UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

3D FACE RECOGNITION WITH DESCRIPTOR IMAGES AND SHALLOW CONVOLUTIONAL NEURAL NETWORKS

JOÃO BAPTISTA CARDIA NETO

ADVISOR: PROF. DR. APARECIDO NILCEU MARANA
CO-ADVISOR: PROF. DR. STEFANO BERRETTI

São Carlos – SP

November/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

3D FACE RECOGNITION WITH DESCRIPTOR IMAGES AND SHALLOW CONVOLUTIONAL NEURAL NETWORKS

JOÃO BAPTISTA CARDIA NETO

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação, área de concentração: Visão Computacional.
Advisor: Prof. Dr. Aparecido Nilceu Marana
Co-Advisor: Prof. Dr. Stefano Berretti

São Carlos – SP

November/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Tese de Doutorado do candidato João Baptista Cardia Neto, realizada em 05/11/2020.

Comissão Julgadora:

Prof. Dr. Aparecido Nilceu Marana (UNESP)

Prof. Dr. João Paulo Papa (UFSCar)

Prof. Dr. Alexandre Luis Magalhães Levada (UFSCar)

Prof. Dr. Agma Juci Machado Traina (USP)

Prof. Dr. Daniel Carlos Guimarães Pedronette (UNESP)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

ACKNOWLEDGMENTS

To UFSCAR for the opportunity to develop my studies.

To CAPES for the sandwich scholarship which allowed me to develop part of my thesis at the University of Florence.

To Dr. Aparecido Nilceu Marana for the guidance, patience, availability, and for helping me to be able to develop my research.

To Dr. Stefano Berretti who welcomed me in Italy and contributed immensely to the development of this thesis.

To the Media Integration and Communication Center (MICC) which helped me to get a better grasp at the most relevant points on my work.

To all the people I have met at MICC, who helped immensely with their knowledge.

To Dr. Claudio Ferrari who gave me tons of insights and helped with some critical key points during my time at Florence.

To the University of Florence for welcoming me during my sandwich period.

To the São Paulo State Technological College (FATEC) for giving me the opportunity to develop the present work.

To NVidia which, with its GPU grant program, gave me the opportunity to run the needed experiments.

RESUMO

A crescente necessidade de sistemas que possam identificar uma pessoa com precisão e rapidez se torna muito evidente nos dias de hoje. Existem algumas aplicações em que a necessidade de descobrir a identidade das pessoas de forma sigilosa é primordial. Pensando nessas aplicações e na utilização de características biométricas a face é uma das características que melhor se adequa á esse tipo de identificação. Isso pois a tecnologia atual é capaz de fornecer imagens faciais 2D de alta resolução capturadas por câmeras de baixo custo a distância e sem a cooperação dos sujeitos. No entanto, em geral, os sistemas biométricos baseados no reconhecimento de face 2D têm sua performance afetada em certos cenários, quando as imagens das faces apresentam variações na pose, iluminação e expressões faciais. Uma maneira de atenuar esse problema é usar dados faciais em 3D, mas os scanners 3D atuais são caros e exigem muita cooperação dos sujeitos. O uso de redes neurais convolucionais profundas é outra forma de mitigar as desvantagens do reconhecimento facial 2D tradicional, mas pode ser inviável, devido à necessidade de grandes conjuntos de dados rotulados para o treinamento das redes e computadores com enorme capacidade de processamento e armazenamento de dados. Portanto, nesta tese, uma abordagem híbrida para reconhecimento de faces 3D é apresentada. Essa abordagem, que tem como foco minimizar a quantidade de dados, o poder computacional e o tempo de processamento necessário na fase de treinamento, é baseada em redes neurais convolucionais rasas e é capaz operar próximo aos métodos do estado da arte e ser capaz de transferir o aprendizado feito em dados de alta resolução para dados de baixa resolução. Outro aspecto importante da abordagem híbrida proposta é a possibilidade de operar nos modos de classificação ou extração de características. Os resultados experimentais obtidos por nossa abordagem híbrida utilizando o dataset EURECOM Kinect Face, com dados de profundidade de baixa resolução, mostraram uma taxa de reconhecimento em *rank-1* de 90,75 % no caso mais difícil do modo de classificação e 73,26 % no modo de extração de características, sendo o desempenho melhor que outras técnicas utilizando o mesmo protocolo e conjunto de dados. Assim, concluímos que a abordagem híbrida proposta ajuda a atenuar as diferenças de resolução e que a utilização de uma entrada construída com dados mais discriminativos, como extratores de característica de baixo nível, permite a utilização de CNN rasas para reconhecimento facial 3D.

Palavras-chave: Biometria, reconhecimento de faces 3D, CNN rasas, 3DLBP, Sigmoid 3DLBP, Descriptor Image, Shallow Learned Feature Representation (SLFR).

ABSTRACT

Nowadays, there is an increasing need for systems that can accurately and quickly identify a person. Traditional identification methods utilize something a person knows or something a person has. This kind of methods has several drawbacks, being the main one the fact that it is impossible to detect an impostor who uses genuine credentials to pass as a genuine person. Besides, in some cases it is necessary to discover the identity of people in a covert manner. One way to deal with these types of problems is to use biometric identification. Face is one of the biometric features that best suit the covert identification since the current technology is able to provide high resolution 2D face images captured by low cost cameras, in a secret way, at a distance and without cooperation from the people being identified. However, in general, biometric systems based on 2D face recognition perform very poorly in certain scenarios when the input images present variations in pose, illumination and facial expressions. One way to mitigate this problem is to use 3D face data, but the current 3D scanners are expensive and require a lot of cooperation from people. The use of deep convolutional neural networks is another way to mitigate the traditional 2D facial recognition drawbacks, but it can be unfeasible, due to their large training data and huge computational power requirements. Therefore, in this thesis, we introduce a hybrid approach, based on Shallow Learned Feature Representation, for 3D face recognition, which is focused on minimizing the amount of data, the computational power and the processing time required in the training stage, while being able to operate close to state-of-the-art methods and being able to transfer the learning made on high-resolution data to low-resolution data. Another important aspect of the proposed hybrid approach is the possibility to operate in both classification or feature-extraction modes. Experimental results obtained by our hybrid approach on EURECOM Kinect Face dataset, a low resolution depth dataset, showed a rank-1 recognition rate of 90.75% on the hardest case of classification mode, and 73.26% on the feature extraction mode, which are better than the rates obtained by related state-of-the-art methods with the same protocol and dataset. So, we conclude that the proposed hybrid approach helps to attenuate the cross-resolution differences and that the utilization of an input built with more discriminative data, such as low-level hand-crafted features, allows the utilization of shallow CNN for 3D face recognition.

Keywords: Biometrics, 3D face recognition, Shallow CNN, 3DLBP, Sigmoid 3DLBP, Descriptor Image, Shallow Learned Feature Representation (SLFR).

LIST OF FIGURES

LIST OF TABLES

CONTENTS

Chapter 1

INTRODUCTION

Nowadays, the necessity for assuring the identity of a person has become one of the most important features in our everyday life. Traditional methods proposed for people identification are based on knowledge (such as a password) or possessions (such as a key-card). However, knowledge can be learned or guessed and possessions can be stolen or lost. Besides, identification systems based on knowledge and possessions are not able to distinguish between an impostor person utilizing genuine credentials from a genuine person (Jain; Ross; Prabhakar, 2004).

The several drawbacks of the traditional identification methods have stimulated the research on biometric identification methods, which are based on biological or behavioral traits of the individuals. Because biometric identification systems use something that persons are, they are more difficult to circumvent (Prabhakar; Pankanti; Jain, 2003).

In surveillance systems, the utilization of biometric characteristics can drastically improve the system performance and, among several other characteristics, face has several advantages (Nguyen et al., 2018). Face is the biometric feature that best suits the covert identification, since the current technology is able to provide high resolution 2D face images captured by low cost cameras, in a secret way, at a distance and without cooperation from the people being identified (Nguyen et al., 2018). However, in general, biometric systems based on 2D face recognition that utilizes hand-crafted features perform poorly in unconstrained environments, common in covert identification scenarios, since the input images present variations in pose, illumination and facial expressions.

The current state-of-the-art has pushed deep learning approaches as alternatives for solving complex pattern recognition problems and have reached almost perfect results in many tasks due to their robustness and great power of abstraction, working with abstract and high-level features, self-learned from training data (Souza et al., 2017). Among the proposed deep learn-

ing architectures, Convolutional Neural Networks (CNN) (Krizhevsky; Sutskever; Hinton, 2012) emerged as one of the most important classes of neural networks able to deal, with great performances, with tasks involving the processing and analysis of two-dimensional signals (usually images). A Convolutional Neural Network consists of several sets of layers containing one or more planes. The input is made of images centered and normalized. Each small local region from this serves as an input to a unit in a plane of the next layer. Each of those layers has a fixed feature detector that is convolved utilizing a local window on the output from the previous layers. These layers are known as convolutional layers (Krizhevsky; Sutskever; Hinton, 2012)

Regarding 2D face recognition, there are several CNN-based approaches with state-of-the-art results. In (Parkhi; Vedaldi; Zisserman, 2015) the experiment carried out on the Labeled Faces in the Wild dataset (Learned-miller et al., 2016) reached 98.91% of rank-1 recognition rate when using 2.6M faces for training, and another experiment carried out on the Youtube Faces Dataset (Wolf; Hassner; Maoz, 2011) obtained 97.3% of rank-1 recognition rate. In (Qian; Deng; Hu, 2019), the experiments carried out on the IJB-A dataset (Klare et al., 2015) reached 96% of rank-1 recognition rate and, in the easiest case, 99.9% of rank-1 recognition rate for the Multi-PIE dataset (Gross et al., 2010). But, when dealing with pose variations bigger than 60 degrees there is a significant loss in performance, from 99% to 81% of rank-1 recognition rate.

In the previously referred works, it is possible to see that the amount of data utilized for training is a crucial part of deep approaches. Both works utilized millions of face images for training, and, for many applications, there may not be enough available data. Other problems with deep approaches are related to the expensive hardware needed to train the deep neural networks, which must provide a huge storage capacity and a very high processing power.

In order to utilize 3D models for face recognition there is an initial problem that needs to be solved. Traditional scanners are expensive and they need a lot of cooperation from people being identified. This last aspect is an obstacle when the goal is to develop a covert 3D face recognition systems, since in those cases it is not possible to account for user cooperation.

In the case of building a system for face recognition that does not need too much user cooperation, the utilization of sensors such as Microsoft Kinect can be a feasible alternative. The main problem with this kind of sensor is when it must be used outdoor. The 3D depth sensor of the Kinect v1 has two components, an infrared emitter and an infrared camera, which estimate the depth data by projecting an array and measuring the distortion caused by the reflected rays captured by the camera. Because of the utilization of infrared, the data is heavily affected by sunlight. Another problem is the maximum distance in which the sensor can generate depth data, even with the Kinect v2 the maximum distance is four meters. Analogous to

2D face recognition, one possibility to increase the performance of 3D face recognition is to utilize Convolutional Neural Networks (CNN). Due to the high abstraction level in the higher CNN layers, it is expected that deep approaches can better generalize data and achieve higher recognition rates.

In order to utilize CNNs, specially deep ones, for face recognition there is a need for a great amount of data. Thinking of building a system that utilizes Kinect data can be troublesome in this context since most of large datasets utilize high-resolution scans. With this in mind, it would be necessary to train the CNN with high-resolution data and use it on low-resolution data, this process is known as cross resolution face recognition (Singh et al., 2018).

In this thesis, we propose a new hybrid approach for 3D face recognition. Our approach utilizes a shallow CNN in conjunction with Descriptor Images (DIs) originated from the depth data. The idea behind this is that, since the DIs are constructed from low-level feature extractors it would be possible to build an ensemble of shallow networks that is modular and needs fewer data and less computational power for training and learning.

We evaluated our approach with two different modes, feature extraction and classification. The classification mode is focused on learning the identities for the known subjects in the dataset while the feature extraction mode is focused on learning the relevant features that are portrait in the dataset. The first mode is more adequate to closed-set scenarios, while the last one for open-set scenarios.

1.1 Hypothesis

This thesis hypothesizes that it is possible to utilize Descriptor Images (DIs) originated from the depth data and shallow CNN to do 3D face recognition, while performing close to the state-of-the-art results obtained by deep CNN-based methods. It is expected that the DIs, which have some low-level features encoded, will allow to utilize shallower CNNs. Utilizing shallow networks is important because it can reduce the amount of processing time and computational power to train the networks and can make feasible the deployment of learning-based biometric systems based on 3D face recognition.

To deal with the problem of the amount of low-resolution data available to train a CNN approach, we hypothesize that utilizing data that has some degree of resolution invariance can help to attenuate this problem since it would be possible to train the CNN with high-resolution data and utilize low-resolution data for 3D face recognition. We believe that our DI has a great deal of resolution invariance and, due to this reason, can be helpful for this particular problem.

1.2 Objectives

The general objective of this thesis is to propose a novel hybrid approach for 3D face recognition that utilizes Descriptor Images (DI), originated from facial cloud points, in conjunction with shallow convolutional neural networks (CNNs). The idea behind this is that, since the DIs are constructed from low-level feature extractors it would be possible to use shallow CNNs, allowing us to build an ensemble of shallow networks that are modular and needs fewer data and less computational power to train the CNN. Our approach can be used for 3D face classification or 3D face feature extraction.

Some specific objectives of this thesis are:

- Propose a new representation for 3D face data, the Descriptor Image, based on a set of low-level features, to be used as input for the shallow CNNs;
- Assess the best shallow CNN architecture to be used, in terms of number of convolutional layers;
- Propose the best ensemble configuration of the shallow CNNs, in order to create more robust 3D face recognition systems;
- Assess the performance of the two modes of operation for our hybrid approach: feature extraction mode and classification mode;
- Assess the effectiveness of the cross-resolution transfer learning in the proposed hybrid approach, based on the shallow CNN when utilizing the proposed Descriptor Images.

1.3 Contribution

This thesis brings the following contributions:

Descriptor Images for 3D face representation: The Descriptor Images (DIs) are a way to encode low-level 3D features into a 2D representation, and to train shallow CNNs for 3D face recognition. Our contribution is in the methodology to generate this type of descriptor, we encode each of the information generated by a hand-crafted feature into different channels, generating 2D projection of the 3D features;

Shallow CNNs for 3D face recognition: As the proposed Descriptor Images (DIs) encode information similarly to the early layers of a CNN, the use of DIs as input allows the utilization of shallow networks for 3D face recognition. Shallow CNNs can be trained much faster than deep CNNs, in much simpler and cheaper hardware, with much less training data;

Transfer Learning from high to low resolution: The 3D face representation by Descriptor Images allows the use of high resolution 3D face data to train a shallow CNN, which can be used later as classifier or as feature extractor for low resolution 3D face recognition, transferring learning from high resolution to low resolution;

Shallow Learned Feature Representation (SLFR): We have proposed the use of shallow CNNs with the Descriptor Images for 3D face recognition in two modes: classification and feature extraction. In the feature extraction mode, we extracted face features from the pre-trained shallow CNN, that we named Shallow Learned Feature Representation (SLFR). SLFR allows learning a feature representation based on different types of DIs. Due to the shallow characteristic, it is possible to fuse different SLFR learned from different DIs and, thus, increase the robustness of the hybrid approach proposed for 3D face recognition.

1.4 Thesis Organization

In addition to this introductory chapter, this thesis has other six chapters:

Chapter 2: Presents the main concepts of Biometrics addressed in this thesis;

Chapter 3: Presents the theoretical background related to the this thesis;

Chapter 4: Presents the new proposed hybrid approach for 3D face recognition;

Chapter 5: Describes the material and the methodology adopted to assess the proposed approach;

Chapter 6: Shows the results obtained with the experiments;

Chapter 7: Presents a discussion and the conclusion of this thesis.

Chapter 2

BIOMETRICS

In this chapter it is discussed what is biometric identification, the benefits of Biometrics over other forms of assuring a person identity, how a biometric system works, and the different types of biometric characteristics. Prior to discussing what biometric identification is and how it can make systems more secure, the topic of personal identification is presented. Understanding such topics is crucial to know the flaws of traditional methods of personal identification and why it is important to research and build biometric systems.

2.1 Personal Identification

The concept of personal identification is very simple. It is the act of linking a person with an identity. It can be categorized into two types: verification and identification (Bolle; Pankanti, 1998). Verification is about refuting or confirming a person's claimed identity, while identification is about establishing a person's identity. When utilizing a set of already known identities to perform this task it is called a closed identification problem, otherwise, it is called an open identification problem (Bolle; Pankanti, 1998).

A traditional way of person identification is utilizing some sort of document or knowledge that the person is supposed to possess or to know (Prabhakar; Pankanti; Jain, 2003). The identification based on possession or knowledge can face several problems. A stolen ID Card can lead to a criminal to be wrongly identified as the true owner of that document and, consequently, gaining access to private information or place. This shows that this traditional form of personal identification is not secure, as the holder of an identification document is not necessarily its owner. Even when combining something the person has (a credit card, for example) with something the person knows (a password, for example), traditional methods of identification are fragile and susceptible to failure. With the use of biometric characteristics, it is possible to

mitigate problems such as those aforementioned.

2.2 Biometrics Characteristics

Biometrics is the automatic recognition of a person identity based on its physical/physiological or behavioral characteristics (Prabhakar; Pankanti; Jain, 2003). Physical and physiological characteristics refer, respectively, to the person body (e.g. fingerprint, iris, face) or to his/her living functions (e.g. hand vein, hand and facial thermogram). Behavioral characteristics, on the other hand, are those traits that can identify a person based on some particular way of doing some kind of task (e.g. gait, keystroke dynamics, signature, voice). In the Figure 2.1 a few examples of physical, physiological and behavioral characteristics are shown.

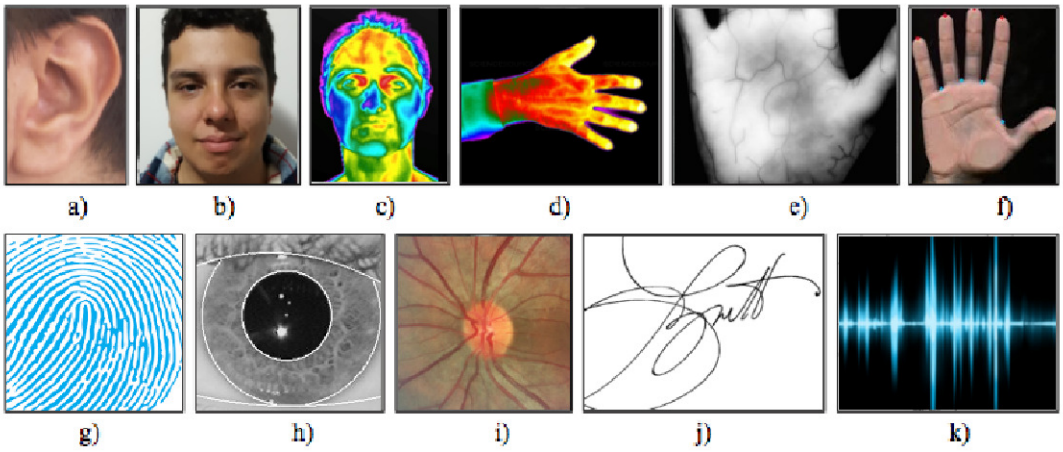


Figure 2.1: Examples of biometrics characteristics: a) ear, b) face, c) facial thermogram, d) hand thermogram, e) hand vein, f) hand geometry, g) fingerprint, h) iris, i) retina, j) signature, and k) voice. Adapted from (Maltoni et al., 2009).

It is possible to use any human characteristic for biometric recognition since it satisfies a few requirements. Maltoni et al. (2009) define them as:

- Universality: Most people need to possess that characteristic;
- Distinctiveness: That characteristic needs to be distinct for distinct persons;
- Permanence: That characteristic should not change over time;
- Collectability: That characteristic must be quantitatively measured;
- Performance: That characteristic must provide high accuracy and robustness while providing low processing time and low computational costs;

- Acceptability: That characteristic must be accepted culturally and socially by the people to be identified;
- Circumvention: That characteristic must be difficult to circumvent.

Table 2.1 shows a comparison between some biometric characteristics in respect to these requirements (Jain; Ross; Prabhakar, 2004).

It is important to emphasize that there is no optimal biometric characteristic. When deciding which biometric trait to utilize, it is essential to take into account the environmental constraints and the characteristics of the group of people to be identified. Due to its characteristic, face is one of the best choice in an application that needs to capture the biometric sample in a covert manner, the same cannot be said for the fingerprint or iris, for instance. This does not mean that one characteristic is superior to the others, but just that it works best in a specific application. Even though security concerns are important, other issues must be taken into account when choosing a biometric feature, such as the reasonableness of the required resources, the danger it can cause to users, the acceptance by the target population, privacy and robustness to fraud. (Maltoni et al., 2009).

Table 2.1: Comparison between biometric characteristics (Jain; Ross; Prabhakar, 2004).

Biometric Trait	Universality	Distinctiveness	Permanence	Collectability	Performance	Acceptability	Circumvention
Face	High	Low	Medium	High	Low	High	Low
Fingerprint	Medium	High	High	Medium	High	Medium	Medium
Iris	High	High	High	Medium	High	Low	High
Hand Geometry	Medium	Medium	Medium	High	Medium	Medium	Medium
Ear	Medium	Medium	High	Medium	Medium	High	Medium
Hand Vein	Medium	Medium	Medium	Medium	Medium	Medium	High
Odor	High	High	High	Low	Low	Medium	High
DNA	High	High	High	Low	High	Low	High
Facial Thermogram	High	High	Low	High	Medium	High	High
Retina	High	High	Medium	Low	High	Low	High
Signature	Low	Low	Low	High	Low	High	Low
Voice	Medium	Low	Low	Medium	Low	High	Low
Gait	Medium	Low	Low	High	Low	High	Medium
Keystroke	Medium	Medium	Low	Medium	Low	Medium	Medium

Analyzing the Table 2.1 one can verify that behavioral characteristics (signature, voice, gait and keystroke) are more susceptible to fraud than most physical and physiological characteristics. This happens because it is easier to mimic the human behavior than its physical or physiological traits.

