Universidade Estadual de Campinas
Instituto de Computação

Allan da Silva Pinto

Analysis of Intrinsic and Extrinsic Properties of Biometric Samples for Presentation Attack Detection

Análise de Propriedades Intrínsecas e Extrínsecas de Amostras Biométricas para Detecção de Ataques de Apresentação

CAMPINAS
2018

Allan da Silva Pinto

Analysis of Intrinsic and Extrinsic Properties of Biometric Samples for Presentation Attack Detection

Análise de Propriedades Intrínsecas e Extrínsecas de Amostras Biométricas para Detecção de Ataques de Apresentação

Tese apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Supervisor/Orientador: Prof. Dr. Anderson de Rezende Rocha
Co-supervisor/Coorientador: Prof. Dr. Hélio Pedrini

Este exemplar corresponde à versão final da Tese defendida por Allan da Silva Pinto e orientada pelo Prof. Dr. Anderson de Rezende Rocha.

CAMPINAS
2018

Agência(s) de fomento e nº(s) de processo(s): CAPES; CNPq, 140069/2016-0; CNPq, 142110/2017-5
ORCID: http://orcid.org/0000-0003-3765-8300

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

P658a

Pinto, Allan da Silva, 1984-
Analysis of intrinsic and extrinsic properties of biometric samples for presentation attack detection / Allan da Silva Pinto. – Campinas, SP : [s.n.], 2018.

Orientador: Anderson de Rezende Rocha.
Coorientador: Hélio Pedrini.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Biometria. 2. Reconhecimento de padrões. 3. Tecnologia da informação - Sistemas de segurança. 4. Ciência forense digital. I. Rocha, Anderson de Rezende, 1980-. II. Pedrini, Hélio, 1963-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Análise de propriedades intrínsecas e extrínsecas de amostras biométricas para detecção de ataques de apresentação
Palavras-chave em inglês:
Biometry
Pattern recognition
Information technology - Security measures
Digital forensic science
Área de concentração: Ciência da Computação
Titulação: Doutor em Ciência da Computação
Banca examinadora:
Anderson de Rezende Rocha [Orientador]
Aparecido Nilceu Marana
Siovani Cintra Felipussi
Sandra Eliza Fontes de Avila
José Mário De Martino
Data de defesa: 06-09-2018
Programa de Pós-Graduação: Ciência da Computação

Universidade Estadual de Campinas
Instituto de Computação

Allan da Silva Pinto

Analysis of Intrinsic and Extrinsic Properties of Biometric Samples for Presentation Attack Detection

Análise de Propriedades Intrínsecas e Extrínsecas de Amostras Biométricas para Detecção de Ataques de Apresentação

Banca Examinadora:

- Prof. Dr. Anderson de Rezende Rocha
  IC/Unicamp
- Prof. Dr. Aparecido Nilceu Marana
  FC/UNESP
- Prof. Dr. Siovani Cintra Felipussi
  CCGT/UFSCar
- Prof.a Dra. Sandra Eliza Fontes de Avila
  IC/Unicamp
- Prof. Dr. José Mário De Martino
  FEEC/Unicamp

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 06 de setembro de 2018

To my wife, Dinéia, who always provided me all support necessary for this long journey. Your words of encouragement and your immense comprehension of the importance of this work in our lives helped us to get to this moment. I love you! Thank you!

Acknowledgments

THERE are many reasons that may motivate a person for pursuing a Ph.D. degree. For some people, the doctoral degree is a valuable asset that increases one's chances of earning higher paid in more satisfying jobs. For other ones, the chance of getting a higher position in their professional career may empower a person for dedicating to this task. For some young people, the perspective of getting a more comfortable future or even their dream profession may be attenuated by their family's background and reality, and so, a Master's and a Ph.D. degrees may represent a valuable passport to change their living conditions through the education.

Although all those reasons may encourage a student to dedicate four or more years of his/her life to have a doctoral degree, I realized that the discovery of a world of research, science and new knowledge can perfectly act as a prime stimulus that makes a student keep moving toward its ultimate goal. In fact, the belief of how the research, science, and education can change a society and provide them with new perspectives of a brilliant future was the real motivation that inspired me to work hard during four years, fighting with myself to surpass several barriers of difficulties found in this journey and my limitations. Undoubtedly, that was the most beautiful finding that could achieve in this thesis, which will be in my heart along for my entire life. Moreover, the most beautiful part of this discovery is that I would have never reached these conclusions only by designing solutions, algorithms, and by writing codes. What really made me understand the valuable work of professors and scientists in our country was the conviviality and interaction with several people, which I had the great opportunity to meet along these years. Without the dedication and guidance of remarkable people that I met during this journey, I certainly would not have found the true meaning of our profession. Additionally, this thesis would not have been possible without the support and collaboration of several people who participated directly or indirectly in the principal moments of my life during these four years. Thus, I dedicate this humble tribute as a form of demonstrating my deep gratitude to all these people.

Firstly, I would like to thank my sweet wife *Dinéia* for being present in all moments of joy and difficulties that we have been through these last years. Thank you for always having words of support and comfort to me. Your wisdom and comprehension of the importance of the hours dedicated to this work was fundamental for me to achieve these results. You were the greatest gift that God could have given me. May God bless you for being an excellent wife, and my best friend. I love you.

I thank my parents, *Laudicéia* and *Nairton* for the dedication and care they always had with my education as a citizen, human being and a professional. It is clear in my memories the moments which you always looked for alternatives to give me a good education, even without enough financial resources. Without your care, I certainly would not have gotten here. May God bless you and give you much joy in your hearts, peace of mind and health. I love you.

I would like to express my gratitude to my dear sister *Vanessa*, my dear sister-in-law

Néria, and my parents-in-law *Lourival* and *Hosana*, who have always supported me with words of encouragement and gestures of affection. You are the best family that anyone could ever have had. I thank God for giving me the honor of living with such wonderful people as you.

I would like to thank my advisor, *Prof. Dr. Anderson de Rezende Rocha*, for the words of support, wisdom, and encouragement given throughout this journey. I thank you for your absolute dedication to always teach me, for your help in times of trouble, for correcting me and for showing me new paths and opportunities of our profession. Undoubtedly, all those gestures were crucial to getting us to this moment. Thank you for always believing in my work, and for showing the real meaning of being a professor. Your love for our profession is an inspiration to me. May God bless you and your family.

I would also like to thank my co-advisor, *Prof. Dr. Hélio Pedrini*, for his dedication as a professor and co-advisor. I am grateful to you for always being ready to help me and hear me whenever I needed. Your wisdom and sensitivity as a human being make you a distinct professional. Thank you for your valuable contributions to this thesis.

I thank *Prof. Dr. William Robson Schwartz*, *Prof. Dr. David Menotti Gomes*, *Prof. Dr. Alexandre Xavier Falcão*, *Prof. Dr. Siome Klein Goldenstein* and *Dr. Giovani Chiachia* for valuable contributions to this thesis and for sharing their wisdom with me during our conversations, meetings, and during the hours of working in the laboratories. Our discussions were valuable to my professional growth.

I would like to thank *Dr. Fernanda Alcântara Andaló*, *Prof. Dr. José Mário De Martino*, and *Prof. Dr. Ricardo da Silva Torres* for valuable contribution to this thesis during the qualification exam process. Your observations, recommendations, and critique were very important for increasing the quality of this work.

I thank *Prof. Dr. Kevin Boyer*, *Prof. Dr. Patrick Flynn*, *Prof. Dr. Walter Scheirer* and *Prof. Dr. Adam Czajka* for hosting me at the University of Notre Dame, USA, during my doctoral internship. I am grateful for all support necessary to accomplish my work and for the excellent hospitality and interaction during my stay at Notre Dame. Thank you for your contributions to this thesis.

I would like to thank my colleagues of the Reasoning for Complex Data (Recod) laboratory for their valuable friendship and discussions that we had along these years. I am grateful for having the opportunity of working with such brilliant minds. I would also like to thank my colleagues from the Computer Vision Research Laboratory (CVRL) from the University of Notre Dame. Your friendship and hospitality made me feel like at home. I thank *Andrey Kuehlkamp*, *Dr. Alexandre Ferreira*, *Aparna Bharati*, *Benedict Becker*, *Dr. Daniel Moreira*, *Joel Brogan*, *Michael Krumdick*, and *Dr. Tiago Carvalho* for their valuable collaborations which certainly made me a better professional. I learned a lot with you guys. A special thanks to *Daniel Moraes*, *Dr. Felipe Louza*, *Luis Augusto Martins*, *Ramon Pires*, and *Samuel Botter Martins*, for their friendship and for our nice conversations during coffee times. You have a friend in me.

I would like to thank the Institute of Computing at the University of Campinas, its faculty, and staff for the excellent work carried out in this university toward providing students with such magnificent professional education. A special thanks to everyone that works hard every day to make the University of Campinas a high-quality educational institution. I also thank this same institution and community for providing me with an education of excellence during my Master's studies, which were fundamental for me to reach the maturity necessary for pursuing a Ph.D. degree.

I thank all my former professors and advisors from the Institute of Mathematics and

Computer Sciences (ICMC) at the University of São Paulo, for all dedication and hard work toward providing me with an excellent professional education during college. I am grateful to all professors and people of that institution that contributed to my formation as professional and citizen. A special thanks to my advisors in the scientific initiation projects that I accomplished at the University of São Paulo, *Prof. Dr. Jorge Luiz e Silva*, and *Prof. Dr. Odemir Martinez Bruno*, for their great patience, for sharing their wisdom.

I thank all my family friends for their friendship and for all moments we had together. I thank you for their comprehension whenever I needed to dedicate to this thesis. A special thanks to Andréia, Clóvis & family, and *Rudnei & family*. I thank God for giving me the opportunity for meeting such nice people in my life. Also, a special thanks to *Anderson Nascimento & family*, for all the help given to my wife and me during our stay in the USA. For our sincere friendship to you and for the moments of joy that we had together. I also thank our friends from Chicago, USA, and from West Lafayette, USA, for all help, support and friendship. May God bless you all.

I am especially grateful to my grandparents, the late Delmira & the late Luís, for valuable advice and care they had with me during their lives.

Finally, I would like to thank the Brazilian federal government agencies, *National Council for Scientific and Technological Development* (CNPq) and *Coordination for the Improvement of Higher Education Personnel* (CAPES), for the financial support during this thesis. I really hope to take the opportunity to reciprocate all this support offered to me by working for the growth of our country.

Resumo

Os recentes avanços nas áreas de pesquisa em biometria, forense e segurança da informação trouxeram importantes melhorias na eficácia dos sistemas de reconhecimento biométricos. No entanto, um desafio ainda em aberto é a vulnerabilidade de tais sistemas contra ataques de apresentação, nos quais os usuários impostores criam amostras sintéticas, a partir das informações biométricas originais de um usuário legítimo, e as apresentam ao sensor de aquisição procurando se autenticar como um usuário válido. Dependendo da modalidade biométrica, os tipos de ataque variam de acordo com o tipo de material usado para construir as amostras sintéticas. Por exemplo, em biometria facial, uma tentativa de ataque é caracterizada quando um usuário impostor apresenta ao sensor de aquisição uma fotografia, um vídeo digital ou uma máscara 3D com as informações faciais de um usuário-alvo. Em sistemas de biometria baseados em íris, os ataques de apresentação podem ser realizados com fotografias impressas ou com lentes de contato contendo os padrões de íris de um usuário-alvo ou mesmo padrões de textura sintéticos. Nos sistemas biométricos de impressão digital, os usuários impostores podem enganar o sensor biométrico usando réplicas dos padrões de impressão digital construídas com materiais sintéticos, como látex, massa de modelar, silicone, entre outros. Esta pesquisa teve como objetivo o desenvolvimento de soluções para detecção de ataques de apresentação considerando os sistemas biométricos faciais, de íris e de impressão digital. As linhas de investigação apresentadas nesta tese incluem o desenvolvimento de representações baseadas nas informações espaciais, temporais e espectrais da assinatura de ruído; em propriedades intrínsecas das amostras biométricas (e.g., mapas de albedo, de reflectância e de profundidade) e em técnicas de aprendizagem supervisionada de características. Os principais resultados e contribuições apresentadas nesta tese incluem: a criação de um grande conjunto de dados publicamente disponível contendo aproximadamente 17K vídeos de simulações de ataques de apresentações e de acessos genuínos em um sistema biométrico facial, os quais foram coletados com a autorização do Comitê de Ética em Pesquisa da Unicamp; o desenvolvimento de novas abordagens para modelagem e análise de propriedades extrínsecas das amostras biométricas relacionadas aos artefatos que são adicionados durante a fabricação das amostras sintéticas e sua captura pelo sensor de aquisição, cujos resultados de desempenho foram superiores a diversos métodos propostos na literature que se utilizam de métodos tradicionais de análise de images (e.g., análise de textura); a investigação de uma abordagem baseada na análise de propriedades intrínsecas das faces, estimadas a partir da informação de sombras presentes em sua superfície; e, por fim, a investigação de diferentes abordagens baseadas em redes neurais convolucionais para o aprendizado automático de características relacionadas ao nosso problema, cujos resultados foram superiores ou competitivos aos métodos considerados estado da arte para as diferentes modalidades biométricas abordadas nesta tese. A pesquisa também considerou o projeto de eficientes redes neurais com arquiteturas rasas capazes de aprender características relacionadas ao nosso problema a partir de pequenos conjuntos de dados disponíveis para o desenvolvimento e a avaliação de soluções para a detecção de ataques de apresentação.

Abstract

Recent advances in biometrics, information forensics, and security have improved the recognition effectiveness of biometric systems. However, an ever-growing challenge is the vulnerability of such systems against presentation attacks, in which impostor users create synthetic samples from the original biometric information of a legitimate user and show them to the acquisition sensor seeking to authenticate themselves as legitimate users. Depending on the trait used by the biometric authentication, the attack types vary with the type of material used to build the synthetic samples. For instance, in facial biometric systems, an attempted attack is characterized by the type of material the impostor uses such as a photograph, a digital video, or a 3D mask with the facial information of a target user. In iris-based biometrics, presentation attacks can be accomplished with printout photographs or with contact lenses containing the iris patterns of a target user or even synthetic texture patterns. In fingerprint biometric systems, impostor users can deceive the authentication process using replicas of the fingerprint patterns built with synthetic materials such as latex, play-doh, silicone, among others. This research aimed at developing presentation attack detection (PAD) solutions whose objective is to detect attempted attacks considering different attack types, in each modality. The lines of investigation presented in this thesis aimed at devising and developing representations based on spatial, temporal and spectral information from noise signature, intrinsic properties of the biometric data (e.g., albedo, reflectance, and depth maps), and supervised feature learning techniques, taking into account different testing scenarios including cross-sensor, intra-, and inter-dataset scenarios. The main findings and contributions presented in this thesis include: the creation of a large and publicly available benchmark containing 17K videos of presentation attacks and bona-fide presentation simulations in a facial biometric system, whose collect were formally authorized by the Research Ethics Committee at Unicamp; the development of novel approaches to modeling and analysis of extrinsic properties of biometric samples related to artifacts added during the manufacturing of the synthetic samples and their capture by the acquisition sensor, whose results were superior to several approaches published in the literature that use traditional methods for image analysis (e.g., texture-based analysis); the investigation of an approach based on the analysis of intrinsic properties of faces, estimated from the information of shadows present on their surface; and the investigation of different approaches to automatically learning representations related to our problem, whose results were superior or competitive to state-of-the-art methods for the biometric modalities considered in this thesis. We also considered in this research the design of efficient neural networks with shallow architectures capable of learning characteristics related to our problem from small sets of data available to develop and evaluate PAD solutions.

List of Figures

List of Tables

List of Abbreviations and Acronyms

Abbreviations and Acronyms	Description
3DMAD	3D Mask Attack Database
ACC	Accuracy
APCER	Attack Presentation Classification Error Rate
ATM	Automated Teller Machine
AUC	Area Under Receiver Operating Characteristic Curve
BPCER	Bona Fide Presentation Classification Error Rate
CF	Color Frequency
CNN	Convolutional Neural Network
DET	Detection Error Trade-off
DNA	Deoxyribonucleic Acid
EER	Equal Error Rate
FAR	False Acception Rate
FPN	Fixed Pattern Noise
FRR	False Rejection Rate
GLCM	Gray-Level Co-occurrence Matrices
HD	High-Definition
HOG	Histogram of Oriented Gradients
HSC	Histograms of Shearlet Coefficients
HTER	Half-Total Error Rate
HVC	Hierarchical Visual Codebook
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LPQ	Local Phase Quantization
PA	Presentation Attack
PAD	Presentation Attack Detection
PAI	Presentation Attack Instrument
PLS	Partial Least Square
PRNU	Photo-Responsiveness of Non-Uniform Light-Sensitive Cells
RER	Relative Error Reduction
ROC	Receiver Operating Characteristic
SfS	Shape-from-Shading

SOTA	State of the art (or the *adjective* State-of-the-Art)
SVM	Support Vector Machine
UVAD	Unicamp Video-Based Attack Database
WLD	Weber Local Image Descriptor

Terms and Glossary

Terms and Glossary	Description
Acquisition process	see “Biometric data capture subsystem”
Acquisition sensor	see “Biometric data capture subsystem”
Attack type	Method used to attack the biometric data capture subsystem
Biometric data capture subsystem	Subsystem responsible for capturing the biometric data
Bona fide presentations	Presentation of a truthly biometric data to the biometric data capture subsystem
Cross-dataset evaluation protocol	see “Inter-dataset evaluation protocol”
Cross-sensor scenario	A testing scenario in which a PAD system is evaluated into a different biometric data capture subsystem from which it was configured
Impostor user	A user without permission for accessing a biometric system
Inter-dataset evaluation protocol	A testing scenario in which a PAD system is evaluation into a different biometric system of the same modality, which means an operation with an different biometric data capture subsystem and template database
Intra-dataset evaluation protocol	A testing scenario in which a PAD system is evaluated into the same biometric system of which it was configured
Legitimate user	A user authorized to access a biometric system
Liveness detection	see “Presentation attack”
PAD system	Module responsible for detection presentation attacks
Presentation attack	A single or multiple attempts of attacking the biometric data capture subsystem aiming to deceive a biometric authentication system
Presentation attack instrument	Instrument used to perform a single or multiple attempts of attacking the biometric data capture subsystem
Spoofing attack	see “Presentation attack”
Template database	A database responsible for storing the templates of the legitimate users enrolled in the biometric system
Valid access	see “Bona fide presentations”

Valid user	see “Legitimate user”

Contents

Chapter 1

Introduction

"One can only see what one observes, and one observes only things which are already in the mind."

—Alphonse Bertillon, *French police officer & biometric researcher (1853–1914)*

"A really intelligent nation might be held together by far stronger forces than are derived from the purely gregarious instincts. A nation need not be a mob of slaves, clinging to one another through fear, and for the most part incapable of self-government, and begging to be led; but it might consist of vigorous self-reliant men, knit to one another by innumerable ties, into a strong, tense, and elastic organisation."

—Sir Francis Galton, *English sociologist, anthropologist, inventor, mathematician, statistician, etc. (1822–1911)*

THIS thesis addresses a still open problem in a subfield of the biometric research related to security aspects of biometric systems, named as presentation attack detection (PAD), which can be briefly described as the ability of detecting attempted attack performed by an impostor user that seeks to deceive the authentication system by presenting to the acquisition sensor a synthetic biometric sample of a legitimate user enrolled in the system.

We organize this thesis as a compilation of articles published (or submitted for publication) in scientific journals containing proposals of algorithms and methodologies designed to protect biometric systems against presentation attacks. In this chapter, we discuss the basic concepts in biometrics and the benefits of its use to a reliable and transparent authentication. Next, we discuss the limitations related to security aspects of such systems, focusing on the presentation attacks vulnerability, as well as the hypotheses established in this thesis, the main contributions of this research and their relationships with the remaining chapters of this document.

1.1 Basic Concepts in Biometrics

The protection of personal data has become a fundamental requirement of security. According to Tipton [250], information security is concerned with the development of methods and tools for protecting information and preserving the value it has for an individual or an organization. For efficient and effective protection, the use of robust authentication mechanisms is paramount.

Knowledge-based methods (e.g., password, secret question) and token-based methods (e.g., smart cards, token codes) are probably the most used authentication mechanisms to date. However, these methods have a critical feature: at the time of authentication, the system does not verify who is requesting access, but rather what the users know or possess. This aspect renders the system vulnerable since that knowledge or an object can be easily lost, shared or manipulated. As an alternative, biometrics is an authentication mechanism considered more natural and reliable as it focuses on verifying who is the person requesting the access [109].

Biometrics provides methods to recognize humans automatically based on their behavior, physical or chemical traits, being the fingerprint, face, iris, hand geometry, hand vein and DNA, the most common traits used for deploying such systems [109]. According to the nature of the application, we might have one or more traits more suitable to be used in a biometric system. In forensic applications such as corpse identification, the DNA might be a better choice to identify a victim due to the natural body's deterioration. On the other hand, in government applications (e.g., driver's license, voting, border control), face and fingerprint traits might be more suitable for authenticating users of the system due to the ease in measuring these characteristics. Jain et al. [110] describe seven factors that could help to determine the suitability of a trait to a biometric system: universality, uniqueness, permanence, measurability, performance, acceptability, and circumvention.

Independent of the traits used to recognize a person, a biometric system can operate in two modes, namely verification and identification modes. In the first mode of operation, the system recognizes a user, recovering the template that was previously extracted and stored in a database, at the time of enrollment, and compares it with the template extracted at the time of the authentication claim (query). This comparison is performed by a matching algorithm that produces a similarity score that indicates whether the user is who he claims to be, in the case in that the similarity score is higher than a threshold preset for each user. In the second mode of operation, the system compares the query of the user with N templates enrolled in the database. In this case, the matching algorithm produces N similarity scores, such that a higher score indicates the identity of the user. Figure 1.1 illustrates the architecture of a generic biometric system.

1.2 Presentation Attacks in Biometrics

Although several traits can be used in an authentication process, researchers are constantly looking for biometric traits with low acquisition and storage costs, that are less invasive, present a high degree of uniqueness and are stable. However, the static nature

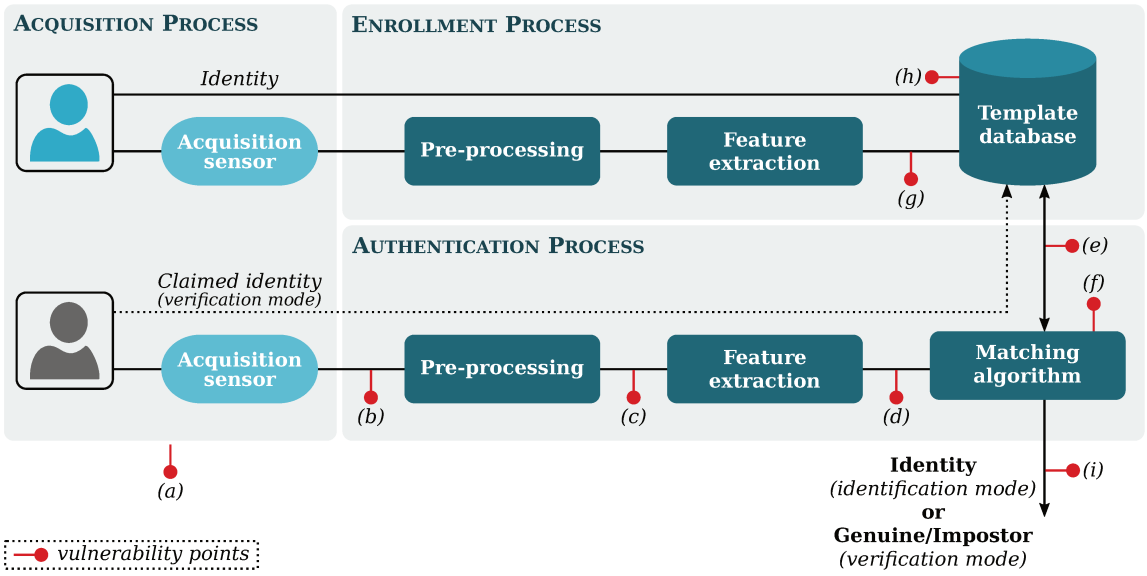


Figure 1.1: Overview of the authentication process in a generic biometric system. The first stage consists of the enrollment process, in which a biometric sample of the user is captured by the sensor, and submitted to the pre-processing module for noise removal, enhancement of the data, among other features. Next, the pre-processed sample is used by the feature extraction module responsible for extracting the main patterns and features relevant to the authentication, generating a specific user template which is stored in the database. In the second stage, when the user tries to authenticate, the template is submitted to the matching algorithm that compares the input template with the ones stored in the database. The red points represent some vulnerabilities when the biometric system is unprotected against such threats (adapted from [202]).

of a stable biometric trait suggests "the paradox of secure biometrics" [167]:

> *"An authenticator must be stable and distinctive to be considered a good authenticator. But, stability leaves no option for compromise recovery, since users cannot change their biometric trait if stolen. Moreover, since a biometric clue is not secret, its information can be learned and copied."*

Despite a stable biometric trait being an ideal authenticator, in practice, its use would not work if it were learned or copied. Therefore, researchers have striven to develop methods that detect whether a biometric sample presented to the acquisition sensor is a replica of the original sample. In the literature, the act of presenting a synthetic biometric sample of some genuine user to the acquisition sensor in order to authenticate itself as a legitimate user is known as presentation attack. Figure 1.1 shows a general biometric authentication system without any protective measure and some points of vulnerabilities. In point (a), we have a threat resulting from an attack on the biometric sensor, presenting a synthetic biometric sample (fake). In points (b), (c), (d), and (e), we have threats resulting from re-submission of a biometric latent signal previously stored in the communication channel. In (f), it can occur an attack on the matching algorithm in order to produce a higher or lower score. In point (g), we have a threat on the communication channel

between the enrollment center and the database (the control of this channel allows an attacker to overwrite the template that is sent to the biometric database). In (h), we have a threat on the database itself, which could result in corrupted models, denial of service to the person associated to the corrupted model, or fraudulent authorization of an individual. Finally, in point (i), we have a threat that consists of overwriting the output of the matching algorithm, bypassing the authentication process. Other abuses in biometric systems can be found in more details in Campisi [34].

Although effective in many situations, a biometric system should have protection against presentation attacks. Erdogmus et al. [66] evaluated the behavior of a face biometric system protected with an anti-spoofing solution proposed in [151] and Microsoft's Kinect motion sensor under attempted attacks performed with static 3-D masks. Kose et al. [127] demonstrated that the face verification system is vulnerable to presentation attacks. In addition, the same authors evaluated the anti-spoofing method proposed in [151], which was proposed to detect photo-based attacks.

Besides these evidences made in laboratories about vulnerability of biometric systems against presentation attacks, real cases confirm the problem. In the city of Ferraz de Vasconcelos, São Paulo state, Brazil, a medical of the SAMU (service of mobile health care and urgency) was caught red-handed by the police using silicone fingers to bypass the authentication system and marking presences at work for co-workers [147]. A similar case it has been investigated by the Federal Police, in which workers of the Paranaguá Harbor, Paraná state, Brazil, where suspects used silicone fingers to circumvent the biometric system for registering the worked hours of employees [35]. In Germany, the biometrics hacking team of the Chaos Computer Club has managed to hack Apple's iPhone Touch ID [10], demonstrating that a biometric system without an adequate protection is unsuitable as an access control method. Other cases of impostors bypassing surveillance systems with 3D masks can be found in [210, 211].

In face of such overwhelming evidence of the need for more secure biometric systems and effective detection of attacks, in this thesis, we investigate the presentation attack problem considering the three different biometric traits: fingerprint, iris, and face. Among several forms of biometrics, face recognition is of paramount importance with outstanding solutions presented thus far such as deformable models [262], texture-based representations [2], and shape-based representations [145]. Although effective in many cases, according to Maltoni et al. [153], face, signature and voice are the easiest biometric signals to be circumvented. For instance, presentation attacks can be successfully accomplished in a face biometric system if an impostor obtains access by presenting to the acquisition sensor a photography, a digital video or a 3D model of the target person [109]. Even with recent advances in biometrics, information forensics and security, the vulnerability of facial biometric systems against presentation attacks is still an open problem.

Iris-based biometrics is another important modality used to recognize people. The high accuracy rate and speed of iris-based recognition systems make them a promising modality [109]. However, such modality is not free of presentation attacks, which are normally performed using printed iris images [229] or, more interestingly, cosmetic contact lenses [32, 264]. Finally, in fingerprint-based biometric systems, the most common presentation attack method consists of using synthetic samples [83] created in a cooperative

way, where a mold of the fingerprint is acquired with the cooperation of a valid user and is used to replicate the user's fingerprint with different materials, including gelatin, latex, play-doh or silicone. Latent fingerprints left on a surface may also be used and further enhanced after acquisition with a digital camera.

1.3 Research Vision: New Insights for the Presentation Attack Detection Problem

During the production of synthetic biometric data, inevitably, there are noise information and telltales added to the biometric signal that can be captured and further processed to pinpoint attacks. In fact, in the manufacturing process of a synthetic sample, there are, at least, two re-quantization and re-sampling steps of the original biometric signal. For example, in the photo- and mask-based face presentation attacks, the continuous signal is quantized and sampled during the digitization process. Then, this digital version is re-quantized and re-sampled due to the printing process with 2D and 3D printers and again digitized during the presentation of the synthetic data to the acquisition sensor. In print-based iris presentation attacks, the process of generation of the synthetic samples are comparable to the photo-based face attempted attacks and, therefore, we also may have noise and telltales upon synthetic data. In video-based face presentation attacks, the continuous signal is digitized and recaptured by the acquisition sensor during an attempted attack. The same reasoning can be applied to the fingerprint-based recognition systems, considering the user cooperation scenario, or by using latent fingerprints. Examples of artifacts and telltales of presentation attacks added in synthetics biometric sample include blurring effects, printing artifacts, banding effects, distortions, flickering, and Moiré patterns. Henceforward, we will refer to the noise and artifacts contained in a presentation attack sample as a noise signature.

1.3.1 Problems with the Existing Approaches

The existing works in the literature for presentation attack detection basically explore three lines of investigation: methods based on texture analysis [117, 125, 139, 141, 151, 152, 181, 214, 245], based on motion and clues of the scene analysis [8, 41, 160, 176, 177, 253, 267, 274], and methods based on image quality [71, 72, 74, 75].

Approaches based on motion and clues of the scene explore the motion information inside the face region (e.g., eye blinks, small movements in the face region) and outside of the face region (e.g., background). These methods estimate the motion information present in these two regions and analyze them over time. Although this approach achieved good results to detect photo-based presentation attacks, the constraints made by these methods (e.g., static background) may render the detection of the video-based attempted attacks unfeasible, in practice. Moreover, the assumption of a background previously known can restrict the use of the approach since in many applications (e.g., web and mobile applications) the data acquisition is performed remotely and, therefore, we cannot assume that. Besides, motion is easily simulated by rotating or bending the photographs,

which can potentially deceive the motion detectors. Another disadvantage of approaches based on motion analysis is that the additional time required to capture some face motions prevents a fast spoofing detection. For example, a type of motion analysis extensively explored in the literature is the action of eye blink that occurs once every four or six seconds. However, this rate can decrease to an average of three to eight every six seconds due to psychological factors [140]. In this case, at least 20 seconds are required to detect eye blinking.

Regarding methods based on texture analysis, their main drawback consists of presenting poor results when we consider high-definition attempted attacks such as print-based images with high-quality printouts and video attacks performed with high-resolution screens. The attempted attacks performed with low-resolution printed images [36] produce a considerable amount of artifacts, which can be easily detected via texture analysis from the original image space. In the video-based attempted attack, these coarse artifacts are practically absent and, therefore, the proposal of more robust methods for detecting these artifacts is necessary. This evidence was confirmed in the 2nd Competition for Face Spoofing Attack [42] and is explored in this thesis.

Although existing PAD algorithms showed good performance results for the intra-dataset evaluation protocol, in which the training and test partitions are from the same dataset (i.e., same domain), the main drawbacks of the current PAD algorithms appear when we consider more challenging protocols and testing scenarios such as inter-dataset and cross-sensor scenarios. The inter-dataset protocol proposes the validation of a PAD algorithm considering the training and testing stages in a completely different scenario by establishing partitions for training and testing whose samples come from different sources. In this case, we have sets for training and testing built in different conditions (e.g., different sensors, different environments, without overlapping of identities, different illumination conditions, different presentation attack instruments (PAI), different attack types, among others). Similarly, the cross-sensor scenario considers testing sets comprising data captured with a sensor, or set of sensors, different from those used for capturing the training data. In this thesis, we also investigate the cross-sensor and inter-dataset scenarios, since such evaluation protocols are more challenging and suitable for reflecting a real operating scenario.

1.3.2 Hypothesis Statements

This thesis introduces alternatives for detecting presentation attacks, taking into account our understanding of limitations of the current approaches and our understanding of the problem itself. The thesis hypotheses presented in this section reflect directions of new perspectives to deal with this problem:

Hypothesis 1: *Noise signatures extracted of biometric samples contain meaningful tell-tales for an effective presentation attack detection, in comparison to existing methods, due to artifacts added to the synthetic biometric samples during its manufacturing process.*

Hypothesis 2: *Supervised feature learning techniques provide an effective representation of data for detecting presentation attacks in different biometric modalities, in comparison with existing techniques published in the literature.*

Hypothesis 3: *Facial surfaces estimated with a shape-from-shading technique contain distinguishable artifacts for a robust presentation attack detection, in comparison to existing methods, which are added during the manufacture of synthetic samples and magnified throughout the surface reconstruction.*

Hypothesis 4: *Shallow Convolutional Neural Networks trained with visual characterizations of the intrinsics properties of facial surfaces (i.e., albedo, reflectance, and depth maps) provide an effective representation of data for detecting presentation attacks in cross-domain scenarios, e.g., the inter-dataset scenario, in comparison to existing architectures published in the literature.*

1.3.3 Novelties and Rationales Brought in this Thesis

The rationales that motivate us to formulate these hypotheses came from a careful analysis of different results from the literature, which will be briefly explained in the next sections and deeply discussed in the remaining chapters of this thesis.

Investigation of approaches based on temporal and spectral analysis

As mentioned before, synthetic biometric samples inevitably contain noise and artifacts generated during their manufacture and recapture might be different from any pattern found in real biometric samples. According to Tan et al. [245] and Määttä et al. [151], there is a deterioration of the facial information and, consequently, a loss of some high frequency components during the manufacture of photographs to be used in spoofing attacks. In our prior work [193], we highlighted the fact that there is a significant increase of the low frequency components due to the blurring effect added during the recapture process of the biometric sample displayed in tablets, smartphones and laptop screens. Besides the blurring effect, other artifacts are added such as flickering, Moiré patterns, and banding effect [14].

These facts motivated us to propose a solution that takes advantage of the noise signature contained on such presentation attack samples (**Hypothesis 1**). **Chapters 2 and 3** propose a Fourier analysis of the noise signature to capture the information encoded in the frequency, phase and amplitude of the component sinusoids [236]. In these chapters, we use Fourier spectrum to quantify the following artifacts:

- *Blurring artifact:* In both production and recapture processes, inevitably we have a decrease in the details of biometric samples due to re-quantization and re-sampling of the original signal. This reduction of details is reflected in the increase of low frequency components and can be observed in the Fourier domain;

- *Flickering effect:* It corresponds to the horizontal and vertical lines equally spaced that appear during the recapture process of samples shown to the acquisition sensor with the display device. When this artifact appears in biometric samples, there are peak lines at abscissa and ordinate axes of the Fourier spectrum when the display device is aligned with the acquisition sensor;

- *Moiré patterns:* They are patterns that can appear when a display device is used to perform an attempted attack. In general, this effect is generated when two similar patterns are overlaid with a small difference in rotation from each other. As a result, we also have the appearance of peaks in different locations in the Fourier spectrum depending on the frequency and direction of the sinusoid in the spatial domain [236].

The main novelty brought in these studies are two-fold: (i) a new perspective of using Fourier analysis to detect presentation attacks, which is performed upon the noise signature extracted from the biometric sample under analysis; and (ii) novel techniques for describing spectral information over time, which led us to propose some time-spectral descriptors. Previous attempts of using the Fourier analysis in this problem [139,141,245], considered its use on the original image representation, which turns this analysis highly sensitive to illumination conditions and image resolutions [141] since these factors cause a considerable disturbance in the frequency components of the Fourier spectrum. On the other hand, our studies demonstrated that the spectral analysis is a powerful mathematical tool for detecting presentation attacks if we consider the noise signature and discard the contents of the images under analysis.

Investigation of supervised feature learning techniques for the PAD problem across different biometric modalities

Inspired by Pinto et al. [196] and Bergstra et al. [21], this research investigated the **Hypothesis 2** considering three approaches to building convolutional networks: the architecture optimization-based approach, filter optimizations, and an interplay between these two techniques.

The architecture optimization consists of exploring thousands of candidate models by considering a search space for the parameters related to the architectural aspect (hyper-parameters) of the networks. To enable a fast evaluation of a set of candidate models, we skipped the filter learning process to the detriment of using filters with random weights. In this work, we evaluate two strategies for searching good hyper-parameters, the random search strategy and the Tree-structured Parzen Estimator Approach (TPE) [21]. On the other hand, the filter optimization process is a traditional method for building convolutional networks that consists of learning proper weights for the filters of the convolutional layers.

Chapter 4 aims to investigate shallow convolutional neural networks (CNN) by considering these two strategies and their combination (**Hypothesis 4**). This study was pioneered in exploring CNN for the presentation attack detection problem and the novelty of this work consists of finding a shallow CNN, named *SpoofNet*, suitable for the three modalities considered in this thesis (**Hypothesis 2**).

