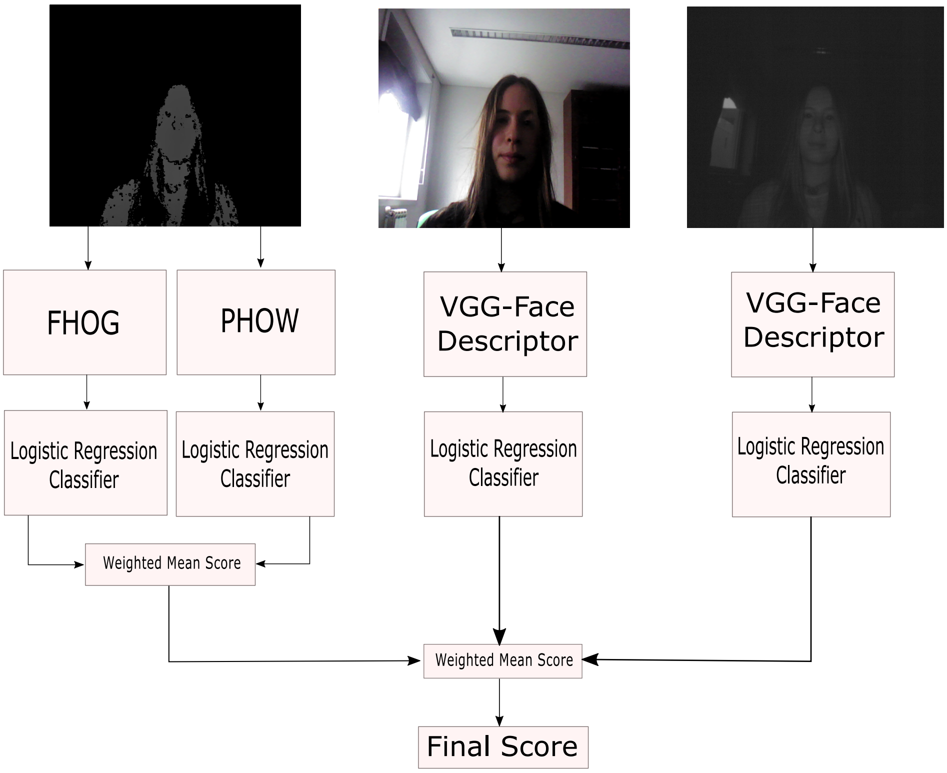

FHOG
PHOW
VGG-Face Descriptor
VGG-Face Descriptor
Logistic Regression Classifier
Logistic Regression Classifier
Logistic Regression Classifier
Logistic Regression Classifier
Weighted Mean Score
Weighted Mean Score
Final Score

Figure 4.6: Scheme that resumes the pipeline of the multimodal proposed framework.

Therefore, Equation 4.2 is used, being combined with the proposed sigmoid function. In cases where illumination conditions are good enough, w_3 is maintained, in poor conditions, $w_3 = Weight_{RGB} \times w_3$. The weight loss from the original w_3 is then divided equally between the other available modalities.

The performance results for all combinations with 3D modality are presented in Tables 4.10 and 4.11. Additionally, Cumulative Match Curves are also presented in Figures 4.8 and 4.9.

In Eurecom, multi-modality does not have a big impact in performance with an improvement of only 0.15% in global performance. Despite that, we can see that the results were improved when compared to the preliminary results using the extension proposed by Monteiro and Cardoso (2015), in both 3D and multimodal (presented in Table 2.5 in Chapter 2).

Due to its severe variation in illumination conditions, in RealFace dataset, the use of all the modalities outperforms their single modality counterparts. We can also see that, with the proposed logistic function for dealing illumination, the performance boost is above to 5% when compared to other combinations. This proves the advantage of using this technique, taking advantage of all the modalities even with varying illumination. Therefore, this method allows us to take full advantage of RGB robust features in good illumination conditions, and discard them in poor illumination

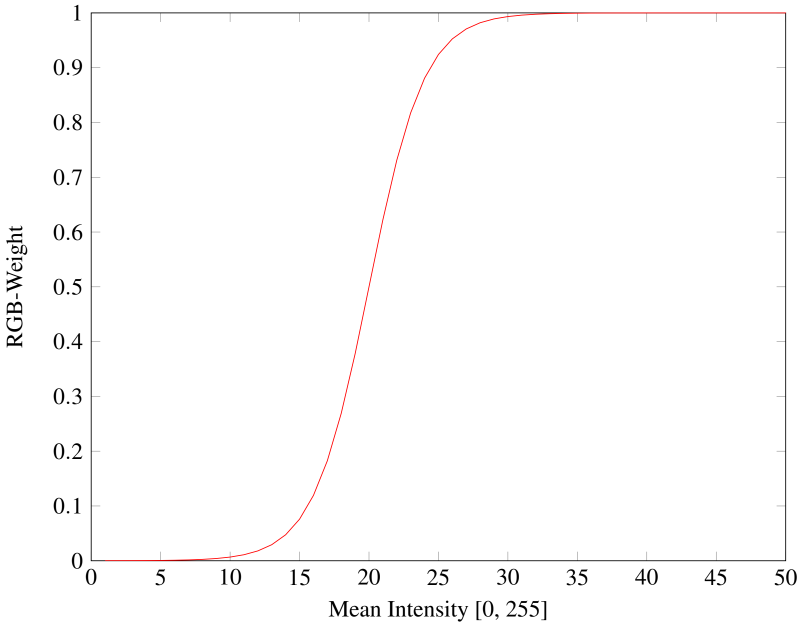


Figure 4.7: Sigmoid function which determines the weight for RGB modality depending on its mean illumination.

conditions.

We can also conclude that, even when one of the extra-modalities (RGB or IR) is missing, the fusion performance is still superior to all the single modalities. Despite that, the best performance is when all the 3 modalities are combined.

Table 4.10: Accuracy Results (%) on 3D (using PHOW+FHOG), RGB (using VGG-Face features) and Multi-modality in the Eurecom dataset.

Modality	LO	OE	OM	OP	MO	S	Global
3D	97.12	93.27	50.96	24.04	93.27	100	76.44
RGB	100	98.08	95.19	96.16	100	100	98.24
RGB + 3D	100	98.08	96.15	96.15	100	100	98.39

Table 4.11: Accuracy Results (%) on 3D (using PHOW+FHOG), IR (using VGG-Face features), RGB (using VGG-Face features) and Different Multi-modal combinations in the RealFace Dataset.