Another important thing to notice is that face has higher universality than the fingerprint, higher collectability than iris, and higher acceptability than fingerprint and iris. On the other hand, face is more susceptible to fraud, has lower performance, permanence, and distinctiveness than the aforementioned characteristics.

Since it is easy and cheap to capture high-quality face images with the current technology, face stands out when discretion or little user collaboration is required in a biometric identification system.

2.3 Biometric Systems

A pattern recognition system that utilizes a feature vector based on any biometric characteristic to assure an individual identity is a biometric system. A Biometric system will normally operate in one of two modes: identification or verification (Prabhakar; Pankanti; Jain, 2003).

Verification mode is when the user claims to be a certain person and the system compares the biometric sample probed with the user's template stored in the database. In this mode, it is possible to answer the question "Is this person who she or he claims to be?". Identification mode is when given a biometric sample the system searches through all the templates stored in the database trying to find a match. In this mode, it is possible to answer the question "Who is this person?". These two distinct modes are typically used for positive or negative person recognition, respectively (Prabhakar; Pankanti; Jain, 2003).

Before recognizing an individual, it is necessary to enroll him/her in the gallery. The enrollment starts with the subject providing a biometric sample and the system extracting and compressing that data into a template. It is important to assure that the captured sample has an acceptable quality, then the system checks for it and, if necessary, asks for another sample. The template can be stored in a central database or in a type of removable media (e.g. flash drive).

In verification mode, the user provides the biometric sample and claims to be an individual. The system then recovers the template from the claimed user and compares both samples. If there is a match, the system validates the individual identity. The method for feature extraction has to be the same as in the enrollment stage.

In identification mode, the individual only needs to provide his/her biometric data, the system will be responsible for finding the associated identity. The biometric is processed and compared with all the templates in the database. The system will indicate an identity that is most similar to the sample or will indicate that there is no such individual in the database. In

Figure 2.2, the three stages of a biometric system (enrollment, verification and identification) are illustrated.

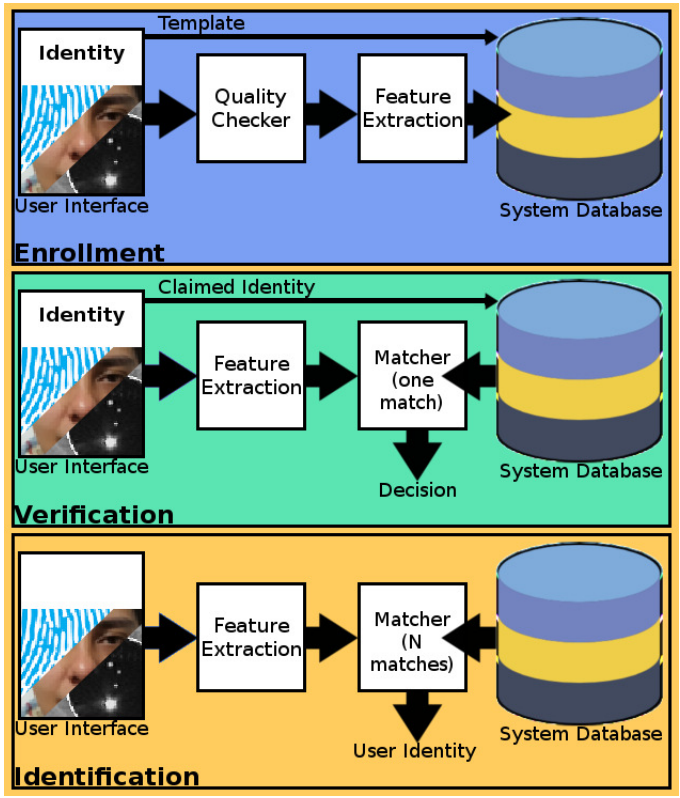


Figure 2.2: Enrollment, verification and identification stages of a biometric system (Prabhakar; Pankanti; Jain, 2003).

A biometric system can be classified into one of the seven categories defined by Wayman (2002):

- Cooperative versus Non-cooperative: Does the user wish to be identified?
- Overt versus Covert: Does the user knows that he or she is being identified?
- Habituated versus Non-habituated: Does the user often submits to the identification?
- Attended versus Non-attended: There is a human operator helping the system?
- Standard environment: What is the environment that the system will operate?
- Public versus private: The users are employees (private) or clients (public)?
- Closed versus open: Is the system utilizing a set of already known identities to perform the task?

Due to noise, environmental conditions, changes in the person traits, and even how the user interacts with the sensor, two samples of a biometric characteristic from the same person will never be the same. Therefore, the system will have a matching score and this score will determine if the comparison succeeds or fails. To do so, a threshold t must be set to regulate the system decision. If the score is greater or equal to t then the characteristics belong to the same individual. Otherwise, they do not.

Due to their nature, biometric systems can present two types of errors (Prabhakar; Pankanti; Jain, 2003):

- False match: A biometric from two different persons are considered to come from the same person;
- False non-match: A genuine biometric comparison is taken as an impostor.

In order to work properly, a biometric system must make a trade-off between these two error rates. If the system designer decides to decrease the FMR (false match rate) then the FNMR (false non-match rate) will increase. On the other hand, if the designer tries to facilitate the user login, the FMR will increase. Figure 2.3 shows the correlation between these two error rates.

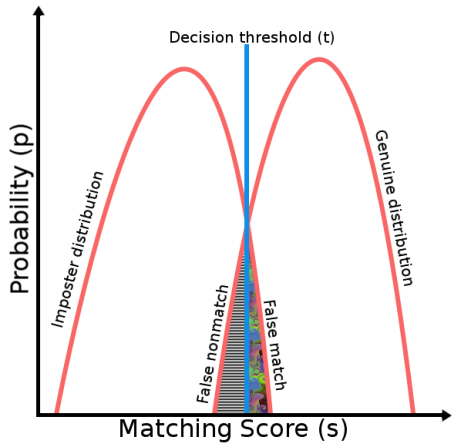


Figure 2.3: Biometric system error rates: FMR (False Match Rate) and FNMR (False Non-Match Rate) (Prabhakar; Pankanti; Jain, 2003).

Other error rates are FAR (False Accept Rate) and FRR (False Rejection Rate), they are analogous to FMR and FNMR. FAR is the rate at which a false subject is accepted as genuine and FRR is the rate in which a genuine match is categorized as an impostor.

Given both error rates, FAR and FRR, it is possible to calculate the Equal Error Rate (EER), which is a security level measurement that identifies a threshold when FAR and FRR have the same value.

While FMR and FNMR are intrinsic system errors, there are others that can happen due to conditions that cannot be controlled: the failure to capture (FTC) and fail to enroll (FTE) (Prabhakar; Pankanti; Jain, 2003).

Since a biometric system has to make a trade-off between two errors rates (FNMR x FMR), it is not possible to assess its performance with a single number. In order to evaluate this kind of system, a performance curve is necessary (Martin et al., 1997). The Receiver Operating Characteristics (ROC) curve is one way to understand the performance of a biometric system. A ROC curve is a two-dimensional graph where the true positive rate is plotted on the Y-axis, while the false positive rate is plotted on the X axis (Fawcett, 2006). In a ROC curve for evaluating biometric system, usually, the Genuine Accept Rate (GAR) is plotted against the False Accept Rate (FAR) (Jain; Ross; Nandakumar, 2011).

For forensic applications the FNMR is more important than the FMR, because, normally, this kind of application deals with criminal identification and it is very important not letting a suspect pass unidentified, even if it is needed to manually select the right identity from a group of matched subjects. For high-security applications occurs the opposite, a person wrongly identified cannot gain access to its own system. For civilian applications, a balance between the two rates is the desired scenario.

Another way to assess the performance of a biometric system is to utilize the Cumulative Match Characteristic (CMC) Curve. To build a CMC Curve it is necessary to order the matching score sets from a probe image to know identities in the gallery, with this it is possible to calculate

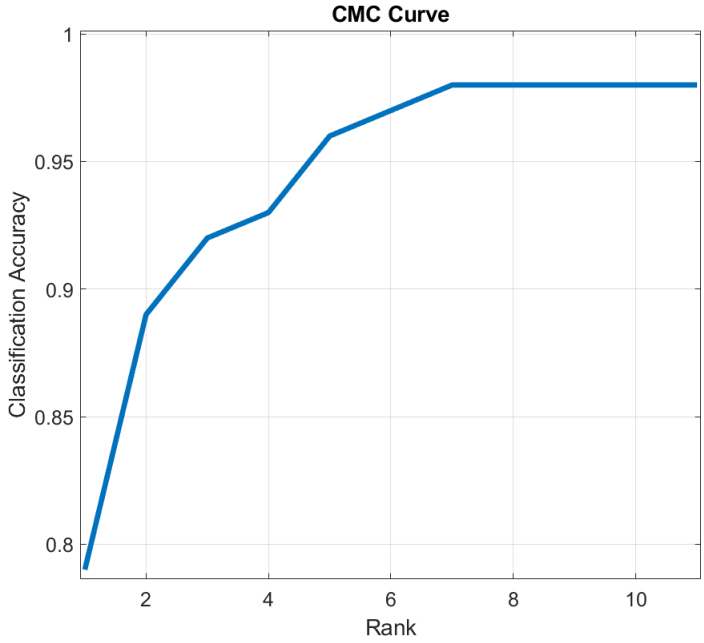


Figure 2.4: An example of a CMC Curve.

the probability in which the right subject appears in the first K positions in the ordered matching score set. The CMC Curve is a visual representation of the top K ranks for a biometric system, with $K \leq N$ and N being the number of identities. The CMC Curve is a *rank-based* metric whilst the ROC Curve is a *aggregate-based* metric (Decann; Ross, 2013). Figure 2.4 shows an example of a CMC Curve.

It is also important to understand the relation between the ROC and CMC Curve. Spite the CMC Curve not having a threshold the way that the matching is performed can relate to the FAR and FRR. This relation is given by understanding that for every genuine score search the value of the imposter score can be considered as a virtual threshold. With this in mind it is possible to construct a CMC Curve given the impostor and genuine distribution in respect to a $1:1$ matcher (Bolle et al., 2005). This means that, if the CMC Curve points towards a good performance it is reasonable to expect that the ROC Curve will have a similar behavior.

Chapter 3

THEORETICAL BACKGROUND AND RELATED WORKS

In this chapter, the theoretical background needed for the current research and some related works are presented.

3.1 Theoretical Background

First, the main concepts of artificial and convolutional neural networks are presented. After, the transfer learning and cross-resolution concepts are defined. Finally, the 3DLBP hand-crafted feature is explained in details, since we utilize it to propose a new Descriptor Image (DI) for 3D face representation and description.

3.1.1 Artificial Neural Networks

Inspired by the human brain, artificial neural networks (ANN) are defined as a collection of units that are connected together. This connection can be made in different ways, forming different structures, such as the feed-forward networks and the recurrent networks. The properties of an ANN are defined by its units and topology (Russell, 2010).

Artificial neural networks have attracted interest due to their ability to learn, their distributed computing and their robustness in handling noisy input data (Russell, 2010). To understand how to create and utilize an ANN, it is necessary to study its units and how it is possible to propagate its activation pattern throughout the whole network. Figure 3.1 exhibits the mathematical model of an ANN unit, referred to as neuron in this context.

To form a neural network, each neuron has to be connected by directed links. The connec-

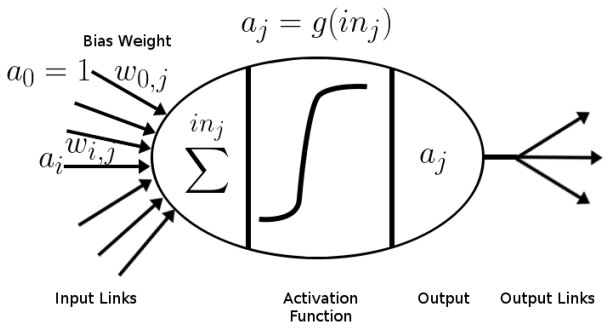


Figure 3.1: A mathematical model of a neuron (Russell, 2010).

tion of a neuron i to a neuron j is necessary to propagate the activation of the units. Each input link has a weight $w_{i,j}$ that controls its degree of influence in the neuron activation. All neurons have a dummy input a_0 that has always the value of 1 and the weight $w_{0,j}$ (Russell, 2010). To calculate the input value of a neuron j, the following equation is utilized (Russell, 2010):

$$in_j = \sum_{i=0}^{n} w_{i,j} a_i \tag{3.1}$$

being in_j the input to the activation function of unit j, n the number of input links, $w_{i,j}$ the weight on the link from unit i to unit j, and a_i the output activation of unit i. The activation function g is used to calculate the output of a neuron (Russell, 2010):

$$a_j = g(in_j) = g\left(\sum_{i=0}^{n} w_{i,j} a_i\right) \tag{3.2}$$

being a_j the output from the unit j.

The activation function can be a hard threshold, in which the neuron is known as perceptron, a logistic function, or a rectified linear unit (ReLU) (Nair; Hinton, 2010).

The ANN learning procedure consists of updating the weights of the network based on its error rate to minimize loss. Since most of the problems in this scenario are non-linear, it is necessary to use the gradient descent to find the best values for several weights (Russell, 2010).

Gradient descent deals with the general problem of optimizing a loss function knowing only its first-order gradient evaluation (Wu; Ward; Bottou, 2018). The equation for calculating the new value for a weight with gradient descent is:

$$w_{j+1} = w_j - \eta_j \nabla f(w_j) \tag{3.3}$$

with η being the learning rate and f the loss function (Wu; Ward; Bottou, 2018). The learning rate

is the size of the step that the gradient does in the weight space. A loss function is defined as the size of the loss when predicting wrongly a result (Russell, 2010).

An ANN can be connected to a feed-forward or recurrent network. A feed-forward network does not have an internal state, it only has a representation of its current input. This behavior is due to the characteristic of not feeding its output back to its input (Russell, 2010).

A recurrent ANN feeds its output to its inputs and, with this, it is possible to experience some sort of dynamic behavior in its activation pattern. This characteristic enables the network to have some sort of short-term memory (Russell, 2010).

An ANN can be arranged in layers. A layer is a set of units that receives input only from the previous set of units. It is said hidden when it is between the input and output layers and cannot be accessed directly from outside the network (Russell, 2010). The quantity of layers in an ANN defines its depth (Goodfellow; Bengio; Courville, 2016). Figure 3.2 shows a feed-forward ANN composed of an input layer, a hidden layer and an output layer.

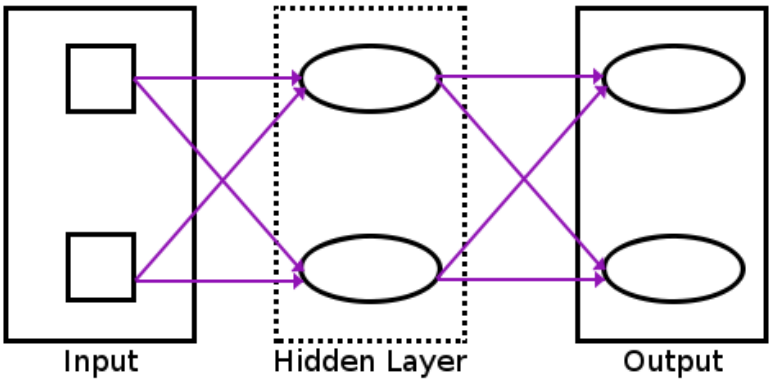


Figure 3.2: Example of a feed-forward network with an input layer, a hidden layer, and an output layer.

Depending on the number of layers an ANN can be considered deep or shallow. The term deep-learning, which is utilized as a synonym for Deep Neural Networks, conceives the concept of an approach for a representation-learning method that combines several layers of representations. These layers are non-linear modules that transform raw data into high-level abstract representations (Lecun; Bengio; Hinton, 2015).

In this thesis, the main interest is on Convolutional Neural Newtork, which is a kind of feed-forward neural network.

3.1.2 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is an ANN with one or more convolutional layers (Skansi, 2018). A convolutional layer passes a logistic regressor (called local receptive field) with input size n through the whole image (Skansi, 2018). The regressor is moved one component at a time but it is possible to increase how much it is moved, this is done utilizing the stride for the layer (Skansi, 2018).

Considering the input size of the regressor the output will be smaller than the original data if there is the necessity to put both data at the same size the procedure of padding can be used. Padding is the process of filling missing data with 0 or 1.

A convolutional layer can receive, as input, 1D data (1D Convolutional Layer), 2D data (2D Convolutional Layer), 3D data (Spatial Convolutional Layer), 4D or higher dimensional data (Hyperspatial Convolutional Layer) (Skansi, 2018). It is common to utilize a Rectified Linear Unit (ReLU) as the activation function for this type of layer. The ReLU function returns the maximum value between 0 and x (Skansi, 2018).

Aside from convolutional layers, the convolutional neural networks normally utilize feature maps and pooling. The idea behind feature maps is the generation of smaller but deeper images from the original data. Taking a 12×12 grayscale image with a 3×3 local receptive field, it will build a 10×10 output image, but if the image has 3 channels (such as RGB) the idea behind a feature map is to pass n local receptive fields over each channel. In the aforementioned example with $n = 5$ a 10×10 image with 15 channels will be constructed, this can greatly help the accuracy of the network (Skansi, 2018).

The next step is pooling. Formally, its function is to progressively reduce the spatial size of the representation in order to reduce the amount of parameters and the computation in the network. The most common form of pooling is max pooling. Max pooling is done in part to help over-fitting by providing an abstracted form of the representation. The max-pooling receives the size of a window as a parameter, the image is divided into a grid utilizing that value. Then, in the output image, each pixel is the maximum value of each window region in the input image. Other strategies for pooling are averaging pixels or utilizing the minimum value (Skansi, 2018).

Figure 3.3 shows an example of a Convolutional Neural Network. It is possible to see the effect of the 3×3 local receptive field and the pooling.

Another important aspect when dealing with neural networks is the possibility to re-utilize past experiences in new scenarios. This concept is known as transfer learning.

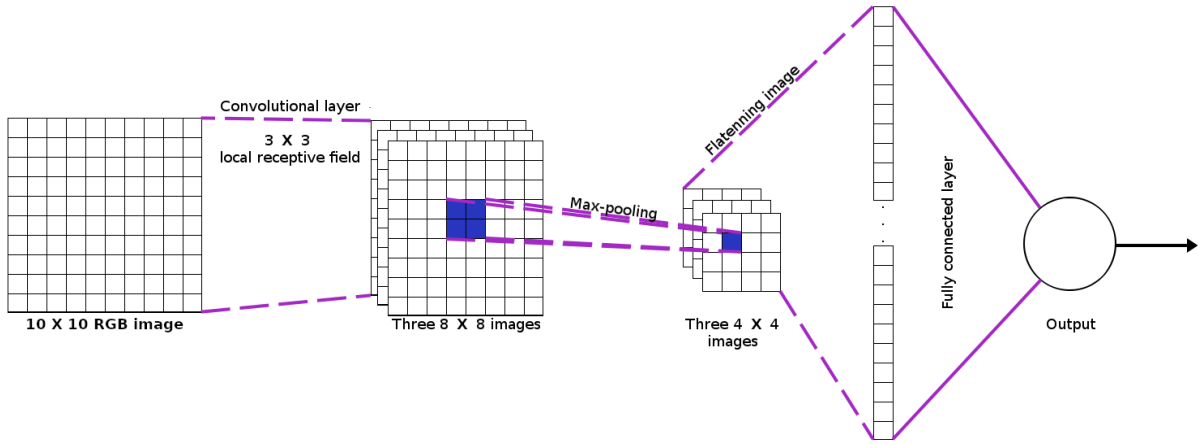


Figure 3.3: Example of a Convolutional Neural Network with a convolutional layer, pooling, and a fully connected layer. The output is given by a single neuron at the end (Skansi, 2018).

3.1.3 Transfer Learning

When previous knowledge is utilized to learn new concepts in a more efficient way, this is known as transfer learning (Yang; Hanneke; Carbonell, 2013). Normally, this is a way to utilize fewer data in adapting ANNs to new, but related, scenarios (*e.g.* fine-tuning a CNN from 2D to 3D data).

Initial layers in an ANN trained on images tend to learn low-level features (Lecun; Bengio; Hinton, 2015), sometimes being similar to Gabor filters (Yosinski et al., 2014). This was observed with different datasets and different contexts (Yosinski et al., 2014).

The features from the initial stages of an ANN can be defined as *general* because they are learned regardless the image dataset and the cost function (Yosinski et al., 2014). This is different from the last layer. In a classification task, a CNN will output data that is relevant to the specific scenario in which it is being trained, a softmax in a face recognition system, for instance, will learn the identities of the dataset, the last layers will output learned features that are relevant to that specifc task and not for a general task, this is why it is possible to define the last layer features as *specific features* (Yosinski et al., 2014).

Transfer learning will be effective under the condition that the features being transferred are general enough to suit both datasets and tasks. Usually, this approach utilizes the initial n layers from an ANN trained on a larger amount of data and transfer them, changing the last m layers, to a more specific, normally with a smaller quantity of data, and retrain. This mimics a real world scenario in which, with less information, it is possible to master a new subject if you have some prior knowledge related to it (Yang; Hanneke; Carbonell, 2013).