Finally, **Chapter 5** introduces a methodology to adapt the Visual Geometry Group (VGG) network [233], for automatically learning features for the presentation attack problem also considering the three modalities examined in this thesis (**Hypothesis 2**). The main purpose of this chapter is to investigate the feature learning process, using a deep CNN architecture, under the intra- and inter-dataset scenarios.

Investigation of telltales of presentation attacks presented in the facial surface reconstructed using shape-from-shading method

Chapter 6 presents an investigation of a novel approach to PAD based on the optical and physical properties of the scene captured by the acquisition sensor (**Hypothesis 3**). Our method takes advantage of the depth, reflection and albedo information, associating them with light properties of the scene to detect an attempted attack. For estimating these properties, we use a classic technique in computer vision known as shape-from-shading (SfS) [99], which aims to reconstruct the surface of an object based on the shading information present in its surface.

The law of refraction [249] (also known as *Snell's law* or *Snell–Descartes law*) establish an understanding of the physical mechanism of the light refraction, in terms of absorption and irradiation of the light incident on a surface. Complementary, the reflection's law governs the reflection of the incident light and states that the incident ray, the reflected ray, and the normal to smooth conducting surfaces (e.g., mirror or polished metal) all lie in the same plane [249]. According to refraction and reflection laws, the beam of light that affects a flat surface may be absorbed, transmitted, and reflected, and the directions of the light refracted and reflected can be predicted considering the refraction index of the material and the roughness of its surface, i.e., the smoothness or texture of the surface.

When a beam of light affects a truly flat surface, each incident ray is reflected at the same angle that we have between the surface normal and such incident ray, but on the opposite side of the surface normal. In contrast, when a beam of light affects rough surfaces, the incident light is reflected in several different directions. An ideal diffuse reflecting surface that reflects the incident light in all directions is said to exhibit a Lambertian reflection. These two processes are known as specular and diffuse reflection, respectively. Although many materials can exhibit both types of reflection, some materials reflect the light in a way that is more diffuse than specular way (e.g., paper fibers, non-absorbing powder such as plaster, poly-crystalline material such as white marble, among others) [70, 116, 165, 252], which makes this property very promising to our problem. **Chapter 6** discusses this and other properties in detail.

The main novelties brought in this line of investigation include: (i) a novel PAD algorithm able to detect evidence of attempted attacks, taking into account artifacts present in the reconstructed surface; (ii) the use of the depth information without using any extra device, which enables the use of our approach in biometric systems equipped with a single RGB camera; (iii) promising performance results for detection through the intra- and inter-dataset scenarios; and (iv) a novel shallow CNN architecture for learning features from the albedo, reflectance and depth maps (**Hypothesis 4**, which was also an object of investigation in **Chapter 6**).

1.4 Key Contributions

Figure 1.2 summarizes the main contributions of this thesis, as well the main challenges and advancements achieved in each step of this research. The other contributions obtained along this thesis will be discussed in their respective chapters.

Contribution 1. A new algorithm that takes advantage of the spectral analysis of the noise signature of the video under analysis and the summarization technique, namely visual rhythm, to generate discriminative texture maps able to reveal attempted attacks. To the best of our knowledge, this is the first method published in the literature that uses noise signature information to detect face presentation attacks, giving new alternatives to solve the problem. We summarize the main progresses obtained in this study in **Advancement 2 and 4**, described in Fig. 1.2 (see **Chapter 2** for more detail).

Contribution 2. An effective method for face presentation attack detection able to recognize different types of attacks, including video-, photo- and 3-D mask-based attempted attacks. We present an effective algorithm based on spectral analysis of the noise signature that takes advantage of the spectral and temporal information that outperforms the current state-of-the-art methods to detect face presentation attacks based on texture and motion analysis. The foremost advance achieved in this work was the design of an effective time-spectral descriptor, showing that the noise signal analysis is in fact an important source of evidence for designing new anti-spoofing algorithms. The most expressive progresses achieved in this study are summarized in **Advancement 3 and 4** (see **Chapter 3**).

Contribution 3. A unified framework able to detect presentation attacks in face-, fingerprint- and iris-based biometric systems based on deep learning techniques. In this work, we show effective learning representation strategies for the modalities of biometrics we consider herein. The proposed unified framework was effective when tested with small datasets. We proposed three deep feature characterization architectures based on Convolutional Neural Networks, one for each modality, that outperform several anti-spoofing methods published in the literature. We summarized the main progresses obtained in this study in **Advancement 5 and 7** (see **Chapter 4** for more details).

Contribution 4. The proposal of an anti-spoofing solution for the three modalities widely employed to design biometric systems, i.e., face, iris, and fingerprint, based on VGG network architecture, a popular deep network originally proposed for the object recognition problem. In this study, we showed a methodology to adapt the VGG network to the two-class presentation attack problem, which was evaluated by considering the most challenging scenarios such as classification across different attack types, biometric sensors, and qualities of samples used during the attacks, besides presenting performance results for the intra- and inter-dataset scenarios. The main progresses achieved in this study are presented in **Advancement 6 and 7** (see **Chapter 5** for more details).

Contribution 5. An effective PAD solution that takes advantage of intrinsic properties of the objects of interest (in our case, we deal with faces) such as reflectance, albedo and depth information to a reveal attempted attack. We propose the use of a shape-from-shading technique to reconstruct facial surfaces in order to highlight telltales of presentation attacks present in the biometric samples. Our results show that the proposed algorithm works well for detecting different attack types even considering more challenging

operation scenarios, simulated with tests considering the multi-sensor scenario and the inter-dataset protocol evaluation. We summarized the main progresses achieved in this study in **Advancement 8 and 9** (see **Chapter 6** for more details).

Contribution 6. An effective shallow CNN architecture suitable for learning meaningful features from the surfaces reconstructed with a shape-from-shading method. A comparison between the proposed shallow CNN architecture and some popular deep architecture that demonstrated the effectiveness of our CNN architecture for learning useful patterns to our problem, even considering the multi-sensor and inter-dataset scenarios. The main expressive progresses achieved in this study are summarized in **Advancement 10 and 11** (see **Chapter 6** for more details).

Contribution 7. The creation of a large and publicly available benchmark considering several display devices and different acquisition sensors. This dataset contains 808 valid access videos and 16,268 videos of video-based presentation attacks, all in full high-definition quality, captured from 404 people[1]. This is the largest and the first multi sensor-based dataset to evaluate face anti-spoofing methods, whose valid access and attempted attack videos were captured with six sensors of different manufacturers. This dataset allowed us to achieve an important conclusion, not yet reported in the literature, regarding the possible impact/influence of the sensor upon the performance of a presentation attack detection system. We summarized the main progress obtained in this study in **Advancement 1** (see **Chapter 2** for more details).

1.5 Thesis Organization

We organized this thesis as a compilation of articles published (or submitted for publication) in international scientific venues in the area of Information Forensics and Security, Image Processing and Neural Networks and Learning Systems. In total, we published the results achieved in this thesis in three journals, one chapter of a book, and one manuscript submitted for publishing, also in a journal. **Chapter 2** presents a proposed method for face presentation attack detection based on a summarization technique, namely visual rhythm, which takes advantage of the spectral and temporal information by analyzing the noise signatures generated by the video acquisition process. **Chapter 3** describes an effective time-spectral descriptor for face presentation attack detection able to characterize telltales present in both spectral and temporal domains, which proved to be a very effective approach to detect attempted attacks performed with different type of attacks (e.g., printed photos, replay attack and 3D mask). **Chapters 4 and 5** present a unified solution able to detecting attempted attacks in face-, iris- and fingerprint-based biometric systems, while **Chapter 6** introduces a PAD algorithm that takes advantage of intrinsic properties of the facial surfaces. Finally, **Chapter 7** presents conclusions and final remarks of the research presented in this thesis.

[1]The users present in the database formally authorized the release of their data for scientific purposes (see **Appendix A**).

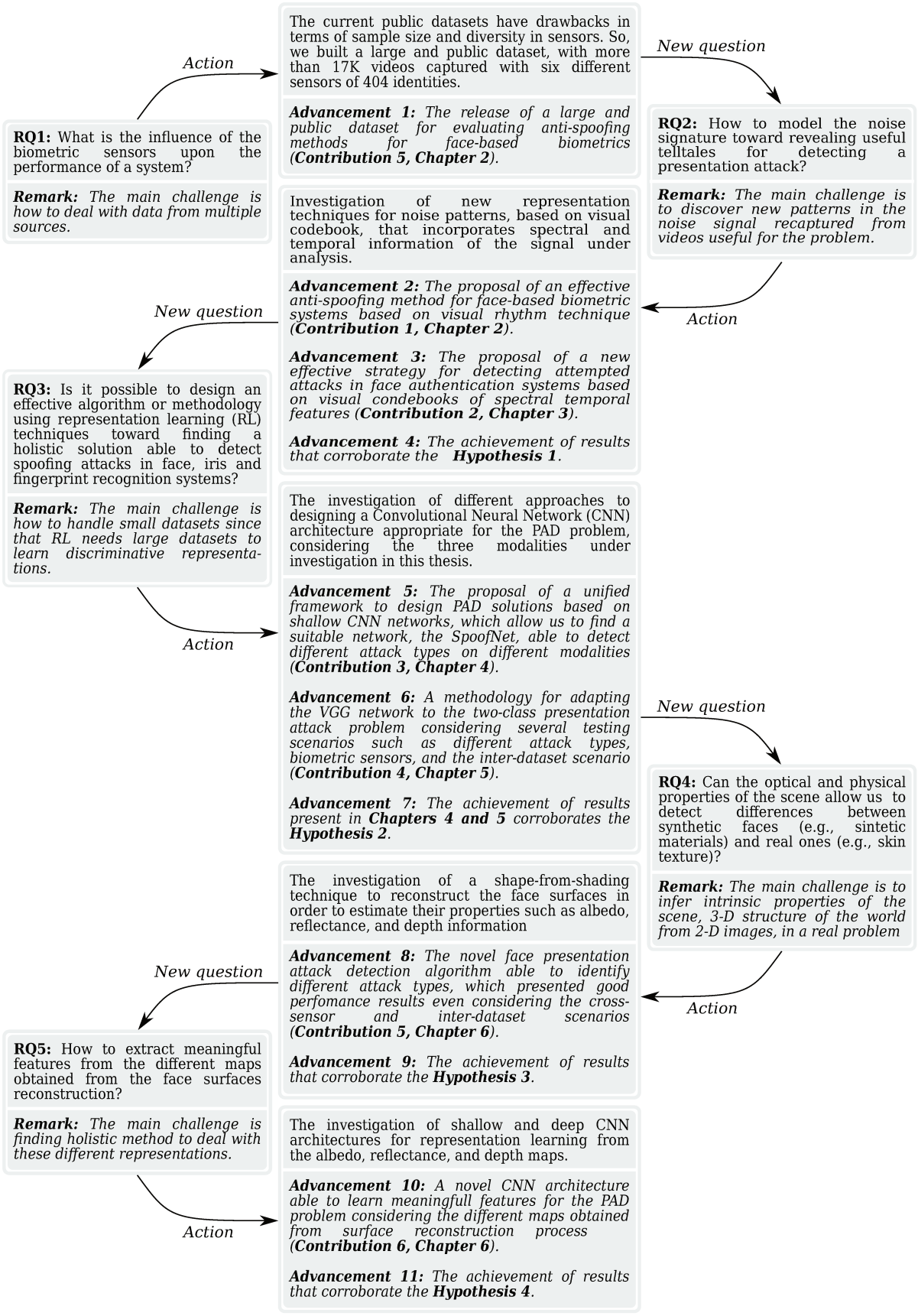


Figure 1.2: This roadmap shows the main contributions of this thesis, Research Questions (RQs) driving our research, main challenges and advancements achieved in our work.

Chapter 2

Using Visual Rhythms for Detecting Video-based Facial Spoof Attacks

"When you run by yourself, you go fast; But when you run with others, you can go so far."

—Maggie MacDonnell, *Global Teacher Prize 2017 Winner*

"The art and science of asking questions is the source of all knowledge."

—Thomas Berger, *American novelist (1924–2014)*

Abstract

Spoofing attacks or impersonation can be easily accomplished in a facial biometric system wherein users without access privileges attempt to authenticate themselves as valid users, in which an impostor needs only a photograph or a video with facial information of a legitimate user. Even with recent advances in biometrics, information forensics and security, vulnerability of facial biometric systems against spoofing attacks is still an open problem. Even though several methods have been proposed for photo-based spoofing attack detection, attacks performed with videos have been vastly overlooked, which hinders the use of the facial biometric systems in modern applications. In this paper, we present an algorithm for video-based spoofing attack detection through the analysis of global information which is invariant to content, since we discard video contents and analyze content-independent noise signatures present in the video related to the unique acquisition processes. Our approach takes advantage of noise signatures generated by

Pre-print of article that will appear in T-IFS, vol.10, no.5, pp.1025-1038, May 2015.
The published article is available on http://dx.doi.org/10.1109/TIFS.2015.2395139
See permission to use the copyrighted material in **Appendix C**.

the recaptured video to distinguish between fake and valid access videos. For that, we use the Fourier spectrum followed by the computation of video visual rhythms and the extraction of different characterization methods. For evaluation, we consider the novel Unicamp Video-Attack Database (UVAD) which comprises 17,076 videos composed of real access and spoofing attack videos. In addition, we evaluate the proposed method using the Replay-Attack Database, which contains photo-based and video-based face spoofing attacks.

2.1 Introduction

BIOMETRIC authentication is an important mechanism for access control that has been used in many applications. Traditional methods, including the ones based on knowledge (e.g., keywords, secret question) or based on tokens (e.g., smart cards), might be ineffective since they are easily shared, lost, stolen or manipulated. In contrast, the biometric access control has been shown as a natural and reliable authentication method [107].

Access control can be seen as a verification problem wherein the authentication of a user is performed by reading and comparing the input biometric data captured by an acquisition sensor (query) with the biometric data of the same user previously stored in a database (template). The comparison between the query and the template is performed by a matching algorithm which produces a similarity score used to decide whether or not the access should be granted to the user.

Although biometric authentication is considered a secure and reliable access control mechanism, it becomes an easy target for attacks if protective measures are not implemented. Fig. 2.1 shows a general biometric authentication system without any protective measure and some points of vulnerabilities. Buhan et al. [33] provide more details about abuses in biometric systems.

Spoofing attack is a type of attack wherein an impostor presents a fake biometric data to the acquisition sensor with the goal of authenticating oneself as a legitimate user (this action can be seen as an impersonation attack), illustrated in Fig. 2.1(a). Depending on the biometric trait used by the system, this mode of attack can be easily accomplished because some biometric data can be synthetically reproduced without much effort. Face biometric systems are highly vulnerable to such attacks since facial traits are widely available on the Internet, on personal websites and social networks such as Facebook[1], MySpace[2], YouTube[3]. In addition, we can easily collect facial samples of a person with a digital camera.

In the context of face biometrics, a spoofing attack can be attempted by presenting to the acquisition sensor a photograph, a video or a 3D face model of a legitimate user enrolled in the database. If an impostor succeeds in the attack using any of these approaches, the uniqueness premise of the biometric system or its *raison d'être* is violated, making the system vulnerable [107].

[1] http://www.facebook.com
[2] http://www.myspace.com
[3] http://www.youtube.com

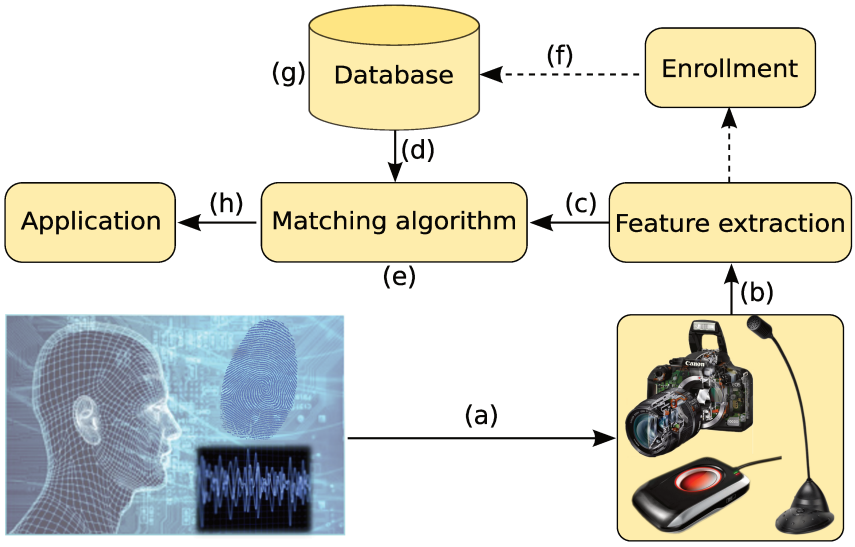


Figure 2.1: General biometric system and its vulnerability points. (a) a threat resulting from an attack on the biometric sensor, presenting a synthetic biometric data (fake); (b), (c) and (d) represent threats resulting from re-submission of a biometric latent signal previously stored in the communication channel; (e) attack on the matching algorithm in order to produce a higher or lower score; (f) an attack on the communication channel between the enrollment center and the database (the control of this channel allows an attacker to overwrite the template that is sent to the biometric database); (g) an attack on the database itself, which could result in corrupted models, denial of service to the person associated to the corrupted model, or fraudulent authorization of an individual; (h) an attack that consists of overwriting the output of the matching algorithm, bypassing the authentication process.

Several methods have been proposed in the literature to detect spoofing attacks based on photographs, whereas attacks performed with videos and 3D models have been overlooked. Many methods aim at distinguishing real from fake biometric data based on the fact that artifacts are inserted into the printed samples due to printing process, therefore allowing one to explore attributes related to such artifacts including color, shape and texture [151, 214, 245]. Since photographs are static, another approach is to detect small movements in the face [140, 176, 263]. Recent works [8, 177] investigate context information of the scene (e.g., background information) to detect face liveness.

We believe that the aforementioned approaches are not suitable for detecting video-based attacks directly, especially in high resolution videos. The difficulty in detecting spoofing performed by video lies in the fact that it is easier to deceive an authentication system through a video since the dynamics of the video makes the biometric data more realistic. Furthermore, the content of a video is less affected by degradations in terms of color, shape or texture, unlike the printed images. Finally, we have less artifacts generated during quantization and discretization of the image captured by the imaging sensor in high resolution videos.

In this paper, we present a method for detecting video-based face spoofing attacks under the hypothesis that fake and real biometric data contain different acquisition-related

noise signatures. To the best of our knowledge, this is the first attempt of dealing with video-based face spoofing using analysis of global information that is invariant to the video content. Our solution explores the artifacts added to the biometric samples during the viewing process of the videos in the display devices and noise signatures added during the recapture process performed by the acquisition sensor of the biometric system. Through the spectral analysis of the noise signature and the use of visual rhythms, we designed a feature characterization process able to incorporate temporal information of the behavior of the noise signal from the biometric samples.

In a previous work [193], we introduced an anti-spoofing solution that was evaluated in an extended version of the Print-Attack database [8] given that, in the literature, there was no specific database to video-based face spoofing attacks. Originally, the Print-Attack database was developed to be used in the evaluation of photograph-based spoofing attack detection. As our aim in that work was also at video-based spoofing detection, we simulated attempts of spoofing attacks using 100 videos of valid access in six monitors, generating 600 attempted attack videos. We reported near-perfect classification results (AUC $\approx$ 100%). That is due to the low resolution of original videos, which favored the high performance of our method, since noise signal was the main information used in it. Furthermore, in a more realistic attack, an impostor probably would create fake biometric samples with the highest quality possible in order to minimize the differences between real and fake biometric samples.

To contemplate a more realistic scenario, this work extends upon our previous work [193] and also introduces the Unicamp Video-Based Attack Database (UVAD) [4], specifically developed to evaluate video-based attacks in order to verify the following aspects:

- The behavior of the method for attempted attacks with high resolution videos;
- The influence of the display devices in our method;
- The influence of the biometric sensor in the proposed method;
- The best feature characterization to capture the video artifacts;
- Comparison with one of the best anti-spoofing methods for photo-based spoofing attack of notice.

Such verifications can be accomplished due to the diversity of the devices used to create the database which comprises valid access and attempted attack videos of 404 different people. Each user was filmed in two sections in different scenarios and lighting conditions. The attempted attack videos were produced using seven different display devices and six digital cameras from different manufacturers. The database has 808 valid access videos and 16,268 videos of video-based attempted spoofing attacks, all in full high definition quality.

In summary, the main contributions of this work are:

[4]This database will be make public and freely available. Users present in the database formally authorized the release of their data for scientific purposes (see **Appendix A**).

(i) An efficient and effective method for video-based face spoofing attack detection able to recognize attempted attacks carried out with high-resolution videos;

(ii) The evaluation of the video characterization process considering different image features such as the Gray-Level Co-occurrence Matrices (GLCM), Histograms of Oriented Gradients (HOG) and Local Binary Patterns Histogram (LBP) feature descriptors;

(iii) The creation of a large and publicly available benchmark to evaluate anti-spoofing methods performed with videos considering several display devices and different acquisition sensors;

(iv) A detailed study of the video-based spoofing attack problem that yielded important conclusions that certainly will be useful for the proposition of new anti-spoofing methods for video-based attacks not only in the biometric domain but also in other applications analyzing video recapture footprints.

We organize the remainder of this paper into five sections. Section 2.2 discusses state-of-the-art methods for detecting spoofing attacks to face biometrics. Section 2.3 presents the proposed method. Section 2.4 gives details regarding the proposed video-attack database while Section 2.5 shows and discusses the experimental results. Finally, Section 2.6 presents the conclusions obtained with this work.

2.2 Related Work

According to Pan et al. [175], there are four major categories of anti-spoofing methods: data-driven characterization, user behavior modeling, user interaction need, and the presence of additional devices. Solutions that require extra devices are limited due to their high cost, which can prevent large-scale use (e.g., deployment of an anti-spoofing solution on all ATMs of a banking network). The user cooperation during the biometric authentication can also be used to facilitate spoofing attack detection, however, this procedure lessens the transparency and inserts an additional time in the authentication process. Finally, the user behavior modeling approach (e.g., eye blinking, small face movements) has been considered in the literature for photo-based face spoofing detection, nevertheless, this approach might not work well for video-based spoofing attack detection due to the high dynamics present in video scenes. Solutions based on data-driven characterization explore biometric data by thoroughly searching for evidence and artifacts useful to detect attempted attacks.

In this section, we review the literature on user behavior modeling and data-driven characterization methods, since such methods are preferable in practice because they are non-intrusive and do not require extra devices or human interaction. Therefore, they are easily integrable with existing face recognition systems. In this category, there are several methods for photo-based spoofing attack detection that explore clues such as motion and frequency analysis, scene information, and texture. Before going any further, however, we first present some available face-related spoofing databases in the literature since most of the methods use one or some of such reference benchmarks.

2.2.1 Existing Databases

NUAA Database

The NUAA Photograph impostor database [245] comprises 5,105 valid access images and 7,509 fake images collected with a generic webcam. The images of valid access were collected of 15 identities in three sections in different places and illumination conditions, all with 640×480 pixel resolution. The production of the fake samples were done by taking high resolution photographs of 15 identities with a Canon digital camera. The authors simulated two attack modes: (1) printing photographs on photo paper; and (2) printing the photographs on A4 paper using an HP color printer.

Print-Attack Database

The Print-Attack database [8] contains short videos of valid access and photo-based spoofing attacks of 50 identities. The valid access videos were generated in controlled and uncontrolled illumination conditions. All videos are in 320×240 pixel resolution, 25 frames per second (fps) and 15 seconds of duration. The attempted attack videos were generated by taking two high resolution photographs with a Canon PowerShot digital camera of the 50 identities printed on common A4 papers. The attempted attack videos were produced showing the photographs to a webcam considering two attack modes: (1) hand-based attacks wherein the impostor user presents the photographs using her own hands; and (2) fixed-support attacks in which the photographs were glued on a wall so that they do not move during the attempted attacks. In total, 200 access valid videos and 200 attempted attack videos were generated.

CASIA Database

The CASIA database [274] comprises 600 video clips of 50 identities. The videos were filmed in a natural scene with three cameras: a new and an old USB camera both with 640×480 pixel resolution and a Sony NEX-5 digital camera with $1,920 \times 1,080$ pixels of resolution. The database contains three attack modes: (1) warped photo attack; (150 640×480-attempted attack videos); (2) cut photo attack (150 640×480-attempted attack videos); and (3) video playback using an iPad (150 $1,280 \times 720$-attempted attack videos). Some limitations of this database include: the authors failed to prevent the downsizing of the videos shown during the simulation of the video-based spoofing attacks. Such downsizing adds artifacts to the attempted attack videos that are not present in the valid access videos, creating an artificial data separability. Furthermore, the small amount of data and the use of only one device in the creation of the video-based spoofing attacks prevent more refined investigations.

Replay-Attack Database

The Replay-Attack database [41] contains short video recordings of valid access and attempted attacks of 50 identities. Similar to the Print-Attack [8], the videos were generated with a low resolution webcam with 320×240 pixel resolution, 25 fps and 15 seconds of

duration and the video capture process is the same as described in [8]. However, different from [8], two other attempted attack modes are considered: (1) mobile attacks where the impostor user displays photographs and videos in an iPhone screen produced with the same iPhone; and (2) high-definition attacks where the impostor user shows high resolution photographs and videos produced with a Canon PowerShot digital camera using the screen of a 1024×768-pixel resolution iPad.

2.2.2 Motion Analysis and Clues of the Scene

Motion analysis of the face region was an early approach used to detect the liveness of biometric samples. In [176], Pan et al. investigated the action of eye blinking to detect attacks performed with photographs. The authors proposed the use of the undirected conditional random field framework to model the action of opening and closing eyes. Tests were performed in a database with 80 videos and 20 identities using a webcam. The authors reported a false alarm rate smaller than 1%. Similarly, Li et al. [140] proposed a method for detecting a person's eye blink based on the fact that edges vary homo-responsively to the behavior of eye blink over some scales and orientations. Analyzing the trends of Gabor response waves in multi-scale and multi-orientation, the authors choose the five most homo-responsive Gabor response waves to the behavior of eye blink.

In [263], Xu et al. proposed a method for detecting the eye states formulated as a binary classification problem in which the closed state represents the positive class and the open state the negative class. The authors scan the region of the eyes with N blocks of different sizes for each biometric sample. For each block, three different feature vectors were extracted by using variants of the Local Binary Pattern Histogram method, generating three sets with N feature vectors. The authors collected $11,165$ images from which $5,786$ were used in the training stage. The best reported detection rate was 98.3%.

Tronci et al. [253] explored the motion information and clues that are extracted from the scene considering static and video-based analyses. A static analysis consists of capturing spatial information of the still images using different visual features as color and edge directivity descriptor, fuzzy color and texture histogram among others. The analysis is motivated by the loss of quality and by the addition of noise in the biometric samples during the manufacturing process of the photographs. Video-based analysis is performed as a combination of simple measures of motion such as eye blink, mouth movement, facial expression change among others. In the end, a classifier is trained for each feature with the aid of a fusion scheme for determining spoofing attacks.

Pan et al. [177] extended upon [176] by including context information of the scene. The authors analyzed clues such as eye blink in the face region. They extracted a set of key points and calculated a Local Binary Pattern Histogram (LBP) around such points and used . the χ^2 distance function to compare histograms to reference patterns previously calculated.

Anjos et al. [8] proposed a database and a method for photo-based spoofing attack detection assuming a stationary facial recognition system. In this case, the intensity of the relative motion between the region of the face and the background can be used as a clue to distinguish valid access of attempted attacks. The authors calculate a measure of

motion for each video frame obtaining a one-dimensional signal, which is described by the extraction of five measures to form a feature vector. The authors validated the method through the Print-Attack database (c.f., Sec. 2.2.1).

Yan et al. [267] proposed a method for liveness detection based on three scene clues in both spatial and temporal spaces. According to the authors, the non-rigid facial motion and the face-background consistency incorporate temporal information that can help the decision-making process regarding the face liveness. The authors seek a pattern of non-rigid motion in the face region using the batch image alignment method. The face-background consistency is based on the fact that if the face is real, its motion must be totally independent of the background and is performed by separating the region of the face from background and analyzing the motion. Finally, the authors perform a banding artifact analysis, which are treated as additive noise. For that, the authors calculated the first order wavelet decomposition of the image. The authors validated the method through the Print-Attack database (c.f., Sec. 2.2.1) as well as others created by them. Good results were reported.

2.2.3 Texture and Frequency Analysis

Li et al. [141] proposed an anti-spoofing method for photo-based attempted attacks under the assumption that the faces present in photographs are smaller than the real faces and that the expressions and poses of the faces in the photographs are invariant. The detection of an attack through photographs is performed by analyzing the 2-D Fourier spectrum of the samples and calculating the energy rate of the high frequency components, which is used as a threshold to decide whether the biometric sample came from a fake face or not.

In [245], Tan et al. dealt with printed photographs attacks by assuming that the surface roughness of real and photo-attack classes are different. The authors proposed the use of the Variational Retinex-based and Logarithmic Total Variation methods for estimating the luminance and reflectance of an input image, respectively. The authors modeled the detection problem as a binary classification problem and evaluated the use of the Sparse Logistic Regression and Sparse Low Rank Bilinear Logistic Regression methods for classifying the luminance, reflectance, and Fourier spectrum images previously estimated. The authors validated the method through the NUAA Photograph impostor database (c.f., Sec. 2.2.1). Peixoto et al. [181] extended upon [245] by incorporating methods for dealing with different illumination conditions. The reported results showed that the proposed extension reduced the misclassification in more than 50% to attempted attacks with high resolution photographs of the NUAA database.

Määttä et al. [151] proposed a method for photo-based spoofing based on the fact that real and fake biometric facial samples differ: (1) in how these objects reflect light (human faces are 3D objects while printed faces are planar objects); (2) in the pigmentation; and (3) in the quality due to printing defects contained in the photographs. The authors used the LBP method for capturing micro-texture information. Thet evaluated the algorithm through the NUAA database (c.f. Sec. 2.2.1), obtaining an AUC of 99%. In [152], the same authors extended their algorithm for considering Histogram of Oriented Gradient (HOG) and the Gabor wavelet descriptors.

Schwartz et al. [214] proposed an anti-spoofing solution for photo-based attacks exploring different properties of the face region (texture, color and shape) to obtain a holistic face representation. Considering only the face region, for each frame of the video containing the facial information, we generate a feature vector formed by combining different low-level feature descriptors as Histogram of Oriented Gradients (HOG), Color Frequency (CF), Gray Level Co-occurrence Matrix (GLCM), and Histograms of Shearlet Coefficients (HSC). Then, the feature vectors are combined into one feature vector containing a rich spatial-temporal information of the biometric sample and fed to a Partial Least Square classification technique.

In [117], Kim et al. explored two key observations: (1) the difference in the existence of 3D shapes leads to the difference in low frequency regions which is closely related to the luminance component; and (2) the difference between real and fake faces generates a disparity in the high frequency information. The motivation for using texture information lies in the fact that printed faces tend to loose the richness of texture details. Their method extracts a feature vector from each biometric sample by transforming the images to the frequency domain and calculating their respective Fourier spectrum on logarithmic scale, from which average values of the energy of 32 concentric rings are extracted.

Recently, Zhang et al. [274] proposed a simple algorithm for detecting photo-based attempted spoofing attacks based on the fact that fake faces present lower quality compared with real faces. For a given image captured by the acquisition sensor, four Difference of Gaussian filters (DoG) with different values of σ were used to extract high frequency information, generating four new images that were concatenated and used as input of a binary classifier trained using the Support Vector Machine (SVM) technique.

In [41], Anjos et al. conducted a study to investigate the potential of texture descriptors based on Local Binary Pattern (LBP), such as $\text{LBP}^{u2}_{3\times3}$, transitional (tLBP), direction-coded (dLBP) and modified LBP (mLBP). From the histograms generated from the descriptors mentioned above, the authors evaluated a simple manner to classify them based on histogram comparisons through χ^2 distance. A set of classifiers was considered, such as Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) with a radial basis function as kernel. Evaluations were performed on the NUAA, Print-Attack, and Replay-Attack databases (c.f., Sec. 2.2.1).

2.2.4 Other Approaches

Optical flow analysis has also been considered in the literature for photo-based spoofing attack detection. Bao et al. [13] proposed an anti-spoofing solution based on the analysis of the characteristics of the optical flow field generated for a planar and 3D object.

Unlike the faces contained in photographs, which are regular planar objects, real faces are irregular and 3D objects, which lead to a differentiation between the optical flow fields generated for real and fake faces. In [123], Kollreider et al. analyzed the trajectory of three parts of the face: the region between eyes and nose, left ear, and right ear. Using optical flow patterns and a model based on Gabor decomposition, the authors note that, in real faces, these parts of the face move differently from fake faces.

Marsico et al. [62] proposed an anti-spoofing solution based on the theory of 3D pro-

jective invariants. By the fundamental theorem of the invariant geometry, it is possible to show that the cross ratio of five points on the same plane are invariant to rotations if and only if the these points satisfy specific collinearity or co-planarity constrains. Thus, six cross-ratio measures are computed to different configurations of points located in non-coplanar regions of the face (e.g. center of eyes, nose tip and chin). If a pose of the face located in front of the acquisition sensor changes, but the computed cross ratio remains constant, the points must be coplanar (i.e., they belong to a planar fake face).

Finally, recent works have been developed in order to evaluate spoofing attacks in multi-modal biometric systems including [3, 4, 25, 26, 157]. In these works, the authors investigate robust fusion schemes for spoofing attacks considering face and fingerprint biometric traits.

2.2.5 Problems with the Existing Approaches

Approaches based on clues of the scene have strong constraints that make sense only to photo-based spoofing attacks. In the case of attacks performed by video, such constraints certainly will fail due to the dynamic nature of the scene in this type of media (e.g., motion). The static background assumption made in some works described earlier is limited since the face moves independently of the background in a video-based attempted spoofing attack. Moreover, the assumption of a background previously known restricts the use of the method since in many applications (e.g., web and mobile applications) the data acquisition is performed remotely in an environment and, therefore, we can not assume a previously known background. Finally, we can easily change the background of an image through image manipulation.

In approaches based on optical flow and motion analysis, motion is easily simulated by rotating or bending the photographs. Moreover, such methods should be evaluated by considering video-based attempted spoofing attacks since these media carries motion information and, therefore, has potential to deceive such methods. Another disadvantage of approaches based on motion analysis is that the additional time required to capture some face motions prevents a fast spoofing detection. For example, a type of motion analyses extensively explored in the literature is the action of eye blink that occurs once every four or six seconds. However, this rate can be reduced to an average of three to eight every six seconds due to psychological factors [140]. In this case, at least 20 seconds are required to detect eye blinking.

Finally, methods based on texture analysis should consider attempted attacks performed with high resolution videos. Photo-based spoofing attacks have a characteristic that facilitates the detection of this type of attack, which is absent in video-based spoofing attacks: the decrease of quality of the biometric sample due to the printing process, since printers have limitations both in terms of resolution and number of colors that can be produced, which directly influence the texture of the biometric sample, being easily captured by texture information.

Finally, the method proposed in this work aims at overcoming such difficulties by capturing acquisition-related noise information features generated by the video recapture. As the noise signal is independent of the image signal, our method explores this fact by

isolating the noise, so it tends to be less dependent of the video content. Furthermore, our method requires only 50 frames ($\approx$ 2 seconds) for detecting an attempted attack.

2.3 Proposed Method

Here, we present an algorithm for video-based attempted spoofing attack detection. Our solution relies on the fact that the addition of a noise pattern in the samples is inevitable during the acquisition step of the facial biometric samples.

The acquisition process is performed by a camera that has an imaging sensor with thousands of photosensitive transducers that convert light energy into electrical charges, which are converted into a digital signal by an A/D converter. In [150], Lukäs et al. define two types of noise that can be present in an image: the fixed pattern noise (FPN) and the noise resulting from the photo-responsiveness of non-uniform light-sensitive cells (PRNU). The noise pattern has been widely explored in digital document forensics as in the problem of identifying the specific camera that acquired a document [150, 215].

During a video-based spoofing attack, we have the insertion of artifacts in the biometric samples captured by the acquisition sensor, such as distortions, flickering, mooring, and banding effect [14]. Such artifacts, loosely referenced in this paper as *noise*, are added during the generation and viewing process of the attack video frames in display device screens. Thus, the biometric sample extracted from an attack video will probably contain more noise than the real biometric samples. With this in mind, we design a feature characterization process based on noise signatures along with video summarization methods that are used by a classification algorithm to find a separation decision boundary between real and fake biometric data. Fig. 2.2 summarizes the steps of the proposed method, which are explained in detail in the following sections.

