Revisiting face processing with light field images

Valeria Chiesa

▶ **To cite this version:**

Valeria Chiesa. Revisiting face processing with light field images. Signal and Image processing. Sorbonne Université, 2019. English. NNT : 2019SORUS059 . tel-03020423

HAL Id: tel-03020423
https://theses.hal.science/tel-03020423v1

Submitted on 23 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

EURECOM
Sophia Antipolis

SORBONNE UNIVERSITÉ

DISSERTATION

In Partial Fulfilment of the Requirements for the
Degree of Doctor of Philosophy from Sorbonne University
Specialization: Signal Processing

Revisiting face processing with light field images

Valeria CHIESA

Defense scheduled on 04/06/2019 before a committee composed of:

President	**Prof. Bernard MERIALDO**, EURECOM, Sophia Antipolis, France
Reviewer	**Prof. Alice CAPLIER**, GipsaLab, Grenoble, France
Reviewer	**Prof. Enrico VEZZETTI**, Politecnico di Torino, Torino, Italy
Examiner	**Prof. Paulo Lobato CORREIA**, Instituto de Telecomunicações, Lisboa, Portugal
Thesis Advisor	**Prof. Jean-Luc DUGELAY**, EURECOM, Sophia Antipolis, France

Abstract

Nowadays, in a time where cities contain millions of people and where travelling across the world is becoming easier and easier, the necessity of automatically identifying a person is starting to be compelling. The physical appearance and the behavioural characteristics have been discovered useful to univocally describe a person. The analytic study of the human body measures with the aim of recognising or verifying the identity of a person, is called *biometrics*, literally "life measure".

In the last century, several biometric traits have been investigated according to the most updated technologies available at the moment, improving recognition, computational time and memory capacity. Starting from the 90's, research on biometrics has received a huge boost thanks to the interest raised by academic institutions, government agencies and private companies. Moreover, the diffusion of new instruments, able to perform faster analyses, and to store more data, simplifies the development of biometric systems. Together with the advancement of processing machines, innovative acquisition devices providing non conventional data have been developed. The investigation of the impact of new technologies applied to specific topics is a mandatory step in order to improve the performances on biometrics.

The main goal of this thesis is to present a non-conventional acquisition technology as light field, to study face analysis performances using images collected with a specific camera, to compare the results with those obtained using data from similar devices and to prove the major benefit provided by the use of up-to-date device over standard cameras in biometric field.

When this thesis started, the literature on face analysis with light field data was bare. The scarcity of biometric data (and particularly of human face images) collected with plenoptic cameras has been tackled with a systematic acquisition of a light field face database, now publicly available. Thanks to the collected data, it has been possible to design and develop experiments on face analysis. Moreover, an exhaustive baseline of a comparison between two RGB-D technologies has been created to sustain future studies.

During the period of this thesis, the interest on light field technology applied on face analysis has grown and the necessity of a survey on algorithm customized for plenoptic images has become compulsory. Thus, a complete overview on existent methods has been compiled.

Abstract

All the algorithms designed and developed have been tested within the context of the H2020 European Projects with the aim of making faster and more user friendly automatic security border controls.

Résumé

Aujourd'hui, à une époque où les villes comptent des millions d'habitants et où il est de plus en plus facile de voyager à travers le monde, la nécessité d'identifier automatiquement une personne commence à s'imposer. L'aspect physique et les caractéristiques comportementales sont utiles pour décrire une personne de manière univoque. La biométrie, littéralement "mesure de la vie", est l'étude analytique des mesures du corps humain dans le but de reconnaître ou de vérifier l'identité d'une personne.

Au cours du dernier siècle, plusieurs traits biométriques ont été étudiés selon la technologie la plus moderne disponible actuellement, améliorant la reconnaissance, le temps de calcul, les problèmes de mémoire rencontrés. Depuis les années 90, la recherche en biométrie a connu un essor considérable auprés des institutions universitaires, des agences gouvernementales et des entreprises privées. De plus, la diffusion de nouveaux equipements, capables d'effectuer des analyses plus rapides et de stocker davantage de données, simplifie le développement des systèmes biométriques. Avec l'évolution des machines de traitement, des dispositifs d'acquisition innovants fournissant des données non conventionnelles se sont développés. L'étude de l'impact des nouvelles technologies appliquées à des sujets spécifiques est une étape obligatoire afin d'améliorer les résultats de la reconnaissance.

L'objectif principal de cette thèse est de présenter une technologie d'acquisition non conventionnelle, d'étudier les performances d'analyse du visage en utilisant des images collectées avec une caméra spéciale, de comparer les résultats avec ceux obtenus en élaborant des données à partir de dispositifs similaires et de démontrer le bénéfice apporté par l'utilisation de dispositifs modernes par raport à des caméras standards utilisées en biométrie.

Au début de la thèse, la littérature sur l'analyse du visage à l'aide de données "light field" a été étudiée. Le probléme de la rareté des données biométriques (et en particulier des images de visages humains) recueillies à l'aide de caméras plénoptiques a été résolu par l'acquisition systématique d'une base de données de visages "light field", désormais accessible au public. Grâce aux données recueillies, il a été possible de concevoir et de développer des expériences en analyse du visage. De plus, une base de référence exhaustive pour une comparaison entre deux technologies RGB-D a été créée pour appuyer les études en perspective.

Pendant la période de cette thèse, l'intérêt pour la technologie du plenoptique appliquée

à l'analyse du visage s'est accrue et la nécessité d'une étude d'un algorithme dédié aux images "light field" est devenue inconturnable. Ainsi, une vue d'ensemble complète des méthodes existantes a été élaborée.
Tous les algorithmes conçus et développés ont été testés dans le cadre d'un projet européen H2020 dans le but de rendre plus rapides les contrôles automatiques de sécurité aux frontières.

Contents

List of Figures

List of Tables

List of Abbreviations

ACER Average Classification Error Rate

AIPA Active Impostor Presentation Attack

APCER Attack Presentation Classification Error Rate

BF Bona Fide

BPCER Bona fide Presentation Classification Error Rate

EER Equal Error Rate

EPI EPIpolar images

FAR False Acceptance Rate

FMR False Match Rate

FMR10000 lowest FNMR for FMR minor of 0.1%

FNMR False Non Match Rate

FRR False Rejected Rate

HOG Histogram of Oriented Gradients

ICAO International Civil Aviation Organization

ICPA Identity Concealer Presentation Attack

ISO International Organization for Standardization

JMD Joint Mini Database

KFD EURECOM Kinect Face Database

LBP Local Binary Pattern

LBP3D Local Binary Pattern for 3D data

LDF Landmark Depth Features

LF Light Field

LFALBP Light Field Angular Local Binary Pattern

LFFD IST-EURECOM Light Field Face Database

LGBP Local Gabor Binary Pattern

LiFFID Light Field Face and Iris Database

MTF Modular Transfer Function

OF Open Face

PA Presentation Attack

PAD Presentation Attack Detection

PCA Principal Component Analysis

PLDF Principal Landmark Depth Features

PR PRecision

PSF Point Spread Function

ROI Region Of Interest

RSD Relative Standard Deviation

SL Structure Light

SLBP Spacial Local Binary Pattern

SNR Signal to Noise Ratio

SVM Support Vector Machine

ToF Time of Flight

ZeroFMR lowest FNMR for FMR equal to 0%

Chapter 1

Introduction

1.1 Context and motivation

In 1996, the access to the Olympic Games Village in Atlanta was protected by a hand geometry system. In 2009, in India the Aadhaar program started with the aim to provide all participants with a personal number based on iris pattern and face image to benefit from healthcare. In 2011, Motorola introduced in its last smartphone model the possibility to be unlocked with fingerprints. In 2017, 120 countries all over the world were releasing biometric passports to their citizens. Nowadays, some important banks allow money transfers verifying the identity of the person by remote fingerprints.

The use of automatic recognition systems based on biometrics in a world populated by billiards of people is, progressively, entering in the daily life of everyone, changing our life quality and raising new ethical issues.

Biometric traits, difficult to steal and impossible to forget, are nowadays used for several applications, from smart objects (phones, laptops, cars) to interface with private informations or to access protected zones. In several countries, the health system is moving through a biometric identification for unconscious patients, in order to identify them and to follow therapeutic paths. In crime forensics, biometry is already extensively used to recognize possible witnesses or criminals on a crime scene from videos, fingerprints, DNA. In university residences, where lodgers often change, keys are replaced by biometric features.

The hypothetical biometric system benefits are unquestionable. In reality, research has to deal with available technology and has to collaborate on developing devices and sensors able to extent to the most possible biometry potentiality. That goes through a systematic study of the technology impact on different applications.

The goal of this Ph.D. dissertation is the investigation of the influence of light field imagery on face analysis. In fact, although light field cameras have been barely used in biometric, I believe that it has high potential thanks to its ability to catch the 3D information of close objects. Moreover, its relatively simple structure and data post processing may be a plus with respect to other well known comparable technologies such as structured light devices.

Some research questions that this work aims to answer are:

1. *Which are the characteristics required in a database that includes light field information? Which protocols should be used for its realization?*
2. *In which context does light filed technology perform better than structured light one?*
3. *Which is the impact of defocusing in soft biometrics analysis?*
4. *How can light field information enhance face recognition rate?*
5. *Can light field technology lead toward more robust anti-spoofing strategies?*

1.2 Thesis contributions

This Ph.D. work aims to study the impact of the light field technology on face analysis. The contribution spans several aspects of face analysis research.

- **Data acquisition**: Three light field face data collections have been carried out during the period of this thesis. The first, in collaboration with *Instituto de Telecomunicações, Instituto Superior Técnico, Universitade de Lisboa*, in Portugal, has been presented in [6]. The database, now publicly available, consists in a collection of images representing 100 individuals acquired during two sessions, each one including 20 face variations (occlusions, emotions, poses, illuminations). At the date of writing this thesis, the *IST-EURECOM Light Field Face Database* has been used to train and test several algorithms customized for light field data.
  Two multimodal databases have been collected in the context of PROTECT project. Data related to 3D face have been acquired, processed and elaborated in order to fuse the results with other biometric traits. The results of the elaboration of the first data collection have been presented in [7]: all biometric traits have been evaluated separately and as a multimodal system, proving the good performances of Lytro Illum camera for face recognition.

- **Comparison and modelling**: The potential of light field data for face recognition have been compared with evaluations done on structured light images in order to create, through exhaustive experiments, a baseline and to suggest the best

technology according to the scenario to deal with ([8], [9]). The impact of light field images on soft biometric traits has been investigated and modelled ([10]).

- **Algorithm**: Two innovative algorithms customized for light field images have been proposed. The first one, presented in [11], tackles face recognition problem using the plenoptic data property to be rendered as sub-aperture pictures. The second, described in [12], shows a method to detect face presentation attacks exploiting the pair RGB-D images.

- **Survey**: By the end of this Ph.D. thesis, literature on the use of light field technology in face analysis has been growing. Thus, in collaboration with *Instituto de Telecomunicações, Instituto Superior Técnico, Universitade de Lisboa*, in Portugal, *Hochschule* of Darmstadt, in Germany and *Institut de recherche en informatique et systèmes aléatoires* (IRISA), in France, a survey on the use of plenoptic data for landmark estimation, face recognition and presentation attack detection has been compiled ([13]).

1.3 Thesis outline

The thesis is divided in two main parts foreshorten by an introduction about biometrics. In Part I, advanced sensors for face analysis are presented, while in Part II, the impact of light field on face analysis is studied.

In the introductory part (Chapter 2), a brief overview of biometrics is provided. Some selected key events in the history of biometrics are listed in order to introduce the motivation of this thesis. General biometric challenges are shown, with a particular attention to face as biometric trait. Multimodal biometric systems are described introducing a data acquisition held in the context of the European Project PROTECT. Difference between hard and soft biometrics is mentioned with the aim of developing the concept of age estimation and gender recognition. For a complete understanding of the impact of biometrics on daily life, some main ethical issues are commented.

In Part I, light field sensors and in particular Lytro devices are presented. A theoretical discussion about the used technology and the possible representation of plenoptic data is reported to provide the reader with basic knowledge to fully comprehend the following studies. Two of the most known face light field databases are introduced and described (Chapter 3). Then Lytro camera is opposed to other 3D sensors used in face analysis. Previous researches found out that structured light sensors are more performant for face analysis, thus, a deeper comparison is carried out between Kinect V1 camera and Lytro device (Chapter 4).

In Part II, the impact of light field camera on soft biometric traits, such as gender recognition and age estimation from face, is investigated (Chapter 5). After a careful compilation of state of the art based on light field data for face recognition and presentation attacks, two methods customised for plenoptic images are proposed and proved to be more performant than current algorithms (Chapter 6, Chapter 7).

1.4 Author comments

Starting from 2012, the popularity of light field imaging in research has grown: several international conferences and journals have organized workshops and special issues on light field imaging. National and European projects finalise to investigate the properties of plenoptic images have received consistent grants and even a JPEG standard[1] has been customized for light field data.

Most of the works presented until now are based on Lytro technology. Even though Lytro Inc, was going to release Lytro Immerge, a device to create virtual reality, it has been acquired in 2018 and it ceased its activities. However, its employed and technologies have been incorporated in a bigger company.

Since Lytro Power, Lytro Desktop and Matlab Light Field Toolbox software are not officially available any more, it is not possible to provide links and references. Nevertheless, it is still possible to find softwares and codes stored in repository.

At the date of writing this manuscript, it is possible to purchase professional light field devices from Raytrix GmbH company.

[1] https://jpeg.org/jpegpleno/lightfield.html

Chapter 2

Introduction to biometrics

In this chapter, a biometric scenario is presented. An overview of the most important steps in biometric history are listed in order to understand the actual context. Biometric characteristics and challenges are introduced, highlighting the main tools used in recognition systems. Face as biometric trait is described more in details because of its main role in this thesis. The relevance of face as biometric trait is demonstrated by the amount and the variety of face databases present in literature.
All known biometric traits is perfect: some of them are difficult to collect, others, such as face, do not provide perfect results. For this reason, the fusion among several biometric measures is done to improve the performances. An example of multimodal database is the one acquired in the context of H2020 project PROTECT. In Section 2.3, the elaboration and the performances of analysis on *PROTECT multimodal database* are shown. The section is inspired by the work presented in [7].
Finally, a short introduction on soft biometric traits and on the difference between hard and soft biometrics are described. Some ethical issues related to the use of biometric systems risen during the last century are mentioned.

2.1 History of biometry[1]

The word *biometrics* is defined as "the use of detailed information about someone's body, for example the patterns of colour in their eyes, in order to prove who they are" by Cambridge Dictionary and it is composed by "$\beta\acute{\iota}o\omega\varsigma$" and "$\mu\acute{\epsilon}\tau\rho o\nu$", life and measure in ancient Greek. Nowadays, biometrics is often associated with automatic recognition systems, but it has not been always like this.

Contrary to what is generally claimed, biometric has been used since the beginning of

[1]This section has been inspired by https://www.biometricupdate.com/201802/history-of-biometrics-2

civilization. Even not considering the human beings ability of recognizing each other through face characteristics, it is possible to find evidence of fingerprints used as signature in Babylon clay tiles created around 500 B.C. In Egypt, physical characteristics are annotated to identify trusted traders and to guarantee their reputation. Joao de Barros, a Portuguese traveller and writer of XVI century reports the use of fingerprints in Chinese business transactions and in children identification. Comments about person verification from fingerprints are found also in the Persian book *Jaamehol-Tawarikh*, attributed to Rashid-al-Din Hamadani, a statesman, physician and historian lived between 1247 and 1318 [14].
In 1685, Govard Bidloo describes friction ridge skin details in *Anatomy of the Human Body*. Marcello Malpighi collects his studies on epidermal ridges in *De externo tactus organo anatomica observatio* in 1665 [15] and in 1684 Nehemiah Grew analyses the ridges pattern [16]. The first scientist theorising the finger print unicity is Johann Christoph Andreas Mayer in 1788 in *Anotomical Copper-plates with Appropriate Explanations* [17].

During the second half of 19th century, the population grows and urban agglomerations increase thanks to the industrial revolution. The introduction of railways facilitates the mobility of goods and people. The idea of justice changes and harsher punishments are considered for habitual criminals. In this scenario, authorities can not rely on their ability to distinguish individuals and the need of a formal identification becomes a necessity.

Several approaches are used. Probably, the most known is the method proposed by Bertillon in 1870 in France, based on anthropometric measures. The procedure of acquiring pictures of criminals is already employed, but Bertillon standardises it in 1888. A different approach is used in India in 1858, where Sir William Herschel records the handprint of the employees in order to identify the persons during the payday [18]. Herschel provides the evidence of this procedure to solve a dispute with the Scottish physician Henry Faulds[2] about the authorship of fingerprint use for identification. Despite the similar idea, while Herschel is using recognition for commercial purpose, Faulds proposes a judicial technique to find criminals as described in his publication on *Nature* in 1880. In order to have a support in the scientific community, Faulds contacts Francis Galton, a British anthropologist, that publishes several works on the probability that different individuals have the same fingerprints. The collaboration leads to many discussions between the two scientists. Galton's research, joint with work of Sir Edward Henry, Inspector General of the Bengal Police, paves the way to the scientific development and the application of fingerprint recognition in justice domain, without acknowledging the colleague [14]. Azizul Haque and Hem Chandra Bose, Henry's workers, elaborate a method to store fingerprints in order to facilitate the search. The Henry Classification

[2] An exhaustive biography on Faulds life can be found here: http://galton.org/fingerprints/faulds.htm

System is the precursor to Federal Bureau of Investigation (FBI) fingerprint search AFIS (Automated Fingerprint Identification System). In 1901 the Henry System of Fingerprint Classification is used to create the Fingerprint Branch at New Scotland Yard.

Also in USA fingerprint identification is used at the end of 19th century. In the Mark Twain's tale *A Thumb-Print and What Come Of It* is described how a man claims his innocence comparing his fingerprint with the one left on a crime scene, proving the existence of this procedure in 1883 [19].
In 1903 New York State Prisons start to use fingerprint to identify and to store the information related to prisoners. After New York, other states adopt the same system. On July 1st 1921, an Act of Congress establishes the Identification Division of the FBI. Meanwhile, the Bertillon system based on anthropometric traits collapses: two men, with identical aspect, are sentenced in Kansas. The Bertillon measures result to be the same for both persons, demonstrating the inefficiency of the system (Figure 2.1).

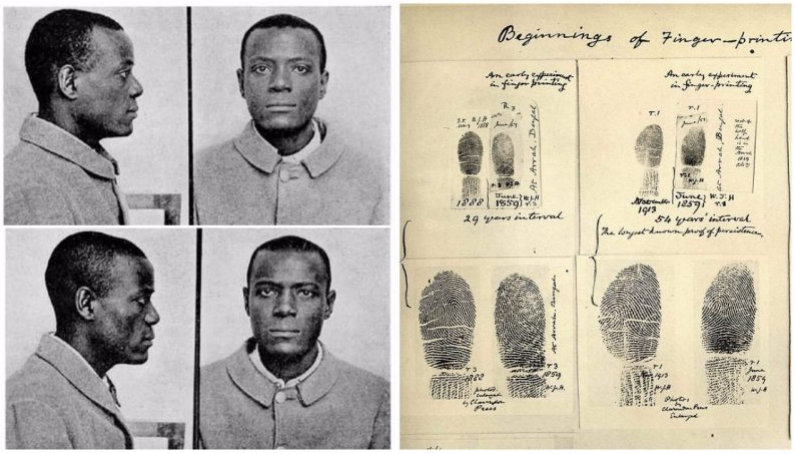


Figure 2.1 – In 1903, Will West arrives at Leavenworth Penitentiary denying previous imprisonment. When the clerk run the Bertillon recognition system, he finds the file of William West, an inmate that is serving a life sentence in the same prison. Since Bertillon system fails, fingerprints are checked, finding a substantial difference[3].

The use of iris pattern as biometric trait has been proposed by the ophthalmologist Frank Burch in 1936. It takes 50 years to formalize the concept that no two iris are alike: only in 1987 Drs Leonard Flom and Aran Safir get a patent for the iris identification concept [20]. With the collaboration of Dr John Daugman, Flom and Safir develop and deliver a prototype for the Defence Nuclear Agency in the United States in 1995.

Although face is one of the most obvious biometric trait to consider, it is not easy to acquire and process. In 1960s the first semi-automatic face recognition systems are developed by Woodrow Bledsoe for US Government [21]. The identification of keypoints in the image is still managed by an human operator, but the process is computed by an

automatic system.

In the same period in Sweden, Gunnar Fant studies the physiological components of acoustic speech, a central concept for speaker recognition [22]. The model elaborated is expanded by Joseph Perkell, including the motion of tongue and jaw in the behavioural analysis. The first prototype system for speaker recognition is developed in 1976 for the US Air Force.
In the 1970s, the only commercial automatic biometric recognition system is based on fingerprints and in 1974 it is flanked by hand geometry system used for a fast access control. The patent for dynamic signature acquisition, hand identification and vascular pattern recognition are awarded in 1977 the first and 1985 the others, respectively by Veripen, Inc., David Sidlauskas and Joseph Rice.

In 1980s the National Institute of Standards and Technology (NIST), that is already studying the fingerprint recognition problem, creates a speech processing group to investigate speaker recognition potentialities. The first standard for exchange of fingerprint minutiae data is published in 1986 by the National Bureau of Standards in collaboration with the American National Standard for Information Systems (ANSI).
The Biometric Consortium held its first meeting in October 1992. The role of this Consortium is to investigate testing methods, to develop standards, interoperability and government cooperation. At first it involves only government agencies, than other organizations, such as the InterNational Committee for Information Technology Standards (INCITS) and the International Organization for Standardization (ISO), join the Consortium in order to increase the impact.

In 1987 Kirby and Sirovich develop an algorithm for face recognition based on principal component analysis [23]. Thanks to that a few years later Turk and Pentland make a real time face recognition [24].
In 1990s several university and research institutes as well as government agencies study face recognition. Each group works on its own database, usually small and with few variations. In 1993 the Defence Advanced Research Products Agency (DARPA) and the DoD Counterdrug Technology Development Program Office sponsor the FacE REcognition Technology (FERET) project [25]. It has the goal of creating a common database representing human faces and of establishing a baseline. The database is collected in three phases under the leadership of Harry Wechsler at George Mason University and Jonathan Phillips at the Army Research Laboratory in Adelphi, Maryland, for a total of 14,126 facial images representing 1199 different individuals. After each acquisition phase, the main existent face recognition algorithms are tested with the same protocol to evaluate their performances.

In 1994, biometrics starts to be used for border and access controls and the Immigration and Naturalization Service Passenger Accelerated Service System (INSPASS[4]) is implemented. The system allows authorized passengers to bypass US immigration inspections presenting the card where is encoded the hand geometry information. The same biometric trait is used in 1996 to control and protect the access to the Olympic Village in Atlanta. During the duration of the Game, over 65000 individuals are enrolled for over 1 million transactions.

In 1998 the FBI starts to collect a DNA in Combined DNA Index System (CODIS) for law enforcement purpose. The follow year, the International Civil Aviation Organization (ICAO) starts to investigate the compatibilities of the new biometric technologies with the international standards.
The FBI's Integrated Automated Fingerprint Identification System (IAFIS) becomes operational in 1999. The standards associated to IAFIS allow comparing fingerprint collected by different systems. A national network is established in US in order to check historical background and identify persons. Moreover, fingerprints are checked with the one discovered on crime scene of all States. The system is still operational.

The terrorist attacks of 2001 trigger the necessity of enforced controls, specially in US, where a significant amount of founding is allocated for biometric research. During the Super Bowl in Tampa, Florida, in 2001 face recognition is used to identify particular individuals, rising privacy concerns in the general public. In 2003 the International Organization for Standardization (ISO) establishes the standardization of generic biometric technologies ISO/IEC JTC1 Subcommittee 37 (JTC1/SC37). In 2003 ICAO introduces biometric identification information into passports and chooses face as the globally interoperable biometric. In the same year, the European Biometric Forum is created as independent European organization. In 2003 the first eGate are installed in US airports among with New York, Washington and Houston.
In 2004 the United State Visitor and Immigrant Status Indication Technology (US-VISIT) program becomes operational. Biometric traits such as face image and fingerprints are collected for each person crossing US borders to match the identity and verify the whether the travel is admissible or not. A personal identification card provided by biometric information is required to all US government employees and contractors to access to government buildings. Some US states create statewide palm print database for law enforcement purpose.

In 2008, 60 countries all over the world (including USA and most of the European countries) are adopting biometric passport. From 2006, machines able to read facial

[4]For more information: https://web.archive.org/web/20070203233102/http://www.biometrics.org/REPORTS/INSPASS.html

images from the passports are mandatory on border controls in all European countries and, from 2009, also fingerprints are included.

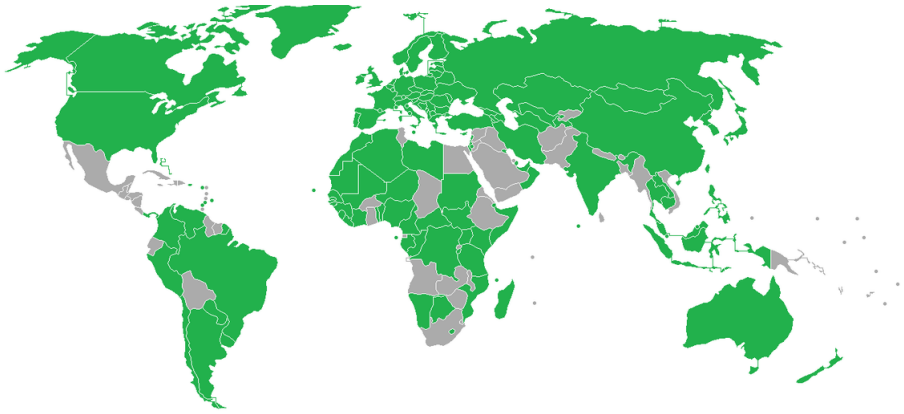


Figure 2.2 – Countries adopting biometric passport in 2018 (Image from https://en.wikipedia.org/wiki/Biometric_passport)

The collection of the world biggest biometric database started in India in 2009. Aadhaar is an identification number based on biometric data and provided by Unique Identification Authority of India (UIDAI) to all Indian residents. In early 2000s, the Indian government faces an people identification issue. In fact, many citizens do not have any identification paper while others are found with several passports with different details. The necessity to create a unique identification system give place to a massive data collection. Face imaging, fingerprint and iris pattern are acquired for each participant in order to facilitate the access and to control fraud attempts in health, education and government services. Even if the enrolment is voluntary, it became indirectly mandatory: bank institutes and phone companies start to link personal accounts to Aadhaar number. In 2015, the Supreme Court raise doubts about the respect of the fundamental right of privacy, confirming and ensuring the registration optionality. In addition to privacy, concerns for security are shown. The collection of a so huge number of data gives place to occasional errors such as acquisition mistakes and fraud attempts. Lately Aadhaar database is found vulnerable to hacking attacks, compromising biometric data of billion of people. In 2018 the program involves over than 90% of Indian population[5].

In 2011, biometrics meets smart technology: Motorola introduces the first fingerprint scanner in a mobile phone but without attracting users. Only a few years later Apple invents Touch ID on iPhone 5S. Thanks to the popularity of the brand, the system obtains the expected success. Galaxy S5, realized by Samsung in 2014, is the first Android smartphone with fingerprint sensor. In 2016 some financial companies such as MasterCard and Visa allow a fingerprint authentication for money transfers.

[5] https://uidai.gov.in/aadhaar_dashboard/index.php

In 2018, it is possible to count more that 3000 eGates all over the world.

2.2 Biometrics in a nutshell

The need for recognizing an individual in a group of people is a well known problem. Among all the possible solutions that have been found so far, three main categories are identified.

- *Token-based*: The ownership of a specific object can allow the access to the system and to confirm the identity of a person (badge, key).
- *Knowledge-based*: The knowledge of a password can allow the access to the system and to confirm the identity of a person (pin, code).
- *Biometric-based*: The physical and behavioural aspect of the individual can allow the access to the system and to confirm the identity of a person (face, iris).

All three possible solutions have pros and cons: a token can be stolen, can be given and it is not possible to be sure about the identity of the person carrying it. A password can easily be forgotten or learned by an other individual. However, a pin can be changed frequently without additional monetary costs.
Often, in order to grant an higher security level, the recognition systems are based of a combination of token, knowledge and biometric: for example, ATM withdrawal requires both a card (token) and a pin (knowledge).
Biometric-based systems are generally more secure and user friendly. Contrarily to other identification tools, biometry can not be forgotten or lost and it is less subject to spoofing.

Biometric systems are generally used for *identification* or *verification*. Even if the ideas can appear similar, the reasons for identifying an individual are quite different from the ones for verification.
In the former case, the goal of the system is, starting from a subject sample, to find a correspondence in a set of identities. The set can be limited (company employees or wanted criminals) or really large (like in Aadhaar program). Normally, identification systems are developed to be used without user collaboration and, often, they have to deal with challenging environment problems. A bigger amount of possible matching makes the identification process slower or computationally more complex than the verification one.
In the latter case, two samples are compared and the identity correspondence is verified. This kind of systems are becoming used to control the access to specific information or sectors, as for border controls, where the identity of travellers is matched with passport data. Usually, users want to be recognised and the environment where the verification is performed is known. Only in the last years, more challenging scenarios are considered:

the arrival of smartphone with cameras and finger print sensors pushes companies (such as banks) to rely on verification through mobile apps, where the environment can not be controlled.

According to the different application, biometric recognition systems face a wide range of challenges.

- *Environment* - A known and controlled scenario increases system performances: background, illumination, distance and occlusions are all important factors to be considered.
- *Sensor* - A system based on a known sensor is significantly more reliable than a system created on a generic device. Different camera brands create different noises and have different characteristics. Verification processes based on voice suffer particularly from this kind of problems because of the large variety and not uniformity of available microphones.
- *User* - User is also important to design a recognition system. A biometric trait acquired by a trained operator is often preferable than a measure done by a person using the system for the first time. A reluctant user can even tries to hide his or her identity with occlusions making more challenging the process. Face expressions, excited voice, sweat fingerprints can also damage the sample collection.

Each biometric trait is more or less impacted by the described problems. In particular, biometric characteristics are divided into *physiological* or *behavioural* or a *mixture* of both. The first category includes biometric traits that are not depending by the willing of the subject, such as DNA, fingerprints, finger veins or iris patterns. Signature or keystroke are considered behavioural traits because they are not related to physical aspect of an individual but they are impacted by his behaviour. These traits are more difficult to acquire without the consent of the subject and they are easier to spoof. Most of the biometrics are a mixture of both: voice and gait, for example, are determinate by the physical vocal chords characteristics and by legs length but also by the emotional status of the individual, by the voice tone or the walk speed.

Besides physical and behavioural distinction, biometric traits are classified according to other criteria. The choice of one trait respect to an other one should be done after an accurate analysis of the application. In [26], Jain et al. define seven factors that have to be considered in the choice of a biometric trait.

- *Universality* - Every individual accessing the application should possess the trait. Face has high universality, while signature is not considered completely universal because it is possible for a user to not be able to sign.

- *Uniqueness* - The trait should be different across individuals. Iris pattern is reliable to highly differentiate across people, while keystroke may be similar for several persons.
- *Permanence* - The trait should be invariant over a sufficient period of time with respect to the matching algorithm. Voice print change, specially during adolescence, while fingerprints do not naturally variate in all life long.
- *Collectability* - The acquisition of the trait should be feasible with the available technology and it should not cause any inconvenience to the individual. The collection of a retina scan can be not comfortable, while the digitalization of hand geometry is quite straightforward.
- *Performance* - The recognition accuracy and the resources required should be compliant with the used system. Usually face recognition requires high computational power to achieve a modest accuracy, while fingerprints-based systems have better recognition rate.
- *Acceptability* - The target population should be willing to present the trait to the system. Individuals are more willing to provide signature than retina scan.
- *Circumvention* - The trait should not be easily reproducible or imitable using artefact. Face is a easy target, while hand veins are less sensible to spoofing attacks.

	Universality	Uniqueness	Permanence	Collectability	Performance	Acceptability	Circumvention
Face	high	low	medium	high	low	high	low
Fingerprint	medium	high	high	medium	high	medium	high
Hand geometry	medium	medium	medium	high	medium	medium	medium
Keystrokes	low	low	low	medium	low	medium	medium
Hand veins	medium	medium	medium	medium	medium	medium	high
Iris	high	high	high	medium	high	low	high
Retinal scan	high	high	medium	low	high	low	high
Signature	low	low	low	high	low	high	low
Voice print	medium	low	low	medium	low	high	low

Table 2.1 – Classification of some biometric traits according with the factors identified by Jain et al. in [26].

As shown in Table 2.1, none of the existent biometric traits is expecting to meet all the requirements described. Thus, a compromise between criteria has to be considered: low uniqueness and performance could be accepted in order to have high collectability and acceptability. In these cases, biometric recognition could be combined with token or password identification to increase the performances and the security.

2.2.1 Face analysis

Face is one of the most important and used human biometric traits. The observation of physical and emotional aspects is crucial in human society in order to recognize people and to comprehend their feelings (angry, sadness, happiness etc). In 90's, several medical

studies [27, 28] have shown that face recognition process is performed in a particular area of human brain, notably in the inferior temporal cortex. The specific functionality of this brain area is still under investigation, but it is proved that the detection and the recognition of a face is not related to the identification of other biometric traits (such as voice) or objects. In 2008, Gruter et al. [29] suggest that 2.5% of United States population is effected of prosopagnosia, a congenital or acquired brain disease that prevents individual recognising faces. Subjects suffering prosopagnosia can still identify a known person thanks to other physical characteristics such as voice, hair colour or clothes but are not able to recognise even their own face, while all other cognitive skills stay unvaried. As some individuals have deficit in face recognition, others have better-than-average face recognition ability. In 2015 London Metropolitan Police creates a team of super-recogniser police officers to help in wanted criminals identification.

The natural human approach leads to deep research in automatic face recognition. As mentioned in [26], face is highly accepted as recognition trait in most of the cultures. The acquisition of face images does not require a strong user collaboration, like for retina scans or voice collection, and it is totally not invasive. Even if during the first 20 years of human life, the face may be subjected to big variations, it is always more or less recognisable. Face as biometry has also some drawbacks: it may be not unique for an automatic system (and sometimes even for a human beings). Relative and specially homozygous twins have really similar physiognomy while other biometric traits are different (as iris or fingerprints). Moreover, the performance of recognition systems based on face images are lower than processes based on other biometrics traits.

The study of face analysis is not restricted identification: a human face has to be detected, the image has to be polished in order to minimize different illumination or pose variation effects, a liveness control can protect the system from impostor attacks. Ageing and make-up impact has to be investigate to improve the system reliability. Images are acquired with a wide range of sensors, starting from 2D conventional cameras, passing through micro-cameras embedded in wearable devices, to 3D sophisticated systems.

Nowadays, face is one of the three biometric traits recognized by ICAO[6], together with iris and fingerprint, and it is the only one mandatory in biometric passports in more than 120 countries all over the world.

The wide literature and research developed in the last 20 years are proven by the amount of face databases publicly available today.

Currently, there are over 100 publicly available face databases. Table 2.2 includes an overview of a set of selected prominent existing face databases (a more complete list

[6] https://www.icao.int/publications/Documents

can be found in [30]). This information is complemented in Table 2.3, which tabulates the variations addressed in these databases, sorted according to their release date. The tables here presented have been compiled for the work published in [6].

The selection is guided by the willingness to include data collected with innovative sensors. For instance, mobile devices (smartphone and laptop) are used to collect the *MObile BIOmetry* (MOBIO) database [31]. Both visible and infrared spectra are acquired in *Surveillance Cameras face* (SCface) database [32] in an uncontrolled indoor environment. Facial expressions in dynamic 3D space are analysed in *Binghamton University 3D Facial Expression* (BU-3DFE) [33]. The *EURECOM Kinect Face database* [34] provides RGB-D face images, captured by Kinect sensor, to evaluate how face recognition technology can benefit from this imaging sensor. A more explicative description and use of *EURECOM Kinect Face database* is provided in Section 4.3.

2.3 Multimodal system

The use of a single biometric trait to recognize an individual could not be sufficient to obtain the desired results. As shown in Table 2.1, none of the most acceptable traits have high performances in recognition. For this reason, several biometric traits can be integrated to create multimodal systems, in contrast with unimodal processes. The possible failure of one feature are balanced by the performances of the others.

In order to create a multimodal system, chimeric databases are often used: biometrics from different individuals and different databases are assigned arbitrarily to the same identity. This databases are quite common but they do not allow the analysis of possible correlation between biometric traits from the same subject.