There are two possible ways to transfer learning, *fine-tuning* or *freezing* the weights (Yosinski

et al., 2014). Fine-tuning is defined when, during the training in a new task with transferred weights, the errors of the trained are backpropagated. Depending on the size of the new dataset and the new task this can lead to overfitting (Yosinski et al., 2014). In the freezing weights strategy, during the training in a new task with the transferred weights, only the specific features are trained (Yosinski et al., 2014).

3.1.4 Cross-Resolution Recognition

In 2D face recognition scenarios, for instance, it is common that low-resolution probe images need to be matched with high-resolution gallery images. This type of matching is defined as cross-resolution face recognition (Singh et al., 2018; Gao et al., 2020).

Cross-resolution recognition can be normally observed in system that utilize probe images from surveillance footage, since the distance in which the subject's image is captured and the quality of the camera can generate important differences in quality and size of the probe images (Gao et al., 2020). This type of problem can reduce significantly the performance of biometric identification systems (Singh et al., 2018; Gao et al., 2020).

The approaches that tries to deal with cross-resolution scenarios can be broadly categorized into *transformation based techniques* and *non-transformation based techniques* (Singh et al., 2018). When dealing with transformation based techniques there will be a transformation function that will be applied to data, in feature or image level, whilst in non-transformation based techniques the focus is to learn features, or classifiers, that are invariant to resolution differences (Singh et al., 2018).

Spite most of the works in the literature focus this problem in 2D environments, this phenomenon also happens with 3D data. Several types of devices can capture depth information and they can vary significantly concerning the quality of data. Table 3.1 shows the accuracy for several different types of 3D scanners, it is possible to see a big gap in the accuracy comparing high-resolution with low-resolution scanners. The difference in accuracy leads to holes, and spikes in the captured data severely impact face recognition systems, for instance.

In the present thesis, an intermediate feature representation is proposed for 3D face recognition. This intermediate representation, named as Descriptor Image (DI), has some level of resolution invariance. This is discussed in Chapter 6.

Table 3.1: Comparison between different 3D scanners (Li et al., 2013).

Device	Accuracy (mm)
3dMD	< 0.2
Minolta	≈ 0.1
Artec Eva	≈ 0.5
3D3 HDI R1	≈ 0.3
SwissRanger	≈ 10
DAVID SLS	≈ 0.5
Kinect V1	$\approx 1.5 - 50$

3.1.5 3D Local Binary Pattern

Local Binary Pattern (LBP) is a type of visual descriptor used in many applications of Computer Vision. LBP was first described in 1994 by Ojala, Pietikainen and Harwood (Ojala; Pietikainen; Harwood, 1994) and, since then, has been considered a powerful feature for texture classification.

Originally, the LBP operator takes a 3×3 sliding window throughout an image and utilizes the window region to calculate the operator for its central pixel. To this end, the region is thresholded with the central pixel value and, then, represented by a binary code. Finally, the binary code is converted into a decimal number, which is utilized as the label for the central pixel of the window region.

Ojala, Pietikainen e Maenpaa (2002) proposed an extended version for the LBP operator that modifies the size and shape of the window (local) region, being possible to select a circular region with radius R and P neighbor points. In this extended version, when a neighbor point falls on integer coordinates the intensity value of the pixel on that integer coordinates is utilized, otherwise, a bi-linear interpolation of the intensity values from the surrounding pixels is used.

The LBP operator only takes into consideration the signal of the comparison between a region and its kernel and cannot deal with the behavior of depth values. In the LBP, when two central points on different samples have highest (or lowest) depth values than their neighbors, they will have the same operator value, even if they are from different subjects (Huang; Wang; Tan, 2006). This would be common on points belonging to the nose tip of a face subject, for instance.

To deal with situations like this, Huang, Wang e Tan (2006) proposed the 3D Local Binary Patterns (3DLBP). This variation of the original operator considers not only the signal of the difference, but also the absolute depth difference. Huang, Wang e Tan (2006) state that, for face, more than 93% of all depth differences (DD) with $R = 2$ are smaller than 7. Due to this property,

the absolute value of the DD is stored in three binary units $(i_2 i_3 i_4)$. Therefore, it is possible to affirm:

$$|DD| = i_2 \cdot 2^2 + i_3 \cdot 2^1 + i_4 \cdot 2^0 \tag{3.4}$$

There is also i_1, a binary unit defined by:

$$i_1 = \begin{cases} 1 & \text{if DD} \geq 0; \\ 0 & \text{if DD} < 0. \end{cases} \tag{3.5}$$

Those four binary units are separated into four layers and, for each of those layers, four decimal numbers are obtained: P_1, P_2, P_3, P_4. The value of the P_1 has the same value as the original LBP. Figure 3.4 shows the process for the generation of the 3DLBP, given an image.

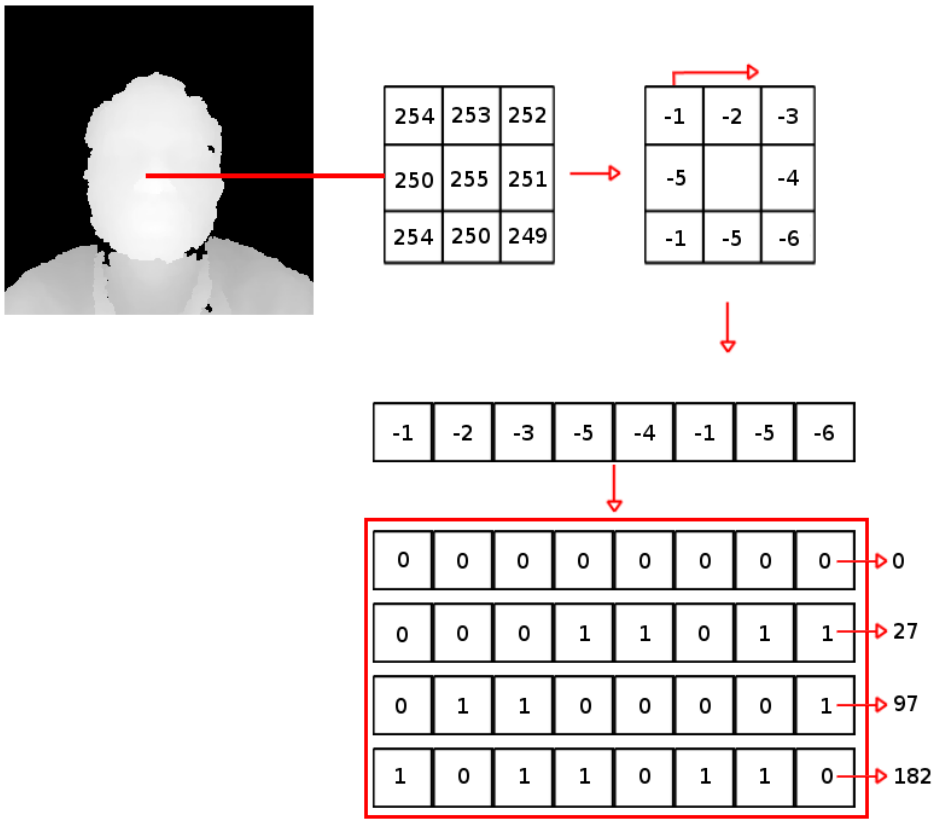


Figure 3.4: The full process of the 3DLBP proposed by (Huang; Wang; Tan, 2006).The absolute depth differences are encoded into the three layers (layer 2, 3 and 4) and the signal into the layer 1.

3.2 Related Works

Over the years, the literature on 3D face recognition has been focusing on methods that describe surfaces, specifically for capturing geometric properties of the facial geometry (Drira et al., 2013; Faltemier; Bowyer; Flynn, 2008; Mian; Bennamoun; Owens, 2007; Berretti; Bimbo; Pala, 2010; Kakadiaris et al., 2007; Spreeuwers, 2011). Even with several advances, most works that have been done so far utilize high-resolution data captured by costly devices in controlled environments. Some of the works that utilize Kinect-like devices try to increase the resolution of the data exploiting the temporal redundancy of frames in a sequence (Bondi et al., 2016; Drosou; Moschonas; Tzovaras, 2013; Hernandez; Choi; Medioni, 2012). The main problem of those approaches is that, in general, they are unable to operate in real time.

In (Li; Sun; Chen, 2017), the authors utilize an approach for 3D face recognition that explores deep representation patterns (DRP) that are sensitive to face location. For this task, first it is necessary to detect the nose-tip, to crop the face region, and to do pose normalization. In the next step, it is necessary to generate 2D projection from the 3D model for the generation of geometry images, in this process three images of the face normal components are estimated and fed into a pre-trained deep face network to generate deep representation. A location sensitive sparse representation classifier (LS-SRC) is utilized to measure the similarity of the deep normal patterns that are associated with each face. In the end, each of the representation goes through a score-level fusion for final identity decision. This work obtained rank-1 recognition rates of 98.01%, 97.60% and 96.13%, respectively, on the FRGC v2 [1], Bosphorus [2], and BU-3DFE [3] datasets.

Only a handful of methods perform face recognition directly from low-resolution data. In the work proposed by Min *et al.* (Min et al., 2012) a real-time 3D face identification system that receives a depth sequence as input is built. Initially, the region of the face is detected and segmented by utilizing a threshold on the depth values. In the next step, the face images are cropped and reduced to common resolutions and the matching is obtained by registering a probe with several intermediate references in the gallery with the EM-ICP (Granger; Pennec, 2002) algorithm. Mantecón et al. (2014) proposed the Depth Local Quantized Pattern as a modification of the original LBP operator. This modification introduces a quantification step that allows the descriptor to distinguish between different patterns. The descriptor was used to train and test an SVM classifier. In another work, Mantecón et al. (2016) proposed an algorithm for face

[1] https://www.nist.gov/programs-projects/face-recognition-grand-challenge-frgc
[2] http://bosphorus.ee.boun.edu.tr/default.aspx
[3] http://www.cs.binghamton.edu/ lijun/Research/3DFE/3DFE_Analysis.html

recognition based on an image descriptor called bag of dense derivative depth patterns. Dense spatial derivatives are first computed and quantized in a face-adaptive fashion to encode the 3D local structure. Then, a multi-bag of words creates a compact vector description from the quantized derivatives.

In (Neto; Marana, 2014) the authors propose a way of doing 3D face recognition utilizing data extracted from Kinect v1 sensors. Since the Kinect data have a great deal of noise, a pre-processing step was utilized, aiming to increase the point cloud density and, after that, feature extraction methods were used and their scores were fused. HAOG (Histogram of Averaged Oriented Gradients) and 3DLBP descriptors were utilized for classification, after training a SVM for each characteristic type and observing the results, the score of both decision were fused, utilizing a weight for each feature decision. The fusion achieved around 97% of rank-1 identification rate in the EURECOM Kinect Face Dataset (Huynh; Min; Dugelay, 2012) [4] and was competitive with the state-of-the-art methods.

The development of deep architectures that deal with 3D data has had a slower expansion than the image-based counterpart, mainly because of the data representation problem. While CNNs were designed to work with 2D images, the wide variety of 3D data (*e.g.*, point-clouds, triangular meshes, etc.) makes it difficult to work in the same standardized way without making significant modifications in the whole framework.

An example of a possible way to make use of existing deep architectures for 3D face recognition is the work proposed by Kim *et al.* (Kim et al., 2017), in which the authors utilized a pre-trained version of the VGG-Face and fine-tuned it for depth data. To deal with the shortage of depth data for training, the authors expanded the dataset by generating expressions and occlusions.

With the necessity of large amounts of data, Gilani *et al.* (Gilani; Mian, 2018) proposed a synthetic data generation technique that they used to build a dataset of ≈3M scans. The authors utilized such data to train a deep architecture that follows a VGG-like structure and consists of 13 convolutional layers, 3 fully connected layers and the final softmax layer. One of the conclusions reached in their work is that, because of the smooth nature of the face surface, there is the need for larger kernels for the convolutional filters with respect to the ones commonly used.

Lee et al. (2016), proposed a face recognition system based on deep learning that utilizes face images captured with a consumer-level RGB-D camera. For this task, three steps are performed: depth image recovery, deep learning for feature extraction, and joint classification.

[4] http://rgb-d.eurecom.fr/

To alleviate the problem of the limited size of available RGB-D data for deep learning, the deep network is firstly trained with a standard RGB face dataset and later fine-tuned on depth face images for transfer learning. The main difference between this work and the aforementioned ones is that it focuses on low-resolution data instead of high-resolution.

A hybrid solution exploiting both the RGB and depth information was presented in (Jiang; Zhang; Deng, 2019), in which a CNN is trained, guided by the supervision of an additional loss, called attribute-aware loss, that attempt to cluster the face images based on attribute information such as gender or age.

Along these lines, another recent work dealing specifically with low-resolution depth scans was developed by Mu et al. (2019), in which a lightweight CNN equipped with a multi-scale feature fusion layer is employed to fill the gap between high- and low-resolution depth scans.

In order to train a Deep Network for face recognition it is necessary a huge amount of labeled data, namely face of subjects. One way to deal with this problem is to create a set of artificial faces utilizing 3D face rendering. The main problem with this is the bias between the 2D rendered and real faces. To deal with such problem, in (An; Deng; Hu, 2017) the authors utilize deep transfer network to reduce the dataset bias. In their work, the authors generate several face images (with pose, neutral, expressions, and several others) utilizing the 3DMM (Morphable Models) face model, utilize the Inception-Resnet-V1 as the benchmark model, and optimize the deep transfer network with a maximum mean discrepancy. Their results are competitive with the state-of-the-art approaches.

The work proposed in (Yin; Liu, 2018) evaluates multi-task learning (MTL) for face recognition. For such a complicated task the authors propose a CNN in which classifying the subjects is the main task and the other problems commonly faced in 2D face recognition (such as pose and illumination) are secondary tasks. In the next step, the authors propose a dynamical weight adjustment for each side of the learning task. Finally, a CNN which groups similar poses to extract pose-specific features is proposed. The paper exhibits results on in-the-wild datasets (such as LFW) which compete with state-of-the-art results. This particular work does not utilize face reconstruction, but CNNs.

Looking at the related works it is possible to identify some gaps in which it would be possible to develop new research topics. More specifically Shallow CNNs, cross-resolution scenarios (training on high-resolution and testing on low-resolution), and face recognition with reconstructed 3D data to name a few.

Chapter 4

PROPOSED METHOD

In this thesis, we propose a novel hybrid approach for 3D face recognition that utilizes descriptor images (DI), also proposed in this thesis, calculated from facial cloud points, in conjunction with shallow convolutional neural networks (CNNs). The idea behind this is that, since the proposed DIs are constructed from low-level feature extractors, it would be possible to use shallow CNNs, allowing us to build an ensemble of shallow networks that are modular and needs fewer data and less computational power to train the CNNs. Our hybrid approach can be used in two modes: 3D face classification and 3D face feature extraction. Both modes have an initial module in which descriptor images are calculated. While the second module of the classification mode uses shallow CNNs as classifiers to identify the 3D input face, the second module of the feature extractor module uses shallow CNNs to extract higher-level features from the 3D input face. Details of the Descriptor Image Generation and of the Classification and Feature Extraction Modes are presented in Sections 4.1, 4.2 and 4.3, respectively. The whole code for this proposed method, including the shallow CNNs models, the pre-trained weights, the pre-processing pipeline, and the DIs generation module is publicly available at github [1].

4.1 Descriptor Image Generation

Depth images obtained from consumer-like cameras (*e.g.* Kinect Sensor) can be an interesting solution to build systems for 3D face recognition, which operate seamlessly as 2D consumer cameras. In this case, the main problem is that, normally, this type of devices obtain data with fewer details of the face compared to those acquired by high-resolution scanners. In this sense, training from scratch a Deep Convolutional Network (DCNN) on such data is difficult for two main reasons: *(i)* depth data in general, and low-resolution depth data in particular, present

[1] https://github.com/jbcnrlz/biometricprocessing

more acquisition noise than RGB images; *(ii)* large volumes of labelled depth faces are difficult to collect. Besides, for these data, it is not possible to collect samples in the same way as one would with 2D scanners. Thus, one workaround found in the literature is taking a Deep CNN architecture pre-trained on RGB data and fine-tuning it with a small set of depth images (Mu et al., 2019; Parkhi; Vedaldi; Zisserman, 2015).

In this thesis, we propose a different approach, in which we utilize a low-level feature extractor to generate an intermediate feature representation for 3D data, this representation is called as Descriptor Image (DI). Our hypothesis is that this type of data allows the use of shallow networks to do 3D face recognition in two different modes, feature extraction and classification. We also hypothesize that the proposed DI has some degree of resolution invariance that allows us to train shallow CNNs in high resolution 3D face data and use it to recognize low resolution 3D faces.

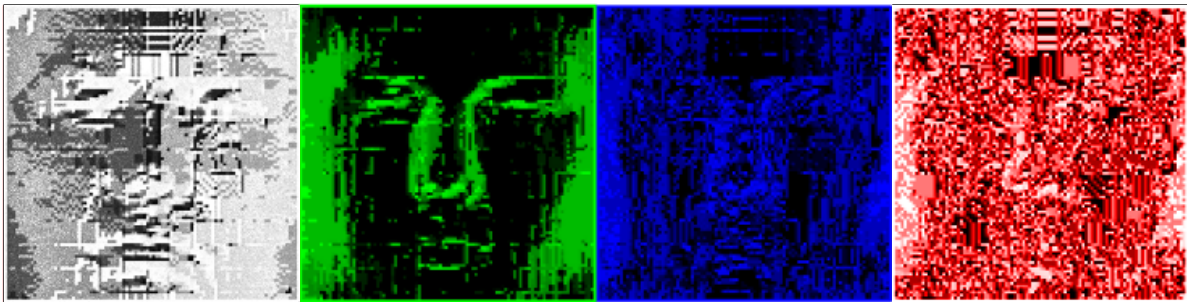
4.1.1 Descriptor Image Generated From 3DLBP

3DLBP descriptor (Huang; Wang; Tan, 2006), described in Section 4.1.1, was adopted in our work as low-level face features due to its computational efficiency and the fact that it has been proven to be effective in describing 3D depth images of faces. Another reason we utilized a low-level hand-crafted feature for the DIs is our belief that generating a representation that encodes this type of feature would allow us to utilize shallower networks, once this type of feature is the same type that initial layers of a CNN would encode.

In 3DLBP, the depth differences are encoded as a feature. With the [-7,+7] range, 15 different values are to be encoded, which results in a four-bit representation. Each bit is regarded as a separate channel: the first channel encodes the sign of the difference, that is 0 if the difference is negative, 1 otherwise; the other three channels encode the absolute value of the difference transformed in a binary code of 3 bits. Figure 3.4 shows the generation of a 3DLBP descriptor for a pixel with a 3×3 neighborhood region.

In our approach, differently from the original 3DLBP descriptor (composed by the concatenation of the histograms of the four channels), the 3DLBP descriptors of the whole depth are transformed into an image, in which each one of the four bits of the 3DLBP is regarded as a separate channel of the final descriptor image (DI). To encode these four bits, we use a four-channel RBGA image, with the last channel being the alpha channel.

In the generated DI, each channel has a different behavior. The first one encodes the sign and it describes if the local neighborhood is increasing or decreasing with respect to the central



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

Figure 4.1: The channels of the DIs generated with the 3DLBP. (a) sign of the difference (b), (c), and (d) the three bits of the absolute difference value.

point (*e.g.*, a local minimum would be encoded as 255, that is all bits are 1). The last three channels encode the absolute depth difference between each center point and its neighbors. Figure 4.1, shows the four channels of a DI obtained from the 3DLBP descriptor of a face. The difference between each of the channels is a direct consequence of the encoding procedure.

As reported, the first channel encodes the sign of the difference, and changes in the values appear to occur smoothly. This happens mainly because faces are smooth surfaces and shifts in values do not occur abruptly. The second channel receives the encoding of the most significant bit of the absolute depth differences. The value of the difference can only be an integer that goes from zero to seven; with this in mind, values of this channel are 1 for differences bigger or equal to four (most significant bit). This does not happen as often on the face because of its smooth surface, as one can see in Figure 4.2, which shows the distribution of depth differences in human faces. However, there are some regions in which it is possible to see more abrupt changes in the difference values, such as the nose and ocular areas. Looking at the last two channels, we note that they are generally noisier. The reason behind this is that they encode less significant bits, thus the changes occur more rapidly. The third channel (second bit) changes for differences of two, while the fourth one (least significant bit) changes for depth differences of one. This generates high frequency information.

4.1.2 Descriptor Image Generated From Sigmoid 3DLBP

A limitation of the standard 3DLBP descriptor is that, within the [-7,+7] interval, negative or positive difference values share the same binary code, except for the sign bit. This implies that some regions of the resulting DI might have the same values on three out of four channels. One way to account for this is to incorporate a sigmoid function in the computation of the 3DLBP operator. Figure 4.3 shows a sigmoid function, which is defined in equation 4.1.