Modality	NN	NMO	NS	NG	AMO	AS	AG	DN	DMO	DS	DG	Global
3D	78.57	69.05	47.62	71.43	89.29	44.05	73.81	91.67	73.81	48.81	59.52	67.97
IR	100	96.43	76.19	78.57	96.43	75.00	73.81	98.81	95.24	61.90	71.43	83.98
RGB	98.81	96.43	80.95	89.29	100	88.10	95.24	-	-	-	-	92.69
IR + 3D	100	96.43	78.57	80.95	96.43	75.00	76.19	98.81	95.24	61.90	75.00	84.96
RGB + 3D	100	97.62	80.95	90.48	100	88.10	95.24	95.24	75.00	47.62	59.52	84.52
IR + 3D + RGB	100	100	89.29	91.67	98.81	90.48	95.24	98.81	95.24	63.1	76.19	90.80

The 3D results show that similar conditions in occlusion and facial expression lead to different performances in different illumination conditions. Although the system is still able to assess the subject identities in different conditions, the depth map seems to differ in different illumination

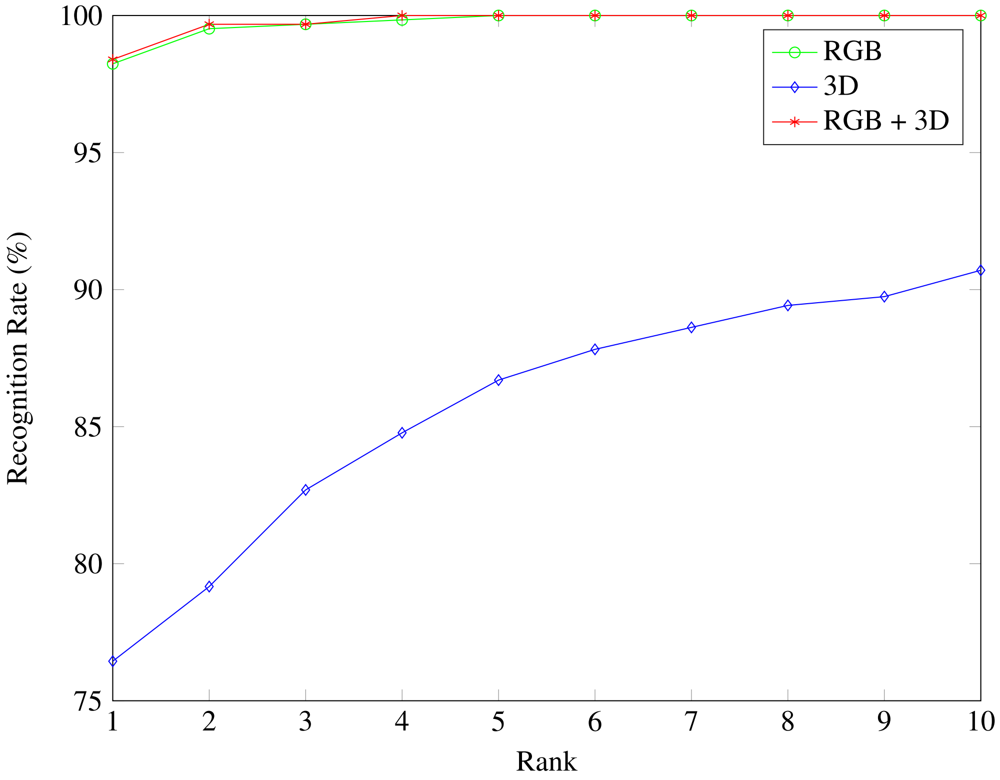


Figure 4.8: Recognition Rates at different Ranks in Eurecom dataset in different modalities: RGB (using VGG-Face features) in green, 3D (using PHOW+FHOG) in blue and 3D + RGB in red.

conditions. The depth estimation depends on the IR emitter, and we can observe that the IR performance also varies in different illumination conditions, where in high illumination the performance seems to increase. Since both streams are correlated, this variation in illumination performance could justify the variations of depth face recognition in different illumination environments.

To evaluate the robustness of the final system we should also test the system to pose variations, which will be the focus of the next section.

4.6 Evaluation of the framework against pose variations

To evaluate the robustness of the developed 3D and multimodal framework, some additional experiences were performed in the RealFace dataset. This dataset offers more images of pose variations (in Eurecom only 4 profile images/per subject are provided), and thus presents a more complete challenge to the developed algorithms. Therefore, maintaining 2 neutral images in artificial illumination for training, the framework was assessed for all the remaining 70 test images/subject. Figure 4.10 shows the performance of individual modalities compared to the multimodal framework in the tested poses (-90°, -45°, 0°, 45° and 90°).

It can be seen that, for pose variations, the 3D framework is the one with lower generalization capability to different pose variations, not being able to extrapolate to new conditions. In both IR and RGB modalities, VGG-Face seem to be robust to pose variations, although, naturally, it exists

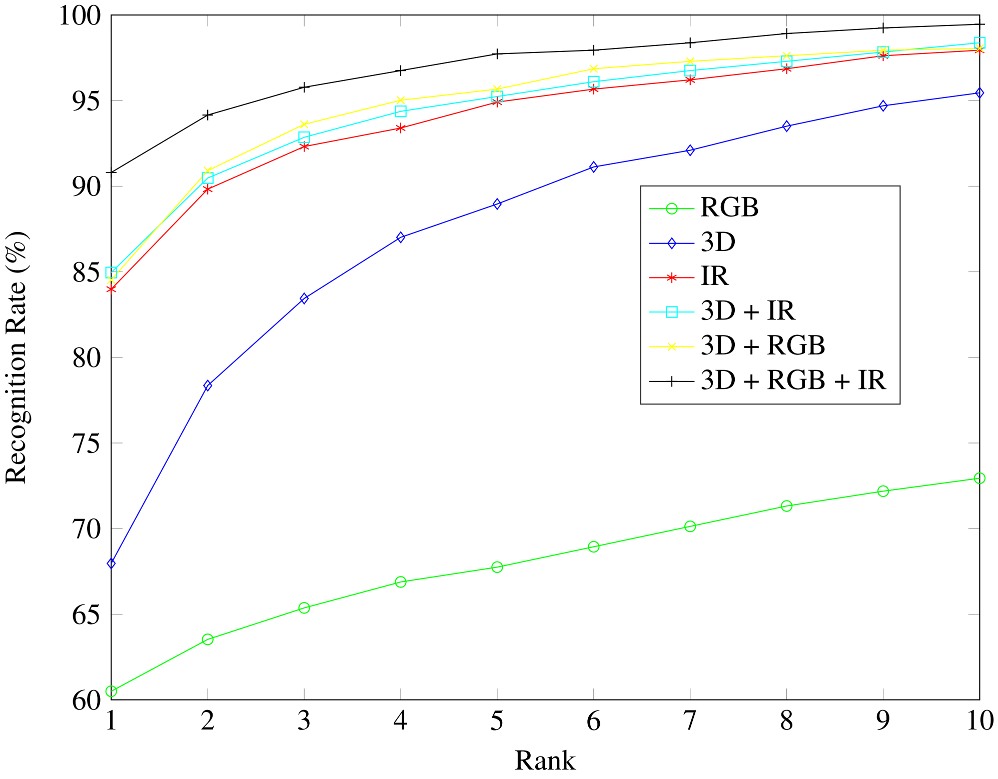


Figure 4.9: Recognition Rates at different Ranks in RealFace dataset in different modalities: RGB (using VGG-Face features) in green, 3D (using PHOW+FHOG) in blue, IR (using VGG-Face features) in red, 3D + IR in cyan, 3D + RGB in yellow and finally 3D + IR + RGB in black.

some decrease in accuracy when profile images are tested. Multi-modality seems to increase the performance in all pose conditions thus improving the robustness of the system.

It is interesting to evaluate whether performance results would be improved if profile images were added to training data. Therefore, additionally to the 2 neutral images, one left and one right profile image in artificial illumination with neutral expression were added to the training set, leading to 4 training images and 68 testing images for each individual. The results for each of the modalities in these conditions are presented in Figure 4.11.