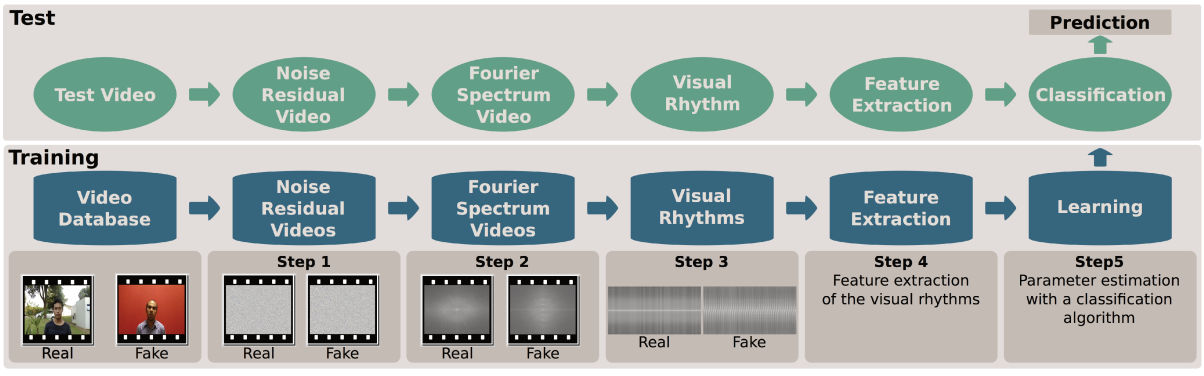


Figure 2.2: Proposed method. Given a training set consisting of videos of valid accesses, video-based spoofs and a test video, we first extract a noise signature of every video (training and testing) and calculate the Fourier Spectrum on logarithmic scale for each video frame and then summarize each video by means of its visual rhythm. Considering the training samples, we train a classifier using a summarized version of the visual rhythms obtained by the estimation of the gray level co-occurrence matrices, as features. With a trained classifier, we are able to test a visual rhythm for a given video under investigation and point out whether it is a valid access or a spoof.

2.3.1 Calculation of the Residual Noise Videos

The first step of the algorithm is to isolate the noise information contained in the videos that were captured by the acquisition sensor, hereinafter referred to as input video ν. A video ν in the domain $2D + t$ can be defined as a sequence of t frames, where each frame is a function $f(x, y) \in \mathbb{N}^2$ of the brightness of each pixel in the position (x, y) of the scene.

The extraction of the noise signal of the input video ν is performed as follows. The frames in video ν are converted into gray-scale and an instance of ν_{Gray} is submitted to a filtering process using a low-pass filter in order to eliminate noise, generating a filtered video $\nu_{Filtered}$. Then, a frame-by-frame subtraction between the ν_{Gray} and $\nu_{Filtered}$ is performed, generating a new video that contains, mostly, the noise signal in which we are interested, hereinafter named as Residual Noise Video (ν_{NR}), as formalized in Equation 2.1.

$$\begin{cases} \nu^{(t)}_{Filtered} = f(\nu^{(t)}_{Gray}) \\ \\ \nu^{(t)}_{NR} = \nu^{(t)}_{Gray} - \nu^{(t)}_{Filtered} \quad \forall\ t\ \in T = \{1, 2, \ldots, t\}, \end{cases} \tag{2.1}$$

where $\nu^{(t)} \in \mathbb{N}^2$ is the t-th frame of ν and f a filtering operation.

2.3.2 Calculation of the Fourier Spectrum Videos

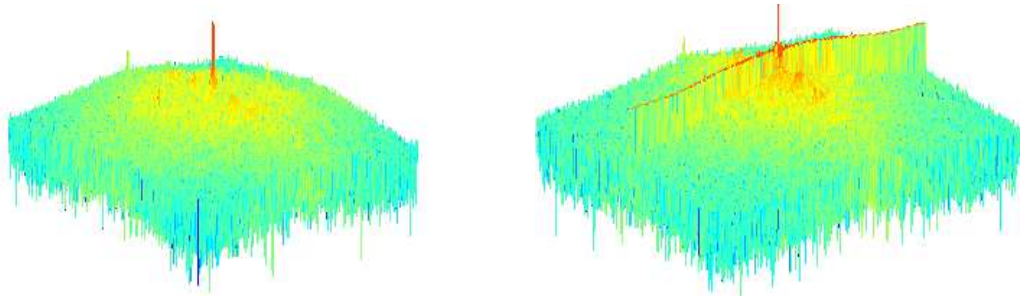
The analysis of the noise pattern and possible artifacts contained in the biometric samples is performed by applying a 2D discrete Fourier transform to each frame of the Noise Residual Video (ν_{NR}) using Equation 2.2. Next, the Fourier spectrum is computed on logarithmic scale and with origin at the center of the frame (Equation 2.3). As a result of this process, we end up with a video of the spectra, further on in this document referred to as Fourier Spectrum Videos ν_{FS}. Fig. 2.3(a) and 2.3(b) depict the logarithm of the Fourier spectrum of a video frame obtained from a valid access video and from an attempted attack video, respectively.

$$\mathcal{F}(v, u) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \nu_{NR}(x, y) e^{-j2\pi[(vx/M)+(uy/N)]} \tag{2.2}$$

$$\begin{aligned} |\mathcal{F}(v, u)| &= \sqrt{\mathcal{R}(v, u)^2 + \mathcal{I}(v, u)^2} \\ \nu_{FS}(v, u) &= \log(1 + |\mathcal{F}(v, u)|) \end{aligned} \tag{2.3}$$

2.3.3 Calculation of the Visual Rhythms

In order to capture the temporal information contained in the Fourier Spectrum Videos (ν_{FS}) and summarize their content, we employ the visual rhythm technique [45]. Visual rhythm is a simplification of a video content in a 2D image obtained by sampling regions of the video. Applications of this concept can be found in the work by Chun et al. [44]



(a) Valid video. (b) Attack video.

Figure 2.3: Example of a video frame of the spectra generated from (a) a valid video and (b) an attack video.

that use visual rhythms for fast text caption localization on video, and Guimarães et al. [89] who propose a method for gradual transition detection in videos. The use of visual rhythm in our work is crucial since it allows us to capture patterns that are present in the Fourier Spectrum Videos providing an effective way of viewing a video as a still image.

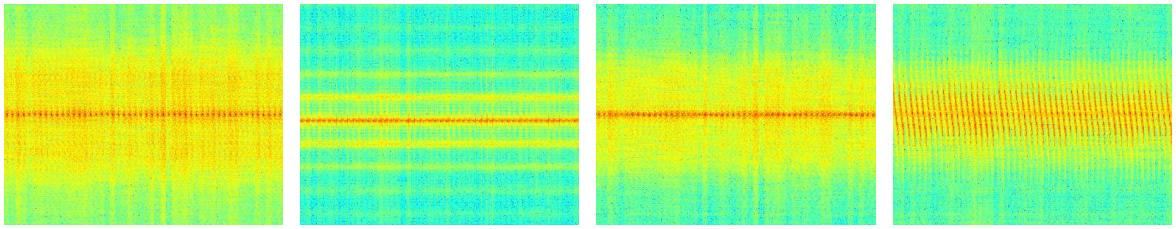
Considering a video ν in the $2D+t$ domain with t frames of dimensions $W \times H$ pixels, the visual rhythm I_{ν_R} is a representation of the video ν, in which regions of interest of each frame are sampled and aggregated to form a new image, called visual rhythm. The regions of interest must be carefully chosen to capture the patterns contained in ν_{FS}. Formally, a visual rhythm I_{ν_R} of a video ν can be defined by

$$I_{\nu_R}(z,t) = \nu(x(z), y(z), t), \tag{2.4}$$

where $x(z)$ and $y(z)$ are functions of the independent variable z. The visual rhythm is a two-dimensional image whose vertical z axis consists of a certain group of pixels extracted from video ν and the samples are accumulated along the time t. Therefore, according to the mapping of $x(z)$ and $y(z)$, we can generate several types of visual rhythms [45]. For instance, the sampling of the central vertical pixels can be performed by applying $I_{\nu_R}(z,t) = \nu(x(\frac{W}{2}), y(z), t)$. Similarly, the central horizontal pixels can be extracted by applying $I_{\nu_R}(z,t) = \nu(x(z), y(\frac{H}{2}), t)$.

Given that the lower responses are mainly concentrated at the abscissa and ordinate axes [236] of the Fourier spectrum (see Fig. 2.3), initially we consider two regions of interest in the frames that form the spectrum video in the construction of two types of visual rhythms: (i) the horizontal visual rhythm formed by central horizontal lines and (ii) the vertical visual rhythm formed by central vertical lines. In both cases, we can summarize relevant content of the spectrum video in a single image. Fig. 2.4 depicts the visual rhythms generated by two regions of interest considering a valid (Fig. 2.4(a) and 2.4(c)) and an attack video (Fig. 2.4(b) and 2.4(d)).

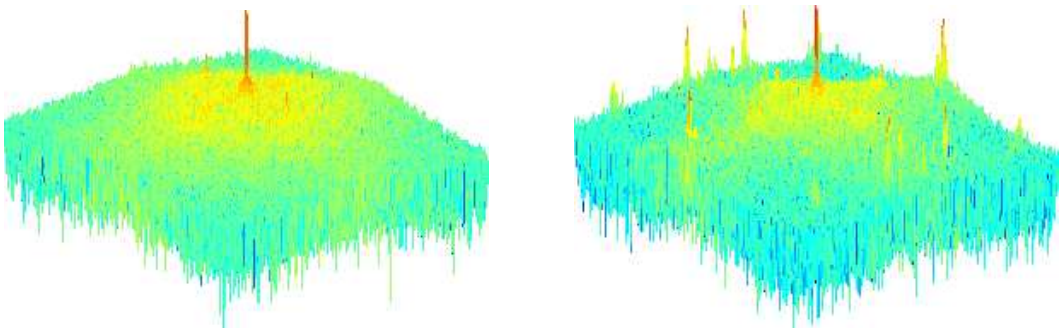
Even though the visual rhythms are different for valid and attack videos, their construction disregards the highest responses that are not at the abscissa and ordinate axes and, in some cases, such information is important to make a better distinction between valid access and attempted attack videos, as shown in Fig. 2.5. With this in mind, we



(a) Valid video. (b) Attack video. (c) Valid video. (d) Attack video.

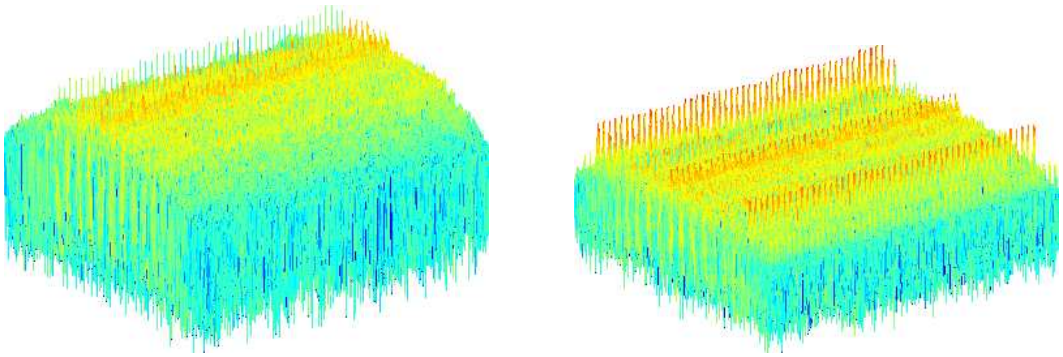
Figure 2.4: Visual rhythms constructed from (a)-(b) central horizontal lines and from (c)-(d) central vertical lines. Note that the visual rhythm obtained from horizontal lines has been rotated 90 degrees for visualization purposes.

extract a third type of visual rhythm by traversing along the frames of Fourier Spectrum Videos (ν_{FS}) in a zig-zag scheme. Fig. 2.6 shows the zig-zag visual rhythm generated for a valid access video and an attempted attack video.



(a) Valid video. (b) Attack video.

Figure 2.5: Examples of spectra whose highest responses are not only at the abscissa and ordinates axes.



(a) Valid video. (b) Attack video.

Figure 2.6: Examples of visual rhythms constructed in a zig-zag traversal.

2.3.4 Feature Extraction

Once the visual rhythms are computed, we can use machine learning techniques to train a classifier to decide whether a biometric sample is fake or not. However, if the intensity

of the pixels composing the visual rhythms are directly considered, the dimensionality of the feature space will be extremely high and most of the traditional classification methods will not work properly. Therefore, we need to extract a compact set of feature descriptors that best discriminate the visual rhythms generated from the fake and valid videos.

In this work, we evaluate the use of three feature descriptors: Gray Level Co-occurrence Matrices (GLCM) [93], Local Binary Patterns (LBP) [168] and Histogram of Oriented Gradients (HOG) [56]. The choice for using GLCM and LBP descriptors is motivated by the fact that the visual rhythms can be interpreted as texture maps (see Fig. 2.4). Moreover, if we consider the intensity values of the pixels of the visual rhythms as height and edge artifacts represented along the maps, we see (Fig. 2.6) that such images have different edge forms, a property that can be reasonably explored by the HOG descriptor.

GLCM It is a structure that describes the frequency of gray level occurrence between pairs of pixels. When normalized, the co-occurrence matrix becomes an estimation of joint probabilities between pairs of pixels at a distance d in a given orientation θ. After calculating the co-occurrence matrix for four different orientations, we extracted 12 measures to summarize the textural information of each matrix: angular second-moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, and directionality.

LBP The LBP operator [168] provides a robust way to describe local binary patterns. Basically, a window of size 3 pixels is thresholded by the value of the central pixel. The pixel values are then multiplied by binomial weights and summed to obtain an LBP number to this window. Thus, LBP can produce up $2^8 = 256$ different texture patterns, and a histogram with 256 bins is calculated and used as a texture descriptor.

HOG The basic idea of this descriptor relies on the fact that the local appearance of the objects and shape can be well characterized by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. Basically, the image is divided into small spatial regions, referred to as cells, and for each cell is calculated a histogram of gradient directions. A set of cells is grouped into a block and the concatenation of the descriptors extracted from each cell followed by a normalization results in the HOG descriptor.

2.3.5 Learning

We evaluate the proposed characterization process using two machine learning techniques: Support Vector Machine (SVM) and Partial Least Square (PLS) that are used in the construction of a binary classifier to decide whether a sample is fake or not.

The SVM algorithm [50] uses either a linear or a non-linear mapping, depending on the type of space used to transform the original data onto a higher dimensional one.

PLS regression method [100] is based on the linear transformation of a large number of descriptors to a new space based on a small number of orthogonal projection vectors. In other words, the projection vectors are mutually independent linear combinations of

the original descriptors. These vectors are chosen to provide maximum correlation with the dependent variables, which are the labels of the training classes.

2.4 Database Creation

This section presents the Unicamp Video-Attack Database (UVAD) specifically built for evaluation of the video-based spoofing attack detection methods. The UVAD contains valid access and attempted attack videos of 404 different identities. All videos were created at Full HD quality, with 30 frames per second and are nine seconds long.

The generation of valid access videos was performed by filming each participant in two sections considering different backgrounds, lighting conditions, and places (indoors and outdoors). As each person is recorded by only one camera, then there is no identity overlap between video from different camera. In total, 808 videos that represent valid accesses were generated with six different cameras: a 9.1 megapixels Sony CyberShot DSC-HX1, a 10.0 megapixels Canon PowerShot SX1 IS, a 10.3 megapixels Nikon Coolpix P100, a 14.0 megapixels Kodak Z981, a 14.0 megapixels Olympus SP 800UZ, and a 12.1 megapixels Panasonic FZ35 digital camera. We used a tripod to avoid disturbance in the videos during the recordings. The generated videos were cropped to maintain a resolution of $1,366 \times 768$ and allow the faces to be positioned at the center of the video frame. No resampling was performed whatsoever.

The attempted attack videos were generated by using the same digital cameras utilized to generate the valid access videos and seven different display devices with a $1,366 \times 768$ pixel resolution. The valid access videos were displayed on seven display devices and recaptured with the same digital cameras used previously. Each display device was positioned in front of each camera at a distance of 90 ± 5cm supported in a tripod, so that to ensure each video with $1,366 \times 768$ resolution after cropping.

As the valid access videos were cropped to maintain a $1,366 \times 768$ resolution, we guarantee that there was no scaling transformations during their exhibition. In total, we have generated $16,268$ attempted attack videos and 808 valid access videos. Fig. 2.7 and 2.8 depict real and fake video frame examples, respectively.

Figure 2.7: Examples of valid access video frames for outdoor (first and second images on the left) and indoor (three images on the right) scenes.

Table 2.1 shows a comparison between the proposed UVAD database and some other reference benchmarks in the literature. The diversity of display devices and acquisition sensors used in the generation of UVAD is an important characteristic that is not found in the other databases, which was essential to a better comprehension of the problem and for a precise evaluation of the methods as we will show in Section 2.5.

Figure 2.8: Examples of attempted attack video frames for outdoor (first and second images on the left) and indoor (three images on the right) scenes using Sony (first and second columns), Canon (third and fourth columns) and Nikon (last column) cameras.

Table 2.1: Comparison of the proposed UVAD database and other available reference benchmarks in the literature.

Database	Number of subjects	Number of valid accesses	Number of attacks by photo	Number of attacks by video	Number of devices used to create the attack videos
NUAA [245]	15	5,105	7,509	–	–
Print-Attack [8]	50	200	200	–	–
CASIA [274]	50	150	300	150	3 cameras and 1 display device
Replay-Attack [41]	50	200	200	800	2 cameras and 2 display devices
UVAD (proposed)	404	808	––	16,268	6 cameras and 7 display devices

2.5 Experimental Results

In this section, we show the details of the experiments and performance evaluations of the developed method. We first consider the UVAD database which was introduced in Section 2.4 (Experiments I-IV). The diversity of devices used allows us to answer important questions regarding some strengths and limitations of the proposed method. In addition, we also evaluate the proposed method with respect to the literature (Experiment V) and through the Replay-Attack Database (c.f., Sec. 2.2.1) (Experiment VI).

2.5.1 Protocols for the UVAD Database

In this section, we define appropriate protocols for each experiment.

Protocol I. The aim of this protocol is at finding the best configuration of the proposed method. In this protocol, we divide the dataset into two sets, hereintofore referred to as training and test sets. During partition, we guarantee that there is no overlap of data from the same capture and display devices between training and test sets, so that we have a proper comparison without experimental bias.

The valid access videos from six cameras were divided into two subsets, A and B. The valid access videos in set A were again divided to form two sets of valid access videos: (i) real training set, composed of videos generated by three cameras chosen arbitrarily (Sony, Canon, and Kodak) and (ii) real test set, composed of videos generated by the remaining three cameras (Nikon, Olympus, and Panasonic).

In sequence, the valid access videos in set B were used to generate two sets of attempted attack videos: (i) the fake training set, in which videos in B generated by the Sony, Canon, and Kodak cameras were displayed on three display devices and recaptured by the same three cameras, and (ii) the fake test set, whose videos in B generated by the Nikon,

Olympus, and Panasonic cameras were displayed on the remaining three display devices and recaptured by the same cameras.

Protocol II. The aim of this protocol is at checking the influence of the biometric sensor on the proposed method. Similarly to the previous protocol, we divide the dataset into two sets, training and test sets. However, we create nine training and test sets, changing the cameras that compose such sets. Again, we guarantee that there is no overlap of data from the same cameras and display devices. Our goal with these partitions is to train a classifier with videos from three cameras and test it with the videos from other three cameras that never were used or seen by the classifier.

Protocol III. The aim of this protocol is at checking the influence of the display devices over the detection method. In this protocol, we divide the videos from each camera into two sets, A and B. Set A contains attempted attacks performed with three display devices and set B comprises attempted attacks performed with the three complementary display devices. The partition considering different display devices for both attack sets was carried out to avoid that a classifier takes biased conclusions regarding videos coming from devices already seen during the training step. The classification results are given in terms of mean of the results obtained in two rounds of experiments by using the set A to train a classifier and B to test it, and vice versa.

2.5.2 Parameters for the Filtering Process, Visual Rhythm Analysis and Classification

To extract signal noise signature of the videos, as Equation 2.1 shows, we consider the use of spatial linear and non-linear filters: a Gaussian filter with $\mu = 0$, $\sigma = 2$, and size 7×7 and a Median filter with size 7×7, respectively. These parameters were obtained empirically in [193] on a different dataset.

After calculating the noise signature using Equations 2.2 and 2.3, we extract the visual rhythms (horizontal and vertical) of each video considering the first 50 frames and a block of either 30 columns (vertical) pixels or 30 lines (horizontal). Since the visual vertical and horizontal rhythms of each video carry different temporal information, we evaluate the two types of visual rhythms along with their combinations. The horizontal visual rhythms (H) are in a dimensional space of $1,366 \times 1,500$-d while the vertical visual rhythms (V) are in $768 \times 1,500$-d. To generate the zig-zag visual rhythms (Z), we also consider the first 50 frames of the Fourier Spectrum transformed videos. We extract block lines of 30 pixels through the traversal of the frames, from left to right, top to bottom. Thus, we obtained visual rhythms that are in a dimensional space of $17,482 \times 1,500$-d.

The high dimensionality and large amount of visual rhythms prevent us from using pixel intensities directly as features. Therefore, we consider the visual rhythms as texture maps and calculate their texture patterns using different characterization methods. For instance, for the standard configuration, we considered the GLCM descriptor with directions $\theta \in \{0^o, 45^o, 90^o, 135^o\}$, distance $d = 1$ and 16 bins. Table 2.2 shows the dimensionality information of each feature. In order to evaluate the robustness of the extracted features, we can use them to train a classifier and generate a model capable of distinguish-

Table 2.2: Number of features (dimensions) using either the direct pixel intensities as features or the features extracted by image description methods.

Name	Descriptor Dimensionality		
	V	H	Z
Pixel Intensity	1,152,000	2,049,000	26,223,000
LBP	256	256	256
GLCM	48	48	48
HOG	36	36	36

ing valid and attack videos, and test the model effectiveness. In this paper, we use two classification techniques: SVM and PLS. For SVM, we use the LibSVM [38] implementation and we analyze the radial basis function kernel, whose parameters were found using LibSVM's built-in grid search algorithm. For PLS, we use the DetectorPLS method [225] and we analyze different numbers of factors. The factors are latent variables that give us the best predictive power and they are extracted from a set of independent variables and are used to predict a set of dependent variables. The interested reader may refer to [225] for more details on factor choices in PLS.

2.5.3 Experiment I: Finding the Best Configuration

The objective here is to find the best configuration of our method and to evaluate the classifiers, visual rhythm setups and filters through the analysis of variance to assess which of these parameters present higher in influence. In addition, we evaluate other important feature characterization methods found in the literature, namely Local Binary Pattern Histogram (LBP) and Histogram of Oriented Gradient (HOG) descriptors. Although we have considered the visual rhythms as texture maps, it is worth analyzing the use of shape descriptors such as HOG as well. With this experiment, it is possible to discover whether considering the visual rhythms as texture maps is the best choice. We carried out these experiments using the *Protocol I* and considering the sets of attacks with videos recaptured by all cameras.

After performing statistical analysis with ANOVA and Tukey's HSD (Honestly Significant Difference) test in the results shown in Table 2.3, the following conclusions can be drawn: (1) GLCM descriptor performance is statistically different from its HOG and LBP counterparts, as shown in Fig. 2.9. As it outperforms the other descriptors with statistical significance, we can conclude that GLCM was able to extract the most discriminative information from the visual rhythms as texture maps better than its counterparts; (2) both Gaussian and Median filters used in this work to generate Noise Residual Videos (ν_{NR}) did not produce statistically different results (figure now shown here); (3) methods for building the visual rhythms did not present results with differences statistically significant (See Fig. 2.10); and (4) with respect to the classification algorithm used in this work, we do not find statistical differences between the use of the SVM and PLS algorithms (figure not shown here). It is noteworthy that both ANOVA and TukeyHSD's tests allow us to reject the hypothesis of equality between comparisons, but not accept

the hypothesis that they are equal, in cases that no statistical differences were found. Therefore, the best configuration considered is the one using Median filter, Horizontal and Vertical visual rhythms combined, GLCM descriptor to extract texture information from visual rhythms, and the PLS classification algorithm.

Table 2.3: Results (AUC) of the experiment in which we find the best configuration of our method considering all possible setups.

		PLS		**SVM**	
Desc.	V. Rhythms	Gaussian	Median	Gaussian	Median
GLCM	V	59.65%	84.33%	68.57%	74.86%
	H	76.27%	86.29%	76.09%	72.55%
	V + H	77.74%	**91.43**%	74.90%	65.28%
	Z	90.92%	80.23%	83.22%	63.59%
LBP	V	61.21%	72.29%	56.06%	65.95%
	H	62.75%	63.55%	70.02%	67.76%
	V + H	64.61%	70.81%	70.97%	73.44%
	Z	67.44%	55.70%	64.65%	57.36%
HOG	V	68.75%	54.68%	67.86%	67.90%
	H	54.68%	64.76%	50.61%	66.88%
	V + H	57.73%	73.72%	66.96%	73.54%
	Z	65.54%	65.54%	52.35%	52.35%

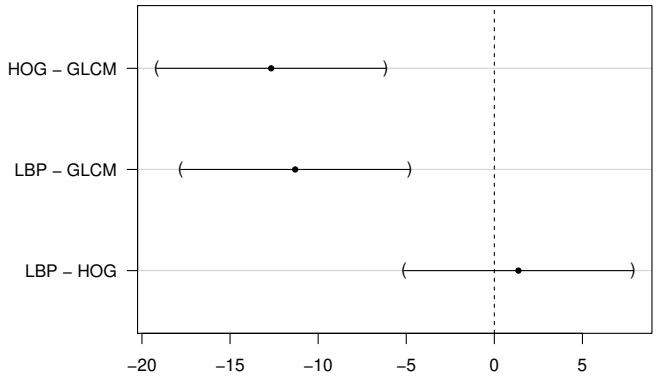


Figure 2.9: Differences in mean levels of the results obtained by the different descriptors used in this work and their confidence intervals for 95% family-wise confidence level. There are statistical difference between the comparisons whose confidence intervals do not include zero.

2.5.4 Experiment II: Influence of the Biometric Sensors

This experiment aims at checking whether the presented method works well in different facial biometric systems (biometric sensors). Experiments performed with only one kind of biometric sensor does not guarantee a broad evaluation of our method. Although

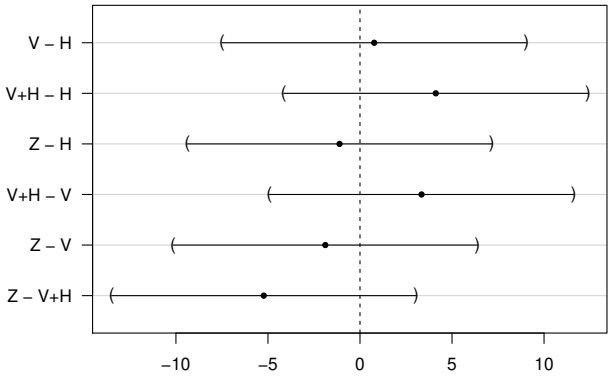


Figure 2.10: Differences in mean levels of the results obtained by the visual rhythms considered in this work and their confidence intervals for 95% family-wise confidence level. There are statistical difference between the comparisons whose confidence intervals do not include zero.

this is not a common practice in the literature, we believe that experiments with several biometric sensors is an essential practice to evaluate countermeasure methods, because the artifact levels inserted into the biometric samples depend, among other factors, on the quality of the acquisition sensor. Using the *Protocol II*, we evaluate the proposed method in its best configuration (see Table 2.4).

Table 2.4: Results (AUC) of the experiment analyzing the influence of the biometric sensors using a PLS Classifier and Median Filter.

Training	Sony Canon Kodak	Sony Canon Panasonic	Sony Canon Olympus	Sony Kodak Panasonic	Sony Kodak Olympus	Sony Olympus Panasonic	Canon Olympus Panasonic	Canon Olympus Kodak	Canon Kodak Panasonic
Test	Nikon Olympus Panasonic	Nikon Olympus Kodak	Nikon Kodak Panasonic	Nikon Canon Olympus	Nikon Canon Panasonic	Canon Kodak Nikon	Sony Kodak Nikon	Sony Nikon Panasonic	Sony Nikon Olympus
AUC	91.43%	90.48%	86.89%	89.66%	96.12%	91.85%	81.07%	86.84%	84.25%

When we vary the cameras used in the training, we have a variation in the method generalization. For instance, considering the best and worst results shown in the Table 2.4, we have a relative error reduction of 79.50%. Though it is evident the in influence of the biometric sensor with this variation in the classification results, we performed the Wilcoxon Signed Rank test to prove this influence, with which we obtained a p-value of 0.0039 and hence confirmation that the values shown in Table 2.4 are indeed statistically different.

2.5.5 Experiment III: Influence of the Display Devices

The aim of this experiment is to check whether the presented method is able to detect attacks with different display devices, that is, whether the display devices produce different amounts of display artifacts (the main artifacts produced are flickering, mooring and banding effect). This is an important question to be answered because if the method is not robust to different devices, learning techniques considering an open scenario could be

considered [222], given that in this case the classifier should be able to recognize attacks with display devices for which it has no prior knowledge.

Considering *Protocol III*, this experiment was performed in two rounds: firstly, we train a classifier with attacks performed with three display devices and tested it with the other three display devices to evaluate the model found by the classifier. Secondly, we switch the sets and redo the analysis. In both cases, we considered the best configuration of our method. The results reported in Table 2.5 correspond to the average ($\overline{x}$) and stdev (s) of the results obtained in the two rounds for each configuration of the method.

Table 2.5: Results (AUC) of the experiment analyzing the influence of the display devices using a PLS Classifier and Median Filter.

Sony	**Canon**	**Nikon**	**Kodak**	**Olympus**	**Panasonic**
✓	✗	✗	✓	✗	✓
$p\text{–}value = 0.0$	$p\text{–}value = 1.0$	$p\text{–}value = 1.0$	$p\text{–}value = 0.0$	$p\text{–}value = 0.574$	$p\text{–}value = 0.015$
$\overline{x} = 92.70\%$	$\overline{x} = 99.34\%$	$\overline{x} = 98.61\%$	$\overline{x} = 96.42\%$	$\overline{x} = 84.57\%$	$\overline{x} = 97.53\%$
$s = 0.23\%$	$s = 0.91\%$	$s = 1.36\%$	$s = 0.76\%$	$s = 14.33\%$	$s = 2.81\%$

The influence of the display devices are evidenced when the results obtained in the two rounds of experiments are discrepant or whether they are statistically different, indicating that the method was not able to detect attempted attacks performed with unknown display devices. To verify whether the differences in the results are statistically significant, we carried out a hypothesis test for two unpaired or independent samples. Once the sample values are nominal, the most appropriate statistical test is χ^2 test for two samples whose values are also shown in all tables, considering a confidence level of 95%. The p-value produced for the χ^2 tests evaluate whether two samples are statistically different (p-value < 0.05). According to results shown in Table 2.5, we have obtained a p-value lower than $\alpha = 0.05$, for some cameras. In these cases, the differences were statistically significant, which leads us to the conclusion that the display device plays an important role in the spoofing detection task.

2.5.6 Experiment IV: Comparison to a State-of-the-Art Method for Photo-Based Spoofing Attack Detection

In the final round of experiments concerning the UVAD database, we compare our method to the one proposed in [214]. We considered the *Protocol I* to compare both methods. It was not possible to run the algorithm by Schwartz et al. by using the same parameters described in [214] due to the high dimensionality of the data their method produces, even on a machine with 48GB of RAM. The dimensionality of the feature vector generated by the original algorithm is higher than five million dimensions for each video frame.

In order to reduce the dimensionality of the feature vectors, we applied the HOG descriptor with blocks of sizes 16×16 and 32×32 with strides of 16 and 32 pixels, respectively. The other parameters were set as described in [214]. With this, we were able to reduce the feature vector dimensionality to $8,880$ dimensions. Table 2.6 shows the results obtained by using the algorithm in [214] and our method, considering the configuration that yielded the lowest classification error. Furthermore, the computational

time spent by the algorithm in [214] was ≈ 237 hours to process all the data, whereas the method proposed in this work spent ≈ 72 hours. According to McNemar statistical test, the result obtained by the methods are statistically different. All experiments were conducted on an Intel Xeon E5620, 2.4GHz quad core processor with 48GB of RAM under Linux operating system.

With this experiment, we can conclude that our method better characterized video-based attacks while being more efficient and suitable for different classification techniques, once it provides more compact feature representations.

Table 2.6: Comparison between Schwartz's approach and the method proposed in this work in its best setup (using combined visual rhythm, Median filter and a PLS Classifier).

	AUC (%)
Schwartz et al. [214]	90.52%
Our method	**91.43**%
Error Reduction	9.60%

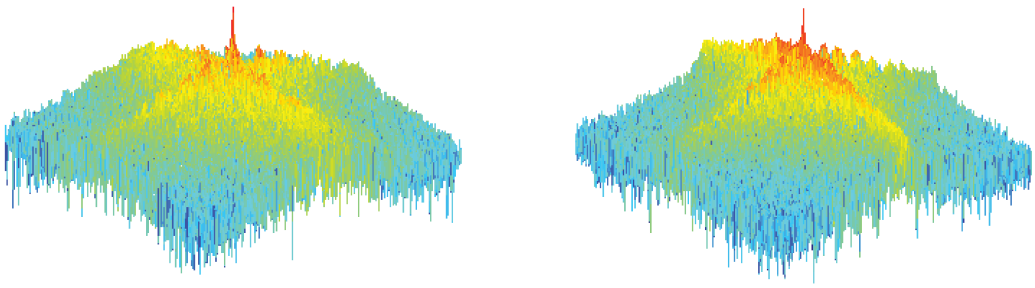
2.5.7 Experiment V: Evaluation of the Method in the Replay-Attack Database

In this experiment, we evaluate our method on the Replay-Attack database (c.f., 2.2.1) which contains photo-based and video-based spoofing attacks. The goal of this experiment is to verify the effectiveness of our method on these several types of attacks. We use the experimental protocol described in [41], whose results are shown in Table 2.7. Although our method is designed for video-based spoofing attack detection, we have obtained a promising AUC of ≈ 93%. For reference, in [41], the authors reported a Half Total Error Rate (HTER) of 34.01% and 15.16%, using a χ^2 and SVM classifier, respectively, to classify $\text{LBP}^{u2}_{3\times2}$ features, while our method yields an HTER of 14.27%. We use a Gaussian filter with $\mu = 0$, $\sigma = 0.5$ and size 3×3, and a Median filter with size 3×3. These parameters were empirically obtained by using the Replay-Attack Database. With this experiment, we can conclude that the proposed method is able not only to detect video-based spoof attacks but also video print-attacks.

Finally, one can notice that, in particular, the zig-zag characterization method does not lead to the best result in this dataset. We believe the reason is that the Replay-Attack [41] is a dataset based on print photograph recaptures (still image attacks) which, when recaptured, tend to concentrate visual information in the center of the Fourier transformed domain as depicted in Fig. 2.11. This tends to favor the vertical and horizontal visual rhythms as they concentrate on these areas. The contrary happens with video attacks since the peaks in the Fourier transformed domain will be more spatially spread over each frame, as shown in Fig. 2.5. Result obtained by the TukeyHSD' test confirm that difference between V+H and Z visual rhythms are statistically significant (p-value=0.03).

Table 2.7: Results (AUC) for the test set of the Replay-Attack database

Visual Rhythms	PLS classifier		SVM classifier	
	Median	Gaussian	Median	Gaussian
V	83.99%	89.01%	86.26%	91.56%
H	81.98%	85.66%	80.67%	73.36%
V + H	90.69%	92.98%	92.01%	91.81%
Z	78.39%	85.35%	86.56%	77.72%



(a) (b)

Figure 2.11: Example of a video frame of the spectra generated from (a) a valid access video of the Replay-Attack database and (b) a video of an attempted attack of the same dataset. Note a concentration of information on the center rather than spread over as for the videos case shown in Fig. 2.5.

2.6 Conclusions and Future Work

Biometric authentication systems have been shown to be vulnerable to spoofing attacks in the sense that impostors can gain access privileges to resources as valid users. Spoofing attacks to a face recognition system can be performed by presenting it a photograph, a video, or a face mask of a legitimate user.

This paper proposed and evaluated a spatio-temporal method for video-based face spoofing detection through the analysis of noise signatures generated by the video acquisition process, which can be used to distinguish between valid and fake access videos. Noise properties are captured using Fourier spectrum for each frame of the video. A compact representation, called visual rhythm, is employed to detect temporal information in the Fourier spectrum. Three different video traversal strategies were considered to form the visual rhythms, of which horizontal and vertical combined was shown to be the most effective. Features were extracted from the visual rhythms through GLCM, LBP and HOG descriptors to allow a proper distinction between fake and real biometric data. The GLCM method was shown to be the most discriminative and compact feature representation for visual rhythm description.

An extensive data set, containing real access and spoofing attack videos, was created to evaluate the proposed method, as well as the state-of-the-art approaches. Through the conducted experiments, it is possible to conclude that the display devices and biometric sensors play an important role in the spoofing detection task. These findings are very important in making the future anti-spoofing methods more effective and guiding the

development of new databases which must be more realistic, as the UVAD Database proposed in this paper. The proposed anti-spoofing method provided competitive or even superior results in the tests when compared to state-of-the-art approaches.

Although this paper represents a step toward solving the spoofing problem, it makes it clear that the problem is not fully-solved yet and poses new questions on future methods regarding how to better handle and tackle with new attacks due to the ever-growing market of acquisition and display devices such as hight quality monitors, hand-held and smartphone devices. In this sense, the dataset provided in this paper will be available at the IEEE Information Forensics and Security Technical Committee website (http://tinyurl.com/pas4t9r) and also registered with a proper DOI through FigShare (http://figshare.com/) in order to advance the frontier of research in spoofing detection.

Future research efforts branch out into devising other spatio-temporal descriptors that capture motion telltales associated with the recapture process as well as verifying other liveness detection problems other than face recognition such as video recapturing, piracy detection, among others [22].

Acknowledgments

We thank the support of CAPES through the DeepEyes project, CNPq (#304352/2012-8 and #477662/2013-7), FAPESP (#2010/05647-4 and #2011/22749-8), FAPEMIG and Microsoft for the support.

Chapter 3

Face Spoofing Detection Through Visual Codebooks of Spectral Temporal Cubes

"Profound study of nature is the most fertile source of mathematical discoveries."