2.3.1 Multimodal systems: PROTECT multimodal database

An example of multimodal non-chimerical database is the *PROTECT multimodal database* presented in [7]. The database, collected in the contest of Horizon 2020 European project PROTECT (described in Section 2.3.2), includes 9 biometric traits (Figure 2.3): 2D face, 3D face, thermal face, iris, voice, finger veins, hand veins, anthropometrics.

The 47 participants are representative of a wide range of age, gender (57% male and 43% female) and ethnicity (English, Arabic, Chinese, Hindi, Italian, Polish, Punjabi, Spanish, Turkish among others). In Figure 2.4 some statistics about age and ethnicity are shown. The acquisitions are done in order to simulate a border control biometric corridor where

Table 2.2 – Overview of a few prominent face databases with different characteristics.

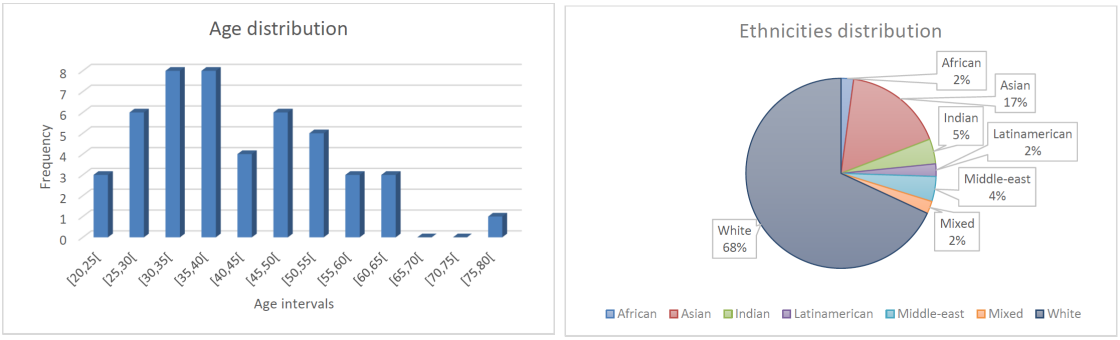
Database name	Year	N. Subjects	Image type	Image modality	Spacial resolution
Yale B [35]	2001	28	Grayscale	2D	640x480
FERET [25]	2003	1199	Grayscale/Color	2D	256x384
MIT-CBCL [36]	2004	10	Color	2D	768x576
FEI [37]	2006	200	Color	3D	640x480
FRAV3D [38]	2007	106	Color	2D texture 2.5D range data 3D VR models	N/A
Bosphorus [39]	2008	105	Color	2D texture 3D coordinate map	1600x1200
Multi-PIE [40]	2009	337	Color	2D	3072x2048
MOBIO [41]	2010	150	Color	2D	Different resolutions up to 2048x1536
3DFRD [42]	2010	118	Color	2D 3D range data	751x501
SCface [43]	2011	130	Color/IR	2D	Different resolutions up to 426x320
BU-3DFE [44]	2013	100	Color	3D 3D dynamic	1040x1329
Kinect Face DB [34]	2014	52	Color	2D - 2.5D 3D - Video	640x480
LiFFID [45]	2016	112	Grayscale	2D 2D rendered	1054x1054 120x120
IST-EURECOM LFFD [6]	2016	100	Color	4D light field 2D rendered 2D depth map	15x15x434x625 2022x1404 2022x1404

Table 2.3 – Overview of variations in a few prominent face databases.

Database name	**Facial occlusions**	**Facial expressions**	**Views/Poses**	**Different dates**	**Illuminations**
Yale B [35]	N	N	9 poses	N	64 levels
FERET [25]	Glasses - Hair	2 expressions	10 poses	1 year	Standard Low
MIT-CBCL [36]	N	N	Frontal - ±30°	N	Y
FEI [37]	N	Neutral - Smile	10 poses	N	Standard Low
FRAV3D [38]	N	Max 2 gesture	Frontal X,Y,Z axis turn	N	N
Bosphorus [39]	Glasses Hair - Hand	35 expressions	14 poses	N	N
Multi-PIE [40]	N	Neutral - Surprise Squint - Smile Disgust -Scream	15 poses	N	19 levels
MOBIO [41]	N	Different expressions (uncontrolled)	Uncontrolled	N	Uncontrolled LED
3DFRD [42]	N	Smile - Talk Mouth and/or eyes opened/closed	N	N	N
SCface [43]	N	N	Uncontrolled	N	Uncontrolled LED
BU-3DFE [44]	N	25 expressions	Frontal ±45°	N	N
Kinect Face DB [34]	Eye - Nose - **finire**	Neutral - Smile Open moth	Frontal Right /left profile	Up to 15 days	Y
LiFFID [45]	N	Neutral - Smile	6 poses	N	Uncontrolled LED
IST-EURECOM LFFD [6]	Glasses - Mask Sunglasses - Hat Eye/Mouth	Neutral - Smile Surprise - Angry	7 poses	Up to 6 mouth	Standard Low - High

a traveller could be identified while walking. The database is provided upon request[7] and it can be used only for research purposes.

Figure 2.3 – Samples from the *PROTECT Multimodal DB* database [7]: 2D face, anthropometrics, 3D face, thermal face, iris, hand veins, finger veins.



(a) (b)

Figure 2.4 – Age (5-year intervals distribution) (Figure 2.4a) and ethnicities distributions (Figure 2.4b) of subjects in the corpus.

Each biometric is evaluated singularly in order to create a baseline and then fused to test the multimodal performances of the database. Voice print, despite of is present in the collection, has not been processed for reasons not related to the research.

Unimodal Evaluation Methods

2D face: The 2D face dataset contains videos captured from three cameras, which are set up along the length of the recording area, roughly the same distance apart and with variations in height, yaw and pitch. The frames with better image quality are manually

[7] For details see projectprotect.eu

selected from all videos. Some criteria used are: full-frontal face, no glasses, person looking directly to the camera if possible. Due to limited number of such frames, only one image per user is selected as gallery sample. For the probe samples, an automatic detection tool is used to select the frames where the face could be detected as the majority of the frames contain faces in low quality. A commercial software by Visage Technologies[8] with state of the art face tracking and analysis is used for evaluating the recognition performance (which underlining algorithms cannot be disclosed).

3D face: The 3D face raw data is processed with the Lytro Power software which allows the RGB picture and the corresponding depth map to be automatically extracted. An alignment algorithm is applied to all RGB data. Depth maps are aligned with the same transformation applied to the corresponding texture image. Because of the complexity of the database due to occlusions or challenging position, only 88% of the faces are detected. To perform the face recognition, the OpenFace [46] features are chosen which are based on the deep neural network described in [47] (implemented on Python and Torch). The algorithm used is explained more in details in Section 6.3.

Thermal face: Thermal facial recognition is based on the analysis of individual heat patterns emitted by the human face, on the form of an image presenting a map of apparent values of temperature on its surface. The thermal image contains sufficient amount of information for distinguishing individuals. Emission dominated, passive imaging does not require additional illuminator and is independent from illumination non-uniformities. The thermal face recognition process is composed of various stages comprising alignment, face detection, feature extraction and comparison. The face detection is performed using the Viola-Jones algorithm [48]. Several feature extraction methods are investigated and the best results are obtained with Local Binary Patterns [49] which perform well with thermal facial images, combined with various distance metrics.

Iris: The iris images collected present a good iris pattern quality for light pigmented irises but a lower quality for dark irises despite the additional lighting used which in turn caused specular reflections. The method for segmentation is the best ranked in the MICHE I competition [50]. For the feature extraction and comparison, it is used a novel approach designed for iris recognition on smartphones submitted to MICHE II. The FIRE method [51] is chosen for its good performance with mobile low-quality images. Among the three possible comparisons: left, right or left-left and right-right eye patterns, the best results are obtained for the right eye pattern comparisons.

Finger-vein: The finger-vein images are collected from both right and left index and middle fingers. The Region of Interest (ROI) extraction is done manually and then the

[8] http://visagetechnologies.com/

images are pre-processed in order to improve the visibility using High Frequency Emphasis Filtering, Circular Gabor Filter and simple Local Histogram Equalisation (CLAHE). For the performance evaluation, some well-established finger-vein recognition schemes are used. The Maximum Curvature (MC) [52] combined with the correlation-based comparison approach proposed by Miura et al. [52] achieved the best results. For more details see [53].

Hand-vein: Dorsal hand-vein images of both hands are acquired under different illumination conditions: two reflected light illuminators (850nm and 950nm) and one trans-illumination light source (850nm). The same processing tool-chain as for finger-vein is used to conduct the hand-vein performance evaluation. In addition, a rotation correction is adopted in the comparison step. The best results are obtained with the MC for the 950nm reflected light acquisition scheme.

Anthropometrics: The collected anthropometrics data include both physiological and behavioural features of an identification subject. Behavioural features, which include parameters such as average step length, are calculated from time-based signals extracted by a network of Kinect sensors. Physiological features include parameters such as height, arm/leg length. The method used for the recognition process applies an artificial neural network to estimate the similarity between two feature vectors. The network is based on siamese architecture [54]. A representative subset of the acquired data is used for the network training and validation purposes.

Multimodal Fusion

The multimodal evaluation is carried out using the MATLAB BOSARIS Toolkit[9] which is a collection of functions and classes that can be used to calibrate, fuse and plot scores for biometric recognition. For each biometric trait, two distance matrices, namely the DEV matrix and EVAL matrix are computed and used as follow.

DEV matrix: facilitates the tuning of the weights that will then be used on the EVAL matrix. It contains the scores originating from the comparison of 47 enrolled samples (one for each subject) against 47 development samples.

EVAL matrix: is made up of scores originating from the comparison of the 47 enrolled samples against 47 evaluation samples (different from the development ones). In a closed-set setup as all 47 subjects are both in the gallery (enrolled samples) and in the probe (testing samples) sets. However, the weights for fusion are computed on the DEV matrix and used on unseen test samples for final performance evaluation.

[9] https://sites.google.com/site/bosaristoolkit/

This protocol is adopted for all traits except for finger/hand-veins, where left hand is used to build the DEV matrix and the right hand for the EVAL matrix. Whenever more than one baseline is tested, the best performing configuration is chosen for fusion.

The BOSARIS toolkit capability to integrate the samples' quality scores for multimodal fusion is used. Thus, prior to fusion, all score matrices have been normalized using MinMax technique so that all scores range in $[0, 1]$ interval.

Results and discussion

The metrics used are defined upon False Non Match Rate (FNMR) and False Match Rate (FMR) as standardisation documents ISO/IEC 19795-1:2006 [55].

- *Equal Error Rate* (EER): error obtained when $FMR = FNMR$
- *FMR1000*: lowest FNMR for $FMR \leq 0.1\%$
- *ZeroFMR*: lowest FNMR for $FMR = 0\%$

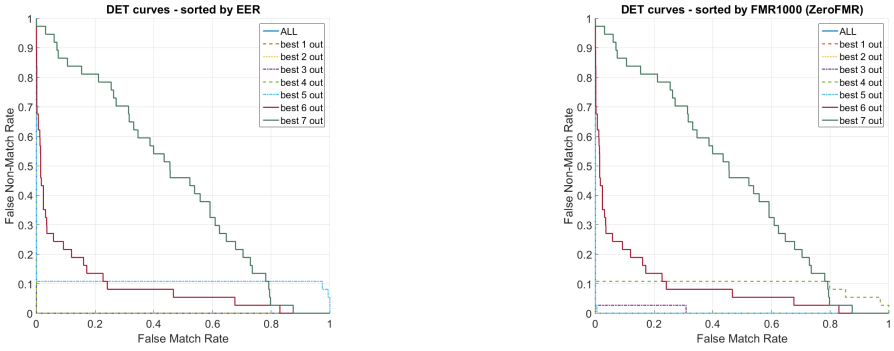
The results obtained for the unimodal recognition evaluation are depicted in Table 2.4. The first three columns show the results obtained by the benchmarking methods with all the available data. The last three columns show the results obtained by the BOSARIS method with the DEV and EVAL matrices that are the input for the fusion method. The best recognition results are obtained for 3D Face RGB and, on the opposite side, 3D Face DF, followed by thermal face and iris lead to the poorest results.

Table 2.4 – Unimodal recognition results (results in %). RGB and depth images from light field data are considered separately and fused together as if they were different biometric measures.

Biometric Trait	**Benchmark evaluation (all data)**			**DEV and EVAL data**		
	EER	FMR1000	ZeroFMR	EER	FMR1000	ZeroFMR
2D Face	9.12	28.10	41.09	2.69	1.28	10.81
3D Face only RGB	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
3D Face only Depth	39.37	100	100	44.27	82.50	97.30
Thermal Face	10.88	73.91	72.13	5.08	0.00	5.41
Iris VIS Mobile	15.32	45.96	65.25	16.17	4.86	70.27
Finger Veins	9.75	11.83	56.80	5.13	9.73	5.41
Hand Veins	0.12	0.25	0.25	9.76	4.77	10.81
Anthropometrics	0.88	4.44	18.66	0.47	0.00	24.32

The results for multimodal fusion are computed according to the leave-best-n-out scheme. After sorting the EER, ZeroFMR and FMR1000 values, the n-best performing biometric traits are excluded from fusion, for $n = 0, 1, ..., N - 1$ with $N = 8$. The DET curves

depicted in Figure 2.5 show how much the fusion results are impacted by the highest performing traits with a notorious decay in performance as the number of best performing traits excluded increases.



(a) (b)

Figure 2.5 – Multimodal recognition results (leave-best-n-out fusion scheme).

2.3.2 Horizon 2020 projects: examples of multimodal systems

In the last decades, travelling is become a usual activity, both for business and pleasure. The amount of people passing through the border controls is increasing, causing a wider need of faster and more accurate identity checking. For this reason, European Union allocates some money for improving the results on this field among others. The research presented in this Ph.D. thesis has been partially financed by two Horizon 2020 projects, ITEA IDEA4SWIFT and PROTECT.

Both projects have as main purpose the automatization of Border eGate for frequent travellers thanks to the analysis and fusion of several biometric traits.

IDEA4SWIFT

IDEA4SWIFT[10] (Identity management, secure Documents, interoperable Exchange and citizens Authentication FOR Systems for Worldwide Interconnection of Frequent Travellers) is a 36 months project financed by Horizon 2020. It started in September 2012 and, after an extension of six months, it successfully finished in March 2017.

The Consortium made by French and Turkish partners, both academic and industrial, worked with the aim of creating a complete, efficient and secure eGate dedicated to travellers.

[10] http://idea4swift.eurecom.fr/

The project can be seen as a three-block: 1. developing a traditional eGate, with higher security levels, controlled by 3rd generation passport thanks to biometric measures, 2. establishing a communication system between the eGate and a centralized storage database able to preserve privacy and 3. creating an enrolment station for frequent travellers.

EURECOM, together with Telecom SudParis, as academic partners, takes care about research in biometrics, in particular on RGB-D face recognition and soft biometric traits. Moreover, EURECOM focuses on several kind of attacks as plastic surgery and presentation attack detections.

PROTECT

PROTECT[11] (Pervasive and UseR Focused BiomeTrics BordEr ProjeCT) is a 36 months project financed by Horizon 2020. It started in September 2016 and it will end in August 2019. The Consortium includes academic and industrial partners from United Kingdom, Poland, Austria, France and Belgium.

PROTECT project has the goal of investigating emerging biometrics in order to expand the technologies considered nowadays by ICAO, to prove the security, accessibility and acceptability of the proposed biometric traits in a controlled environment by using real locations and to offer a reliable solution to law-enforcement agencies.

EURECOM is in charge of 3D face analysis, from image acquisition to fusion, passing through data processing and evaluation. The analysis applied for PROTECT project are done using the algorithms described in Chapter 6 and Chapter 7.

2.4 Soft biometric

Face is not only useful to recognise the identity of a person but it may be exploit also to investigate other features of a person. Gender, age, eye colour and even hair style can be automatically detected from face analysis. The study of characteristics useful to define the belonging of a subject to a specified category is called soft biometrics, in contrast to hard biometrics, finalised to recognise a specific individual.

The applications of soft biometric traits are sometimes complementary with hard biometrics: in a recognition process, where the system has to identify a target person in a huge pool of subjects, the knowledge of the gender may make faster the computation, since the algorithm evaluates the matching score in a smaller set of persons (for example,

[11] http://projectprotect.eu/

considering only males).
Soft biometry can be used also in other scenarios not related with identity recognition: an online video game provider may want to restrict under age teenagers to play, a clothes shop may want to adapt the advertisement according to the customer's gender.

Because of their categorizing properties, soft biometric traits are more impacted from ethical concerns. The ability for a machine to distinguish (and differently customize the system) males and females, or individuals with particular racial physical traits could be controversial.

2.5 Ethical issue on biometrics

The benefit due to biometrics use in the daily life is unquestionable. Screen unlock for smartphone or computers, easy access to working place and fast line for automatic passport controls are only few examples. In many countries, biometric data are used to access welfare services in order to limit fraud attends and to compensate for paper document lost. In emergency situation, the possibility to identify a person without his cooperation could be useful for medical and security reasons. Accessing a bank account with the acquisition of a fingerprint is easier than remembering a PIN.

Nevertheless, biometric systems may raise some issues. Generally people do not have problems to provide their name, place and date of birthday to be identified. It is a well known process and it is considered as legitimate. On the contrary, the acquisition of fingerprints or iris image is often seen as a suspicious activity, even a privacy invasion. From one side, the insufficient knowledge of biometric systems makes them not trustable, from the other side, the criminal use of biometric data can be actually problematic. If any biometric data is stolen it can not be substituted with a new version. While in case of stolen password, the user can just choose a new character combination and reset the system, that is not possible if the access is controlled by face imaging.

The strictly personal use of a biometric system is an efficient method to avoid impostors pretending to be the legitimate owner. But also, it prevents the possibility to borrow the device protected by fingerprints authentication. Moreover, some biometric traits are not so personal as are generally considered. Two brothers have part of DNA in common, mother and daughter may have similar faces, similar ears and, sometimes, even similar iris patterns. In some cases, providing personal biometric data can even damage the privacy of someone else.

Nowadays, privacy protection of biometric data is regulated by rigid rules. A wide research branch is working on systems able to recognize an individual having access to

only encrypted information, in order to prevent privacy violations.

Part I

A - Advanced sensors for face analysis

Introduction

Over the last century, standard cameras have been flanked by various technologies able to process complementary information. Thermal cameras allow to detect heat sources in the dark and they may be useful in surveillance context. Near infra-red sensors are often used for iris recognition because of their ability to collect the iris pattern details. High dynamic range cameras have been proposed to have high quality images in an environment with high light variations. That are only a few examples of advanced sensors used in biometrics and, in general, in image processing.

3D face analysis is quite popular and it is often used to improve recognition performances over the results obtained with standard sensors, to protect the systems from spoofing attacks and to remove eventual occlusions.

The first part of this manuscript is focus on the presentation of the most popular 3D advanced devices used for face analysis. In particular, Chapter 3 introduces the concept of light field imaging and presents the data that will be used for the analysis shown in part II. Two databases of light field face images are described and compared in order to answer to the research questions: *Which are the characteristics required in a database that includes light field information? Which protocols should be used for its realization?*

In Chapter 4, an overview on other 3D technologies is compiled and the performances on face recognition of light field and structured light devices are compared in order to answer to the research question: *In which context does light filed technology perform better than structured light one?*

Part of the presented work has been published in [6], [8, 9] and [13].

Chapter 3

Light field

3.1 Introduction

The purpose of this chapter is to describe the light field (LF) technology, which will be used for the analysis proposed in the following chapters. Some samples of images acquired with Lytro camera are shown to introduce LF data to the reader. An overview on LF databases (already presented in [6]) is compiled with a particular attention to *Light Field Face and Iris Database* [45] and *IST-EURECOM Light Field Face Database* [6].

IST-EURECOM Light Field Face Database represents one contribution of this thesis. It has been collected in the *Multimedia Signal Processing-Lx (MSP-Lx)* at *Instituto de Telecomunicações, Instituto Superior Técnico, Universitade de Lisboa*, in Portugal, and in the Imaging Security Lab at EURECOM, at the SophiaTech Campus, Nice, France. It has been presented to the public in the 5^{th} International Workshop on Biometrics and Forensics (IWBF) in 2017 [6].

The research questions that this chapter aims to answer are: *Which are the characteristics required in a database that includes light field information? Which protocols should be used for its realization?*

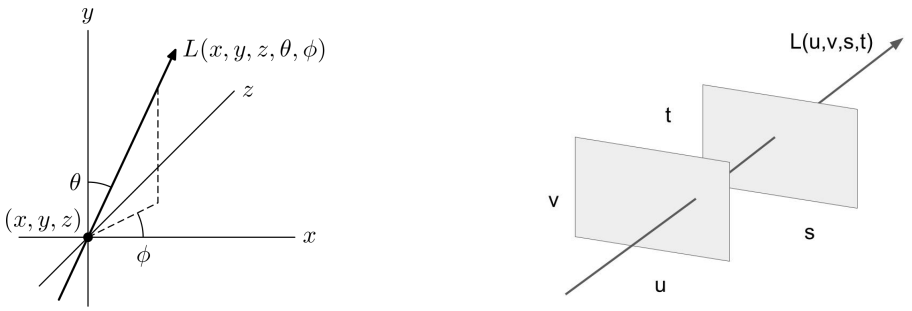
3.2 Light field technology

LF terminology is coined by Andrey Gershun in 1936 [56] to describe a concept already proposed by Michael Faraday in the middle of the previous century. LF is defined as the radiance of light rays emitted by points in a 3D scene along different orientations. The idea is that each point in the space can be represented by the 7-parameters *plenoptic*

function described in Equation (3.1).

$$L = f(x, y, z, \theta, \phi, \lambda, t) \tag{3.1}$$

Where (x, y, z) is the location of the point respect to a Cartesian system, (θ, ϕ) is the light direction, λ is the wavelength of the light and t the time. In this dissertation, only still images in visible spectrum are considered. For this reason Equation (3.1) can be reduced to a 5-parameters function (Figure 3.1a).



(a) (b)

Figure 3.1 – 5D plenoptic function (Figure 3.1a) and light slab representation proposed in [1] (Figure 3.1b)

In 1996, Levoy et al. [1] propose to adopt the representation shown in Figure 3.1b to describe the LF. Two planes are defined and placed at arbitrary position (u, v) and (s, t) and the light ray (called *light slab*) is defined by the function $L(u, v, s, t)$. The goal is to create a 4D model for LF from a set of 2D images and to create new 2D images from the 4D LF model. In fact, thanks to the suggested representation, each 2D image can be considered as a slice of 4D light. Authors visualize the 4D structure either as a (u, v) array of (s, t) images or as a (s, t) array of (u, v) samples (Figure 3.2).

The target applications include long range depth estimation, augmented or virtual reality with immersive content. According with uses and possibilities, LF cameras may have different structures. Camera rigs are employed to capture the set of views, offering a high spatial resolution for each view but a low angular resolution (i.e. large baseline) [57]. Single cameras mounted on moving gantries capturing the scene at regular time intervals are also tested [58]. While camera rigs can be quite bulky and not easy to use, moving gantries are limited to LFs of static scenes.

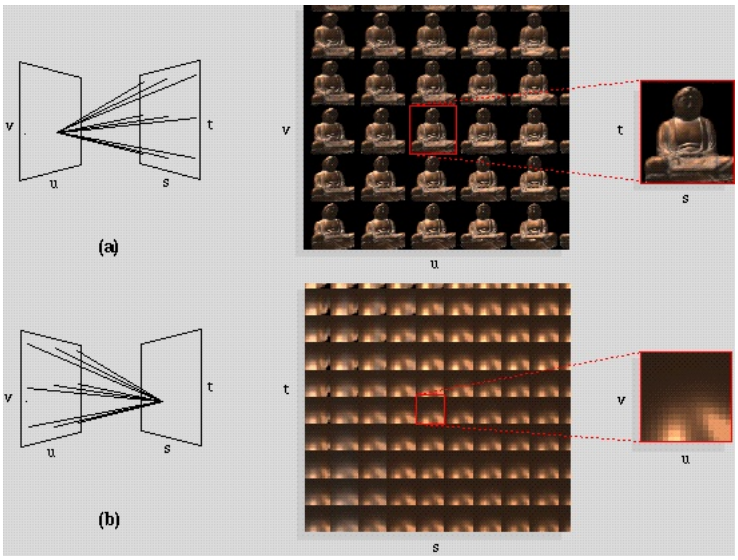
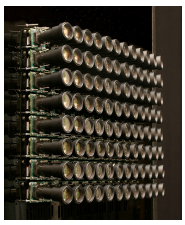


Figure 3.2 – Two possible visualization of 4D LF proposed in [1]



(a)

(b)

Figure 3.3 – The Stanford multi-camera array (Figure 3.3a) and the Adobe LF camera prototype (Figure 3.3b)

3.3 Lytro camera

In 2006, Ren Ng, under the supervision of Marc Levoy, defends his Ph.D. thesis entitled *Digital Light Field Photography* [59]. In his dissertation, Ng illustrates the working principles of LF photography and presents experimental validation with a LF prototype camera. He also presents a complete and detailed description of post processing operations.

Ng's starting point is the 4D light slab representation proposed by Levoy et al. in [1]. He considers the main lens and the sensor as reference planes (Figure 3.4), so that each point in the space is represented as a stripe in (u, x) coordinates. In a standard camera, all rays that contribute to a single pixel are represented by a vertical lines, destroying the directional information provided by u.

In order to preserve the directional information, the film plane is separated from the x plane. As shown in Figure 3.5, two points located at different distances create oblique

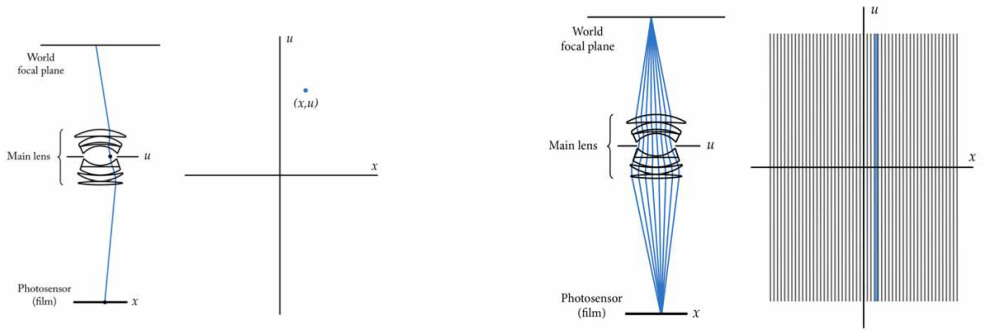


Figure 3.4 – 2D model of light ray in LF cameras. Even if the representation is bi-dimensional, it is straightforward to generalize to a 4D model. The images are extracted from [1].

projection. The angular coefficient of the line in (u, x) plane tells if the considered point is either closer or farther to the camera respect to the virtual focal plane.

In a LF camera, a microlens array carries out the function of x plane. In Figure 3.5, is shown the different representations of objects situated at different distances from the camera.

In [59] three different interpretations of LF data are described.

3.3.1 Raw data

LF raw data consists of the information acquired by the RGB sensor upon the microlens array. The image obtained is apparently similar to a standard image but, when zooming, it is possible to identify circular pattern, due to microlens (Figure 3.6). Each pixel of this image can be described with two values, (x, y), representing the spatial or pixel coordinates, and with other two parameters, (u, v) representing the angular or viewpoints coordinates.

3.3.2 Sub-aperture images

Given the information stored in the raw data, it is possible to obtain several sub-aperture images. One pixel is extracted from the sensor region covered by each microlens, in order to compose one sub-aperture image (Figure 3.7a). Each sub-aperture image is seen from a slightly different point of view (Figure 3.7b). This format is efficient and easy to implement and it has proved to be useful for face analysis, even if the angular resolution is achieved at the expense of a decreased spatial resolution when compared with classical 2D cameras. Despite of the small disparities between views (small baseline), they turn

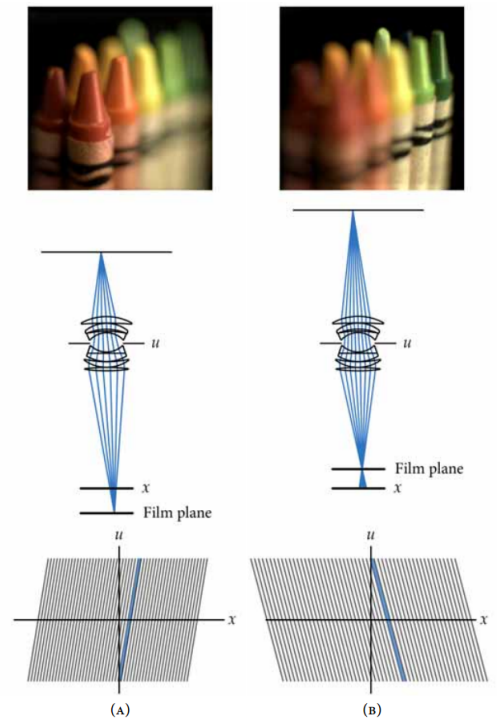


Figure 3.5 – Schematic representation of impact of distance object-sensor in LF camera. The images are extracted from [1].

out to be suitable also for 3D reconstruction.

3.3.3 Epipolar images

The idea behind epipolar image comes from the epipolar concept used in stereo vision, proposed in 1987 by Bolles et al. [60], to create a 3-dimensional description of a static scene from a set of images. An epipolar image represents a spatio-angular 2D slice of the 4D LF cut through a horizontal or vertical set of sub-apertures (e.g. the slice corresponding to the horizontal blue line in Figure 3.8a). The EPI representation shown in Figure 3.8 gives a LF sample at a constant y-value corresponding to the blue line. The y^{th} row of each aperture is extracted and shown one below the other in order to obtain Figure 3.8b. The slope of the line created in this way is a measure of the distance between the object and the camera.

3.3.4 Depth maps

The information stored in LF data enable 3D scene reconstruction (Figure 3.9). Depth map estimation techniques from LF data can be divided into two groups, depending on

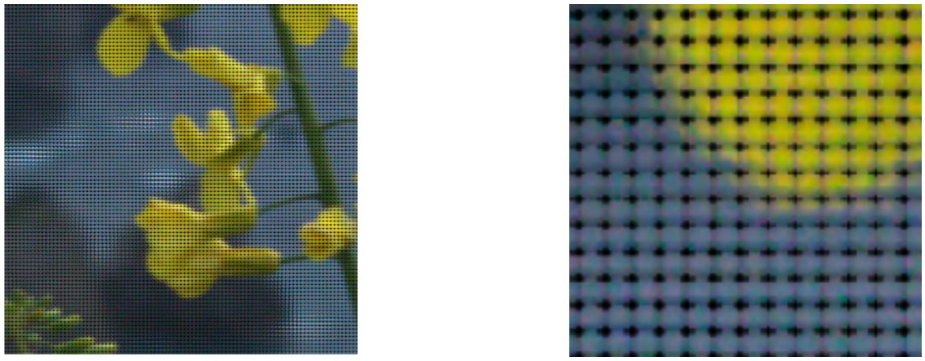
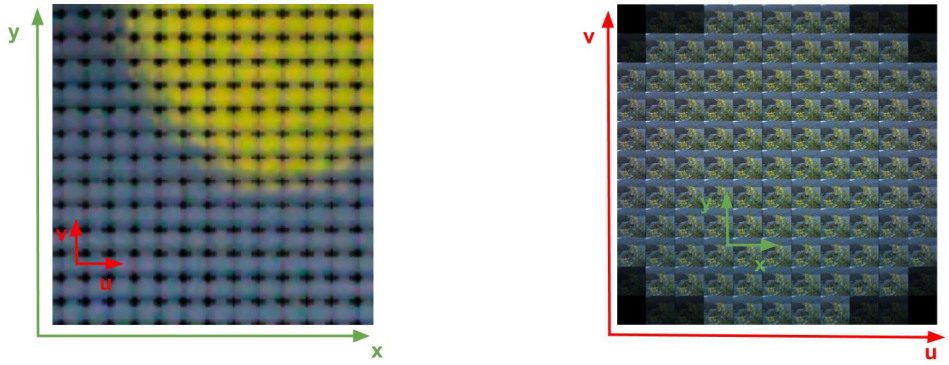


Figure 3.6 – Sample of raw Lytro image



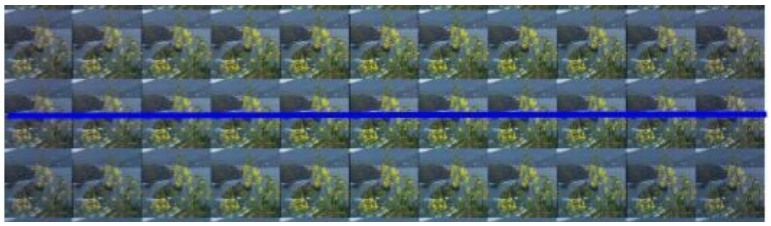
(a) (b)

Figure 3.7 – Sample of sub-aperture representation of a Lytro image.

disparities between views. Patch-based block matching is used in the first case [61, 62, 63], while depth is extrapolated from the linear structure of EPI images for dense LFs [64, 65]. In fact, the angular coefficient of each line in EPI image is proportional to the disparity value [60]. Note that, while stereo methods allow to estimate larger disparities, EPI based methods are only suitable for densely sampled LFs.

3.3.5 Re-focusing from light fields

The structure adopted by Lytro camera enables a densely sampling of LF and allows to render images with different focusing depths, called *focal stack* (Figure 3.10). LF re-focusing consists in defining a new LF $L'(x, y, u, v) = L(x - us, y - vs, u, v)$, where s is the re-focus parameter defined in such a way that the regions of disparity $d = s$ in the LF L have zero disparity in the LF L'. A refocused image I_s with parameter s is



(a)

(b)

Figure 3.8 – Sample of epipolar representation of a Lytro image.

Figure 3.9 – Sample of depth map representation of a Lytro image.

computed by integrating the light rays over the angular dimension as [66].

$$I_s(x,y) = \int_{\mathbb{R}} \int_{\mathbb{R}} L(x - us, y - vs, u, v)\psi(u,v)\mathrm{d}u\mathrm{d}v. \tag{3.2}$$

where $\psi(u,v)$ represents the aperture of the imaging system, equal to 1 in the case of a full aperture. Therefore, re-focusing is conceptually a mere summation of shifted versions of the sub-aperture images over the entire $\psi(u,v)$ aperture.

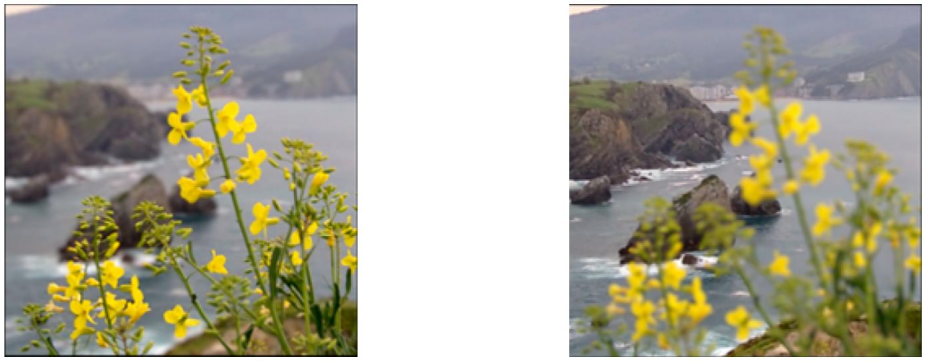


Figure 3.10 – Sample of different focusing depth of a Lytro image.

3.4 Light Field Face Databases

LF imaging is a relatively new topic. Thus, only few databases have been made available. At the date of writing, the following databases containing the full LF data are identified, none of them focused on biometric traits. The MMSPG dataset, created at EPFL, contains 118 LF images captured by a Lytro ILLUM camera [67]. The dataset is organized into ten different categories, covering a wide range of potential usages. The *IRISA Lytro First Generation Dataset* is a collection of 30 images including indoor and outdoor scenarios, some of them taken with motion blur or long exposure times [68]. In [69], a LF database for material recognition composed by 1200 images acquired using a Lytro ILLUM camera is presented. The *Light Field Saliency Dataset* (LFSD) consists of 100 images taken with a first generation Lytro camera, targeting saliency detection [70]. Some of the LF image databases are created with the purpose of studying the acquisition process or developing 2D image rendering algorithms; for this reason, the size, resolution, content and even the provided metadata are very different. The Computer Graphics Laboratory at Stanford University creates a 22 images LF archive using a multi-camera acquisition methodology [71]. The *Northwestern University database* is composed of 30 images of varied scenes, captured with a first generation Lytro camera, with the main goal to evaluate a dictionary learning based color demosaicing algorithm [30]. *The Heidelberg Collaboratory for Image Processing* (HCI) project database contains two categories of images: seven artificial LF rendered with a Blender, some of them with segmentation information, and six real-world LF acquired with a gantry [31]. Finally, a synthetic LF database with 18 images, including transparencies, occlusions and reflections, is presented in [32].