Interestingly, the 3D framework seems to be the one which is able to get better performance in this setup, although it clearly is not robust enough to classify the intermediate case between profile and frontal images, by a margin of approximately 20%. In 3D, the trained models seem to be overfitted to the training data, not being able to generalize to intermediate positions. The remaining modalities seem to increase the robustness in profile poses while maintaining similar performances when compared to the results with only frontal training images. Multi-modality, although getting similar performance for frontal and 45° poses, doubles the performance for profile poses. We can also see that multimodality, although globally better in terms of performance, presents a small performance drop, relatively to the 3D results. This may occur due to big differences in individual modalities performances in this scenario. In frontal poses, RGB and IR seem to have better performance after adding profile training images, which results in higher performance

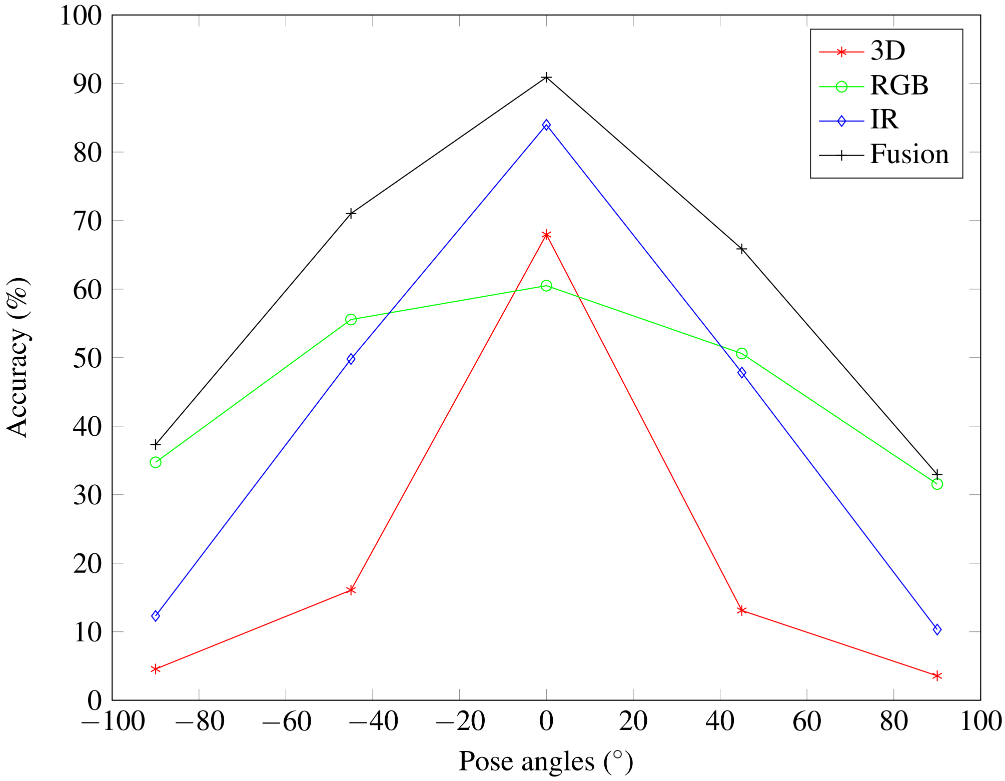


Figure 4.10: Global Accuracy in RealFace dataset in different pose angles using neutral images as training: RGB (using VGG-Face features) in green, 3D (using PHOW+FHOG) in red, IR (using VGG-Face features) in blue, 3D + IR in cyan, 3D + RGB in yellow and finally 3D + IR + RGB in black.

in this modalities, and an improvement of 4% in multimodal framework.

Training with profile poses seems to improve the global accuracy, but it still does not solve all the problems. It important to refer that, additional to performance losses due to pose variations, there is also the decrease of performance due to natural occlusions (hair mainly), facial expression and disguises. Therefore, the recognition task in these conditions is not trivial and the obtained global performance of 77.38 % is a fair value in these conditions. Despite that, a pose correction or a face alignment algorithm introduced in 3D, could significantly improve the results and should be able to attenuate losses in performance. Additionally, due to its potential, it would be a considerable improvement to create a CNN for depth face recognition, although this is a difficult task due to the lack of available data. Even for IR, a fine adaptation of the pre-trained VGG-face network adapted to IR images, could result in a good improvement in performance, since this CNN was not specifically designed for this type of images.

4.7 Comparison With Other State-Of-The-Art Methodologies

To end the analysis of the proposed framework it is important to compare the results with some state of the art methods in the Eurecom dataset. Therefore, a comparison in terms of performance

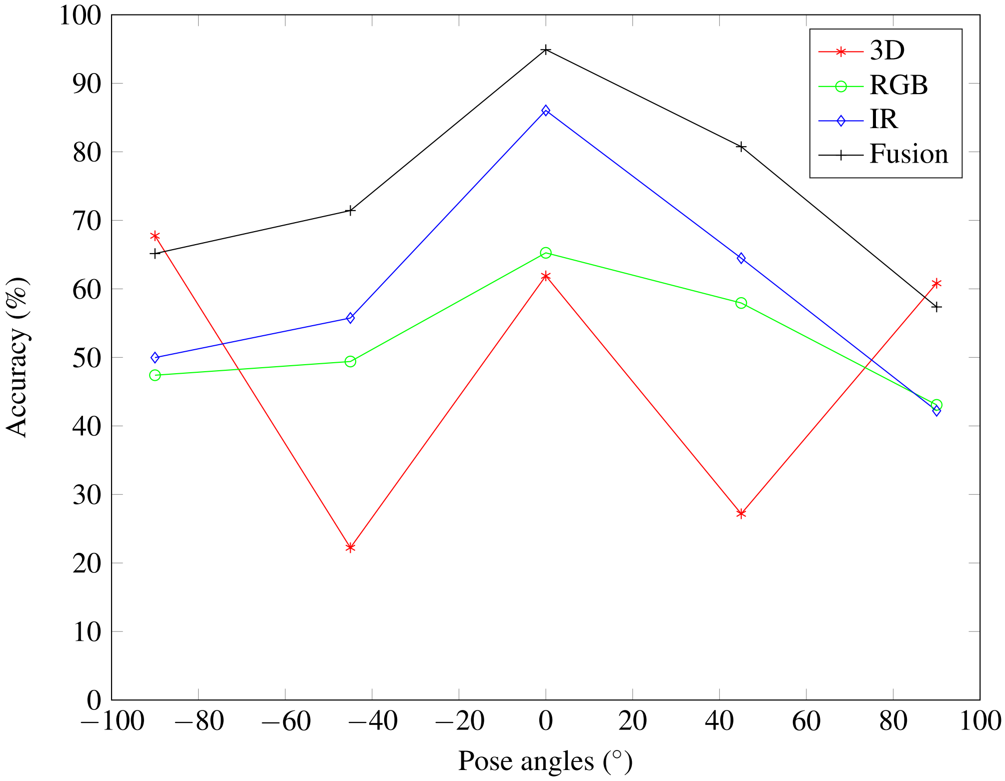


Figure 4.11: Global Accuracy in RealFace dataset in different pose angles using neutral and profile images as training: RGB (using VGG-Face features) in green, 3D (using PHOW+FHOG) in red, IR (using VGG-Face features) in blue, 3D + IR in cyan, 3D + RGB in yellow and finally 3D + IR + RGB in black.

with some methods are presented in Table 4.12.

Table 4.12: Comparison of the obtained Accuracy Results (%) with some state-of-the art methodologies performances in the Eurecom dataset.