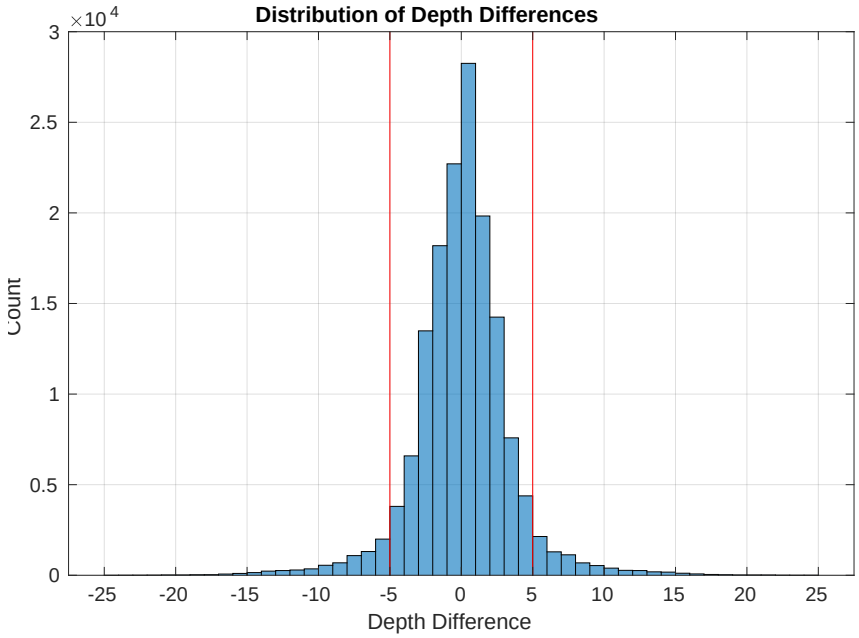


Figure 4.2: Distribution of the depth difference in the FRGC dataset. The red vertical lines indicate the [-5,+5] range. It is clear that the vast majority of the depth differences falls in this range.

$$f(x) = \frac{1}{1+e^{-x}}. \tag{4.1}$$

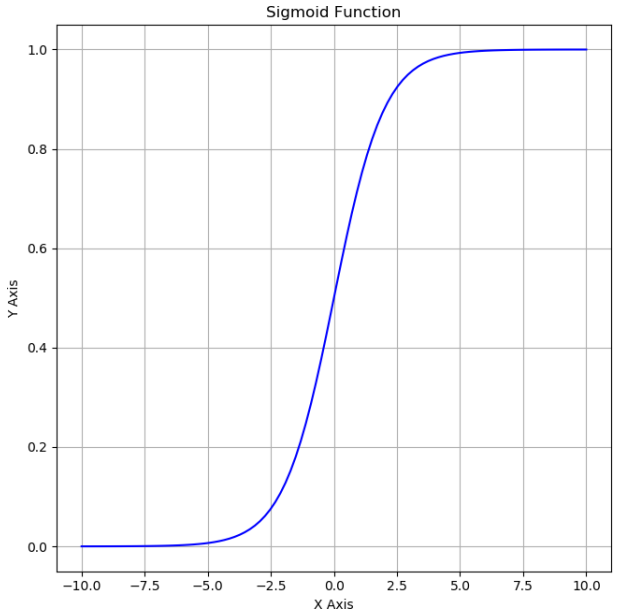


Figure 4.3: The S shaped curve of a sigmoid function.

The sigmoid function was formulated in the 19th century to describe population growth and the course of auto-catalytic chemical reactions (Cramer, 2002). It is a bounded, differentiable, real function that is defined for all real input values and has a non-negative derivative at each point and exactly one inflection point (Han; Moraga, 1995).

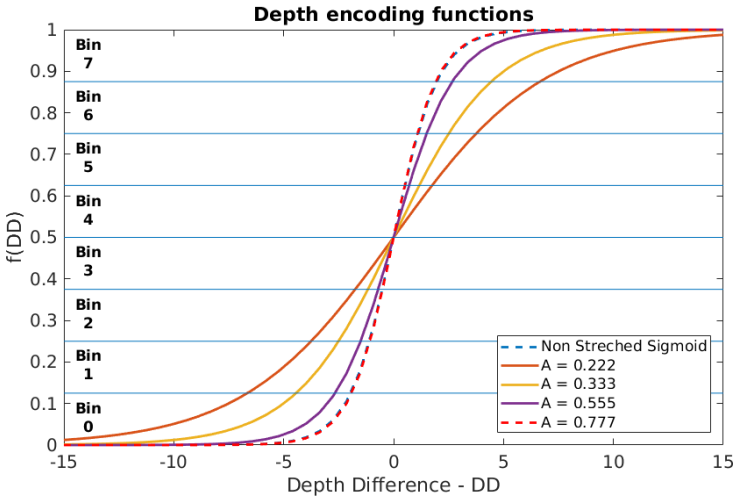


Figure 4.4: Different curves of each sigmoid utilized in the encoding of the depth differences. It is possible to visualize the bins utilized to encode the values, alongside with where each value will fall. It is possible to observe that the curve for the non stretched sigmoid and with $A = 0.777$ have very similar behavior, almost overlapping.

Instead of using four bits for representing values in the [-7,+7] range and truncating the exceeding values, with the use of a sigmoid function it is possible to map larger intervals to four bits. To encode the sigmoid values, 8 bins are defined in the interval between 0 and 1. Then, each $f(x)$ value is mapped to its closest bin, in a histogram-like fashion. Note that, even though the sign channel is maintained into the four-channel image, $f(x)$ is computed considering the depth difference along with the sign, so that same values with opposite signs are put into different bins. This has the advantage of encoding a larger range of variations, though with a coarser resolution. This also ensures that different maps with respect to the classic 3DLBP are generated. Spite the fact that Eq. (4.1) encodes the same absolute value in different bins depending on the sign, it is only effective if most of the values fall in the range [-4,+4] (see Figure 4.2). In order to encode a broader range of values equation 4.2 was utilized

$$f(x) = \frac{1}{1 + e^{-A*\ln(2+\sqrt{3})*x}}, \tag{4.2}$$

in which x is the depth difference value, and A is a scalar value that stretches the sigmoid function, making it possible to change the range in which the values are encoded. Figure 4.4 shows how the function of equation 4.2 behaves given different stretch values.

4.1.3 Descriptor Image Generation Module

The novel hybrid approach for 3D face recognition proposed in this thesis utilizes 3DLBP-based descriptor images (DI), calculated from facial cloud points. Figure 4.5 shows a diagram

of the proposed descriptor image generation module. One can observe that this module is composed of two main stages: the pre-processing and the 3DLBP feature extraction. First, a cloud point obtained from the face goes through a pre-processing pipeline in order to improve the depth data quality and to attenuate differences that are caused by differences in resolution. The output of the pre-processing stage is a depth image that, by its nature, can be considered a Descriptor Image. Next, in the depth image generated by the pre-processing is used as input to the hand-crafted 3DLBP-based feature extraction. The extracted 3DLBP features are, finally, utilized to build an image representation called Descriptor Image (DI).

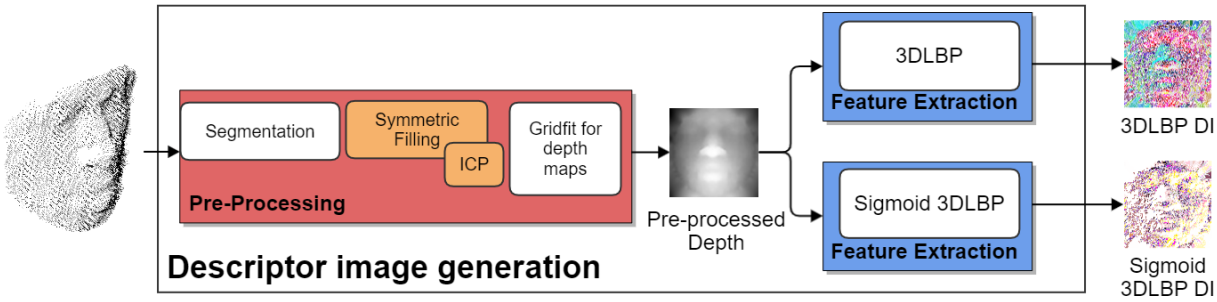


Figure 4.5: Block diagram of the descriptor image generation module. A cloud point is utilized as an input. First, it goes through a pre-processing pipeline to improve data quality and to attenuate differences that are caused by differences in resolution. The output of the pre-processing is a depth image that, by its nature, can be considered a Descriptor Image. The next step is the hand-crafted feature extraction from the pre-processed depth data. The extracted 3DLBP-based feature is then utilized to build an image representation called Descriptor Image.

In the pre-processing stage, there are four main steps, which are necessary to fill the holes in the point cloud and to increase the data quality:

Segmentation: Starting with the nosetip the face is centered at the origin and a circle with radius R is segmented. Figure 4.6 shows how this step works;

Symmetric Filling: The Symmetric Filling technique, proposed by Li et al. (2013), utilizes the left side of the face to increase point density by including the set of mirrored points

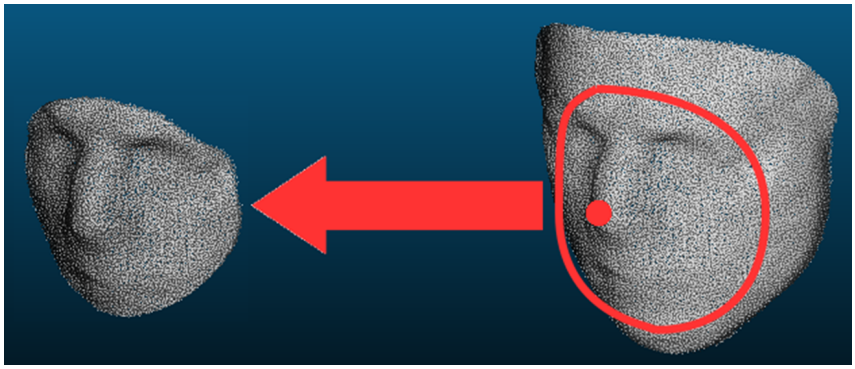


Figure 4.6: Face segmentation process.

from the right side of the face, and vice-versa. However, not all the mirrored points are useful because the goal is to fill only in the missing data (occluded regions, for instance). Likewise in (Li et al., 2013), the strategy to be used in our method is to add the mirrored point only if there is no neighboring point at that location. During this process, if the Euclidean distance from a mirrored point to its neighbors in the original point cloud is greater than a threshold value δ, then that point will be added to the original face data. Figure 4.7 shows how this step works;

Iterative Closest Point: The Iterative Closest Point (ICP), proposed by Besl e McKay (1992), is a well known solution for the problem of registration. It tries to find a rigid transformation that minimizes the least-square distance between two points. Given two 3D point sets (A and B), ICP performs the following three basic steps:

1. Pair each point of A to the closest point in B;
2. Compute the motion that gives the lowest Mean Squared Error (MSE) between the points;
3. Apply the motion to the point set A and update the MSE.

The three aforementioned steps are performed until the MSE is lower than a threshold τ. A complete description of this method can be found in (Besl; Mckay, 1992) and (Chetverikov; Stepanov; Krsek, 2005);

Generation of Depth Map: In the proposed method, depth maps from cloud points must be generated. Cloud points will be obtained from the reconstructed 3D face models. In order to generate depth maps from cloud points, a circular region with radius R is cropped centered at the nose tip. Then, the cropped image goes through the symmetric filling process. Finally, the resulting face image is fitted to a smooth surface using an approximation method, implemented by an open source code[2] written in Matlab. The result of this process is a 100×100 matrix, as illustrated in Figure 4.5. This pre-processed depth map is considered a DI. Figure 4.8 shows the input and output of our whole process.

Figure 4.9 compares the results of our pre-processing pipeline in high- and low-resolution data. It is possible to see that the pipeline helps to attenuate resolution differences, specially when removing holes and spikes from the low-resolution data.

After the preprocessing steps, the pre-processed depth map is utilized as an input for the 3DLBP-based feature extraction (traditional 3DLBP or sigmoid 3DLBP). Instead of generating

[2] http://mathworks.com/matlabcentral/fileexchange/8998-surface-fitting-using-gridfit

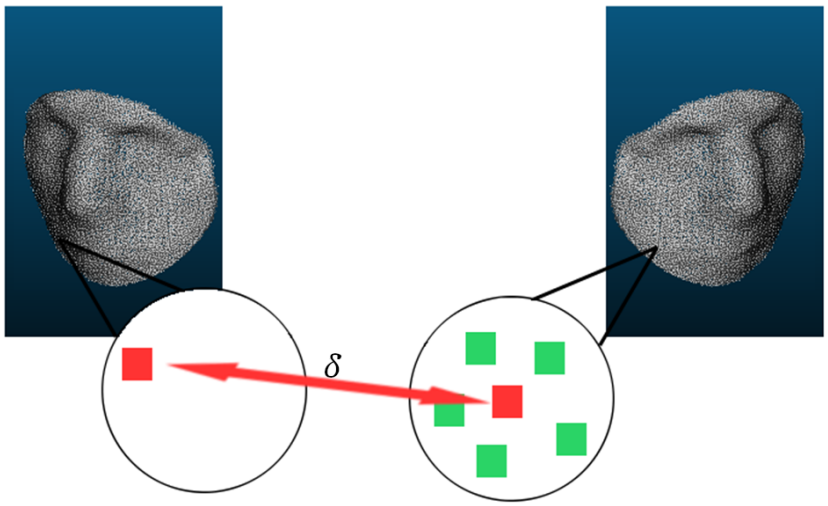


Figure 4.7: Symmetric filling process.

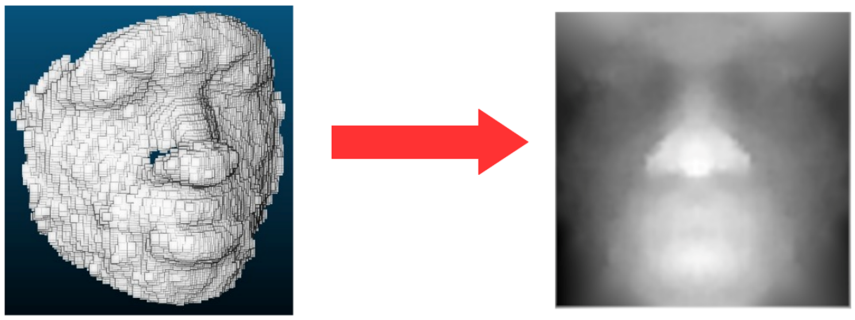
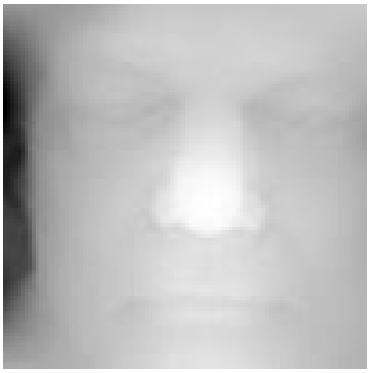
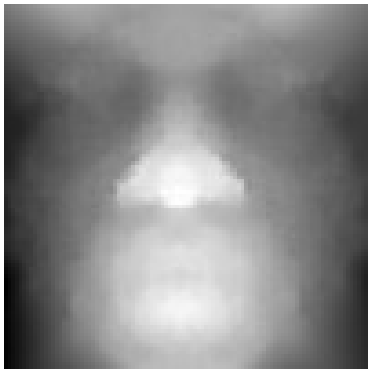


Figure 4.8: Input and output for our pre-processing pipeline.

a histogram as the descriptor, the idea is to generate a 2D representation that can be utilized as input to a CNN. This representation is generated by substituting each central point in a



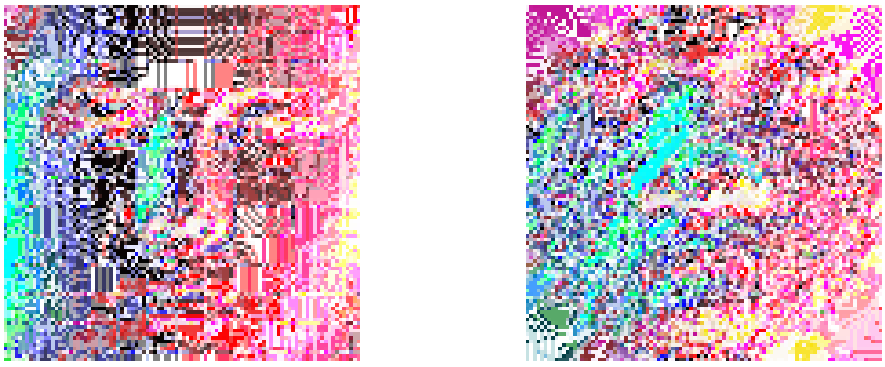
(a) High-resolution pre-processed depth map



(b) Low-resolution pre-processed depth map

Figure 4.9: Comparison between a high-resolution pre-processed depth map (a) and a low-resolution pre-processed depth map (b). Even though there are visible differences between resolution, landmarks of the face are visually perceptible even in the depth map obtained from low-resolution data.

neighborhood region with the value of the four components generated by the 3DLBP. Being more specific, in a region, the central depth value is substituted with the four values generated by the 3DLBP operator. In this scenario, an image that has only one channel, with the depth value, will be represented as an RGBA image, with each channel being the value of each layer of the 3DLBP operator. Figure 4.10 shows a comparison between 3DLBP DI from high- and low-resolution. In the regions where the red color is more predominant it means that the majority of regions have the central point depth value smaller than the majority of neighbors, it also means that the depth difference from the initial points (*e.g.* the upper left most neighbor) are closer to 1, this means that the value for the first layer, which corresponds to the red channel, are bigger than the rest. In this scenario, the face region in the DI has a reddish tone. The regions where the blue color is more predominant normally have the depth different from its initial points closer to -2, -3, -6, and -7. This means that the red channel will have a smaller value while the blue channel, corresponding to the second layer from the absolute depth difference, will have bigger values giving a bluish tone to the image region.



(a) High-resolution 3DLBP DI (b) Low-resolution 3DLBP DI

Figure 4.10: Comparison between a high-resolution 3DLBP DI (a) and a low-resolution 3DLBP DI (b).

4.2 Classification Mode

In the approach proposed in this thesis for 3D face recognition, 3DLBP-based DIs, computed from low-resolution cloud point data, as described in Section 4.1, are used as input to a shallow CNN in order to find the identities of individuals. Figure 4.11 shows a diagram that illustrates the proposed classification mode.

Since the DIs represent a more elaborated version of the original raw depth data, our hypothesis is that face recognition can be made by CNNs with much less layers than those CNNs necessary to do the same task but using the original raw depth data. Based on this hypothesis,

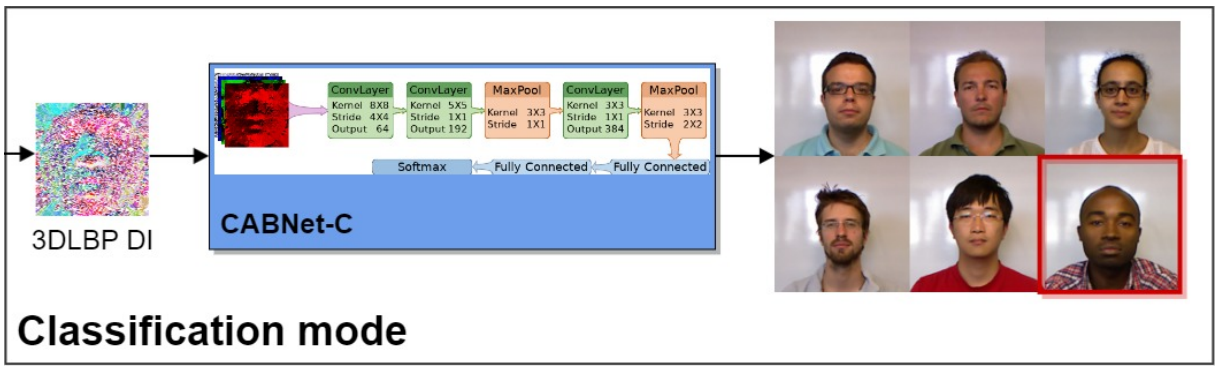


Figure 4.11: Block diagram of the classification mode proposed in this thesis, that uses the 3DLBP-based DI and a shallow CNN.

the neural network architecture designed in this thesis is a very shallow CNN composed only of three convolutional layers, as illustrated in Fig. 4.12. We have named this CNN as CABNet-C.

In the proposed CABNet-C, the first layer operates with 64 filters, an 8×8 kernel and a stride of 4. The next convolutional layer has 192 filters, a kernel of 5×5 and a stride of 1. The last convolutional layer has 384 filters and a kernel of 3×3. Between the second and last convolutional layer, there is a max pooling layer with a kernel of 3×3 and stride of 2. After the last max pooling layer, there are two fully connected layers and a softmax at the output for subject classification. There is also a ReLU activation layer after each convolutional layer.

The input image size is $100 \times 100 \times 4$, and all of the images are normalized before being utilized as input. For normalization, each channel is divided by its max value. Since depth data is generally smoother than 2D data, we utilize larger kernel sizes, as suggested by (Gilani; Mian, 2017). Another reason for utilizing bigger kernels at the beginning of the network is that we believe that the DIs already encode some meaningful information, allowing the use of networks that are more shallow than most of the literature utilizes.

In order to train the CABNet-C, we use the FRGC dataset with 200 epochs. For each face, we apply a 3D rotation in the Y (pitch angle) and X (yaw angle) axis from -30 to 30 degrees. With this dataset augmentation strategy, the total number of images for training is $48,867$.

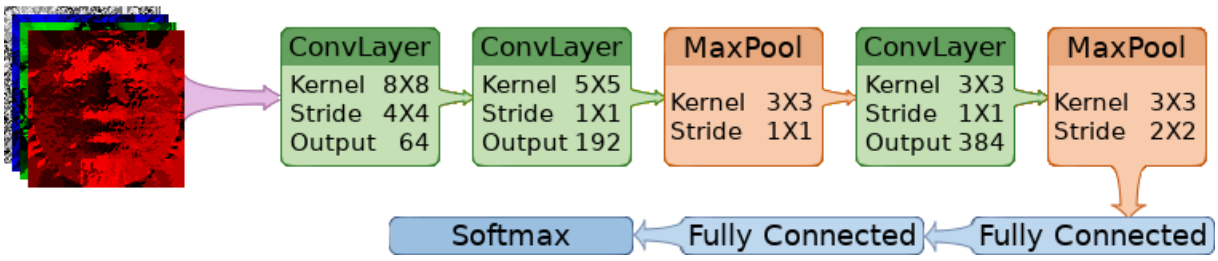


Figure 4.12: Architecture of the proposed Shallow CNN Classifier (CABNet-C). The network takes the DIs as input (four channel images in RGBA format).