3.4.1 Light Field Face and Iris Database

Recent works explore the possibility of creating multiple focus images, rendered from the same LF image acquisition using super-resolution and fusion schemes for the multi-face recognition problem [45, 72, 73, 74]. The results demonstrate the benefits of LF imaging in terms of post-capture refocusing capability and accuracy when compared with conventional images. In order to assess the LF imaging based face recognition methods, the so-called *Light Field Face and Iris Database* (LiFFID) [45] has been proposed in 2016.

This database is composed by 112 subjects, 70 male and 32 female, ageing from 18 to 65 and belonging to different ethnicities. The collection task lasted for a period of one year.

Acquisition protocol

The acquisition protocol can be divided in two steps: 1. collection of enrolment data and 2. collection of probe data.

- **Enrolment data**

  Enrolment data consist in 8 high quality images for each individual, collected in indoor environment with good illumination and uniform background. Canon EOS 550D DSLR camera is chosen to satisfy the following ICAO standards:

  - Uniform illumination;
  - High resolution (at least 90 pixels between the eyes);
  - Limited pose angles;
  - Sufficient focus and depth of field.

  The eight samples represent the subject with neutral face, smiling and small variations of pitch, yaw and roll for the head. Each image is encoded in color JPEG format with a resolution of 5184x3456 pixels.

- **Probe data**

  Probe data are collected with a First Generation Lytro camera. Each image includes multiple subjects (from 2 to 4) at different distances from the camera (from 0.5m to 20m) in order to simulate a videosurveillance situation. Three different protocols are applied:

  - Protocol 1: Indoor scenario with controlled environment. The images are collected in a completely light controlled situation.

- Protocol 2: Indoor scenario with uncontrolled illumination. Individuals stand in a room close to windows, so that natural light can contribute in illumination together with artificial light. The result is a semi-controlled shadowing effect.
- Protocol 3: Outdoor scenario with uncontrolled background. The samples are acquired during a random moment of the day, in different location in order to have more challenging conditions.

The 1327 LF data are collected, for a total of 2986 face samples (1028 for Protocol 1, 447 for Protocol 2 and 1511 for Protocol 3). The multiple acquisition sessions held in a period of one year increase the variability of the data collected.

Lytro raw data are not directly provided to other research institutes. Images are subjected to a post processing operation: each data is rendered at different focusing depth with Lytro Desktop Software. The number of rendered levels is not regular and it may variate from 2 to 9. The represented faces are detected, cropped, converted in grayscale and resized to 120x120 pixels. Then, they are named according with the protocol and the subject identification number. Because of the limited access to raw data, during the work of this thesis, it has not been possible to compute deepest researches on LiFFID.

Figure 3.11 – Different focusing depth levels of the same raw data representing a face

3.4.2 IST-EURECOM Light Field Face Database

To support face recognition research exploiting LF images, the IST-EURECOM Light Field Face Database (LFFD) is designed, collected, elaborated and presented in [6].

The database has been made available in order to be an instrumental for designing, testing and validating LF imaging based recognition systems. The proposed face database includes data captured by a Lytro ILLUM camera in two sessions separated by 1-6 months. 20 samples per each person per session are collected with several facial variations, covering a range of emotions, actions, poses, illuminations, and occlusions. The database includes the raw LF data, 2D rendered images and associated depth maps, along with a rich set of metadata. The LFFD is expected to become a valuable addition to existing face database repositories. The LFFD includes data from 100 volunteers, including 66 males and 34 females, with a total number of 4000 LF face images in the database, corresponding to a

total disk space of about 270 GB. The participants were born between 1957 and 1998, and are from 19 different countries (Figure 3.12).

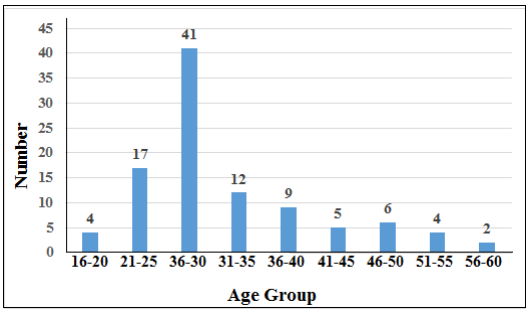


Figure 3.12 – Distribution of LFFD database participant divided per age.

Acquisition protocol

Image acquisition is performed in an indoor environment, using a lenslet-based LF camera: Lytro ILLUM. The Lytro ILLUM camera has a 40 Megaray sensor and a 30-250 mm lens with 8.3x optical zoom and f/2.0 aperture. The acquisition setup, illustrated in Figure 3.13, includes a white backdrop background behind a chair at a fixed distance of 1.25 m to the camera. The scene is illuminated with a three-point lighting kit, including a key light, a fill light and a back light, placed in order to limit shadows and allow ease segmentation of the subject from the background. The image acquisition process is repeated in the two labs with the same predefined setup (Figure 3.13).

Figure 3.13 – Acquisition environment for LFFD in IST and EURECOM labs.

The LFFD includes a total of 20 face variations per person, categorized into 6 dimensions:

1. Neutral image (one image): image captured with standard illumination, frontal pose, neutral emotion, no action, and no occlusion;
2. Emotions (3 images): images with three different emotions, notably happy, angry and surprise;
3. Actions (2 images): images with two different actions, notably closed eyes and open mouth;

4. Poses (6 images): images with different poses, notably looking up, looking down, right half-profile, right profile, left half-profile, left profile;
5. Illumination (2 images): images with different illumination intensities, notably low and high illumination levels;
6. Occlusions (6 images): images with occlusions, notably eye occluded by hand, mouth occluded by hand, with glasses, with sunglasses, with surgical mask and with hat.

Examples of the various face variations considered in the LFFD are illustrated in Figure 3.14. All images are taken under controlled conditions, but there are no restrictions on clothing, make-up and hair style.

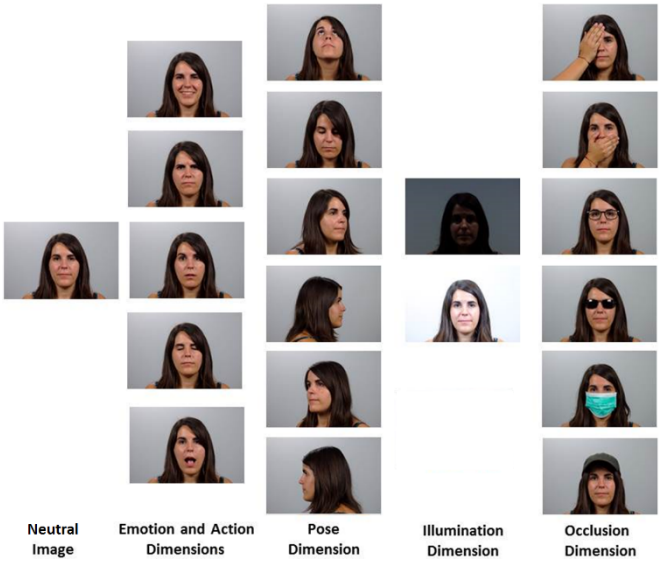


Figure 3.14 – Samples of one individual in IST-EURECOM Light Field Face Database.

LFFD database authors provide LF raw data[1] together with additional information.

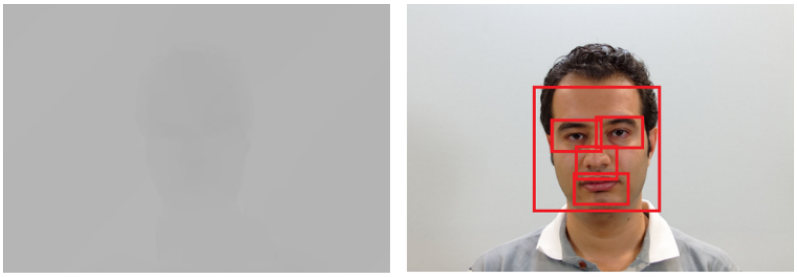


Figure 3.15 – Example of data provided in IST-EURECOM Light Field Face Database.

[1] http://lffd.eurecom.fr; http://www.img.lx.it.pt/LFFD

- **Light Field Images**

  LF images are the most important component of the database; they are stored in the Lytro Illum native format, using the so-called Light Field Raw (LFR) files. LFR files can be used as initial input for both the Lytro camera software i.e., Lytro Desktop Software, or to any other processing library/toolbox, such as the Matlab Light Field Toolbox V0.4 [75]. Face recognition systems working on LF images will typically require to perform some conversion of the LFR files in order to further process them. As an example, the Matlab Light Field Toolbox supports converting the LFR raw data into a sub-aperture array, with dimension $UxVxXxYx4$, where UxV represents the number of views, XxY represents the resolution of each sub-aperture image, and 4 corresponds to the number of components, notably R, G, B and a confidence level matrix associated to each pixel.

- **2D Rendered Images**

  Since LF images are not directly viewable in conventional 2D displays, the proposed database also includes 2D rendered images for the central view of each image variation, generated using the Lytro Desktop Software. It is worth noting that this software automatically performs a number of processing steps, including up-sampling and color correction, to enhance the quality of the output images. As the raw LF images are available, any other rendering solutions may also be used. The 2D rendered face images can be viewed using conventional 2D displays or be further processed.

- **Depth Maps**

  A depth map can be used to bridge the gap between 2D and 3D face recognition. Depth maps (see example in Figure 3.15) can provide geometric information about the position and shape of objects, to be explored by recognition systems. The supplied depth maps are generated using the Lytro Desktop Software.

- **Landmark Information**

  Facial landmarks are relevant for facial region extraction and normalization in face recognition systems. In the LFFD, the facial landmarks information includes the location of the face, left eye, right eye, nose and mouth bounding boxes, as illustrated in Figure 3.15. To obtain these data, the detection solution proposed in [76] is used to identify the selected facial landmarks. The landmark information is extracted for the central view 2D rendered images.

- **Metadata Information**

  Metadata information can be used for the evaluation of automatic face recognition, facial expression recognition, gender classification, and age estimation. The IST-

Table 3.1 – Metadata information in LFFD

Field	Range				
Date taken	Date				
Gender	Male	Female			
Age	Number				
Facial Hair	1-Shaved	2-Unshaved	3-Beard	4-Moustache	5-None
Make-up	1-Regular	2-Heavy	3-None		
Haircut	1-No hair	2-Short hair	3-Long hair	4-Gathered hair	
Earring	Yes	No			
Neckless	Yes	No			
Scarf	Yes	No			
Particular sign	Yes	No			

EURECOM LFFD rich metadata include the image acquisition date, as well as the subject gender, age, facial hair, makeup, haircut and usage of accessories; the range of values for each of these metadata fields is listed in Table 3.1.

- **Calibration Information**

  Calibration data is essential for compensating the specific properties of each camera's sensor. Furthermore, it is a required input for some LF image processing software products, such as the Lytro Desktop Software and the Matlab Light Field Toolbox [75].

Database File Structure

The files composing the database are organized according to a hierarchical structure. The root level of the hierarchy includes the metadata information and facial landmarks for all the subjects and the camera calibration files. The root level also includes a folder for each of the 100 subjects in the database, named using a 3 digit identifier, xxx. Each of these folders contains 3 sub-folders: "LFR files", "2D rendered images" and "Depth map images".

The naming convention for the database LF images is *type_xxx_s_yy_variation* where:

- "type" refers to the type of image, notably "LF" (light field), "2D" (2 dimensional) and "DM" (depth map);
- "xxx" is a three digit integer uniquely identifying the subject, starting from 001; the first 50 subjects have been recorded at IST and the second 50 subjects at EURECOM;
- "s" is a digit indicating the acquisition session number, notably "1" or "2";
- "yy" is a two digit integer indicating the variation number, ranging from 01 to 20, corresponding to different variations illustrated in Figure 3.

- "variation" is a two letter acronym identifying the variation in a format more suitable for human reading, e.g., HF (Happy Face) for the face image with happy emotion.

The proposed database can be used not only in the context of face recognition research but also for other research areas such as emotion recognition, gender classification, age estimation, ethnicity classification and face modelling.

3.4.3 LiFFID vs LFFD

The databases described are created for different purposes and they present different characteristics. While the main goal of LiFFID is to exploit the potentiality of LF cameras in unconstrained environment and targets videosurveilliance research, LFFD should provide a benchmark on face analysis on LF images. The number of subjects included in the databases is comparable, but LFFD contains a considerable bigger number of actions and occlusions. The bigger environment and distance variability present in LiFFD is only partially exploitable because of the impossibility to work on raw data.

The more updated technology used to collect LFFD allows the authors to discard the acquisition of enrolment data. In fact, the resolution of data from Lytro Illum camera and the protocol acquisition fulfil the ICAO standards mentioned in Section 3.4.1.

In Table 3.2, some of the main differences between LiFFID and LFFD databases are listed.

Table 3.2 – Main differences between LiFFID and LFFD Database

Field	**LiFFID**	**LFFD**
Num subjects	112	100
Sessions	3	2
Camera	Lytro first generation Canon EOS 550D	Lytro Illum
Multi face	Yes	No
Enrol images	Yes	No
Provided data	2D rendered images grayscale	LFR 2D colored rendered images Depth map Calibration data
Pose variation	6	6
Actions	0	2
Emotions	2	4
Occlusions	0	6
Illumination	3	2
Distance	0.5m - 20m	1m
Metadata	No	Yes

Chapter 4

Light field VS other 3D sensors

4.1 Introduction

In this chapter the main 3D sensors used for face analysis are presented. Some recent researches prove that structured light cameras are more suitable for face analysis than other 3D devices. For this reason, particular attention is paid on this technology. The images acquired with Lytro Illum and Kinect V1 images are compared in order to create a baseline for future studies on face recognition. The work has been presented in 2018 in the 17^{th} edition of Internation Conference on CyberWorlds in Singapore [8]. An extended version of the same study is currently under revision in Future Generation Computer Systems journal [9].

The research question that this chapter aims to answer is: *In which context does light filed technology perform better than structured light one?*

4.2 3D sensors for face analysis

In literature, several devices have been used to compute 3D analysis. Different technologies have been employed, according with material possibilities (sensor size, projector power) and target scenario. A small set of 3D scanners are based on contact principle: the object is physically in contact with a part of the scanner, such as a mechanic arm. Moving the arm and keeping it in a perpendicular direction respect to the surface, the scanner is able to establish the shape of the scanned object. This kind of techniques are not suitable for face data acquisition, thus they will not be analyzed in this thesis.

Contact-less 3D scanners are more convenient for face analysis. A first natural classification can be done according to the interaction with the environment. Active devices are based on the emission of some kind of waves carried by air (or water in some cases).

Usually, light wavelengths (specially infrared), ultrasound or x-ray are used. Active 3D scanners suffer from common problems such as light interference when the acquisition is performed in an uncontrolled environment or when multiple devices are used at the same time. They may also present artifacts and errors in presence of reflecting or transparent surfaces. Passive devices do not interfere with the environment and collect the light (visible or other wavelength) already present in the ambient. Passive cameras are usually cheaper and require less complex structures. Since there are not active interactions with the environment, several devices can be used simultaneously for recording indifferently indoor and outdoor.

4.2.1 Active - Structured light

Structured light (SL) cameras are one of the most known active devices. Recent studies from Politecnico of Torino have proved the superior performance of SL sensors on other technology in the context of face analysis [77]. For this reason, Section 4.3 is dedicated to give more detailed description of this kind of cameras.

4.2.2 Active - Time of Flight

Time of Flight (ToF) cameras (in Figure 4.1a) are based on the measure of the light speed. A laser (or a LED illumination system) projects one or more light pulses while a sensor evaluates the time required for the pulse to be gathered by the camera lens. The distance of a particular point in the space is proportional to the time of flight. ToF devices are become popular around 2000s, thanks to the speed improvement of semiconductor processes. A known example of ToF camera is Kinect V2[1]. These cameras are usually compact and have low computation power because ToF data can be processed by simple algorithms. No calibration or alignment is required before the data collection and the frame rate per second is quite high, allowing real-time application.

4.2.3 Passive - Stereoscopic

A stereoscopic system (in Figure 4.1b) usually requires two cameras placed in slightly different locations, recording the same object. The collected images are processed in order to find the disparities and evaluate the distance of each point according with a triangulation algorithm. Stereoscopic systems need to be aligned carefully.

[1]For more information on Kinect V2 https://blogs.technet.microsoft.com/microsoft_blog/2013/10/02/collaboration-expertise-produce-enhanced-sensing-in-xbox-one/

(a)

(b)

Figure 4.1 – Sample of ToF camera produced by MultiPix Imaging (https://multipix.com) and of stereoscopic camera from Nerian (https://nerian.com).

Table 4.1 – Comparison of 3D sensors considered. For more information the reader is addressed to [77]

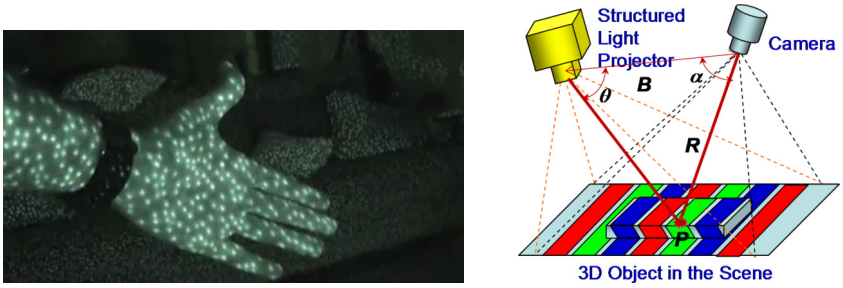
	Active		Passive	
	SL	ToF	Pass Stereo	LF
Frame rate	60 FPS	30 FPS	60 FPS	3 FPS
Max dist	5 m	60 m	15 m	0.985 m^2
Min dist	0.2 m	0.15 m m	0.23	0.285 m^3
Resolution	1280 x 1024	640 x 480	2208 x 1242	2450 x 1634
Size	80 x 20 x 20 mm	65 x 65 x 68 mm	57 x 30.5 x 14.7 mm	86 x 145 x 166 mm

4.3 Structured light vs light field camera in face recognition

The structured light (SL) technology is based on an active illumination system. While an imaging sensor acquires the RGB image, the third dimension of the scene is detected projecting a defined regular 2D-light pattern and collecting it with a customized sensor. The light frequencies can variate according to the considered device. The sensor used in this work projects a infrared dot pattern similar to the one shown in Figure 4.2a. The receiving sensor is located in a known position respect to the projector. In this way, it is possible to estimate the pattern that would be acquired if the surface represented in the image does not present any irregularity. The structured light pattern distortion in the acquired image respect to the expected one provides the information about the shape and the distance of the object. The triangulation principle can be described with Equation (4.1).

$$R = B\frac{sin(\theta)}{sin(\alpha + \theta)} \tag{4.1}$$

Where R, B, θ and α are defined as in Figure 4.2b. According with the projected light frequencies, the pattern shape, the acquiring camera resolution, SL device can be more or less accurate and suitable for different purposes.



(a) (b)

Figure 4.2 – Light pattern projected by SL Kinect V1 (Figure 4.2a) and an a schematic representation of structured light working principle (Figure 4.2b from [78])

4.4 Kinect V1 PrimeSense

Kinect V1 is an example of SL sensor. It is a low cost-effective multimodal sensing device, mostly known as gaming sensor. The camera is composed by two independent systems, one for RGB images and one for depth images (Figure 4.3). The first consists of a conventional RGB sensor located in the middle of the apparatus. The second deals with depth data acquisition and it is based on infrared wavelengths. A laser emits a known pattern of infrared dots and a CMOS sensor sensible to infrared frequencies detects the retrieved information. It has been launched in 2010 with the name of PrimeSense: the definition of the RGB sensor is 640x480 pixels, while the depth sensor has 320x240 pixels.

Figure 4.3 – Kinect V1 PrimeSense camera

RGB-D images acquired with Kinect V1 allow a perfect matching between RGB and depth pixels. While the acquisition process is partially confidential, the triangulation method used to align RGB and depth images is described by Freedman in [79]. Knowing the intrinsic parameters of the camera, it is possible to obtain the depth map from the

disparity image with Equation (4.2)

$$z_{world}^{-1} = \left(\frac{m}{f \times b}\right) \times d' + \left(Z_0^{-1} + \frac{n}{f \times b}\right) \tag{4.2}$$

where z_{world} is the distance between the object and the Kinect camera, d' is the normalized disparity value, n and m are the de-normalization parameters, b and f are the base length and the focal length respectively and Z_0 is the distance between the referenced pattern and the Kinect camera. All the parameters are provided by the manufacturer. The described processes are computed by the Kinect software and the final output consists in a RGB and a depth image with size of 640x480 pixels. Since the standard camera and the depth acquisition system are independent, RGB and depth images are not well aligned, making necessary further computations.

The procedure followed to align the database used in Section 4.7 consists of four steps. First the depth is converted in 3D coordinates $(x_{world}, y_{world}, z_{world})$ using Equation (4.3).

$$\begin{aligned} x_{world} &= -\frac{z_{world}}{f}(x - x_0 + \delta_x) \\ y_{world} &= -\frac{z_{world}}{f}(y - y_0 + \delta_y) \end{aligned} \tag{4.3}$$

where (x_0, y_0) is the centre of the reference system of the depth map and (δ_x, δ_y) represents the distortion of the lens, provided by the manufacturer.

The second step consists in transforming the 3D coordinates obtained from depth map in 3D coordinate system defined by RGB sensor, in order to align the two images (Equation (4.4)).

$$\begin{bmatrix} x'_{world} \\ y'_{world} \\ z'_{world} \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_{world} \\ y_{world} \\ z_{world} \\ 1 \end{bmatrix} \tag{4.4}$$

where $R \in \mathbb{R}^{3x3}$ is the rotation matrix and $T \in \mathbb{R}^{3x1}$ is the translation vector.

The 3D coordinate of the RGB system can be projected in a 2D space, consistent with

the RGB sensor plane. With appropriate transformation due to lens distortion and aberration is finally possible to align the depth image with the one produced by the RGB sensor.

$$\begin{bmatrix} x_{RGB} \\ y_{RGB} \\ 1 \end{bmatrix} = \frac{f_{RGB}}{z'_{world}} VD \begin{bmatrix} x'_{world} \\ y'_{world} \\ z'_{world} \end{bmatrix} \tag{4.5}$$

4.4.1 Kinect in face recognition

Kinect has been widely used in biometric between 2010 and 2016 because of its low cost and user-friendly operability. Several face databases [80, 81, 82, 83, 84, 85] have been created and provided in order to facilitate the study of this topic. The size of these databases variate from 936 images [80] to 845K data [85], collected either with Kinect V1 sensor [80, 82, 83] or with Kinect V2 camera [84, 85]. The database presented in [81] is collected with both technologies at the same time in an uncontrolled environment. All of them include multiple pose variations and different illuminations.

Even if some authors choose to work on super-resolution depth images computed combining depth map of multiple frames of short videos [84], most of the published methods are based on fusion (at feature or score level) between depth and RGB image. In [80], Min et al. propose a baseline evaluated with classical feature extraction algorithms for the analysis of EURECOM Kinect Face database. Goswami et al. [83, 86], develop a method based on Histogram of Gradient (HOG) feature computed on entropy and saliency maps from RGB and depth images, fused together to feed a Random Forest Classifier. The same team trains a neural network able to classify images according to the represented individual [87]. A different problem is tackled in [88], where authors want to define a metric including information from depth images and to use it to organize standard RGB images.

4.5 Databases

In order to perform a fair comparison between LF (Lytro camera) and SL technology (Kinect V2) two databases are selected. The data collection including LF images is the one presented in Section 3.4.2. In fact, LFFD database allows data manipulation

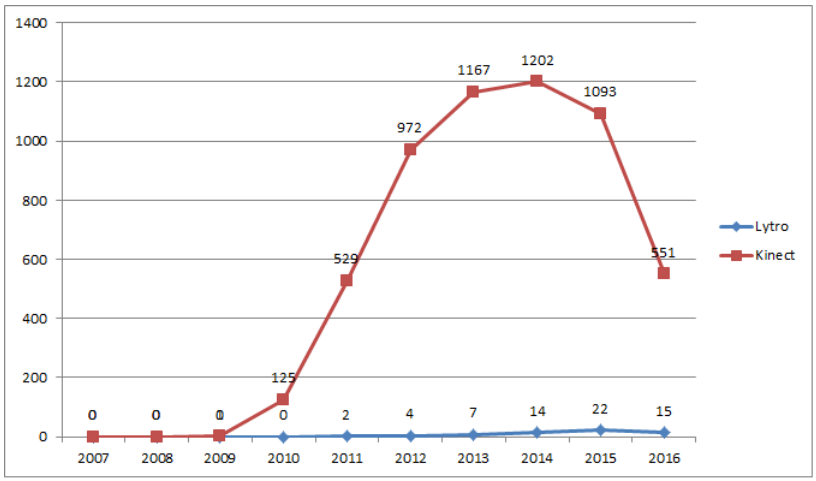


Figure 4.4 – Number of publication about Kinect and Lytro technologies. The literature related to LF cameras is really limited with respect to the amount of studies on Kinect cameras. In last years, works on infrared sensors have become less popular after having achieved a peak in 2014. On the contrary, studies on multiview systems have not yet taken off, leaving room for further research.

and provides RGB-D image pairs. Among all Kinect databases present in literature, *EURECOM Kinect Face Database* (KFD) [80] is selected to carry out on the analyses. The choice is leaded by the several common face variations present in LFFD and KFD and the similar acquisition protocol.

4.5.1 EURECOM Kinect Face Database

KFD is composed by 52 subjects, 38 male and 14 female born between 1974 and 1987, coming from different countries. As for LFFD, each subject participates in two sessions, with a distance of 5-14 days. In each session, for each person, several data are recorded: RGB images, depth images, 3D point cloud and a RGB-D video sequence. The subject is asked to sit in front of the camera and to follow a precise procedure, without imposing any clothing code or hairstyle.

The EURECOM Kinect Face Database is composed by:

- 2D rendered images with size 224x224 pixels;
- Depth maps with size 224x224 pixels, already aligned with RGB images;
- Depth maps of each pixel in the original coordinates not aligned with RGB images;
- 3D object files;
- Coordinates in 3D object space;
- Metadata information, such as gender and age.

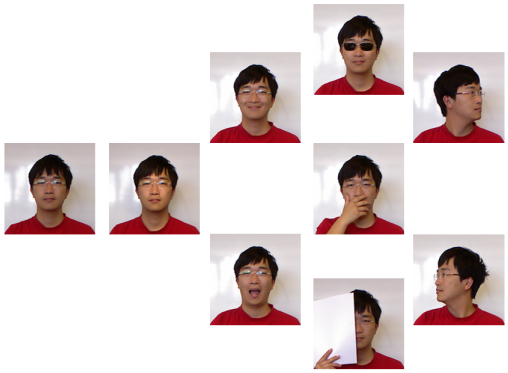


Figure 4.5 – Sample of data from *EURECOM Kinect Face Database*.

4.5.2 Additional Joint Mini Database

The two selected databases do not have any common individual. Moreover, they have been collected in different locations and at different times. An analysis of data collected at simultaneously with both devices could prove the consistency of the results over the different subjects represented. A joint mini-database (JMD) acquired simultaneously with Lytro Illum camera and Kinect V1 sensor is created. 20 individuals are asked to repeat the 6 common variations. Light, background and environment are kept as much similar as possible to the conditions described in [80] and [6]. However, the narrow size of JMD does not allow to base the analyses only on these data: the purpose of this mini-database is to prove that the conclusion attached to the experiments described in Section 4.7 are not influenced by minor variations in databases.

4.5.3 Selected images

While IST-EURECOM Light Field Face Database includes 20 face variations, EURECOM Kinect Face Database incorporates only 9 face variations. For the purpose of this work, only the 6 common variations are chosen. One or more representative samples are selected from each dimension: neutral image (or frontal face), smiling face as emotion, open mouth as action, high illumination, hand on mouth and sunglasses as occlusions. In fact, the analysis is carried out on each variation independently, in order to study the impact of the technology in different situations. Pose variations are skipped as they require more complicated alignment algorithm beyond the scope of this research.

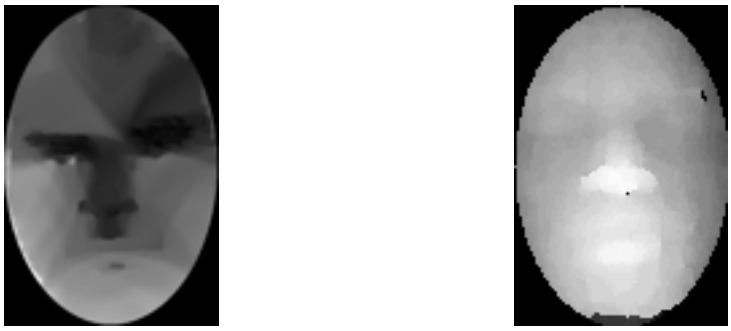
Table 4.2 – Technical comparison between Kinect V1 and Lytro Illum camera

	Kinect V1	Lytro Illum
Price	50 €	600[4] €
Outdoor use	No	Yes
Output	Video RGB-D images	Multi-focus RGB-D Sub-aperture
Alignment required	Yes	No
Resolution RGB/depth images	224x224	2022x1404

4.6 Similarities and differences between Kinect and Lytro data

The principles behind LF and SL sensors are quite different. While the first is based on a totally passive technology, the second interferes actively on the scene projecting a light pattern on the recorded object. Lytro Illum camera is created with a single sensor sensible to visible frequencies, while Kinect V1 needs two sensors to capture different spectra. Both devices require a calibration process that is performed only once in the device life. For Lytro Illum camera, the calibration consists in determining the precise location of each micro-lens, while for Kinect camera, the distance and the angle between projector and infrared sensor have to be known.
Because of different technologies, they also face different challenges. The process for obtaining RGB images does not present specific problems for any of devices. The acquisition of depth maps with Kinect V1 camera can result difficult if it used outdoor or if several devices are activated at the same time. Moreover, reflecting surfaces can create unexpected scattering and a possible movement of the camera could lead to a wrong distortion measures. On the contrary, the simultaneous acquisitions from multiple Lytro Illum cameras do not impact on the result and the use of LF technology is not limited indoor. Although if Lytro Illum camera does not have the possibility to collect videos, the only limitation would be software implementation and memory space. Instead, LF devices struggle to estimate the depth map where the area has uniform color, since it is more difficult to evaluate a disparity image.

In Figure 4.6 example of depth maps from Lytro Illum and Kinect V1 camera are shown. The former is visibly clearer even though more influenced by light.



(a) (b)

Figure 4.6 – Examples of depth image: 4.6a Lytro Illum, 4.6b PrimeSensor Kinect V1

4.6.1 Signal to noise ratio and precision

Because of the mentioned dissimilarities, Lytro Illum and Kinect V1 cameras create depth maps with different characteristics (Figure 4.6). For this reason, a preliminary study is carried out on depth images of the databases. Moreover, the impact of environment factors such as light, background, distance camera-individual, is not completely known. In order to perform a fair comparison, depth images from the JMD are confronted with the data of the other two dataset.

Two values are defined:

- *Signal to noise ratio (SNR)*: This value represents the depth resolution, i.e. the minimum distance in z-direction between two points to be considered different. The usual definition of SNR for a generic signal can be found in Equation (4.6)

$$SNR = \frac{\mu_{sig}}{\sigma_{bkg}} \tag{4.6}$$

  Where μ_{sig} is the average value of the signal and σ_{bkg} is the background standard deviation. Unfortunately, in this case the background of Kinect depth images is set to 0, leading to infinite value of SNR. Assuming a smooth shape of the face, noise information is extracted shifting the image and subtracting it from itself. This operation is done four times in the four main directions (left, right, up and down). The standard deviation of the evaluated error is used as denominator and called σ_{err}. Instead of average value as signal measure (μ_{sig}), the range of values stored in the depth map is employed. In fact, post processing operations (implemented in Lytro Illum software camera and not accessible to final users) assign a narrow range among all possible values (Figure 4.7). This characteristic reliably representative

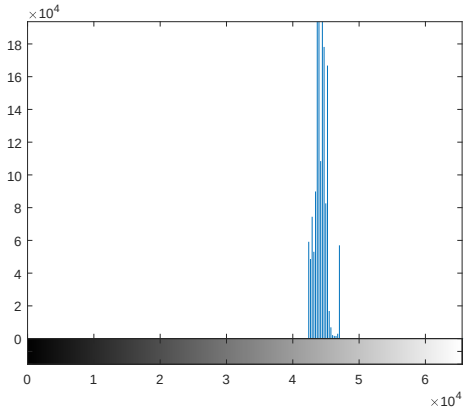


Figure 4.7 – Example of Lytro Illum depth map histogram. It is possible to observe how the distribution is concentrated in a small range of values.

of the signal and $range_{sig}$ is defined as range of depth map's values.

Thus, SNR for depth images is define as in Equation (4.7)

$$SNR_{depthmap} = \frac{range_{sig}}{\sigma_{err}} \tag{4.7}$$

- *Precision (PR):* This quantity represents the capability of reproducing the real depth value. Since no ground-truth is available, a PR measure is proposed. Two assumptions are done: face is perfectly aligned with the camera and it is symmetric with respect to a vertical plane passing through the centre of the image. Thus, dividing a depth picture along the symmetry axis, each pixel on left side should have a corresponding value on the right. SNR has to be considered in this computation: if the resolution is low, PR is not significant. Each pixel lying on left part of the image is compared with its symmetric one. If the difference between them is lower than the SNR, they are considered equal. PR is defined as the percentage of number of similar pixels (with respect to the metric defined by SNR) between the left part and the right part of the face (Figure 4.8).

$$PR = \frac{\sum_{y=0}^{N-1} \sum_{x=0}^{\frac{M-1}{2}} I_{(p_{(-x,y)} - p_{(x,y)})}}{N * M} \tag{4.8}$$

Where M, N are the horizontal and vertical image dimensions, I the identity

function defined in Equation (4.9), $p_{(x,y)}$ is the value of pixel in position (x, y).

$$I_{p_{(-x,y)}-p_{(x,y)}} = \begin{cases} 1 & if \quad p_{(-x,y)} - p_{(x,y)} \leq SNR \\ 0 & otherwise \end{cases} \tag{4.9}$$

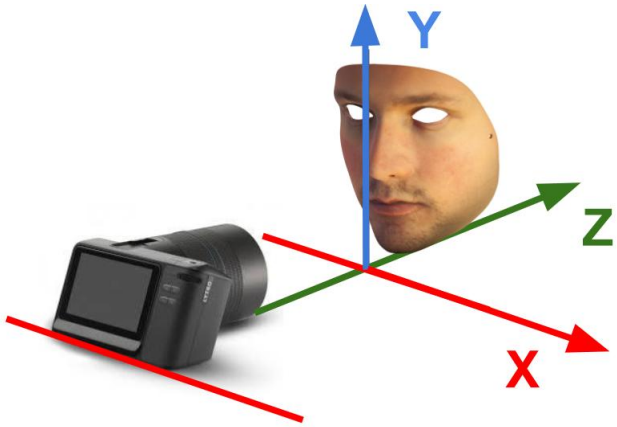


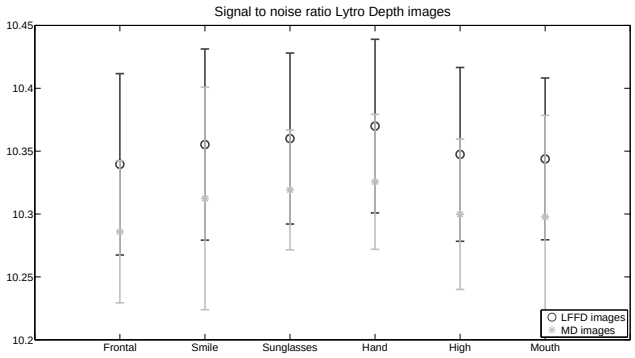
Figure 4.8 – Schematic representation of acquisition protocol.

From one hand, the analysis shows an higher SNR for KFD images (average value 47 vs 10). From the other hand, depth images from LFFD have a better precision (0.97 and 0.76 for LFFD and KFD respectively). The analysis paves the way to further studies devoted to improve the acquisition method, specially with Lytro Illum camera. In fact, the low signal to noise ratio is mostly caused by the narrow range of values in which face information are coded in the depth map, depending probably by subject-camera distance and physical configuration of the camera.