Methodology	Modality	LO	OE	OM	OP	MO	S	Global
LBP (Min et al., 2014)	3D	95.2	78.9	51.9	12.5	81.7	83.7	67.3
LBP (Monteiro and Cardoso, 2015)	3D	79.8	63.5	43.3	10.6	53.8	78.8	55.0
Cardia Neto and Marana (2015)	3D	-	-	-	-	-	-	98.0
Proposed	3D	97.1	93.3	51.0	24.0	93.3	100	76.4
LGBP (Min et al., 2014)	3D + RGB	100	89.4	98.1	84.6	98.1	100	95.0
LBP (Min et al., 2014)	3D + RGB	99	93.4	98.1	81.7	96.2	100	94.7
2D-SIFT + 3D-LBP (Monteiro and Cardoso, 2015)	3D + RGB	100	98.1	95.2	62.5	93.3	99.0	91.4
Ajmera et al. (2014)	3D + RGB	-	-	-	-	-	-	89.3
Goswami et al. (2013)	3D + RGB	-	-	-	-	-	-	88.0
Hsu et al. (2014)	3D + RGB	100	84	99	86	96	100	∼94
Sang et al. (2015)	3D + RGB	100	85	99	86	97	100	∼95
Proposed	3D + RGB	100	98.1	96.2	96.2	100	100	98.4

Before taking conclusions, is important to refer that these results were assessed using different experimental setups. Min et al. (2014) and Hsu et al. (2014) only used one neutral image for training, and the remaining for testing. Cardia Neto and Marana (2015) tested only neutral images, training the models with ope mouth, smile and light on images. Ajmera et al. (2014) uses 4 randomly training images and the remaining for testing. Finally, Sang et al. (2015) used smile, open mouth, light on and neutral images from session 1 as training, and the remaining were used for testing. The remaining methods, used a setup analogous to the previously referred in this dissertation.

Analyzing the 3D methodologies, it is clear that the proposed framework outperforms the state-of-the-art, except for Occlusion Mouth scenario, where Min et al. (2014) seems to outperform our method. A fair comparison could not be totally made since the results for Min et al. (2014) and Cardia Neto and Marana (2015) were assessed in different setups. Despite that, the proposed 3D framework seems to represent a good contribution when compared to the remaining state-of-the-art methods.

As for the multimodal works, our system seems overall to be the more robust and with best performing methodology in Eurecom. Despite that, Hsu et al. (2014) and Min et al. (2014), using less gallery images for training, present a good performance and could come closer to the performance achieved in this dissertation, if the number of training images was 2, as in the proposed experimental. Hsu et al. (2014) also includes an automatic pose correction method, that also is an advantage relatively to the proposed methodology.

The presented results proven that the proposed framework is competitive with the rest of the state-of-the-art methods, achieving good performance in unconstrained environments.

Globally, the proposed framework got good results in both tested databases and showed to be robust to variations in occlusion, facial expression and disguises, while also getting interesting results with pose variations. In unconstrained conditions, the system is able to assess the identity of the query subjects with good performance. The use of different modalities seems to always improve the robustness of the framework, leading to higher performances in the wide array of tested image variations.

The developed framework allowed the creation of a multimodal face recognition prototype, using Intel® RealSense^{TM} F200 model. This prototype will be the focus of the next chapter.

Chapter 5

Prototype Development

During this dissertation, a prototype for face recognition using the new Intel® RealSense$^{\text{TM}}$ F200 model was developed.

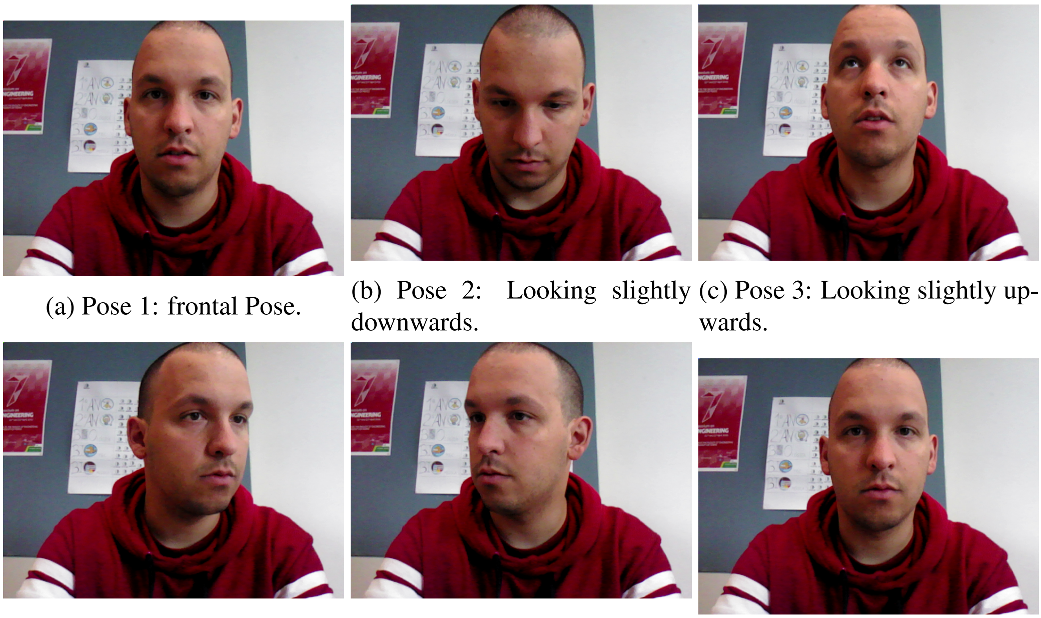
The prototype was created using the same library used for RealFace dataset acquisition, LibrealSense (Intel®, 2015) and was built in C++ language, using the OpenCV library (Cosenza, 2016). This library was used due to its already implemented functionalities designed for real-time computer vision application systems. Additionally to Librealsense and OpenCV, Caffe (Jia et al., 2014), a deep learning framework, was used for the integration of the VGG-Face CNN in the prototype.

Taking advantage of IR, Depth an RGB streams, a preliminary multimodal framework has been built using a similar approach as the described in the previous Chapter.

To develop this prototype a small dataset was acquired for 7 subjects in which 6 images were captured: two frontal, two image taken while looking slightly to the right and left, and two other taken while looking slightly upwards and downwards. This small dataset was acquired to create the first subject models for the real-time system to work with. An example of these six images for one of the subjects is shown in Figure 5.1.

By capturing 30 frames/second, the system starts by detecting the face region in IR and RGB images, using a Viola-Jones cascade face detector (included in OpenCV). To improve the detection in IR images, contrast is improved using CLAHE (Contrast Limited Adaptive Histogram Equalization (Zuiderveld, 1994)). Since the Depth and IR images are aligned, the detected face region in IR image is also used in depth images. If no face is detected in the IR image, the corresponding frame is not used. It is important to refer that only the biggest face in the image is used by the system, and therefore, no multi-person identification at the same time is possible in this prototype version.

The detected face depth regions are then resized for 96 × 96, before feature extraction. IR and RGB are resized to 224 × 224 and inputted in two parallel VGG-Face CNNs for feature extraction, while for the depth image 3DLBP is used. The three acquired feature vectors are, then, compared individually with all the feature vectors of the all images of the dataset (whose feature have been extracted and saved offline), using an Euclidean distance classifier, resulting in



(a) Pose 1: frontal Pose.

(b) Pose 2: Looking slightly downwards.

(c) Pose 3: Looking slightly upwards.

(d) Pose 4: Looking slightly to the left.

(e) Pose 5: Looking slightly to the right.

(f) Pose 6: Frontal Pose.

Figure 5.1: Representative scheme of the poses visited by each subject in each of the conditions for enrollment in the prototype.

a minimum distance to each enrolled individual data. This classifier was used due to its simplicity and to some preliminary tests with good results, while also being able to work in real-time.