Before validating the network, it goes through a fine-tuning process for 100 epochs and the data utilized for the process is augmented in the same way as the FRGC data.

4.2.1 Score Fusion

To perform 3D face recognition we employed a score fusion between the 3DLBP and the Sigmoid 3DLBP DIs, in order to investigate whether a coarser encoding of larger depth intervals (Sigmoid 3DLBP DI) can bring complementary information to the original formulation (3DLBP DI). The fusion is a weighted sum of the classification scores:

$$Score_F = w_{3DLBP} \times Score_{3DLBP} + w_{SIG} \times Score_{SIG} \,, \tag{4.3}$$

being $Score_F$ the final score for a subject, w_{3DLBP} the weight for the 3DLBP approach, $Score_{3DLBP}$ the original softmax score for the 3DLBP, w_{SIG} the weight for the sigmoid 3DLBP, and $Score_{SIG}$ the softmax score for the original sigmoid. We experimented the fusion of scores with the w_{SIG} weight values ranging from 0.1 to 0.9 (step = 0.1) and the $w_{3DLBP} = 1 - w_{SIG}$ during the training on the FRGC dataset. The values that yielded the best result were w_{3DLBP} equal to 0.6 and w_{SIG} equal to 0.4. Figure 4.13 illustrates how the score fusion works.

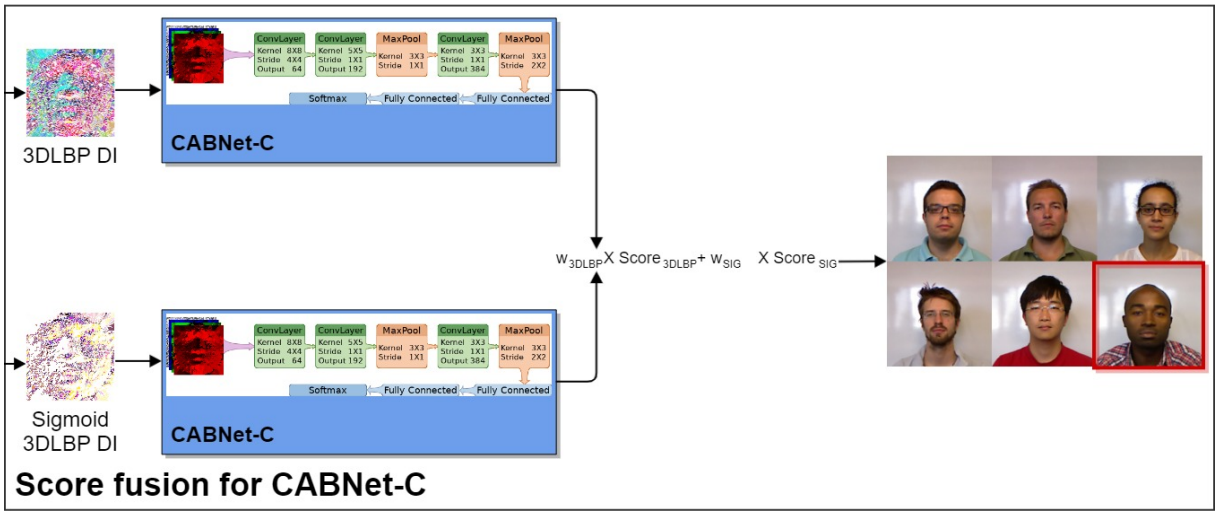


Figure 4.13: Block diagram that illustrates the score fusion between the 3DLBP and Sigmoid 3DLBP DIs.

4.3 Feature Extraction Mode

Figure 4.14 shows a diagram that illustrates the proposed feature extraction mode. One can observe that for the feature extraction mode we propose a shallow CNN architecture that is dif-

ferent from the CABNet-C, we have done this because of the necessity to create a more compact and robust feature representation. This new shallow CNN is composed of four convolutional layers and the input can be DIs of $100 \times 100 \times 4$ or $100 \times 100 \times 3$. This allows flexibility on the type of data to be utilized as input to the network. For our hand-crafted DIs, we utilize four channels, while for the pre-processed depth DIs, we utilize three channels. After each convolutional layer, there is a batch normalization and a ReLU activation function. We have named this shallow CNN as CABNet-FE. Figure 4.15 shows details of the proposed architecture of the CABNet-FE.

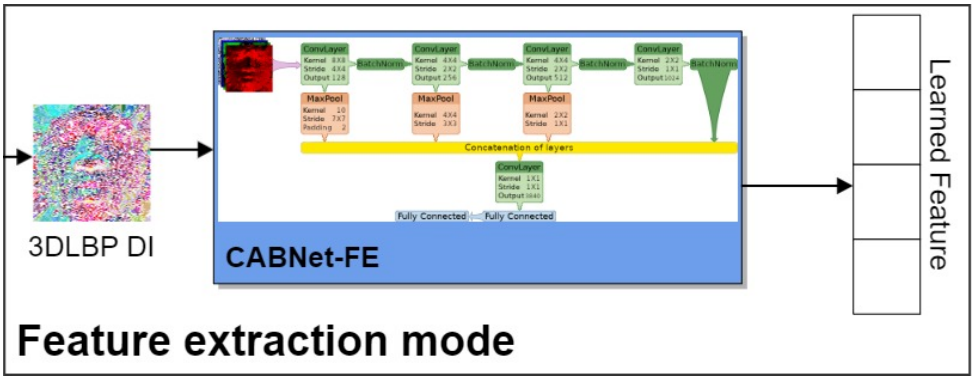


Figure 4.14: Block diagram of the feature extraction mode proposed in this thesis, that uses the 3DLBP-based DI and a shallow CNN.

To build the face feature we concatenate the outputs from the convolutional layers. Since the feature maps from each layer have different output sizes, a max pooling layer is utilized to normalize it. With that in mind, the output from each max pooling is, respectively, $3 \times 3 \times 128$, $3 \times 3 \times 256$, and $3 \times 3 \times 512$. The last convolutional layer does not utilize a max pooling and its output shape is $3 \times 3 \times 1024$.

The reason we utilize feature maps from different parts of the network is that we believe that

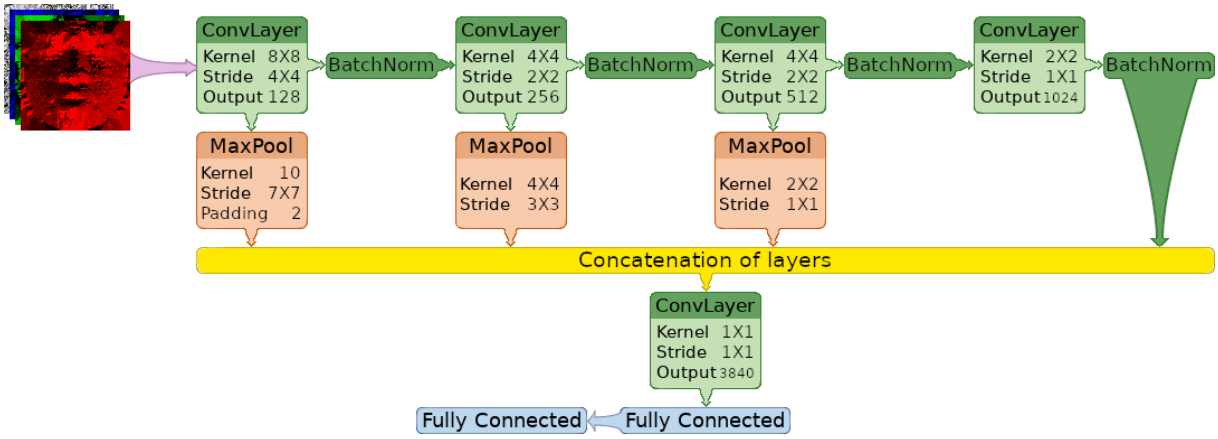


Figure 4.15: Architecture of the proposed CABNet-FE.

there is discriminative information throughout its whole structure. This is a hybrid approach because we start with a previous hand-crafted feature, instead of the raw data (even the pre-processed depth map is a type of feature we generate after the pre-processing pipeline).

After the concatenation process, a feature map of $3 \times 3 \times 1920$ is obtained. This is fed into another convolutional layer with the output of $3 \times 3 \times 3840$. At the end of the network there are two fully connected layers, the feature map from the previous convolutional layer is flattened having the size of 34560. This flattened feature map is fed into the first fully connected layer reducing the dimensionality to 2048 features. It is desired to create a robust and discriminative face representation based on the concatenation of several features. As the network architecture is shallow, this idea becomes viable, since each instance of the network consumes few resources.

4.3.1 Shallow Learned Feature Representation (SLFR)

We have named the features that are learned with the CABNet-FE as Shallow Learned Feature Representation (SLFR). Whilst the CABNet-C has a feature size of 4096, the SLFR has a size of 2048, making it more compact. Another aspect that is clearly shown in the Section 6.3.1 is that the SFLR has shown higher levels of resolution invariance.

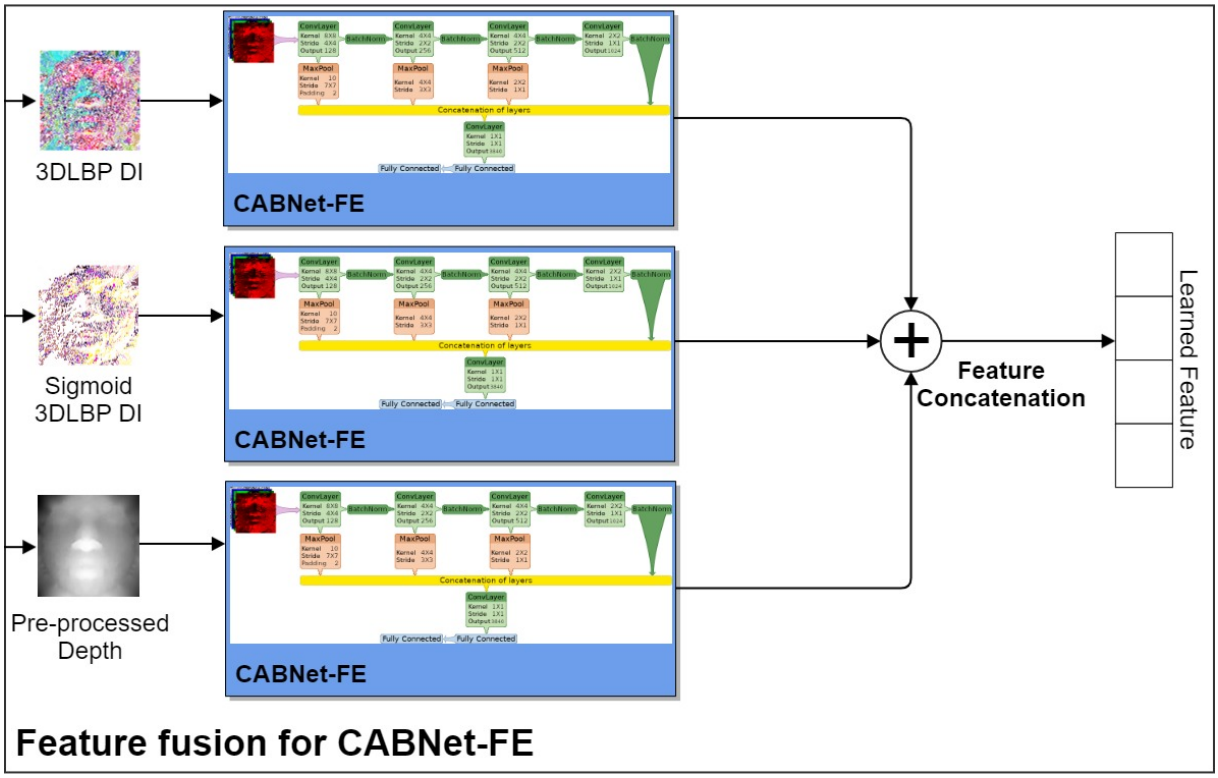


Figure 4.16: Block diagram that illustrates how the feature concatenation generates the fused 3D face descriptor.

4.3.2 Feature Fusion

Instead of using fusion in the score level, as the case with the classification mode, we have employed in the feature-extraction mode a fusion in the feature level. To this end we concatenate different SLFR and utilized it as the 3D face descriptor. Figure 4.16 shows how the feature fusion is carried out.

Chapter 5

MATERIAL AND METHODOLOGY

In order to assess the proposed novel hybrid approach for 3D face recognition that utilizes descriptor images (DI), originated from facial cloud points, in conjunction with shallow convolutional neural networks (CNNs), we carried out some experiments. In sections 5.1 and 5.2 we present, respectively, the material and the methodology adopted in the experiments.

5.1 Material

The datasets used in this work are: *(i)* Face Recognition Grand Challenge v2.0 (FRGC) dataset (Phillips et al., 2006), *(ii)* Bosphorus 3D face dataset (Savran et al., 2008), and *(iii)* EURECOM Kinect Face Dataset (Huynh; Min; Dugelay, 2012). The FRGC is used to train our shallow CNNS from scratch, while tests are conducted on the EURECOM Kinect Face Dataset for low-resolution data, and on the Bosphorus for high-resolution data. When there is the need for fine-tuning, the original softmax layer is discarded and a new one is utilized, this is needed since different datasets have different amounts of subjects and the new softmax layer will have its output with the same size of the number of subjects that the dataset utilized for fine-tuning, while the original softmax has the same size as the subjects from the FRGC. Another aspect when fine-tuning is that it has only utilized images that compose the gallery as part of the process. Details of these three datasets are presented in the following subsections.

5.1.1 FRGC Dataset

The FRGC dataset (Phillips et al., 2006) consists of 50,000 recordings divided into training and validation partitions. The training partition is designed for training algorithms and the validation partition is for assessing performance of an approach in a laboratory setting. The

validation partition consists of data from 4,003 subject sessions. A subject session is the set of four controlled still images, two uncontrolled still images, and one three-dimensional image. The controlled images were taken in a studio setting, are full frontal facial images taken under two lighting conditions and with two facial expressions (smiling and neutral). The uncontrolled images were taken in varying illumination conditions; e.g., hallways, atriums, or outside. Each set of uncontrolled images contains two expressions, smiling and neutral. The 3D image was taken under controlled illumination conditions. The 3D images consist of both a range and a texture image. The 3D images were acquired by a Minolta Vivid 900/910 series sensor. Figure 5.1 shows a subject from the FRGC Face Dataset.

Figure 5.1: An example from a subject in the FRGC Dataset (Phillips et al., 2006), the first row are the RGB images and the second row is our pre-processed depth map obtained from range data.

5.1.2 Bosphorus Dataset

The Bosphorus dataset (Savran et al., 2008) comprises $4,666$ high-resolution scans of 105 individuals. There are up to 54 scans per subject, which include expression variations, rotations and occlusions. The data is acquired utilizing a structured-light system, all subjects were sited at a distance of 1.5 meters from the digitizer and a 1000W halogen lamp was utilized in a dark room for illumination, ensuring that the lighting was homogenous.

Regarding the Bosphorus dataset, the rotated and occluded scans were not used in our experiments to make the comparison with other works more fair. We have utilized a total of $2,902$ neutral and expressive scans. Figure 5.2 shows examples of scans of four individuals from the Bosphorus Face Dataset.

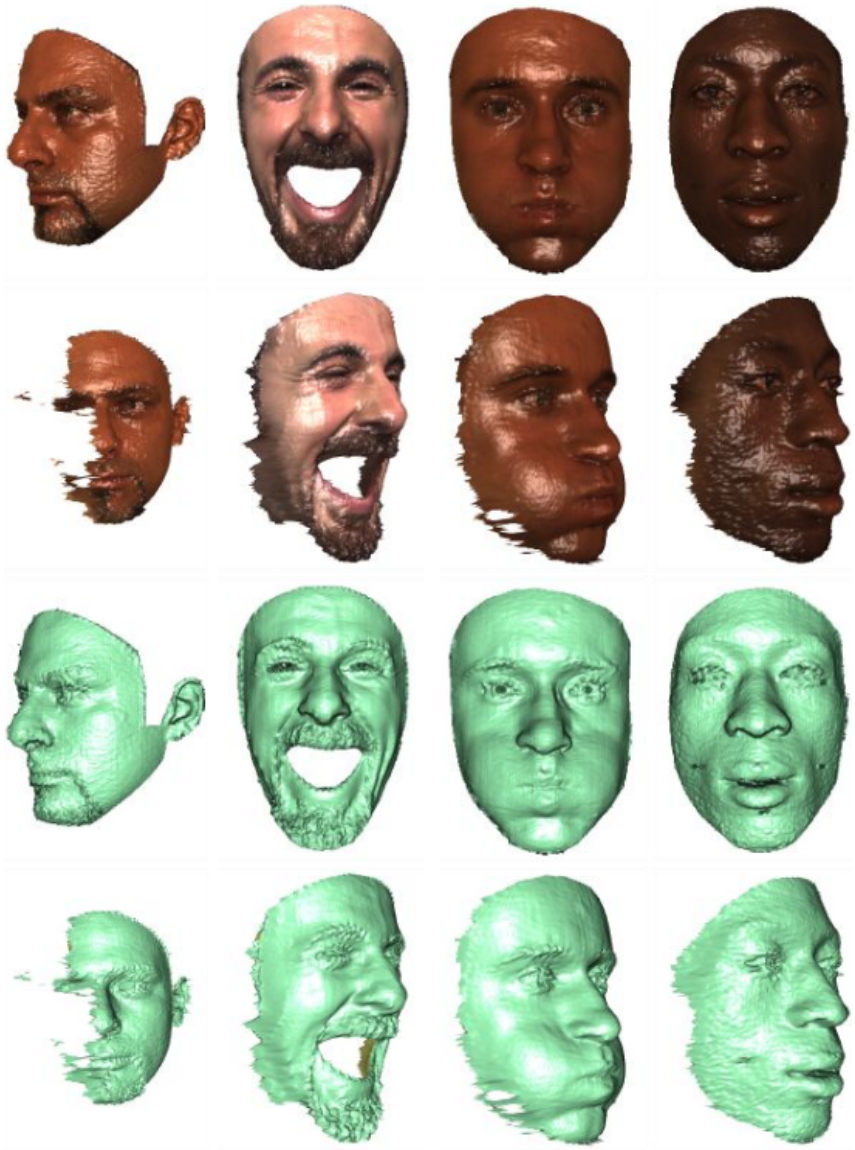


Figure 5.2: Examples of scans of four individuals from the Bosphorus Face Dataset (Savran et al., 2008).

5.1.3 EURECOM Dataset

The EURECOM dataset (Huynh; Min; Dugelay, 2012) collects RGB-D images acquired with a Kinect sensor of 52 subjects, taken in two separate sessions with 7 variations each: neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion and open mouth. This dataset is employed for evaluating our hybrid approach with low-resolution data with three different protocols, as defined in (Huynh; Min; Dugelay, 2012):

(i) Gallery composed of seven variations (neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion, and open mouth) from Session 1, and Probe composed of seven variations (neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion

and open mouth) from Session 2;

(ii) Gallery composed of seven variations (neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion, and open mouth) from Session 1, and Probe composed of three variations (neutral, smile, illumination) from Session 2;

(iii) Gallery composed of seven variations (neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion, and open mouth) from Session 1, and Probe composed of one variation (neutral) from Session 2.

Figure 5.3 shows a subject from the EURECOM Kinect Face Dataset.

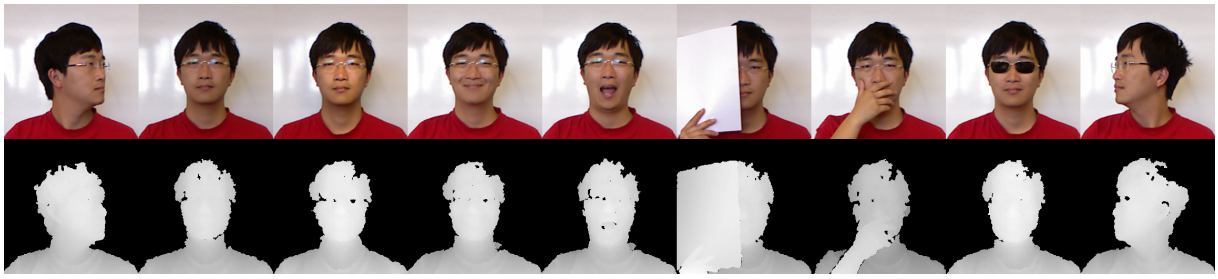


Figure 5.3: A sample from a subject from EURECOM Kinect Face Dataset (Huynh; Min; Dugelay, 2012). In the first row the RGB images are shown and on the second the original depth images.

5.2 Methodology

The methodology of our experiments has three main stages: descriptor image generation, network training, and person identification.

Initially, our novel hybrid approach receives high-resolution cloud points, then the nose-tip is utilized as the center for a circle with radius R of 70 mm, the points that fall inside this circle are segmented and compose the face. High-resolution data does not go through the Symmetric Filling step since the trade-off between performance and data improvement is not worth in this case. Lastly, the segmented face is transformed into a depth map, utilizing a grid fit approach that approximates the points into a grid of size 100×100. This depth map (that we call pre-processed depth map) is the first DI that our method generates.

The next step is to utilize the pre-processed depth map in order to obtain a hand-crafted 3DLBP-based feature, since one of our hypothesis is that by utilizing this type of data we are using the same type of information that the early stages of a Convolutional Neural Network (CNN) would do, hence being able to utilize CNNs with few layers. In this thesis, we have proposed two different ways to obtain 3DLBP-based features: (i) by applying the traditional

3DLBP, as described in subsection 4.1.1, or (ii) by utilizing a sigmoid variation of the 3DLBP, as described in subsection 4.1.2. The sigmoid modification of the 3DLBP focus on changing the way that the depth difference is encoded into the operator. Instead of truncating the values between $[-7, +7]$ we use a sigmoid function in the absolute value of the depth difference. Since the sigmoid result is a real number in the range [0,1] we divide that interval into 8 bins. The final value is the bin number (from 0 to 7) that the value falls into and the signal from the depth difference. In the end, the sigmoid 3DLBP and the traditional 3DLBP are represented as a DI in the same manner: a four channel (RGBA) image.