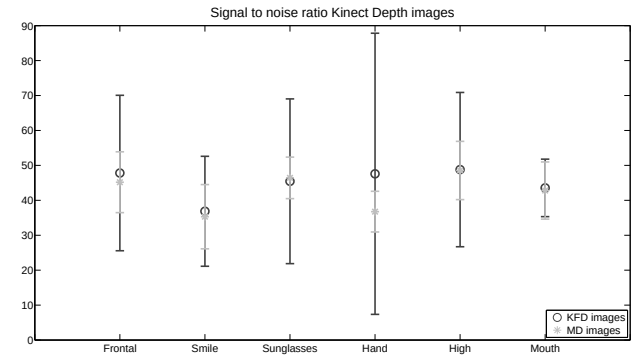
The mean value and standard deviation from LFFD and KFD are computed dividing the dataset according to face variations. A confident intervals are represented in Figure 4.9 and Figure 4.10 as variability measure. The mean value obtained from JMD images results to be in the defined interval for each face variations considered independently. This allows the validity of the analysis to be affirmed, regardless of the environment in which the images are acquired and the subjects represented.

4.7 Experimental setup and results

The preprocessing applied on images is similar to all used databases. The pair RGB and depth images is extracted from the raw data. Exploiting the perfect match between depth and RGB pixels position, both RGB and depth images are aligned and cropped. An oval mask is applied on the face region in order to avoid background interferences. In



(a)



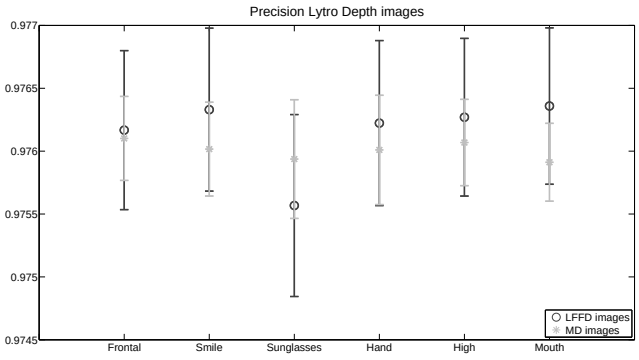
(b)

Figure 4.9 – Error bar representing the statistic of signal to noise ratio for Kinect images (Figure 4.9b) and Lytro images (Figure 4.9a). The mean value obtain from JMD images is always included in the confident interval estimated with LFFD and KFD databases.

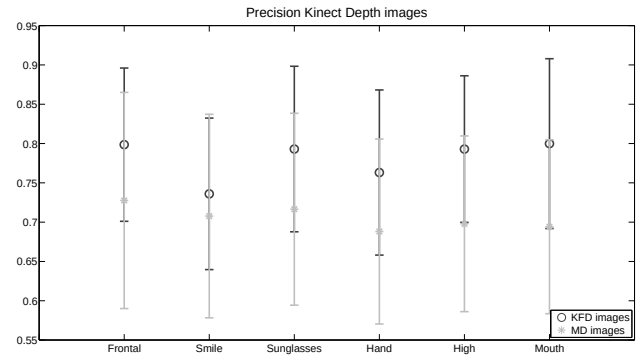
each image, a face alignment is performed using the Python Package DLIB [89]. The implemented algorithm consists in a face detector based on the Histogram of Gradient features (HOG) and able to identify 68 landmarks in significant keypoints on faces and to apply a rigid rotation.

Lytro and Kinect images have different resolutions. After a background removal operation, images from the Lytro camera have a resolution of 864x504 pixels, while the size of an image from the Kinect camera is 120x80 pixels. Although image resolution is an intrinsic property of the sensor, it could deeply influence face recognition algorithms. In [90] the relation between face recognition and image resolution is proved not linear and dependent by the used database. To limit the impact of the resolution disparity, LFFD images resampled with size 120x80 pixels are studied as well as LFFD images with original size.

The depth images from both KFD and LFFD result to be noisy because of some quantization operations or approximations occurred during the post-process. In order to



(a)



(b)

Figure 4.10 – Error bar representing the statistic of precision for Kinect images and Lytro images. The mean value obtained from JMD images is always included in the confident interval estimated with LFFD and KFD databases.

attenuate this problem, Gaussian filters with different sizes and variances are applied, creating two additional sets of depth images for each database.

- Filter 1 - size: 10x10, variance: 10
- Filter 2 - size: 30x30, variance: 60

4.7.1 Baseline techniques and configuration

In order to study the potential of LF and SL camera, three standard methods designed for RGB images (PCA [91], LBP [49] and LGBP [92]) and one customized for depth images (LBP3D [93]) are tested. The identification task is achieved by nearest neighbour (NN) algorithm and χ^2 (for LBP, LGBP and LBP3D) and Euclidean (for PCA) distances are used.

Two experiments are conducted on LFFD and KFD separately. In both the experiments, datasets are split in gallery set and probe set, taking care to have at least one sample for each subject in the gallery group. To each element in the probe set is assigned the identity of the gallery sample with the minimum distance. The performances are evaluated counting the percentage of good matches. In *experiment I*, only the neutral face images acquired during the first session are included into the gallery, while in *experiment II*, neutral expression data for both the sessions are used.

PCA, LBP, LGBP and LBP3D methods are tested on images from KFD and LFFD in order to create a baseline. The tests are generalized through analysis on JMD, in order to prove that results are not deeply dependent on the acquisition environment or on represented subject.

4.7.2 RGB images

RGB images from both databases are analysed. First, LBP, LGBP, PCA and LBP3D features are extracted for each image, then recognition algorithms are applied. Obtained results are shown in Table 4.3.

Experiment I: The performances on pictures acquired during the first session are higher than others for both databases because of small variations in term of light, pose and physical aspect between the gallery and probe images. In fact, the gallery set is composed by images illustrating "Neutral" expression recorded during the first session. The time gap between the two sessions in KFD and LFFD databases is at maximum 14 days and at least 1 month respectively. This difference gives rise to unbalanced results between the databases: it is particularly evident the case of "Sunglasses" occlusion, where the recognition rate on LFFD is 51% using the LBP-based method, while in KFD is higher than 88%. For both KFD and LFFD, the PCA-based method does not suit in cases of occlusions produced by sunglasses or hand on face; whereas, using the LGBP-based method, the recognition rate is always higher than 70%. Only the LBP-based method is influenced by image resizing: performances increase when the image size is reduced. The differences between adjacent pixels change radically when the images are resampled. This phenomenon is particularly evident for "Open mouth" images from the second session, where the gain is around 36%.

Experiment II: The results of second experiment are shown in Table 4.3. In this case the training set is composed by two images for each subject ("Neutral" images from both sessions). As expected, the tests done with the second procedure on images of the first session reveal better performances with respect to *experiment I* for both LFFD and KFD. The additional images in the gallery data increase the possibility of good

Table 4.3 – Percentage of rank-1 recognition rate for the first and second experiment on RGB images. The PCA-based method has really low performances when applied on images with occlusions like "Sunglasses" or "Hand". In second experiment, the results increase for each occlusion with respect to experiment 1, specially for the data acquired during the second session.

Exp 1		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Neutral (2)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
LBP	Kinect	92.30	**94.23**	86.53	96.15	88.46	**98.07**	**96.15**	**88.46**	**76.92**	**94.23**	**82.69**
	Lytro orig	**100**	81	86	97	91	88	82	51	56	79	50
	Lytro resize	**100**	92	**97**	**99**	**92**	95	82	68	68	89	68
LGBP	Kinect	98.07	**96.15**	90.38	98.07	75	**98.07**	**98.07**	**92.30**	80.76	96.15	**78.84**
	Lytro orig	**100**	91	**96**	98	**94**	97	90	76	80	96	72
	Lytro resize	**100**	91	**96**	98	**94**	97	90	74	80	96	73
PCA	Kinect	96.15	15.38	**55.76**	**96.15**	**76.92**	**88.46**	**88.46**	11.53	**48.07**	**78.84**	**50**
	Lytro orig	97	49	52	94	68	81	68	28	32	66	36
	Lytro resize	**98**	**51**	53	94	67	81	69	28	31	67	39
LBP3D	Kinect	94.23	**94.23**	90.38	96.15	88.46	**98.07**	**94.23**	**88.46**	**80.77**	**96.15**	**84.61**
	Lytro orig	**100**	80	88	97	90	89	82	53	56	79	51
	Lytro resize	**100**	91	**97**	**99**	**94**	93	80	67	72	91	62

Exp 2		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Neutral (2)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
LBP	Kinect	**100**	**100**	96.15	98.07	**100**	-	98.07	**98.07**	**98.07**	**100**	**96.15**
	Lytro orig	**100**	96	92	**99**	95	-	**100**	92	88	99	93
	Lytro resize	**100**	98	**98**	**99**	97	-	**100**	96	97	**100**	94
LGBP	Kinect	**100**	96.15	98.07	**100**	**96.15**	-	100	96.15	**98.07**	**100**	**98.07**
	Lytro orig	**100**	**98**	**99**	99	97	-	**100**	94	96	**100**	96
	Lytro resize	**100**	**98**	**99**	99	97	-	**100**	94	97	**100**	96
PCA	Kinect	96.15	21.15	55.76	**98.07**	**78.84**	-	98.07	23.07	**65.38**	92.30	**75**
	Lytro orig	97	**48**	56	93	64	-	**100**	49	62	98	60
	Lytro resize	**98**	**48**	**59**	91	66	-	**100**	**50**	61	**97**	63
LBP3D	Kinect	**100**	**100**	**100**	98.07	98.07	-	98.07	**98.07**	94.23	98.07	**98.07**
	Lytro orig	**100**	96	92	**99**	97	-	**100**	91	88	99	93
	Lytro resize	**100**	99	97	**99**	98	-	**100**	97	**98**	**100**	96

matching. The time gap between acquisition sessions no longer influences the analysis. The considerations done in the previous paragraph about the PCA-based method are still valid: eigenface features are not tailored for recognition in case of occlusions. Image resizing brings no performance improvement while using the procedure described for the second experiment. For both experiments, when the LBP3D-based method is applied, the results on RGB images are similar to the one obtained with the LBP-based method. The results of face recognition on RGB images from LFFD and KFD shows a good recognition rate for all methods except for the PCA-based method when applied on face variation with occlusions. KFD images seem to suffer more this phenomenon than LFFD images, specially for Sunglasses occlusion.

In order to prove the consistency of the results, the experiments described are run on JMD database. Since the size of this database is too small to generate significant rank-1 recognition rate values, distances between samples are analysed. Mean and standard deviation of distances between samples representing the same subject, respectively from LFFD and KFD images, are computed with the aim of creating a confidence interval. Figure 4.11 shows how the same measure evaluated on JMD images falls in the previously defined interval for each face variation for the LBP-based method. The same results are obtained for the LGBP and the LBP3D-based methods. The procedure could not be followed for the PCA-based methods because the space where the distance computed depends on the database.

4.7.3 Depth images

The same experiments performed on RGB images are repeated on depth maps.

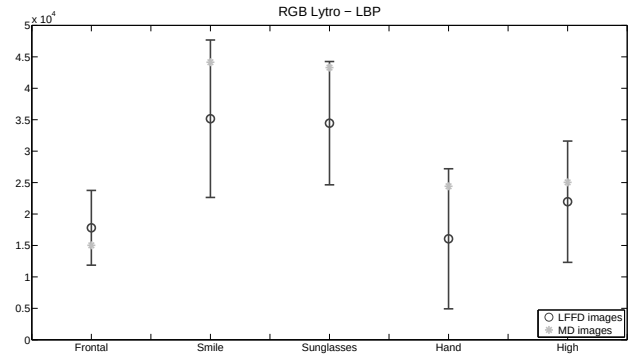
Compared with RGB images, performances on depth images are lower because the proposed algorithms are more suitable for textured images. Nevertheless, the recognition rate is sill superior than random.

Experiment I: In Table 4.4, the results of *experiment I* on depth images are shown. Depth images from LF technology are largely influenced by illumination: this leads to significant low recognition rate for "High illumination" variation acquired with Lytro camera. In fact, only 54% of images are well recognized. Instead, for SL camera, 92% of subjects is associated with the correct training data. In addition to the time gap between the first and the second session, the light variation influences the recognition rate on images obtained during the second session. Even though databases authors tried to reproduce the same conditions, light differences are present between sessions. All methods have lower performances if applied on face variations involving those occlusions that change radically the depth of the image. *Experiment I* conducted on resized images from LFFD

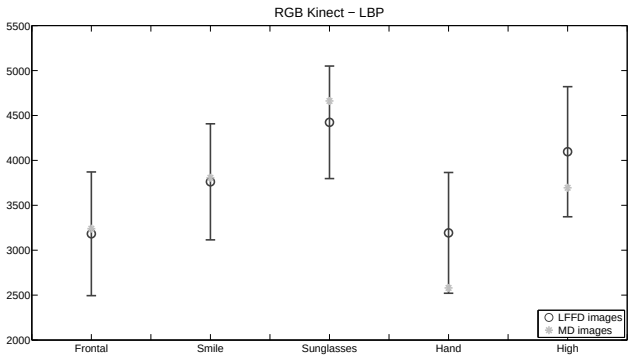
Table 4.4 – Percentage of rank-1 recognition rate for the first and second experiment on depth images. It is possible to observe a huge disparity on results on "High Illumination" images from KFD and LFFD databases.

Exp 1		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Neutral (2)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
LBP	Kinect	**82.69**	**75**	**40.38**	**92.30**	53.84	**84.61**	**55.76**	**57.69**	**25**	**82.69**	**38.46**
	Lytro orig	80	48	30	46	62	25	22	14	10	22	16
	Lytro resize	80	42	37	44	**63**	31	24	14	11	20	16
LGBP	Kinect	82.69	**57.69**	38.46	**78.84**	55.76	**65.38**	**55.76**	**40.38**	**26.92**	**75**	**34.61**
	Lytro orig	**86**	45	**45**	54	**66**	29	26	15	13	28	19
	Lytro resize	85	45	**45**	54	**66**	29	25	16	13	28	19
PCA	Kinect	63.46	**34.61**	9.61	**63.46**	15.38	**48.07**	**28.84**	**19.23**	**9.61**	**40.38**	**13.46**
	Lytro orig	**73**	25	**18**	43	**45**	13	17	8	5	14	7
	Lytro resize	**73**	25	**18**	42	**45**	14	18	7	5	15	7
LBP3D	Kinect	**80.77**	**78.84**	32.70	**88.46**	53.84	**80.77**	**57.69**	**57.69**	**19.23**	**78.84**	**36.53**
	Lytro orig	79	45	**36**	41	**64**	27	22	12	11	20	16
	Lytro resize	79	45	**36**	41	**64**	27	22	12	11	20	16

Exp 2		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Neutral (2)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
LBP	Kinect	**94.23**	**84.61**	**59.61**	**96.15**	**76.92**	-	**90.38**	**84.61**	**53.84**	**92.30**	**69.23**
	Lytro orig	82	44	29	39	58	-	86	31	26	33	58
	Lytro resize	79	44	30	50	58	-	87	35	30	43	59
LGBP	Kinect	80,76	**63,46**	**36,53**	**88,46**	59,61	-	84,61	**61,53**	**46,15**	**84,61**	53,84
	Lytro orig	**84**	47	32	54	59	-	**92**	38	36	52	**61**
	Lytro resize	79	44	30	50	58	-	87	35	30	43	59
PCA	Kinect	61.53	**36.53**	7.69	**67.30**	19.23	-	55.76	**32.69**	3.84	**51.92**	17.30
	Lytro orig	**73**	25	**16**	42	**44**	-	**79**	17	**11**	37	31
	Lytro resize	72	25	**16**	41	**44**	-	**79**	17	10	37	**32**
LBP3D	Kinect	**88.46**	**88.46**	**51.92**	**92.30**	**67.30**	-	84.61	**75**	**38.46**	**92.30**	**69.23**
	Lytro orig	83	43	28	37	59	-	**86**	32	27	34	58
	Lytro resize	80	46	29	44	58	-	**86**	37	33	43	59



(a)

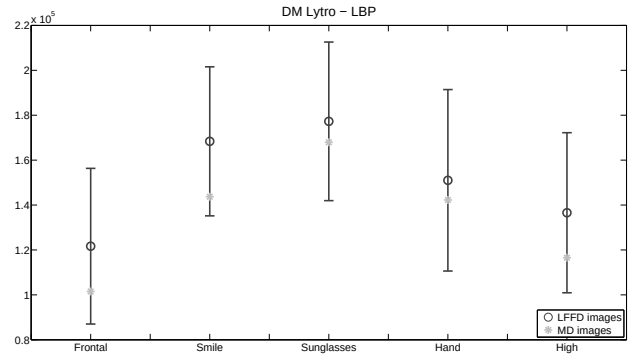


(b)

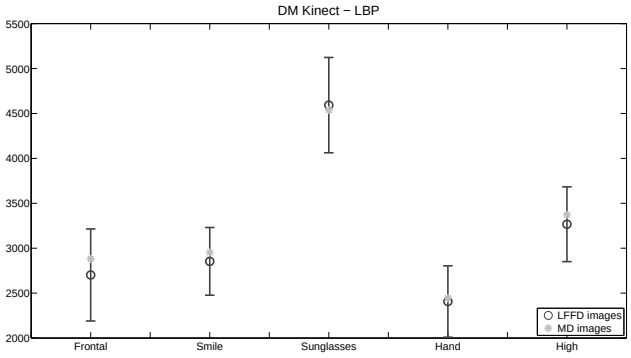
Figure 4.11 – Error bar representing the distribution of distances between LBP features obtained by images with the same subject. The graphics show how the statistics from JMD and LFFD (Figure 4.11a) and JMD and KFD (Figure 4.11b) are comparable for RGB images.

leads to a slight improvement with respect to LBP, LGBP and LBP3D-based methods. Despite of more challenging images, the results of analysis conducted on KFD data using LBP, LGBP and LBP3D-based methods are higher than 80% on "Smile" and "High illumination" variations of the first session. The light dissimilarities impact only slightly on the second session, where the recognition rate of "High illumination" is higher than 75%.

Experiment II: Results of *experiment II* on depth images are shown in Table 4.4. As for RGB images, the percentage of well recognized images increases with respect to *experiment I*, specially on the images acquired during the second session. The improvement is evident on "Smile" expression from LFFD, where the recognition rate increases up to 92% for LGBP-based method. LBP3D-based method does not outperform LBP-based method for depth images. Previous studies [93, 94] proved that better results could be obtained analyzing different parts of the face separately. However, the main purpose here is to



(c)



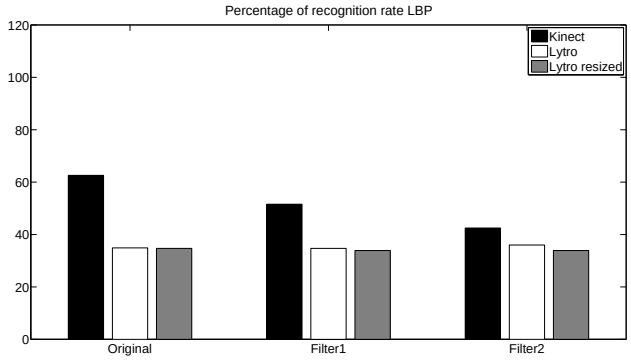
(d)

Figure 4.11 – Error bar representing the distribution of distances between LBP features obtained by images with the same subject. The graphics show how the statistics from JMD and LFFD (Figure 4.11c) and JMD and KFD (Figure 4.11d) are comparable for both RGB and depth images.

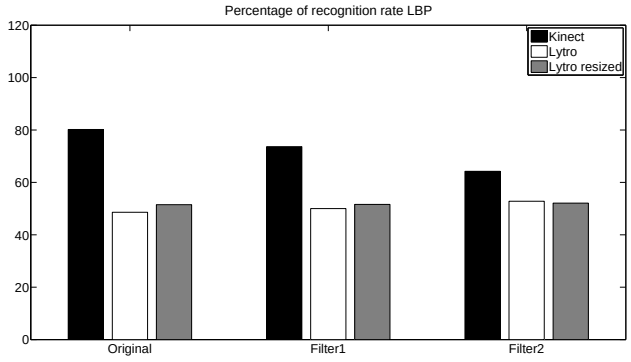
compare the impact of LBP3D features on depth images from Lytro and Kinect. In both cases, the rank-1 recognition rates from LBP and LBP3D features are close.

In Figure 4.12 and Figure 4.14, the percentage of recognition rate over all depth images is shown. Each method is applied on the original depth images and on their smoothed versions. In Figure 4.12 and Figure 4.14 results for LBP and LBP3D-based method of *experiment I* (Figure 4.12a and Figure 4.15a) and *experiment II* (Figure 4.12b and Figure 4.15b) are shown. Both experiments prove that the impact of the filters on KFD is negative, while on LFFD is slightly positive. As mentioned before, a small improvement could be obtained resizing LFFD images. In fact, both during the resizing and the smoothing process, image noise decreases. Results for LGBP-based method are presented in Figure 4.12 related to *experiment I* (Figure 4.13a) and *experiment II* (Figure 4.13b). Also in this case, KFD images are more influenced (negatively) by smoothing filters than LFFD depth images.

Recognition rate on depth images using PCA-based method is low: percentage of well recognized subjects over all database is always lower than 35% for *experiment I* and than 40% for *experiment II*. However, in this case, smoothing depth images from both databases leads to improvements, as shown in Figure 4.14.



(a)

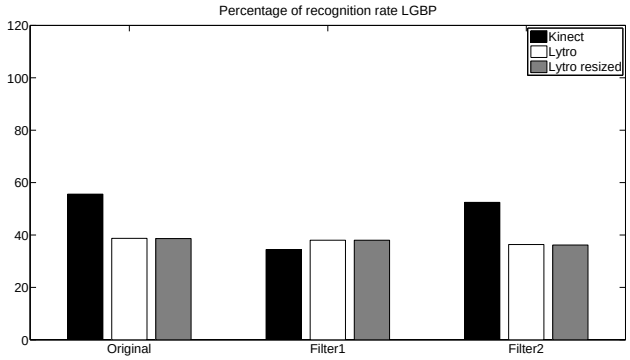


(b)

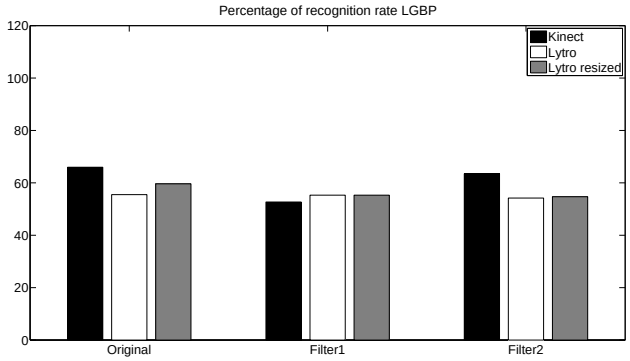
Figure 4.12 – Percentage of recognition over all database for depth map filtered with Gaussian filters (Filter 1: 10x10 pixels, Filter 2: 30x30 pixels) for LBP-based method for *experiment I* (Figure 4.12a) and *experiment II* (Figure 4.12b). The decreasing of performances in case of filters for KFD images is particular evident. In Figure 4.12b, the improvement due to LFFD images resizing is notable.

Analysis on depth images are generalized with the same procedure described for RGB images. Also in this case, the results show (Figure 4.11) how the environment and the different represented subjects do not influence the conclusions.

The different performances obtained for LFFD and KFD images lead to conclude that depth images from LF cameras are less informative for face recognition purposes than the one obtained with SL sensors. Smoothing filters could be applied successfully on LF images. In fact, depth images from Lytro are more noisy, as proved by analysing the SNR.



(a)



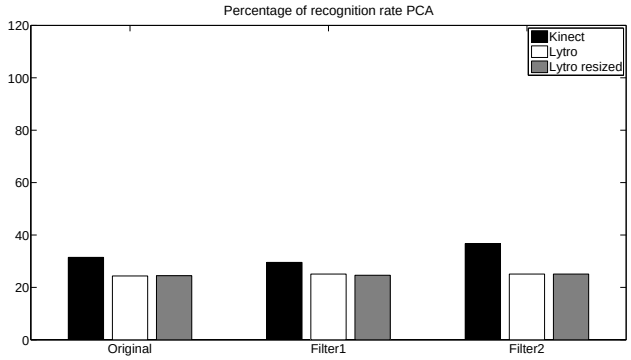
(b)

Figure 4.13 – Percentage of recognition over all database for depth map filtered with Gaussian filters (Filter 1: 10x10 pixels, Filter 2: 30x30 pixels) for LGBP-based method for *experiment I* (Figure 4.13a) and *experiment II* (Figure 4.13b). Decreasing of performances in case of filters for KFD images is particular evident. In Figure 4.13b the improvement due to LFFD images resizing is remarkable.

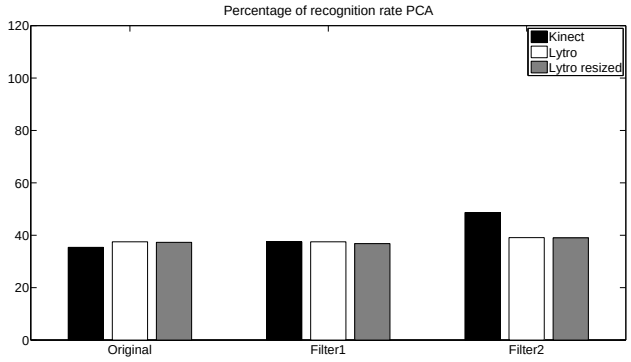
4.7.4 Fusion

The strength of both Kinect and Lytro camera is the double output, RGB and depth image. The analysis of RGB and depth images independently do not provide a general view of the potentiality of these technologies. For this reason, fusion between RGB and depth images at score level is studied. Scores from RGB and depth data are weighted and summed in order to jointly contribute to the recognition process.

The strategy adopted is composed of three steps. First, the dissimilarity values (the scores) for both RGB and depth images are normalized.



(a)

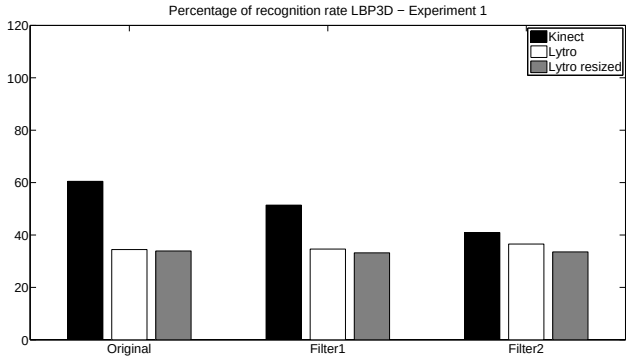


(b)

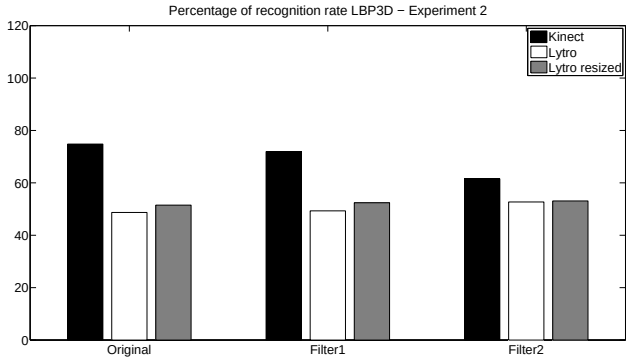
Figure 4.14 – Percentage of recognition over all database for depth map filtered with Gaussian filters (Filter 1: 10x10 pixels, Filter 2: 30x30 pixels) for the PCA-based method for *experiment I* (Figure 4.14a) and *experiment II* (Figure 4.14b). Improvements generate by filtering KFD and LFFD images are remarkable.

$$\begin{aligned} \hat{d}_{RGB} &= \frac{d_{RGB} - \mu_{RGB}}{\sigma_{RGB}} \\ \hat{d}_{DM} &= \frac{d_{DM} - \mu_{DM}}{\sigma_{DM}} \\ d_{fusion} &= w\hat{d}_{RGB} + (1-w)\hat{d}_{DM} \end{aligned} \tag{4.10}$$

For each experiment, mean values and standard deviations of dissimilarity values for RGB and depth images are computed in a different way. In *experiment I*, the μ_{RGB}, μ_{DM}, σ_{RGB} and σ_{DM} are evaluated comparing gallery data and "Neutral" expression recorded during the second session, while for *experiment II* the whole database is considered. In the second step, the weights are fixed with a search grid, choosing the value that



(a)



(b)

Figure 4.15 – Percentage of recognition over all database for depth map filtered with Gaussian filters (Filter 1: 10x10 pixels, Filter 2: 30x30 pixels) for LBP3D-based method for *experiment I* (Figure 4.15a) and *experiment II* (Figure 4.15b).

maximizes the recognition rate. Finally Nearest Neighbor algorithm is evaluated on scores obtained in Equation (4.10).

Figure 4.16 and Figure 4.18, compare the percentage of recognition rate for RGB images, depth images and fusion at score level. Results for LBP-based method are shown in Figure 4.16 for *experiment I* (Figure 4.16a) and *experiment II* (Figure 4.16b). For both databases, fusion improves recognition. On KFD a gain lower than 0.2% is achieved to be compared to a larger 1.6% for LFFD. The LFFD experience shows how depth information could be useful in the face recognition process. No better results are obtained by application of fusion on dissimilarities generated with LGBP features. The improvement obtained on LFFD images is lower than 0.5% and even negative when the method is applied on KFD images. This phenomenon could be explained by the small room for improvements left by the already high percentage of recognition on RGB images. The negative trend is due to the fact that the weights are evaluated in order to maximize the recognition over gallery data and "Neutral" expression of the second

session for *experiment I* and the all database for *experiment II*.

As well as for RGB images, results obtained with LBP3D features are similar to the values achieved with LBP features.

The analysis of the PCA-based results shown in Figure 4.18 is more interesting. The improvement due to fusion in *experiment I* (Figure 4.18a) is 7.8% and 3.2% respectively for KFD and LFFD images. Also for *experiment II* (Figure 4.18b) the gain is remarkable, 2% for both KFD and LFFD.

In Table 4.5 fusion results are described for each face variation for experiment 1. It is possible to observe how fusion improves recognition rate when 3D occlusions (such as "Sunglasses" or "Hand on face" variations) are present. For all databases and sessions, combining RGB and depth images increases the percentage of well recognized subjects on "Sunglasses" and "Hand on face" variations when the method used is based on LBP features as well as on LBP3D features. The case of "High illumination" is of particular interest: results achieved with SL camera are better than the ones obtained with the LF technology, because of depth map sensibility to the light. Results for the LGBP-based method for *experiment I* are presented in Table 4.5. As previously mentioned, recognition rate does not change significantly. Results obtained fusing RGB and depth images using PCA-based method are particularly interesting. Recognition rate on RGB images regarding "Sunglasses" variation is low, specially on KFD images (15%). Fusion with depth images increases of 200% the number of well recognized subjects for the first session and of 160% for the second. The improvements are also remarkable when the method is tested on LFFD.

In Table 4.6, fusion results of *experiment II* are presented. Also in this case, the analysis shows how fusion between RGB and depth images improves face recognition for both KFD and LFFD when of 3D occlusions occur. However, when the face variation considered does not present depth difference from gallery data (such as "Smile" variation), the impact of fusion is less important. Moreover, the high recognition rate for RGB images gives a small room for improvements.

4.8 Conclusion

In this chapter, an overview on 3D sensors used for face analysis is presented. In particular, the potential on face recognition of SL and LF cameras, is investigated. The difference between depth map created with SL and LF devices are studied introducing Signal to Noise Ratio and Precision. In order to study the differences and the similarities between Kinect and Lytro sensors, four baseline algorithms are tested on IST-EURECOM Light

Table 4.5 – Percentage of rank-1 recognition rate for the first experiment on RGB, depth images and fusion of both. For LBP, LGBP and LBP3D-based method, face variations with a recognition rate already high for RGB images are almost no influenced by fusion, or, as in case of "Smile" for the second session are negatively impacted. For LGBP-based method, particularly evident is the gain when the fusion is applied on "Open Mouth" face variation. Results of fusion achieved with PCA-based method indicate how depth images could be used as additional information in case of tridimentional occlusions like "Sunglasses".

LBP		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Neutral (2)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
Kinect Orig	RGB	92.30	94.23	86.53	96.15	88.46	98.07	96.15	88.46	76.92	94.23	82.69
	Depth	82.69	75	40.38	92.30	53.84	84.61	55.76	57.69	25	82.69	38.46
	Fusion	92,30	**96,15**	86,53	**98,07**	82,69	98,07	92,30	88,46	**80,76**	**96,15**	**84,61**
Lytro Orig	RGB	100	81	86	97	91	88	82	51	56	79	50
	Depth	80	48	30	46	62	25	22	14	10	22	16
	Fusion	100	**84**	**87**	97	**95**	88	82	**53**	**58**	**80**	**51**
Lytro resize	RGB	100	92	97	99	92	95	82	68	68	89	68
	Depth	80	42	37	44	63	31	24	14	11	20	16
	Fusion	100	**94**	97	99	**96**	95	81	**70**	**69**	89	67

LGBP		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Neutral (2)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
Kinect Orig	RGB	98.07	96.15	90.38	98.07	75	98.07	98.07	92.30	80.76	96.15	78.84
	Depth	82.69	57.69	38.46	78.84	55.76	65.38	55.76	40.38	26.92	75	34.61
	Fusion	92.30	96.15	86.53	98.07	**80.76**	98.07	92.30	82.69	80.76	96.15	**80.76**
Lytro Orig	RGB	100	91	96	98	94	97	90	76	80	96	72
	Depth	86	45	45	54	66	29	26	15	13	28	19
	Fusion	100	**92**	96	98	**97**	97	89	74	80	**97**	**73**
Lytro resize	RGB	100	91	96	98	94	97	90	74	80	96	73
	Depth	85	45	45	54	66	29	25	16	13	28	19
	Fusion	100	**92**	95	98	**97**	97	89	74	80	**97**	73

PCA		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Neutral (2)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
Kinect Orig	RGB	96.15	15.38	55.76	96.15	76.92	88.46	88.46	11.53	48.07	78.84	50
	Depth	63.46	34.61	9.61	63.46	15.38	48.07	28.84	19.23	9.61	40.38	13.46
	Fusion	**98,07**	**44,23**	53,84	96,15	76,92	**92,30**	**90,38**	**30,76**	38,46	**82,69**	**57,69**
Lytro Orig	RGB	97	49	52	94	68	81	68	28	32	66	36
	Depth	73	25	18	43	45	13	17	8	5	14	7
	Fusion	**98**	**54**	**57**	94	69	80	68	**30**	**34**	66	**41**
Lytro resize	RGB	98	51	53	94	67	81	69	28	31	67	39
	Depth	85	45	45	54	66	29	25	16	13	28	19
	Fusion	98	**55**	**57**	94	**69**	81	70	**31**	**35**	67	**40**

LBP3D		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Neutral (2)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
Kinect Orig	RGB	94.23	94.23	90.38	96.15	88.46	98.07	94.23	88.46	80.77	96.15	84.61
	Depth	80.77	78.84	32.70	88.46	53.84	80.77	57.69	57.69	19.23	78.84	36.53
	Fusion	94.23	**96.15**	90.38	**98.07**	88.46	98.07	92.30	**90.38**	**82.69**	96.15	82.69
Lytro Orig	RGB	100	80	88	97	90	89	82	53	56	79	51
	Depth	80	46	31	47	63	25	23	15	9	23	17
	Fusion	100	**84**	87	97	**95**	88	82	53	**58**	80	51
Lytro resize	RGB	100	91	97	99	94	93	80	67	72	91	62
	Depth	79	45	36	41	64	27	22	12	11	20	16
	Fusion	100	**94**	97	99	**96**	**95**	81	**70**	69	91	**67**

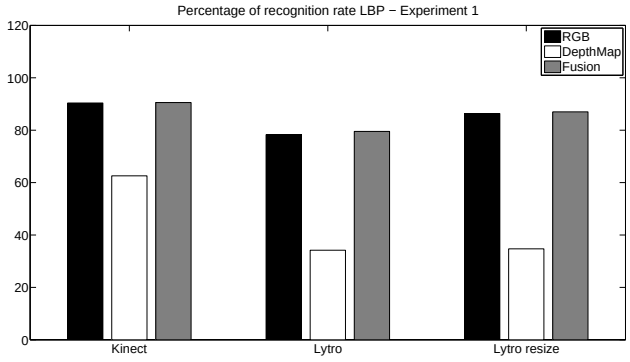
Table 4.6 – Percentage of rank-1 recognition rate for the second experiment on RGB, depth images and fusion of both. For LBP and LGBP-based method, the percentage of subjects well recognized using fusion between RGB and depth images on "Hand" face variation from KFD decrease respect to results obtain from RGB images. Fusion results very effective on PCA-based method on "Sunglasses" and "Open Mouth" face variations, while for "Hand on mouth" variation related to KFD images the impact is negative. Fusion does not result very effective for LBP3D-based method because the small room gives to improvements respect to RGB images analysis.