The mean intensity of the image is then calculated and, using the Equation 4.3, the weight of RGB modality is calculated. In optimal conditions, the RGB stream will have 50% of weight, while in very poor illumination conditions this weight will be reduced with an asymptotic tendency to 0%. The distances from each of the classifiers are then weighed, for each of the subjects in the database for the current frame. The decision for the correspondent identity is not made using solely the current frame, with the mean of the distances of the last 100 frames being used to increase the robustness of the system. It is expected that the percentage of errors is diluted by using data from multiple frames.

Furthermore, the distance between the extracted descriptors relative to the previous frame is also computed. If this distance is above a certain threshold, the distances from the last 100 frames are reset. This allows us to deal with the cases where a new subject appears suddenly in the image. To deal with the presence of no subjects in the image, if no face is detected in 5 consecutive frames, the system also resets the saved distances from the last frames.

The prototype was designed for demonstration purposes. The interface displayed to the user, depicted on Figure 5.2, displays the 3 real-time streams, while displaying the 5 most likely identities, allowing to the users to see the real-time performance of the system. The interface is console-based, informing the user which controls should be used for each interaction, before starting. A display message indicates when no face is detected in the images.

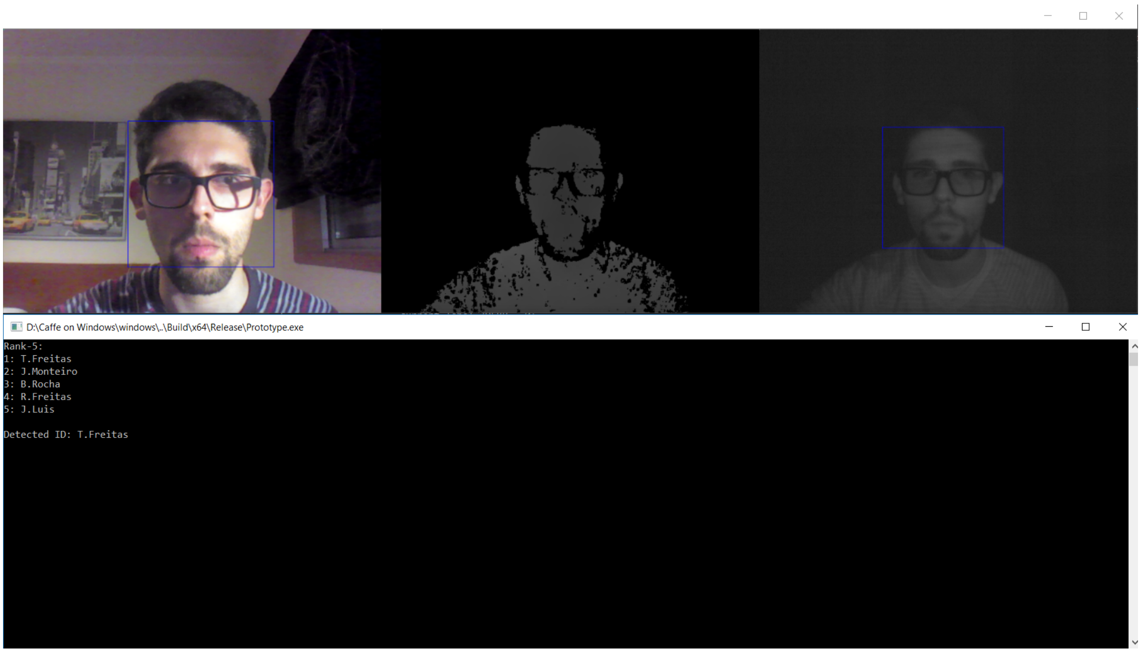
One thing that was crucial for demonstration purposes was online enrollment of new users.

Anytime, new users are allowed to add themselves to the database. The user receives indications to take six images in the same previously described conditions of the individuals of the original database. From the acquired images, feature extraction is performed, and the resulting feature vectors are then added to the original dataset, allowing future identification of the user. Therefore, the system does not save the images, only the extracted feature vectors. The user is added with the desired user name, and all the data is saved in XML files.

The tests performed with the prototype showed that the system is able to assess the identity of different subjects, even when we face fast transitions between subjects in the image and in different illumination conditions. The tests were carried out at INESC-TEC Open Day CTM 2016 (INESC-TEC, 2016), where different users were added to the database and tested the framework.

Despite the positive results there are several improvements that need to be made in order to create a more robust system, closer to a deployable product. First of all, the 3D feature extraction should be changed to a more fit framework than the currently used, replacing 3DLBP by a combination of PHOW and FHOG.

Additionally, it is important to include the possibility of automatically detecting when a user is not included in the dataset, a functionality which is not yet implemented in the current prototype. Simultaneously real-time face recognition of multiple individuals would also be a good feature. Improvements should also be made in terms of face detection, where in high pose variations, the current used face detector does not perform well enough. The creation of a face detector for depth and IR images should resolve this problem. Also, as referred in the previous chapter, the inclusion of pose correction in depth images could also be a good addition.


D:\Caffe on Windows\windows\..\Build\x64\Release\Prototype.exe

```
Rank-5:
1: T.Freitas
2: J.Monteiro
3: B.Rocha
4: R.Freitas
5: J.Luis

Detected ID: T.Freitas
```

Figure 5.2: Console-Based interface of the prototype.

For a real-time system, a log with the detected subjects should be implemented, while also saving the video streams in a live server for security purposes.

Obviously, the interface could be improved to be more user-friendly. Despite not being yet in its final stage, this prototype serves well as a proof of concept that Intel® RealSense$^{\text{TM}}$ sensors (at least the short distance one) could be used for real-time face recognition. It would be interesting, in the future, to develop the same prototype in the long range sensor taking advantage of the longer ranges at which the sensor operates. The long range model is still in its birth stage and needs some improvements to allow this implementation, though.

The developed multimodal framework combined with this prototype allowed to idealize some suggestions to follow up the work developed in this dissertation. In the next chapter we will talk about these suggestions and about the global conclusions of the present work.

Chapter 6

Conclusions and Future Work

This dissertation allowed the investigation of new methods in 3D face recognition, using low-cost sensors. In unconstrained environments, the presence of 3D information could be important to increase the robustness of real-time systems. The integration of depth information with RGB, has proven to always increase the performance of face recognition systems, as is referred in the state-of-the-art (Abate et al., 2007).

With the emergence of the Intel® RealSense$^{\text{TM}}$ depth sensors, and verifying that the current available 3D face recognition datasets using low-cost sensors were all acquired with Kinect Sensors, a new dataset, the RealFace dataset, was created. This dataset comes as an alternative to the scientific community, providing challenging conditions for multimodal face-recognition systems. Additionally, it can be used in other applications, such as face alignment experiences, gender and age prediction as well as face detection experiments in depth and IR modalities. The creation of a real-time face recognition prototype using the F200 model, showed that this sensor is fast enough for real-time applications and can be implemented in real-time systems for 3D and multimodal face recognition.

RealFace dataset allowed the assessment of Intel® RealSense$^{\text{TM}}$ F200 (currently re-named for SR300) in face recognition systems, being robust enough to compete with Kinect sensor. Despite that, the long range R200 model proved to be too noisy for this type of applications. In the captured dataset, the sensor did not perform well enough and still needs to undergo improvements to compete with the remaining sensors (a similar observation was also presented in Song et al. (2015)).