—Joseph Fourier, *French mathematician and physicist (1768–1830)*

"If you want to find the secrets of the universe, think in terms of energy, frequency and vibration."

—Nikola Tesla, *Serbian-American inventor, electrical engineer, physicist, etc. (1856–1943)*

Abstract

Despite important recent advances, the vulnerability of biometric systems to spoofing attacks is still an open problem. Spoof attacks occur when impostor users present synthetic biometric samples of a valid user to the biometric system seeking to deceive it. Considering the case of face biometrics, a spoofing attack consists in presenting a fake sample (e.g., photograph, digital video or even a 3D mask) to the acquisition sensor with the facial information of a valid user. In this paper, we introduce a low-cost and software-based method for detecting spoofing attempts in face recognition systems. Our hypothesis is that during acquisition there will be inevitable artifacts left behind in the recaptured

Pre-print of article that will appear in T-IP, vol.24, no.12, pp.4726-4740, Dec. 2015.
The published article is available on http://dx.doi.org/10.1109/TIP.2015.2466088
See permission to use the copyrighted material in **Appendix C**.

biometric samples allowing us to create a discriminative signature of the video generated by the biometric sensor. To characterize these artifacts, we extract time-spectral feature descriptors from the video, which can be understood as a low-level feature descriptor that gathers temporal and spectral information across the biometric sample and use the visual codebook concept to find mid-level feature descriptors computed from the low-level ones. Such descriptors are more robust for detecting several kinds of attacks than low-level ones. Experimental results show the effectiveness of the proposed method for detecting different types of attacks in a variety of scenarios and datasets including photos, videos and 3D masks.

3.1 Introduction

NOWADAYS, the protection of personal data has become a fundamental requirement of security. According to Tipton [250], information security is concerned with the development of methods and tools for protecting information and preserving the value it has for an individual or an organization. For an efficient and effective protection, the use of robust authentication mechanisms is paramount.

Knowledge-based methods (e.g., password, secret question) and token-based methods (e.g., smart cards, token code) are probably the most used authentication mechanisms to date. However, both methods have a critical feature: at the time of authentication, the system does not verify who is requesting access, but rather what the users know or possess. This renders the system vulnerable, since that knowledge or an object can easily be lost, shared or manipulated. As an alternative, biometrics is an authentication mechanism considered more natural and reliable as it focuses on verifying who is the person requesting the access [109]. Biometrics provides methods for recognizing humans automatically based on behavior, physical or chemical traits, being fingerprint, face, iris, hand geometry, hand vein, voice and DNA, the most common traits used [109].

Although there are several traits that can be used to perform user authentication, researchers are constantly looking for biometric traits with low acquisition and storage costs, that are less invasive, present a high degree of uniqueness and are stable. However, the static nature of a stable biometric trait suggests "the paradox of secure biometrics" [167]:

> *"An authenticator must be stable and distinctive to be considered a good authenticator. But, stability leaves no option for compromise recovery, since users cannot change their biometric trait if stolen. Moreover, since a biometric clue is not secret, its information can be learned and copied."*

Although a stable biometric trait is an ideal authenticator, in practice, its use would not work if it were learned or copied. Therefore, researchers have striven to develop methods that detect whether a biometric sample presented to the acquisition sensor is a replica of the original sample. In the literature, the action of presenting a synthetic biometric sample of some valid user to the acquisition sensor in order to authenticate itself as a legitimate user is known as spoofing attack.

Among several forms of biometric, face recognition is of paramount importance with outstanding solutions presented thus far such as deformable models [262], texture-based representations [2], and shape-based representations [145]. Although effective in many cases, according to Maltoni et al. [153], face, signature and voice are the easiest biometric signals to be circumvented. For instance, spoofing attacks can be successfully accomplished in a face biometric system if an impostor obtains access by presenting to the acquisition sensor a photography, digital video or a 3D model of the target person [109]. Even with recent advances in biometrics, information forensics and security, the vulnerability of facial biometric systems against spoofing attacks is still an open problem.

During the production of the synthetic biometric data, inevitably, there are noise information and telltales added to the biometric signal that can be captured and further processed to pinpoint attacks. In fact, in the manufacturing process of a synthetic sample, there are, at least, two re-quantization steps of the original biometric signal. In photo- and mask-based face spoofing attacks, the continuous signal is quantized during the digitization process. Then, this digital version is re-quantized due to the printing process with 2D and 3D printers and again digitized during the presentation of the synthetic data to the acquisition sensor. In video-based face spoofing attacks, the continuous signal is digitized and recaptured by the acquisition sensor during the attack.

Recent works [151, 193, 245] show that noise and artifacts such as blurring effects, printing artifacts, banding effects, and Moiré patterns are added to the synthetic biometric samples during their manufacture and recapture. In this paper, we propose a spatio-temporal algorithm that captures such effects along time to provide an effective discriminative signature for valid access and spoofing attempts. In summary, the main contributions of this paper are:

- a new method for extracting temporal and spectral information from face biometric samples, referred to as time-spectral descriptors;
- evaluation of the visual codebook model, also referred to as Bag-of-Visual-Word model, for creating a mid-level representation from time-spectral descriptors, referred to as time-spectral visual words; and
- a low-cost solution for spoofing detection, illustrated in Figure 3.1, that does not rely on the user interaction or on extra hardware (e.g., infrared, motion or depth sensors) to detect different types of synthetic samples or attacks (e.g., photos, videos and masks) and is amenable to be implemented in computational devices such as PCs, handheld, and embedded systems.

We organize the remaining of this paper as follows. Section 3.2 discusses state-of-the-art methods for face spoofing attack detection. Section 3.3 presents our method for spoofing attack detection. Section 3.4 shows and discusses the experimental protocol and the obtained results. Finally, Section 3.5 concludes the paper and discusses possible future work.

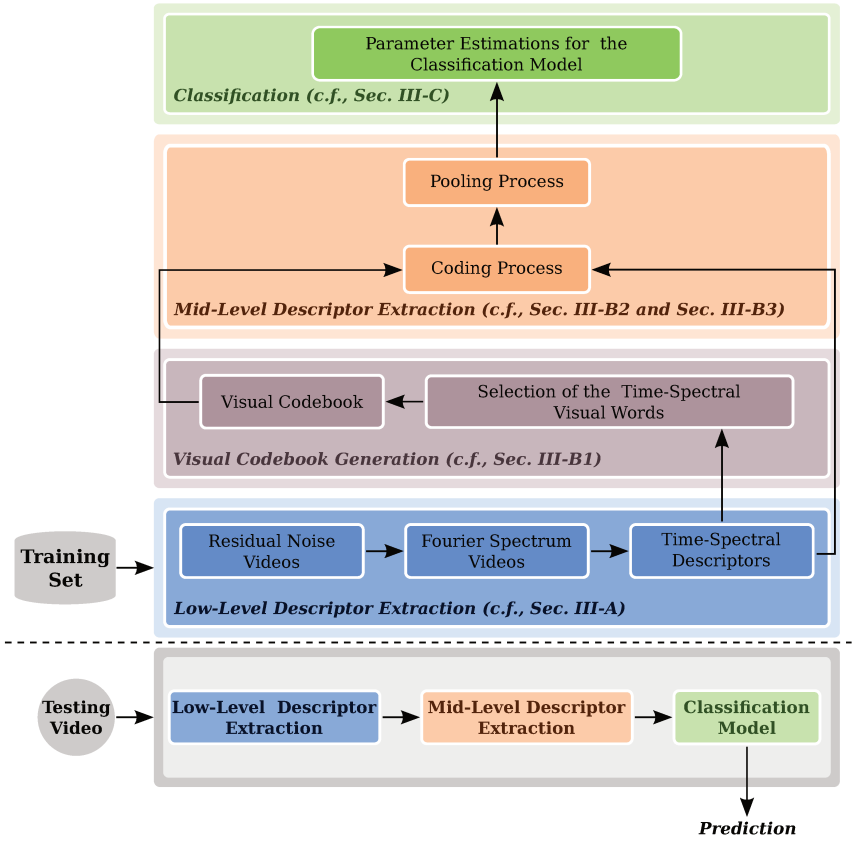


Figure 3.1: Main steps of the proposed method. Given a training set consisting of valid access and attempted attack videos, and also a testing video, we first extract a noise signature from every training video, generating a residual noise video, and calculate its spectrum video. Then, we extract time-spectral descriptors from spectrum videos (low-level representation), which are used to generate a visual codebook. With the visual codebook at hand, we transform the low-level descriptors in time-spectral visual word descriptors (mid-level representation). Finally, these mid-level descriptors are used to find parameters of the classification model, which are employed to predict whether a given testing video is an attempted attack.

3.2 Related Work

The existing techniques for detecting spoofing on face recognition methods can be roughly categorized into four groups: user behavior modeling, user cooperation, methods that require additional hardware and methods based on data-driven characterization. The first aims at modeling the user behavior with respect to the acquisition sensor (e.g., eye blinking or small head and face movements) to decide whether a captured biometric sample is synthetic. Methods based on user cooperation can be used to detect spoofing by means of challenge questions or by asking the user to perform specific movements, which adds extra time and removes the naturalness inherent to facial recognition systems. Techniques that require extra hardware (e.g., infrared cameras or motion and depth sensors) use the additional information generated by these sensors to detect possible clues of an attempted attack. Finally, methods based on data-driven characterization exploit only the data captured by the acquisition sensor looking for evidence and artifacts that may reveal an attempted attack.

In [140,176,263], the authors proposed a solution for detecting photo-based attacks by eye blinking modeling under the assumption that an attempted attack with photographs differs from valid access by the absence of movements. Bao et al. [13] and Kollreider et al. [123] proposed a method based on the analysis of the characteristics of the optical flow field generated for living faces and photo-based attacks. As a living face is a 3D object and a photograph is a planar object, these methods analyze sequential images to detect facial movements, facial expressions or parts of the face such as mouth and eye. Pan et al. [177] extended upon [176] including contextual information of the scene (clues outside of the face) and eye blinking (clues inside the face region).

Methods that use extra hardware have also been considered in the literature. Sun et al. [237] proposed a solution based on thermal IR spectrum modeling the face in the cross-modality of thermal IR and visible light spectrum by canonical correlation analysis. Recently, Erdogmus et al. [66] evaluated the behavior of a face biometric system protected with anti-spoofing solutions [42,151] and the Microsoft's Kinect under attempted attacks performed with static 3D masks. Although these approaches were successful, techniques requiring extra hardware devices have the disadvantage of not being possible to implement in computational devices that do not support them, such as smartphones and tablets.

Turning our attention to the data-driven characterization methods, we can identify three different approaches explored in the literature: methods based on frequency analysis [139, 141, 193], texture analysis [117, 125, 151, 152, 181, 214, 245], and the ones based on motion and clues of the scene analysis [8, 41, 253, 267, 274]. We shall briefly review these approaches in the next sections. For further reading on the problem, we recommend Galbally et al.'s survey [73] and Marcel et al.'s handbook [158].

3.2.1 Frequency-based approaches

Li et al. [141] explored the fact that faces in photographs are smaller than the real ones and that the expressions and poses of the faces in the photographs are invariant to devise a method for detecting photo-based attempted attacks.

Pinto et al. [193] proposed a method for detecting attacks performed with videos using visual rhythm analysis. According to the authors, in a video-based spoofing attack, a noise signature is added to the biometric samples during the recapture of the videos of attacks. The authors isolated the noise signal using a low-pass filter and used the visual rhythm technique to capture the temporal information of the video.

Lee et al. [139] proposed a method based on the frequency entropy of image sequences. The authors used a face verification algorithm to find the face region, normalized the RGB channels using z-score technique, and applied the independent components analysis (ICA) method to remove cross-channel noise caused by interference from the environment. Finally, the authors calculated the power spectrum and analyzed the entropy of the channels individually. Based on a threshold, the authors decide whether a biometric sample is synthetic or real.

3.2.2 Texture-based approaches

Tan et al. [245] proposed a solution for detecting attacks with printed photographs motivated by the difference of the surface roughness of an attempted attack and a real face. The authors estimate the luminance and reflectance of the image under analysis and classify them using Sparse Low Rank Bilinear Logistic Regression methods. Their work was further extended by Peixoto et al. [181] by incorporating measures for different illumination conditions.

Määttä et al. [151] explored micro textures for spoofing detection through the Local Binary Pattern (LBP). To find a holistic representation of the face, able to reveal an attempted attack, Schwartz et al. [214] proposed a method that extracts different information from images (e.g., color, texture and shape of the face). Results of both techniques were reported in the Competition on Counter Measures to 2D Facial Spoofing Attacks [36], with an HTER of 0.00% and 0.63%, respectively, upon the Print Attack Database [8].

Chingovska et al. [41] investigated the use of different variations of the LBP operator used in [151], such as $\text{LBP}^{u2}_{3\times3}$, tLBP , dLBP and mLBP. The histograms generated from these descriptors were classified using χ^2 histogram comparison, Linear Discriminant Analysis and Support Vector Machine.

Face spoofing attacks performed with static masks have also been considered in the literature. Erdogmus et al. [65] explored a database with six types of attacks using facial information of four subjects. To detect attempted attacks, the authors used two algorithms based on Gabor wavelet [262, 273] with a Gabor-phase based similarity measure [90].

Similarly to Tan et al. [245], Kose et al. [128] evaluated a solution based on reflectance to detect attacks performed with masks. To decompose the images into components of illumination and reflectance, the Variational Retinex [5] algorithm was applied.

Pereira et al. [184] proposed a score-level fusion strategy for detecting various types of attacks. The authors trained classifiers using different databases and used the Q statistic to evaluate the dependency between classifiers. In a follow-up work, Pereira et al. [69] proposed an anti-spoofing solution based on the dynamic texture, a spatio-temporal version of the original LBP. Results showed that LBP-based dynamic texture description

has a higher effectiveness than the original LBP, which reinforces the idea that temporal information is of prime importance to detect spoofing attacks.

3.2.3 Motion-based approaches

Tronci et al. [253] explored the motion information and clues that are extracted from the scene by combining two types of processes, referred to as static and video-based analysis. The static analysis consists in combining different visual features such as color, edge, and Gabor textures, whereas the video-based analysis combines simple motion-related measures such as eye blink, mouth movement, and facial expression change.

Anjos et al. [8] proposed a method for detecting photo-based attacks assuming a stationary facial recognition system. According to the authors, the intensity of the relative motion between the face region and the background can be used as a clue to distinguish valid access of attempted attacks, since that motion variations between face and background regions exhibit greater correlation in the case of attempted attacks.

In contrast with the methods described in this section, we present in this work a new anti-spoofing solution based on a temporal characterization of the frequency components from the noise signal extracted from videos. Furthermore, to the best of our knowledge, this was the first attempt of dealing with visual codebooks to find a mid-level representation useful for face spoofing attack detection.

3.3 Proposed Method

In this section, we introduce a method for detecting different forms of face spoofing attacks. The method comprises three main steps: *low-level descriptor extraction*, *mid-level descriptor extraction*, and *classification*. Fig. 3.1 illustrates these steps, which we explain in details in the following sections.

We designed the algorithm based on the fact that synthetic biometric samples contain noise and artifacts generated during their manufacture and recapture that are different from any pattern found in real biometric samples. According to Tan et al. [245] and Määttä et al. [151], there is a deterioration of the facial information and, consequently, a loss of some high frequency components during the manufacture of photographs to be used in spoofing attacks. In our prior work [193], we highlighted the fact that there is a significant increase of the low frequency components due to the blurring effect added during the recapture process of the biometric sample displayed in tablets, smartphones and laptop screens. Besides the blurring effect, other artifacts are added such as flickering, Moiré patterns, and banding effect [14].

These facts motivated us to propose a solution that takes advantage of the noise and artifacts contained on such fake biometric samples, which heretofore we refer to as a noise signature. We perform a Fourier analysis of the noise signature to capture the information encoded in the frequency, phase and amplitude of the component sinusoids [236]. In this paper, we use Fourier spectrum to quantify the following artifacts:

- *blurring artifact:* In both the production and recapture processes, inevitably we have a decrease in the details of biometric samples due to re-quantization of the

original signal. This reduction of details is reflected in the increase of low frequency components and can be observed in the Fourier domain;

- *flickering effect:* It corresponds to the horizontal and vertical lines equally spaced that appear during the recapture process of the samples shown to the acquisition sensor with the display device. When this artifact appears in biometric samples, there are peak lines at abscissa and ordinate axes of the Fourier spectrum when the display device is aligned with the acquisition sensor;

- *Moiré pattern:* They are irregular patterns that can appear when a display device is used to perform an attempted attack. As a result, we also have the appearance of peaks in different locations in the Fourier spectrum depending on the frequency and direction of the sinusoid in the spatial domain [236].

The novelty of our solution is in the two-tier low and mid-level characterization scheme, called time-spectral visual words, that captures patterns present in such noise signatures useful to reveal spoofing attacks. For this, we extract temporal-spectral descriptors from the noise signature transformed to the frequency domain and create a mid-level representation for them using the concept of visual codebooks [12, 235]. Visual codebooks are a method for constructing mid-level representations widely employed in several applications in pattern recognition and computer vision, such as object recognition [254], gesture recognition [96], and information retrieval [182], among others. However, unlike existing methods, we obtain visual informative features from the noise signature present in the videos instead of their raw pixels or from objects in the scene.

3.3.1 Low-Level Descriptor Extraction

In our previous work [193], we found that the noise signal is an important source for low-level discriminative features for spoofing detection. When working with the noise signal and discarding the video content, we minimize possible negative impacts on the method performance. Next, we present the steps of the proposed method to compute the low-level descriptors.

Calculation of the Residual Noise Videos

The low-level representation of the videos is computed through the spectrum analysis of the noise signal in the frequency domain. To isolate the noise signal of a given video V, we filter a copy of V using a Gaussian filter with mean μ, std. σ, and kernel size $k \times k$ to remove the high frequency components, generating a filtered video. Then, we perform a subtraction operation between the input video and its filtered version, generating a new video, called Residual Noise Video (V_{RN}):

$$V_{RN}^{(t)} = V^{(t)} - h(V^{(t)}) \quad \forall\ t\ \in T = \{1, 2, \ldots, t\}, \tag{3.1}$$

where $V^{(t)} \in \mathbb{N}^2$ is the t-th frame of V and h is a filter whose impulse response is a Gaussian function.

Calculation of the Fourier Spectrum Videos

After calculating the residual noise videos, we can analyze the noise pattern and possible artifacts contained in the biometric samples by applying the 2D Discrete Fourier Transform to each frame of the V_{RN} using Eq. 3.3. In this work, we evaluate two important characteristics of the noise signal in the frequency domain, the magnitude and phase of the signal. The analysis of these two characteristics is performed by calculating the magnitude spectrum (Eq. 3.5) and phase spectrum (Eq. 3.6), with the origin at the center of the frame. In both cases, the result is a Fourier spectrum video.

$$\mathcal{F}(V_{\mathrm{RN}}(x,y)) \equiv F(v,u) \tag{3.2}$$

$$F(v,u) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} V_{\mathrm{RN}}(x,y)e^{-j2\pi[(vx/M)+(uy/N)]} \tag{3.3}$$

$$|F(v,u)| = \sqrt{\mathcal{R}(v,u)^2 + \mathcal{I}(v,u)^2} \tag{3.4}$$

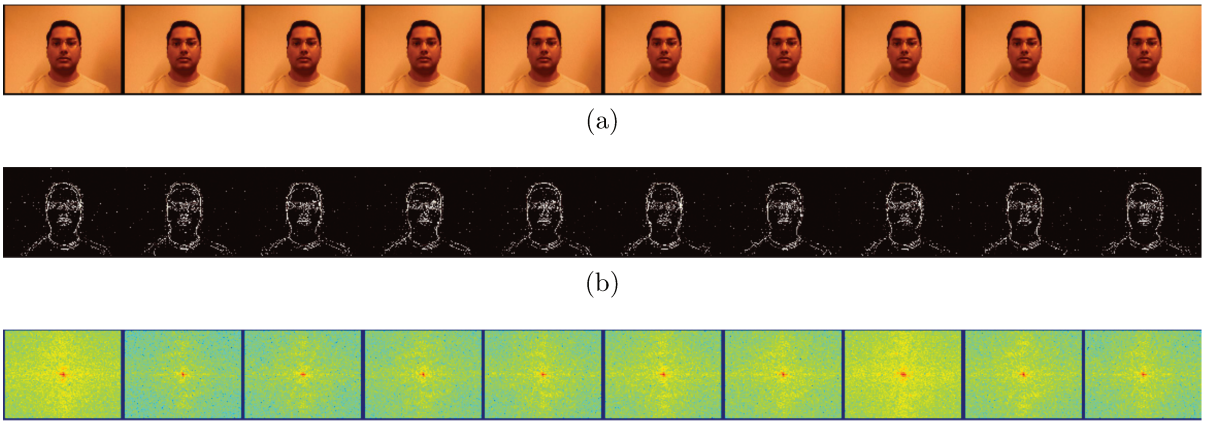
$$V_{\mathrm{MS}}(v,u) = \log(1+|F(v,u)|) \tag{3.5}$$

$$V_{\mathrm{PS}}(v,u) = \arctan\left(\frac{\mathcal{I}(v,u)}{\mathcal{R}(v,u)}\right) \tag{3.6}$$

From the Fourier spectrum video, we can extract spectral and temporal information relevant to the spoofing attack detection. In the case of the spectral information, we need to capture peaks present in the central region caused by artifacts that reduce some details in the scene (e.g., skin marking, edge information) such as blurring effect, defocus, and printing artifacts and peaks present in the peripheral region of the frame caused mainly by artifacts such as the banding effect and Moiré pattern, which appear during the recapture of the biometric information during an attack.

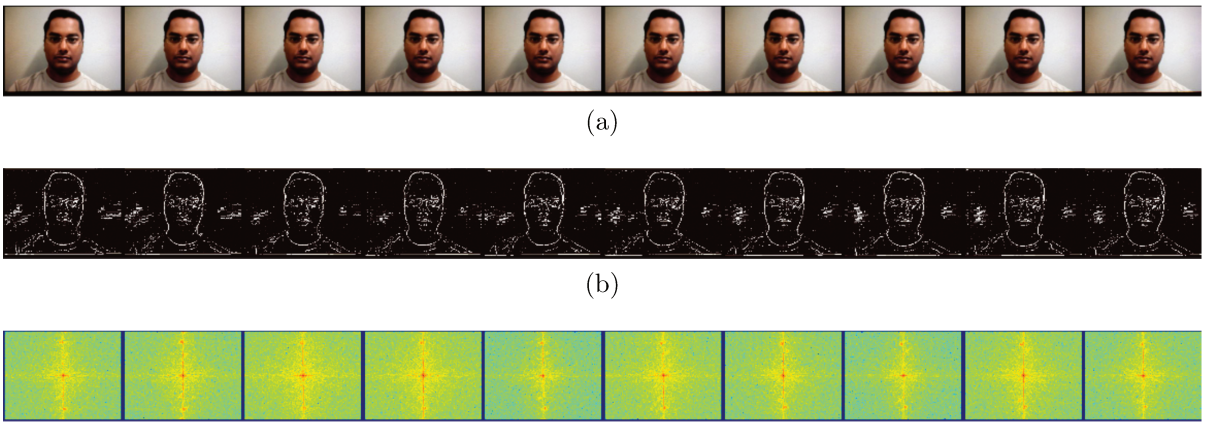
Figs. 3.2 and 3.3 show an attempt to depict the temporal disturbances added to the biometric samples during attacks. In this example, we extract the first ten consecutive frames of an attack video and of a valid video for the same client, and calculate their respective magnitudes spectra from the residual noise video. In addition, Fig. 3.4 shows examples in which we have frames extracted from valid access videos (a) and spoof attack videos (b-c). In this figure, we aim at showing the Moiré and blurring effects found in attempted attacks performed with a mobile device. The blurring effect is present in the magnitude spectrum with an increase of the low frequency components, whereas the Moiré effect is present in the magnitude spectrum with peaks in the horizontal center region of the frames. It is hard to find a direct mapping of the effects to the phase spectra, but we can see clearly that there are disturbances in the phase spectra calculated from attempted attack frames when compared to phase spectra extracted from valid access frames.

It is important to remark that we are not proposing a method for capturing each of the artifacts separately. We believe that the presence of one or more artifacts causes disturbances in the frequency components in the Fourier domain and the proposed method aims at describing and capturing this disturbance in space and time.



(a) (b) (c)

Figure 3.2: (a) Original frames extracted from a valid access video, (b) their respective residual noise frames and (c) magnitude spectra.



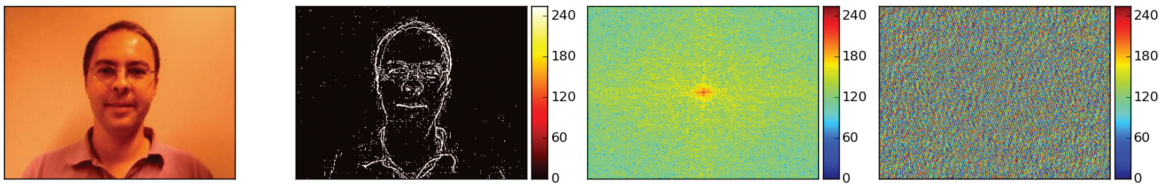
(a) (b) (c)

Figure 3.3: (a) Original frames extracted from an attempted attack video, (b) their respective residual noise frames and (c) magnitude spectra.

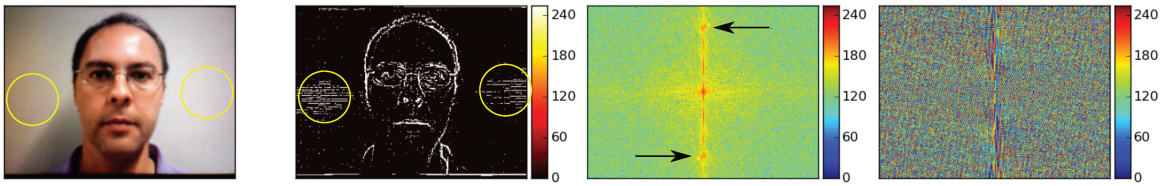
Computation of the time-spectral descriptor

Due to the dynamics involved in the appearance of artifacts and noise in the synthetic biometric samples and the spectral information, the temporal information becomes important to detect spoofing attacks. Therefore, we design a feature descriptor that gathers temporal and spectral information from an input video. We extract n temporal cubes of size of $w \times h \times t$ (blocks of size $w \times h$ of t frames) from the Fourier spectrum video. The idea of temporal cubes has been somewhat explored to quantify temporal information in other tasks in computer vision [6, 119, 209]. In all cases, it always boils down to designing important discriminative features for capturing the event of interest. In this paper, we design new ideas for spoofing detection.

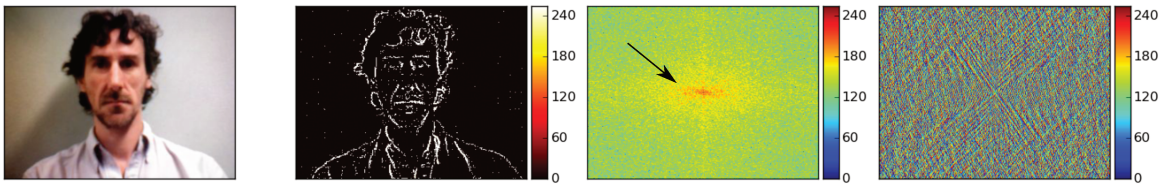
The computation of the measure over temporal cubes can be performed on each frame separately, hereinafter referred to as spatial measures, or between consecutive frames, hereinafter referred to spatio-temporal measures. Examples of spatial measures that can be used are energy and entropy of the signal, which quantify the signal size and amount of information, respectively. As examples of spatial-temporal measures, we can mention



(a) Examples of a frame extracted from valid access video



(b) Examples of frames extracted from a mobile-attack videos. We highlighted the Moiré effect with yellow circle in the original image and its respective residual noise frame. The arrows on the magnitude spectrum indicate the effect of the Moiré effect over Fourier spectrum



(c) Examples of frames extracted from a mobile-attack videos. In this frame, we show a blurring effect in the original image and its effect in the residual noise frame. The arrows on the magnitude spectrum show the impact of this effect over Fourier spectrum

Figure 3.4: Examples of valid access and attempted attack videos. The first column shows the original frame extracted from a video and the second column shows the residual noise frame calculated from the original frames. Finally, the third and fourth columns show the magnitude and phase spectrum, respectively. Note that the phase spectra calculated from valid access frames are different from attempted attack frames.

correlation and mutual information, which are applied to measure dependence between consecutive frames. At the end of this process, we have a set of n time-spectral descriptors of t dimensions, for each video. As spatio-temporal measures are applied on consecutive frames, this process yield n time-spectral descriptors of $(t-1)$ dimensions each.

3.3.2 Mid-Level Descriptor Extraction

To find a robust representation for the low-level feature descriptors, with less sensitivity to the intra- and extra-class variations, we use the Bag-of-Visual-Word (BoVW) model [235], which maps the low-level features onto a more discriminative mid-level representation. Methods based on the BoVW model can be understood in the following steps: visual codebook generation, coding, and pooling.

Visual Codebook Generation

The generation of the visual codebook consists in the selection of time-spectral descriptors that are more frequent and representative considering all descriptors extracted from training videos. The selected descriptors, called time-spectral visual words, form the visual codebook. The selection can be performed using two strategies: (1) random selection, whereby all descriptors are pooled and m visual words are randomly chosen using a uniform distribution; or (2) selection via clustering (e.g., k-means) whereby all descriptors undergo a clustering process and the m centroids found by the algorithm are used to form the visual codebook. In both cases, we end up with a single visual codebook, which is used to encode the low-level time-spectral descriptors from videos.

Instead of pooling all descriptors extracted from videos into a training set to build a single visual codebook, we can build class-based visual codebooks. When creating class-based visual codebooks, we consider the use of valid access and attempted attack video descriptors separately in order to find codebooks in each class. For each class-based codebook, we use the same procedures described above for a single visual codebook creation. The two visual codebooks are concatenated to create the final codebook.

Coding

The coding process performs a pointwise transformation of the low-level descriptors into another representation [31]. There are several strategies for coding being the hard and soft assignments the most common. Given a visual codebook and a low-level descriptor, the hard assignment transforms such descriptor into a binary vector with only one nonzero coefficient representing the visual word closest to it. The soft assignment [80], in turn, gives a real valued vector that represents the descriptor as a linear combination of the visual words of the codebook, whose coefficients give an associativity degree between the descriptor and the visual words of the codebook [146]. In this paper, we evaluate these two strategies for coding the low-level descriptors.

Pooling

The pooling process aims at summarizing the information contained in the set of n mid-level feature descriptors extracted from an input video into only one feature descriptor to obtain its final representation. In the literature, we have two common techniques to do that, known as sum-pooling (Eq. 3.7) and max-pooling (Eq. 3.8). In this paper, we evaluate these two strategies, as well.

$$v_i^{(j)} = \sum_{i=1}^{n} u_i^{(j)} \quad \forall\, j \ \in \{1, 2, \ldots, m\} \tag{3.7}$$

$$v_i^{(j)} = \mathtt{max}_i u_i^{(j)} \quad \forall\, j \ \in \{1, 2, \ldots, m\} \tag{3.8}$$

3.3.3 Classification

After finding a new space representation for the videos in the database, we use machine learning algorithms to find a classification model to decide whether a sample is an attempted attack or a valid access. In this paper, we evaluate the Partial Least Square (PLS) [101] and Support Vector Machine (SVM) [50] algorithms.

3.4 Experiments and Results

In this section, we present and discuss the experimental results and the validation of the proposed method. Section 3.4.1 shows details of the datasets used in the experiments while Section 3.4.2 describes the experimental protocols employed in this work. Section 3.4.3 shows the experimental setup of the proposed method regarding its parameters. The experiments in Section 3.4.4 aim at validating our method and choosing its best parameter setup. In addition, Section 3.4.4 addresses important questions regarding the low- and mid-level descriptor extraction procedures: (1) the best characteristic extracted from Fourier spectrum (e.g., magnitude or phase spectrum); (2) the best measure for spectrum summarization (e.g., energy, entropy, correlation, mutual information, etc); and (3) the visual codebook size most appropriate for the problem; among others. The remaining subsections compare the proposed method with the best methods reported in the literature including a challenging cross-dataset protocol, whereby we train our method using a dataset and test it with another dataset.

3.4.1 Datasets

In this work, we consider four datasets:

- **Replay-Attack Dataset [41]**: This dataset comprises videos of valid accesses and attacks of 50 identities. The videos were generated with a webcam with a resolution of 320×240 pixels and 25 frames per second (fps). This dataset contains 200 valid access videos, 200 print-based attacks, 400 mobile-based attacks using an iPhone, and 400 high-definition attacks using an iPad screen with $1,024 \times 768$ pixel resolution.

- **CASIA Face Anti-Spoofing Dataset [274]**: This dataset contains videos of valid accesses and attacks of 50 identities and considers different types of attacks such as warped photo attacks and cut photo attacks, besides the photos and video attacks. It also considers attacks performed with different image/video quality: (1) low-quality videos captured by a long-time-used USB camera with 480×640 pixel resolution; (2) normal-quality videos captured with a new USB camera with 480×640 pixel resolution; and (3) high-quality videos captured with a Sony NEX-5 camera with $1,920 \times 1,080$ pixel resolution. In total, it comprises 150 valid access videos and 450 video spoofing attacks.

- **UVAD Dataset [190, 191]**[1]: This dataset contains valid access and attempted attack videos of 404 different people, all created at Full HD quality, 30 fps, and nine seconds long. It contains 16,268 attempted attack videos and 808 valid access videos. Seven different display devices were used to simulate the attempted attacks performed upon three acquisition sensors of different manufacturers: a 9.1 megapixel (MP) Sony CyberShot DSC-HX1, a 10.0-MP Canon PowerShot SX1 IS, a 10.3-MP Nikon Coolpix P100, a 14.0-MP Kodak Z981, a 14.0-MP Olympus SP 800UZ, and a 12.1-MP Panasonic FZ35 digital camera. Figs. 3.5 and 3.6 illustrate some examples of this dataset.

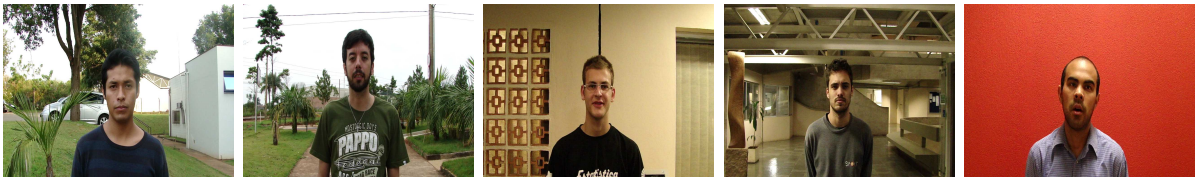


Figure 3.5: Examples of valid access video frames for outdoor (first and second images on the left) and indoor (three images on the right) scenes.

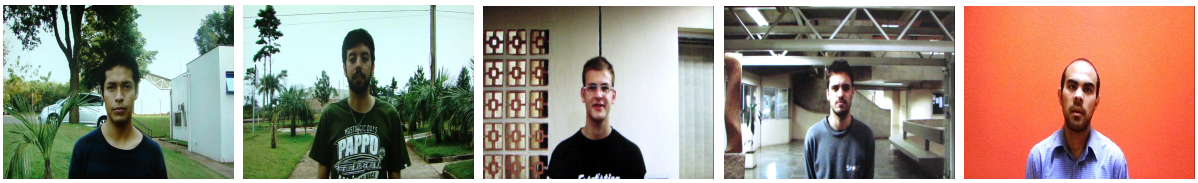


Figure 3.6: Examples of attempted attack video frames for outdoor (first and second images on the left) and indoor (three images on the right) scenes using Sony (first and second columns), Canon (third and fourth columns) and Nikon (last column) cameras.

- **3DMAD Dataset [66]**: This dataset comprises valid access and mask attack videos of 17 different subjects, whose faces were recorded by a Microsoft Kinect sensor. To build a synthetic biometric sample, the authors used frontal and profile face images to make the facial reconstruction. Afterwards, the authors used a 3D printer to build a mask containing facial information of the target person. Spoofing attack simulations were performed by presenting the 3D masks to the same Microsoft Kinect sensor. In total, the authors generated 85 valid access videos and 85 attempted attack videos.

3.4.2 Experimental Protocol

We use two measures for performance evaluation: the area under the curve (AUC) and the half total error rate (HTER). While the former quantifies the overall ability of a classifier to discriminate between attempted attacks and valid accesses, the latter combines the false acceptance rate (FAR) and false rejection rate (FRR) in a specific operating point of the ROC curve into a single measure. HTER is commonly calculated in the operating

[1]This dataset is freely available through FigShare (http://figshare.com/articles/visualrhythm_antispoofing/129545

point in which the FAR is equal to the FRR, known as the Equal Error Rate (EER). We use the freely available toolbox Bob [9] to calculate the AUC and HTER values. Finally, the employed evaluation protocols follow the ones proposed by the authors of the Replay-Attack, CASIA, UVAD and 3DMAD datasets. The source code of all proposed methods are freely available.[2]

Protocol I

In this experimental protocol, we use the Replay-Attack dataset, which is divided into three subsets: a training set with 300 attack videos and 60 valid videos; a development set with 300 attack videos and 60 valid access videos; and a test set with 400 attempted attack videos and 80 valid access videos. The training set is used to fit a classification model, the development set to find the EER, whereas the test set is used to report the final error rates.