For training, we utilize the DIs (3DLBP, sigmoid 3DLBP, or pre-processed depth image) as input and train the proposed shallow CNNs for 200 epochs. The CNNs are implemented with PyTorch library (Paszke et al., 2019). We utilize the stochastic gradient descent as optimizer, with a learning rate of 0.01 and the cross-entropy function loss. The weights that we utilize in production are the weights of the smaller loss from the 200 epochs.

When assessing the performance of our new proposed hybrid approach the data also need to go through the descriptor image generation step, the difference, in this case, is that the symmetric filling is applied since we used low-resolution data (EURECOM dataset). In this thesis, we focused on low-resolution data and for this type of data, the symmetric filling must be used, since it improves the data quality and helps to attenuate the difference caused by resolution changes.

Finally, the person identification, which is the primary goal of our work, varies depending on the used mode:

Classification mode: In this mode, the first step is to fine-tune the shallow CNN for 100 epochs with the gallery images. For the fine-tuning process the original softmax layer is replaced with a new one, whose output size is the same as the number of classes being identified;

Feature extraction mode: In this mode, the last fully connected layer is utilized as the features, the identity of a subject is given by using a 1-NN classifier, with the cosine similarity function. For this end, we calculate the similarity between the current face and all of the faces in the gallery, the subject identity is taken as the identity associated with the face with higher similarity among all gallery face images.

For evaluating our method with the state-of-the-art works following the literature, we have used *rank-based* metrics, with tables comparing rank-1 recognition rate and CMC Curves.

Chapter 6

EXPERIMENTAL RESULTS

In this chapter, we present results obtained with the experiments carried out in order to assess the proposed hybrid approach for 3D face recognition based on Descriptor Images (DI) and shallow Convolutional Neural Networks. We start presenting an ablation study which intended to evaluate how the CNN depth affects the approach's performance. Then, we show results regarding the Descriptor Images (DI) and the Shallow Learned Feature Representation (SLFR) proposed in this thesis for 3D face representation and description.

6.1 Ablation Study

We carried out an ablation study to evaluate the two main components of the proposed approach for 3D face recognition. In the following, we report an analysis regarding the depth of the neural network, that is, the number of convolutional layers, and the sigmoidal encoding for the Descriptor Images.

For evaluating the impact of the neural network depth in the recognition rates, we have experimented three different network configurations: the first one is a very shallow network composed of only one convolutional layer; the second one has two convolutional layers and a max pooling layer; and the third one has three convolutional layers and two max-pooling layers.

In order to assess if the proposed DIs, when originating from low-resolution data, carry enough information to distinguish between distinct 3D faces, in the first experiment, we trained the neural networks with EURECOM Kinect Face dataset.

In the experiment, we utilized Stochastic Gradient Descent (SGD) with learning rate of 0.01, and momentum of 0.5. We utilized the Cross Entropy as our loss function. The gallery data was augmented by rotating the original face depth images from -30 to 30 degrees in the X

and Y axis.

6.1.1 CNN Depth Analysis

In these experiments, carried out in order to evaluate the impact of the number of convolutional layers on the recognition accuracy, the gallery (training set) was composed of depth images from Session 1 of the EURECOM Kinect Face dataset, with all variations, as described in subsection 5.1.3, except the "paper occlusion" one, for a total of 4,035 images. The probe images (test set) come from the same classes, but of Session 2. We utilize two types of data, the 3DLBP and the pre-processed depth, the latter being used as the baseline reference.

Figure 6.1shows the learning curves for the three CNN configurations with, respectively, 1, 2 and 3 convolutional layers. With these curves it is also possible to compare the 3DLBP-based DIs with the pre-processed depth map-based DIs.

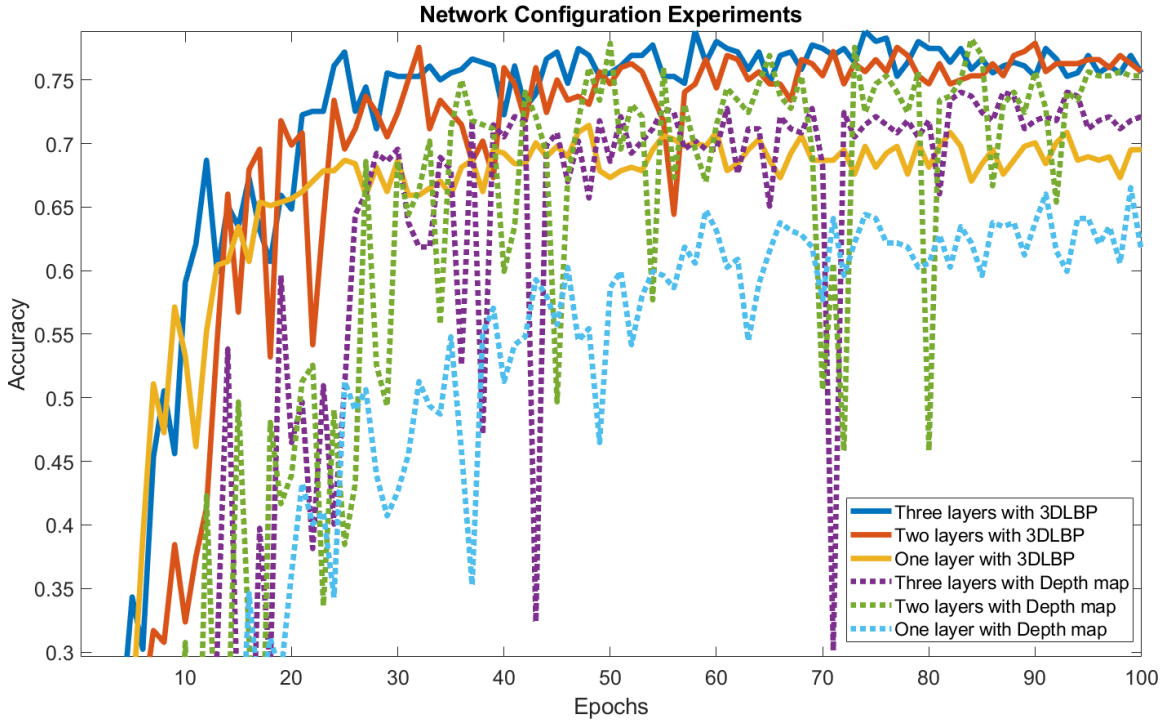


Figure 6.1: Learning curves on the EURECOM dataset for three CNN configurations, with 1, 2 or 3 convolutional layers, respectively. The curves also compare the 3DLBP-based DIs, and pre-processed depth map-based DI.

One can observe in Figure 6.1 that the best result was obtained when three convolutional layers were used. Thus, from now on we will only use this winner CNN architecture for all subsequent experiments in our work.

One can also observe in Figure 6.1 that the accuracy rates obtained by using 3DLBP-based DIs are higher and the curves are better behaved than by using the DIs based on the depth maps.

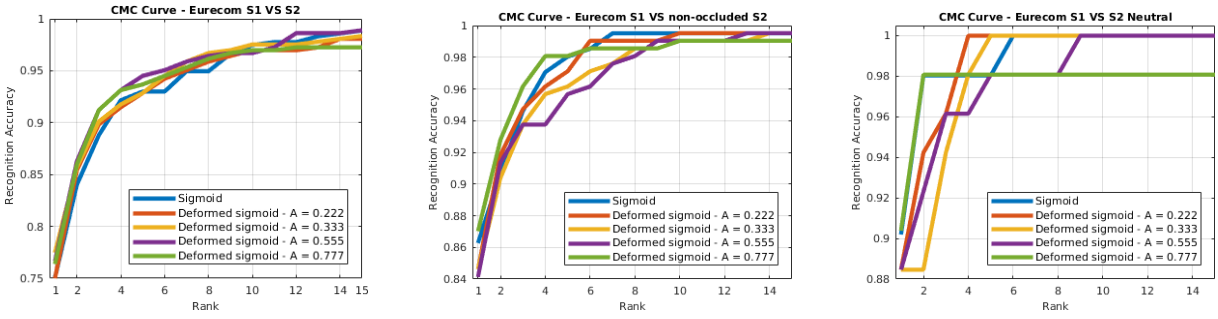
This result corroborates our hypothesis that the 3DLBP-based DIs contain more discriminative information than the DIs based on the pre-processed depth maps.

Finally, one can observe that the gap between the CNN architectures with 2 or 3 convolutional layers is rather slight, making it possible to use the CNN with only two convolutional layers when very low-powered devices must be used, without impacting much on the performance.

6.1.2 Sigmoid Encoding Function

In order to evaluate the sigmoid encoding functions, we performed some experiments using the three protocols for EURECOM Kinect Face dataset (S1 vs. S2, S1 vs. Non-Occluded, S1 vs. Neutral), described in subsection 5.1.3.

Figure 6.2 presents the Cumulative Matching Characteristic (CMC) curves obtained for the three protocols, for different values of parameter A, which stretches the sigmoid function, according to Eq. 4.2.



(a) *Protocol (i)* (b) *Protocol (ii)* (c) *Protocol (iii)*

Figure 6.2: CMC curves obtained by utilizing different types of sigmoid encoding. Experiments were performed using the three protocols defined for the EURECOM dataset: (a) S1 vs. S2; (b) S1 vs. Non-Occluded; (c) S1 vs. Neutral.

Looking at the curves in Figure 4.4, it is possible to see that both these solutions encode depth differences larger than 4 in the same bin. This points out that such bigger differences are less informative, while encoding more precisely the smaller ones is beneficial and that we can use the regular sigmoid function instead of the deformed one without any major differences in the performance. This makes sense in as much as larger differences are likely to occur in case of strong noise, holes or occlusions, *e.g.*, hand or glasses occlusions.

6.1.3 Evaluating Learned Features

In order to assess the importance of the learned features for 3D face recognition using low-resolution data, we carried out experiments using, again, the three protocols for EURECOM Kinect Face dataset (S1 vs. S2, S1 vs. Non-Occluded, S1vs. Neutral), but now with shallow CNN learned features and the DIs as 3D face features. That is, we used the output obtained from the shallow CNNs and the flattened DIs (3DLBP DI and Sigmoid 3DLBP DI) as feature vectors, in each experiment. The matching was done by calculating the cosine similarity between the probe and gallery feature vectors, according to the three protocols. Table 6.1 shows the rank-1 recognition rates obtained with this experiment, which indicate that the use of the shallow CNN learned features improves in more than 100% the rank-1 recognition rates.

Table 6.1: Rank-1 recognition rate results obtained on EURECOM Kinect Face dataset. Comparing the protocols with the DI and the Learned Features from the DIs.

Method	Protocols		
	(i)	*(ii)*	*(iii)*
3DLBP (flattened DI)	27.47%	32.21%	32.69%
Sigmoid 3DLBP (flattened DI)	22.25%	27.40%	26.92%
3DLBP (learned features)	**59.06%**	**67.30%**	**73.07%**
Sigmoid 3DLBP (learned features)	58.79%	65.86%	67.30%

6.2 Classification Mode

In this section, we discuss the CABNet-C being utilized with the proposed DIs in the classification mode, in different resolutions.

Spite the score fusion described in Section 4.2.1 we also assessed the fusion scores from pre-processed depth map with 3DLBP and Sigmoid 3DLBP DIs. In this case, the best weight values were 0.3 for the pre-processed depth map, 0.3 for the sigmoid 3DLBP, and 0.4 for the 3DLBP.

6.2.1 Low-Resolution 3D Face Scans

In this section, the results obtained on low-resolution 3D face scans from the EURECOM Kinect Face dataset are reported. In this experiment, during the fine-tuning process, a new softmax layer was stacked in place of the original one and re-trained, and the matching was performed by a 1-NN (Nearest-Neighbor) classifier, using the cosine similarity between the face features. Results comparing the baseline and state-of-the-art methods are reported in Table 6.2.

Table 6.2: Rank-1 accuracy rates on EURECOM Kinect Face dataset (best results in bold). The gallery is from Session 1, while the probe set is composed of: *(i)* Session 2, *(ii)* the three variants without occlusions from Session 2, *(iii)* neutral scans from Session 2.

Method	*(i)*	*(ii)*	*(iii)*
RGB Data	51.83%	55.71%	60.00%
Pre-processed depth map	79.12%	86.54%	90.38%
Sigmoid 3DLBP	77.8%	89.2%	94.1%
3DLBP	80.9%	90.2%	**96.1%**
3DLBP + Sigmoid 3DLBP	**90.75%**	**98.0%**	**96.1%**
Pre-processed depth map+ 3DLBP + Sigmoid 3DLBP	71.15%	77.88%	78.84%
VGG (Kim et al., 2017)	13.9%	14.8%	13.5%
Lee *et al.* (Lee et al., 2016)	-	80.8%	78.8%

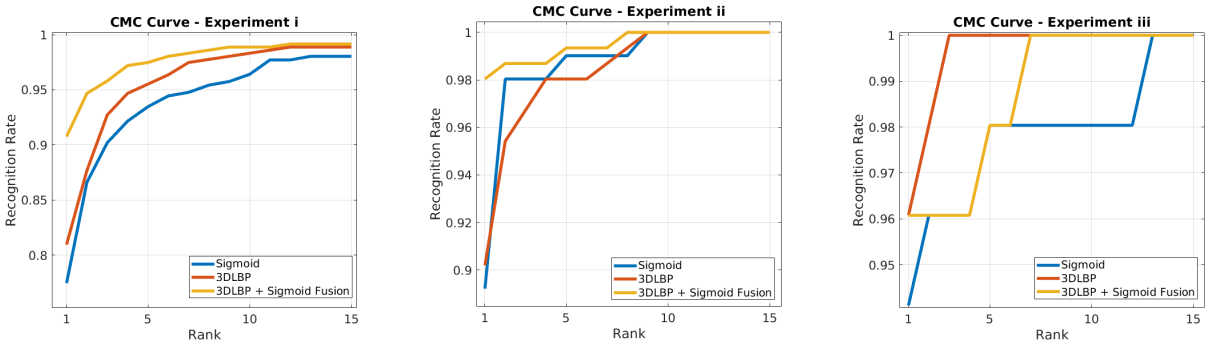
In Lee *et al.* (Lee et al., 2016), a pipeline to include the 3D face shape in a deep network is proposed. It performs depth face image recovery and enhancement, extraction of deep representation, and joint classification for depth and RGB data. Since our work deals only with depth data, for a fair comparison, we reported in Table 6.2 only the results for depth data reported in (Lee et al., 2016).

Since the original work in VGG (Kim et al., 2017) does not deal with low-resolution data we have utilized the pre-trained weights availble to do 3D face recognition on the EURECOM dataset, in this way the reported results here are with low-resolution data.

It is possible to see in Table 6.2 that, in the scenario of this experiment, the advantage of applying the pre-processing pipeline to the original depth map images is more evident, as we are dealing with low resolution and noisy data. It is also possible to see that there was a small advantage of the traditional 3DLBP approach over the sigmoid 3DLBP in this case. However, the score fusion led to a noticeable accuracy improvement. This can be a piece of evidence that, due to the coarser nature of the depth maps, it is beneficial to exploit balanced information from both sides. It is worth to note that the benefit of the fusion strategy is more evident in harder protocols (*(i)* and *(ii)*), and only in the protocol *(iii)* it tied with the 3DLBP approach in rank-1 recognition.

When dealing with low-resolution and complex scenarios, *e.g.* strong occlusions, the RGB data does not carry sufficient information for a satisfactory recognition with such a shallow CNN. Instead, with the proposed pre-processing procedure, using the depth images we obtained higher accuracy for all the three protocols, demonstrating that using 3D information can be beneficial. We can then further improve the performance by means of the proposed 3DLBP and Sigmoid 3DLBP DIs (Descriptor Images).

In the case of low-resolution data, a more balanced score fusion led to more accurate recog-



(a) *Protocol (i)* (b) *Protocol (ii)* (c) *Protocol (iii)*

Figure 6.3: Results on EURECOM Kinect Face dataset. CMC curves obtained using the Classification Mode in experiments carried out according the three protocols: (a) Protocol (i): Gallery from Session 1 vs. probe from Session 2 (7 variants each); (b) Protocol (ii): Gallery from Session 1 vs. probe with the three variants without occlusions from Session 2; and (c) Protocol (iii): Gallery from Session 1 vs. neutral probes from Session 2. The scales on the vertical axis are different.

nition. This can be a piece of evidence that, due to the coarser nature of the depth maps, it is beneficial to utilize more information from both sides (3DLBP and Sigmoid 3DLBP). Figure 6.3 shows the CMC Curves obtained in this experiment.

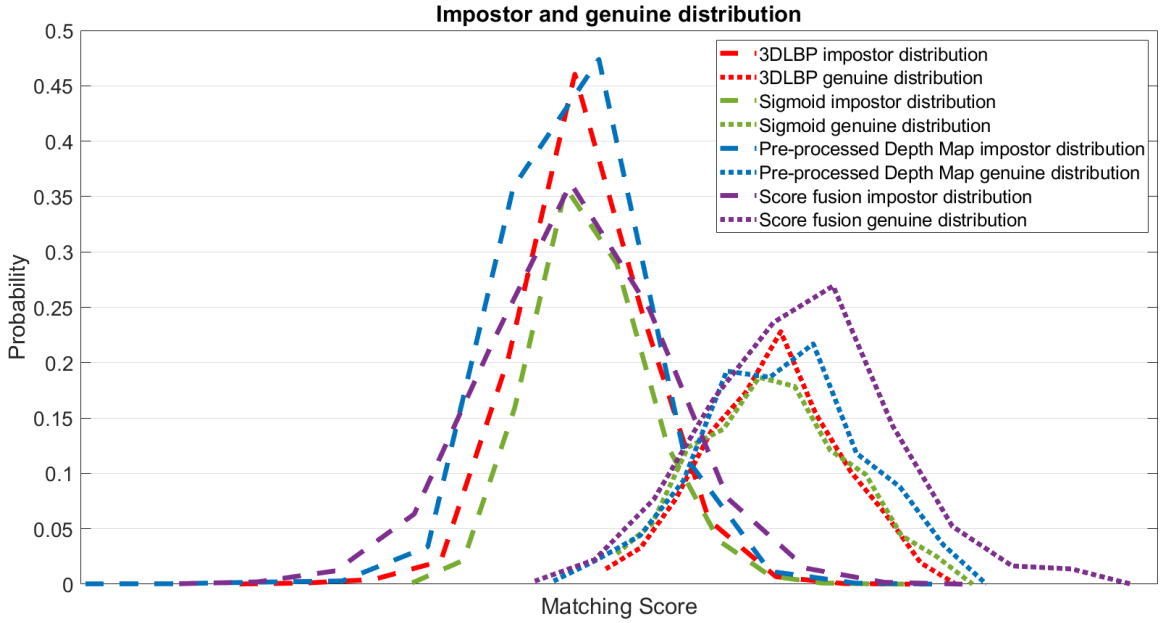


Figure 6.4: Impostor and Genuine distribution curves. This makes clear that there is bigger overlap between impostor and genuine distribution when all of the features are fused. For this the low-resolution images were utilized.

Fusing the pre-processed depth maps with the scores from the other DIs, in the classification mode, does not improve the recognition rates. This is an evidence that the information, in score level, for this type of data does not have complementary information. Figure 6.4 shows the impostor and genuine distribution curves. In this is possible to see that the Sigmoid and 3DLBP

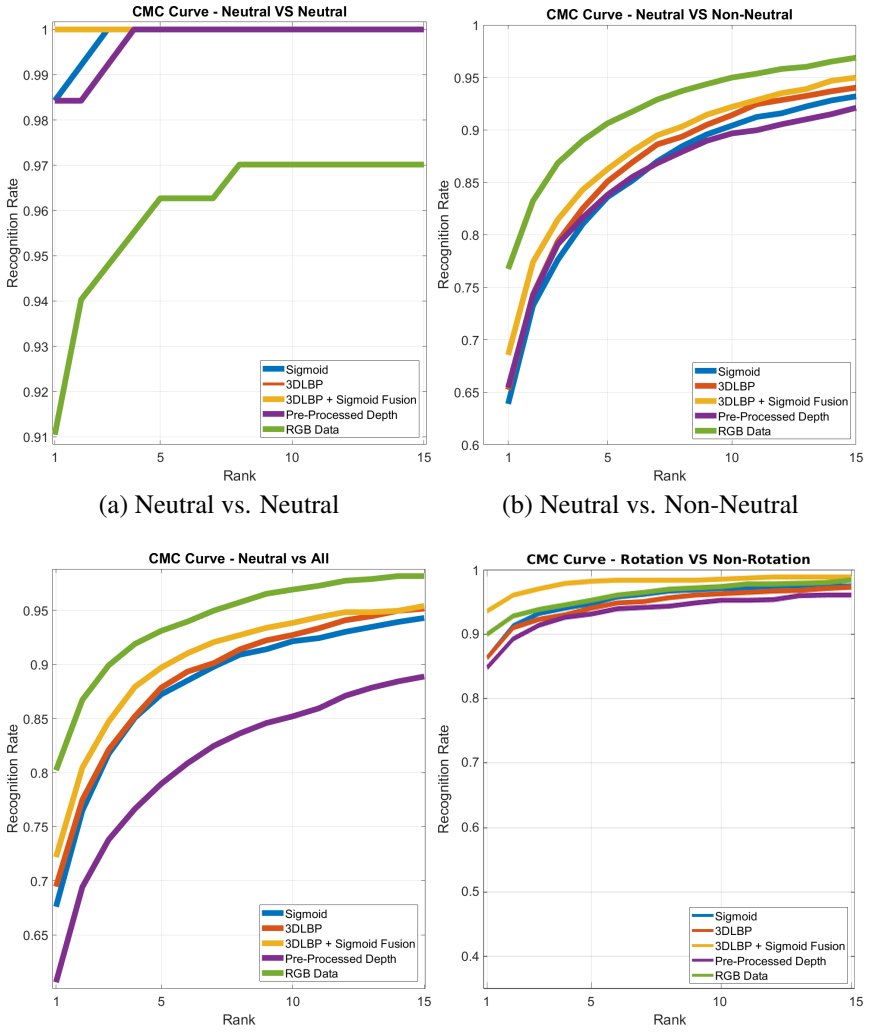
have similar behaviors, but the Pre-processed depth map is more spread across the genuine and impostor distributions. When fusing all of the DIs we got an impostor and genuine curves that have a bigger area of intersection than other features, in this way lowering the systems recognition performance. This is a piece of evidence that the behaviors are not complementary and, in this context, should not be score-fused.