The dataset also allows testing face recognition methodologies in different illumination conditions, evaluating if the developed multimodal algorithms are robust enough to adapt to these variations. Additionally, the created database allowed to verify if 3D face recognition was, in fact, totally independent of the illumination conditions. Notwithstanding the fact that 3D face recognition can be performed in extreme conditions of darkness, the performance did vary in different illumination conditions, not being fully invariant to this factor.

Among the various feature extractors tested, FHOG and PHOW seem to outperform all the remaining tested alternatives, even though they were not designed for depth images. Despite that, the performance in 3D are still not close enough to their RGB counterparts. The developed framework showed to be robust to variations in facial expression, disguises, natural and artificial occlusions and partially to pose variations. The proposed system can be applied with any low-cost depth sensor, being able to integrate additional modalities. The framework seems to be able to compete with other state-of-the art methods, getting similar or superior performances in the Eurecom dataset.

The use of the VGG-Face as a feature extractor in IR images, has proven to be capable of extracting robust features, can be used in face recognition using this modality. Compared to the results in RGB images, in similar conditions, a small drop in performance was noted, as expected, since this CNN was trained with RGB images, not being designed for IR images.

Despite the overall good capability of the developed system to perform 3D and multimodal face recognition in unconstrained environments, some suggestions for improvements and future work have been identified and will be outlined in the next section.

6.1 Future Work

In terms of the developed framework, some improvements can be made, namely the inclusion of a pose normalization and face alignment algorithms, in order to increase performance in scenarios of high pose variation. To improve the recognition in occlusion cases, a good addition would be to adapt the developed framework to include a hierarchical model, similar to the proposed by Monteiro and Cardoso (2015). Since this method depends on a correct face alignment, the previously referred pose normalization and face alignment inclusion would be crucial to this adaptation.

To improve the adaptation of the system to challenging environments, it would be interesting to create a more robust manner to choose the different modality weights, possibly by training a model that receives image quality and illumination features.

The RealFace dataset should also be improved to include more subjects, in order to increase its variability, before being made public to the scientific community.

The developed prototype can be improved in many ways. As referred in the previous chapter, the inclusion of a 3D system similar to the proposed in Chapter 4, should improve its robustness. The inclusion of pose normalization and face alignment system should also be a good addition. The creation of a new face detector, both for depth and for IR, combined with a keypoint detector, could improve the currently used face detection method, based in traditional Viola-Jones.

When the long range depth sensor increases its quality in depth data, it would be also interesting to create a similar prototype for the long range model. It would be positive too to study if one could adapt the trained models in the short sensor, to images acquired with a different sensor like the long range model or even Kinect.

Taking into account the high performances of CNN-based works in computer vision, especially in face recognition, it would be interesting to create a CNN, similar to VGG-Face designed for face recognition in depth images. Although, still not being available for face recognition, some recent works tried to implement this idea to object recognition (Cheng et al. (2015), Gupta et al.

(2014), Sun et al. (2015)). This CNN could be trained from scratch or by fine-tuning the original VGG-Face. The problem, when compared to the RGB CNN, is the lack of data, since the total images in all available databases, does not come even closer to the number of images used in the VGG-Face training (approximately 60 thousand vs 1 million images). Nevertheless, preliminary developments are already being carried out to train such CNN. Using all the data from the referred datasets in Chapter 2 (except Eurecom), the number of images were augmented by flipping, rotating and generating new views of faces by Point Cloud rotations and posterior conversion to depth images. Until this time, the results are not yet conclusive, but some promising results were accomplished in small subsets of the augmented data. Currently, the CNN has not yet been fully trained with good performances for the entire augmented dataset. If a similar CNN is terms of robustness is created, depth recognition rates could come closer to RGB values, which is yet not possible. Additionally, a similar adaptation of VGG-Face to IR images could represent an equally important contribution.

The referred suggestions for future research could complement the developed work in this dissertation, making the developed system more robust for unconstrained 3D face recognition applications.

Bibliography

3dMD (n.d.). 3dmdface system. Available at `http://www.3dmd.com/3dmd-systems/#face`, Accessed on 15 January 2016.

Abate, A. F., M. Nappi, D. Riccio, and G. Sabatino (2007). 2d and 3d face recognition: A survey. *Pattern Recognition Letters 28*(14), 1885–1906.

Ajmera, R., A. Nigam, and P. Gupta (2014). 3d face recognition using kinect. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, pp. 76. ACM.

Albiol, A., D. Monzo, A. Martin, J. Sastre, and A. Albiol (2008). Face recognition using hog–ebgm. *Pattern Recognition Letters 29*(10), 1537–1543.

Amon, C., F. Fuhrmann, and F. Graf (2014). Evaluation of the spatial resolution accuracy of the face tracking system for kinect for windows v1 and v2. In *Proceedings of the 6th Congress of the Alps Adria Acoustics Association*.

Ansari, A., M. Abdel-Mottaleb, et al. (2003). 3d face modeling using two views and a generic face model with application to 3d face recognition. In *Advanced Video and Signal Based Surveillance, 2003. Proceedings. IEEE Conference on*, pp. 37–44. IEEE.

Asus (2011). Xtion pro live. Available at: `https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/overview/`, Accessed on 15 January 2016.

Bennamoun, M., Y. Guo, and F. Sohel (2015). Feature selection for 2d and 3d face recognition. *Encyclopedia of electrical and electronics engineering. Book Chapter*, 1–54.

Berretti, S., A. D. Bimbo, and P. Pala (2012). Superfaces: A super-resolution model for 3d faces. In *ECCV Workshops (1)*, pp. 73–82.

Besl, P. J. and N. D. McKay (1992). Method for registration of 3-d shapes. In *Robotics-DL tentative*, pp. 586–606. International Society for Optics and Photonics.

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

Bondi, E., P. Pala, S. Berretti, and A. Del Bimbo (2015). Reconstructing high-resolution face models from kinect depth sequences acquired in uncooperative contexts. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Volume 7, pp. 1–6. IEEE.

Bosch, A., A. Zisserman, and X. Munoz (2007). Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8. IEEE.

Bowyer, K. W., K. Chang, and P. Flynn (2006). A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer vision and image understanding 101*(1), 1–15.

Cao, C., Y. Weng, S. Zhou, Y. Tong, and K. Zhou (2014). Facewarehouse: a 3d facial expression database for visual computing. *Visualization and Computer Graphics, IEEE Transactions on 20*(3), 413–425.

Cao, Y. and B.-L. Lu (2015). Intensity-depth face alignment using cascade shape regression. In *Neural Information Processing*, pp. 224–231. Springer.

Cardia Neto, J. B. and A. N. Marana (2015). 3dlbp and haog fusion for face recognition utilizing kinect as a 3d scanner. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pp. 66–73. ACM.

Cerna, L., G. Cámara-Chávez, and D. Menotti (2013). Face detection: Histogram of oriented gradients and bag of feature method. In *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, pp. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Chang, K., K. Bowyer, and P. Flynn (2003). Face recognition using 2d and 3d facial data. In *ACM Workshop on Multimodal User Authentication*, pp. 25–32.

Chang, K. I., K. W. Bowyer, and P. J. Flynn (2005). An evaluation of multimodal 2d + 3d face biometrics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27*(4), 619–624.

Cheng, Y., X. Zhao, K. Huang, and T. Tan (2015). Semi-supervised learning and feature evaluation for rgb-d object recognition. *Computer Vision and Image Understanding 139*, 149–160.

Chua, C.-S., F. Han, and Y.-K. Ho (2000). 3d human face recognition using point signature. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 233–238. IEEE.