LBP		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
Kinect Orig	RGB	100	100	96.15	98.07	100	98.07	98.07	98.07	100	96.15
	Depth	94.23	84.61	59.61	96.15	76.92	90.38	84.61	53.84	92.30	69.23
	Fusion	100	100	96,15	98,07	100	**100**	98,07	94,23	100	**100**
Lytro Orig	RGB	100	96	92	99	95	100	92	88	99	93
	Depth	82	44	29	39	58	86	31	26	33	58
	Fusion	100	96	**93**	99	**96**	100	92	**93**	**100**	93
Lytro resize	RGB	100	98	98	99	97	100	96	97	100	94
	Depth	79	44	30	50	58	87	35	30	43	59
	Fusion	100	**100**	98	99	**99**	100	95	97	100	94

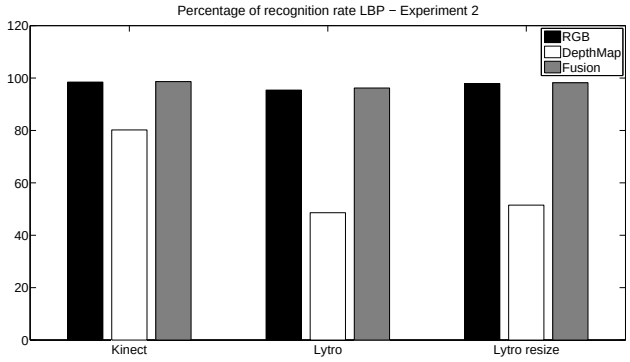
LGBP		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
Kinect Orig	RGB	100	96.15	98.07	100	96.15	100	96.15	98.07	100	98.07
	Depth	80,76	63,46	36,53	88,46	59,61	84,61	61,53	46,15	84,61	53,84
	Fusion	100	**98.07**	92.30	98.07	**100**	100	**98.07**	90.38	100	**100**
Lytro Orig	RGB	100	98	99	99	97	100	94	96	100	96
	Depth	84	47	32	54	59	92	38	36	52	61
	Fusion	100	98	99	99	**98**	100	94	**99**	100	96
Lytro resize	RGB	100	98	99	99	97	100	94	97	100	96
	Depth	84	46	31	55	60	92	38	36	54	60
	Fusion	100	98	99	99	**98**	100	94	**99**	100	96

PCA		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
Kinect Orig	RGB	96.15	21.15	55.76	98.07	78.84	98.07	23.07	65.38	92.30	75
	Depth	61.53	36.53	7.69	67.30	19.23	55.76	32.69	3.84	51.92	17.30
	Fusion	**100**	**48.07**	38.46	98.07	**80.76**	98.07	**40.38**	50	92.30	**76.92**
Lytro Orig	RGB	97	48	56	93	64	100	49	62	98	60
	Depth	73	25	16	42	44	79	17	11	37	31
	Fusion	**98**	**54**	55	93	**68**	100	**51**	62	**99**	**64**
Lytro resize	RGB	98	48	59	91	66	100	50	61	97	63
	Depth	72	25	16	41	44	79	17	10	37	32
	Fusion	98	**53**	55	**92**	**67**	100	**54**	**64**	**99**	**65**

LBP3D		Smile (1)	Sunglasses (1)	Hand (1)	High (1)	Mouth (1)	Smile (2)	Sunglasses (2)	Hand (2)	High (2)	Mouth (2)
Kinect Orig	RGB	100	100	100	98.07	98.07	98.07	98.07	94.23	98.07	98.07
	Depth	88.46	88.46	51.92	92.30	67.30	84.61	75	38.46	92.30	69.23
	Fusion	100	100	98.07	**100**	98.07	**100**	98.07	90.38	98.07	**100**
Lytro Orig	RGB	100	96	92	99	97	100	91	88	99	93
	Depth	83	43	28	37	59	86	32	27	34	58
	Fusion	100	96	93	99	96	100	91	**93**	100	93
Lytro resize	RGB	100	99	97	99	98	100	97	98	100	96
	Depth	80	46	29	44	58	86	37	33	43	59
	Fusion	100	100	97	99	99	100	97	98	100	96



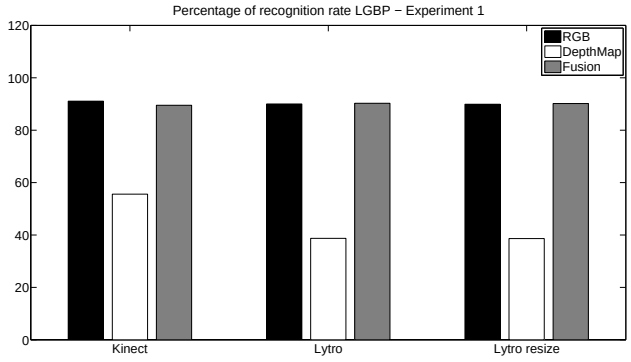
(a)



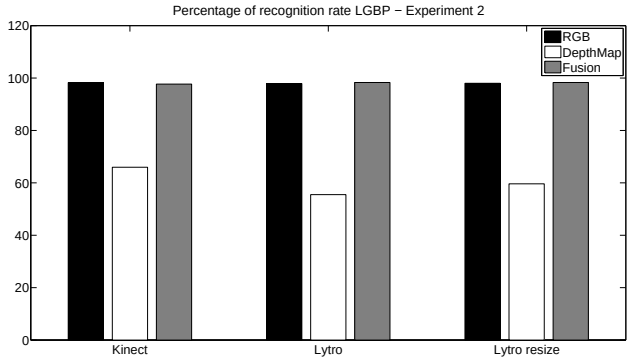
(b)

Figure 4.16 – Comparison between RGB, depth Map and Fusion recognition rate for LBP-based method. From the barplots it is possible to observe the improvements due to fusion with respect to analysis done on RGB images only.

Field Face Database and on EURECOM Kinect Face Database taking advantages of their similar protocol acquisition and data structure. Two experiments are set up, one using as gallery set one picture per subject and one keeping two images per person, the former from the first session, the latter from the second. The analysis shows how the time gap between the sessions could be really influential in face recognition and how rescale interpolation influences the performance of LBP and LBP3D-based methods on RGB images. Tested algorithms are proven more suitable for depth images acquired with SL sensors that LF cameras, although the latter have better resolution. Particular relevant is the fact that depth images from LF cameras are strongly affected by illumination. Gaussian filters are applied on images in order to smooth the noise. Only a small variation on face recognition results is observed on LF images. A mini-database of 20 persons is acquired with both Lytro Illum camera and Kinect sensor at the same time in order to check the consistency of our conclusions. Finally RGB and depth images are fused at score level verifying the improvement in recognition ratio, specially when the image



(a)



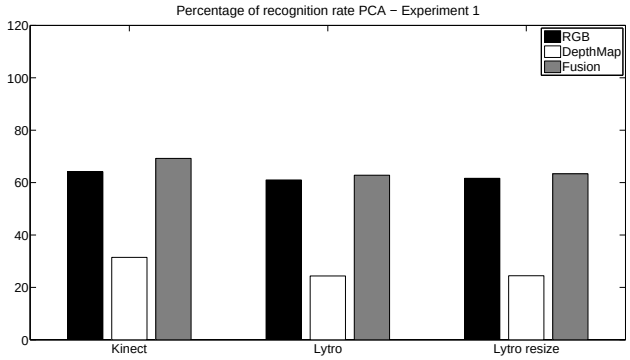
(b)

Figure 4.17 – Comparison between RGB, depth Map and Fusion recognition rate for LGBP-based method. From the barplots it is possible to observe the improvements due to fusion with respect to analysis done on RGB images only.

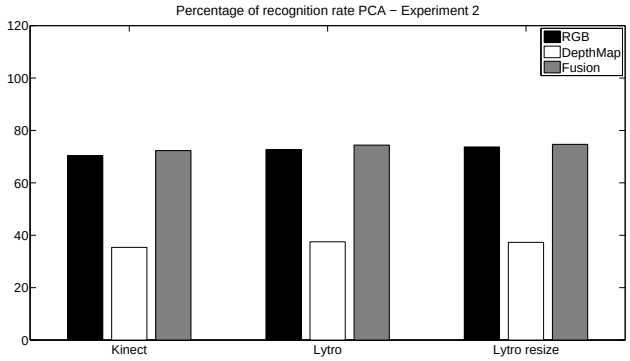
presents tridimentional occlusion like in the case of "Sunglasses" variation.

SNR on LF depth images is strongly influenced by the acquisition process. The information stored in depth map from IST-EURECOM Light Field Face Database do not take all range of possible eligible values. For this reason, understanding how to improve the quality of Lytro image of face data acquisition would be an essential topic still to be discussed. Thanks to the robustness of tested methods, image resolution appears to be not relevant in face recognition problem.

Even if in this work hand-crafted features are used, this result may be useful in further analysis based on learned features. A natural development of this work could be verifying the consistency of the results when using deep-learning algorithms.

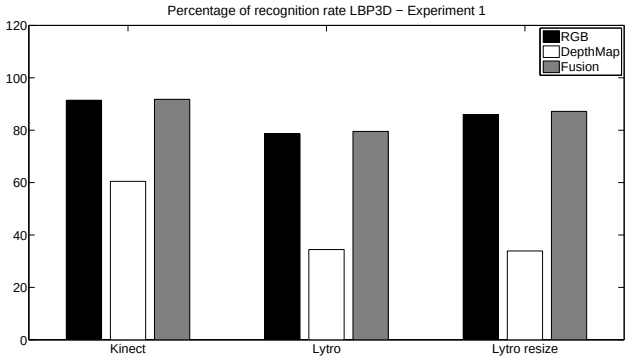


(a)

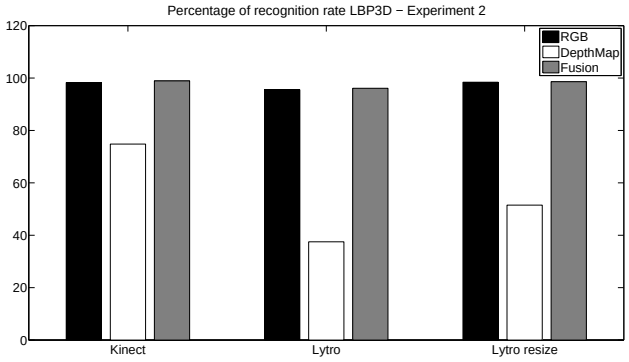


(b)

Figure 4.18 – Comparison between RGB, depth Map and Fusion recognition rate for PCA-based method. From the barplots it is possible to observe the improvements due to fusion with respect to analysis done on RGB images only.



(a)



(b)

Figure 4.19 – Comparison between RGB, depth Map and Fusion recognition rate for LBP3D-based method. From the barplots it is possible to observe the improvements due to fusion with respect to analysis done on RGB images only.

Part II

B - Impact of light field data on face analysis

Introduction

Face analysis is a huge field divided in different branches: starting from an image, the represented faces have to be detected and localized in order to select a specific region. Than, customized post-processing operations are applied to align, enhance and polish the image. Recognition or verification algorithms are used to reach the desired results. In addition to identification, face images can be exploit also to prove the liveness of the data or to extract soft biometric characteristics such as gender or age.

When a new technology is investigated, its impact on all the steps of the pipeline has to be analysed. The purpose of the second part of this Ph.D. dissertation is to present a study of light field cameras influence on soft biometric traits and an analysis of benefits on face recognition and anti-spoofing.

Chapter 2.4 wants to answer to the research question: *Which is the impact of defocusing in soft biometrics analysis?* proposing models of gender recognition and age estimation variations in relation with different focus depths.

In Chapter 6, an identification algorithm customized for plenoptic data is shown and tested, answering to the research question: *How can light field information enhance face recognition rate?*

The last chapter is focus on proposing an innovative method to protect a recognition system from presentation attacks exploiting light field data. The research question that Chapter 7 wants to answer is: *Can light field technology lead toward more robust anti-spoofing strategies?*

Part of the presented work has been published in [10], [11], [12] and [13].

Chapter 5

Gender recognition and age estimation

5.1 Introduction

In video surveillance the estimation of semantic traits as gender and age has always been debated topic because of the uncontrolled environment: while light or pose variations have been largely studied, defocused images are still rarely investigated.

In this chapter, two main topics are tackled: 1. the impact of focusing depths on gender recognition and age estimation from face and 2. the difference between Gaussian blurring and light field (LF) post processing defocusing in evaluating gender and age probability. A preliminary analysis of correlation between focusing depth and both gender recognition and age estimation is carried on. Additional experiments are designed in order to compare the defocusing obtained with LF post-processing and the blurring due to Gaussian filters. The work has been presented in 2016 in San Diego, USA, during SPIE Optical Engineering and Application conference [10].

The research question that this chapter wants to answer is: *Which is the impact of defocusing in soft biometrics analysis?*

While soft biometric traits may be less relevant for identity recognition, they are essential in other contexts such as customized advertisement or videosurveillance. Thus, soft biometric characteristics are often investigated in uncontrolled environment and on low quality data [95, 96, 97]. The study of technology's impact on soft biometrics, such as gender or age, is necessary to develop optimal approaches and tackle specific problems. Whereas gender recognition from images acquired with SL camera has already been studied in [98, 99], the influence of LF sensors has not been explored yet.

As described in Section 3.3.5, LF images can be represented as 2D standard pictures rendered with different focusing depths. In a scenario where several individuals are present

in the same acquired image but at different distances, like described in Section 3.4.1, the change of focusing depth can be useful to preserve privacy. Vice-versa, the knowledge about age or gender of a out-of-focus subject may help in the identification process.

Although out-of-focus is one crucial factor of image quality degradation, its impact in biometrics has not been largely studied and it is often associated to other kind of blurring processes. In [100], Fang Hua et al. analyse the influence of defocusing on face recognition using images from Q-FIRE database [101] generated by turning the camera ring and acquiring multiple shots. The goal of this Chapter is to explore the possibility to change the focusing level from an image acquired in a single shot. This will avoid pose or light variations, motion blur or different camera aberrations, yielding to concentrate on defocusing problem. In order to perform the analysis, LiFFD database presented in Section 3.4.1 is used.

LiFFID database does not include metadata information. Thus, we manually annotate the gender of each individual. Whereas human beings are usually able to recognize the gender of a person, we may fail to guess the exact age. The purpose of the analysis presented in this Chapter is not testing the performances of the used algorithms but studying the impact of different focusing depth on gender recognition and age estimation from face, independently of the truthfulness of the results.

5.2 Techniques

In this section, the techniques selected to perform our studies are contextualized and described. First, the metric used to assess image quality is defined. Then, the algorithms chosen to estimate gender and age are introduced.

5.2.1 Image quality

Image quality is a large research branch in computer vision, spanning in several domains and useful for several applications, such as final user experience in looking at the image or improving the performance of image processing algorithms [102]. In [102], Niu et al. compile a short overview on image quality assessment analysing distorted images in terms of fidelity, aesthetics and visual comfort. Several studies have been carried out to explore the impact of defocusing on image quality [103, 104, 105]. In particular, Hua et al. [100] analyse the influence of out-of-focus on face recognition through Modular Transfer Function (MTF). This function is defined by the Equation (5.1), where $k \in [0, N-1]$ and y_n is the position of the n^{th} pixel. It represents the magnitude of the Optical Transfer Function (OTF) and it may also be associated with the ratio between a defined value of

frequency and low frequencies.

$$MTF = \sum_{n=0}^{N-1} \left(y_n \exp\left(-ikn \frac{2\pi}{N} \right) \right) \tag{5.1}$$

In order to evaluate the quality of LF images rendered with different focusing depths, the MTF evaluated as the ratio between high frequencies and low frequencies is chosen. Thus, higher MTF values will correspond to sharper images.

5.2.2 Gender recognition

The recognition of hard and soft biometric traits in unconstrained environment is still an open challenge: in videosurveillance light and pose variations and low camera resolution hamper recognition algorithms from having good performance. Despite having obtained nearly perfect results of gender recognition in controlled environment [106], the accuracy on more challenging database like *Labeled Faces in the Wild* is still considerably low. In [107], authors combine LBP and SVM reaching an accuracy of 94.81%. Tapia and Perez [108] succeed to classify the 98% of subjects using *Labeled Faces in the Wild* for training and testing. In [109], authors explore a possible correlation between gender recognition and age but they do not outperform the previous works. One of the most performant algorithms employed on unconstrained database is described in [110]: the method is based on Local Binary Pattern features and C-Pegasos classifier and reports an accuracy of 96.86%.

Light Field Face and Iris Database, described in Section 3.4.1 is composed of images collected in unconstrained environment. Thus, a gender recognition algorithm that has achieved high performance in challenging cross-dataset protocols is chosen to carry on our studies. The minimalistic CNN-based model, described in [111], fits well the criteria required.

The used CNN is composed of an ensemble of 3 CNNs obtained by minimising the model proposed by Simonyan and Zisserman in [112].
To create and test the network, the authors used two well known databases, *CASIA WebFace* [113] and *Labeled Faces in the Wild* [114,115]. While the first is used only in the training phase, the second is employed only for testing. Both databases contain images with pose, illumination and resolution variation, resulting in a total of 16324 individuals, 10575 from *CASIA WebFace* and 5749 from *Labeled Faces in the Wild*. Faces contained in the images are detected, cropped and resized to 32x32 pixels.

Each used CNN is composed of convolutional layers with spatial domain of 3x3 pixels and the rectified linear units (ReLU) is used as activation function. The optimal layer width is set to 32 for the first two layers and 64 for the others. The fully-connected layer placed at the end of the process is composed of 16 neurons. The structure can be summarized as follow:

Input: 32x32
Conv: 32@3x3
Conv: 32@3x3
MaxPool: 2x2
Conv: 64@3x3
Conv: 64@3x3
MaxPool: 2x2
Fully-Con: 16

The output of the CNN model is a real number between 0 and 1 that represent the probability for the individual represented in the image to be female. More the number is close to 1, more the face is feminine and, vice versa, more the output is close to 0, more the subject is masculine.

5.2.3 Age estimation

Whereas human beings can easily recognize gender, guessing a person's age is not as intuitive as it might seem. Automatic algorithms tackle the same challenge since ageing process is different for each person and influenced by several environmental factors. Moreover, apparent age and biological age often do not coincide. That makes it more difficult to collect a database of apparent age, and thus, to have a complete analysis of the problem. In 2015, a dataset of faces labelled with apparent age is collected in order to conduct the first *ChaLearn Looking at People* competition and it is enlarged on 2016 [116]. The best performance on both biological and apparent age estimation is achieved using pre-trained CNN models and later fine-tuned on the specified problem [117], [118].

Multiple studies are done in order to better understand the influence of light and pose variation in biometrics. In [119], the authors explore a possible relationship between gender recognition and emotion. In [120] an algorithm designed to estimate age under

different light conditions is described. In [121], Multi-level Local Binary Pattern (MLBP), Gabor filtering, Principal Component Analysis and Support Vector Regression are used to describe the age characteristics of motion blurred images. In [122], the authors study the influence of privacy filters on automatic gender recognition and on a crowdsourcing classification, proving that the robustness of computer vision algorithm is close to human classification.

Generally, it is possible to divide the age estimation algorithms in three categories:

- **Real number**: the method is based on regression approach. The output age is a real number defined by a regression model.
- **0/1 classification**: the method is based on a multi-class classifier, where each class represent an "age" or a range of "ages". According with the model, the input image can belong to one or another age class.
- **Label distribution**: the model is a fuzzy multi-class classifier, where the output is a vector representing the probability to belong to each class.

In order to estimate age from face, the winning algorithm [123] of *2016 ChaLearn LAP* competition on Apparent Age Estimation [116] is chosen. The work is based on VGG-16 convolutional neural network pre-trained for face recognition [124], trained on *IMDB-Wiki dataset* [125] and fine-tuned using the competition dataset. Although the strong point of the winning approach is apparent age estimation on children, this algorithm is one of the most performant on adults too.

The training phase of the algorithm used can be divided in four steps:

- **Step 1**: a pre-trained VGG-16 CNN is trained with a label distribution method for 100 classes (ages between 0 and 99 years old with intervals of 10 years).
- **Step 2**: children age is more difficult to estimate and it requires a separate step. The network obtained in step 1 is fine-tuned using a 0/1 classifier with the purpose to distinguish children between 0 and 12 years old from adults.
- **Step 3**: the model created in step 1 is fine-tuned with a distribution label encoding with adult images. All the images are shuffled and 11-folder cross validation method is used to obtained 11 models.
- **Step 4**: the model created in step 2 is fine-tuned a second time with a 0/1 classification procedure on children images. In this case, only 3-folder cross validation method is used.

In the test phase, input images are preprocessed: face regions are detected and cropped. In order to minimize the impact of minor face alignment errors, 7 versions of the original image are created: mirrored, rotated at $\pm$ 5°, shifted by 5% on the left/right side, scaled in/out by 5%. The 8 obtained images are processed by the 11 models created during the training phase and 88 output vectors of size 100 are obtained. The n^{th} value of the m^{th} vector represents the probability according with the model m to belong to the class n. Thus, the general age is evaluated weighing each possible age for its probability (Equation (5.2)).

$$\text{general age} = \sum_{i=1}^{100} i * p_i \tag{5.2}$$

If the estimated age is below 12 years old, the 8 versions of the input image are processed with the 3 classifiers fine-tuned with children faces. The same weighing procedure is applied to the output.

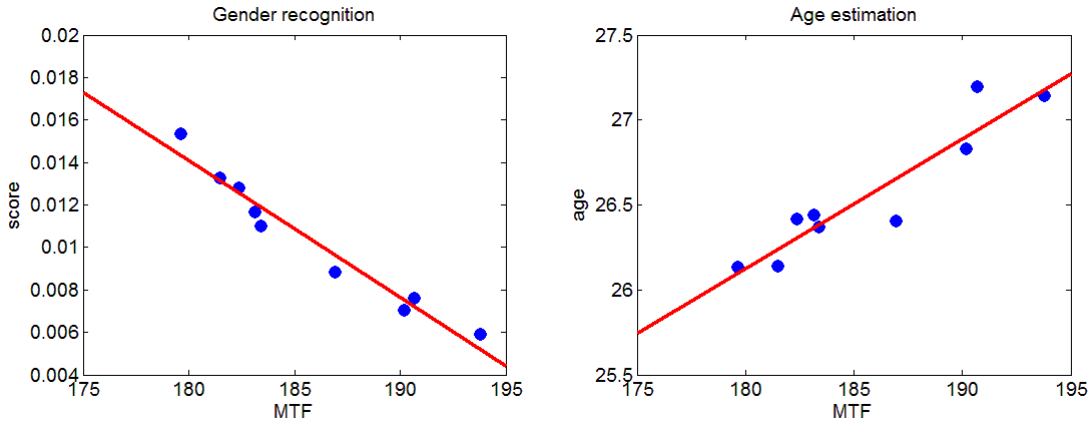
5.3 Preliminary analysis

In order to create a baseline, enrolment data, described in Section 3.4.1, are processed with the gender recognition algorithm, reaching an accuracy of 99.61%. The same analysis done on images extracted from LF images gives an accuracy of 95.25%, proving the high performances of the classifier on LiFFID database. Because of the absence of metadata in the database, it is not possible to create a baseline for age estimation from face.

LiFFID database provides several 2D conventional images rendered at different focusing depths extracted from the same LF data. The number of versions of the same image varies between 2 and 9 according with the data. The MTF described in Section 5.2.1 is used as measure of out-of-focus.

For each version of each LF image, the MTF score, the gender score and the age score are computed. In Figure 9.1, an example of the variation of gender score (Figure 9.1a) and age score (Figure 9.1b) evaluated at different focusing depths compared with MTF score is shown. It is important to highlight that in this example a single raw data, rendered at different focusing depths, is analysed.

The linear correlation between MTF and both gender and age is quite evident in Figure 9.1. In this particular case, the individual represented in the image is a young man. The more sharp is the image, the higher is the probability for the individual in the image to



(a) (b)

Figure 5.1 – Sample of gender (Figure 9.1a) and age (Figure 9.1b) variation in a single LF data rendered at different focusing depths

be classified as male (correct classification of the considered sample) and the older the subject looks like.

5.4 Experimental setup and results

The experiments conducted are based on the analysis of linear regression outputs. Thus, a short overview on linear regression method is presented.

Linear regression is a statistic method that aims modelling the relationship between a scalar value y and one or more variables x_i with $i \in [1, p]$.

$$y = \beta_0 1 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon_p \tag{5.3}$$

Where ϵ is the error or noise.

Knowing y and x_i, coefficients β_i with $i \in [1, p]$ are generally evaluated with the Least Squares algorithm. The variables are then selected according with the following hypothesis test:

$$H_0 : \quad \beta = 0$$
$$H_1 : \quad \beta \neq 0$$

The *p-value* of each hypothesis test has to be studied in order to estimate the significance of the variable. If the *p-value* is low, there are statistical reasons to reject the null hypothesis and, thus, to assume that it is possible to identify a correlation between the considered independent variable (x_i) and the dependent variable (y). The measure of how data fit in the proposed model is R^2. It is the ratio of the explained variation to the total variation and it gives a value between 0 and 1: the higher is the value, the better the data fits in the model.

For each model, it is necessary to find a trade-off between fitting and complexity in order to reach the goal without do overfitting.

5.4.1 Experiment 1: proposal for a model to describe the relationship between gender/age score and MTF

The first experiment aims to suggest a function able to model the relationship between gender (and age) and MTF. It is important to highlight that the goal is not to determine the values of β coefficients because they are related to image content, but rather the model that best describes the data.

The environment where data collection has been done may impact on the analysis. As described in Section 3.4.1, the LiFFID database is collected with three different protocols: 1. indoor scenario with controlled environment, 2. indoor scenario with uncontrolled illumination and 3. outdoor scenario with uncontrolled background. Thus, results are shown separately for each protocol.

Gender recognition

The Figure 9.1a shows that the correlation between gender score and MTF is linear for one set of images rendered from the same raw data. The goal of this experiment is to prove that the same model can describe each image set considered.

The first step is to select the data that are rendered with enough focusing depths in order to have sufficient statistics to perform the analysis. In fact, a small amount of focusing depth levels would not allow to study the real variation of the samples. The minimum amount of focusing depths chosen is 4. Then, for each set, 2 linear regression models are evaluated.

$$\text{Model 1:} \quad y = \beta_0 + \beta_1 x + \epsilon \tag{M11}$$

$$\text{Model 2:} \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \tag{M12}$$

Where the dependent variable (y) is the gender score and the independent variable (x) is the MTF.

For each model, the R^2 distribution is studied. A perfect representation of the model would portray a delta distribution at value 1. The high asymmetry of the histogram suggests that Model M12 fits better the data (Figure 5.2), but, in order to keep as lower as possible the number of regressors, the Model M11 is preferable.

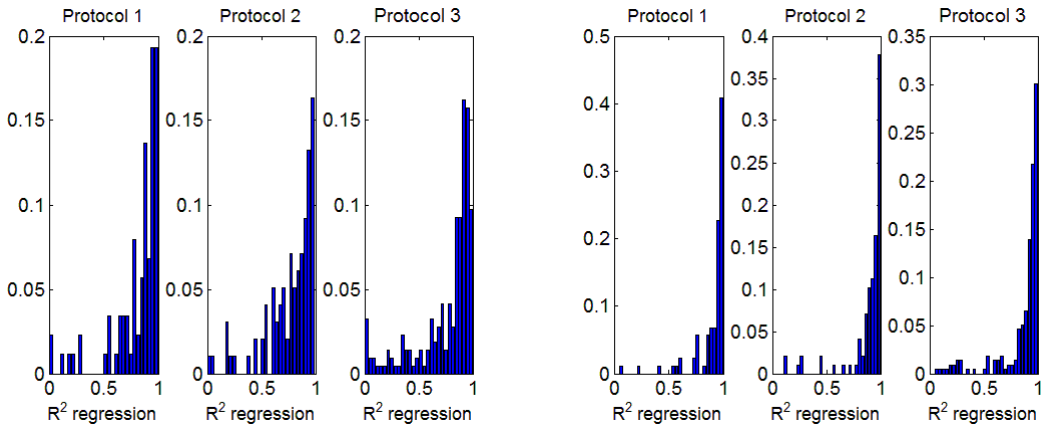
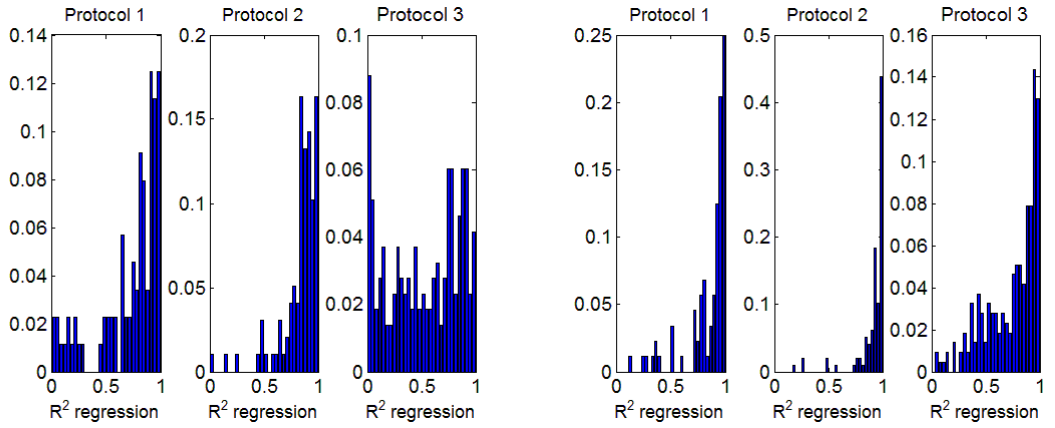


Figure 5.2 – R^2 histograms of Model M11 (Figure 5.2a) and Model M12 (Figure 5.2b) for gender recognition.

Age estimation

The same experiment is carried out on age scores. While in case of the gender recognition the protocol used during image collection has not a particular influence, here the different light conditions change critically the results (Figure 5.3a). An additional study on regression shape shows that the angular coefficient in protocol 1 and 2 is always positive and in protocol 3 only the 6,98% of the cases have negative coefficients. In other words, the sharper the image is, the older the individual looks like. That is not surprising: age markers (like skin spot or wrinkles) are mostly stored in high frequencies and become less visible when the picture is out-of-focus. The particular behaviour observed on protocol 3 could be explained considering the more challenge environment: on one hand, light variations during the acquisition process have almost no influence on gender recognition,



(a) (b)

Figure 5.3 – R^2 histograms of Model M11 (Figure 5.3a) and Model M12 (Figure 5.3b) for age estimation.

on the other hand, age estimation seems to be strongly influenced by environment variations.

5.4.2 Experiment 2: investigation on different impact of defocusing and Gaussian blurring on soft biometrics

In human vision, defocusing and Gaussian blurring are really similar although the generation processes are different. The second experiment is designed with the purpose of studying the possible difference between the impact of defocusing and Gaussian blurring on soft biometric traits. For each image, the best-on-focus version is chosen and seven Gaussian filters with dimension between 3x3 pixels and 15x15 pixels are applied. The experiment, presented in Section 5.4.1, is repeated on the new sets of images.

Gender recognition

In order to compare the relationship between gender score and MTF on defocused images and gender score and MTF on blurred images, different regression models for each set of images are evaluated, considering both the defocused and the blurred versions. Four different models are fitted. For all models, the dependent variable y is the gender score, the independent variable x is the MTF value and the dummy variable d has the value 1 for the blurred images and 0 for the defocused images. A high *p-value* for the coefficients related to dummy variables indicates that the difference between defocusing and blurring is not statistically significant.

$$\text{Model 1:} \quad y = \beta_0 + \beta_1 d + \beta_2 x + \beta_3 xd + \epsilon \tag{M21}$$

$$\text{Model 2:} \quad y = \beta_0 + \beta_1 d + \beta_2 x + \epsilon \tag{M22}$$

$$\text{Model 3:} \quad y = \beta_0 + \beta_1 d + \epsilon \tag{M23}$$

$$\text{Model 4:} \quad y = \beta_0 + \beta_1 x + \epsilon \tag{M24}$$

For this experiment, the percentage of low *p-values* for each coefficient is listed in Table 5.1 for different protocols. In the first model, although the percentage of high correlation coefficient is close to 87%, the two regressors related to the dummy variable have coefficients with values that are more likely equal to 0. In the second model the percentage related to R^2 is inferior to model 1 and again the dummy variable seems to be superfluous. The third model represents two horizontal regression lines and clearly it does not fit the data. The last model suggests that both defocusing and blurring influence in the same way the gender recognition classifier.

Table 5.1 – Gender recognition: percentage of R^2 superior to 0.7 and percentage of low p-value for coefficients evaluated for each model.

		$R^2 > 0.7$	***p-value*** < 0.2 (β_0)	***p-value*** < 0.2 (β_1)	***p-value*** < 0.2 (β_2)	***p-value*** < 0.2 (β_3)
P1	M21	86.36	78.78	37.87	74.24	36.36
	M22	80.30	100	43.93	98.48	/
	M23	0	75.75	42.42	/	/
	M24	71.21	100	98.48	/	/
P2	M21	85.71	79.76	38.09	78.57	30.95
	M22	72.61	97.61	21.42	97.61	/
	M23	0	76.19	33.33	/	/
	M24	65.47	97.61	98.80	/	/
P3	M21	83.74	81.77	50.73	85.22	48.27
	M22	71.92	95.07	50.24	95.07	/
	M23	0	90.14	42.85	/	/
	M24	56.65	92.61	94.58	/	/

Age estimation

The same experiment, described above for gender recognition, is applied on age estimation (Table 5.2). Also in this case, blurred and defocused images have the same regression line. Despite the fact that the percentage of *p-value* related to coefficients of dummy variables lower than 0.2 is higher than in the previous case. The model, defined as Model M24, can still be considered the best. In the third protocol, the percentage of high correlation is

lower than 40% for Model M24, that is due to the instability of age estimation algorithm on unconstrained environment.

Table 5.2 – Age estimation: percentage of R^2 superior to 0.7 and percentage of low *p-value* for coefficient evaluated for each model.

		$R^2 > 0.7$	***p-value*** < 0.2 (β_0)	***p-value*** < 0.2 (β_1)	***p-value*** < 0.2 (β_2)	***p-value*** < 0.2 (β_3)
P1	M21	87.87	59.09	42.42	68.18	36.36
	M22	81.81	84.84	40.90	98.48	/
	M23	1.51	100	54.54	/	/
	M24	74.24	78.78	98.48	/	/
P2	M21	95.23	54.76	47.61	84.52	41.66
	M22	86.90	65.47	35.71	100	/
	M23	0	100	29.76	/	/
	M24	80.95	66.66	98.80	/	/
P3	M21	63.05	75.86	50.24	70.93	46.79
	M22	44.82	84.24	47.78	85.71	/
	M23	0	100	34.97	/	/
	M24	34.97	84.72	84.72	/	/

Additional test

In order to confirm the hypothesis that the impact of Gaussian blurring and defocusing is the same for both gender recognition and age estimation, a model based on only blurred images is created. Then, the percentage of defocused images falling in the prevision interval is computed. The prevision interval is set with a confidence of 95%. Results, illustrated in Table 5.3, show that the 89% of LF data collected with protocol 1 have more than 80% of their versions in the prevision interval. Defocused images and the blurred ones follow a different behaviour when the acquisition is done in uncontrolled environment. The same experiment on age estimation gives coherent results regarding the previous analysis. While for the first two protocols it is possible to claim that blurred and defocused images have the same impact on the gender recognition and age estimation, the data collected with the third protocol gives different results.

Table 5.3 – Percentage of images of which at least of 80% and 50% of defocused versions are present in the prevision interval.

	Gender recognition ($> 80\%$)	Age estimation ($> 80\%$)	Gender recognition ($> 50\%$)	Age estimation ($> 50\%$)
Protocol 1	89.39	80.30	95.45	95.45
Protocol 2	84.34	75.90	95.18	87.95
Protocol 3	60.10	51.23	86.70	72.91

Table 5.4 – Percentage of images where the MTF increases after deblurring filter application.

	Defocused images	Blurred images
Protocol 1	55.77	100
Protocol 2	66.66	100
Protocol 3	70.77	100

5.4.3 Experiment 3: impact of deblurring filter on defocusing and blur images in the context of soft biometrics

In this experiment, the impact of deblurring filters on soft biometrics is studied using defocused images obtained from LF data. For each defocused version of each image, the most similar, in terms of MTF, blurring version of the same image is found. Knowing the window size of the convolution used to blur the image, the best deconvolution window for the defocused image is guessed. A blind deconvolution algorithm based on a convolution between image and Point Spread Function (PSF) is used as deblurring filter. The method tries to find the best PSF so that the resulting image has the higher probability of being an instance of the blurred image, with the assumption of Poisson noise statistics. Blind deconvolution can be used when no information about distortion is available, like in the case of defocused images. Whereas the impact of deblurring filter on blurred images is always positive and increases the MTF, it is effective only for a relatively small percentage of defocused images, as show in Table 5.4. The higher percentage of improvement in protocol 3 respect to the other protocols is due to the lower quality of the images, strongly influenced by the environment.