6.2.2 High-Resolution 3D Face Scans

In this section, the results obtained on high-resolution 3D face scans from the Bosphorus database are reported and compared with the state-of-the-art. For this experiment, three protocols were adopted (gallery vs. probe set): *(i)* Neutral vs. Neutral (N vs. N), *(ii)* Neutral vs. Non-Neutral (N vs. NN), and *(iii)* Neutral vs. All (N vs. A), which includes neutral scans as well. In all the three cases, the gallery is composed of the first neutral image of each subject.

Figure 6.5 shows the Cumulative Matching Characteristic (CMC) curves obtained using the 3DLBP, the sigmoid 3DLBP encoding and a late fusion of both. It can be observed that, in the N vs. N scenario (Figure 6.5 (a)), results were very good and encouraging. However, the performance radically dropped when the probe set included several types of expressions, as emerges from the CMC curves in Figure 6.5 (b)-(c). This behavior is most likely caused by the very low number of images contained in the gallery, which were used to re-train the classification layer. Considering the very low number of images, we got a reasonable result, but far lower from the results reported in the current state-of-the art. To verify that the problem can be ascribed to this lack of data, we included rotated scans in the gallery, for a total of 5,139 gallery images (Figure 6.5 (c)). Even if rotated scans were characterized by missing parts, as consequence of self-occlusions, by adding them to the gallery, the results increased substantially. Note that, even if both sets, the training dataset (FRGC) and the gallery, contain a very limited amount of expressive scans, we are still able to perform rather accurate classification on them. Finally, this result also evidences the usefulness of our proposed pre-processing pipeline.

Table 6.3 shows the results of all the protocols with respect to baseline and state-of-the-art methods. In particular, to demonstrate the effectiveness of the proposed DIs, we compared results obtained using our shallow CNN trained on the RGB face images, on the original and on the pre-processed version of the depth map images. We further compared the hand-crafted features based approach by Li *et al.* (Li et al., 2015), Deng *et al.* (Deng et al., 2020), and Cai *et al.* (Cai et al., 2019). As a first outcome, results show that the proposed pre-processing technique is effective in generating enhanced depth images, with a considerable accuracy improvement above RGB data. Looking at the results for (N+R vs. A), we can observe that the score fusion



(a) Neutral vs. Neutral (b) Neutral vs. Non-Neutral

(c) Neutral vs. All (d) Neutral+Rotated vs. All

Figure 6.5: Results on Bosphorus dataset. CMC curves obtained using the Classification Mode in experiments carried out following all protocols. In (a-c), the standard protocols are reported. In (d), rotated scans and augmented images are added to the gallery in order to compensate the lack of gallery images.

helped in increasing the recognition rate in a significant way. Nevertheless, both the 3DLBP and the Sigmoid 3DLBP encoded images had better results with respect to the pre-processed depth. This indicates that the DIs have some complementary information, allowing both the networks to learn more effectively than pre-processed depth, and to combine them to further increase the accuracy.

We also can observe that, as happened in the low-resolution results, the fusion of the three types of DIs (Pre-proc. Depth, 3DLBP, and Sigmoid) decreases the performance in our approach. We believe that the reason behind this is the non-complementary nature of the information. Figure 6.6 shows the impostor genuine distribution curves. In this case, it is possible to see that the DIs based on low-level features helps to decrease the intersection between the

Table 6.3: Rank-1 accuracy rates on Bosphorus dataset (best results in bold). Comparison with the state-of-the-art using different protocols.

Method	*N vs. N*	*N vs. NN*	*N vs. A*	*N+R vs. A*
RGB data	91.04%	77.5%	80.8%	89.95%
Pre-processed depth map	98.42%	65.40%	60.61%	84.78%
Sigmoid 3DLBP	98.42%	63.89%	67.60%	86.40%
3DLBP	**100.0%**	65.76%	69.68%	86.30%
3DLBP + Sigmoid 3DLBP	**100.0%**	69.42%	72.06%	**93.54%**
Pre-processed depth map + 3DLBP + 3DLBP Sigmoid	91.33%	63.06%	62.85%	-
Li *et al.* (Li et al., 2015)	**100.0%**	-	96.60%	-
Cai *et al.* (Cai et al., 2019)	-	-	**99.75%**	-
Deng *et al.* (Deng et al., 2020)	**100.0%**	**97.60%**	-	-

impostor and genuine curves but when fusing the three DIs the intersection area between them is bigger than the 3DLBP or Sigmoid 3DLBP curves intersection. We did not run the same type of fusion for the N+R vs. A protocol due to poor performance in other experiments but we keep the results for the sake of completeness.

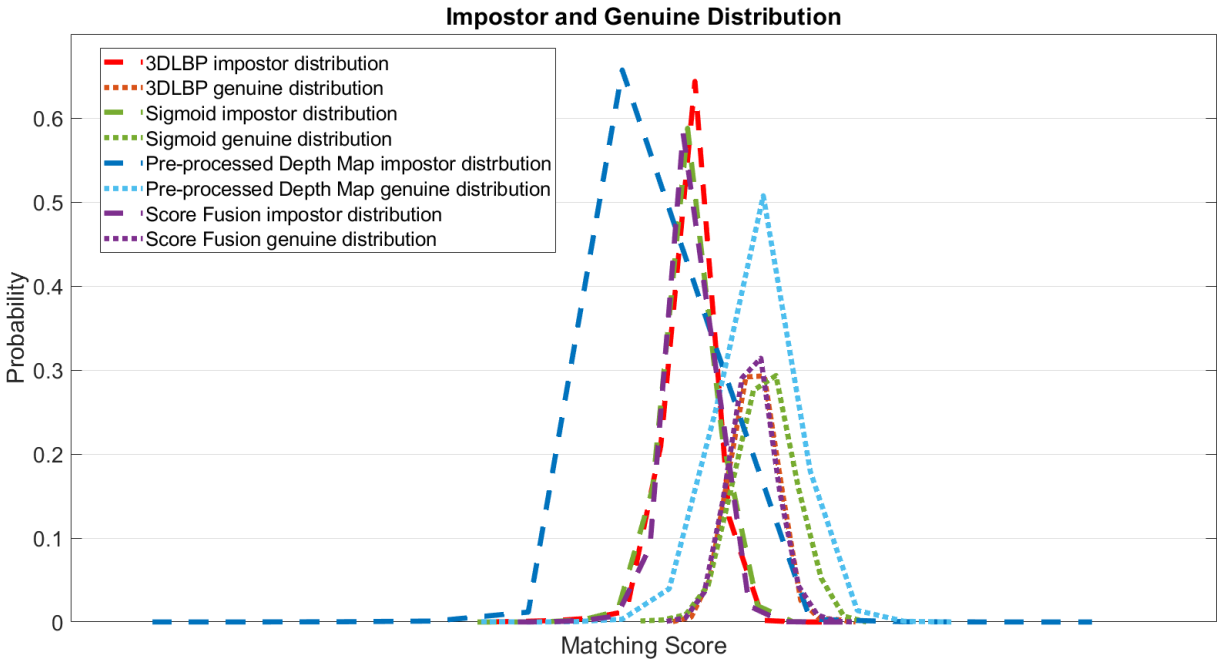


Figure 6.6: Impostor and Genuine distribution curves. In the high-resolution case it is possible to see that the fusion of the three DIs (3DLBP, Sigmoid, and Pre-processed depth map), spite being less spread than in the low-resolution scenario, still yields worst performance.

Lastly, even though there is a huge difference in the data quality, the proposed method demonstrated to be effective on both high and low-resolution data, pointing towards a great deal of resolution invariance.

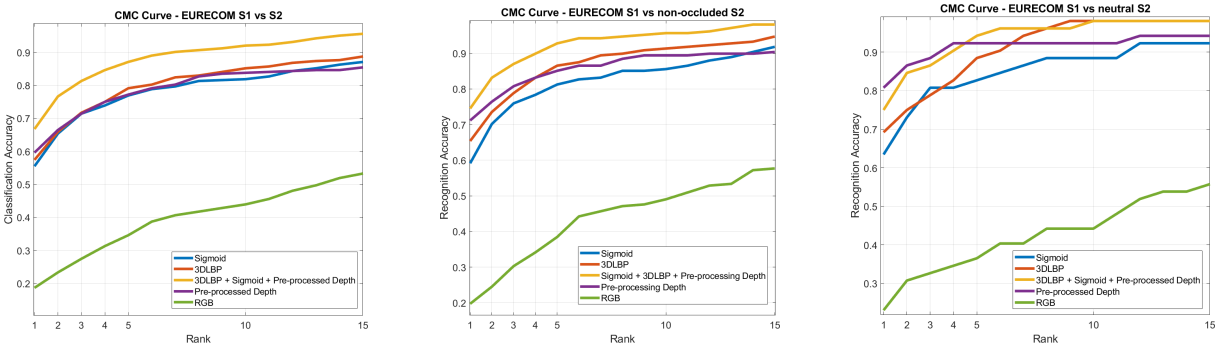
6.3 Feature Extraction Mode

In this section, the results of both CABNet-C and CABNet-FE for Feature Extraction mode are reported. This case is more interesting than the Classification Mode since the shallow CNNs are trained only with high-resolution images and evaluated on very low-resolution images (there is no fine-tuning procedure). The protocols utilized here are the same as the ones reported in section 6.2

Initially, we evaluate the performance with the CABNet-C. This will give us a better understanding about the resolution invariance properties of our approach. Finally, we show the results obtained with the CABNet-FE.

6.3.1 Low-Resolution 3D Face Scans

Figure 6.7 shows the CMC curves obtained with the CABNet-C on the EURECOM Kinect Face dataset, for the same protocols as described in Section 5.1.3.



(a) *Protocol (i)* (b) *Protocol (ii)* (c) *Protocol (iii)*

Figure 6.7: Results on EURECOM Kinect Face dataset. CMC curves obtained utilizing the CABNet-C. The identity for each subject is defined as the sample with the higher cosine similarity among probe and gallery.

Initially, one can see that depth-based data (3DLBP, Sigmoid 3DLBP, and pre-processed depth map) perform better than RGB data. Since the network is only trained with high-resolution data and does not go through any fine-tuning process, this could indicate a larger transmission of knowledge through resolution with depth data. In addition, the EURECOM Kinect Face dataset contains lots of occlusions, which seem to impair the recognition when using RGB imagery. The pre-processed depth map and the proposed 3DLBP-based DIs, instead, demonstrated higher robustness to such nuisances. Looking at the CMC curves, it is possible to see that on rank-1 the fusion of the three DIs had the better performance in the first two experiments from the protocol. In the experiment for S1 vs. Neutral, the pre-processed depth outperforms the

fusion and the 3DLBP-based DIs. However, the fusion of the three DIs outperforms all the other types of features on rank-5 onwards. Practically, having a better performance in the latter ranks allows building a system that, given a list of possible identities, produces a higher level of confidence than the genuine subject as presented there. In any case, results obtained with our CABNet-C on 3D data are much higher than those obtained using RGB images or those obtained using the pre-trained deep VGG16 network of (Kim et al., 2017) as shown in Table 6.4.

Table 6.4: EURECOM dataset: Rank-1 results. The gallery is from Session 1, while the probe set is composed of: *(i)* Session 2, *(ii)* the three variants without occlusions from Session 2, *(iii)* neutral scans from Session 2. The results here are obtained with the CABNet-C.

Method	*(i)*	*(ii)*	*(iii)*
RGB Data	18.68%	19.71%	23.08%
Pre-processed depth map	59.62%	71.15%	**80.76%**
Sigmoid 3DLBP	55.49%	59.13%	63.46%
3DLBP	57.41%	65.38%	69.23%
3DLBP + Sigmoid 3DLBP	58.52%	62.98%	59.62%
Pre-processed depth map + 3DLBP + Sigmoid 3DLBP	**66.76%**	**74.52%**	75.00%
VGG (Kim et al., 2017)	13.9%	14.8%	13.5%
LED 3D (Mu et al., 2019)	34.34%	38.76%	44.8%

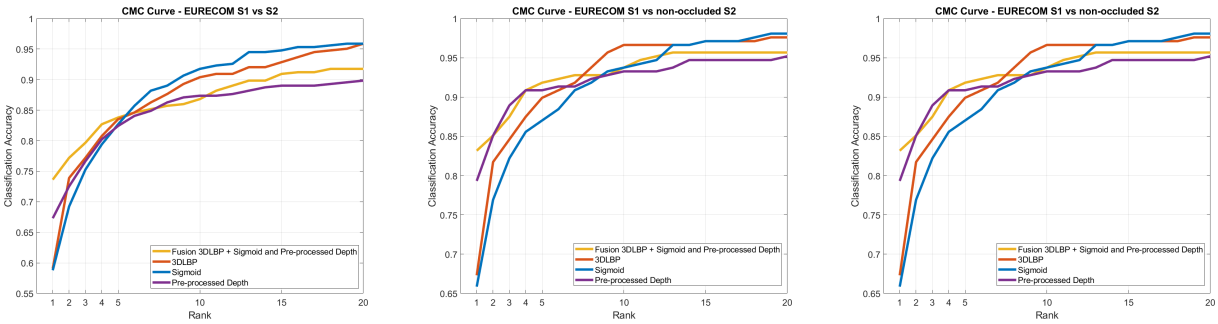
The results showed in Table 6.4 highlight that the resolution differecens matter. Given that no fine-tuning is performed, the results indicate that the proposed DIs maintain a rather pronounced resolution invariance. Another important aspect observed in the CMC curves is that occlusions affect the depth data more than the 3DLBP-based DIs do.

With CABNet-FE, described in Section 4.3 and illustrated in Figure 4.15, we utilize information from different layers to build a more compact and robust feature representation for 3D faces. The results obtained with this shallow CNN are shown in Table 6.5.

Table 6.5: Rank-1 results on EURECOM Kinect Face dataset (best results in bold). The gallery is from Session 1, while the probe set is composed of: *(i)* Session 2, *(ii)* the three variants without occlusions from Session 2, *(iii)* neutral scans from Session 2. This results are achieved with the CABNet-FE.

Method	*(i)*	*(ii)*	*(iii)*
RGB Data	12.91%	13.94%	15.38%
Pre-processed depth map	67.30%	79.36%	82.69%
Sigmoid 3DLBP	58.79%	65.86%	67.30%
3DLBP	59.06%	67.30%	73.07%
3DLBP + Sigmoid 3DLBP	67.58%	75.48%	78.84%
Pre-processed depth map + 3DLBP + Sigmoid 3DLBP	**73.26%**	**83.17%**	**84.61%**
VGG (Kim et al., 2017)	13.9%	14.8%	13.5%
LED 3D (Mu et al., 2019)	34.34%	38.76%	44.8%

When looking for single features, it is possible to see that the pre-processed depth map has the better performance. Another important aspect is that the fusion greatly increases the system performance, in this scenario the fusion is the concatenation of the features and not a weighted sum of scores, as in the Classification Mode approach. This makes sense, whilst in the Classification Mode approach the output are probabilities of a face be from distinct subjects, in the Feature Extraction Mode approach we have features that describe each face. The CMC curves shown in Figure 6.8 highlights this.



(a) *Protocol (i)* (b) *Protocol (ii)* (c) *Protocol (iii)*

Figure 6.8: Results on EURECOM Kinect Face dataset. CMC curves for the three experiments with the Feature Extraction Mode approach: (a) Gallery from Session 1 vs. probe from Session 2 (7 variants each), (b) Gallery from Session 1 vs. probe with the three variants without occlusions from Session 2, and (c) Gallery from Session 1 vs. neutral probes from Session 2. The scales on the vertical axis are different. These results were obtained with the CABNet-FE.

The fairer comparison of our work on Feature Extraction Mode is with the work of Mu et al. (2019). This is because both utilize high-resolution data to train a convolutional neural network and utilizes low-resolution data to validate it. The main difference is that, on its original work (Mu et al., 2019) utilizes a Kinect v2 dataset, whilst we utilize a Kinect v1 dataset. To be able to compare both approaches we have run the approach in (Mu et al., 2019) with the EURECOM dataset (Kinect v1 dataset). The results presented in Table 6.5 show that with 3D models having a great deal of noise our approach performed better.

Table 6.6 shows the best results obtained with both shallow CNNs architectures proposed for the Feature Extraction Mode. It is possible to see that, in a cross-resolution scenario, the CABNet-FE outperforms the CABNet-C. This is a clear evidence of how the features obtained from throughout the whole architecture plays an important role in cross-resolution 3D face recognition.

Table 6.6: Rank-1 results on EURECOM dataset comparing between both shallow CNNs proposed as feature extractors, using fusion of pre-processed depth map, 3DLBP and Sigmoid 3DLBP.

CNN Architecture	*(i)*	*(ii)*	*(iii)*
CABNet-C	66.76%	74.52%	75.00%
CABNet-FE	**73.26%**	**83.17%**	**84.61%**

6.3.2 High-Resolution 3D Face Scans

Tables 6.7 shows the results obtained with the proposed Feature Extration Mode on a high-resolution 3D face dataset, the Bosphorus dataset.

Comparing the results from Tables 6.7 and 6.3, it is possible to see that the accuracy values obtained using the CABNet-C as classifier or feature extractor are quite similar in this case. The lower accuracy in the N-vs-NN and N-vs-A protocols, differently from the Classification Mode, can be ascribed to the lack of expressive scans in the training data (FRGC). As said previously, Kim *et al.* (Kim et al., 2017), utilized pose variations and synthetic expressions on the fine-tuned process of the deep VGGFace (Parkhi; Vedaldi; Zisserman, 2015) network, this is to specifically deal with the lack of expression on the training data. However, obtaining such a larger accuracy when expressive scans are included in the probe set required collecting a huge amount of data. In addition, we recall that our input images are of size 100×100, against 224×224 of the VGG. In any case, our shallow CNN using the DIs demonstrated to be more accurate in the N-Vs-N scenario, indicating a very promising capability of capturing the relevant identity traits.

As was the case with the Classification Mode, we also can observe that the fusion of the three types of DIs (Pre-processed depth map, 3DLBP, and Sigmoid 3DLBP) decreases the performance. We did not run the same type of fusion for the N+R vs. A due to poor performance in other experiments, but we keep the results for the sake of completeness.

Lastly, we show the results obtained with the CABNet-FE on a high-resolution scenario. Looking at Tables 6.8 we can see that the CABNet-FE, focused on cross-resolution feature

Table 6.7: Rank-1 results on Bosphorus dataset using CABNet-C as a feature extractor. Protocols: *Neutral vs. Neutral*, *Neutral vs. Non-Neutral*, and *Neutral vs. All*.

Method	*N vs. N*	*N vs. NN*	*N vs. A*	*N+R vs. A*
RGB data	85.82%	71.57%	72.31%	**94.03%**
Pre-processed depth map	98.42%	63.97%	65.78%	60.60%
Sigmoid 3DLBP	99.21%	58.91%	61.03%	61.92%
3DLBP	**100.0%**	63.24%	65.16%	67.11%
3DLBP + Sigmoid 3DLBP	**100.0%**	66.81%	68.55%	70.53%
3DLBP + Sigmoid 3DLBP + Pre-processed depth map	92.13%	62.76	64.29%	-
VGG (Kim et al., 2017)	99.2%	**95.0%**	**95.3%**	-

Table 6.8: Rank-1 results on Bosphorus dataset using the CABNet-FE as a feature extractor. Protocols: *Neutral vs. Neutral*, *Neutral vs. Non-Neutral*, and *Neutral vs. All*.

Method	*N vs. N*	*N vs. NN*	*N vs. A*
RGB data	88.06%	64.26%	65.50%
Pre-processed depth data	97.64%	57.96%	60.04%
Sigmoid 3DLBP	95.28%	50.63%	52.98%
3DLBP	98.42%	55.17%	57.44%
3DLBP + Sigmoid 3DLBP	**100.0%**	60.18%	62.27%
3DLBP + Sigmoid 3DLBP + Pre-processed depth data	92.13%	56.56	58.42%
VGG (Kim et al., 2017)	99.2%	**95.0%**	**95.3%**

extraction, underperforms the CABNet-C. This can be explained by the fact that we have created the CABNet-FE architecture focused on extracting information throughout the layers of the whole CNN, and in a high-resolution scenario this process can bring redundant information that impacts the system. Nonetheless, our idea with this approach is focused on cross-resolution scenarios.

Comparing the results showed in Tables 6.7, 6.8 and 6.9 we can see that the fusion with the pre-processed depth map does not contribute for increasing the performance. This result differs from the low-resolution scenario and can mean that the data in pre-processed depth, in high-resolution scenarios, is not complementary to the 3DLBP-based DIs. This is different for the fusion of our low-level hand-crafted features (3DLBP and Sigmoid).