Protocol II

In this protocol, we use CASIA dataset, divided into two disjoint subsets: training and test sets. Due to the absence of a development set to estimate a threshold to be applied in the test set and afterwards to calculate the HTER, the official protocol of this dataset recommends to use the training set to build a classifier and then use the test set to report the EER value. To report the results in terms of HTER, the original training set was divided into two subsets, named as training and development sets, in the proportion of 80% and 20%, respectively. We use the new training set to find the classification model and the development set to estimate the threshold that gives us the EER, whereas the official test set is used to report the final results in terms of HTER.

Protocol III

In this protocol, we use the UVAD dataset, which contains six subsets comprising valid access and attempted attack videos. Each subset considers attacks against one acquisition sensor: Sony, Kodak, Olympus, Nikon, Canon and Panasonic. Here, we train a classifier using the sensors Sony, Kodak and Olympus, and we test it with videos (valid access and attempted attacks) from three other different manufacturers: Nikon, Canon and Panasonic.

Protocol IV

Here, we use the 3DMAD dataset to evaluate spoofing detection of attacks using 3D masks. The dataset contains 85 RGB videos that represent valid access and 85 RGB videos that represent attempted spoofing attacks. As this dataset does not contain explicit subsets, we randomly partitioned the data into three subsets: training, development and testing, and we use Protocol I for testing.

[2]The source code is freely available for scientific purposes on GitHub (https://github.com/allansp84/spectralcubes), along with this article.

Table 3.1: After the statistical analysis, we have found that the factors highlighted with † are the ones that did not present statistical significance when configuring our method, whereas the levels highlighted in bold are the chosen levels.

Factor	Levels	Description
LGF	C, and **W**	Strategies for extracting the low-level features from video of phase spectrum or video of magnitude spectrum: extraction considering a central region (crop) in each frame (C) and the entire/whole frames (W).
M	PE, PH, ME, MH, PMI, MMI, PC, and **MC**	Characteristics of the frequency spectrum evaluated that can be the phase (P) or magnitude (M) and the measures used for summarizing the spectral information that can be energy (E), entropy (H), mutual information (MI), or correlation (C).
CS	R, and **K**	Mode of selection of the visual words that compose the visual codebooks: Random (R) or using k-means clustering algorithm (K).
SDD†	S and D	Strategies for generating the visual codebooks: a single visual codebook (S) and class-based visual codebooks (D), one for each data class (spoofing vs non-spoofing).
DS†	80, 120, 160, 200, 240, 280, 320, and 360	Visual codebook sizes. This is an important parameter because the visual codebook size gives us visual codebooks with different degrees of specificities because large visual codebooks can incorporate small clusters of data that appear sometimes in specific cases.
CP	hardsum, hardmax and **softmax**	We evaluate the combination of two strategies in the coding process (hard-assignment and soft-assignment) and two strategies in the pooling process (max-pooling and sum-pooling).
C	**SVM** and PLS	Classification algorithms.

3.4.3 Method Parameterization

For reproducibility purposes, this section discuss the parameters whose values are constant in the setup of our method.

We extract the noise signature from RGB videos using a Gaussian filter with $\mu = 0$, $\sigma = 0.5$, and kernel size 3×3 (Eq. 3.1). These values were obtained empirically in [193]. Next, we extract cuboids of size $32 \times 32 \times 8$ from the Fourier spectrum videos (Eqs. 3.5 and 3.6), whose spatio-temporal location is chosen randomly based on a uniform distribution.

The use of spatial measures produces low-level 8-dimensional descriptors per channel, whereas the use of spatio-temporal measures produces low-level 7-dimensional descriptors per channel, which gives us a final low-level descriptor of 24-dimensional and 21-dimensional, respectively. Finally, the number of cubes extracted from videos is determined by dividing the volume of the video with respect to the cube.

Regarding the mid-level descriptors, the only parameters with constant values are the ones that define the Gaussian kernel used in the soft-assignment coding technique, whose values are $\mu = 0$ and $\sigma = 0.04$. Finally, the SVM parameters are found through grid search in the training data.

3.4.4 Experimental Design and Analysis

To find the best method configuration, we performed a factorial experiment with replication ($N = 3$) followed by an analysis of variance (ANOVA) [257]. Each experimental unit

is represented as a tuple of n objects, each one with a level of a factor. Considering the replications, we have a total of $9,216$ tuples, which are used to instantiate the proposed method. The instances of the proposed method are evaluated through the measurement of the value of the system response variable, the AUC value, after running such instances using the Replay-Attack dataset and Protocol I, using the development set. Next, we collect obtained AUC values and then we performed an ANOVA test to analyze the significance of the effects of the parameters on the classification results.

With this approach, we can discover which parameters significantly affect the system response variable and also the best configuration of the method [94]. Henceforth, the method parameters are referred to as *factors* and their values as *levels*. Table 3.1 shows a brief description of the factors and their respective levels we consider herein.

Low-Level Descriptor Extraction Parameter Analysis (LGF and M)

The low-level feature extraction has two important parameters: the frequency characteristics of the signal (phase or magnitude), and the function used to summarize the information of the temporal cubes extracted from a video. In this work, we evaluate measures that describe spatial information of the temporal cubes (energy an entropy), and measures that describe the temporal behavior of the cubes (mutual information and correlation across time).

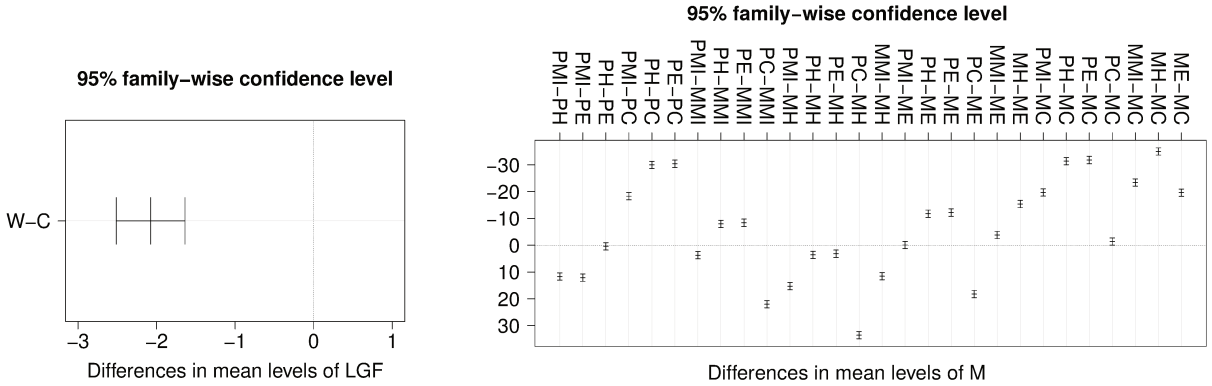
To find which levels are statistically different for each factor, we perform the Tukey's HSD test (see Fig. 3.7). In Figs. 3.7(a)-(b), the pairs in comparison whose confidence intervals do not intercept the zero value are statistically different. Considering the top-5 method configuration obtained in this experiment, we conclude that the whole frame for extracting features is more interesting than any cropped region in the center of the frame. In addition, the characteristic extracted of the Fourier spectrum and the summarization measure used to generate the low-level feature descriptors have a great impact in the method discriminability (Fig. 3.7(b)), as several comparisons in pairs of features are statistically significant.

Mid-Level Descriptor Extraction Parameter Analysis (CS, SDD, DS, and CP)

To construct a discriminative visual codebook, we need to choose the best strategy for selecting the words that compose the visual codebooks (CS) as random or clustering-based, the visual codebook size (DS), the policy to create the visual codebooks (SDD) as single or class-based, and the pooling and coding strategies (CP).

Fig. 3.8 shows the results of the post-hoc test with Tukey's HSD. In Fig. 3.8(a), we have the results of the statistical analysis for DS parameter (dictionary size), to which was not found statistical significance. Therefore, we recommend that dictionary size parameter to be optimized according to the application of interest. In turn, Fig. 3.8(b) shows that different pooling and coding processes causes statistically significant impacts on the response variable, and softmax is the recommended choice.

In addition, Fig. 3.8(c) shows that the method used to select the words that compose the visual codebook (random vs. clustering-based selection) also presents results that are statistically significant with k-means being the recommend choice due to the high



(a) LGF denotes the region of the frame for extracting the low-level features.

(b) M refers to summarization measure and the video of spectra used to generated the low-level descriptors.

Figure 3.7: Confidence interval on the differences between the means of the levels of the factors (a) LGF, and (b) M. For each comparison, the Tukey's HSD test provides an estimation of the differences between mean pairs and their respective confidence intervals, as well the p-value for each comparison. All comparisons whose confidence intervals do not contain zero value have a p-value lower than 0.05 and, therefore, are statistically different with a 95% confidence level. (See Table 3.1 to see the description of levels.)

performance achieved by models built with visual codebooks generated using k-means, during this experiment. Finally, Fig. 3.8(d) shows that the visual codebook creation strategy (single visual codebook vs. class-based codebooks) does not present statistical difference and, therefore, should also be considered in a future optimization process during the implementation of the method in a real application.

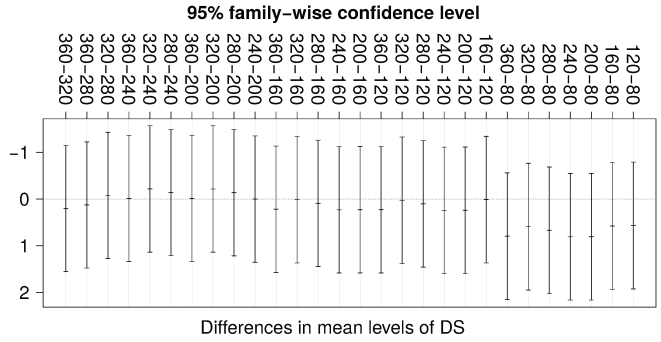
Classification Step Parameter Analysis (C)

The SVM classifier outperformed PLS classifier with a statistically significant difference ($p\text{-}value = 0.00$). We believe that this happened because of the non-linearity of the data as we use a non-linear version of SVM the a linear version of PLS.

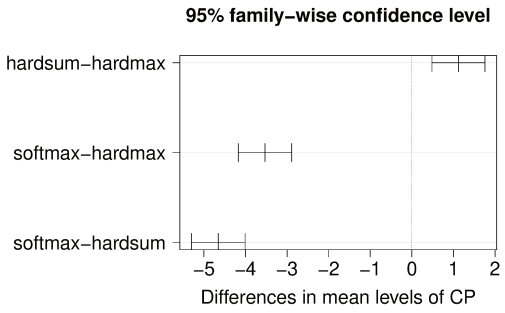
Analysis of Interaction Effects and Choice of the Best Configuration

After analyzing each factor in isolation, we examine whether there is significant interaction between factors. In this case, if a small p-value is obtained in the interaction effect analysis between two factors, then we can conclude that these factors do not operate independently of each other [94]. Otherwise, there is no evidence of an interaction effect.

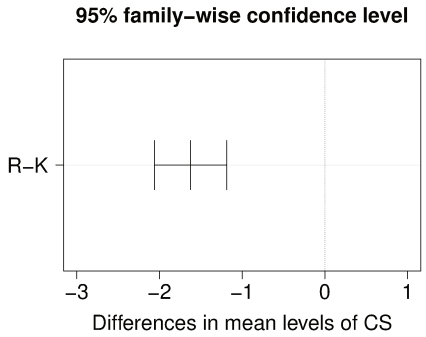
First of all, we can see that there is a relationship between the region from which the low-level time-spectral features are extracted (factor LGF) and the spectral information used in the generation of time-spectral descriptors (factor M). When analyzing the magnitude spectrum of the Fourier transform, we see that there is a concentration of low frequency components in the abscissa and ordinate axes. Fig. 3.9 shows that this interaction between factors LGF and M exists. In addition: (i) we have an increase in the mean of AUC values for measures MH, PH, PE and PMI, when these measures are calculated in the center region of the frames; (ii) we have a decrease in the mean of



(a) DS refers to the number of time-spectral visual words present in the visual codebook.



(b) CP denotes the coding and pooling strategies used to build the mid-level descriptors.



(c) CS denotes the selection mode of the time-spectral visual words for composing the visual codebook.

(d) SDD denotes the strategies for generating the visual codebooks, single or class-based visual codebooks.

Figure 3.8: Confidence interval of the differences between the means of the levels of the factors (a) DS, (b) CP, (c) CS and (d) SDD. All comparisons whose confidence intervals do not contain zero value have a p-value lower than 0.05 and, therefore, are statistically different with a 95% confidence level as it is the case for the comparisons indicated on the (a) x axis and (b-d) y axis. (See Table 3.1 for the description of levels.)

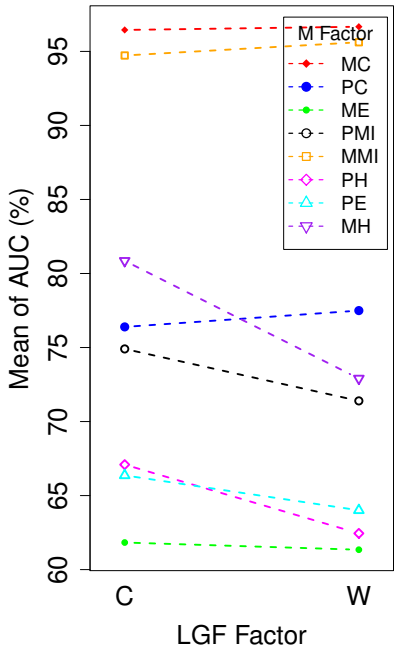
AUC for measure PC; and (iii) we have very small changes in mean values of AUC for MC and MMI when we compare the two strategies for feature extraction.

Finally, the form of selecting the visual words (factor CS) and the method of coding and pooling used in the construction of the dictionaries (factor CP) also presents an interesting interaction. Both factors significantly influence the results, but not in isolation. Fig. 3.9(b) shows that the results obtained with hardmax, hardsum and softmax are worse when the visual words are chosen randomly instead of through clustering.

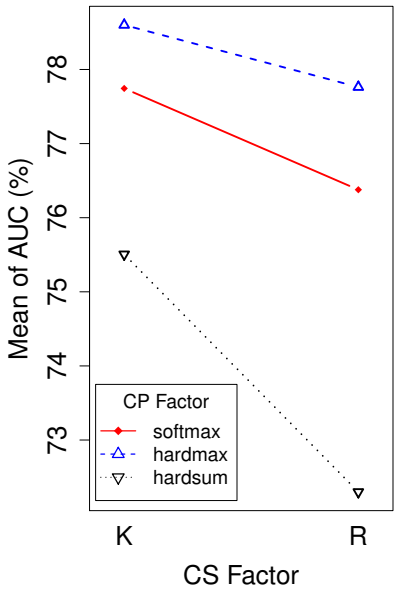
3.4.5 Summary After Analyzing Different Factors and Levels

The proposed method presents better results using time-spectral features extracted from magnitude spectrum videos considering the whole frames of a video and using the correlation measure from time-spectral features for generating the time-spectral descriptors.

The class-based codebooks outperform the single codebook and the selection strategy of the visual words that best fits to the spoofing detection problem is the k-means clustering. The most appropriate size for codebooks is 320 visual words and the softmax outperformed the other coding and pooling strategies. With this configuration, we



(a) LGF and M interaction. Note that we have a slump in the mean of AUC when setting LGF to W and decrease the number of low-level features.



(b) CS and CP interaction. We have a significant increase in the mean of AUC with the hard-assignment when we use k-means to select the visual words of the codebooks.

Figure 3.9: Interaction plots between pairs of factors (a) LGF×M and (b) CS×CP. The factor LGF denotes the region in the frame considered for extracting the low-level features, while factor M denotes the statistical measures considered for describing the information of the temporal cubes. Finally, the factor CS denotes the mode of selection of the visual words from visual codebooks and the factor CP refers to the strategies used in the coding and pooling process. (See Table 3.1 to see the description of levels.)

obtained an AUC of 99.46% and an HTER of 2.75%, considering the test set of the Replay-Attack dataset [41]. Next, we show experiments and results for this method using this final configuration.

3.4.6 Results

This section compares the proposed method with others in the literature for the Replay-Attack [41], CASIA [274] and 3DMAD [65] datasets. In all experiments, we used the best configuration of the proposed method as discussed in the last section. The parameters that did not present statistical significance (DS and SDD), were fine-tuned for each dataset.

Replay-Attack Dataset

We first consider the validation Protocol I (c.f., Sec. 3.4.2) and the Replay-Attack dataset. Table 3.2 shows the results for the three types of attacks available in this set. Fig. 3.10(a) shows that attacks performed with high-quality samples are more difficult to detect (HTER of 5.94%). This result was expected as high-quality fake samples usually contain less artifacts revealing an attack.

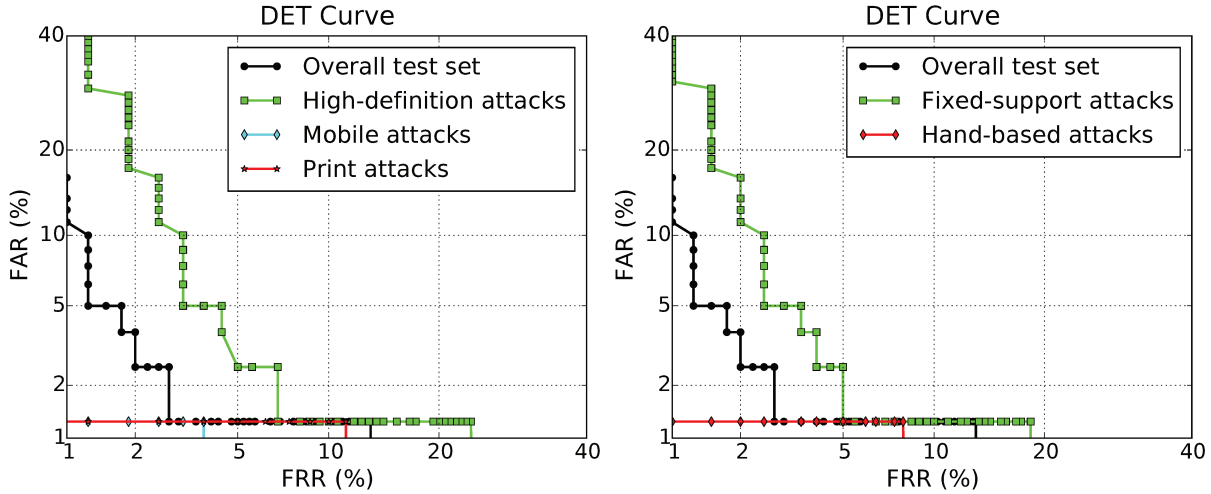


Figure 3.10: Results obtained on Replay-Attack dataset for each type of attack using fixed-support (a) in contrast with hand-based attacks (b).

In turn, video-based and photo-based attacks were easily detected (HTER of 0.63%). Note that video-based spoofing attacks are more susceptible to blurring effects, whereas the photo-based attacks show a large amount of flickering effects due to printing defects. Fig. 3.10(b) shows results obtained considering fixed-support and hand-based attacks, separately. We believe that hand-based attacks are easier to be detected given that small movements of the impostor user during the attack generate more artifacts in the biometric sample causing more disturbances in the frequency components.

Table 3.2: Performance results for the Replay-Attack Dataset.

Dataset	FAR	FRR	HTER	AUC
High-definition attack	10.63	1.25	5.94	98.77
Mobile attack	0.00	1.25	0.63	99.95
Print attack	0.00	1.25	0.63	99.86
Hand-based attack	1.00	1.25	1.13	99.87
Fixed-support attack	7.50	1.25	4.38	99.03
Overall test set	**4.25**	**1.25**	**2.75**	**99.46**

CASIA Face Anti-Spoofing Dataset

In this experiment, we evaluate the proposed method using the Protocol II (c.f., Sec. 3.4.2) and CASIA dataset.

Table 3.3 shows the results obtained for the seven scenarios of attacks available in this dataset. Fig. 3.11(a) shows that video-based and warp-photo spoofing attacks are easier to be detected by the proposed method (HTER of $\approx 8\%$). On the other hand, the cut-based spoofing attacks are more difficult to be detected (HTER of 22.22%). One possible reason for cut-based attacks to be more difficult for detecting is that during an attempted attack based on cut-photos, the photographs are practically in the same position during all the

time, generating fewer artifacts along time, whereas for the attempted attacks based on warped-photos, the photographs are bent during the attack to simulate facial motion. In addition, we believe that video-based attacks were easier to be detected because of the inevitable downsize of the high-resolution samples by the screen device used during attack, as also reported by CASIA's authors [274]. In this case, many evidences of attempted attacks are generated and added to the fake sample.

As for the quality of the acquisition (Fig. 3.11(b)), the proposed method showed better results for attacks carried out with low-quality videos. An interesting result is the best performance of the method to deal with high-resolution videos than normal quality videos. We believe that any conclusion would be precipitous because many factors can influence the noise level of a sensor such as sensor imperfections (e.g., appearance of hot pixels, dead pixels, as well as pixel traps under different acquisition conditions). Several works in the literature have explored these issues. For instance, thermal action has a considerable impact over pattern noise of a digital camera and appearance of defective pixels [39, 150, 216]. As we do not assure that the captures/recaptures happened under similar acquisition conditions, it is wiser only to point out the existence of classification differences in this case.

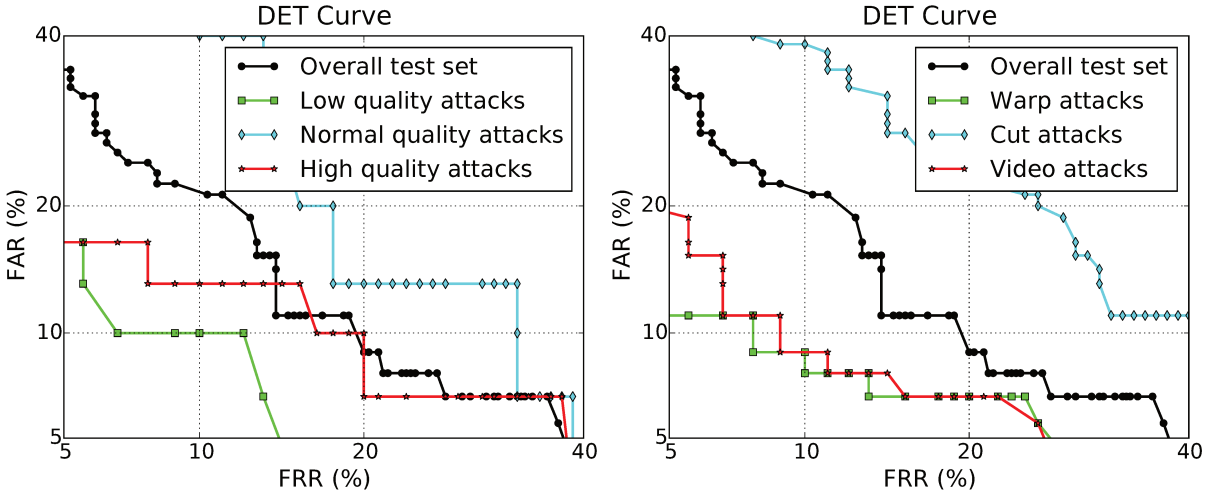
Table 3.3: Performance results for the CASIA dataset.

Dataset	FAR	FRR	HTER	AUC
Low quality	10.00	10.00	10.00	98.11
Normal quality	17.78	20.00	18.89	87.67
High quality	13.33	13.33	13.33	95.04
Warp photo attack	7.78	8.89	8.33	96.05
Cut photo attack	22.22	22.22	22.22	87.27
Video attack	8.89	8.89	8.89	96.41
Overall Attack	**14.07**	**14.44**	**14.26**	**93.25**

3DMAD Dataset

We now turn our attention to evaluate the proposed method for mask-based spoofing attack detection using the Protocol IV (c.f., Sec. 3.4.2). Using the official dataset protocol, the proposed method obtained an AUC of 96.16% and an HTER of 8.0%.

Erdogmus et al. [66] reported an HTER of 0.95% using block-based LBP features (local features) and the Linear Discriminant Analysis (LDA) classifier. This performance difference is somewhat explained due to the different validation protocol used. Erdogmus et al. used an 1000-fold cross validation method and, in each fold, the clients from the dataset were randomly assigned into training, development and test sets. In our case, we randomly divided the clients from dataset and assigned them into training, development and test set only once. Even so, the proposed method outperforms other techniques using global LBP, whose HTERs reported by Erdogmus et al. were all above 10.0%.



(a) (b)

Figure 3.11: Results obtained on CASIA dataset for the three type of attacks (a) and for the three quality of attack (b).

UVAD Dataset

In this experiment, we evaluate the proposed method using the Protocol III (c.f., Sec. 3.4.2) and UVAD dataset. We also evaluate the proposed method considering LBP-based and motion-based countermeasure methods.

According to Pereira et al. [184], the correlation method presents an HTER of 11.79% on Replay-Attack. In turn, $\text{LBP}^{u2}_{8,1}$ [41] was effective to characterize the artifacts embedded in the attack videos on Replay-Attack obtaining an HTER of 15.16%. In the UVAD dataset, however, both methods obtained a more modest performance as Table 3.4 shows. With respect to $\text{LBP}^{u2}_{8,1}$ method, for instance, the proposed method reduces the classification error in about 36%.

Table 3.4: Comparison among LBP-based approach, motion-based approach and the proposed method on the UVAD dataset.

Methods	FAR (%)	FRR (%)	HTER (%)
Correlation (motion-based approach) [8]	81.60	14.56	48.06
$\text{LBP}^{u2}_{8,1}$ [41]	27.41	66.04	46.72
Proposed Method	**44.73**	**15.00**	**29.87**

Comparison with State-of-the-Art Methods for CASIA and Replay-Attack Datasets

In this section, we compare the proposed method with others available in the literature for Replay-Attack and CASIA datasets. Table 3.5 shows results for the Replay-Attack Dataset. The proposed method outperforms the ones based on texture analysis [41,69,151] and also methods based on motion analysis [8]. It was also more effective than methods

based on fusion schemes reported by Pereira et al. [184] and Komulainen et al. [124], with a relative error reduction (RER) of 67.69% and 46.18%, respectively.

Table 3.5: Comparison among the existing methods. The first column shows the HTERs reported by the authors, whereas the second column shows the Relative Error Reduction (RER) obtained with the proposed method. The reported HTERs were obtained using the original Replay-Attack Dataset protocol. The results highlighted with † and ‡ were reported by Chingovska et al. and Pereira et al., respectively.

Methods	HTER (%)	RER (%)
Chingovska et al. [41]	15.16	81.86
Allan Pinto et al. [190]	14.27	80.73
Määttä et al. [151]	13.87†	80.17
Anjos and Marcel [8]	11.79‡	76.68
Pereira et al. [184]	8.51	67.69
Pereira et al. [69]	7.60	63.82
Komulainen et al. [124]	5.11	46.18
Proposed Method	**2.75**	0

Table 3.6 shows a comparison among the proposed method and others reported in the literature for CASIA dataset. The proposed method is on par with the best ones in the literature.

Table 3.6: Comparison among the proposed method and others available in the literature. According to the authors of the proposed methods, EERs reported were obtained using the original CASIA Dataset protocol.

Methods	EER (%)
DoG Baseline. [274]	17.0
$LBP^{u2}_{8,1}$. [69]	16.0
$LBP\text{-}TOP^{u2}_{8,8,8,1,1,1}$. [69]	10.0
Proposed Method	**14.0**

Analysis of the Minimum Detection Time

We now analyze the impact of the video length over the method discriminability for CASIA, Replay-Attack and 3DMAD datasets. This experiment evaluates: the minimum number of frames required for the method to operate; and the method stability, in terms of HTER(%), for the three different datasets.

Fig. 3.12 indicates that HTER values vary only slightly when we change the video length for the three datasets and that the proposed method uses about two seconds to detect an attempted attack, thus not compromising the transparency of the authentication process.

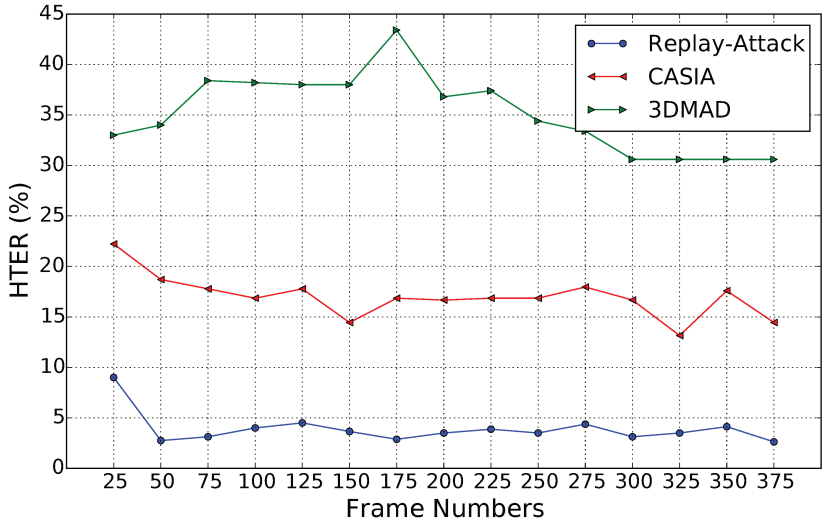


Figure 3.12: Results in terms of HTER (%) of the proposed method for different video input length for Replay-Attack, CASIA and 3DMAD datasets.

Cross-Dataset Evaluation

In this section, we discuss the performance of the proposed method considering a more difficult scenario (cross-dataset), in which the proposed method is trained with one dataset but it is tested on a different dataset with different acquisition conditions. In this experiment, all datasets used during the training were randomly divided into training and development sets in a proportion of 80% and 20%, respectively. The development set is used to estimate the EER threshold that is necessary to calculate the HTER during the test.

Table 3.7 shows the results using the cross-dataset protocol. The results indicate that the proposed method presents better generalization when trained with CASIA, with a mean HTER of 40.17%. We believe this occurred due to more variability of the type of attacks and video quality in the CASIA dataset, which enriches the training. This dataset contains warped-, cut- and video-based attacks performed with spoofed samples of different quality: low, normal and high quality. Such characteristics enables a better generalization of the method when CASIA is used for training.

In turn, the best performance when testing the CASIA and 3DMAD datasets was obtained when training with UVAD dataset, another rich dataset for training. Although this dataset contains only video-based spoofing attacks, it has comprises different sensors (for capturing and recapturing the biometric samples) and display devices used during the attempted attacks. We believe that such variability adds different sensor-intrinsic noise levels to the training samples, which contribute to build a more robust classification model.

With regard to the more modest generalization presented during the test of the 3DMAD dataset, we believe that it is due to the absence of some artifacts that are commonly found in samples from photo-based and video-based attacks (e.g., blurring, flickering effects) that were not found in the attempted attack video from 3D masks. In addition, spoofing attacks performed with masks are less likely to add temporal distur-

bances similar to those added when the impostor presents the fake samples, by hand, using a monitor or a photo.

Finally, Table 3.8 shows a comparison among the obtained results reported in the literature. Except for the correlation method, all others present a better performance when they are trained with CASIA. Once again, we believe that our method performs better when training with CASIA because such dataset is more heterogeneous than Replay-Attack. The Correlation [8] and LBP-TOP [41] methods aim to characterize temporal information, similarly to the proposed method, and the results of both methods emphasize the difficulty in characterizing such information completely. In this protocol, besides handling data from different sensors, all methods have to deal with different lighting conditions and background.

Table 3.7: Results obtained with the cross-dataset protocol and using the overall test sets of each dataset.

Train	Test	FAR (%)	FRR (%)	HTER (%)	Mean HTER (%)
CASIA	3DMAD	88.00	4.00	46.00	40.17%
	Replay-Attack	32.50	36.25	**34.38**	
	UVAD	38.61	41.67	**40.14**	
Replay-Attack	3DMAD	52.00	44.00	48.00	47.45%
	CASIA	0.00	100.0	50.00	
	UVAD	5.74	83.33	44.54	
UVAD	3DMAD	84.00	4.00	**44.00**	41.76%
	CASIA	13.70	63.33	**38.52**	
	Replay-Attack	79.25	6.25	42.75	

3.5 Conclusions and Future Work

In this paper, we proposed an algorithm for detecting spoofing attacks that takes advantage of noise and artifacts added to the synthetic biometric samples during their manufacture and recapture. We showed that the analysis of the behavior of the noise signature, in the frequency domain, is proper to reveal spoofing attacks. For this, we proposed the use of time-spectral features as low-level descriptors, which gather temporal and spectral information in a single feature descriptor. To handle several types of attacks and to obtain a feature descriptor with a suitable generalization, we also proposed the use of the visual codebook concept to find a mid-level representation from time-spectral descriptors.

The experimental results showed that the magnitude is an important characteristic from a signal, in frequency domain, for spoofing attack detection. We also showed how to use the visual codebook concept effectively in order to find a more robust space representation to the different kinds of attacks and with a good generalization. The obtained results demonstrated the effectiveness of the proposed method in detecting different types of attacks (photo-, video-, and 3D-mask-based ones).

We believe that the frequency-based approach used is effective because we have a decrease in low frequency components due to information loss caused during manufacture of the fake samples (e.g., information loss during printing) and recapture (e.g., blurring

effect) and an increase in some high frequency components in the fake samples during recapture due to some artifacts added to the fake samples (e.g., printing artifacts, banding effect, noise added by the imaging sensor). Moreover, these disturbances in the composition of the components of frequencies are best characterized as we analyze the biometric sample in the frequency domain rather than spatial domain and along time instead of on isolated frames or still images.

Table 3.8: Comparison among different anti-spoofing methods considering cross-dataset protocol.

Methods	**Train**	**Test**	**HTER (%)**
Proposed Method	Replay-Attack	CASIA	50.00
	CASIA	Replay-Attack	34.38
Correlation	Replay-Attack	CASIA	48.28
	CASIA	Replay-Attack	50.25
LBP-TOP$^{u2}_{8,8,8,1,1,1}$	Replay-Attack	CASIA	61.33
	CASIA	Replay-Attack	50.64
LBP$^{u2}_{8,1}$	Replay-Attack	CASIA	57.90
	CASIA	Replay-Attack	47.05

Regarding the important cross-dataset validation, the performed experiments demonstrated that the proposed method and other approaches available in the literature still have modest generalizations. This is of particular importance for the research community as it shows that the problem is still far from solved and cross-dataset validation must be considered from now on when designing and deploying spoofing detection techniques.

As discussed earlier, we observed that different biometric sensors present different properties. Therefore, it is important to train a classifier considering this variability. UVAD dataset comes in hand for this purpose and will surely serve the community in this regard with more than 15k samples of hundreds of clients and diverse sensors.

Finally, it is worth mentioning that we do not claim to introduce the best method out there for spoofing detection. On the contrary, our very objective in this paper was to show that capturing spatio, spectral and temporal features from biometric samples can be successfully considered in the spoofing detection scenario. That being said, it is likely that the proposed approach, when combined with existing ones in the literature, may as well boost the performance since they will likely rely on complementary features for solving the problem.

Directions for future research include the investigation of new approaches to transforming low-level descriptors into mid-level descriptors as Fisher vectors [186] and Bossa Nova [12]. These strategies for finding mid-level representations could also be exploited by methods that use texture-based descriptors. In such cases, the goal would be to investigate whether the representation space found by the texture descriptors used in the literature for detecting face spoofing attacks (e.g., LBP, LBP-TOP, and their variants) could be transformed in a new representation space better adapted to the face spoofing problem in a scenario with different types of attacks.

Acknowledgments

We thank CAPES (DeepEyes Project), FAPESP (#2010/05647-4 and #2011/22749-8), CNPq (#307113/2012-4, #304352/2012-8, #477662/2013-7, #487529/2013-8 and #477457/2013-4), FAPEMIG (APQ-01806-13 and APQ-00567-14), and Microsoft Research for the financial support.

Chapter 4

Deep Representations for Iris, Face, and Fingerprint Spoofing Detection

"The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspires."

—William Arthur Ward, *American administrator, writer, pastor, and teacher (1921–1994)*

"Persistence can change failure into extraordinary achievement."

—Matthew Nicholas Biondi, *American swimmer*

Abstract

Biometrics systems have significantly improved person identification and authentication, playing an important role in personal, national, and global security. However, these systems might be deceived (or "spoofed") and, despite the recent advances in spoofing detection, current solutions often rely on domain knowledge, specific biometric reading systems, and attack types. We assume a very limited knowledge about biometric spoofing at the sensor to derive outstanding spoofing detection systems for iris, face, and fingerprint modalities based on two deep learning approaches. The first approach consists of learning suitable convolutional network architectures for each domain, while the second approach focuses on learning the weights of the network via back-propagation. We consider nine biometric spoofing benchmarks — each one containing real and fake samples of a given biometric modality and attack type — and learn deep representations for each benchmark

Pre-print of article that will appear in T-IFS, vol.10, no.4, pp.864-879, April 2015.
The published article is available on http://dx.doi.org/10.1109/TIFS.2015.2398817
See permission to use the copyrighted material in **Appendix C**.

by combining and contrasting the two learning approaches. This strategy not only provides better comprehension of how these approaches interplay, but also creates systems that exceed the best known results in eight out of the nine benchmarks. The results strongly indicate that spoofing detection systems based on convolutional networks can be robust to attacks already known and possibly adapted, with little effort, to image-based attacks that are yet to come.

4.1 Introduction

BIOMETRICS human characteristics and traits can successfully allow people identification and authentication and have been widely used for access control, surveillance, and also in national and global security systems [108]. In the last few years, due to the recent technological improvements for data acquisition, storage and processing, and also the scientific advances in computer vision, pattern recognition, and machine learning, several biometric modalities have been largely applied to person recognition, ranging from traditional fingerprint to face, to iris, and, more recently, to vein and blood flow. Simultaneously, various *spoofing attacks* techniques have been created to defeat such biometric systems.