The Table 5.5 illustrates the percentage of improvements on gender recognition and age estimation. The improvement for gender recognition can be assessed straightforwardly by comparing the obtained results to the ground truth. For age estimation, the increasing of estimated age is considered as improvement. As expected, classification on defocused images is less subjected to deblurring filtering than classification on blurred images. Moreover, gender recognition appears more effected by deblurring respect to age estimation, possibly due to artefacts created by deblurring algorithm.

5.5 Conclusion

In this chapter, LF images from LiFFID database, described in Section 3.4.1, are used for two main goals: 1. to investigate how gender recognition and age estimation algorithms are effected by focusing depths and 2. if Gaussian blur and LF post processing defocusing

Table 5.5 – Percentage of images where deblurring improved the recognition of soft biometric traits.

	Gender recognition defocused images	Gender recognition blurred images	Age estimation defocused images	Age estimation blurred images
Protocol 1	61.55	78.73	39.11	62.33
Protocol 2	68.38	81.00	47.42	67.89
Protocol 3	67.77	82.60	64.84	71.29

have a difference impact on soft biometric methods.

After presenting the techniques used to assess image quality and to perform gender recognition and age estimation, three experiments are carried out on LF data.
First, the linear correlation between focus and gender recognition is verified: the sharper an image is, the more accurate is the classification. The same relationship is revealed for age estimation. However, since the ground truth for age is not provided with the database, it is impossible to evaluate the actual performance of the classifier. The experiment demonstrate that the sharper is the image, the higher the apparent age is: this is not surprising considering that age signs, as wrinkles and skin spots, are mainly present in high frequencies.
The hypothesis are tested on different environments and, although in sunlight condition the linearity is less obvious, the linear model is confirmed for all the acquisition protocols presented in the database.

In the second experiment, a unique model is proposed for defocused and blurred images. Regression with dummy variable and the study of prevision interval confirm that it is possible to assume the same behaviour for both distortions on gender recognition problem. Age estimation classifier looks more influenced by light variation and asserting that blurred and defocused images share the same model in the condition presented in protocol 3 would not be correct.

Finally, the application of deblurring filters on defocused images is tested. The quality improvement of out-of-focus LF images are found less evident than the one obtained for blurred images. Moreover, gender recognition algorithm is proved more effected by image quality compared to age estimation.

Chapter 6

Face recognition

6.1 Introduction

Face recognition is one of the main branch of face analysis. However, the improvements due to light field (LF) technology on this research topic are studied in literature only in recent years and not extensively. In this chapter, a short overview on the existent methods for face recognition based on LF images is compiled. Then, a specific algorithm designed to exploit the LF data characteristic to be rendered as sub-aperture images is deeply described.

The original contribution of this chapter consists in two parts. The first is a detailed compilation of face recognition methods based on LF images, currently under revision in Sensor journal Special Issue on Advanced Sensor for Face Analysis [13]. The work has been carried out in collaboration with *Instituto de Telecomunicações, Instituto Superior Técnico, Universitade de Lisboa*, in Portugal, *Hochschule* of Darmstadt, in Germany, and *Institut de recherche en informatique et systèmes aléatoires* (IRISA), in France. The second is a description of an innovative algorithm customized for sub-aperture images, presented in the 26^{th} edition of European Signal Processing Conference held in Rome in 2018 [11].

The research question that this chapter aims to answer is: *How can light field information enhance face recognition rate?*

6.2 State of the art in face recognition with light field

While most of the algorithms proposed for LF face analysis are customized to be applied on data acquired with plenoptic cameras (in particular from technologies released by Lytro), some works presented before 2006 use as database images collected with conventional

sensors. In [126, 127, 128], Gross et al. suggest to exploit pose and illumination variation collected in CMU PIE [129] and FERET [25] databases to create a LF model of individual faces. Pixels belonging to face are detected and LF is computed as shown in Figure 6.1. The recognition analysis is based on eigen LFs evaluated solving with least-squares method the equation described in Equation (6.1) where I represents the original image and $W(\theta, \phi)$ is the new base (eigen LFs in [126, 128] and Fisher LFs in [127]). The same concept is reconsidered in 2011 by Wibowo et al. [130] to perform face recognition from video sequences.

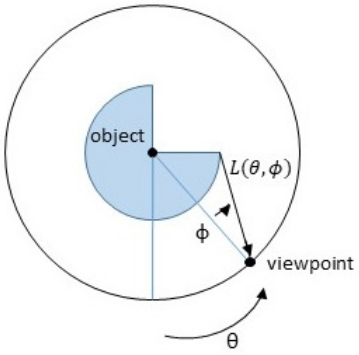


Figure 6.1 – A visual description of the method used to define light field from 2D images

$$I - \sum_{i=1}^{d} \lambda_i W_i(\theta, \phi) = 0 \tag{6.1}$$

In [131], Zhou et al. integrate the Lambertain reflectance model to the method proposed by Gross et al. in order to take in account illumination variations in addition to different poses.

6.2.1 Multi-focus based methods

The first studies on face recognition on LF images are based on plenoptic data rendered at different focusing depths. In 2013, Raghavendra et al. [2] published an innovative technique to extract the best-on-focus images from a set of different focus pictures. In addition to one of the first version of a LF database for biometrics, the authors introduce an approach to detect, select and extract features from LF data (Figure 6.2).

- The images in the database presented in [132] are processed with the Viola-Jones

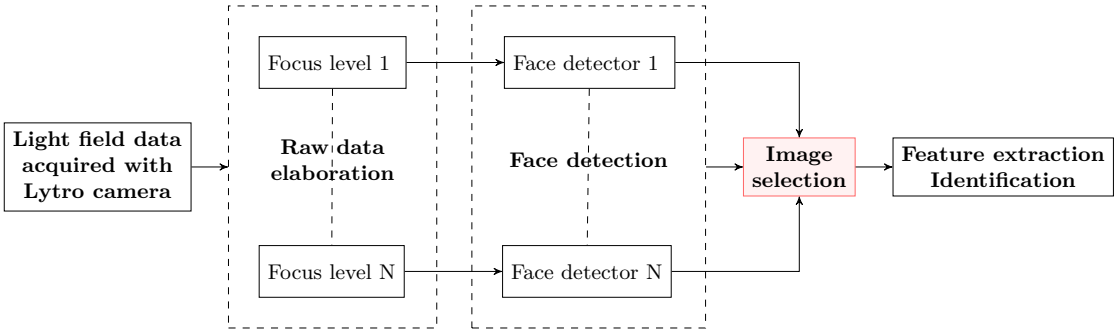


Figure 6.2 – Workflow proposed in [2]

face detector [48] trained with 2429 face images and 3000 non-face samples. The algorithm is applied on all data files rendered with different focusing depths. For each file, the image containing more detected faces is chosen and used to define the regions of the faces.

- Once that the faces are detected and cropped, the best image for each individual is selected according to energy. The authors chose as energy measure the 2D Discrete Wavelet Transform (DWT) with Haar wavelet because of its robustness to noise and its content-independent property. The rendered version associated with larger energy is chosen.

- Local Binary Pattern (LBP) features [49] and Log-Gabor (LG) features [133] are extracted separately and used as input of Kernel Discriminant Analysis (KDA) [134] and Sparse Reconstruction Classifier (SRC) [135].

During the creation of the database, images are acquired with both LF and conventional cameras. The purpose of Raghavendra et al. [2] is to compare the recognition rate obtained on LF data with the one achieved on standard 2D-images. The results show that it is possible to increase the identification rate of more than 7%.

In [73], the same authors improve the selection step fusing the face images rendered at different focus level to create a super-resoluted image. In this preliminary work, state of the art super-resolution techniques [136, 137, 138, 139, 140] are used to create super-resoluted images. Recognition algorithms result having better performances when applied on super-resoluted images than when applied on all-in-focus ones.

A novel weighted image fusion scheme is proposed by Raghavendra et al. in [141]. While the face detection is the same of the first work [2], the image selection is based on the measure of entropy. For each focus version of each detected face, the 2D-Discrete Wavelet

Transform (DWT) is applied and the log-entropy is evaluated through Equation (6.2)

$$E = -\sum_{i=1}^{K} \left(log_2 \left(W_i \right)^2 \right) \tag{6.2}$$

where W_i is the wavelet coefficient obtained for $i \in [1 : K]$.

Among all focusing depth levels, only the samples with positive entropy are kept. The remained image entropy is normalized and sorted in decreasing order so that the most in focus image appears as first. The difference between adjacent entropy values (see Equation (6.3)) is used to assign a weight to each image.

$$D_j = |Sor_{j+1} - Sor_j| \quad for \quad j \in [1 : \#images]\} \tag{6.3}$$

$$w_j = \begin{cases} (0.5 + (0.5 * D_j)) * Max_w & if \quad D_j \geq Th \\ \frac{Max_w}{2} & otherwise \end{cases} \tag{6.4}$$

where Th is empirically set to 0.2 and Max_w is initially equal to 1 and then updated for each sample j with the weight value of the sample $j + 1$.

Image samples are then fused with the weighted sum rule (see Equation (6.5)).

$$W_f = \sum_j W_j * w_j \tag{6.5}$$

where W_j is the j^{th} image in the discrete wavelet domain and w_j the corresponding weight. The final result W_f is converted in the spacial domain and used to extract features for performing face recognition.

The results are compared here with the performances obtained considering only the image with largest entropy. The identification rate achieved with the fusion scheme is higher in all the considered scenarios.

Method	Image selection	Baseline
[2]	Higher energy over multi-focus set	Conventional images
[73]	Super-resolution state-of-the-art methods	All-on-focus images
[141]	Fusion according with entropy energy	Higher entropy energy images
[142]	Substitution of lower sub-band image with super-resoluted image	Super-resoluted images
[141]	Fusion of two higher entropy energy images	All-in-focus

Table 6.1 – Summary of multi-focus based methods.

The scheme is further improved in [142] where a new hybrid resolution enhancement technique is proposed. Also in this case the first step of face detection is unvaried. As in [141], 2D-DWT is performed on images by applying filtering and downsampling with high-pass (H) and low-pass (L) filters on rows and columns. This process produces four sub-images: I_{LL}, I_{LH}, I_{HL} and I_{HH}. For each sub-image the wavelet energy is calculated and the image version with largest energy is selected to represent the sample. The sub-image I_{LL}, containing the lower-frequency band is replaced by a super-resoluted version of the original image obtained with a state of the art method [136,137,138,139,140]. The obtained results show that the algorithm outperforms other well-known super-resolution techniques in terms of identification rate.

In 2015, Raja et al. [143] study the problem of fusion in depth. Following an approach inspired by [141], they consider in the fusion scheme only the two images with highest energy. As for [142], the energy is calculated from the energy sum of three sub-images obtained with low-pass and high-pass filtering.

6.2.2 Sub-aperture based methods

The improvements achieved by using sub-aperture data on face recognition have already been studied using several technologies different from LF. Usually, views are collected with different devices at the same time [144] or with the same sensor with different shots [145]. In both cases, the data acquisition can be complex or require high degree of cooperation from the subject. Similar studies are presented in [146] where authors tackle face recognition problem creating sub-aperture representation from RGB-D images collected with Kinect sensor. Data are processed with a deep-learning algorithm in order to investigate how viewspace partitioning impacts on face recognition performance. Also in [147] a novel approach based on sub-aperture properties is used to recognize subjects illustrated in different poses. Synthetic face images are generated to imitate the other pose variations, thus helping in the recognition process under different perspectives.

With regard to LF technology, the publication of the *IST-EURECOM Light Field Face Database* (LFFD) [6] gives room to the development of approaches based on the full information provided by LFs data. The access to raw data allows, for example, the

investigation of sub-aperture representation impact on face recognition.

The first method based on sub-aperture visualization of LF data is proposed by Sepas-Moghaddam et al. [148] and is inspired by Local Binary Pattern (LBP) algorithm [49]. While the computation of the classic LBP feature vector requires adjacent pixels, the Light Field Local Binary Pattern (LFLBP) is composed by an additional component that includes the information stored in side-view images (Figure 6.3).

- Spacial Local Binary Pattern (SLBP): the first component is the LBP feature vector extracted from the central view of the LF image considered.

  Let (x, y) be the reference pixel sample, a the starting angle, p the number of selected values and $L_{0,0,x,y}$ the central view of a LF data

$$SLBP_{r,a,p} = \sum_{i=1}^{p} sign\left(L_{0,0,(x+k),(y+l)} - L_{0,0,x,y}\right) * 2^{i-1} \tag{6.6}$$

  where

$$\begin{cases} k &= \left[rsin\left(a + \frac{360^\circ}{p} * (i-1)\right)\right] \\ l &= \left[rcos\left(a + \frac{360^\circ}{p} * (i-1)\right)\right] \end{cases} \tag{6.7}$$

  and

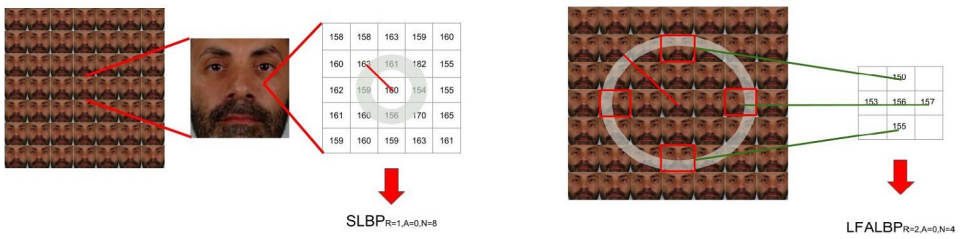
$$sign(x) = \begin{cases} 1, & if \quad x \geq 0 \\ 0, & otherwise \end{cases} \tag{6.8}$$

- Light Field Angular Local Binary Pattern (LFALBP): the second component is a LBP variation customized for LF images. Let (x, y) be the reference sample and R the radius representing the distance of the selected side-view from the central view,

$$LFALBP_{R,A,N}(x, y) = \sum_{j=1}^{N} sign\left(L_{u,v,x,y} - L_{0,0,x,y}\right) * 2^{j-1} \tag{6.9}$$

where

$$\begin{cases} u = \left[Rsin\left(A + \frac{360°}{N} * (j-1)\right)\right] \\ v = \left[Rcos\left(A + \frac{360°}{N} * (j-1)\right)\right] \end{cases} \tag{6.10}$$



(a) (b)

Figure 6.3 – Visual representation of SLBP Figure 6.3a and LFALBP Figure 6.3b used in [148]

6.2.3 Deep learning algorithms

In 2018, Sepas-Moghaddam et al. use for the first time a deep learning approach on LF images to deal with face recognition problem [3]. In this work, the authors choose to fuse several representations of the raw data in order to exploit as much as possible the information stored in the image. Features are extracted from three VGG-Face neural networks [149] and fused to feed a Support Vector Machine classifier (SVM).

- Central view is the input of a pre-trained VGG-Face model;
- The disparity map calculated from the sub-aperture representation is used to fine-tune a second VGG-Face model;
- A third VGG-Face model is fine-tuned on depth maps.

The analysis of the rank-1 recognition rate shows how the proposed algorithm outperform of 1.5% deep learning methods based on 2D-RGB images, 2D-RGB + disparity and 2D-RGB + depth map.

The same authors develop in a second work a double-deep spatio-angular learning structure [4] based on the analysis of several pictures obtained from the sub-aperture representation. Each considered view is processed with a pretrained VGG-Face network in order to extract a 4096-dimensional feature vector. The output of this first elaboration is used as input for a Long Short-Term Memory (LSTM) Recurrent Neural Network [150]

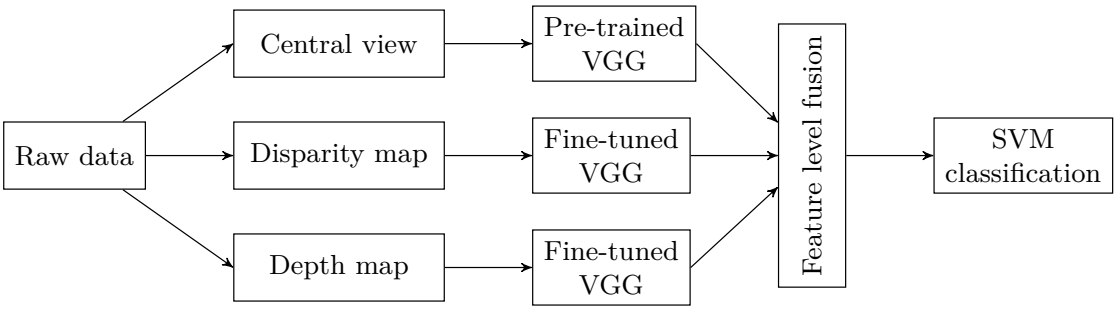


Figure 6.4 – Workflow proposed in [3]

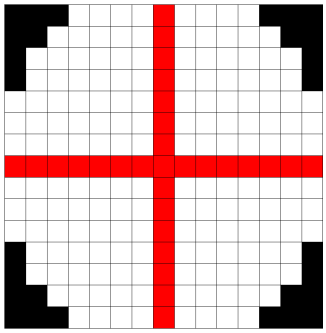


Figure 6.5 – The red squares represent the position of the view selected in the mid-density horizontal and vertical configuration chosen in [4].

that extracts the angular dependencies across the views. The last step is a Soft-Max classifier able to suggest the most probable identity of the individual represented in the image.

The use of all views extracted from a LF image would be computationally expensive and not necessary. The authors focus the study on the comparison between the application of this method on different views selected with different order. The configuration called "Mid-density horizontal and vertical" where the views considered are the one highlighted in Figure 6.5. So far, this method is the algorithm achieving better performance for face recognition for LF images.

6.3 OpenFace features

OpenFace (OF) features are a set of values extracted from a 2D standard image representing a human face. These features are able to map each image in a 128-dimensional space where data belonging to the same person are clustered automatically, facilitating face recognition process. The code used in this work has been provided by Amos et al. [46] and it is inspired by the paper published in 2015 by Schroff et al. [47].

The innovation of Schroff et al. is the learning approach used based on triplet loss minimization. Authors embed the function in a 128-dimensional Euclidean space and impose it to have results on a 128-dimensional hypersphere. Moreover, having three face images, data representing the same individual have to be closer each other respect to data representing different subjects.

Let x_i^a the anchor image, x_i^p a positive sample (image representing the same person in the anchor image) and x_i^n a negative sample (image representing a different person respect to the anchor image), the loss function is defined as Equation (6.11)

$$L = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ \tag{6.11}$$

Where $f(x) \in \mathbb{R}^{128}$ and $\|f(x)\|_2 = 1$.

The deep neural network used is GoogleNet Inception [151].

6.4 Proposed method

The proposed method can be inserted in the group of multi-view-based algorithm together with [148]. The main purpose of this study is not a comparison among recognition algorithms but rather a preliminary analysis to prove the additional value of LF data over 2D images. Since it would not be possible to compare face recognition performances without having customized methods for 3D images, a simple but effective algorithm based on sub-aperture images is proposed. At the time of this analysis only few algorithms were published. Thus, a complete comparison is out of the goal of the contribution.

In order to be able to exploit the whole additional information stored in LF data, the *IST-EURECOM Light Field Face Database* (LFFD) presented in Section 3.4.2 is used. Among the 20 face variation par subject, 6 significant classes (for a total of 1200 images) are selected avoiding that sever occlusions or strong pose distortions effect the results.

6.4.1 Preprocessing

The raw data are first processed with Lytro Power Tool. Among other functions and tools, this software is able to manipulate raw Lytro data and to render LF images as sub-aperture RGB images, each one collected from a slightly shifted point of view as described in Section 3.3.2. Contrary to MATLAB Toolbox, user is allowed to chose the

number of views to render and the desired shift between them. A small number of views would not be useful in analysing significantly the problem, a big number would lead to small variations difficult to detect. For this reason, a trade-off between view resolutions and number of selected views is evaluated: each data is transformed in a 5x5 RGB view matrix, each view with size 2022x1404 pixels (Figure 6.6).

Figure 6.6 – Example of sub-aperture representation of LFFD data

Each view is initially processed as conventional 2D picture. The face represented is detected, aligned and cropped with a pre-trained model based on Histogram of Oriented Gradients features available on the free library DLIB [89].

To perform face analysis, 3 sets of features are selected. Two of them are classic hand-crafted features, Local Binary Pattern (LBP) [49] and Local Gabor Binary Pattern (LGBP) [92], whereas the latter is based on deep-learning algorithm, OpenFace[1] (OF) [46]. All of them are selected due to their good performances on face recognition.

6.4.2 Preliminary analysis

Before to introduce the proposed recognition algorithm, we want to prove the presence of complementary information useful for face recognition in each view. In fact, most of the algorithms present in literature are developed with the purpose to be stable to small variations.

A simple test on the database is carried out in order to evaluate features behaviour. Each LF data is associated with classical 2D RGB image considering only the central view. One sample for each subject is kept as reference data and all other images are used as probe set. OF, LBP and LGBP features are extracted and compared with χ^2 and L^2 distance. Several classifiers are defined according with the threshold used to divide Matches and Non Matches. Results are shown in Table 6.2. Performances are evaluated and compared with Equal Error Rate (EER), False Non Match Rate for False Match Rate equal to $\frac{1}{1000}$ (FMR1000) and the lower False Non Match Rate which no

[1] OF features are described more in details in Section 6.3

False Matches occur (ZerosFMR).

	OF	LBP	LGBP
EER	0.0156	0.2099	0.1760
FMR1000	0.0938	0.7692	0.5623
ZeroFMR	0.2981	0.8773	0.7271

Table 6.2 – EER, FMR1000 and ZeroFMR related to OF, LBP and LGBP-based methods on central views obtained from LFFD. OF-based method outperforms handcrafted algorithms.

This preliminary analysis shows how recognition achieved with OF features outperforms the results obtained with LBP and LGBP features. That result has already been proved [152] and the comparison of handcrafted features and deep-learning methods is not the main purpose of this work.

Because of the structure of the device used for the acquisition, the differences between views are minimal, as shown in Figure 6.6. Before applying a method able to exploit multi-view properties, it is necessary to study how the change of perspective impacts on features computation. With this aim, OF, LBP and LGBP features are computed on all views of the same raw data separately, as if they were independent 2D conventional pictures. Distances among feature vectors of central view and all other views are computed.

The Table 6.3 reports the statistics related to the Relative Standard Deviation (RSD). The mean value and standard deviation are evaluated on all views of the same raw data and successively summarized on the whole considered database to be shown. In this preliminary step, all modalities and sessions of LFFD are used indifferently because intra-image differences (i.e. differences between views from the same plenoptic image) are not influenced by face variation.

	intra images	OF	LBP	LGBP
RSD	mean	**0.4411**	0.2858	0.2975
	var	0.0019	0.0434	0.04

Table 6.3 – Relative Standard Deviation statistics: the average value obtained with OF features is higher than the others. Therefore, a stronger impact of sub-aperture representation when using OF-based algorithm is expected.

The higher value of the RSD related to OF features indicates a stronger average variance between views of the same plenoptic image when they are represented by these features. This consideration suggests better results on sub-aperture fusion if performed with OF features.

The distance (in terms of face recognition) between the top left corner view and all

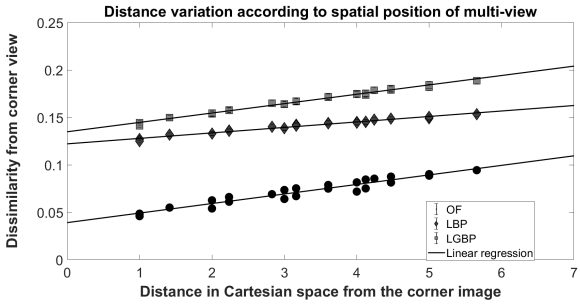


Figure 6.7 – Normalized Euclidean distance of each view respect to corner view in the same image v.s. space distance between the considered views. The solid lines are linear fit to the data, showing a linear relation between view shift and recognition algorithms.

the other views is computed to study the maximum variation achievable from a single data. Considering the structure of Lytro Illum camera, horizontal and vertical shifts are considered equally. Thus, pitch and yaw angles are not processed separately. The average normalized dissimilarity scores achieved in features space are evaluated as a function of the view shift. View shift is calculated as $D = \sqrt[2]{(i-1)+(j-1)}$ where i and j are the view index in the image array, so that the central view will be at $D = 2$. The linear trend between distances in features space and view shift is illustrated in Figure 6.7.

Better performances of OF features based methods respect to LBP or LGBP are expected because of the resulting steeper slope. Preliminary analyses show that the information stored in the sub-aperture representation of a human face are richer than a standard RGB image, especially when faces are mapped in the 128-dimensional hypersphere defined by OF features. Thus, further analyses are carried out in parallel with all described feature extraction algorithms but, since the conclusions are similar for all methods, only results related to OF features are commented.

6.5 Experimental setup and results

Two experiments are performed to study the impact of LF images on face recognition over 2D conventional images. While the comparison of a pair of 2D picture is well known, the matching operation between plenoptic data is not immediate. In the first case a feature vector is extracted for each image and a vector distance (in this case χ^2 and L^2 distances) is used. In the second case, one feature vector is extracted from each view of each image. Thus, the comparison between vectors becomes a comparison between sets of feature vectors.

To easily deal with this problem, a simple procedure is proposed: 1. computing conventional distance between all elements in set A and all elements in set B 2. reducing down

the pool of distances to one single value.

Eight functions able to extract a distance measure between sets are tested.

Let A and B be the sets of feature vectors describing the views from two raw data and d_e the conventional distance (selected according with the feature extraction method) between single elements. Thus, it is possible to define:

- Baseline: distance between the central view of two images.
- Min distance: minimum value through all the possible cross distances. $d_{min} = min_{a \in A, b \in B}(d_e(a, b))$
- Min distance corners: the minimum value through all the possible distances between corner views. $d_{minc} = min_{a \in A_c, b \in B_c}(d_e(a, b))$
- Mean pseudo-distance: the average value of all the possible cross distances. $d_{mean} = mean(d_e(a, b) \quad \forall a \in A, \forall b \in B)$
- Mean pseudo-distance corners: average value through all possible distances between corner views. $d_{meanc} = mean(d_e(a, b) \quad \forall a \in A_c, \forall b \in B_c)$
- Max pseudo-distance: the maximum value through all the possible cross distances. $d_{max} = max_{a \in A, b \in B}(d_e(a, b))$
- Max pseudo-distance corners: the maximum value through all the possible distances between corner views. $d_{maxc} = max_{a \in A_c, b \in B_c}(d_e(a, b))$
- Hausdorff mean distance: $d_{Hmean} = \frac{1}{\#A+\#B} \{\sum_{a \in A} min_{b \in B} d_e(a, b) + \sum_{b \in B} min_{a \in A} d_e(a, b)\}$
- Hausdorff max distance: $d_{Hmax} = Max \{max_{a \in A} min_{b \in B} d_e(a, B), max_{b \in B} min_{a \in A} d_e(b, A)\}$

It is important to highlight that not all the tested functions are distances: in fact, a function can be define "distance" in the space S if and only if:

$$d(x, y) = d(y, x) \quad \forall x, y \in S \tag{P1}$$

$$d(x, y) \geq 0 \quad \forall x, y \in S \tag{P2}$$

$$d(x, x) = 0 \quad \forall x \in S \tag{P3}$$

The last property is not fulfilled in pseudo-distances. In this work all the proposed functions are called "distances" according with the effective use of them. Both d_{min}

and d_{max} distances are studied. In fact, considering the minimum value of dissimilarity, the distance between matching samples decreases together with the distance between mismatching samples. Vice versa, when d_{max} distance is applied, the dissimilarity increases. In the first case, the probability of false matching increases, whereas in the second, the probability of false non-matching raises. Distances evaluated using only corner views are studied in order to develop computationally less expensive algorithms.

Once that the difference (in term of face recognition) between views is proved, the improvement of recognition capability due to 3D information over 2D conventional images is investigated. Two verification experiments are set up. The first follows a closed-set protocol where each subject considered during the validation phase is also used for testing. The second is an open-set experiment where the system is created with samples of 80 subjects and tested on the remaining 20 individuals. A cross-validation model is applied in order to generalize the conclusions.

In both cases, *frontal face* images from the first session of LFFD are used as reference data. Distances are computed between each plenoptic image of the probe set and each reference data. During the validation phase, some classifiers are defined (notably the distance representing the threshold below which the probe sample is matched with the reference one). Performances are shown comparing False Acceptance Rate (FAR) and False Rejected Rate (FRR) evaluated on the test set.

All analyses are compared with the results obtained evaluating the same experiment protocols on the central view of each data as if it was a 2D conventional image.

6.5.1 Closed-set experiment

In the closed-set experiment, the application of the suggested technique on face verification over different time span is studied. Thus, the validation and test sets are selected from the different acquisition sessions of the database.

Validation phase: Images acquired during the first session of LFFD are used. *Frontal face* variation is considered as reference (one image for each subject for a total of 100 raw data) and all other variations from the first session are used as validation set (one image for each subject for each modality, for a total of 500 raw data). The features distances between all reference and validation samples are computed in order to define nine (eight proposed and baseline) pools of linear classifiers, one for each distance definition. From each pool, the classifier corresponding to Equal Error Rate (EER) is chosen.

Test phase: The test set is composed of all data from the second session considered in this work (600 raw data).

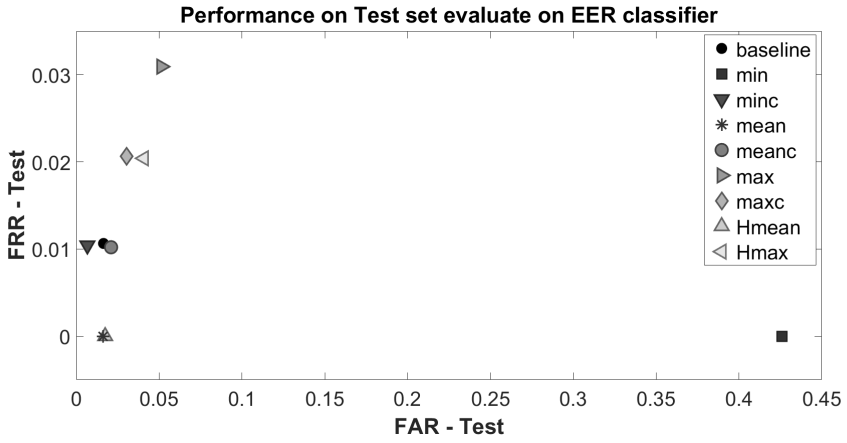


Figure 6.8 – Ex 1: Performance of classifiers evaluated on test set represented as False Acceptance Rate (FAR) vs False Rejected Rate (FRR). The best performances are obtained with d_{mean} and d_{Hmean} distances

Results: In Figure 6.8 the False Acceptance Rate (FAR) vs False Rejection Rate (FRR) on the test set of the nine classifiers created with the different distances from OF feature vectors is represented. While the classifier based on d_{min} distance obtains low FRR in spite of high FAR, the one with d_{max} distance shows the opposite results. Both d_{mean} and d_{Hmean} classifiers outperform baseline classifier obtaining accuracy respectively equal to 99.20% and 99.13% v.s. 98.65%, notwithstanding the fact that the latter already perform at an high level of accuracy.

6.5.2 Open-set experiment

The purpose of the open-set experiment is to demonstrate how the described algorithms can be successfully tested on subjects that are not considered during the validation phase.

Validation phase: Only 80% of LFFD subjects are considered during the validation process. All face variations of the 80 subjects are used to define the classification threshold (880 raw data). A pool of classifiers for each distance is created and the one corresponding to EER is chosen.

Test phase: All face variations illustrating the remaining subjects are included in the test set (220 raw data).

Cross-validation: With the aim of improving analysis stability, a cross-validation algorithm is applied. Validation and test phases are repeated 100 times splitting the database randomly. In Figure 6.9, a representation of EER distributions for OF, LBP and LGBP evaluated on validation set is shown. As suggested in Section 6.4.2, OF features have

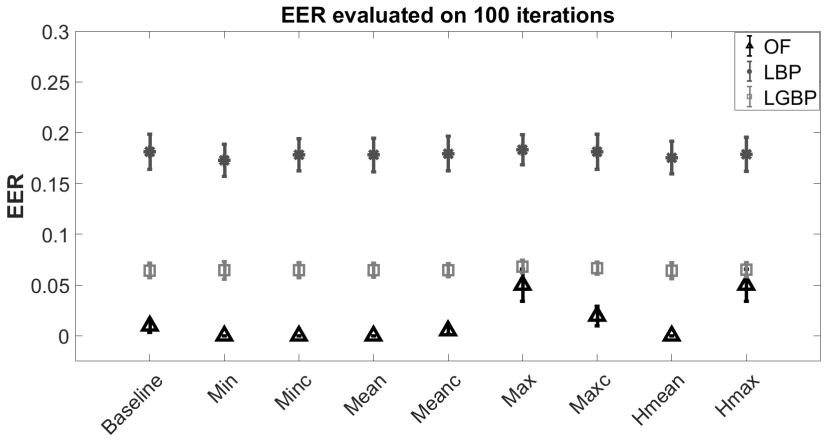


Figure 6.9 – Ex 2: EER evaluated on the validation set for different distances and features. While EERs relative to LBP and LGBP features have a stable behavior among different distances, EERs obtained with OF features classifiers present a higher variance

not only lower EER but also a higher variance among different distance classifiers.

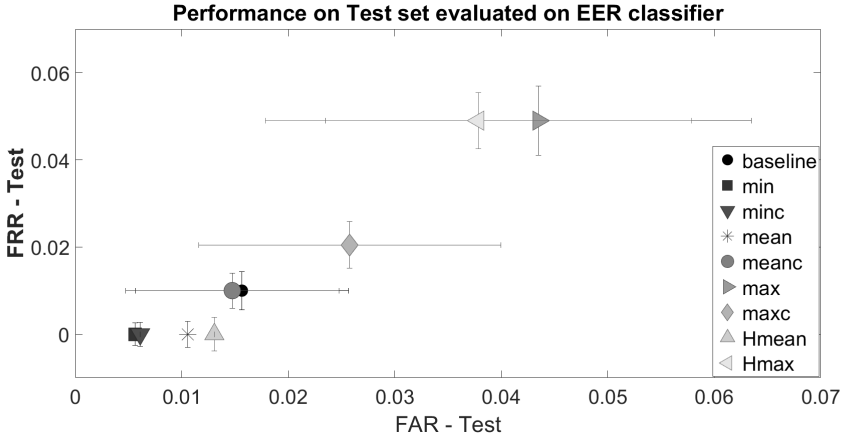


Figure 6.10 – Ex 2: Performance of classifiers evaluated on test set represented as False Acceptance Rate (FAR) vs False Rejected Rate (FRR). The best performances are obtained with d_{min} and d_{minc} distances

Results: In Figure 6.10, results obtained with OF features are illustrated. Like in the first experiment, d_{min} and d_{minc} classifiers outperform baseline with a respective accuracy of 99.80% and 99.78% v.s. 99.27%. d_{max}, d_{Hmax} and d_{maxc} classifiers do not improve verification results.

6.6 Conclusion

In this chapter, the impact of LF technology on face recognition problem is presented. After describing the state of the art on LF face recognition, an new method customized for sub-aperture images is proposed.

The images from *IST-EURECOM Light Field Face Database* are rendered as a 5 by 5 2D standard pictures array for data analysis. A preliminary study proves how LF data are richer than RGB images. Then, cross pseudo-distances among views of two images are computed. Eight new distances designed for reducing down multiple values to a single one are defined. Two verification experiments are carried out. The close-set experiment shows how d_{mean} and d_{Hmean} distances can improve face verification performances on images collected in a different period than the samples used during the validation phase. The open-set experiment demonstrates how d_{min} and d_{minc} distances could be successfully tested on subjects that are not considered during the validation phase.

The analysis paves the way to a more exhaustive study on impact of sub-aperture fusion on face recognition and verification when bigger pitch and yaw angles are applied between subject face axes and camera. This can be achieved with Lytro Illum by taking closer snapshots or with different kind of LF cameras (e.g. Raytrix).

Chapter 7

Face presentation attack detection

7.1 Introduction

In this Chapter, the concept of face presentation attack detection is introduced and an overview on the the use of LF images to perform anti-spoofing is presented. An innovative technique based on depth map analysis is shown and compared with other state of the art algorithms through three experiments.