Table 6.9: Rank-1 results on Bosphorus dataset comparing both networks being utilized as feature extractor. The results here are the best for each experiment.

CNN Architecture	*N vs. N*	*N vs. NN*	*N vs. A*
CABNet-C	**100.0%**	**66.81%**	**68.55%**
CABNet-FE	**100.0%**	60.18%	62.27%

Chapter 7

CONCLUSIONS

In this thesis, we addressed the problem of 3D face recognition. We did so because spite 2D face recognition has achieved state-of-the-art results, there are still some problems that impact negatively the biometric systems based on face recognition, such as occlusion and pose variations.

Another important aspect is that the state-of-the-art results obtained by 2D methods are achieved, in most cases, with Deep CNNs, which brings a scenario where huge computational power and data are needed.

In a scenario where user interaction is not possible the utilization of traditional 3D scanners is not feasible. In this aspect, devices like Microsoft Kinect, which works in a similar way as a standard 2D camera, can be a solution.

The problem with data captured by devices like Kinect is that they are very low-resolution and most of the approaches in the literature are based on high-resolution data. There is also a shortage of depth-based data (*e.g.*, 3D models, cloud points) to allow training effectively a Deep CNN.

With this in mind, we came up with our main hypothesis that using an intermediate feature representation for depth-based data would allow us to use shallow CNNs for 3D face recognition. Ideally, this feature representation should have a higher degree of resolution invariance, allowing us to train on high-resolution and use it on low-resolution scenarios seamlessly. This cross-resolution characteristic is desired because there is more high-resolution data publicly available than low-resolution.

Therefore, we proposed a hybrid approach for 3D face recognition, which is focused on minimizing the amount of data, the computational power and the processing time required in

the training stage, while being able to operate close to state-of-the-art methods and being able to transfer the learning made on high-resolution data to low-resolution data. The proposed hybrid approach also allows us to operate in classification or feature-extraction modes.

Experimental results obtained by our hybrid approach on EURECOM Kinect Face dataset, a low resolution depth dataset, showed a rank-1 recognition rate of 90.75% with the CABNet-C on the hardest case of classification mode, and 73.26% on the feature extraction mode with the SLFR, which are better than the rates obtained by related state-of-the-art methods with the same protocol and dataset.

Therefore, we concluded that the proposed hybrid approach helps to attenuate the cross-resolution differences and that the utilization of an input built with more discriminative data, such as low-level hand-crafted features, allows the utilization of shallow CNN for 3D face recognition.

7.1 Main Contributions

The original contributions of this thesis are related to: (i) the proposition of new Descriptor Images (DI) for 3D representation; (ii) the proposition of shallow CNN architetures that togheter with the proposed DIs allow 3D face classification in closed-set and open-set scenarios; (iii) the proposition of an approach to transfer learning from high to low resolution 3D face data; and (iv) the proposition of a shallow learned feature representation to allow learning a feature representation based on different types of DIs.

7.1.1 Descriptor Images For 3D Representation

The utilization of DIs has been proved as an effective way to represent 3D data and utilize it to train a CNN. There are some advantages in utilizing this approach. The first one is the possibility of creating different types of representations from the same subject data and doing an ensemble of shallow networks. This can be an alternative to the standard data augmentation approach.

In this thesis, we utilized two types of hand-crafted features to generate the DIs, the 3DLBP and a variation proposed by us which is the Sigmoid 3DLBP. In both modes the fusion of these data increased the accuracy rates of the 3D face recognition. This corroborates our ideas of utilizing the DIs and the feasibility of building an ensemble of shallow networks.

7.1.2 Shallow CNNs For 3D Face Recognition

The results obtained in the experiments (Sections 6.2 and 6.3) show that it is possible to do 3D face recognition utilizing shallow CNNs. We tailored two different CNNs, CABNet-C and CABNet-FE, and showed that, even with a small amount of low-resolution 3D face data and a very simple CNNs architecture, it is possible to obtain results close to the state-of-the-art.

When considering open-set scenarios, the CABNet-FE and the SLFR (Shallow Learned Feature Representation) showed great potential. Even though our method has limitations, mainly related to facial expressions, the results obtained highlight the capacity demonstrated by the learned features for cross-resolution transfer learning. Since the FRGC dataset does not contain many faces with expressions a way to deal with this limitation would be to utilize faces from other high-resolution dataset with expressions.

The results obtained with the SLFR have shown that concatening different features improved the 3D face recognition rates. This is an evidence that fusing different types of features can be a way to overcome some limitations of our approach. Exploring other types of DIs and assuring that they are more robust to changes in facial expressions can help to build, with feature-fusion, a more robust 3D face descriptor.

7.1.3 Transfer Learning From High To Low-Resolution

We concluded that the DIs help to attenuate differences caused by the resolution of the 3D models. This becomes more evident when we look at the results with the feature extraction mode. In the classification mode, there is a fine-tuning phase which helps the network to adapt to a more low-resolution domain, this is also the reason for the classifier performing better than the feature extraction. Looking at the results of the feature extraction mode (Section 6.3) it is possible to see that, even though the performance is worst, the network performs fairly well, given the fact that we dont fine-tune our network and the amount of data is smaller than a Deep approach would use.

When dealing with low-resolution images, in the classification mode, the proposed approach performs close to the state-of-the-art methods that uses either DCNN or hand-crafted features. With high-resolution data, the hybrid approach can compete with state-of-the-art methods, but only on the Neutral vs. Neutral setting; this is mainly because the lack of expressions on the training data for the FRGC.

Despite this latter limitation, to the best of our knowledge our approach is the first one capa-

ble of obtaining competitive performance on depth data that span a large variety of resolutions. In addition to this, we proved our framework can be trained efficiently even with small amount of data, thus making it a viable solution for applications where training data is limited.

It is also important to note that the hand-crafted feature behind the generation of DIs plays an important role in the result. Since 3DLBP and Sigmoid 3DLBP describe low-level features they can act as a way to summarize information to a CNN. Utilizing a hand-crafted feature that does not describe low-level features will not have the same effect. Taking into consideration that the face is a smooth surface the utilization of local descriptors makes the DIs more robust to noise.

Comparing both modes it is possible to conclude that, on scenarios where the number of subjects is known, the classification approach is more adequate. In these scenarios, there is the need for a fine-tuning process and this is only possible if the number of subjects is previously established. However, every time the number of people changes, it is necessary to redo the fine-tuning process.

7.1.4 Shallow Learned Feature Representation (SLFR)

The feature extraction mode seems more adequate when the conditions are not stable, since it is trained in one dataset but it can be utilized to recognize DIs generated from other sources, not necessarily from the same original type of data.

Looking at the results for the feature extraction mode (Section 6.3) there is strong evidence for higher levels of resolution invariance. The comparison with the deep VGG16 highlights this; the drop in performance from Bosphorus to EURECOM highlights how much the change in resolution affects VGG while our method still performed in an acceptable way. Taking into consideration, the limited amount of data for training points towards a great deal of information on the DIs and that the shallow CNN can learn discriminative information from them.

In the feature extraction mode, the SLFR works more adequately in the cross-resolution scenario. The better performance in the high-resolution feature extraction mode for the CABNet-C means that there is some loss of resolution invariance.

From the latter results, we also concluded that our architecture has yet some limitations when dealing with expressions, this can be explained by the lack of expressions on the training data. However, the higher performance with respect to other solutions based on deep networks suggests interesting future perspectives that can pave the way for the development of smaller but still effective networks for 3D face recognition systems.

7.2 Future Work

The contributions and results presented in this thesis pointed to the need for further studies. The following are some suggestions for topics that can be addressed in future work:

- Investigation of other types of DIs;
- Evaluation of the performance of the proposed method on reconstructed face models from 2D images;
- Investigation of other ways for fusing scores and features;
- Evaluation of ways to select more discriminative features from different layers of CABNet-FE architecture.

7.3 Published Papers

Publications of scientific articles that reflected the level of development of research and contributions to the literature were made continuously throughout the doctorate. The papers published to date are presented below:

- CARDIA NETO, João Baptista; MARANA, Aparecido Nilceu; FERRARI, Claudio; BERETTI, Stefano; DEL BIMBO, Alberto. Deep Learning from 3DLBP Descriptors for Depth Image Based Face Recognition. In: 2019 International Conference on Biometrics (ICB). IEEE, 2019. p. 1-7.
- CARDIA NETO, João Baptista; MARANA, Aparecido Nilceu; FERRARI, Claudio; BERETTI, Stefano; DEL BIMBO, Alberto. Depth-Based Face Recognition by Learning from 3D-LBP Images. In: 12th Eurographics Workshop on 3D Object Retrieval (3DOR). 2019. p. 55-62.
- CARDIA NETO, João Baptista; MARANA, Aparecido Nilceu. 3D Face Recognition with Reconstructed Faces from a Collection of 2D Images. In: Iberoamerican Congress on Pattern Recognition. Springer, Cham, 2018. p. 594-601.
- CARDIA NETO, João Baptista; MARANA, Aparecido Nilceu. Utilizing deep learning and 3DLBP for 3D face recognition. In: Iberoamerican Congress on Pattern Recognition. Springer, Cham, 2017. p. 135-142.

During the doctorate, we have collaborated with other research projects, not related to this thesis, whose results were published in the following papers:

- TAVARES, Henrique Leal; CARDIA NETO, João Baptista; PAPA, João Paulo; COLOMBO, Diego; MARANA, Aparecido Nilceu. Tracking and Re-identification of People Using Soft-Biometrics. In: 2019 XV Workshop de Visão Computacional (WVC). IEEE, 2019.
- TAVARES, Henrique Leal; CARDIA NETO, João Baptista; PAPA, João Paulo; COLOMBO, Diego; MARANA, Aparecido Nilceu. People Identification Based on Soft Biometrics Features Obtained from 2D Poses. In: Brazilian Conference on Intelligent Systems. 2020.

REFERENCES

AN, Z.; DENG, W.; HU, J. Deep transfer network for face recognition using 3d synthesized face. In: *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017. p. 1–4.

BERRETTI, S.; BIMBO, A. D.; PALA, P. 3d face recognition using isogeodesic stripes. *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 32, n. 12, p. 2162–2177, 2010.

BESL, P. J.; MCKAY, N. D. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 14, n. 2, p. 239–256, fev. 1992. ISSN 0162-8828. Disponível em: <http://dx.doi.org/10.1109/34.121791>.

BOLLE, R.; PANKANTI, S. *Biometrics, Personal Identification in Networked Society: Personal Identification in Networked Society*. Norwell, MA, USA: Kluwer Academic Publishers, 1998. ISBN 0792383451.

BOLLE, R. M. et al. The relation between the roc curve and the cmc. In: Ieee. *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, 2005. p. 15–20.

BONDI, E. et al. Reconstructing high-resolution face models from kinect depth sequences. *IEEE Trans. on Information Forensics and Security*, v. 11, n. 12, p. 2843–2853, Dec 2016.

CAI, Y. et al. A fast and robust 3d face recognition approach based on deeply learned face representation. *Neurocomputing*, Elsevier, v. 363, p. 375–397, 2019.

CHETVERIKOV, D.; STEPANOV, D.; KRSEK, P. Robust euclidean alignment of 3d point sets: the trimmed iterative closest point algorithm. *Image and Vision Computing*, v. 23, n. 3, p. 299 – 309, 2005. ISSN 0262-8856. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0262885604001179>.

CRAMER, J. *The Origins of Logistic Regression*. [S.l.], dez. 2002. Disponível em: <https://ideas.repec.org/p/tin/wpaper/20020119.html>.

DECANN, B.; ROSS, A. Relating roc and cmc curves via the biometric menagerie. In: Ieee. *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013. p. 1–8.

DENG, X. et al. A multi-scale three-dimensional face recognition approach with sparse representation-based classifier and fusion of local covariance descriptors. *Computers & Electrical Engineering*, Elsevier, v. 85, p. 106700, 2020.

DRIRA, H. et al. 3D face recognition under expressions, occlusions, and pose variations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, v. 35, n. 9, p. 2270–2283, Sept 2013.

DROSOU, A.; MOSCHONAS, P.; TZOVARAS, D. Robust 3D face recognition from low resolution images. In: *Int. Conf. of the BIOSIG Special Interest Group*, 2013. p. 1–8.

FALTEMIER, T. C.; BOWYER, K. W.; FLYNN, P. J. A region ensemble for 3-D face recognition. *IEEE Trans. on Information Forensics and Security*, v. 3, n. 1, p. 62–73, March 2008.

FAWCETT, T. An introduction to {ROC} analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861 – 874, 2006. ISSN 0167-8655. {ROC} Analysis in Pattern Recognition. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016786550500303X>.

GAO, G. et al. Cross-resolution face recognition with pose variations via multilayer locality-constrained structural orthogonal procrustes regression. *Information Sciences*, Elsevier, v. 506, p. 19–36, 2020.

GILANI, S. Z.; MIAN, A. Learning from millions of 3d scans for large-scale 3d face recognition. *CoRR*, abs/1711.05942, 2017. Disponível em: <http://arxiv.org/abs/1711.05942>.

GILANI, S. Z.; MIAN, A. Learning from millions of 3D scans for large-scale 3D face recognition. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. p. 1896–1905.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016.

GRANGER, S.; PENNEC, X. Multi-scale em-icp: A fast and robust approach for surface registration. In: Springer. *European Conference on Computer Vision*, 2002. p. 418–432.

GROSS, R. et al. Multi-pie. *Image Vision Comput.*, Butterworth-Heinemann, Newton, MA, USA, v. 28, n. 5, p. 807–813, maio 2010. ISSN 0262-8856. Disponível em: <http://dx.doi.org/10.1016/j.imavis.2009.08.002>.

HAN, J.; MORAGA, C. The influence of the sigmoid function parameters on the speed of backpropagation learning. In: MIRA, J.; SANDOVAL, F. (Ed.). *From Natural to Artificial Neural Computation*, 1995. p. 195–201. ISBN 978-3-540-49288-7.

HERNANDEZ, M.; CHOI, J.; MEDIONI, G. Laser scan quality 3-D face modeling using a low-cost depth camera. In: *European Signal Processing Conf. (EUSIPCO)*, 2012. p. 1995–1999.

HUANG, Y.; WANG, Y.; TAN, T. Combining statistics of geometrical and correlative features for 3d face recognition. In: *Proceedings of the British Machine Vision Conference*, 2006. p. 90.1–90.10. ISBN 1-901725-32-4. Doi:10.5244/C.20.90.

HUYNH, T.; MIN, R.; DUGELAY, J.-L. An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In: *ACCV 2012, Workshop on Computer Vision with Local Binary Pattern Variants, Daejeon, Korea, November 5-9, 2012 / Published also as LNCS, Vol 7728, PART 1*, 2012. Disponível em: <http://www.eurecom.fr/publication/3849>.

JAIN, A. K.; ROSS, A.; PRABHAKAR, S. An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, v. 14, p. 4–20, 2004.

JAIN, A. K.; ROSS, A. A.; NANDAKUMAR, K. *Introduction to Biometrics*. [S.l.: s.n.], 2011. ISBN 0792383451.

JIANG, L.; ZHANG, J.; DENG, B. Robust rgb-d face recognition using attribute-aware loss. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, 2019.

KAKADIARIS, I. A. et al. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 29, n. 4, p. 640–649, abr. 2007. ISSN 0162-8828. Disponível em: <http://dx.doi.org/10.1109/TPAMI.2007.1017>.

KIM, D. et al. Deep 3D face identification. In: *IEEE Int. Joint Conf. on Biometrics (IJCB)*, 2017. p. 133–142.

KLARE, B. F. et al. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015. p. 1931–1939.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F. et al. (Ed.). *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012. p. 1097–1105. Disponível em: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.

LEARNED-MILLER, E. et al. Labeled faces in the wild: A survey. In: ____. *Advances in Face Detection and Facial Image Analysis*. Cham: Springer International Publishing, 2016. p. 189–248. ISBN 978-3-319-25958-1. Disponível em: <https://doi.org/10.1007/978-3-319-25958-1$_{8¿}$.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Research, v. 521, n. 7553, p. 436–444, 2015. ISSN 0028-0836. Disponível em: <http://dx.doi.org/10.1038/nature14539>.

LEE, Y. et al. Accurate and robust face recognition from rgb-d images with a deep learning approach. In: *British Machine Vision Conf. (BMVC)*, 2016. p. 123.1–123.14.

LI, B. et al. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In: *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, 2013. p. 186–192. ISSN 1550-5790.

LI, H. et al. Towards 3d face recognition in the real: a registration-free approach using fine-grained matching of 3d keypoint descriptors. *Int. Journal of Computer Vision*, Springer, v. 113, n. 2, p. 128–142, 2015.

LI, H.; SUN, J.; CHEN, L. Location-sensitive sparse representation of deep normal patterns for expression-robust 3d face recognition. In: *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017. p. 234–242.

MALTONI, D. et al. *Handbook of fingerprint recognition*. [S.l.]: Springer Science & Business Media, 2009.

MANTECÓN, T. et al. Depth-based face recognition using local quantized patterns adapted for range data. In: *IEEE Int. Conf. on Image Processing (ICIP)*, 2014. p. 293–297.

MANTECÓN, T. et al. Visual face recognition using bag of dense derivative depth patterns. *IEEE Signal Processing Letters*, v. 23, n. 6, p. 771–775, June 2016.

MARTIN, A. et al. The det curve in assessment of detection task performance. In: , 1997. p. 1895–1898.

MIAN, A.; BENNAMOUN, M.; OWENS, R. An efficient multimodal 2d-3d hybrid approach to automatic face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 29, n. 11, p. 1927–1943, nov. 2007. ISSN 0162-8828. Disponível em: <http://dx.doi.org/10.1109/TPAMI.2007.1105>.

MIN, R. et al. Real-time 3D face identification from a depth camera. In: *Int. Conf. on Pattern Recognition (ICPR)*, 2012. p. 1739–1742.

MU, G. et al. Led3d: A lightweight and efficient deep approach to recognizing low-quality 3d faces. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. p. 5773–5782.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 2010. (ICML'10), p. 807–814. ISBN 978-1-60558-907-7. Disponível em: <http://dl.acm.org/citation.cfm?id=3104322.3104425>.

NETO, J. B. C.; MARANA, A. N. *3D Face Recognition Using Kinect*. Dissertação (Mestrado) — Universidade Estadual Paulista Júlio de Mesquita Filho - UNESP, 2014.

NGUYEN, V. et al. How to choose deep face models for surveillance system? In: ____. *Modern Approaches for Intelligent Information and Database Systems*. Cham: Springer International Publishing, 2018. p. 367–376. ISBN 978-3-319-76081-0.

OJALA, T.; PIETIKAINEN, M.; HARWOOD, D. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Ieee. *Proceedings of 12th International Conference on Pattern Recognition*, 1994. v. 1, p. 582–585.

OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 24, n. 7, p. 971–987, 2002. ISSN 0162-8828.

PARKHI, O. M.; VEDALDI, A.; ZISSERMAN, A. Deep face recognition. In: , 2015.

PASZKE, A. et al. Pytorch: An imperative style, high-performance deep learning library. In: WALLACH, H. et al. (Ed.). *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019. p. 8024–8035. Disponível em: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.

PHILLIPS, P. J. et al. Preliminary face recognition grand challenge results. In: Ieee. *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, 2006. p. 15–24.

PRABHAKAR, S.; PANKANTI, S.; JAIN, A. Biometric recognition: Security and privacy concerns. *IEEE Security and Privacy*, IEEE Computer Society, Los Alamitos, CA, USA, v. 1, p. 33–42, 2003. ISSN 1540-7993.

QIAN, Y.; DENG, W.; HU, J. Unsupervised face normalization with extreme pose and expression in the wild. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. p. 9851–9858.

RUSSELL, P. N. S. *Artificial Intelligence: A Modern Approach*. 3rd. ed. Prentice Hall, 2010. (Prentice Hall Series in Artificial Intelligence). ISBN 0136042597, 9780136042594. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=C37CE53726FB431E5815F9B1E573BFD6>.

SAVRAN, A. et al. Bosphorus database for 3d face analysis. In: Springer. *European Workshop on Biometrics and Identity Management*, 2008. p. 47–56.

SINGH, M. et al. Identity aware synthesis for cross resolution face recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

SKANSI, S. *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*. Springer International Publishing, 2018. (Undergraduate Topics in Computer Science). ISBN 9783319730042. Disponível em: <https://books.google.com.br/books?id=5cNKDwAAQBAJ>.

SOUZA, G. B. de et al. Deep texture features for robust face spoofing detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, v. 64, n. 12, p. 1397–1401, 2017.

SPREEUWERS, L. Fast and accurate 3D face recognition. *Int. Journal of Computer Vision*, v. 93, n. 3, p. 389–414, July 2011.

WAYMAN, J. Technical testing and evaluation of biometric identification devices. In: JAIN, A.; BOLLE, R.; PANKANTI, S. (Ed.). *Biometrics*. [S.l.]: Springer US, 2002. p. 345–368. ISBN 978-0-387-28539-9.

WOLF, L.; HASSNER, T.; MAOZ, I. *Face recognition in unconstrained videos with matched background similarity*. [S.l.]: IEEE, 2011.

Wu, X.; Ward, R.; Bottou, L. WNGrad: Learn the Learning Rate in Gradient Descent. *ArXiv e-prints*, mar. 2018.

YANG, L.; HANNEKE, S.; CARBONELL, J. A theory of transfer learning with applications to active learning. *Machine learning*, Springer, v. 90, n. 2, p. 161–189, 2013.

YIN, X.; LIU, X. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, v. 27, n. 2, p. 964–975, Feb 2018. ISSN 1057-7149.

YOSINSKI, J. et al. How transferable are features in deep neural networks? In: GHAHRAMANI, Z. et al. (Ed.). *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014. p. 3320–3328. Disponível em: <http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf>.