There are several ways to spoof a biometric system [204,205]. Indeed, previous studies show at least eight different points of attack [76,203] that can be divided into two main groups: *direct* and *indirect* attacks. The former considers the possibility to generate synthetic biometric samples, and is the first vulnerability point of a biometric security system acting at the sensor level. The latter includes all the remaining seven points of attacks and requires different levels of knowledge about the system, e.g., the matching algorithm used, the specific feature extraction procedure, database access for manipulation, and also possible weak links in the communication channels within the system.

Given that the most vulnerable part of a system is its acquisition sensor, attackers have mainly focused on direct spoofing. This is possibly because a number of biometric traits can be easily forged with the use of common apparatus and consumer electronics to imitate real biometric readings (e.g., stampers, printers, displays, audio recorders). In response to that, several biometric spoofing benchmarks have been recently proposed, allowing researchers to make steady progress in the conception of anti-spoofing systems. Three relevant modalities in which spoofing detection has been investigated are iris, face, and fingerprint. Benchmarks across these modalities usually share the common characteristic of being image- or video-based.

In the context of irises, attacks are normally performed using printed iris images [229] or, more interestingly, cosmetic contact lenses [32,264]. With faces, impostors can present to the acquisition sensor a photography, a digital video [41], or even a 3D mask [66] of a valid user. For fingerprints, the most common spoofing method consists of using artificial replicas [83] created in a cooperative way, where a mold of the fingerprint is acquired with the cooperation of a valid user and is used to replicate the user's fingerprint with different materials, including gelatin, latex, play-doh or silicone.

The success of an anti-spoofing method is usually connected to the modality for which

it was designed. In fact, such systems often rely on expert knowledge to engineer features that are able to capture acquisition telltales left by specific types of attacks. However, the need of custom-tailored solutions for the myriad possible attacks might be a limiting constraint. Small changes in the attack could require the redesign of the entire system.

In this paper, we do not focus on custom-tailored solutions. Instead, inspired by the recent success of Deep Learning in several vision tasks [46, 47, 130, 173, 242], and by the ability of the technique to leverage data, we focus on two general-purpose approaches to build image-based anti-spoofing systems with convolutional networks for several attack types in three biometric modalities, namely iris, face, and fingerprint. The first technique that we explore is hyperparameter optimization of network architectures [18, 195] that we henceforth call *architecture optimization*, while the second lies at the core of convolutional networks and consists of learning filter weights via the well-known back-propagation [135] algorithm, hereinafter referred to as *filter optimization*.

Fig. 4.1 illustrates how such techniques are used. The architecture optimization (AO) approach is presented on the left and is highlighted in blue while the filter optimization (FO) approach is presented on the right and is highlighted in red. As we can see, AO is used to search for good architectures of convolutional networks in a given spoofing detection problem and uses convolutional filters whose weights are set at random in order to make the optimization practical. This approach assumes little a priori knowledge about the problem, and is an area of research in deep learning that has been successful in showing that the architecture of convolutional networks, by themselves, is of extreme importance to performance [15, 18, 20, 194, 195, 221]. In fact, the only knowledge AO assumes about the problem is that it is approachable from a computer vision perspective.

Still in Fig 4.1, FO is carried out with back-propagation in a predefined network architecture. This is a longstanding approach for building convolutional networks that has recently enabled significant strides in computer vision, specially because of an understanding of the learning process, and the availability of plenty of data and processing power [130, 233, 242]. Network architecture in this context is usually determined by previous knowledge of related problems.

In general, we expect AO to adapt the architecture to the problem in hand and FO to model important stimuli for discriminating fake and real biometric samples. We evaluate AO and FO not only in separate, but also in combination, i.e., architectures learned with AO are used for FO as well as previously known good performing architectures are used with random filters. This explains the crossing dotted lines in the design flow of Fig 4.1.

As our experiments show, the benefits of evaluating AO and FO apart and later combining them to build anti-spoofing systems are twofold. First, it enables us to have a better comprehension of the interplay between these approaches, something that has been largely underexplored in the literature of convolutional networks. Second, it allows us to build systems with outstanding performance in all nine publicly available benchmarks considered in this work.

The first three of such benchmarks consist of spoofing attempts for iris recognition systems, Biosec [219], Warsaw [51], and MobBIOfake [228]. Replay-Attack [41] and 3DMAD [66] are the benchmarks considered for faces, while Biometrika, CrossMatch, Italdata, and Swipe are the fingerprint benchmarks here considered, all them recently

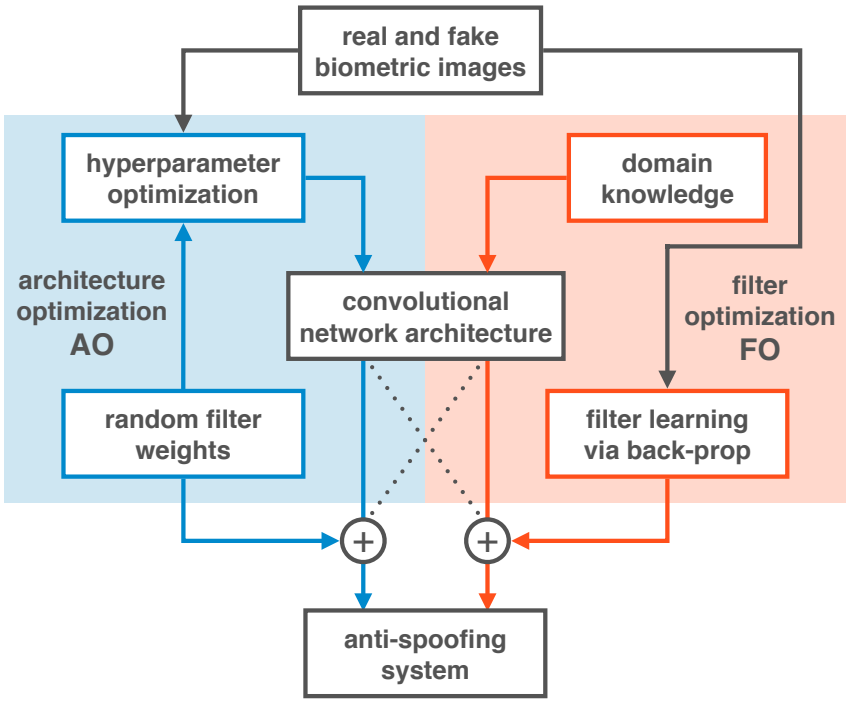


Figure 4.1: Schematic diagram detailing how anti-spoofing systems are built from spoofing detection benchmarks. Architecture optimization (AO) is shown on the left and filter optimization (FO) on the right. In this work, we not only evaluate AO and FO in separate, but also in combination, as indicated by the crossing dotted lines.

used in the 2013 Fingerprint Liveness Detection Competition (LivDet'13) [83].

Results outperform state-of-the-art counterparts in eight of the nine cases and observe a balance in terms of performance between AO and FO, with one performing better than the other depending on the sample size and problem difficulty. In some cases, we also show that when both approaches are combined, we can obtain performance levels that neither one can obtain by itself. Moreover, by observing the behaviour of AO and FO, we take advantage of domain knowledge to propose a single new convolutional architecture that push performance in five problems even further, sometimes by a large margin, as in CrossMatch (68.80% *v.* 98.23%).

The experimental results strongly indicate that convolutional networks can be readily used for robust spoofing detection. Indeed, we believe that data-driven solutions based on deep representations might be a valuable direction to this field of research, allowing the construction of systems with little effort even to image-based attack types yet to come.

We organized the remainder of this work into five sections. Section 4.2 presents previous anti-spoofing systems for the three biometric modalities covered in this paper, while Section 4.3 presents the considered benchmarks. Section 4.4 describes the methodology adopted for architecture optimization (AO) and filter optimization (FO) while Section 4.5 presents experiments, results, and comparisons with state-of-the-art methods. Finally, Section 4.6 concludes the paper and discusses some possible future directions.

4.2 Related Work

In this section, we review anti-spoofing related work for iris, face, and fingerprints, our focus in this paper.

4.2.1 Iris Spoofing

Daugman [60, Section 8 – Countermeasures against Subterfuge][1] was one of the first authors to discuss the feasibility of some attacks on iris recognition systems. The author proposed the use of Fast Fourier Transform to verify the high frequency spectral magnitude in the frequency domain.

The solutions for iris liveness detection available in the literature range from active solutions relying on special acquisition hardware [113,137,174] to software-based solutions relying on texture analysis of the effects of an attacker using color contact lenses with someone else's pattern printed onto them [259]. Software-based solutions have also explored the effects of cosmetic contact lenses [32,64,120,264]; pupil constriction [102]; and multi biometrics of electroencephalogram (EEG) and iris together [115], among others.

Galbally et al. [75] investigated 22 image quality measures (e.g., focus, motion, occlusion, and pupil dilation). The best features are selected through sequential floating feature selection (SFFS) [198] to feed a quadratic discriminant classifier. The authors validated the work on the BioSec [68,219] benchmark. Sequeira et al. [226] also explored image quality measures [75] and three classification techniques validating the work on the BioSec [68, 219] and Clarkson [224] benchmarks and introducing the MobBIOfake benchmark comprising 800 iris images from the MobBIO multimodal database [228].

Sequeira et al. [227] extended upon previous works also exploring quality measures. They first used a feature selection step on the features of the studied methods to obtain the "best features" and then used well-known classifiers for the decision-making. In addition, they applied iris segmentation [164] to obtaining the iris contour and adapted the feature extraction processes to the resulting non-circular iris regions. The validation considered five datasets (BioSec [68,219], MobBIOfake [228], Warsaw [51], Clarkson [224] and NotreDame [63].

Textures have also been explored for iris liveness detection. In the recent MobILive[2] [229] iris spoofing detection competition, the winning team explored three texture descriptors: Local Phase Quantization (LPQ) [171], Binary Gabor Pattern [272], and Local Binary Pattern (LBP) [169].

Analyzing printing regularities left in printed irises, Czajka [51] explored some peaks in the frequency spectrum were associated to spoofing attacks. For validation, the authors introduced the Warsaw dataset containing 729 fake images and 1,274 images of real eyes. In [224], The First Intl. Iris Liveness Competition in 2013, the Warsaw database was also evaluated, however, the best reported result achieved 11.95% of FRR and 5.25% of FAR by the University of Porto team.

Sun et al. [238] recently proposed a general framework for iris image classification based

[1] It also appears in a lecture of Daugman at IBC 2004 [58].
[2] MobLive 2014, Intl. Joint Conference on Biometrics (IJCB).

on a Hierarchical Visual Codebook (HVC). The HVC encodes the texture primitives of iris images and is based on two existing bag-of-words models. The method achieved state-of-the-art performance for iris spoofing detection, among other tasks.

In summary, iris anti-spoofing methods have explored hard-coded features through image-quality metrics, texture patterns, bags-of-visual-words and noise artifacts due to the recapturing process. The performance of such solutions vary significantly from dataset to dataset. Differently, here we propose the automatically extract vision meaningful features directly from the data using deep representations.

4.2.2 Face Spoofing

We can categorize the face anti-spoofing methods into four groups [214]: user behavior modeling, methods relying on extra devices [270], methods relying on user cooperation and, finally, data-driven characterization methods. In this section, we review data-driven characterization methods proposed in literature, the focus of our work herein.

Määttä et al. [151] used LBP operator for capturing printing artifacts and micro-texture patterns added in the fake biometric samples during acquisition. Schwartz et al. [214] explored color, texture, and shape of the face region and used them with Partial Least Square (PLS) classifier for deciding whether a biometric sample is fake or not. Both works validated the methods with the Print Attack benchmark [8]. Lee et al. [139] also explored image-based attacks and proposed the frequency entropy analysis for spoofing detection.

Pinto et al. [193] pioneered research on video-based face spoofing detection. They proposed visual rhythm analysis to capture temporal information on face spoofing attacks.

Mask-based face spoofing attacks have also been considered thus far. Erdogmus et al. [65] dealt with the problem through Gabor wavelets: local Gabor binary pattern histogram sequences [273] and Gabor graphs [262] with a Gabor-phase based similarity measure [90]. Erdogmus & Marcel [66] introduced the 3D Mask Attack database (3DMAD), a public available 3D spoofing database, recorded with Microsoft Kinect sensor.

Kose et al. [127] demonstrated that a face verification system is vulnerable to mask-based attacks and, in another work, Kose et al. [126] evaluated the anti-spoofing method proposed by Määttä et al. [151] (originally proposed to detect photo-based spoofing attacks). Inspired by the work of Tan et al. [245], Kose et al. [128] evaluated a solution based on reflectance to detect attacks performed with 3D masks.

Finally, Pereira et al. [184] proposed a score-level fusion strategy in order to detect various types of attacks. In a follow-up work, Pereira et al. [69] proposed an anti-spoofing solution based on the dynamic texture, a spatio-temporal version of the original LBP. Results showed that LBP-based dynamic texture description has a higher effectiveness than the original LBP.

In summary, similarly to iris spoofing detection methods, the available solutions in the literature mostly deal with the face spoofing detection problem through texture patterns (e.g., LBP-like detectors), acquisition telltales (noise), and image quality metrics. Here, we approach the proplem by extracting meaningful features directly from the data regardless of the input type (image, video, or 3D masks).

4.2.3 Fingerprint Spoofing

We can categorize fingerprint spoofing detection methods roughly into two groups: hardware-based (exploring extra sensors) and software-based solutions (relying only on the information acquired by the standard acquisition sensor of the authentication system) [83].

Galbally et al. [71] proposed a set of feature for fingerprint liveness detection based on quality measures such as ridge strength or directionality, ridge continuity, ridge clarity, and integrity of the ridge-valley structure. The validation considered the three benchmarks used in LivDet 2009 – Fingerprint competition [159] captured with different optical sensors: Biometrika, CrossMatch, and Identix. Later work [72] explored the method in the presence of gummy fingers.

Ghiani et al. [82] explored LPQ [171], a method for representing all spectrum characteristics in a compact feature representation form. The validation considered the four benchmarks used in the LivDet 2011 – Fingerprint competition [266].

Gragnaniello et al. [86] explored the Weber Local Image Descriptor (WLD) for liveness detection, well suited to high-contrast patterns such as the ridges and valleys of fingerprints images. In addition, WLD is robust to noise and illumination changes. The validation considered the LivDet 2009 and LivDet 2011 – Fingerprint competition datasets.

Jia et al. [112] proposed a liveness detection scheme based on Multi-scale Block Local Ternary Patterns (MBLTP). Differently of the LBP, the Local Ternary Pattern operation is done on the average value of the block instead of the pixels being more robust to noise. The validation considered the LivDet 2011 – Fingerprint competition benchmarks.

Ghiani et al. [81] explored Binarized Statistical Image Features (BSIF) originally proposed by Kannala et al. [114]. The BSIF was inspired in the LBP and LPQ methods. In contrast to LBP and LPQ approaches, BSIF learns a filter set by using statistics of natural images [104]. The validation considered the LivDet 2011 – Fingerprint competition benchmarks.

Recent results reported in the LivDet 2013 Fingerprint Liveness Detection Competition [81] show that fingerprint spoofing attack detection task is still an open problem with results still far from a perfect classification rate.

We notice that most of the groups approach the problem with hard-coded features sometimes exploring quality metrics related to the modality (e.g., directionality and ridge strength), general texture patterns (e.g., LBP-, MBLTP-, and LPQ-based methods), and filter learning through natural image statistics. This last approach seems to open a new research trend, which seeks to model the problem learning features directly from the data. We follow this approach in this work, assuming little a priori knowledge about acquisition-level biometric spoofing and exploring deep representations of the data.

4.2.4 Multi-modalities

Recently, Galbally et al. [74] proposed a general approach based on 25 image quality features to detect spoofing attempts in face, iris, and fingerprint biometric systems. Our work is similar to theirs in goals, but radically different with respect to the methods. Instead of relying on prescribed image quality features, we build features that would be

hardly thought by a human expert with AO and FO. Moreover, here we evaluate our systems in more recent and updated benchmarks.

4.3 Benchmarks

In this section, we describe the benchmarks (datasets) that we consider in this work. All them are publicly available upon request and suitable for evaluating countermeasure methods to iris, face and fingerprint spoofing attacks. Table 4.1 shows the major features of each one and in the following we describe their details.

4.3.1 Iris Spoofing Benchmarks

Biosec

This benchmark was created using iris images from 50 users of the BioSec [219]. In total, there are 16 images for each user (2 sessions $\times$ 2 eyes $\times$ 4 images), totalizing 800 valid access images. To create spoofing attempts, the original images from Biosec were preprocessed to improve quality and printed using an HP Deskjet 970cxi and an HP LaserJet 4200L printers. Finally, the iris images were recaptured with the same iris camera used to capture the original images.

Warsaw

This benchmark contains 1274 images of 237 volunteers representing valid accesses and 729 printout images representing spoofing attempts, which were generated by using two printers: (1) a HP LaserJet 1320 used to produce 314 fake images with 600 dpi resolution, and (2) a Lexmark C534DN used to produce 415 fake images with 1200 dpi resolution. Both real and fake images were captured by an IrisGuard AD100 biometric device.

MobBIOfake

This benchmark contains live iris images and fake printed iris images captured with the same acquisition sensor, i.e., a mobile phone. To generate fake images, the authors first performed a preprocessing in the original images to enhance the contrast. The preprocessed images were then printed with a professional printer on high quality photographic paper.

4.3.2 Video-based Face Spoofing Benchmarks

Replay-Attack

This benchmark contains short video recordings of both valid accesses and video-based attacks of 50 different subjects. To generate valid access videos, each person was recorded in two sessions in a controlled and in an adverse environment with a regular webcam. Then, spoofing attempts were generated using three techniques: (1) *print attack*, which presents to the acquisition sensor hard copies of high-resolution digital photographs printed with

Table 4.1: Main features of the benchmarks considered herein.

Modality	Benchmark/Dataset	Color	Dimension	# Training			# Testing			# Development		
			$cols \times rows$	Live	Fake	Total	Live	Fake	Total	Live	Fake	Total
Iris	Warsaw [51]	No	640×480	228	203	431	624	612	1236			
	Biosec [219]	No	640×480	200	200	400	600	600	1200			
	MobBIOfake [228]	Yes	250×200	400	400	800	400	400	800			
Face	Replay-Attack [36]	Yes	320×240	600	3000	3600	4000	800	4800	600	3000	3600
	3dMad [42]	Yes	640×480	350	350	700	250	250	500	250	250	500
Fingerprint	Biometrika [83]	No	312×372	1000	1000	2000	1000	1000	2000			
	CrossMatch [83]	No	800×750	1250	1000	2250	1250	1000	2250			
	Italdata [83]	No	640×480	1000	1000	2000	1200	1000	2000			
	Swipe [83]	No	208×1500	1221	979	2200	1153	1000	2153			

a Triumph-Adler DCC 2520 color laser printer; (2) *mobile attack*, which presents to the acquisition sensor photos and videos taken with an iPhone using the iPhone screen; and (3) *high-definition attack*, in which high resolution photos and videos taken with an iPad are presented to the acquisition sensor using the iPad screen.

3DMAD

This benchmark consists of real videos and fake videos made with people wearing masks. A total of 17 different subjects were recorded with a Microsoft Kinect sensor, and videos were collected in three sessions. For each session and each person, five videos of 10 seconds were captured. The 3D masks were produced by `ThatsMyFace.com` using one frontal and two profile images of each subject. All videos were recorded by the same acquisition sensor.

4.3.3 Fingerprint Spoofing Benchmarks

LivDet2013

This dataset contains four sets of real and fake fingerprint readings performed in four acquisition sensors: Biometrika FX2000, Italdata ET10, Crossmatch L Scan Guardian, and Swipe. For a more realistic scenario, fake samples in Biometrika and Italdata were generated without user cooperation, while fake samples in Crossmatch and Swipe were generated with user cooperation. Several materials for creating the artificial fingerprints were used, including gelatin, silicone, latex, among others.

4.3.4 Remark

Images found in these benchmarks can be observed in Fig. 4.5 of Section 4.5. As we can see, variability exists not only across modalities, but also within modalities. Moreover, it is rather unclear what features might discriminate real from spoofed images, which suggests that the use of a methodology able to use data to its maximum advantage might be a promising idea to tackle such set of problems in a principled way.

4.4 Methodology

In this section, we present the methodology for architecture optimization (AO) and filter optimization (FO) as well as details about how benchmark images are preprocessed, how AO and FO are evaluated across the benchmarks, and how these methods are implemented.

4.4.1 Architecture Optimization (AO)

Our approach for AO builds upon the work of Pinto et al. [195] and Bergstra et al. [20], i.e., fundamental, feedforward convolutional operations are stacked by means of hyperparameter optimization, leading to effective yet simple convolutional networks that do not

require expensive filter optimization and from which prediction is done by linear support vector machines (SVMs).

Operations in convolutional networks can be viewed as linear and non-linear transformations that, when stacked, extract high level representations of the input. Here we use a well-known set of operations called (i) *convolution* with a bank of filters, (ii) rectified linear *activation*, (iii) spatial *pooling*, and (iv) *local normalization*. **Appendix B** provides a detailed definition of these operations.

We denote as *layer* the combination of these four operations in the order that they appear in the left panel of Fig. 4.2. Local normalization is optional and its use is governed by an additional "yes/no" hyperparameter. In fact, there are other six hyperparameters, each of a particular operation, that have to be defined in order to instantiate a layer. They are presented in the lower part of the left panel in Fig. 4.2 and are in accordance to the definitions of **Appendix B**.

Considering one layer and possible values of each hyperparameter, there are over 3,000 possible layer architectures, and this number grows exponentially with the number of layers, which goes up to three in our case (Fig. 4.2 right panel). In addition, there are network-level hyperparameters, such as the size of the input image, that expand possibilities to a myriad potential architectures.

The overall set of possible hyperparameter values is called *search space*, which in this case is discrete and contains variables that are only meaningful in combination with others. For example, hyperparameters of a given layer are just meaningful if the candidate architecture has actually that number of layers. In spite of the intrinsic difficulty in optimizing architectures in this space, *random search* has played and important role in problems of this type [18, 195] and it is the strategy of our choice due to its effectiveness and simplicity.

We can see in Fig. 4.2 that a three-layered network has a total of 25 hyperparameters, seven per layer and four at network level. They are all defined in **Appendix B** with the exception of *input size*, which seeks to determine the best size of the image's greatest axis (rows or columns) while keeping its aspect ratio. Concretely, random search in this paper can be described as follows:

1. Randomly — and uniformly, in our case — sample values from the hyperparameter *search space*;

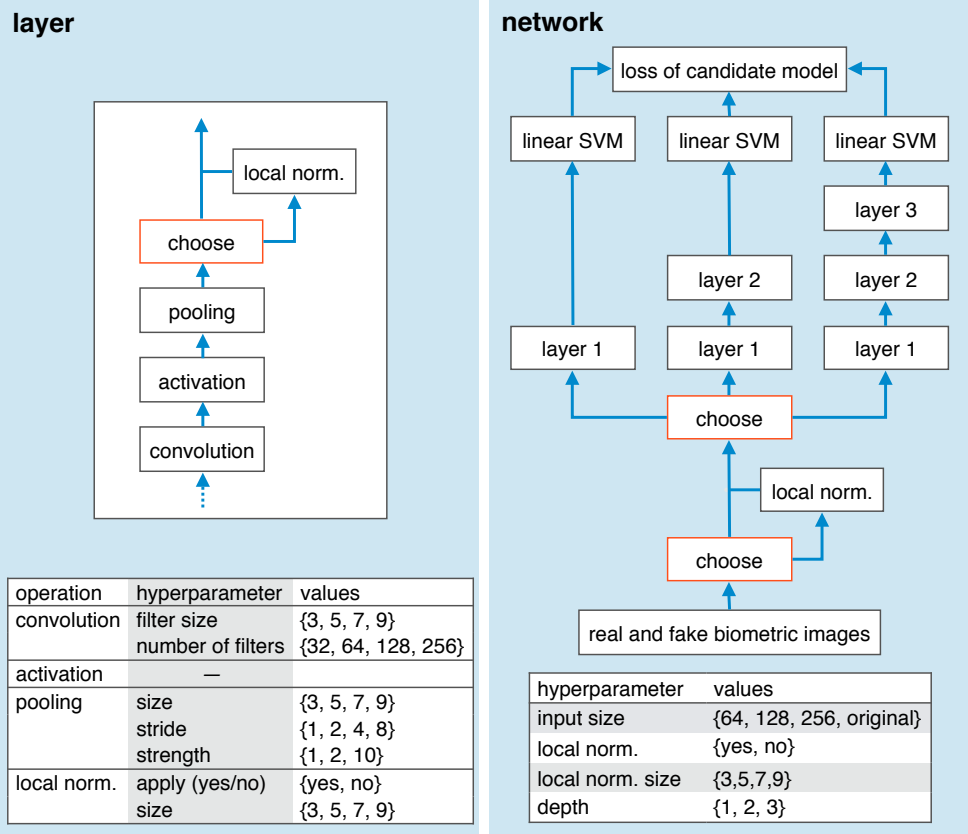
2. Extract features from real and fake training images with the candidate architecture;

3. Evaluate the architecture according to an *optimization objective* based on linear SVM scores;

4. Repeat steps 1–3 until a *termination criterion* is met;

5. Return the best found convolutional architecture.

Even though there are billions of possible networks in the search space (Fig. 4.2), it is important to remark that not all candidate networks are valid. For example, a large



operation	hyperparameter	values
convolution	filter size	{3, 5, 7, 9}
	number of filters	{32, 64, 128, 256}
activation	—	
pooling	size	{3, 5, 7, 9}
	stride	{1, 2, 4, 8}
	strength	{1, 2, 10}
local norm.	apply (yes/no)	{yes, no}
	size	{3, 5, 7, 9}

hyperparameter	values
input size	{64, 128, 256, original}
local norm.	{yes, no}
local norm. size	{3,5,7,9}
depth	{1, 2, 3}

Figure 4.2: Schematic diagram for architecture optimization (AO) illustrating how operations are stacked in a layer (left) and how the network is instantiated and evaluated according to possible hyperparameter values (right). Note that a three-layered convolutional network of this type has a total of 25 hyperparameters governing both its architecture and its overall behaviour through a particular instance of stacked operations.

number of candidate architectures (i.e., points in the search space) would produce representations with spatial resolution smaller than one pixel. Hence, they are naturally unfeasible. Additionally, in order to avoid very large representations, we discard in advance candidate architectures whose intermediate layers produce representations of over 600K elements or whose output representation has over 30K elements.

Filter weights are randomly generated for AO. This strategy has been successfully used in the vision literature [111, 194, 195, 221] and is essential to make AO practical, avoiding the expensive filter optimization (FO) part in the evaluation of candidate architectures. We sample weights from a uniform distribution $U(0, 1)$ and normalize the filters to zero mean and unit norm in order to ensure that they are spread over the unit sphere. When coupled with rectified linear activation (**Appendix B**), this sampling enforces sparsity in the network by discarding about 50% of the expected filter responses, thereby improving the overall robustness of the feature extraction.

A candidate architecture is evaluated by first extracting deep representations from real and fake images and later training hard-margin linear SVMs (C=10^5) on these representations. We observed that the sensitivity of the performance measure was saturating with traditional 10-fold cross validation (CV) in some benchmarks. Therefore, we opted

for a different validation strategy. Instead of training on nine folds and validating on one, we train on one fold and validate on nine. Precisely, the *optimization objective* is the mean detection accuracy obtained from this adapted cross-validation scheme, which is maximized during the optimization.

For generating the 10 folds, we took special care in putting all samples of an individual in the same fold to enforce robustness to cross-individual spoofing detection in the optimized architectures. Moreover, in benchmarks where we have more than one attack type (e.g., Replay-Attack and LivDet2013, see Section 4.3), we evenly distributed samples of each attack type across all folds in order to enforce that candidate architectures are also robust to different types of attack.

Finally, the *termination criterion* of our AO procedure simply consists of counting the number of valid candidate architectures and stopping the optimization when this number reaches 2,000.

4.4.2 Filter Optimization (FO)

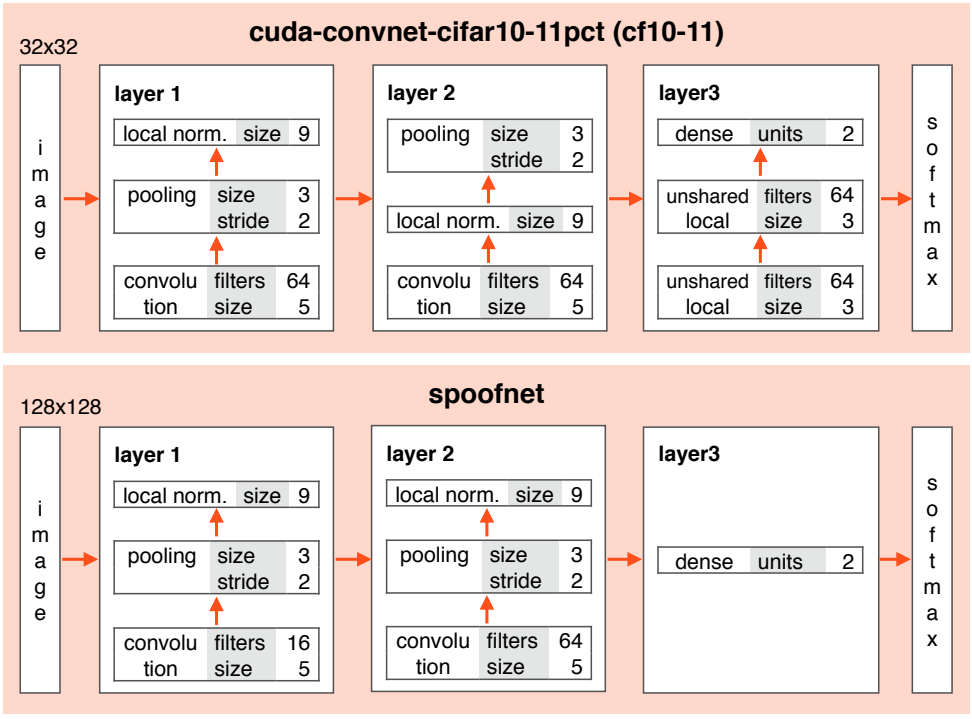


Figure 4.3: Architecture of convolutional network found in the Cuda-convnet library and here used as reference for filter optimization (*cf10-11*, top). Proposed network architecture extending upon *cf10-11* to better suiting spoofing detection problems (*spoofnet*, bottom). Both architectures are typical examples where domain knowledge has been incorporated for increased performance.

We now turn our attention to a different approach for tackling the problem. Instead of optimizing the architecture, we explore the filter weights and how to learn them for better characterizing real and fake samples. Our approach for FO is at the origins of

convolutional networks and consists of learning filter weights via the well-known back-propagation algorithm [135]. Indeed, due to a refined understanding of the optimization process and the availability of plenty of data and processing power, back-propagation has been the gold standard method in deep networks for computer vision in the last years [130, 233, 271].

For optimizing filters, we need to have an already defined architecture. We start optimizing filters with a standard public convolutional network and training procedure. This network is available in the Cuda-convnet library [129] and is currently one of the best performing architectures in CIFAR-10,[3] a popular computer vision benchmark in which such network achieves 11% of classification error. Hereinafter, we call this network *cuda-convnet-cifar10-11pct*, or simply *cf10-11*.

Fig. 4.3 depicts the architecture of *cf10-11* in the top panel and is a typical example where domain knowledge has been incorporated for increased performance. We can see it as a three-layered network in which the first two layers are convolutional, with operations similar to the operations used in architecture optimization (AO). In the third layer, *cf10-11* has two sublayers of unshared local filtering and a final fully-connected sublayer on top of which softmax regression is performed. A detailed explanation of the operations in *cf10-11* can be found in [129].

In order to train *cf10-11* in a given benchmark, we split the training images into four batches observing the same balance of real and fake images. After that, we follow a procedure similar to the original[4] for training *cf10-11* in all benchmarks, which can be described as follows:

1. For 100 epochs, train the network with a learning rate of 10^{-3} by considering the first three batches for training and the fourth batch for validation;

2. For another 40 epochs, resume training now considering all four batches for training;

3. Reduce the learning rate by a factor of 10, and train the network for another 10 epochs;

4. Reduce the learning rate by another factor of 10, and train the network for another 10 epochs.

After evaluating filter learning on the *cf10-11* architecture, we also wondered how filter learning could benefit from an optimized architecture incorporating domain-knowledge of the problem. Therefore, extending upon the knowledge obtained with AO as well as with training *cf10-11* in the benchmarks, we derived a new architecture for spoofing detection that we call *spoofnet*. Fig. 4.3 illustrates this architecture in the bottom panel and has three key differences as compared to *cf10-11*. First, it has 16 filters in the first layer instead of 64. Second, operations in the second layer are stacked in the same order that we used when optimizing architectures (AO). Third, we removed the two unshared local filtering operations in the third layer, as they seem inappropriate in a problem where object structure is irrelevant.

[3] http://www.cs.toronto.edu/~kriz/cifar.html
[4] https://code.google.com/p/cuda-convnet/wiki/Methodology.

Table 4.2: Input image dimensionality after basic preprocessing on face and fingerprint images (highlighted). See Section 4.4.3 for details.

Modality	Benchmark	Dimensions *columns* × *rows*
Iris	Warsaw [51]	640 × 480
	Biosec [219]	640 × 480
	MobBIOfake [228]	250 × 200
Face	Replay-Attack [36]	**200 × 200**
	3DMAD [42]	**200 × 200**
Fingerprint	Biometrika [83]	**312 × 372**
	CrossMatch [83]	**480 × 675**
	Italdata [83]	**384 × 432**
	Swipe [83]	**187 × 962**

These three modifications considerably dropped the number of weights in the network and this, in turn, allowed us to increase of size of the input images from 32×32 to 128×128. This is the fourth and last modification in *spoofnet*, and we believe that it might enable the network to be more sensitive to subtle local patterns in the images.

In order to train *spoofnet*, the same procedure used to train *cf10-11* is considered except for the initial learning rate, which is made 10^{-4}, and for the number of epochs in each step, which is doubled. These modifications were made because of the decreased learning capacity of the network.

Finally, in order to reduce overfitting, data augmentation is used for training both networks according to the procedure of [130]. For *cf10-11*, five 24×24 image patches are cropped out from the 32×32 input images. These patches correspond to the four corners and central region of the original image, and their horizontal reflections are also considered. Therefore, ten training samples are generated from a single image. For *spoofnet*, the procedure is the same except for the fact that input images have 128×128 pixels and cropped regions are of 112×112 pixels. During prediction, just the central region of the test image is considered.

4.4.3 Elementary Preprocessing

A few basic preprocessing operations were executed on face and fingerprint images in order to properly learn representations for these benchmarks. This preprocessing led to images with sizes as presented in Table 4.2 and are described in the next two sections.

Face Images

Given that the face benchmarks considered in this work are video-based, we first evenly subsample 10 frames from each input video. Then, we detect the face position using Viola & Jones [255] and crop a region of 200×200 pixels centered at the detected window.

Fingerprint Images

Given the diverse nature of images captured from different sensors, here the preprocessing is defined according to the sensor type.

1. *Biometrika*: we cropped the central region of size in columns and rows corresponding to 70% of the original image dimensions.

2. *Italdata* and *CrossMatch*: we cropped the central region of size in columns and rows respectively corresponding to 60% and 90% of the original image columns and rows.

3. *Swipe*: As the images acquired by this sensor contain a variable number of blank rows at the bottom, the average number of non-blank rows M was first calculated from the training images. Then, in order to obtain images of a common size with non-blank rows, we removed their blank rows at the bottom and rescaled them to M rows. Finally, we cropped the central region corresponding to 90% of original image columns and M rows.

The rationale for these operations is based on the observation that fingerprint images in LivDet2013 tend to have a large portion of background content and therefore we try to discard such information that could otherwise mislead the representation learning process. The percentage of cropped columns and rows differs among sensors because they capture images of different sizes with different amounts of background.

For architecture optimization (AO), the decision to use image color information was made according to 10-fold validation (see Section 4.4.1), while for filter optimization (FO), color information was considered whenever available for a better approximation with the standard cf10-11 architecture. Finally, images were resized to 32×32 or 128×128 to be taken as input for the cf10-11 and spoofnet architectures, respectively.

4.4.4 Evaluation Protocol

For each benchmark, we learn deep representations from their training images according to the methodology described in Section 4.4.1 for architecture optimization (AO) and in Section 4.4.2 for filter optimization (FO). We follow the standard evaluation protocol of all benchmarks and evaluate the methods in terms of detection accuracy (ACC) and half total error rate (HTER), as these are the metrics used to assess progress in the set of benchmarks considered herein. Precisely, for a given benchmark and convolutional network already trained, results are obtained by:

1. Retrieving prediction scores from the testing samples;

2. Calculating a threshold τ above which samples are predicted as attacks;

3. Computing ACC and/or HTER using τ and test predictions.

The way that τ is calculated differs depending on whether the benchmark has a development set or not (Table 4.1). Both face benchmarks have such a set and, in this

case, we simply obtain τ from the predictions of the samples in this set. Iris and fingerprint benchmarks have no such a set, therefore τ is calculated depending on whether the convolutional network was learned with AO or FO.

In case of AO, we calculate τ by joining the predictions obtained from 10-fold validation (see Section 4.4.1) in a single set of positive and negative scores, and τ is computed as the point that lead to an equal error rate (EER) on the score distribution under consideration. In case of FO, scores are probabilities and we assume $\tau = 0.5$. ACC and HTER are then trivially computed with τ on the testing set.