The main contributions are two: 1. a summary of the state of the art algorithms for presentation attack detection with LF images, currently under revision in Sensor journal Special Issue on Advanced Sensor for Face Analysis [13]. The work has been carried out in collaboration with *Instituto de Telecomunicações, Instituto Superior Técnico, Universitade de Lisboa*, in Portugal, *Hochschule* of Darmstadt, in Germany, and *Institut de recherche en informatique et systèmes aléatoires* (IRISA), in France, and 2. the description of an innovative algorithm based on depth map representation presented in the 17^{th} edition of International Conference of the Biometrics Special Interest Group (BioSig) in Darmstadt in 2018 [12].

The research question that this chapter aims to answer is: *Can light field technology lead toward more robust anti-spoofing strategies?*

7.2 Presentation attack detection with light field images

With the diffusion of unsupervised face recognition systems to protect private accesses and to identify individuals, the risk of impostor attacks is becoming a major concern. In the last decades, academic and industry research has been working in order to detect subversive activities. *ISO/IEC 30107 Biometric presentation attack detection* [153, 154], a document released by the biometric community, tries to uniform the presentation attack

concept. In this document, *presentation attack* is defined as a presentation, directed to the biometric data capture subsystem, with the goal of interfering with the operation of the biometric system.

In literature, two types of attacks are generally considered. The first is the *Active Impostor Presentation Attack* (AIPA) in which the attacker aims to be recognized as a specific subject known to the system (e.g. an impersonation attack), spoofing the biometric capture. The second is called *Identity Concealer Presentation Attack* (ICPA) where the attacker wants to avoid being matched with its own identity. While for ICPA, a non-conform behaviour (such as extreme facial expressions) can be sufficient to escape the identification, in AIPA the attacker needs an artificial object or representation of biometric characteristic or synthetic biometric patterns to interact with the capture device. According with the technology used to acquire the biometric measure, the object used to spoof the system can differ: the set of possible attack artefacts can be composed by several materials among which artificial displays, pictures or 3D masks.
In this work we focus on AIPA attacks, discarding the analysis of ICPA attacks.

Figure 7.1 – Example of active face presentation attack with printed paper

Generally, two metrics are used to evaluate the Presentation Attack Detection (PAD) system: 1. Attack Presentation Classification Error Rate (APCER) that defines the ratio of presentation attacks incorrectly classified as Bona Fide (BF) presentations and 2. Bona Fide Presentation Classification Error Rate (BPCER), that defines the ration of BF samples incorrectly classified as presentation attacks.

Let RES_i be

$$RES_i = \begin{cases} 1 & \text{if the } i^{th} \text{ sample is classified as PA} \\ 0 & \text{if the } i^{th} \text{ sample is classified as BF} \end{cases} \tag{7.1}$$

The APCER and BPCER can be calculated as follows:

$$APCER = \frac{1}{N_{PA}} \sum_{i=1}^{N_{PA}} (1 - RES_i) \tag{7.2}$$

$$BPCER = \frac{1}{N_{BF}} \sum_{i=1}^{N_{BF}} RES_i$$

Where, N_{PA} is the number of attack presentations and N_{BF} is the number of BF presentations. The mean value of APCER and BPCER is defined as Average Classification Error Rate (ACER) and it is often used to summarize the system performances in a single value.

The vulnerability of face recognition capture devices for presentation attacks has intensively being discussed in literature. Given the easiness to render high resolution video material on low cost tablets, the limits to detect such attacks with a conventional 2D face capture device are easy to imagine.
Soon after LF capture devices became available, they have been investigated as means of defence against presentation attacks with 2D presentation attack instruments. The motivation is straightforward as the LF provides a superset of data acquired from the capture subject, which allows not only focus analysis at various depths but also disparity exploitation.

Kim et al. [155] investigate edge features with inner and outer binary as well as a ray difference feature to distinguish Bona Fide presentations from attack presentations conducted with printed PAD an high resolution tablet. The work reports an ACER in the range of 0.89% to 4.10% for the edge feature and 2.5% to 4.22% for the ray difference feature.
In Raghavendra et al. [156] the focus variation between images at multiple depths is analysed. Similarly, presentation attack instruments that are either high quality printed facial photos (both laser and ink jet printers are used) or an electronic display are employed. A measure expressing the variation in focus is classified by a Support Vector Machine (SVM). On a dedicated dataset, detection accuracy is reported with ACER from 4.01% to 5.27% depending on the presentation attack instrument.
Later Ji et al [157] propose a PAD subsystem that is based on a Light Field Histogram of Gradients (LFHoG) descriptor and includes the gradients in three directions (vertical, horizontal and depth). The features are again classified by a SVM classifier. The approach integrates the distribution of color intensity and the distribution of scene depth simultaneously. Detection accuracy is reported as 99.75%, while omitting details on

APCER and BPCER.
Also convolutional neural networks (CNN), implemented with the Tensorflow framework, are applied recently for PAD on LF data [158]. The approach is based on two different features: the microlens image and the ray difference image. The CNN is configured to process an input of size 250 x 250 x 3, where 3 corresponds to the RGB color channels. The results are reported using a benchmark with printed photos, warped photos and screen displays. The best ACER reached for the microlens features is 0.028%.

In 2018, Sepas-Moghaddam et al. [5, 159] present an overview on LF based PAD and develop novel methods exploiting the disparity information available in LF image. They analyse the variations observed for multiple directions in which the light is captured. Compliant to the ISO/IEC standard [153] they report BPCER at a fixed 1% APCER. Detection accuracy for the set of presentation attack instruments (including laptop, tablet, mobile and paper) ranges from 0% to 0.45%. The two described methods are selected to be reproduced in this thesis because they have been previously applied on the database chosen to test the proposed method.

The algorithm presented in [5] is quite similar to the one described in Section 6.2.2: variation of the classical LBP algorithm customized for LF images is defined. Instead of considering adjacent pixels, values from different views transposed in the HVS and YC_bC_r color spaces are used. In this scheme, two classifiers are created, one trained with LBP_{HVS} features and one with $LBP_{YC_bC_r}$. Scores are merged in order to give the final classification result. Figure 7.2 shows a schematic representation of the framework.

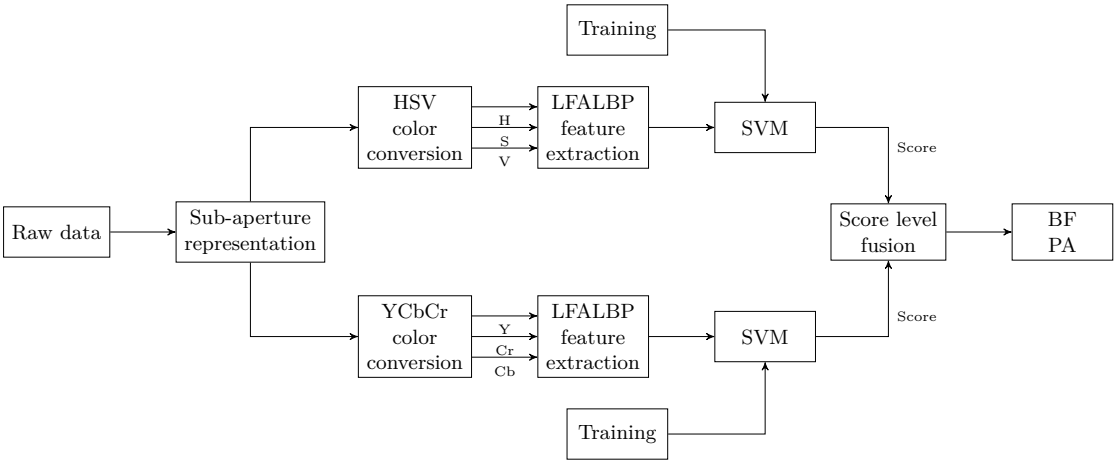


Figure 7.2 – Schematic representation of the workflow to compute $LBP_{HVS+YC_bC_r}$ features [5].

The method shown in [159] is inspired to [5] but, instead of LBP algorithm, it emulates the well known Histogram of Oriented Gradients (HOG) method [160]. Raw data are preprocessed with MATLAB Light Field Toolbox: thus, each image is represented

as a 4-dimension matrix ($L(u, v, x, y)$) where the first two values (u, v) indicate the considered sub-aperture and the last two values (x, y) the position of the pixel in the sup-aperture image. The horizontal ($G_x(x, y)$) and vertical ($G_y(x, y)$) gradients are defined by Equation (7.3).

$$\begin{cases} G_x(x, y) = L(u_1, v_1, x, y) - L(u_2, v_2, x, y) \\ G_y(x, y) = L(u_3, v_3, x, y) - L(u_4, v_4, x, y) \end{cases} \tag{7.3}$$

Authors prove empirically that the most suitable sup-aperture images to perform PAD are: $(u_1 = 15, v_1 = 8)$, $(u_2 = 1, v_2 = 8)$, $(u_3 = 8, v_3 = 15)$, $(u_4 = 8, v_4 = 1)$[1]. Then, the disparity gradient and the orientation are evaluated as in Equation (7.4).

$$\begin{cases} |\bigtriangledown I(x, y)| = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \\ \theta(x, y) = arctan(\frac{G_x(x,y)}{G_y(x,y)}) \end{cases} \tag{7.4}$$

The quantization, the normalization and the concatenation are performed following the standard HOG algorithm [159].

What is lacking for LF PAD research is a comparative benchmark on a database that includes sophisticated silicone mask [161], which should be considered now state of the art in face presentation attacks.

7.3 Database presentation

In order to investigate antispoofing methods customised for LF images, the *IST Lenslet Light Field Face Spoofing Database* (IST LLFFSD) is presented by Sepas-Moghaddam et al [162]. Starting from the neutral face collected in LFFD database[2] [6], the authors create a database including several presentation attacks collected with Lytro Illum camera. The 2D RGB picture is printed or displayed with different devices and a new LF image is acquired (Figure 7.3). The involved presentation attacks are the following:

- *Printed paper*: image printed on a A4 white paper with a colour laser printed. The paper is placed on a flat surface;

[1] The sub-aperture image in the top-left corner has indexes $(1, 1)$

[2] Only the first 50 subjects of LFFD database are considered.

- *Wrapped printed paper*: image printed on a A4 white paper with a colour laser printed. The paper is wrapped around a cylindrical object;
- *Laptop*: image displayed on MacBook Pro 13";
- *Tablet*: image displayed on iPad Air2 9.7";
- *Mobile 1*: image displayed on iPhone 6S;
- *Mobile 2*: image displayed on Sony Xperia Z2.

Figure 7.3 – IST LLFFSD sample: Original image, printed paper attack, wrapped printed paper attack, laptop attack, tablet attack, mobile 1 attack and mobile 2 attack.

7.4 Proposed feature

The contribution of this thesis on presentation attack detection using light field technology consists in the proposition of a new method based on depth map representation.

Raw data are elaborated with the proprietary software Lytro Desktop and for each acquisition, a pair of RGB and depth images is extracted. The software is able to set a perfect matching between RGB image and depth map so that a depth value is associated to each pixel of texture image. First, a landmark detection is performed on RGB image with a method based on HOG. The implementation is described in [163] and the model is trained on the database described in [164]. The algorithm identifies 68 landmarks for each human face in the image and detects a rectangular Region Of Interest (ROI). The

size of the ROI depends on the dimension of the represented face and, in this case, it does not impact in features computation. Then, for each landmark, the associated depth value is taken in account. In order to smooth some eventual noise, the depth map is convoluted with a 7x7 pixels average filter. Landmark Depth Features (LDF) are defined as the set of depth values associated to the 68 landmarks (Figure 7.4).

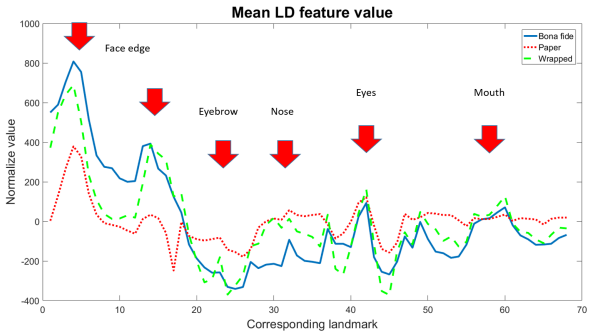


Figure 7.4 – Average LDF value. In order to be represented in the same graph, the mean value of each curve has been subtracted. On x axis is reported the number of corresponding landmark (landmarks order is not indicative. The corresponding map can be find in [163]).

The depth of specific landmarks could be more effective than others for detecting a presentation attack. Thus, with the purpose of investigating the linear combination of the landmark depth, a Principal Component Analysis (PCA) is performed. Fifty sets of images composed of an equal number of BF samples and of PA samples are randomly created, equally distributed among the possible attacks. The application of PCA to the whole database proves that more than 99.99% of information (evaluated as cumulative sum of eigenvalues) is stored in the first 10 principal components. Thus, Principal Landmark Depth Features (PLDF) are defined as the first 10 principal components computed as previously described.

The two sets of features (LDF and PLDF) are tested separately.

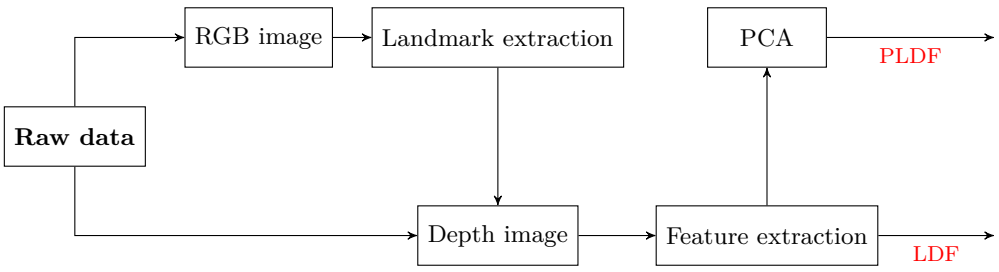


Figure 7.5 – Workflow of proposed PAD method

7.5 A Preliminary Study

A preliminary study on PLDF shows an evident data clusterization according with attack type (Figure 7.6). The analysis carried out in Chapter 4.3 demonstrate that the LF depth information is deeply influenced by light conditions.

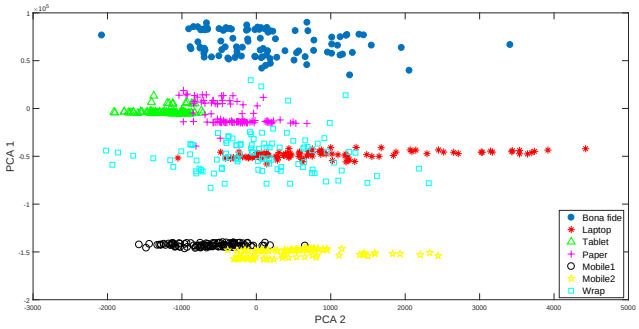
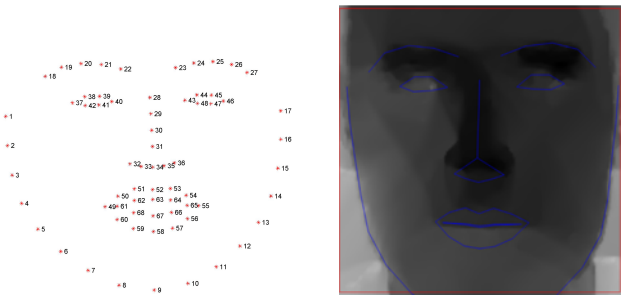


Figure 7.6 – Samples represented in 2D space created with the first two principal components. It is possible to observe a clear clusterization according with the first principal component.

Nevertheless, a focused investigation of screen backlight impact has not been carried out yet. Although the reason of this clusterization is not completely explained, it may be ascribed to different distances between the subject and the camera during data acquisition. This information can be important in PAD. In fact, when the attack is performed with devices such as smartphone, the identification of spoofed images is straightforward: the size of the screen is considerate being smaller than a human face and the device must be held closer to the camera in order to present a credible image. Contrary to standard cameras, LF sensors can easily detect the proximity analysing the depth map values. Thus, if face recognition system is based on plenoptic cameras, an eventual impostor tackles an additional challenge: the size of spoofed face has to be plausible.



(a) (b)

Figure 7.7 – Landmark map (Figure 7.7a) and landmark on depth image (Figure 7.7b)

7.6 Experiment

Experiment performances are evaluated considering Bona fide Presentation Classification Error Rate (BPCER), Attack Presentation Classification Error Rate (APCER) and the average value of them called Average Classification Error Rate (ACER) [55]. During the training phase of all the experiments, a optimal threshold corresponding to Equal Error Rate (EER) point is defined. The reported values are evaluated on the test set using the threshold chosen in the training phase.

7.6.1 Experiment 1: detection of a specific presentation attack instrument

In the first experiment, data are randomly split in training (75% of samples) and test set (25% of samples). The samples used in training phase are not considered in the test set. The number of bona fide and artifact images is kept equal in the training set to avoid unbalanced results. A C-SVM with linear kernel [165] is used as classifier. One attack modality per time is processed; thus, six classifiers are created. The experiment is repeated 50 times with different compositions of training and test set in order to prevent overfitting. In Table 7.1, the ACER is presented for each spoofing attack and for each considered feature set. ACER value is lower than 5% for all combinations of features and presentation attack modalities. LDF-based method is able to perfectly classify most of the attacks and it has a small percentage of failure ($< 0.75\%$) only in facing wrapped printed paper spoofing .

Features	Pap	Tab	Lap	Mob1	Mob2	Wrap
$LBP_{HSV+YC_bC_r}$	0.68%	1.87%	0%	2.11%	0%	1.14%
HDG	0.80%	2.58%	0.16%	4.08%	2.27%	**0.28%**
LDF (Proposed method)	**0%**	**0%**	**0%**	**0%**	**0%**	0.46%
PLDF (Proposed method)	**0%**	**0%**	**0%**	**0%**	**0%**	0.72%

Table 7.1 – Average ACER value evaluated over 50 runs of experiment 1.

7.6.2 Experiment 2: detection of several known presentation attack instruments

In real world scenarios, it is often not possible to know in advance in which way the recognition system will be attacked. The second experiment is designed to create a classifier able to recognize all the attacks present in the database at the same time. As for *experiment 1*, the training set contains an equal number of BF samples and artefact. Conversely, all the spoofing attack modalities in the database are considered together as artefact samples. The procedure is applied 50 times with different split of training

and test set in order to avoid overfitting. The performances of the classifier (a C-SVM as described in *experiment 1*) are reported for each attack. In Figure 7.8, DET curves for each presentation attack modality are shown. Average ACER values are presented in Table 7.2. As for experiment 1, LDF method outperforms the algorithms proposed in [162] and [160].

Features	Pap	Lap	Tab	Mob1	Mob2	Wrap
$LBP_{HSV+YC_bC_r}$	2.09%	3.15%	3.98%	2.04%	4.17%	6.56%
HDG	3.96%	4.04%	19.08%	3.99%	3.96%	8.17%
LDF (Proposed method)	**0%**	**0%**	**0%**	**0%**	**0%**	**0.74%**
PLDF (Proposed method)	**0%**	**0%**	**0%**	**0%**	**0%**	0.8%

Table 7.2 – Average ACER value evaluated over 50 runs of experiment 2.

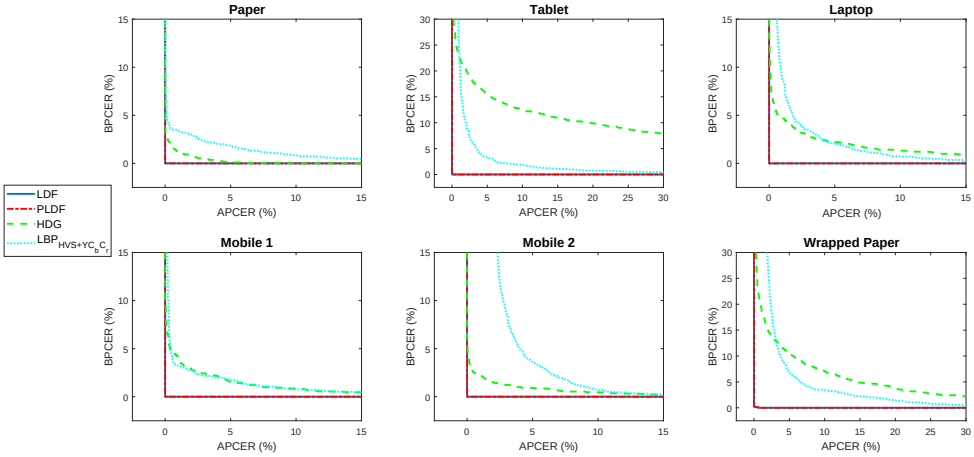


Figure 7.8 – DET curve for experiment 2. The proposed method (LDF) outperforms the other algorithms tested when evaluated on all considered presentation attacks. Also the performances of the presented reduced version (PLDF) reaches good results.

7.6.3 Experiment 3: detection of unknown presentation attack instruments

In the last experiment, no spoofed images are involved in training phase. A One-Class SVM trained only with bona fide images is used to identify all the presentation attacks included in the database. Starting from the elaboration of the information stored in only one class (in this case bona fide images) the One-Class classifier is able to discriminate test samples in two classes: belonging and not belonging to a given training class. As in experiments described in *experiment 1* and *experiment 2*, the kernel of the SVM is linear; the parameter ν is chosen empirically for each set of features. Since the One-Class SVM implementation does not provide output scores, LBP_{HSV} and $LBP_{YC_bC_r}$ feature sets are tested separately. The size of the training set impacts significantly on the stability of

One-Class SVM. The percentage of ACER varying the training set size is represented in Figure 7.9. While for the proposed method (LDF) a training set composed by 30 bona fide samples leads to high accuracy, HDG-based method reaches a stable accuracy with a bigger training set, because of the higher number of features used (23400). The state of the art methods here represented do not perform as well as LDF and they may not be as versatile as the proposed algorithm when changing the presentation attack. In particular LBP_{HSV} and $LBP_{YC_bC_r}$ performances decrease when the methods are seen separately [162].

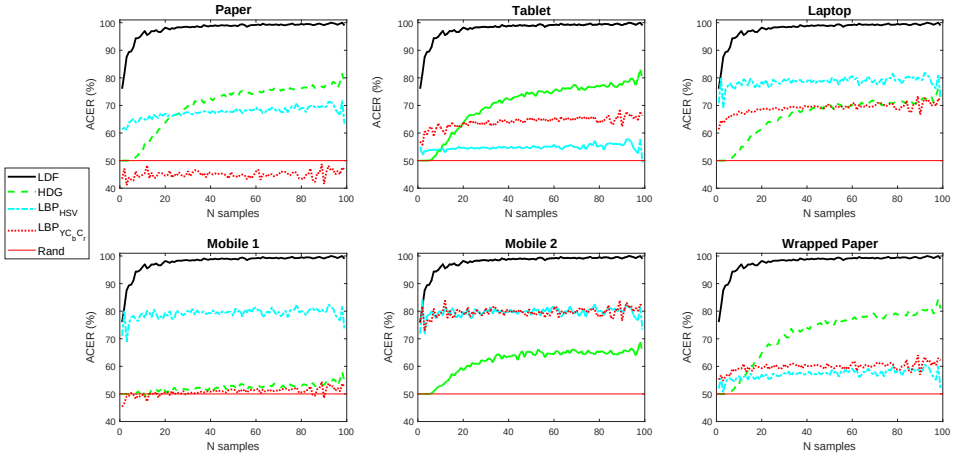


Figure 7.9 – Average ACER value varying training size. Convergence is reached with a small training set for all methods. In all studied cases, LDF algorithm obtains the higher ACER.

7.6.4 Complexity analysis

Although the performances of classifiers trained with the proposed features (LDF) are not always higher than the state of art alternatives, the complexity is clearly lower. In fact, the feature size of the proposed method is 68 or 50 for the reduced version (PLDF) while in HDG algorithm 24300 features are created for each image. The number of features used for $LBP_{HVS+YC_bC_r}$ method is comparable with the size of our algorithm (96), but it requires two classifiers trained and tested separately.

7.7 Conclusion

In this Chapter, the problem of presentation attack detection in face recognition systems is tackled for plenoptic images. After compiling a state of the art works on PAD with LF image, a new method based on Lytro images property of being rendered as pair of

RGB and depth map is presented. A preliminary analysis shows how distance between subject and camera can impact on the performance of a system able to identify spoofed images. Thus, it is easy to detect a presentation attack realized with a smaller (or bigger) support than an average face: in fact, the impostor has to hold the fake representation at a different distance from the camera to show a credible face. Three experiments are carried out in order to study the characteristics of the proposed features in different training conditions. While in the first experiment, each presentation attack is separately analyzed, in the second all assaults are considered during the classifier training. The last experiment tests the possibility to use only bona fide sample to train a system able to detect any attack. All the experiments lead to almost perfect classification results obtained with the proposed method, outperforming the state of the art approaches. This work shows how presentation attacks designed for standard 2D cameras are not effective in spoofing plenoptic sensors. Thus, it paves the way for the study of attacks customized for LF sensors. Moreover, a deeper investigation of distance subject-camera effect is becoming necessary to improve the analysis on face recognition on pleniptic images. With this aim, a new database acquired with a particular attention to distance subject-camera should be collected.

Chapter 8

Conclusions

In last 30 years, the presence of biometric systems in the daily life has become more and more important. Automatic border controls, smart objects, security accesses are only a few examples of how biometrics impacts on our habits and how it can change the quality of our lives. Like in every scientific fields, research on biometrics depends on technologies: computers with higher computational power, bigger storage capacity and better acquisition devices have to be investigated in order to fully exploit biometric trait potentiality.

This Ph.D. thesis has two main goals: 1. to present a technology so far poorly investigated in biometrics domain and 2. to explore the benefits on face analysis provided by this technology.

The scarcity of face databases with light field images has been the first problem that I had to tackle during this Ph.D. work. The absence of light field raw data representing human faces pushed me to collaborate on systematic image collection, which resulted in the publication of one of the first face databases acquired with Lytro Illum camera. Because of its variations, the dataset can be used to investigate several topics, among which face expressions, occlusions removing and landmark detection. Thanks to this experience, I could contribute to other two multimodal data collections held in the contest of an European project PROTECT, in which I was responsible of 3D face data.

As starting point, a comparison between light field and structured light data has been carried out. Taking advantage of the possibility to render both kinds of data as RGB-D pair, well known face recognition algorithms are tested. The results are analysed in parallel and they have shown, for each sensor how depth map characteristics can differ, how image resolution can be changed without decreasing performances and how light field images are more suitable than structured light data in case of three-dimensional

occlusions but suffer from light variations.

Before proposing new algorithms, I studied the impact of out-of-focus created by post processing operations on light field images on gender recognition and age estimation from face. From this analysis, a strong similarity between Gaussian blur and focusing depths on estimation of soft biometric traits from face images has been found.

Once light field data have been studied and compared to images obtained with other RGB-D technologies, I was able to successfully propose two algorithms for plenoptic data, one to deal with face recognition problem and one with presentation attack detection. In the first work, I proved that the presented method outperforms the recognition over 2D standard images. In the second work, the proposed algorithm is compared with the state of the art presentation attack methods customized for light field data.

In conclusion of my PhD period, I have been coordinating the writing of a survey on the application of light field sensor on face analysis. The work includes an overview of existent methods for landmark detection, face recognition and presentation attack detection based on plenoptic images.

8.1 Limitations and future directions

Like all technologies, light field cameras are not perfect. In particular, the camera used during this work has been created for an amateur uses. For this reason, the access to software is limited: user's community has developed a Python library to interact with the internal software but it is unstable and badly documented. The closure of Lytro company in 2018 interrupted all the developing projects. The analysis performed in this thesis has proven the potentiality of light field technology on face analysis. The use of devices developed for biometric purposes could improve further the results.

This thesis is finalised to study the impact of light field data on face analysis and it is not focused on raw data elaboration. Thus, the algorithms used for rendering the images are not investigated. As mentioned in Section 3.3.4, depth maps can be estimated with several methods. A deep analysis on the optimal algorithm to exploit face shape would be of benefit.

8.1.1 Short term perspective

This thesis can play the role of baseline for several studies:

- Chapter 3: A new data collection could be carried out with varying the distance

subject-camera. In fact, in Figure 4.7, the histogram of depth map values is shown. The concentration of the distribution suggests that image acquisition technique has still to be improved. Decreasing the distance between subject and camera may help in spreading the depth values over a wider range.

- Chapter 4: Recent studies carried out by Politecnico of Torino proved that, among several 3D sensors, the most suitable for face analysis is structured light. Thus, in this thesis, light field device has been compared only with structured light camera. A possible work improvement could be a study including other kinds of RGB-D sensors.

- Chapter 5: The work described in this chapter has been done before the collection of IST-EURECOM Light Field Database [6]. All the analysis described may be repeated on the new database, taking advantage of raw data availability: focusing depths could be systematically evaluated and considered in the study.

- Chapter 6: The studies conducted on [4, 166] could inspire a different approach for the choice of views.

- Chapter 7: Also in this work the different distances between subject (or object) and camera could play a key role in recognizing a presentation attack. However, since the analysis conducted in Section 4.3 has shown a particular light sensibility in light field cameras, the impact of screen illumination in presentation attacks could be an interesting topic.

8.1.2 Long term perspective

In addition to the analyses presented in this manuscript, I have investigated other topics among which gender recognition and age estimation starting from sub-aperture representation of face plenoptic data. All the algorithms used have been revealed quite stable to the small variations present in different views of the same data and the obtained results have not outperformed the baseline. In fact, gender and age do not required an high level of details to be analysed and they are not deeply impacted by the point of view change. However, sub-aperture representation can be useful to study other facial characteristics such as human micro-expressions: a plenoptic video could record from different point of view micro facial variations that may be difficult to detect with a conventional camera.

Lytro Illum camera is designed for amateur use and it exploits only part of light field technology potential. As described in Chapter 3, the parameter λ and t of the plenoptic function have been kept as constant. The variation of these parameters in order to obtain

videos or images in different spectra could be of benefit for face analysis. Recently, Raytrix GmbH has released plenoptic devices sensible to near infra-red light. The integration and the fusion of light field information from different spectra recorded at different moments may be an interesting topic to investigate.

Chapter 9

Publications

- Chiesa, Valeria; Dugelay, Jean-Luc, *A comparison between Kinect and Lytro in RGB-D face recognition*, Submitted to Special Issue on CyberSecurity & Biometrics for a better Cyberworld, Future Generation Computer Systems journal
- Chiesa, Valeria; Galdi, Chiara; Busch, Christoph; Correia, Paulo Lobato; Dugelay, Jean-Luc; Guillemot, Christine, *Light fields for Face Analysis*, Submitted to Special Issue on Sensor Applications on Face Analysis, Sensor journal
- Chiesa, Valeria; Dugelay, Jean-Luc, *Kinect vs lytro in RGB-D face recognition*, Cyberworlds 2018, International Conference, 3-5 October 2018, Singapore
- Chiesa, Valeria; Dugelay, Jean-Luc, *Advanced face presentation attack detection on light field images*, BIOSIG 2018, 17th International Conference of the Biometrics Special Interest Group, 26-29 September 2018, Darmstadt, Germany
- Sequeira, Ana F.; Chen, Lulu; Ferryman, James; Galdi, Chiara; Chiesa, Valeria; Dugelay, Jean-Luc; Maik, Patryk; Gmitrowicz, Piotr; Szklarski, Lukasz; Prommegger, Bernhard; Kauba, Christof; Kirchgasser, Simon; Uhl, Andreas; Grudzien, Artur; Kowalski, Marcin, *PROTECT Multimodal DB: a multimodal biometrics dataset envisaging border control*, BIOSIG 2018, 17th International Conference of the Biometrics Special Interest Group, September 26-29, 2018, Darmstadt, Germany
- Chiesa, Valeria; Dugelay, Jean-Luc, *On multi-view face recognition using lytro images*, EUSIPCO 2018, 26th European Signal Processing Conference, 3-7 September 2018, Rome, Italy
- Sepas-Moghaddam, Alireza; Chiesa, Valeria; Lobato Correia, Paulo; Pereira, Fernando; Dugelay, Jean-Luc, *The IST-EURECOM light field face database*, IWBF 2017, 5th International Workshop on Biometrics and Forensics, 4-5 April 2017, Coventry, UK

- Chiesa, Valeria; Dugelay, Jean-Luc, *Impact of multi-focused images on recognition of soft biometric traits*, Optics and Photonics 2016, SPIE Optical Engineering + Applications, 28 August-1 September 2016, San Diego, USA

Bibliography

[1] M. Levoy and P. Hanrahan, "Light field rendering," in *23rd Annual Conference on Computer Graphics and Interactive Techniques*, (New York, NY, USA), pp. 31–42, ACM, 1996.

[2] R. Raghavendra, B. Yang, K. B. Raja, and C. Busch, "A new perspective — face recognition with light-field camera," in *International Conference on Biometrics (ICB)*, pp. 1–8, June 2013.

[3] A. Sepas-Moghaddam, P. Correia, K. Nasrollahi, T. Moeslund, and F. Pereira, "Light field based face recognition via a fused deep representation," pp. 1–6, September 2018.

[4] A. Sepas-Moghaddam, M. A. Haque, P. L. Correia, K. Nasrollahi, T. B. Moeslund, and F. Pereira, "A double-deep spatio-angular learning framework for light field based face recognition," *CoRR*, vol. abs/1805.10078, 2018.

[5] A. Sepas-Moghaddam, L. Malhadas, P. Correia, and F. Pereira, "Face spoofing detection using a light field imaging framework," *IET Biometrics*, vol. 7, no. 1, pp. 39–48, 2018.

[6] A. Sepas-Moghaddam, V. Chiesa, P. L. Correia, F. Pereira, and J. Dugelay, "The ist-eurecom light field face database," in *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6, April 2017.

[7] A. F. Sequeira, L. Chen, J. Ferryman, C. Galdi, V. Chiesa, J.-L. Dugelay, P. Maik, P. Gmitrowicz, L. Szklarski, B. Prommegger, C. Kauba, S. Kirchgasser, A. Uhl, A. Grudzien, and M. Kowalski, "Protect multimodal db: a multimodal biometrics dataset envisaging border control," in *17th International Conference of the Biometrics Special Interest Group, BIOSIG*, (Darmstadt, Germany), September 2018.

[8] V. Chiesa and J.-L. Dugelay, "Kinect vs lytro in rgb-d face recognition," in *Cyberworlds 2018, International Conference*, (Singapore), October 2018.

[9] V. Chiesa and J.-L. Dugelay, "Rgb-d face recognition: a comparative study of kinect and lytro," *Future Generation Computer Systems*, Submitted.

[10] V. Chiesa and J.-L. Dugelay, "Impact of multi-focused images on recognition of soft biometric traits," in *Optical Engineering + Applications (SPIE)*, (San Diego, USA), August 2016.

[11] V. Chiesa and J.-L. Dugelay, "On multi-view face recognition using lytro images," in *EUSIPCO 2018, 26th European Signal Processing Conference, 3-7 September 2018, Rome, Italy*, (Rome, Italy), September 2018.

[12] V. Chiesa and J.-L. Dugelay, "Advanced face presentation attack detection on light field images," in *17th International Conference of the Biometrics Special Interest Group, BIOSIG*, (Darmstadt, Germany), September 2018.

[13] C. Galdi, V. Chiesa, C. Busch, P. L. Correia, J.-L. Dugelay, and C. Guillemot, "Light field for face analysis," *Special Issue on Sensor Applications on Face Analysis, Sensor*, Submitted.

[14] I. R. M. A. IRMA, *Computer vision: Concepts, methodologies, tools, and applications*. 01 2018.

[15] G. Minelli, *All'origine della biologia moderna. La vita di un testimone e protagonista: Marcello Malpighi nell'Università di Bologna*. 1987.

[16] N. Grew, *The description and use of the pores in the skin of the hands and feet*. 1684.

[17] J. C. A. Mayer, *Anatomische Kupfertafeln nebst dazu gehörigen Erklärungen*. 1783-1788.

[18] J. Berry and D. Stoney, "The history and development of fingerprinting," pp. 1–40, 06 2001.

[19] M. Twain, *Life on the Mississippi*. 1883.

[20] L. Flom and A. Safir, *Iris recognition system*. 1987.

[21] W. W. Bledsoe, *The model method in facial recognition*. 1964.

[22] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. 1970.

[23] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am. A*, vol. 4, pp. 519–524, March 1987.

[24] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–591, June 1991.

[25] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090–1104, October 2000.

[26] R. Bolle and S. Pankanti, *Biometrics, Personal Identification in Networked Society: Personal Identification in Networked Society*. Kluwer Academic Publishers, 1998.

[27] J. Sergent, S. Ohta, and B. Macdonald, "Functional neuroanatomy of face and object processing – a positron emission tomography study," *Brain : a journal of neurology*, vol. 115 Pt 1, pp. 15–36, March 1992.

[28] N. G. Kanwisher, J. Mcdermott, and M. M. Chun, "The fusiform face area: A module in human extrastriate cortex specialized for face perception," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 17, pp. 4302–11, July 1997.