Colombo, A., C. Cusano, and R. Schettini (2011). Umb-db: A database of partially occluded 3d faces. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 2113–2119. IEEE.

Cook, J., V. Chandran, S. Sridharan, and C. Fookes (2004). Face recognition from 3d data using iterative closest point algorithm and gaussian mixture models. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pp. 502–509. IEEE.

Cosenza, M. (2016). Opencv (open source computer vision). Available at: `http://opencv.org/`, Accessed on 15 January 2016.

Cyberware (n.d.). Cyberware model 3030 color 3d scanhead. Available at: `http://cyberware.com/products/scanners/3030Specs.html`, Accessed on 15 January 2016.

Dai, X., S. Yin, P. Ouyang, L. Liu, and S. Wei (2015). A multi-modal 2d+ 3d face recognition method with a novel local feature descriptor. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pp. 657–662. IEEE.

Dal Mutto, C., P. Zanuttigh, and G. M. Cortelazzo (2012). *Time-of-Flight Cameras and Microsoft Kinect^{TM}*. Springer Science & Business Media.

Dalal, N. and B. Triggs (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Volume 1, pp. 886–893. IEEE.

Diamond, R. and S. Carey (1986). Why faces are and are not special: an effect of expertise. *Journal of Experimental Psychology: General 115*(2), 107.

Elaiwat, S., M. Bennamoun, F. Boussaid, and A. El-Sallam (2015). A curvelet-based approach for textured 3d face recognition. *Pattern Recognition 48*(4), 1235–1246.

Fanelli, G., M. Dantone, J. Gall, A. Fossati, and L. Van Gool (2013, February). Random forests for real time 3d face analysis. *Int. J. Comput. Vision 101*(3), 437–458.

Farah, M., G. Humphreys, and H. Rodman (1999). Object and face recognition. *Fundamental neuroscience, ed. MJ Zigmond, FE Bloom, SC Landis, JL Roberts & LR Squire. Academic Press.[aFVDV].*

Farah, M. J. (1996). Is face recognition 'special'? evidence from neuropsychology. *Behavioural brain research 76*(1), 181–189.

Felzenszwalb, P. F., R. B. Girshick, D. McAllester, and D. Ramanan (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*(9), 1627–1645.

Galoogahi, H. K. and T. Sim (2012). Inter-modality face sketch recognition. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pp. 224–229. IEEE.

Gordon, G. G. (1991). Face recognition based on depth maps and surface curvature. In *San Diego,'91, San Diego, CA*, pp. 234–247. International Society for Optics and Photonics.

Goswami, G., S. Bharadwaj, M. Vatsa, and R. Singh (2013). On rgb-d face recognition using kinect. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pp. 1–6. IEEE.

Goswami, G., M. Vatsa, and R. Singh (2014). Rgb-d face recognition with texture and attribute features. *Information Forensics and Security, IEEE Transactions on 9*(10), 1629–1640.

Guardian, T. (2013). Apple buys 3d sensor company primesense for $350m. Available at: `http://www.theguardian.com/technology/2013/nov/25/apple-buys-primesense-microsoft-kinect`, Accessed on 15 January 2016.

Gupta, S., K. R. Castleman, M. K. Markey, and A. C. Bovik (2010). Texas 3d face recognition database. In *Image Analysis & Interpretation (SSIAI), 2010 IEEE Southwest Symposium on*, pp. 97–100. IEEE.

Gupta, S., R. Girshick, P. Arbeláez, and J. Malik (2014). Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, pp. 345–360. Springer.

Gutfeter, W. and A. Pacut (2015). Face 3d biometrics goes mobile: Searching for applications of portable depth sensor in face recognition. In *Cybernetics (CYBCONF), 2015 IEEE 2nd International Conference on*, pp. 489–494. IEEE.

Hayat, M., M. Bennamoun, and A. A. El-Sallam (2015). An rgb–d based image set classification for robust face recognition from kinect data. *Neurocomputing*.

Heseltine, T., N. Pears, and J. Austin (2004a). Three-dimensional face recognition: A fishersurface approach. In *Image Analysis and Recognition*, pp. 684–691. Springer.

Heseltine, T., N. Pears, and J. Austin (2004b). Three-dimensional face recognition: An eigensurface approach. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, Volume 2, pp. 1421–1424. IEEE.

Hesher, C., A. Srivastava, and G. Erlebacher (2002). Principal component analysis of range images for facial recognition. In *Proc. of the International Conference on Imaging Science, Systems, and Technology*, pp. 62–68.

Hg, R., P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund, and G. Tranchet (2012). An rgb-d database using microsoft's kinect for windows for face detection. In *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, pp. 42–46. IEEE.

Hiremath, P. and H. Manjunatha (2013). 3d face recognition based on depth and intensity gabor features using symbolic pca and adaboost. *International Journal of Signal Processing, Image Processing and Pattern Recognition 6*(5), 1–12.

Hsia, C.-H. (2015). Improved depth image-based rendering using an adaptive compensation method on an autostereoscopic 3-d display for a kinect sensor. *Sensors Journal, IEEE 15*(2), 994–1002.

Hsu, G.-S. J., Y.-L. Liu, H.-C. Peng, and P.-X. Wu (2014). Rgb-d-based face reconstruction and recognition. *Information Forensics and Security, IEEE Transactions on 9*(12), 2110–2118.

Hu, Y., D. Jiang, S. Yan, L. Zhang, and H. Zhang (2004). Automatic 3d reconstruction for face recognition. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pp. 843–848. IEEE.

Huang, D., M. Ardabilian, Y. Wang, and L. Chen (2012). 3-d face recognition using elbp-based facial description and local feature hybrid matching. *Information Forensics and Security, IEEE Transactions on 7*(5), 1551–1565.

Huang, Y., Y. Wang, and T. Tan (2006). Combining statistics of geometrical and correlative features for 3d face recognition. In *BMVC*, pp. 879–888. Citeseer.

Huynh, T., R. Min, and J.-L. Dugelay (2013). An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data. In *Computer Vision-ACCV 2012 Workshops*, pp. 133–145. Springer.

I3DU (n.d.). Primesense$^{\text{TM}}$ 3d sensors. Available at: `http://www.i3du.gr/pdf/primesense.pdf`, Accessed on 15 January 2016.

INESC-TEC (2016). Inesc-tec open day ctm 2016. Available at: `http://opendayctm.inesctec.pt/`, Accessed on 15th June 2016.

Intel® (2015). Intel® realsense$^{\text{TM}}$ cross platform api. Available at: `https://github.com/IntelRealSense/librealsense`, Accessed on 1st June 2016.

Intel (2015a). Intel® realsense$^{\text{TM}}$ development kit. Available at: `https://software.intel.com/en-us/RealSense/Devkit/`, Accessed on 15 January 2016.

Intel (2015b). Introducing the intel® realsense$^{\text{TM}}$ r200 camera (world facing). Available at: `https://software.intel.com/en-us/articles/realsense-r200-camera`, Accessed on 15 January 2016.

Jégou, H., M. Douze, C. Schmid, and P. Pérez (2010). Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3304–3311. IEEE.

Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Krishnan, P. and S. Naveen (2015). Rgb-d face recognition system verification using kinect and frav3d databases. *Procedia Computer Science 46*, 1653–1660.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105.

Kussul, E., T. Baidyk, C. Conde, I. Martin de Diego, and E. Cabello (2013). Face recognition improvement with distortions of images in training set. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pp. 1–6. IEEE.