It is worth noting that the Warsaw iris benchmark provides a supplementary testing set that here we merge with the original testing set in order to replicate the protocol of [224]. Moreover, given face benchmarks are video-based and that in our methodology we treat them as images (Section 4.4.3), we perform a score-level fusion of the samples from the same video according to the max rule [217]. This fusion is done before calculating τ.

4.4.5 Implementation

Our implementation for architecture optimization (AO) is based on Hyperopt-convnet [17] which in turn is based on Theano [19]. LibSVM [37] is used for learning the linear classifiers via Scikit-learn.[5] The code for feature extraction runs on GPUs due to Theano and the remaining part is multithreaded and runs on CPUs. We extended Hyperopt-convnet in order to consider the operations and hyperparameters as described in **Appendix B** and Section 4.4.1 and we will make the source code freely available in [40]. Running times are reported with this software stack and are computed in an Intel i7 @3.5GHz with a Tesla K40 that, on average, takes less than one day to optimize an architecture — i.e., to probe 2,000 candidate architectures — for a given benchmark.

As for filter optimization (FO), Cuda-convnet [129] is used. This library has an extremely efficient implementation to train convolutional networks via back-propagation on NVIDIA GPUs. Moreover, it provides us with the cf10-11 convolutional architecture taken in this work as reference for FO.

4.5 Experiments and Results

In this section, we evaluate the effectiveness of the proposed methods for spoofing detection. We show experiments for the architecture optimization and filter learning approaches along with their combination for detecting iris, face, and fingerprint spoofing on the nine benchmarks described in Section 4.3. We also present results for the *spoofnet*, which incorporates some domain-knowledge on the problem. We compare all of the results with the state-of-the-art counterparts. Finally, we discuss the pros and cons of using such approaches and their combination along with efforts to understand the type of features learned and some effeciency questions when testing the proposed methods.

[5] http://scikit-learn.org

Table 4.3: Overall results considering relevant information of the best found architectures, detection accuracy (ACC) and HTER values according to the evaluation protocol, and state-of-the-art (SOTA) performance.

modality	benchmark	architecture optimization (AO)					our results		SOTA results		Ref.
		time (secs.)	size (pixels)	layers (#)	features (#)	objective (%)	ACC (%)	HTER (%)	ACC (%)	HTER (%)	
iris	Warsaw	52+35	640	2	$10 \times 15 \times 64$ (9600)	98.21	**99.84**	0.16	97.50	—	[51]
	Biosec	80+34	640	3	$2 \times 5 \times 256$ (2560)	97.56	98.93	1.17	**100.00**	—	[75]
	MobBIOfake	18+37	250	2	$5 \times 7 \times 256$ (8960)	98.94	98.63	1.38	**99.75**	—	[229]
face	Replay-Attack	69+15	256	2	$3 \times 3 \times 256$ (2304)	94.65	98.75	**0.75**	—	5.11	[124]
	3DMAD	55+15	128	2	$5 \times 5 \times 64$ (1600)	98.68	100.00	**0.00**	—	0.95	[65]
fingerprint	Biometrika	66+25	256	2	$2 \times 2 \times 256$ (1024)	90.11	96.50	3.50	**98.30**	—	[83]
	Crossmatch	112+12	675	3	$2 \times 3 \times 256$ (1536)	91.70	**92.09**	8.44	68.80	—	[83]
	Italdata	46+27	432	3	$16 \times 13 \times 128$ (26624)	86.89	97.45	2.55	**99.40**	—	[83]
	Swipe	97+51	962	2	$53 \times 3 \times 32$ (5088)	90.32	88.94	11.47	**96.47**	—	[83]

4.5.1 Architecture Optimization (AO)

Table 4.3 presents AO results in detail as well as previous state-of-the-art (SOTA) performance for the considered benchmarks. With this approach, we can outperform four SOTA methods in all three biometric modalities. Given that AO assumes little knowledge about the problem domain, this is remarkable. Moreover, performance is on par in other four benchmarks, with the only exception of Swipe. Still in Table 4.3, we can see information about the best architecture such as time taken to evaluate it (feature extraction + 10-fold validation), input size, depth, and dimensionality of the output representation in terms of *columns* $\times$ *rows* $\times$ *feature maps*.

Regarding the number of layers in the best architectures, we can observe that six out of nine networks use two layers, and three use three layers. We speculate that the number of layers obtained is a function of the problem complexity. In fact, even though there are many other hyperparameters involved, the number of layers play an important role in this issue, since it directly influences the level of non-linearity and abstraction of the output with respect to the input.

With respect to the input size, we can see in comparison with Table 4.2, that the best performing architectures often use the original image size. This was the case for all iris benchmarks and for three (out of four) fingerprint benchmarks. For face benchmarks, a larger input was preferred for Replay-Attack, while a smaller input was preferred for 3DMAD. We hypothesize that this is also related to the problem difficulty, given that Replay-Attack seems to be more difficult, and that larger inputs tend to lead to larger networks.

We still notice that the dimensionality of the obtained representations are, in general, smaller than 10K features, except for Italdata. Moreover, for the face and iris benchmarks, it is possible to roughly observe a relationship between the optimization objective calculated in the training set and the detection accuracy measure on the testing set (Section 4.4.4), which indicates the appropriateness of the objective for these tasks. However, for the fingerprint benchmarks, this relationship does not exist, and we accredit this to either a deficiency of the optimization objective in modelling these problems or to the existence of artifacts in the training set misguiding the optimization.

4.5.2 Filter Optimization (FO)

Table 4.4 shows the results for FO, where we repeat architecture optimization (AO) results (with random filters) in the first column to facilitate comparisons. Overall, we can see that both networks, *cf10-11* and *spoofnet* have similar behavior across the biometric modalities.

Surprisingly, *cf10-11* obtains excellent performance in all four fingerprint benchmarks as well as in the MobBIOFake, exceeding SOTA in three cases, in spite of the fact that it was used without any modification. However, in both face problems and in two iris problems, *cf10-11* performed poorly. Such difference in performance was not possible to anticipate by observing training errors, which steadily decreased in all cases until training was stopped. Therefore, we believe that in these cases FO was misguided by the lack of training data or structure in the training samples irrelevant to the problem.

Table 4.4: Results for filter optimization (FO) in *cf10-11* and *spoofnet* (Fig. 4.3). Even though both networks present similar behavior, *spoofnet* is able to push performance even further in problems which *cf10-11* was already good for. Architecture optimization (AO) results (with random filters) are shown in the first column to facilitate comparisons.

		filter			
modality		random	optimized		
(metric)	benchmark	AO	*cf10-11*	*spoofnet*	SOTA
iris	Warsaw	**99.84**	67.20	66.42	97.50
(ACC)	Biosec	98.93	59.08	47.67	**100.00**
	MobBIOfake	98.63	99.13	**100.00**	99.75
face	Replay-Attack	**0.75**	55.13	55.38	5.11
(HTER)	3DMAD	**0.00**	40.00	24.00	0.95
fingerprint	Biometrika	96.50	98.50	**99.85**	98.30
(ACC)	Crossmatch	92.09	97.33	**98.23**	68.80
	Italdata	97.45	97.35	**99.95**	99.40
	Swipe	88.94	98.70	**99.08**	96.47

To reinforce this claim, we performed experiments with filter optimization (FO) in *spoofnet* by varying the training set size with 20%, 40%, and 50% of fingerprint benchmarks. As expected, in all cases, the less training examples, the worse is the generalization of the *spoofnet* (lower classification accuracies). Considering the training phase, for instance, when using 50% of training set or less, the accuracy achieved by the learned representation is far worse than the one achieved when using 100% of training data. This fact reinforces the conclusion presented herein regarding the small sample size problem. Maybe a fine-tuning of some parameters, such as the number of training epochs and the learning rates, can diminish the impact of the small sample size problem stated here, however, this is an open research topic by itself.

For *spoofnet*, the outcome is similar. As we expected, the proposed architecture was able to push performance even further in problems which *cf10-11* was already good for, outperforming SOTA in five out of nine benchmarks. This is possibly because we made the *spoofnet* architecture simpler, with less parameters, and taking input images with a size better suited to the problem.

As compared to the results in AO, we can observe a good balance between the approaches. In AO, the resulting convolutional networks are remarkable in the face benchmarks. In FO, networks are remarkable in fingerprint problems. While in AO all optimized architectures have good performance in iris problems, FO excelled in one of these problems, MobBIOFake, with a classification accuracy of 100%. In general, AO seems to result in convolutional networks that are more stable across the benchmarks, while FO shines in problems in which learning effectively occurs. Considering both AO and FO, we can see in Table 4.4 that we outperformed SOTA methods in eight out of nine benchmarks. The only benchmark were SOTA performance was not achieved is Biosec, but even in this case the result obtained with AO is competitive.

Understanding how a set of deep learned features capture properties and nuances of a problem is still an open question in the vision community. However, in an attempt to understand the behavior of the operations applied onto images after they are forwarded through the first convolutional layer, we generate Fig. 4.4a that illustrates the filters learned via backpropagation algorithm and Figs. 4.4b and 4.4c showing the mean of real and fake images that compose the test set, respectively. To obtain output values from the first convolutional layer and get a sense of them, we also instrumented the *spoofnet* convolutional network to forward the real and fake images from the test set through network. Figs 4.4d and 4.4e show such images for the real and fake classes, respectively.

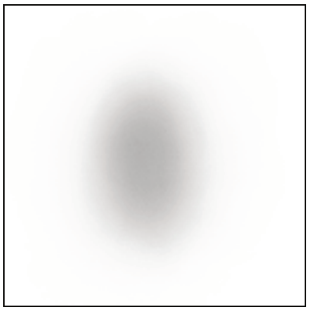
We can see in Fig. 4.4a that the filters learned patterns resemble textural patterns instead of edge patterns as usually occurs in several computer vision problems [130, 173]. This is particularly interesting and in line with several anti-spoofing methods in the the literature which also report good results when exploring texture information [83, 151].

In addition, Fig. 4.4b and 4.4c show there are differences between real and fake images from test, although apparently small in such a way that a direct analysis of the images would not be enough for decision making. However, when we analyze the mean activation maps for each class, we can see more interesting patterns. In Figs. 4.4d and 4.4e, we have sixteen pictures with 128×128 pixel resolution. These images correspond to the sixteen filters that composing the first layer of the *spoofnet*. Each position (x, y) in these 128×128 images corresponds to a 5×5 area (receptive field units) in the input images. Null values in a given unit means that the receptive field of the unit was not able to respond to the input stimuli. In contrast, non-null values mean that the receptive field of the unit had a responsiveness to the input stimuli.

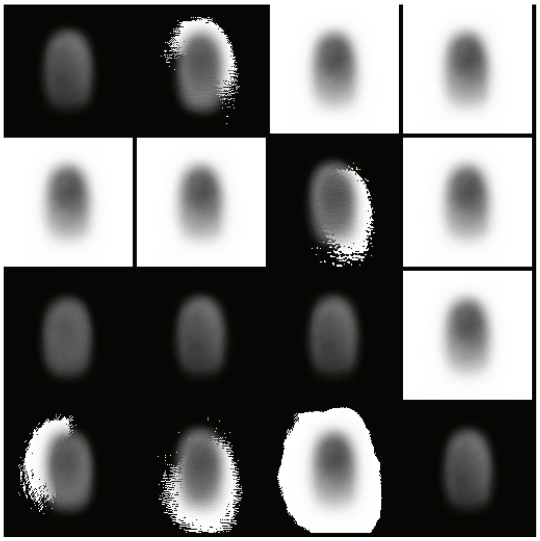
We can see that six filters have a high responsiveness to the background information of the input images (filters predominantly white) whilst ten filters did not respond to background information (filters predominantly black). From left to right, top to bottom, we can see also that the images corresponding to the filters 2, 7, 13, 14 and 15 have high responsiveness to information surrounding the central region of the sensor (usually where fingerprints are present) and rich in texture datails. Although these regions of high and low responsiveness are similar for both classes we can notice some differences. A significant difference in this first convolutional layer to images for the different classes is that the response of the filters regarding to fake images (Fig 4.4e) generates a blurring pattern, unlike the responses of the filters regarding to real images (Fig 4.4d) which generate a sharper pattern. We believe that the same way as the first layer of a convolutional network has the ability to respond to simple and relevant patterns (edge information) to a problem of recognition objects in general, in computer vision, the first layer in the *spoofnet* also was able to react to a simple pattern recurrent in spoof problems, the blurring effect, an artifact previously explored in the literature [74]. Finally, we are exploring visualisation only of the first layer; subsequent layers of the network can find new patterns in these regions activated by the first layer further emphasizing class differences.

(a) Filter weights of the first convolutional layer that were learned using the backpropagation algorithm.

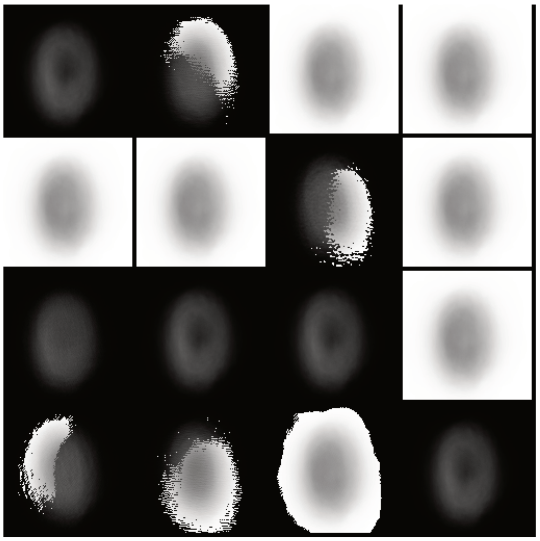
(b) Mean for real images (test set).



(c) Mean for fake images (test set).



(d) Activation maps for real images.



(e) Activation maps for fake images.

Figure 4.4: Activation maps of the filters that compose the first convolutional layer when forwarding real and fake images through the network.

4.5.3 Interplay between AO and FO

In the previous experiments, architecture optimization (AO) was evaluated using random filters and filter optimization (FO) was carried out in the predefined architectures *cf10-11* and *spoofnet*. A natural question that emerges in this context is how these methods would perform if we (i) combine AO and FO and if we (ii) consider random filters in *cf10-11* and *spoofnet*.

Results from these combinations are available in Table 4.5 and show a clear pattern. When combined with AO, FO again exceeds previous SOTA in all fingerprint benchmarks and performs remarkably good in MobBIOFake. However, the same difficulty found by FO in previous experiments for both face and two iris benchmarks is also observed here. Even though *spoofnet* performs slightly better than AO in the cases where SOTA is exceeded (Table 4.4), it is important to remark that our AO approach may result in architectures with a much larger number of filter weights to be optimized, and this may

Table 4.5: Results for architecture and filter optimization (AO+FO) along with *cf10-11* and *spoofnet* networks considering random weights. AO+FO show compelling results for fingerprints and one iris benchmark (MobBIOFake). We can also see that *spoofnet* can benefit from random filters in situations it was not good for when using filter learning (e.g., Replay-Attack).

		filter			
modality		optimized	random		
(metric)	benchmark	AO	*cf10-11*	*spoofnet*	SOTA
iris	Warsaw	59.55	87.06	96.44	**97.50**
(ACC)	Biosec	57.50	97.33	97.42	**100.00**
	MobBIOfake	99.38	77.00	72.00	**99.75**
face	Replay-Attack	55.88	5.62	**3.50**	5.11
(HTER)	3DMAD	40.00	8.00	4.00	**0.95**
fingerprint	Biometrika	**99.30**	77.45	94.70	98.30
(ACC)	Crossmatch	**98.04**	83.11	87.82	68.80
	Italdata	**99.45**	76.45	91.05	99.40
	Swipe	**99.08**	87.60	96.75	96.47

have benefited *spoofnet*.

It is also interesting to observe in Table 4.5 the results obtained with the use of random filters in *cf10-11* and *spoofnet*. The overall balance in performance of both networks across the benchmarks is improved, similar to what we have observed with the use of random filters in Table 4.3. An striking observation is that *spoofnet* with random filters exceed previous SOTA in Replay-Attack, and this supports the idea that the poor performance of *spoofnet* in Replay-Attack observed in the FO experiments (Table 4.4) was not a matter of architecture.

4.5.4 Runtime

We estimate time requirements for anti-spoofing systems built with convolutional networks based on measurements obtained in architecture optimization (AO). We can see in Table 4.3 that the most computationally intensive deep representation is the one found for the Swipe benchmark, and demands 148 (97+51) seconds to process 2,200 images. Such a running time is only possible due to the GPU+CPU implementation used (Section 4.4.5), which is critical for this type of learning task. In a hypothetical operational scenario, we could ignore the time required for classifier training (51 seconds, in this case). Therefore, we can estimate that, on average, a single image captured by a Swipe sensor would require approximately 45 milliseconds — plus a little overhead — to be fully processed in this hypothetical system. Moreover, the existence of much larger convolutional networks running in realtime in budgeted mobile devices [258] also supports the idea that the approach is readily applicable in a number of possible scenarios.

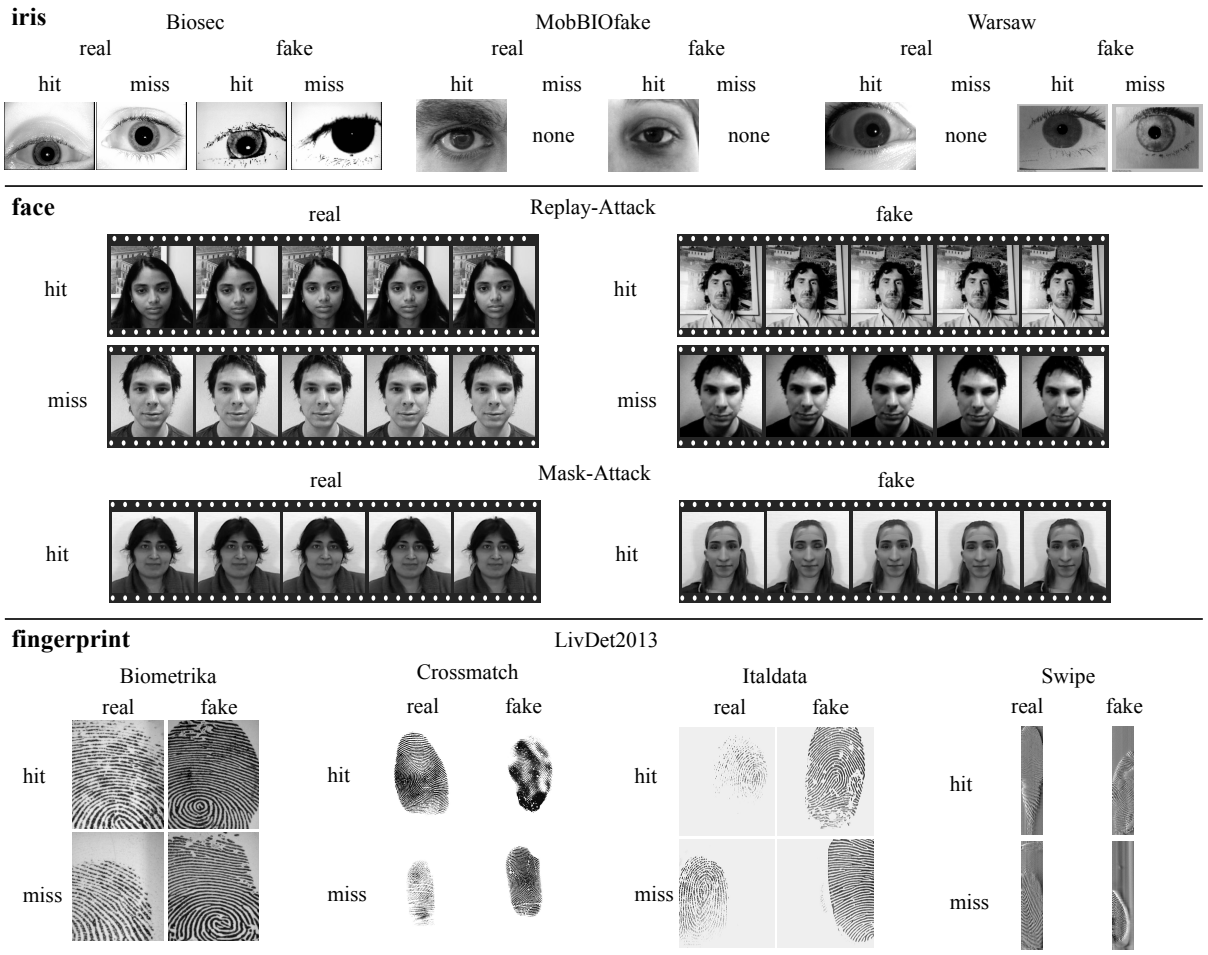


Figure 4.5: Examples of hit and missed testing samples lying closest to the real-fake decision boundary of each benchmark. A magnified visual inspection on these images may suggest some properties of the problem to which the learned representations are sensitive.

4.5.5 Visual Assessment

In Fig. 4.5, we show examples of hit and missed testing samples lying closest to the real-fake decision boundary of the best performing system in each benchmark. A magnified visual inspection on these images may give us some hint about properties of the problem to which the learned representations are sensitive.

While it is difficulty to infer anything concrete, it is interesting to see that the real missed sample in Biosec is quite bright, and that skin texture is almost absent in this case. Still, we may argue that a noticeable difference exists in Warsaw between the resolution used to print the images that led to the fake hit and the fake miss.

Regarding the face benchmarks, the only noticeable observation from Replay-Attack is that the same person is missed both when providing to the system a real and a fake biometric reading. This may indicate that some individuals are more likely to successfully attack a face recognition systems than others. In 3DMAD, it is easy to see the difference between the real and fake hits. Notice that there was no misses in this benchmark.

A similar visual inspection is much harder in the fingerprint benchmarks, even though

the learned deep representations could effectively characterize these problems. The only observation possible to be made here is related to the fake hit on CrossMatch, which is clearly abnormal. The images captured with the Swipe sensor are naturally narrow and distorted due to the process of acquisition, and this distortion prevents any such observation.

4.6 Conclusions and Future Work

In this work, we investigated two deep representation research approaches for detecting spoofing in different biometric modalities. On one hand, we approached the problem by learning representations directly from the data through architecture optimization with a final decision-making step atop the representations. On the other, we sought to learn filter weights for a given architecture using the well-known back-propagation algorithm. As the two approaches might seem naturally connected, we also examined their interplay when taken together. In addition, we incorporated our experience with architecture optimization as well as with training filter weight for a given architecture into a more interesting and adapted network, *spoofnet*.

Experiments showed that these approaches achieved outstanding classification results for all problems and modalities outperforming the state-of-the-art results in eight out of nine benchmarks. Interestingly, the only case for which our approaches did not achieve SOTA results is for the Biosec benchmark. However, in this case, it is possible to achieve a 98.93% against 100.0% accuracy of the literature. These results support our hypothesis that the conception of data-driven systems using deep representations able to extract semantic and vision meaningful features directly from the data is a promising venue. Another indication of this comes from the initial study we did for understanding the type of filters generated by the learning process. Considering the fingerprint case, learning directly from data, it was possible to come up with discriminative filters that explore the blurring artifacts due to recapture. This is particularly interesting as it is in line with previous studies using custom-tailored solutions [74].

It is important to emphasise the interplay between the architecture and filter optimization approaches for the spoofing problem. It is well-known in the deep learning literature that when thousands of samples are available for learning, the filter learning approach is a promising path. Indeed, we could corroborate this through fingerprint benchmarks that considers a few thousand samples for training. However, it was not the case for faces and two iris benchmarks which suffer from the small sample size problem (SSS) and subject variability hindering the filter learning process. In these cases, the architecture optimization approach was able to learn representative and discriminative features providing comparable spoofing effectiveness to the SOTA results in almost all benchmarks, and specially outperforming them in three out of four SOTA results when the filter learning approach failed. It is worth mentioning that sometimes it is still possible to learn meaningful features from the data even with a small sample size for training. We believe this happens in more well-posed datasets with less variability between training/testing data as it is the case of MobioBIOfake benchmark in which the AO approach achieved

99.38% just 0.37% behind the SOTA result.

As the data tell it all, the decision to which path to follow can also come from the data. Using the evaluation/validation set during training, the researcher/developer can opt for optimizing architectures, learn filters or both. If training time is an issue and a solution must be presented overnight, it might be interesting to consider an already learned network that incorporates some additional knowledge in its design. In this sense, *spoofnet* could be a good choice. In all cases, if the developer can incorporate more training examples, the approaches might benefit from such augmented training data.

The proposed approaches can also be adapted to other biometric modalities not directly dealt with herein. The most important difference would be in the input type of data since all discussed solutions directly learn their representations from the data.

For the case of iris spoofing detection, here we dealt only with iris spoofing printed attacks and some experimental datasets using cosmetic contact lenses have recently become available allowing researchers to study this specific type of spoofing [32, 264]. For future work, we intend to evaluate such datasets using the proposed approaches here and also consider other biometric modalities such as palm, vein, and gait.

Finally, it is important to take all the results discussed herein with a grain of salt. We are not presenting the final word in spoofing detection. In fact, there are important additional research that could finally take this research another step forward. We envision the application of deep learning representations on top of pre-processed image feature maps (e.g., LBP-like feature maps, acquisition-based maps exploring noise signatures, visual rhythm representations, etc.). With an n-layer feature representation, we might be able to explore features otherwise not possible using the raw data. In addition, exploring temporal coherence and fusion would be also important for video-based attacks.

Acknowledgment

We thank UFOP, Brazilian National Research Counsil – CNPq (Grants #303673/2010-9, #304352/2012-8, #307113/2012-4, #477662/2013-7, #487529/2013-8, #479070/2013-0, and #477457/2013-4), the CAPES DeepEyes project, São Paulo Research Foundation – FAPESP, (Grants #2010/05647-4, #2011/22749-8, #2013/04172-0, and #2013/11359-0), and Minas Gerais Research Foundation – FAPEMIG (Grant APQ-01806-13). D. Menotti thanks FAPESP for a grant to acquiring two NVIDIA GeForce GTX Titan Black with 6GB each. We also thank NVIDIA for donating five GPUs used in the experiments, a Tesla K40 with 12GB to A. X. Falcão, two GeForce GTX 680 with 2GB each to G. Chiachia, and two GeForce GTX Titan Black with 6GB each to D. Menotti.

Chapter 5

Counteracting Presentation Attacks in Face, Fingerprint, and Iris Recognition

"A friend is one that knows you as you are, understands where you have been, accepts what you have become, and still, gently allows you to grow."
—William Shakespeare, *(1564–1616)*

"A sweet friendship refreshes the soul."
—Bible, *Proverbs 27:9*

Abstract

This chapter explores data-driven approaches to presentation attack detection for three biometric modalities: face, iris and fingerprint. The primary aim of this chapter is to show how pre-trained deep neural networks can be used to build classifiers that can distinguish between authentic images of faces, irises and fingerprints and their static imitations. The most important, publicly available benchmarks representing various attack types were used in a unified presentation attack detection framework in both same-dataset and cross-dataset experiments. The pre-trained VGG neural networks, being the core of this solution, tuned independently for each modality and each dataset present almost perfect accuracy for all three biometric techniques. In turn, low classification accuracies achieved in cross-dataset evaluations show that models based on deep neural networks are sensitive not only to features specific to biometric imitations, but also to dataset-specific properties of samples. Thus, such models can provide a rapid solution in scenarios in which properties of imitations can be predicted but appropriate feature engineering is difficult. However, these models will perform worse if the properties of imitations being detected are unknown. This chapter includes also a current literature review summarizing up-to-date data-driven solutions to face, iris and finger liveness detection.

5.1 Introduction

BIOMETRIC authentication is a technology designed to recognize humans automatically based on their behavior, physical and chemical traits. Recently, this technology emerged as an important mechanism for access control in many modern applications, in which the traditional methods including the ones based on knowledge (*e.g.*, keywords) or based on tokens (*e.g.*, smart cards) might be ineffective since they are easily shared, lost, stolen or manipulated [109]. Biometric technologies are increasingly used as the main authenticating factor for access control and also jointly with traditional authentication mechanisms, as a "step-up authentication" factor in two- or three-factor authentication systems.

In this context, face, iris and fingerprint are the most commonly-used biometric traits. In fact, the choice of the trait to be used takes into account some issues such as universality, easiness to measure the biometric characteristics, performance, or difficulty to circumvent the system [109]. However, a common disadvantage of these traits is that an impostor might produce a synthetic replica that can be presented to the biometric sensor to circumvent the authentication process. In the literature, the mechanisms to protect the biometric system against this type of attack are referred to as *spoofing detection*, *liveness detection* or *presentation attack detection* (PAD). Hereinafter, we will use the most generic term, presentation attack detection (PAD), which was initially proposed by SC37 experts in ISO/IEC 30107 – Presentation Attack Detection – Framework (Part 1), Data Formats (Part 2), and Testing and Reporting (Part 3).

The idea of spoofing biometric recognition is surprisingly older than biometrics itself. A careful reader of the Old Testament can find an impersonation attempt described in the Book of Genesis, based on presentation of a goat's fur put on Jacob's hand to imitate properties of Esau's skin, so that Jacob would be blessed by Isaac. A fictitious example that is surprisingly realistic is the description of how to copy someone's fingerprint using a wax mold and gelatin presented by Austin Freeman in his crime novel "The Red Thumb Mark". The novel appeared in 1907, and the technique described is still used almost 100 years later to spoof fingerprint sensors. Note that this description appeared only four years after fingerprints were adopted by Scotland Yard, and long before the first fingerprint sensor appeared on the market.

Recent scientific studies and open challenges such as LivDet (`www.livdet.org`) suggest that presentation attacks are still an open problem in biometrics. Phan and Boulkenafet [27,187] suggest that face recognition systems are vulnerable to presentation attacks with an equal error rate (related to distinguishing presentation attacks from genuine samples) reaching as high as 9%. Fingerprint-based recognition systems still face the same problem, with an average classification error rate achieving 2.9% [166]. Iris-based authentication, considered by many to be one of the most reliable biometrics, awaits efficient PAD methodology. Recent proposals in this area still report an average classification error rate around 1% [199].

Besides the laboratory testing of the biometric system's vulnerability to attack, a few real cases also confirm the problem. In the small city of Ferraz de Vasconcelos, in the outskirts of São Paulo, Brazil, a physician of the service of mobile health care and urgency

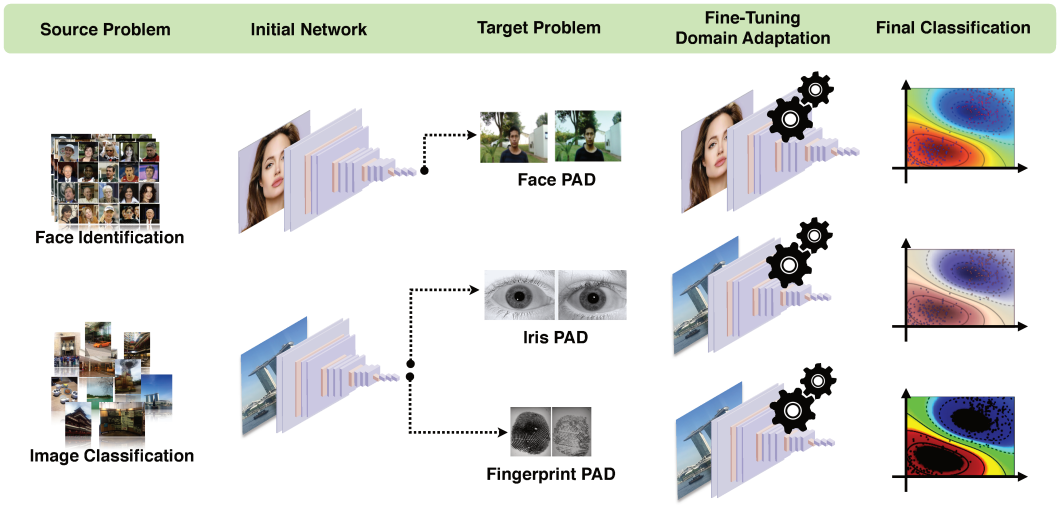


Figure 5.1: General pipeline exploited in this work. Initial network architectures, originally proposed for other problems, are independently fine-tuned with appropriate PAD examples from different datasets leading to discriminative features. Ultimately, classifiers are trained to separate between authentic images of faces, irises and fingerprints from their presentation attack versions.

was caught red-handed by the police in a scam that used silicone fingers to bypass an authentication system and confirm the presence several colleagues at work [147]. A similar case has been investigated by the Brazilian Federal Police in 2014, when workers at the Paranaguá Harbor in the Brazilian southern state of Paraná, were suspected of using silicone fingers to circumvent a time attendance biometric system [35]. In Germany, the biometric hacking team in the Chaos Computer Club managed to hack Apple's iPhone Touch ID [10] a few days after its launch, demonstrating that a biometric system without an adequate protection is unsuitable as a reliable access control method. Other cases of spoofing surveillance systems with 3-D masks to change their apparent age or race can also be found in [210, 211].

Considering the three aforementioned modalities, when we look at the literature and analyze the algorithms to prevent presentation attacks, we observe that the most promising in terms of errors and minimum effort of implementation or cost often share an interesting feature: they belong to a group of algorithms referred to as data-driven characterization algorithms. According to Pinto *et al.* [189], methods based on data-driven characterization exploit only the data that comes from a standard biometric sensor looking for evidence of artifacts in the already acquired biometric sample. Such approaches are preferable in practice because they are easily integrable with the existing recognition systems, as there is no extra requirement in terms of hardware nor is there the need of human interaction to detect attempted attacks.

Although the existing methods following this idea have led to good detection rates, we note that some aspects still need to be taken into account when evaluating a PAD approach, *e.g.*, different types of attack, variety of devices to perform attempted attacks, and attacks directed to different sensors. Another aspect that is normally overlooked is

that most detection methods are custom-tailored to specific types of presentation attacks, in what we refer to as hand-crafting of the features. With the emergence of deep learning methods and their success in tasks such as image classification, voice recognition and language translation, in this chapter, we set forth the objective of exploiting deep learning solutions for detecting presentation attacks, using data-driven solutions. In these cases, the biometric designer is responsible for choosing an appropriate architecture for PAD and training solely from the existing data available. We believe that this type of solution is the next step when designing robust presentation attack detectors and also that they can, if carefully designed, better deal with the challenging cross-dataset scenario. The cross-dataset scenario arises when the system is trained with a dataset from one sensor or one scenario, and then later tested on data from a different sensor or scenario. Figure 5.1 depicts the general pipeline we exploit in this chapter. We start with pre-trained deep neural networks and tune them independently for each modality (face, iris and fingerprint) with different datasets before building the final classifiers to distinguish between authentic images of faces, irises and fingerprints from their static counterparts.

We organize the rest of this Chapter as follows. Section 5.2 discusses state-of-the-art methods for PAD considering the three modalities considered in this chapter (face, iris and fingerprint). Section 5.3 details the data-driven PAD solution that we advocate as very promising for this problem, while Section 5.5 shows the experiments and validations for different biometric spoofing datasets. We close the chapter with some final considerations in Section 5.6.

5.2 Related Work

In this section, we review some of the most important presentation-attack detection methods published in the literature for iris, face and fingerprint.

5.2.1 Face Presentation Attack Detection

The existing face anti-spoofing techniques can be categorized into four groups [214]: user behavior modeling [176, 263] (*e.g.*, eye blinking, small face movements), methods that require additional hardware [66] (*e.g.*, infrared cameras and depth sensors), methods based on user cooperation (*e.g.*, challenge questions) and, finally, data-driven characterization approaches, which is the focus of our work herein.

We start this section reviewing frequency-based approaches, which are methods that rely on analyzing artifacts that are better visible in the frequency domain. Early studies followed this idea [141], and nowadays we have several works that support the effectiveness of this approach in detecting face spoofing. In [141], Li *et al.* proposed a face spoofing detection that emerged from the observation that the faces in photographs are smaller than the real ones and that the expressions and poses of the faces in photographs are invariant. Based on these observations, the authors devised a threshold-based decision method for detecting photo-based attempted attacks based on the energy rate of the high frequency components in the 2-D Fourier spectrum. The major limitation of the technique proposed by Li *et al.* is that the high frequency components are affected by

illumination, which makes this frequency band too noisy [141,245]. To reduce that effect, Tan *et al.* [245] exploited the difference of image variability in the high-middle band. This is done using Difference of Gaussian (DoG) bandpass filtering, which keeps as much detail as possible without introducing noisy or aliasing artifacts.

In [193], Pinto *et al.* introduced an idea seeking to overcome the illumination effect when working in the frequency domain. In that work, the authors proposed a face anti-spoofing method for detecting video-based attempted attacks based on Fourier analysis of the noise signature extracted from videos, instead of using the image pixel values directly. Basically, after isolating the noise signal present in the video frames, the authors transformed that information to the Fourier domain and used the visual rhythm technique to capture the most important frequency components to detect an attempted attack, taking advantage of the spectral and temporal information. In a more recent work [190], the same authors expanded upon this technique taking advantage of the spectral, temporal and spatial information from the noise signature, using the concept of visual codebooks. According to the authors, the new method enabled them to detect different types of attacks such as print- and mask-based attempted attacks as well.

Lee *et al.* [139] proposed an anti-spoofing technique based on the cardiac pulse measurements using video imaging [197]. The authors extended upon previous work proposed by Poh *et al.* [197] by adding a threshold-based decision level based on the entropy measure. It was calculated from the power spectrum obtained from normalized RGB channels after eliminating the cross-channel noise, caused by the environment interference, using the Independent Component Analysis (ICA).