[29] T. Grüter, M. Grüter, and C.-C. Carbon, "Neural and genetic foundations of face recognition and prosopagnosia," *Journal of neuropsychology*, vol. 2, pp. 79–97, April 2008.

[30] X. Huang and O. Cossairt, "Dictionary learning based color demosaicing for plenoptic cameras," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 455–460, June 2014.

[31] S. Wanner, S. Meister, and B. Goldlücke, "Datasets and benchmarks for densely sampled 4d light fields," in *Vision, Modeling and Visualization*, pp. 225–226, 2013.

[32] W. Gordon, *Light field archive*, 2016. MIT, [Online]. Available: http://web.media.mit.edu/ gordonw/SyntheticLightFields/index.php.

[33] M. Grgic and C. Delac, *Face recognition homepage*, 2016. [Online]. Available: http://www.face-rec.org/databases/.

[34] R. Min, N. Kose, and J. Dugelay, "Kinectfacedb: A kinect database for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, pp. 1534–1548, November 2014.

[35] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 643–660, June 2001.

[36] B. Weyrauch, B. Heisele, J. Huang, and V. Blanz, "Component-based face recognition with 3d morphable models," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 85–85, June 2004.

[37] C. E. Thomaz and G. A. Giraldi, "A new ranking method for principal components analysis and its application to face image analysis," *Image and Vision Computing*, vol. 28, no. 6, pp. 902 – 913, 2010.

[38] C. Conde, A. Serrano, and E. Cabello, "Multimodal 2d, 2.5d amp; 3d face verification," in *2006 International Conference on Image Processing*, pp. 2061–2064, October 2006.

[39] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, e. B. Akarun, Lale, N. C. Juul, A. Drygajlo, and M. Tistarelli, "Bosphorus database for 3d face analysis," in *Biometrics and Identity Management*, (Berlin, Heidelberg), pp. 47–56, Springer Berlin Heidelberg, 2008.

[40] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–8, September 2008.

[41] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernocký, N. Poh, J. Kittler, A. Larcher, C. Lévy, D. Matrouf, J. Bonastre, P. Tresadern, and T. Cootes, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *IEEE International Conference on Multimedia and Expo Workshops*, pp. 635–640, July 2012.

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

[43] M. Grgic, K. Delac, and S. Grgic, "Scface – surveillance cameras face database," *Multimedia Tools and Applications*, vol. 51, pp. 863–879, February 2011.

[44] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692 – 706, 2014.

[45] R. Raghavendra, K. B. Raja, and C. Busch, "Exploring the usefulness of light field cameras for biometrics: An empirical study on face and iris recognition," *IEEE Transactions on Information Forensics and Security*, vol. 11, pp. 922–936, May 2016.

[46] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," tech. rep., CMU-CS-16-118, CMU School of Computer Science, 2016.

[47] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," pp. 815–823, June 2015.

[48] P. Viola and M. Jones, "Robust real-time face detection.," vol. 57, p. 747, 01 2001.

[49] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 2037–2041, Dec 2006.

[50] M. Haindl and M. Krupička, "Unsupervised detection of non-iris occlusions," *Pattern Recognition Letters*, vol. 57, pp. 60 – 65, 2015. Mobile Iris CHallenge Evaluation part I (MICHE I).

[51] C. Galdi and J. Dugelay, "Fusing iris colour and texture information for fast iris recognition on mobile devices," in *23rd International Conference on Pattern Recognition (ICPR)*, pp. 160–164, December 2016.

[52] N. Miura, A. Nagasaka, and T. Miyatake, "Extraction of finger-vein patterns using maximum curvature points in image profiles," *IEICE Transactions*, vol. 90-D, pp. 1185–1194, 2005.

[53] C. Kauba, J. Reissig, and A. Uhl, "Pre-processing cascades and fusion in finger vein recognition," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–6, September 2014.

[54] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 539–546 vol. 1, June 2005.

[55] ISO, "Iso/cei 19795-1:2006: Information technology - biometric performance testing and reporting - part 1: Principles and framework," iso, International Organization for Standardization, Geneva, Switzerland, 2006.

[56] A. A. Gershun, "The light field," *Unify Technical Press, translated by P. Moon and G. Timoshenko in Journal of Mathematics and Physics in 1939*, 1936.

[57] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, a. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, pp. 765–776, July 2005.

[58] J. Unger, S. Gustavson, P. Larsson, and A. Ynnerman, "Free Form Incident Light Fields," *Computer Graphics Forum*, 2008.

[59] R. Ng, *Digital Light Field Photography*. PhD thesis, Stanford, CA, USA, 2006. AAI3219345.

[60] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, vol. 1, pp. 7–55, Mar 1987.

[61] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1547–1555, 2015.

[62] C.-T. Huang, "Empirical bayesian light-field stereo matching by robust pseudo random field modeling," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 1–1, February 2018.

[63] X. Jiang, M. Le Pendu, and C. Guillemot, "Depth estimation with occlusion handling from a sparse set of light field views," in *IEEE Int. Conf. on Image Processing (ICIP)*, pp. 634–638, 2018.

[64] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 606–619, August 2013.

[65] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Journal of Computer Vision and Image Understanding*, vol. 145, pp. 148–159, April 2016.

[66] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a handheld plenoptic camera," *Technical Report CTSR 2005-02, Stanford University*, 2005.

[67] R. Rerabek and T. Ebrahimi, "New light field image dataset," in *QoMEX, Lisbon, Portugal*, June 2016.

[68] A. Mousnier, E. Vural, and C. Guillemot, "Partial light field tomographic reconstruction from a fixed-camera focal stack," *CoRR*, vol. abs/1503.01903, 2015.

[69] T. Wang, J. Zhu, H. Ebi, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4d light-field dataset and CNN architectures for material recognition," *CoRR*, vol. abs/1608.06985, 2016.

[70] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1605–1616, August 2017.

[71] V. V. et al., *The (new) Stanford light field archive*, 2016. Computer Graphics Laboratory, Stanford University, 2008. [Online]. Available: http://lightfield.stanford.edu/index.html.

[72] R. Raghavendra, B. Yang, K. B. Raja, and C. Busch, "A new perspective — face recognition with light-field camera," in *2013 International Conference on Biometrics (ICB)*, pp. 1–8, June 2013.

[73] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch, "Improved face recognition at a distance using light field camera and super resolution schemes," in *SIN*, 2013.

[74] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch, "Comparative evaluation of super-resolution techniques for multi-face recognition using light-field camera," in *18th International Conference on Digital Signal Processing (DSP)*, pp. 1–6, July 2013.

[75] D. Dansereau, *Light Field Toolbox V. 0.4*, 2016. MATLAB, [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/49683-lightfield-toolbox-v0-4.

[76] W. Ben Soltana, D. Huang, M. Ardabilian, L. Chen, and C. Ben-Amar, "Comparison of 2d/3d features and their adaptive score level fusion for 3d face recognition," in *3D Data Processing, Visualization and Transmission (3DPVT)*, (Paris, France), pp. 1–8, May 2010.

[77] L. Ulrich, E. Venzetti, S. Moos, and F. Marcolin, "3d face analysis: identification of the most suitable sensor technology for supporting different facial usage scenarios," 2019.

[78] J. Geng, "Structured-light 3d surface imaging: a tutorial," *Advances in Optics and Photonics*, vol. 3, pp. 128–160, June 2011.

[79] Freedman, "Primesense patent application us 2010/0290698," 2010.

[80] R. Min, N. Kose, and J. Dugelay, "Kinectfacedb: a kinect database for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, July 2014.

[81] P. Chhokra, A. Chowdhury, G. Goswami, M. Vatsa, and R. Singh, "Unconstrained kinect video face database," *Information Fusion*, vol. 44, pp. 113 – 125, 2018.

[82] R. I. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund, and G. Tranchet, "An rgb-d database using microsoft's kinect for windows for face detection," in *8th International Conference on Signal Image Technology and Internet Based Systems*, pp. 42–46, November 2012.

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

[84] D. Huang, Y. Wang, and J. Sun, "Lock3dface: A large-scale database of low-cost kinect 3d faces," in *International Conference on Biometrics (ICB)*, pp. 1–8, June 2016.

[85] H. Zhang, H. Han, J. Cui, S. Shan, and X. Chen, "Rgb-d face recognition via deep complementary and common feature learning," in *13th IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pp. 8–15, May 2018.

[86] G. Goswami, M. Vatsa, and R. Singh, "Rgb-d face recognition with texture and attribute features," *IEEE Transactions on Information Forensics and Security*, vol. 9, pp. 1629–1640, October 2014.

[87] A. Chowdhury, S. Ghosh, R. Singh, and M. Vatsa, "Rgb-d face recognition via learning-based reconstruction," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–7, September 2016.

[88] X. Xu, W. Li, and D. Xu, "Distance metric learning using privileged information for face verification and person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, pp. 3150–3162, December 2015.

[89] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[90] M. A. Anjum and M. Y. Javed, "Face recognition vs image resolution," in *International Conference on Information and Communication Technologies*, pp. 109–112, August 2005.

[91] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, pp. 71–86, January 1991.

[92] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition," in *10th IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, pp. 786–791 Vol. 1, October 2005.

[93] Y. Huang, Y. Wand, and T. Tan, "Combining statistics of geometrical and correlative features for 3d face recognition," in *British Machine Vision Conference (BMVA)*, pp. 90.1–90.10, 2006.

[94] J. B. C. Neto and A. N. Marana, "Face recognition using 3dlbp method applied to depth maps obtained from kinect sensors," in *Workshop de Visao Computacional (WVC)*, 2014.

[95] D. Moctezuma, C. Conde, I. Diego, and E. Cabello, "Soft-biometrics evaluation for people re-identification in uncontrolled multi-camera environments," *EURASIP Journal on Image and Video Processing*, vol. 2015, p. 28, August 2015.

[96] M. Demirkus, K. Garg, and S. Guler, "Automated person categorization for video surveillance using soft biometrics," *Proceedings of SPIE*, April 2010.

[97] G. Antipov, S.-A. Berrani, N. Ruchaud, and J.-L. Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," in *23rd ACM Multimedia Conference*, (Brisbane, Australia), October 2015.

[98] T. Huynh, R. Min, and J.-L. Dugelay, "An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data," in *Computer Vision* (J.-I. Park and J. Kim, eds.), (Berlin, Heidelberg), pp. 133–145, Springer Berlin Heidelberg, 2013.

[99] E. Boutellaa, A. Hadid, M. Bengherabi, and S. Ait-Aoudia, "On the use of kinect depth data for identity, gender and ethnicity classification from facial images," *Pattern Recognition Letters*, vol. 68, pp. 270 – 277, 2015. Special Issue on "Soft Biometrics".

[100] F. Hua, P. Johnson, N. Sazonova, P. Lopez-Meyer, and S. Schuckers, "Impact of out-of-focus blur on face recognition performance based on modular transfer function," in *5th IAPR International Conference on Biometrics (ICB)*, pp. 85–90, March 2012.

[101] P. A. Johnson, P. Lopez-Meyer, N. Sazonova, F. Hua, and S. Schuckers, "Quality in face and iris research ensemble (q-fire)," in *4th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–6, Setptember 2010.

[102] Y. Niu, Y. Zhong, W. Guo, Y. Shi, and P. Chen, "2d and 3d image quality assessment: A survey of metrics and challenges," *IEEE Access*, vol. 7, pp. 782–801, 2019.

[103] N. D. Kalka, J. Zuo, N. A. Schmid, and B. Cukic, "Image quality assessment for iris biometric," 2006.

[104] H. J. Galiyawala and R. Chaudhari, "Hand geometry- and palmprint-based biometric system with image deblurring," in *Information and Communication Technology for Competitive Strategies* (S. Fong, S. Akashe, and P. N. Mahalle, eds.), (Singapore), pp. 591–604, Springer Singapore, 2019.

[105] Z. Shen, T. Xu, J. Zhang, J. Guo, and S. Jiang, "A multi-task approach to face deblurring," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, p. 23, January 2019.

[106] I. Ullah, M. Hussain, G. Muhammad, H. Aboalsamh, G. Bebis, and A. M. Mirza, "Gender recognition from face images with local wld descriptor," in *19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 417–420, April 2012.

[107] C. Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 431 – 437, 2012. Intelligent Multimedia Interactivity.

[108] J. E. Tapia and C. A. Perez, "Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape," *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 488–499, March 2013.

[109] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela, "Robust gender recognition by exploiting facial attributes dependencies," *Pattern Recognition Letters*, vol. 36, pp. 228 – 234, 2014.

[110] S. Jia and N. Cristianini, "Learning to classify gender from four million images," *Pattern Recognition Letters*, vol. 58, pp. 35 – 41, 2015.

[111] G. Antipov, S.-A. Berrani, and J.-L. Dugelay, "Minimalistic cnn-based ensemble model for gender prediction from face images," *Pattern Recognition Letters*, vol. 70, pp. 59 – 65, 2016.

[112] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[113] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. abs/1411.7923, 2014.

[114] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[115] G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," Tech. Rep. UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[116] S. Escalera, M. T. Torres, B. Martínez, X. Baró, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, M. Oliu, M. A. Bagheri, and M. Valstar, "Chalearn looking at people and faces of the world: Face analysisworkshop and challenge 2016," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 706–713, June 2016.

[117] R. Rothe, R. Timofte, and L. Gool, "Dex: Deep expectation of apparent age from a single image," *IEEE International Conference on Computer Vision Workshops (ICCV)*, pp. 10–15, January 2015.

[118] H.-F. Yang, B.-Y. Lin, K.-Y. Chang, and C.-S. Chen, "Automatic age estimation from face images via deep ranking," in *Proceedings of the British Machine Vision Conference (BMVC)* (X. Xie, M. W. Jones, and G. K. L. Tam, eds.), pp. 55.1–55.11, BMVA Press, September 2015.

[119] K. Zibrek, L. Hoyet, K. Ruhland, and R. McDonnell, "Evaluating the effect of emotion on gender recognition in virtual humans," August 2013.

[120] K. Ueki, M. Sugiyama, and Y. Ihara, "Perceived age estimation under lighting condition change by covariate shift adaptation," in *20th International Conference on Pattern Recognition (ICPR)*, pp. 3400–3403, August 2010.

[121] T. D. Nguyen, S. R. Cho, T. D. Pham, and K. R. Park, "Human age estimation method robust to camera sensor and/or face movement," in *Sensors*, 2015.

[122] N. Ruchaud, G. Antipov, P. Korshunov, J.-L. Dugelay, T. Ebrahimi, and S.-A. Berrani, "The impact of privacy protection filters on gender recognition," in *Optical Engineering + Applications, Applications of Digital Image Processing XXXVIII (SPIE)*, (San Diego, USA), August 2015.

[123] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Apparent age estimation from face images combining general and children-specialized deep learning models," in *29th IEEE Conference on Computer Vision and Pattern Recognition Workshops, (CVPRW)*, (Las Vegas, USA), June 2016.

[124] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *Proceedings of British Machine Vision Conference*, 2015.

[125] R. Rothe, R. Timofte, and L. V. Gool, "Dex: Deep expectation of apparent age from a single image," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[126] R. Gross, I. Matthews, and S. Baker, "Eigen light-fields and face recognition across pose," in *5th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 3–9, May 2002.

[127] R. Gross, I. Matthews, and S. Baker, "Fisher light-fields for face recognition across pose and illumination," in *German Symposium on Pattern Recognition (DAGM)*, September 2002.

[128] R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light-fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 449–465, April 2004.

[129] R. Gross, I. Matthews, J. Cohn, and T. a. Kanade, "Multi-pie," in *IEEE International Conference on Automatic Face and Gesture Recognition*, IEEE Computer Society, September 2008.

[130] M. E. Wibowo and D. Tjondronegoro, "Face recognition across pose on video using eigen light-fields," in *International Conference on Digital Image Computing: Techniques and Applications*, pp. 536–541, December 2011.

[131] S. Zhou and R. Chellappa, "Illuminating light field: image-based face recognition across illuminations and poses," in *6th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 229–234, May 2004.

[132] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch, "Guclf: a new light field face database," *Proc SPIE*, 11 2013.

[133] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America*, vol. 4, pp. 2379–2394, December 1987.

[134] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Transactions on Neural Networks*, vol. 14, pp. 117–126, January 2003.

[135] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 210–227, February 2009.

[136] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical Models and Image Processing*, vol. 53, no. 3, pp. 231 – 239, 1991.

[137] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *Journal of the Optical Society of America. A, Optics and image science*, vol. 6, pp. 1715–26, December 1989.

[138] R. W. Gerchberg, "Super-resolution through error energy reduction," *Journal of Modern Optics - J MOD OPTIC*, vol. 21, pp. 709–720, September 1974.

[139] A. Papoulis, "A new algorithm in spectral analysis and band-limited extrapolation," *IEEE Transactions on Circuits and Systems*, vol. 22, pp. 735–742, September 1975.

[140] A. Zomet, A. Rav-Acha, and S. Peleg, "Robust super-resolution," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–I, December 2001.

[141] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch, "A novel image fusion scheme for robust multiple face recognition with light-field camera," pp. 722–729, January 2013.

[142] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch, "Comparative evaluation of super-resolution techniques for multi-face recognition using light-field camera," in *18th International Conference on Digital Signal Processing (DSP)*, pp. 1–6, July 2013.

[143] K. B. Raja, R. Raghavendra, F. Alaya Cheikh, and C. Busch, "Evaluation of fusion approaches for face recognition using light field cameras," July 2015.

[144] M. Y. Shams, A. S. Tolba, and S. H. Sarhan, "A vision system for multi-view face recognition," *CoRR*, vol. abs/1706.00510, 2017.

[145] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: A deep model for learning face identity and view representations," vol. 1, pp. 217–225, January 2014.

[146] D. Kim, B. Comandur, H. Medeiros, N. M. Elfiky, and A. C. Kak, "Multi-view face recognition from single rgbd models of the faces," *Computer Vision and Image Understanding*, vol. 160, pp. 114 – 132, 2017.

[147] K. Niinuma, H. Han, and A. K. Jain, "Automatic multi-view face recognition via 3d model based pose regularization," September 2013.

[148] A. Sepas-Moghaddam, P. L. Correia, and F. Pereira, "Light field local binary patterns description for face recognition," in *IEEE International Conference on Image Processing (ICIP)*, pp. 3815–3819, September 2017.

[149] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, September 2014.

[150] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, December 1997.

[151] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Computing Research Repository, CoRR*, vol. abs/1409.4842, 2014.

[152] S. Setiowati, Zulfanahri, E. L. Franita, and I. Ardiyanto, "A review of optimization method in face recognition: Comparison deep learning and non-deep learning methods," in *9th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 1–6, October 2017.

[153] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 30107-1. Information Technology - Biometric presentation attack detection - Part 1: Framework*. International Organization for Standardization, 2016.

[154] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC FDIS 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting*. International Organization for Standardization, 2017.

[155] S. Kim, Y. Ban, and S. Lee, "Face liveness detection using a light field camera," *Sensors (Basel, Switzerland)*, vol. 14, pp. 22471–22499, November 2014.

[156] R. Raghavendra, K. Raja, and C. Busch, "Presentation attack detection for face recognition using light field camera," *IEEE Transaction on Image Processing*, vol. 24, no. 3, pp. 1060–1074, 2015.

[157] Z. Ji, H. Zhu, and Q. Wang, "Lfhog: A discriminative descriptor for live face detection from light field image," in *IEEE International Conference on Image Processing (ICIP)*, pp. 1474–1478, September 2016.

[158] M. Liu, H. Fo, Y. Wei, Y. Rehman, L. Po, and W. Lo, "Light field-based face liveness detection with convolutional neural networks," *SPIE, Electronic Imaging*, vol. 28, April 2019.

[159] A. Sepas-Moghaddam, F. Pereira, and P. Correia, "Light field based face presentation attack detection: Reviewing, benchmarking and one step further," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 7, pp. 1696–1709, 2018.

[160] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886–893 vol. 1, June 2005.

[161] S. Bhattacharjee, A. Mohammadi, and S. Marcel, "Spoofing deep face recognition with custom silicone masks," in *9th International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–8, IEEE Computer Society, October 2019.

[162] A. Sepas-Moghaddam, L. Malhadas, P. L. Correia, and F. Pereira, "Face spoofing detection using a light field imaging framework," *IET Biometrics*, vol. 7, no. 1, pp. 39–48, 2018.

[163] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, June 2014.

[164] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: database and results," vol. 47, January 2016.

[165] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[166] S. S. Farfade, M. J. Saberian, and L. Li, "Multi-view face detection using deep convolutional neural networks," *CoRR*, vol. abs/1502.02766, 2015.

Résumé en français

9.1 Introduction

Dans ce travail de doctorat, l'impact de la technologie plenoptique sur l'analyse du visage a été étudié. Les contributions couvrent plusieurs aspects de la recherche en l'analyse des visages.

- **Acquisition de données:** Trois collectes de données "light field" de visage ont été réalisées au cours de la période couverte par cette thèse. La premiere, réalisée en collaboration avec l'Instituto de Telecomunicações, Instituto Superior Técnico, Universitade de Lisboa, a été présentée dans [6]. La base de données, désormais accessible au public, consiste en une collection d'images représentant 100 individus acquises lors de deux sessions, chacune comportant 20 variations de visages (occultations, émotions, poses, illuminations).
  Deux bases de données multimodales ont été collectées dans le cadre du projet PROTECT. Les données relatives au visage 3D ont été acquises et traitées afin de fusionner les résultats avec d'autres biométries. Les résultats de la première collecte de données sont présentés dans [7].

- **Comparaison et modélisation**: Le potentiel du "light field" en reconnaissance faciale est comparé aux évaluations faites sur des images lumière structurée afin de créer, par des expériences exhaustives, une base de référence et de suggérer la meilleure technologie en fonction du scénario à traiter avec [8], [9]. L'impact des images plénoptiques sur les caractéristiques biométriques douces est étudié et modélisé [10].

- **Algorithme**: Deux algorithmes innovants, adaptés aux images plenoptiques, ont été proposés. Le premier, présenté dans [11], s'est attaqué au problème de la reconnaissance faciale en utilisant la propriété plénoptique des données à rendre sous forme d'images en "sub-aperture". La deuxième, décrite dans [12], montrait une méthode pour détecter les attaques de présentation de visage exploitant les

images RGB-D.

- **État de l'art**: A la fin de cette thèse de doctorat, la littérature sur l'utilisation de la technologie plenoptiques en l'analyse de visage a été développée. Ainsi, en collaboration avec l'Instituto de Telecomunicações, Instituto Superior Técnico, Universitade de Lisboa, Hochschule de Darmstadt et l'Institut de recherche en informatique et systèmes aléatoires (IRISA), une revue sur l'utilisation des données plénoptiques pour l'estimation des points clés du visage, la reconnaissance faciale et la détection des attaques de présentation a été élaborée [13].

9.2 Contribution

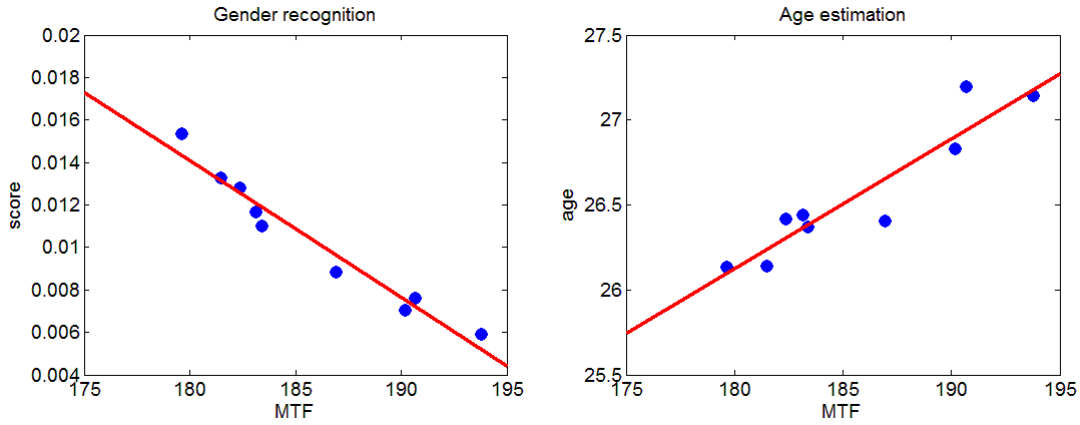
Afin de résumer les contributions de cette thèse, un bref aperçu de chaque article publié est fourni.

9.2.1 Impact des images multifocales en reconnaissance de caractéristiques biométriques douces

Dans le domaine de la vidéosurveillance, l'estimation des traits sémantiques comme le genre et l'âge a toujours été considéré comme compliqué en raison de l'environnement incontrôlé: si les variations de lumière ou de pose ont été largement étudiées, les images défocalisées l'ont été rarement. Récemment, l'émergence de nouvelles technologies, comme les caméras plénoptiques, permet de faire face à ces problèmes en analysant des images multi-focales. Grâce à un réseau de microlentilles disposées entre le capteur et l'objectif principal, les caméras plénoptiques sont en mesure d'enregistrer non seulement les valeurs RGB mais aussi les informations relatives à la direction des rayons lumineux: les données supplémentaires permettent de restituer l'image dans un plan focal différent par rapport à l'acquisition. Pour nos expériences, nous utilisons la base de données GUC Light Field Face Database qui comprend des images de la caméra Lytro de première génération. Tirant parti des images "light field", nous explorons l'influence de la défocalisation sur les problèmes de reconnaissance du genre et d'estimation de l'âge.

9.2.2 Base de donnée IST-EURECOM Light Field Face

Les caméras plenoptiques deviennent de plus en plus puissantes pour acquérir des représentations de scènes riches qui fournissent des images uniques pour l'analyse et la représentation. Certains travaux récents ont montré la puissance et l'utilité dé informations enrichies obtenues par l'imagerie "light field", notamment pour la reconnaissance faciale. Toutefois, il est encore difficile d'évaluer pleinement comment la technologie de



(a) (b)

Figure 9.1 – Exemple de variation selon le genre (Figure 9.1a) et l'âge (Figure 9.1b) sur un même jeu de données plenoptiques rendues à différentes profondeurs de focalisation

reconnaissance faciale peut bénéficier de ces nouveaux capteurs d'images, notamment en raison l'absence du données appropeiées. Pour soutenir la recherche sur la reconnaissance faciale à l'aide d'images plenoptiques, la base de données de visage IST-EURECOM avec donnée plenoptiques (IST-EURECOM LFFD) est présentée dans ce document. L'objectif est de rendre compte de la disponibilité publique d'une base de données de visage "light field" qui devrait servir à concevoir, tester et valider des systèmes de reconnaissance basés sur l'imagerie plénoptique. La base de données proposée comprend des données de 100 sujets, capturées par Lytro ILLUM en deux séances séparées de 1 à 6 mois, avec 20 échantillons pour chaque personne par session. Pour simuler plusieurs scénarios, les images sont acquises avec plusieurs variations faciales, couvrant toute une gamme d'émotions, d'actions, de poses, d'illuminations et d'occlusions. La base de données comprend les images plénoptiques brutes, le rendu 2D et les cartes de profondeur associées, ainsi qu'un ensemble asses complet d'images et de métadonnées. Le IST-EURECOM LFFD devrait devenir un ajout précieux aux référentiels de base de données de visage existants.

9.2.3 Reconnaissance faciale RGB-D : une étude comparative de Kinect et Lytro

Récemment, les caméras plénoptiques sont devenues plus abordables et de plus en plus populaires grâce à des capacités plus élevées par rapport aux caméras ordinaires pour capturer l'information d'une scène. Même si le principe associé aux capteurs de lumière structurés est très différent de celui des caméras "light field", les données fournies par ces technologies sont similaires. Dans le but de comparer le potentiel des capteurs Kinect et Lytro sur la reconnaissance faciale, une analyse préliminaire sur la base de données

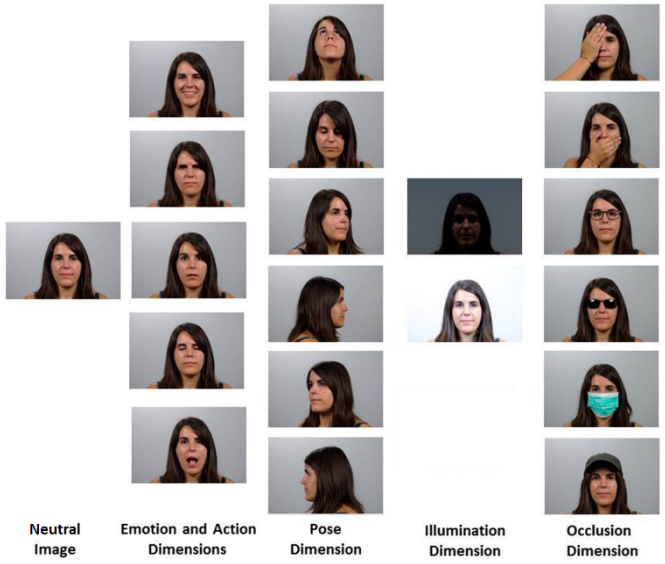


Figure 9.2 – Exemple de données de IST-EURECOM LFFD

IST-EURECOM Light Field Database et sur la base de données EURECOM Kinect Face est présentée. Les méthodes de base sont testées sur des images RGB et des cartes de profondeur des deux bases de données. En s'inspirant de travaux antérieurs, les améliorations générées par les informations de profondeur sont étudiées et comparées entre les deux technologies. Les résultats sont intégrés à une module de fusion au niveau du score entre les images RGB et les images de profondeur. Les analyses sont validées par des tests sur une mini-base de données collectées avec les deux technologies en même temps. Les améliorations dues à l'intégration des cartes de profondeur sont prouvées par nos résultats, bien que les cartes de profondeur plénoptiques soient plus sensibles aux variations de lumière par rapport aux capteurs de lumière structurée. En particulier, l'introduction d'informations de profondeur dans les données RGB s'avère plus efficace que l'imagerie bi-dimensionnelle standard en cas d'occultationes.

(a) (b)

Figure 9.3 – Exemple de carte de profondeur du capteur Kinect (Figure 9.3a) et Lytro (Figure 9.3b)

9.2.4 Sur la reconnaissance faciale multi-vues utilisant des images Lytro

Dans ce travail, une approche simple et efficace de reconnaissance des visages à partir d'images plénoptiques, notamment de la caméra Lytro Illum, est proposée. La méthode suggérée est basée sur la propriété des images "light field", restituées par une représentation multi-vues. Dans l'analyse préliminaire, les vecteurs de caractéristiques extraits de différentes vues d'une même image Lytro sont suffisamment différents pour fournir des informations complémentaires utiles à la reconnaissance faciale. A partir d'un ensemble de vues multiples pour chaque donnée, le problème de vérification faciale est abordé et les résultats sont comparés à ceux obtenus avec des images 2D classiques simulées en utilisant une seule vue, c'est-à-dire la vue centrale. Deux expériences sont décrites et, dans les deux cas, la méthode présentée montre des performances supérieures aux algorithmes standards adoptés par les capteurs d'imagerie classiques.

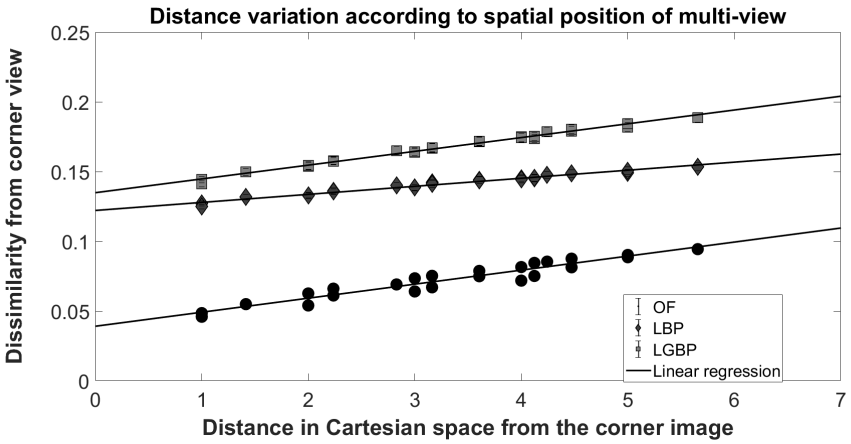


Figure 9.4 – Distance euclidienne normalisée de chaque vue par rapport à la vue d'angle dans la même image v.s. espace distance entre les vues considérées. Les lignes pleines montrant une relation linéaire entre les algorithmes de décalage de vue et de reconnaissance.

9.2.5 Détection avancée d'attaque de présentation de visage sur des images plénoptiques

Au cours des dernières années, plusieurs travaux se sont concentrés sur l'impact des nouveaux capteurs sur la reconnaissance faciale. Un intérêt particulier a été porté aux technologies capables de détecter la profondeur de la scène comme est le cas des caméras plenoptiques. Parallèlement aux algorithmes d'identification de personnes, de nouvelles méthodes "anti-spoofing" adaptées à des dispositifs spécifiques doivent être étudiées. Dans ce travail, un nouvel algorithme pour la détection d'attaque de présentation sur la base de données de visage "light field" est proposé. Bien que la distance entre le sujet et l'appareil

photo ne soit pas une information pertinente pour les attaques d'usurpation d'identité 2D standard, elle pourrait être importante lors de l'utilisation de caméras 3D. Nous prouvons à travers trois expériences que la méthode proposée basée sur l'élaboration de cartes de profondeur surpasse les algorithmes existants en matière de détection d'attaque de présentation sur des images plenoptiques.

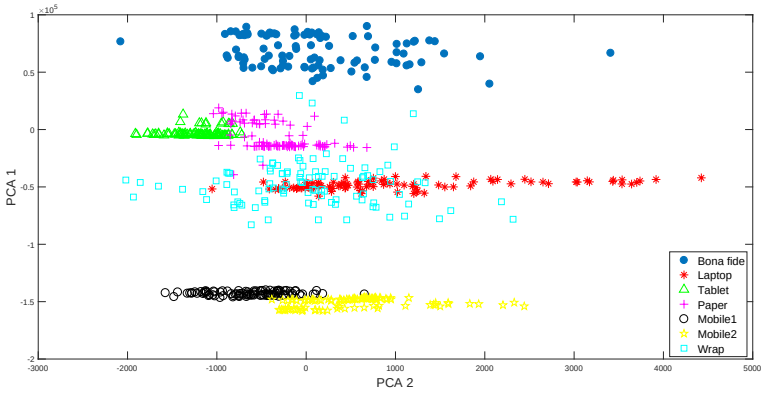


Figure 9.5 – Exemple de clustering pour defférentes attaques de présentation de visage

9.2.6 PROTECT Multimodal DB : un ensemble de données biométriques multimodales dans un contexte de contrôle aux frontières

Ce travail présente une nouvelle base de données multimodale comprenant le visage en 3D, le visage en 2D, le visage en thermique, l'iris visible, les veines des doigts et des mains, la voix et l'anthropométrie. Cet ensemble de données constituera une ressource précieuse pour la recherche grâca a son nombre et sa variété de traits biométriques. Acquis dans le cadre du projet EU PROTECT, l'ensemble des données permet plusieurs combinaisons de traits biométriques et d'empreintes digitales et envisage des applications telles que le contrôle aux frontières. D'après les résultats des données unimodales, une fusion a été appliquée pour déterminer le potentiel de reconnaissance de la combinaison de ces caractéristiques biométriques dans un système multimodale. En raison de la variabilité du pouvoir discriminatif des traits de caractère, il convient de ne pas tenir compte de l'importance de ces derniers, la technique de fusion n-meilleurs out a été appliquée pour obtenir différents résultats de reconnaissance.

9.3 Conclusion

Dans cette thèse, l'impact de la technologie plénoptique sur l'analyse du visage a été discuté. Une comparaison détaillée entre la caméra Kinect et la caméra Lytro permet

Figure 9.6 – Exemple d'acquisition de données de la base de données multimodale PROTECT

d'établir une base de référence pour les travaux futurs. Une base de données du visage "light field" a été mise à disposition à des fins de recherche. La défocalisation créée par le post-traitement des images plénoptique est étudiée et la différence entre "défocalisé" et "flou" due au filtre gaussien est analysée. L'impact des les deux techniques sur les traits biométriques doux comme le genre et l'âge est comparable.

L'amélioration de la reconnaissance faciale et de la détection d'attaque de présentation attribuable aux données plénoptiques est prouvée par des algorithmes novateurs adaptés aux données "light field".

Un travail sur l'analyse du visage avec des images plenoptiques a été compilé afin de rassembler en un seul document les recherches les plus récentes.

Toutes les réalisations atteintes ont été (ou vont être) publiées dans des actes de conférences internationales ou dans des articles de revues.