Lachat, E., H. Macher, M. Mittet, T. Landes, and P. Grussenmeyer (2015). First experiences with kinect v2 sensor for close range 3d modelling. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS).*

Le, A. V., S.-W. Jung, and C. S. Won (2014). Directional joint bilateral filter for depth images. *Sensors 14*(7), 11362–11378.

Li, B. Y., A. Mian, W. Liu, and A. Krishna (2013). Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pp. 186–192. IEEE.

Lu, C.-Y., H. Min, J. Gui, L. Zhu, and Y.-K. Lei (2013). Face recognition via weighted sparse representation. *Journal of Visual Communication and Image Representation 24*(2), 111–116.

Lu, X., R.-L. Hsu, A. K. Jain, B. Kamgar-Parsi, and B. Kamgar-Parsi (2004). Face recognition with 3d model-based synthesis. In *Biometric Authentication*, pp. 139–146. Springer.

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

Mian, A. S., M. Bennamoun, and R. Owens (2008). Keypoint detection and local feature matching for textured 3d face recognition. *International Journal of Computer Vision 79*(1), 1–12.

Microsoft (2010). Kinect for windows sensor components and specifications. Available at: `https://msdn.microsoft.com/en-us/library/jj131033.aspx`, Accessed on 15 January 2016.

Microsoft (2014). Kinect for xbox one. Available at: `http://www.xbox.com/en-us/xbox-one/accessories/kinect-for-xbox-one#fbid=AN5dWEs2zxt`, Accessed on 15 January 2016.

Min, R., J. Choi, G. Medioni, and J.-L. Dugelay (2012). Real-time 3d face identification from a depth camera. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 1739–1742. IEEE.

Min, R., N. Kose, and J.-L. Dugelay (2014). Kinectfacedb: A kinect database for face recognition. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on 44*(11), 1534–1548.

Minolta, K. (2006). Konica minolta 3d laser scanners. Available at: `http://www.3dscanco.com/products/3d-scanners/3d-laser-scanners/konica-minolta/`, Accessed on 15 January 2016.

Mohammadzade, H. and D. Hatzinakos (2013). Iterative closest normal point for 3d face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 35*(2), 381–397.

Monteiro, J. C. and J. S. Cardoso (2015). A cognitively-motivated framework for partial face recognition in unconstrained scenarios. *Sensors 15*(1), 1903–1924.

Monteiro, J. C., J. S. Cardoso, and T. S. Freitas (2016). Multimodal hierarchical face recognition using information from 2.5d images. *U.Porto Journal of Engineering*.

Moreno, A. and A. Sanchez (2004). Gavabdb: a 3d face database. In *Proc. 2nd COST275 Workshop on Biometrics on the Internet, Vigo (Spain)*, pp. 75–80.

Moreno, A. B., A. Sánchez, J. F. Vélez, and F. J. Díaz (2003). Face recognition using 3d surface-extracted descriptors. In *Irish Machine Vision and Image Processing Conference*, Volume 2003. Citeseer.

Mracek, S., M. Drahansky, R. Dvorák, I. Provaznik, and J. Vána (2014). 3d face recognition on low-cost depth sensors. In *Biometrics Special Interest Group (BIOSIG), 2014 International Conference of the*, pp. 1–4. IEEE.

Naveen, S. and R. Moni (2015). A robust novel method for face recognition from 2d depth images using dwt and dct fusion. *Procedia Computer Science 46*, 1518–1528.

Naveen, S., S. S. Nair, and R. Moni (2015). 3d face recognition using optimised directional faces and fourier transform. In *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, pp. 1856–1861. IEEE.

of Applied Sciences of the Royal Military Academy, F. (n.d.). 3d_rma : 3d database. Available at: `http://www.sic.rma.ac.be/\simbeumier/DB/3d_rma.html`, Accessed on 15 January 2016.

of York, T. U. (n.d.). The 3d face database. Available at: `https://www-users.cs.york.ac.uk/~nep/research/3Dface/tomh/3DFaceDatabase.html`, Accessed on 15 January 2016.

Papatheodorou, T. and D. Rueckert (2004). Evaluation of automatic 4d face recognition using surface and texture registration. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pp. 321–326. IEEE.

Parkhi, O. M., A. Vedaldi, and A. Zisserman (2015). Deep face recognition. *Proceedings of the British Machine Vision 1*(3), 6.

Phillips, P. J., P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek (2005). Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, Volume 1, pp. 947–954. IEEE.

Razavian, A., H. Azizpour, J. Sullivan, and S. Carlsson (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813.

Reynolds, D. A., T. F. Quatieri, and R. B. Dunn (2000). Speaker verification using adapted gaussian mixture models. *Digital signal processing 10*(1), 19–41.

Sang, G., J. Li, and Q. Zhao (2015). Pose-invariant face recognition via rgb-d images. *Computational Intelligence and Neuroscience 2016*.

Savran, A., N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun (2008). Bosphorus database for 3d face analysis. In *Biometrics and Identity Management*, pp. 47–56. Springer.

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

SofKinetic (2007). Sofkinetic depthsense® 325. Available at: `http://www.softkinetic.com/Store/ProductID/6`, Accessed on 15 January 2016.

Song, S., S. P. Lichtenberg, and J. Xiao (2015). Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576.

Structure (2013). Structure sensor st01. Available at: `http://structure.io/embedded`, Accessed on 15 January 2016.

Sun, Y., D. Liang, X. Wang, and X. Tang (2015). Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*.

Taigman, Y., M. Yang, M. Ranzato, and L. Wolf (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708.

Tanaka, H. T., M. Ikeda, and H. Chiaki (1998). Curvature-based face surface recognition using spherical correlation. principal directions for curved object recognition. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 372–377. IEEE.

Tang, Y., X. Sun, D. Huang, J.-M. Morvan, Y. Wang, and L. Chen (2015). 3d face recognition with asymptotic cones based principal curvatures. In *Biometrics (ICB), 2015 International Conference on*, pp. 466–472. IEEE.

Tepper, M. and G. Sapiro (2012). L1 splines for robust, simple, and fast smoothing of grid data. *arXiv preprint arXiv:1208.2292*.

Tomasi, C. and R. Manduchi (1998). Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pp. 839–846. IEEE.

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

Tsalakanidou, F., D. Tzovaras, and M. G. Strintzis (2003). Use of depth and colour eigenfaces for face recognition. *Pattern Recognition Letters 24*(9), 1427–1435.

Tsao, D. Y. and M. S. Livingstone (2008). Mechanisms of face perception. *Annual review of neuroscience 31*, 411.

Vijayanagar, K. R., M. Loghman, and J. Kim (2014). Real-time refinement of kinect depth maps using multi-resolution anisotropic diffusion. *Mobile Networks and Applications 19*(3), 414–425.

Wang, Y., G. Pan, Z. Wu, and S. Han (2004). Sphere-spin-image: A viewpoint-invariant surface representation for 3d face recognition. In *Computational Science-ICCS 2004*, pp. 427–434. Springer.

Wright, J., A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma (2009). Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 31*(2), 210–227.

Yin, L., X. Chen, Y. Sun, T. Worm, and M. Reale (2008). A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference On*, pp. 1–6. IEEE.

Yin, L., X. Wei, Y. Sun, J. Wang, and M. J. Rosato (2006). A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pp. 211–216. IEEE.

Zhang, B., Y. Gao, S. Zhao, and J. Liu (2010). Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. *Image Processing, IEEE Transactions on 19*(2), 533–544.

Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pp. 474–485. Academic Press Professional, Inc.