Another expressive branch of face anti-spoofing algorithms reported in the literature consists of texture-based approaches. In general, those algorithms exploit textural cues inserted in the fake biometric samples during its production and presentation to the biometric sensor under attack (*e.g.*, printing defects, aliasing and blurring effects). Tan *et al.* [245] proposed a texture-based approach to detect attacks with printed photographs based on the difference of the surface roughness of an attempted attack and a real face. The authors estimate the luminance and reflectance of the image under analysis and classify them using Sparse Low Rank Bilinear Logistic Regression methods. Their work was extended upon by Peixoto *et al.* [181], who incorporated measures for different illumination conditions.

Similar to Tan *et al.* [245], Kose *et al.* [128] evaluated a solution based on reflectance to detect attacks performed with printed masks. To decompose the images into components of illumination and reflectance, the Variational Retinex [5] algorithm was applied.

Määttä *et al.* [151,152] relied on micro textures for face spoofing detection, inspired by the characterization of printing artifacts and by differences in light reflection when comparing real samples and presentation attack samples. The authors proposed a fusion scheme based on the Local Binary Pattern (LBP) [168], Gabor wavelets [61], and Histogram of Oriented Gradients (HOG) [56]. Similarly, to find a holistic representation of the face able to reveal an attempted attack, Schwartz *et al.* [214] proposed a method that employs different attributes of the images (*e.g.*, color, texture and shape of the face).

Chingovska *et al.* [41] investigated the use of different variations of the LBP operator used in [151]. The histograms generated from these descriptors were classified using χ^2

histogram comparison, Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM).

Face spoofing attacks performed with static masks have also been considered in the literature. Erdogmus *et al.* [65] explored a database with six types of attacks using facial information of four subjects. To detect attempted attacks, the authors used two algorithms based on Gabor wavelet [138] with a Gabor-phase based similarity measure [90].

Pereira *et al.* [184] proposed a score-level fusion strategy for detecting various types of attacks. The authors trained classifiers using different databases and used the Q statistics to evaluate the dependency between classifiers. In a follow-up work, Pereira *et al.* [69] proposed an anti-spoofing solution based on the dynamic texture, which is a spatiotemporal version of the original LBP.

Garcia *et al.* [77] proposed an anti-spoofing method based on detection of the Moiré patterns, which appear due to the overlap of the digital grids. To find these patterns, the authors used a peak-detector algorithm based on maximum-correlation thresholding, in that strong peaks reveal an attempted attack. Similar to [77], Patel *et al.* [179] proposed a presentation attack detection technique also based on the Moiré pattern detection, which uses the the multi-scale version of the LBP descriptor (M-LBP).

Tronci *et al.* [253] exploited the motion information and clues that are extracted from the scene by combining two types of processes, referred to as static and video-based analysis. The static analysis consists of combining different visual features such as color, edge, and Gabor textures, whereas the video-based analysis combines simple motion-related measures such as eye blink, mouth movement, and facial expression change.

Anjos *et al.* [8] proposed a method for detecting photo-based attacks assuming a stationary facial recognition system. According to the authors, the intensity of the relative motion between the face region and the background can be used as a clue to distinguish valid access of attempted attacks, since the motion variations between face and background regions exhibit greater correlation in the case of attempted attacks.

Wen *et al.* [260] proposed a face spoof detection algorithm based on image distortion analysis (IDA), describing different features such as specular reflection, blurriness, chromatic moment, and color diversity. These features are concatenated in order to generate feature vectors, which are used to generate an ensemble classifier, each one specialized to detect a type of attempted attack.

Kim *et al.* [118] proposed a method based on the diffusion speed of a single image to detect attempted attacks. The authors define the local patterns of the diffusion speed, namely local speed patterns via Total Variation (TV) flow [218], which are used as feature vectors to train a linear classifier, using the SVM, to determine whether the given face is fake. In turn, Boulkenafet *et al.* [29] proposed an anti-spoofing technique using a color texture analysis. Basically, the authors perform a micro-texture analysis considering the color-texture information from the luminance and the chrominance channels by extracting feature descriptions from different color spaces.

Different from the previous methods, which focus on defining a presentation attack detection that does not leverage the identity information present in the gallery, Yang *et al.* [268] proposed a person-specific face anti-spoofing approach, in which a classifier was built for each person. According to the authors, this strategy minimizes the interferences

among subjects.

Virtually all previous methods exploit handcrafted features to analyze possible clues related to a presentation attack attempt. Whether these features are related to texture, color, gradients, noise or even reflection, blurriness, and chromatic moment, they always come down to the observation of specific artifacts present in the images and how they can be captured properly. In this regard, LBP stands out as the staple of face-based spoofing research thus far. Departing from this hand-crafted characterization modeling strategy, a recent trend in the literature has been devoted to designing and deploying solutions able to directly learn, from the existing available training data, the intrinsic discriminative features of the classes of interest, the so-called data-driven characterization techniques, probably motivated by the huge success these approaches have been showing in other vision-related problems [131,241]. Out of those, the ones based on deep learning solutions stand out right away as very promising for being highly adaptive to different situations.

Menotti *et al.* [162] aimed at hyperparameter optimization of network architectures [16, 196] (architecture optimization) and on learning filter weights via the well-known back-propagation algorithm [136] (filter optimization) to design a face spoofing detection approach. The first approach consists of learning suitable convolutional network architectures for each domain, whereas the second approach focuses on learning the weights of the network via back propagation.

Manjani *et al.* [154] proposed an anti-spoofing solution based on a deep dictionary learning technique originally proposed in [246] to detect attempted attacks performed using silicone masks. According to the authors, deep dictionary learning combines concepts of two most prominent paradigms for representation learning, deep learning and dictionary learning, which enabled the authors to achieve a good representation even using a small data for training.

5.2.2 Fingerprint Presentation Attack Detection

Fingerprint PAD methods can be categorized into two groups: hardware-based and software-based solutions [83]. Methods falling into the first group use information provided from additional sensors to gather artifacts that reveal a spoofing attack that is outside of the fingerprint image. Software-based techniques rely solely on the information acquired by the biometric sensor of the fingerprint authentication system.

Based on several quality measures (*e.g.*, ridge strength or directionality, ridge continuity), Galbally *et al.* [71, 72] proposed a set of features aiming at fingerprint presentation attack detection, which were used to feed a Linear Discriminant Analysis (LDA) classifier.

Gragnaniello *et al.* [86] proposed an anti-spoofing solution based on Weber Local Descriptor (WLD) operating jointly with other texture descriptors such as Local Phase Quantization (LPQ) and Local Binary Pattern Descriptor (LBP). The experimental results suggest that WLD and LPQ complement one another, and their joint usage can greatly improve their discriminating ability, even when compared individually or combined with LBP.

Inspired by previous works based on LBP descriptor, Jia *et al.* [112] proposed a spoofing detection scheme based on Multi-scale Block Local Ternary Patterns (MBLTP) [244].

According to the authors, the computation of the LTP descriptor is based on average values of block subregions rather than individual pixels, which makes it less sensitive to noise, since the computation is based on a 3-value code representation and on average values of block subregions, rather than on individual pixels.

Ghiani *et al.* [81] proposed the use of Binarized Statistical Image Features (BSIF), a textural binary descriptor whose design was inspired by the LBP and LPQ methods. Basically, the BSIF descriptor learns a filter set by using statistics of natural images [104], leading to descriptors better adapted to the problem. The same authors also explored the LPQ descriptor to find a feature space insensitive to blurring effects [82].

In [85], Gottschlich proposed another idea based on filter learning convolution comparison pattern. To detect a fingerprint spoofing, the authors compute the discrete cosine transform (DCT) from rotation invariant patches, and compute their binary patterns by comparing pairs of DCT coefficients. These patterns are gathered in a histogram, which was used to feed a linear SVM classifier.

Rattani *et al.* [206] introduced a scheme for automatic adaptation of a liveness detector to new spoofing materials in the operational phase. The aim of the proposed approach is to reduce the security risk posed by new spoof materials on an anti-spoofing system. The authors proposed a novel material detector specialized to detect new spoof materials, pointing out the need for retraining the system with the new material spotted.

Similar to [206], Rattani *et al.* [207] proposed an automatic adaptation anti-spoofing system composed of an open-set fingerprint spoofing detector and by a novel material detector, both based on Weibull-calibrated SVM (W-SVM) [223]. The novel material detector was built with a multi-class W-SVM, composed by an ensemble of pairs of 1-Class and binary SVMs, whereas the open set fingerprint spoofing detector was trained with features based on textural [82], physiological [155] and anatomical [243] attributes.

Gragnaniello *et al.* [88] proposed a fingerprint spoofing detection based on both spatial and frequency information, in order to extract local amplitude contrast, and local behavior of the image, which were synthesized by considering the phase of some selected transform coefficients generated by the short-time Fourier transform (STFT). This information generates a bi-dimensional contrast-phase histogram, which was used to train a linear SVM classifier.

Kumpituck et al. [134] exploited an anti-spoofing schema based on wavelet decomposition and LBP operator. In this work, the authors extract LBP histograms from several wavelet sub-band images, which were concatenated and used to feed an SVM classifier. The authors also evaluated a more conventional approach that consists of calculating the energy from wavelet sub-bands instead of the LBP histograms. Experimental results show that wavelet-LBP descriptor achieved a better discrimination than wavelet-energy and LBP descriptors used separately, besides achieving competitive results with the state-of-the-art methods.

Finally, also departing from the traditional modeling, which uses basically texture patterns to characterize fingerprint images, Nogueira *et al.* [166] proposed a fingerprint anti-spoofing technique based on the concept of pre-trained convolutional neural networks. Basically, the authors use well-known CNN architectures in the literature such as AlexNet [131] and VGG [233] as their starting point for learning the network weights for

fingerprint spoofing detection.

Marasco *et al.* [156] investigated two well-known CNN architectures, the GoogLeNet [239], CaffeNet [131], in order to analyze their robustness in detecting unseen spoof materials and fake samples from new sensors. As mentioned before, Menotti *et al.* [162] also proposed hyperparameter optimization of network architectures along with filter optimization techniques for detecting fingerprints presentation attacks.

5.2.3 Iris Presentation Attack Detection

Early work on iris spoofing detection dates back to the 1990's, when Daugman [59] discussed the feasibility of some attacks on iris recognition systems. In that work, he proposed to detect such attempts using the Fast Fourier Transform to verify the high frequency spectral magnitude.

According to Czajka [53], solutions for iris liveness detection can be categorized into four groups, as Cartesian product of two dimensions: type of measurement (passive or active) and type of model of the object under test (static / dynamic). Passive solutions mean that the object is not stimulated more than it is needed to acquire an iris image for recognition purpose. Hence, it typically means that no extra hardware is required to detect an attempted attack. Active solutions try to stimulate an eye and observe the response to that stimuli. It means that typically some extra hardware elements are required. In turn, the classification between static and dynamic objects means that the algorithm can detect an attempted attack using just one (static) image from the biometric sensor or needs to use a sequence of images to observe selected dynamic features. In this section, we review only passive and static methods, which is the focus of this chapter.

In [174], Pacut *et al.* introduced three iris liveness detection algorithms based on the analysis of the image frequency spectrum, controlled light reflection from the cornea and pupil dynamics. These approaches were evaluated with paper printouts produced with different printers and printout carriers, and shown to be able to spoof two commercial iris recognition systems. A small hole was made in the place of the pupil, and this trick was enough to deceive commercial iris recognition systems used in their study. The experimental results obtained on the evaluation set composed of 77 pairs of fake and live iris images showed that the controlled light reflections and pupil dynamics achieve zero for both False Acceptance Rate and False Rejection Rate. In turn, two commercial cameras were not able to detect 73.1% and 15.6% of iris paper printouts and matched them to biometric references of authentic eyes.

Galbally *et al.* [75] proposed an approach based on 22 image quality measures (*e.g.*, focus, occlusion, and pupil dilation). The authors use sequential floating feature selection [198] to single out the best features, which were used to feed a quadratic discriminant classifier. To validate the proposed approach, the authors used the BioSec [68,219] benchmark, which contains print-based iris spoofing attacks. Similarly, Sequeira *et al.* [226] also exploited image quality measures [75] and three different classification techniques, validating the work on BioSec [68, 219] and Clarkson [224] benchmarks and introducing the MobBIOfake benchmark comprising 800 iris images. Sequeira *et al.* [227] extended upon previous work using a feature selection step to obtain a better representation to detect

an attempted attack. The authors also applied iris segmentation [164] to obtain the iris contour and adapted the feature extraction processes to the resulting non-circular iris regions.

In [259], Wei *et al.* addressed the problem of iris liveness detection based on three texture measures: iris edge sharpness (ES), iris-texton feature for characterizing the visual primitives of iris texture (IT) and using selected features based on co-occurrence matrix (CM). In particular, they used fake iris wearing color and textured contact lenses. The experiments showed that the ES feature achieved comparable results to the state of the art methods at that time, and the IT and CM measures outperformed the state of the art algorithms.

Czajka [51] proposed a solution based on frequency analysis to detect printed irises. The author associated peaks found in the frequency spectrum to regular patterns observed for printed samples. This method, tuned to achieve close-to-zero false rejection rate (*i.e.,* not introducing additional false alarms to the entire system), was able to detect 95% of printed irises. This paper also introduced the Warsaw LivDet-Iris-2013 dataset containing 729 fake images and $1,274$ images of real eyes.

Texture analysis has also been explored for iris spoofing detection. In the MobILive [229] iris spoofing detection competition, the winning team relied upon three texture descriptors: LBP [169], LPQ [171] and Binary Gabor Pattern (BGP) [272]. Sun *et al.* [238] recently proposed a general framework for iris image classification based on a Hierarchical Visual Codebook (HVC). The HVC encodes the texture primitives of iris images and is based on two existing bag-of-words models. The method achieved a state-of-the-art performance for iris spoofing detection, among other tasks related to iris recognition.

Doyle *et al.* [64] proposed a solution based on modified Local Binary Patterns (mLBP) [170] descriptor. In this work, the authors show that although it is possible to obtain good classification results using texture information extracted by the mLBP descriptor, when lenses produced by different manufacturers are used, the performance of this method drops significantly. They report 83% and 96% of correct classification when measured on two separated datasets, and a significant drop in accuracy when the same method was trained on the one dataset and tested on the other dataset: 42% and 53%, respectively. This cross-dataset validation has been shown to be very challenging and seems to be recommended in several validation setups for presentation attack detection. Yadav *et al.* [264] extended upon the previous work by analyzing the effect of soft and textured contact lenses on iris recognition.

In [200], Raja *et al.* proposed an anti-spoofing method based on Eulerian Video Magnification (EVM) [24], which was applied to enhance the subtle phase information in the eye region. The authors proposed a decision rule based on cumulative phase information, which was applied by using a sliding window approach upon the phase component for detecting the rate of the change in the phase with respect to time.

Raghavendra *et al.* [199] proposed a novel spoofing detection scheme based on a multi-scale version of the Binarized Statistical Image Features (BSIF) and linear Support Vector Machine (SVM). Gupta *et al.* [91] proposed an anti-spoofing technique based on local descriptors such as LBP [168], HOG [56], and GIST [172], which provide a representation space by using attributes of the images such as color, texture, position, spatial frequency,

and size of objects present in the image. The authors used the feature vectors produced by the three descriptors to feed a nonlinear classifier and decide whether an image under analysis is fake.

Czajka [52] proposed an iris spoofing detection based on pupil dynamics. In that work, the author used the pupil dynamics model proposed by Kohn and Clynes [121] to describe its reaction after a positive light stimuli. To decide whether the eye is alive, the author used variants of the SVM to classify feature vectors that contain the pupil dynamic information of a target user. This work has been further extended to a mixture of negative and positive light stimuli [53] and presented close-to-perfect recognition of objects not reacting to light stimuli as expected for a living eye.

Finally, Lovish *et al.* [148] proposed a cosmetic contact lens detection method based on Local Phase Quantization and Binary Gabor Patterns, which combines the benefits of both LBP and Gabor filters [272]. The histograms produced for both descriptors were concatenated and used to build a classification model based on SVM algorithm.

Similarly to the approaches tackling the presentation attack problem in fingerprint and faces, handcrafted texture features seem to be the preferred choice in iris spoofing detection. Methods inspired by LBP, visual codebooks and quality metrics are the most popular methods so far. In this sense, the works of Menotti *et al.* [162] and Silva *et al.* [232], which exploit data-driven solutions for this problem, are sufficiently different from the previous methods and present very promising results.

5.2.4 Unified Frameworks to Presentation Attack Detection

Galbally *et al.* [74] proposed a general approach based on 25 image quality features to detect attempt attacks in face, iris and fingerprint biometric systems simultaneously. Evaluations performed upon popular benchmarks for three modalities show that this approach is highly competitive, considering the state-of-the-art methods dedicated for single modalities.

In [87], Gragnaniello *et al.* evaluated several local descriptors for face-, fingerprint- and iris-based biometrics in addition to the investigation of promising descriptors using the Bag-of-Visual-Word (BoVW) model [235], Scale-Invariant Feature Transform (SIFT) [149], DAISY [251], and the Shift-Invariant Descriptor (SID) [122].

Menotti *et al.* [162] (mentioned earlier in this section) showed that the combination of architecture optimization and filter optimization provides better comprehension of how these approaches interplay for face, iris and fingerprint PAD, and also outperforms the best known approaches for several benchmarks.

In this chapter, we decided to explore data-driven solutions for spoofing detection in different modalities based on deeper architectures than the one used in [162] and evaluate the effects of such decision. Our objective is to show the potential of this approach but also highlight its limitations, especially related to cross-dataset experiment.

5.3 Methodology

In this section, we present the convolutional neural network that we adopted to PAD for face, fingerprint and iris. Our objective is simply to show that this new trend in the literature is also relevant for the task of presentation attack detection and that research in this direction needs to be considered. At the same time, we also show that even when adopting a powerful image classification technique such as deep neural networks, we still cannot deal effectively with the very challenging cross-dataset problem. As a result, it is clear that the research community now needs to shift its attention to cross-dataset validation setups (or, more general, open-set classification) as they are closer to real-world operational conditions when deploying biometric systems.

5.3.1 Network Architecture

For this work, we adopted the VGG network architecture proposed by [234]. However, that network was first proposed for object recognition and not presentation attack detection. Therefore, for each problem of interest (PAD in face, iris and fingerprint), we adapt the network's architecture as well as fine-tune its weights to our two-class problem of interest. Training the network from scratch to our problem is also a possibility if enough training samples (normal and presentation attack samples) are available. However, as this is not often the case in this area, it is recommended to start the network weights with a related (source) problem and then adapt these weights with training examples of a target problem.

Figure 5.2 depicts the network architecture we adopted in this work. During training, the network's input consists of fixed-size 224×224 RGB images which go through a stack of convolutional layers comprising filters with a very small receptive field (3×3). In this network, the convolution stride is fixed to one pixel and the spatial zero-padding for convolutional operation is also of one pixel. There are five max-pooling layers in this network (carefully placed after some convolution layers). The max-poolings are performed over a 2×2 pixel window, with stride 2.

The stack of convolutional layers is followed by three fully-connected (FC) layers: the first two have 4,096 units each, while the the third layer performs the 2-way spoofing classification problem of our interest (originally this was an FC layer with 1,000 units for the ImageNet 1,000-way classification problem). The final layer is the soft-max layer translating the outputs of 2-unit layer into a posterior probabilities of class membership. Each unit in the hidden layers has a rectified linear (ReLU) activation function [132]. The depth of convolution layers or, in other words, their number of channels, starts with 64 and is iteratively doubled after each max-pooling layer to a maximum of 512.

5.3.2 Training and Testing

For training, we start with the network trained to a source problem whenever it is possible. To detect presentation attacks with faces, we initialize the network with the weights learned for face recognition [178]. However, the closest problem we had for iris and fingerprints was general image classification. Therefore, presentation attack detection for iris and fingerprints is performed by the network initialized with the weights pre-computed

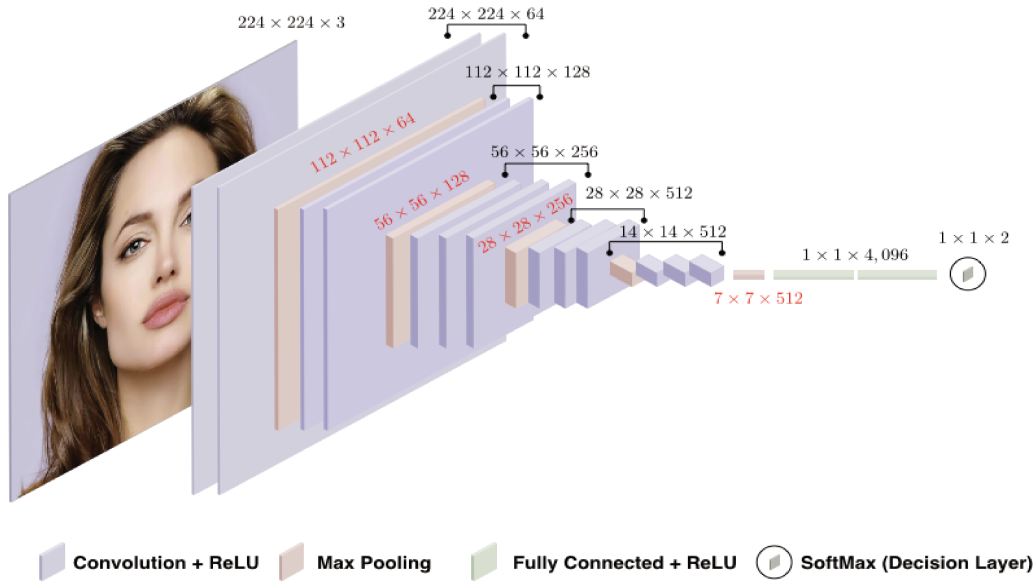


Figure 5.2: Adopted network architecture, originally proposed for object recognition by the Visual Geometry Group and thus referred to as VGG network.

for the ImageNet classification problem. The first convolutional layers act mostly as a generic feature detectors (such as edges) and are suitable for different computer vision tasks. However, each next convolutional layer is more context-focused and extracts features that are task-related. Hence, using last layers trained for general object recognition in visual spoofing detection is not optimal, and a large improvement may be achieved by specializing the network. Certainly, a preferable solution is to initialize the weights with those used in networks solving iris- and fingerprint-related tasks, as the network would have been specialized to this type of imagery. However, since training of such networks from the scratch requires a lot data and effort, it is still a good move to initialize own network with image-related weights than just purely at random and tune its weights if there is not enough available training data.

Once a source set of weights to initialize the network is chosen, the fine-tuning follows a standard procedure: selects the training set of the target domain and uses it to perform forward passes and back-propagation in the network. The test for an input image is straightforward. Just resize it to the network's input size and feed it to the network. As the network has been fully adapted to the target problem of interest, it will already produce a two-class output.

More specifically, for the cases of fingerprints, the input images in a dataset are center-cropped and resized to 224 × 224 pixels, which is the standard input size of the VGG network. The centering happens through calculating the average of black pixels in the binary fingerprint image and keeping all the rows/columns with a density of black pixels greater than the global image average plus or minus 1.8 standard deviations of each respective row/column. This is used to eliminate the borders without any useful information. For optimizing the network in the fingerprint case, we use the standard SGD solver implemented in Caffe with the following hyperparameters: base learning rate of 0.0001,

step lr policy, step size of 2,000, momentum of 0.9, weight decay of 0.0002, gamma of 0.5 and maximum of 2,001 iterations.

In the case of faces, we center-cropped the images based on the eye coordinates calculated with the aid of Face++[1]. Upon center-cropping, the image is resized to 224×224 pixels. For optimizing the network in the face case, we use the standard SGD solver implemented in Caffe with the following hyper parameters: base learning rate of 0.001, step lr policy, step size of 1,000, momentum of 0.9, weight decay of 0.0005, gamma of 0.001 and maximum number of iterations of 4,000.

For irises, we resize the images to the network's standard input size of 224×224 pixels and employ the same parameters as for the face optimization problem.

5.3.3 Memory Footprint

The chosen network has an average size of 140 MB. Most of its parameters (and memory) are in the convolution and fully-connected layers. The first FC layer contains 100M weights, out of a total of 134M for the entire adapted network.

5.4 Metrics and Datasets

In this section, we describe the benchmarks (datasets) and selected accuracy estimators considered in this work. All datasets used in this chapter were freely available to us and we believe that it is the case for other researchers upon request sent directly to their creators. Datasets composing our testing environment are the most commonly used benchmarks to evaluate presentation attack detection for face, iris and fingerprints. Since all the benchmarks have been already divided by their creators into training and testing subsets, we decided to follow these divisions. Each training subset was divided by us into two disjoint subsets multiple times to perform cross-validation-based training to increase generalization capabilities of the winning model and to minimize an overfitting. The results reported further in this chapter are those obtained on testing sets. The next subsections characterize briefly all datasets and Table 5.1 shows their major features, in particular the number of samples in each benchmark and their assignment to training and testing subsets.

5.4.1 Video-based Face Spoofing Benchmarks

In this chapter, we use two benchmarks used to evaluate the performance of PAD algorithms for face modality, Replay-Attack [41] and CASIA Face Anti-Spoofing [274] datasets. These datasets contain five types of attempted attacks performed with fake samples presenting different qualities.

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

Replay-Attack

This benchmark contains short video recordings of both valid accesses and video-based attacks of 50 different subjects. To generate valid access videos, each person was recorded in two sessions in a controlled and in an adverse environment with a regular webcam. Then, spoofing attempts were generated using three techniques:

- *print attack*: hard copies of high-resolution digital photographs were presented to the acquisition sensor; these samples were printed with a Triumph-Adler DCC 2520 color laser printer;
- *mobile attack*: videos displayed on an iPhone screen were presented to the acquisition sensor; these videos were taken also with the iPhone;
- *high-definition attack*: high resolution photos and videos taken with an iPad were presented to the acquisition sensor using the iPad screen.

CASIA

This benchmark was based on samples acquired from 50 subjects. Genuine images were acquired by three different sensors presenting different acquisition quality (from low to high): "long-time-used USB camera", "newly bought USB camera", and Sony NEX-5. Pixel resolution of images was either 640×480 (both webcams) or 1920×1080 (Sony sensor). Sony images were cropped to 1280×720 by the authors. During the acquisition, subjects were asked to blink. Three kinds of presentation attacks were carried out:

- *warped photo attack*: high quality photos were printed on a copper paper and videos were recorded by Sony sensor; the printed images were intentionally warped to imitate face micro-movements,
- *cut photo attack*: eyes were cut from the paper printouts and an attacker hidden behind an artifact imitated the blinking behavior when acquiring the video by the Sony sensor,
- *video attack*: high quality genuine videos were displayed on an iPad screen of 1280×720 pixel resolution.

The data originating from 20 subjects was selected for a training set, while remaining samples (acquired for 30 subjects) formed the testing set.

5.4.2 Fingerprint Spoofing Benchmarks

Two datasets used in Liveness Detection Competitions (LivDet, www.livdet.org) were employed in this chapter. LivDet is a series of international competitions that compare presentation attack methodologies for fingerprint and iris using a standardized testing protocol and large quantities of spoof and live samples. All the competitions are open to all academic and industrial institutions which have software-based or system-based

biometric liveness detection solutions. For fingerprints, we use datasets released in 2009 and 2013.

The LivDet2009 benchmark [159] consists of three subsets of samples acquired by Biometrics FX2000, CrossMatch Verifier 300 LC and Identix DFR2100. Both the spatial scanning resolution and pixel resolution vary across subsets, from 500 DPI to 686 DPI, and from 312×372 to 720×720 pixels, respectively. Three different materials were used to prepare spoofs: Play-Doh, gelatin and silicone.

The LivDet2013 benchmark [83] contains four subsets of real and fake fingerprint samples acquired by four sensors: Biometrika FX2000, Italdata ET10, Crossmatch L Scan Guardian, and Swipe. Inclusion of samples from the Swipe sensor is especially interesting, since it requires – as the name suggests – swiping a finger over the small sensor. This makes the quality of spoofs relatively different when compared to the regular, flat sensors requiring only touching the sensor by the finger. For a more realistic scenario, fake samples acquired by Biometrika and Italdata were generated without user cooperation, while fake samples acquired by Crossmatch and Swipe were generated with user cooperation. Several materials for creating the artificial fingerprints were used, including gelatin, silicone, latex, among others. The spatial scanning resolution varies from a small 96 DPI (the Swipe sensor) to 569 (the Biometrika sensor). The pixel resolution is also heterogeneous: from relatively non-standard 208×1500 to pretty large 800×750. This makes the cross-subset evaluation quite challenging.

5.4.3 Iris Spoofing Benchmarks

To evaluate our proposed method in detecting iris presentation attack, we used two benchmarks: AVTS [219] and a new dataset Warsaw LivDet2015, which is an extension of Warsaw LivDet2013 [51]. These datasets contain attempted attacks performed with printed iris images, which were produced using different printers and paper types.

AVTS

This benchmark was based on live samples collected for 50 volunteers under the European project BioSec (Biometrics and Security). To create spoofing attempts, the authors tested two printers (HP Deskjet 970cxi and HP LaserJet 4200L), various paper types (*e.g.*, cardboard as well as white, recycle, photo, high resolution and butter papers), and a number of pre-processing operations. The combination that gave the highest probability of image acquisition by the LG IrisAccess EOU3000 sensor used in the study was selected for a final dataset collection. The authors printed their samples with the inkjet printer (HP Deskjet 970cxi) on a high resolution paper and applied an Open-TopHat pre-processing to each image prior printing. The pixel resolution of each image was 640×480, which is recommended by ISO/IEC as a standard resolution for iris recognition samples.

Warsaw LivDet2015

This dataset is an extension of the LivDet-Iris 2013 Warsaw Subset [51] and was used in 2015 edition of LivDet-Iris competition (`www.livdet.org`). It gathers 2854 images of

Table 5.1: Main features of the benchmarks considered herein.

Modality	Benchmark	Color	Dimension	# Training			# Testing		
			cols × *rows*	Live	Fake	Total	Live	Fake	Total
Face	Replay-Attack	Yes	320 × 240	600	3000	3600	4000	800	4800
	CASIA	Yes	1280 × 720	120	120	240	180	180	360
Iris	Warsaw LivDet2015	No	640 × 480	852	815	1667	2002	3890	5892
	AVTS	No	640 × 480	200	200	400	600	600	1200
Fingerprint	LivDet2009: CrossMatch	No	640 × 480	500	500	1000	1500	1500	3000
	LivDet2009: Identix	No	720 × 720	375	375	750	1125	1125	2250
	LivDet2009: Biometrika	No	312 × 372	500	500	1000	1500	1500	3000
	LivDet2013: Biometrika	No	312 × 372	1000	1000	2000	1000	1000	2000
	LivDet2013: CrossMatch	No	800 × 750	1250	1000	2250	1250	1000	2250
	LivDet2013: Italdata	No	640 × 480	1000	1000	2000	1000	1000	2000
	LivDet2013: Swipe	No	208 × 1500	1250	1000	2250	1250	1000	2250

authentic eyes and 4705 images of the paper printouts prepared for almost 400 distinct eyes. The photographed paper printouts were used to successfully forge an example commercial iris recognition system (*i.e.*, samples used in real and successful presentation attacks). Two printers were used to generate spoofs: HP LaserJet 1320 and Lexmark C534DN. Both real and fake images were captured by an IrisGuard AD100 biometric device with liveness detection functionality intentionally switched off. To get a free copy of this dataset follow the instructions given at Warsaw's lab webpage http://zbum.ia.pw.edu.pl/EN/node/46.

5.4.4 Error Metrics

In this chapter we use the error metrics that are specific to presentation attack detection, and partially considered by ISO/IEC in their PAD-related standards [105].

Attack Presentation Classification Error Rate (APCER): proportion of *attack presentations* incorrectly classified as *bona fide (genuine) presentations* at the PAD subsystem in a specific scenario. This error metric is analogous to false match rate (FMR) in biometric matching, that is related to false match of samples belonging to two different subjects. As FMR, the APCER is a function of a decision threshold τ.

Bona Fide Presentation Classification Error Rate (BPCER): proportion of *bona fide (genuine) presentations* incorrectly classified as *presentation attacks* at the PAD subsystem in a specific scenario. This error metric is analogous to false non-match rate (FNMR) in biometric matching, that is related to false non-match of samples belonging to the same subject. Again, the BPCER is a function of a decision threshold τ.

Half Total Error Rate (HTER): combination of APCER and BPCER in a single error rate with a decision threshold as an argument:

$$\mathrm{HTER}(\tau) = \frac{\mathrm{APCER}(\tau) + \mathrm{BPCER}(\tau)}{2} \tag{5.1}$$

5.5 Results

In this section, we present and discuss the experimental results of the proposed method. Sections 5.5.1, 5.5.2 and5.5.3 show the performance results and the experimental protocols employed to validate the performance of the proposed methodology.

5.5.1 Face

In this section, we present the results of our proposed PAD for face modality. The experiments are conducted considering the original protocol of the datasets used in this chapter (cf., Section 5.4), as well cross-dataset protocol, hereafter referred to as same-dataset and cross-dataset protocols, respectively. In general, a prime requirement of most machine learning algorithms is that both training and testing sets are independent and identically distributed. But, unfortunately, it does not always happen in practice – subsets can be identically distributed (*e.g.*, captured using the same sensor and in the same environment conditions), but totally dependent due to adding of bias in the data (*e.g.*, some dirt in the biometric sensor used to capture both subsets, identities present in two subsets, artifacts added during the attack simulations, etc.). In addition, the effects of the closed-world assumption [223] may mislead us to believe that a given approach is perfect when in fact its performance can be disastrous when deployed in practice for unknown presentation attacks. In this context, both same-dataset and cross-dataset are key experimental protocols in determining more accurate detection rates of an anti-spoofing system when operating in less controlled scenarios with different kinds of attacks and sensors.

Same-dataset results. Table 5.2 shows the results for Replay-Attack and CASIA datasets, considering that training and testing is performed on the same dataset. The VGG network was able to detect all kinds of attempted attacks present in the Replay-Attack dataset, and also to detect two methods of attempted attacks (hand-based and fixed-support attacks), which were confirmed by the perfect classification result (HTER of 0.0%). Considering the CASIA dataset, the proposed method obtained an HTER of 6.67%. The performance achieved by the proposed method on this dataset can be explained by the high degree of variability present in the CASIA dataset (*e.g.*, different kinds of attack and resolution) that makes this dataset more challenging. In both datasets, we use the k-fold cross-validation technique ($k = 10$) to built a classification model using the training set, and also the development set whether it is available. Figures 5.3 and 5.4 present empirical distributions of the difference between two CNN output nodes and the corresponding ROC curves.

Cross-dataset results. Table 5.3, Fig. 5.5 and Fig. 5.6 show the results obtained in cross-dataset evaluation protocol. We can clearly see a dramatic drop in the performance when we train and test on different datasets. Several sources of variability between the datasets may contribute to this result. The first one is that the datasets contain different kinds of attack. The Replay-Attack dataset contains three kinds of attacks (high

Table 5.2: Performance results obtained in the **same-dataset** evaluations of the **face PAD**. Pointers to plots presenting Receiver Operating Characteristics (ROC) and empirical Probability Distribution Functions (ePDF) are added in the last column.

	APCER (%)	BPCER (%)	HTER (%)	ROC and ePDF
Replay-Attack	0.00	0.00	0.00	Fig. 5.3
CASIA	0.00	13.33	6.67	Fig. 5.4

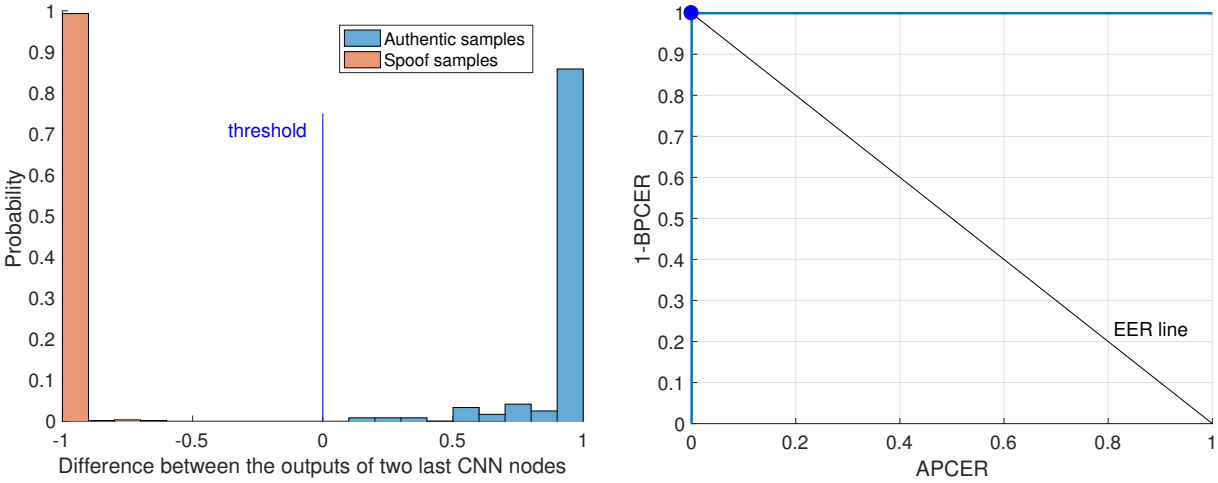


Figure 5.3: **Left:** Empirical probability distributions (ePDF) of the difference between two CNN output nodes (after softmax) obtained separately for authentic and spoof **face** samples. **Right:** ROC curve. Variant: training on **Replay-Attack**, testing on **Replay-Attack**. The threshold shown in blue color on the left plot and the blue dot on the ROC plot correspond to the approach when the predicted label is determined by the node with the larger output.

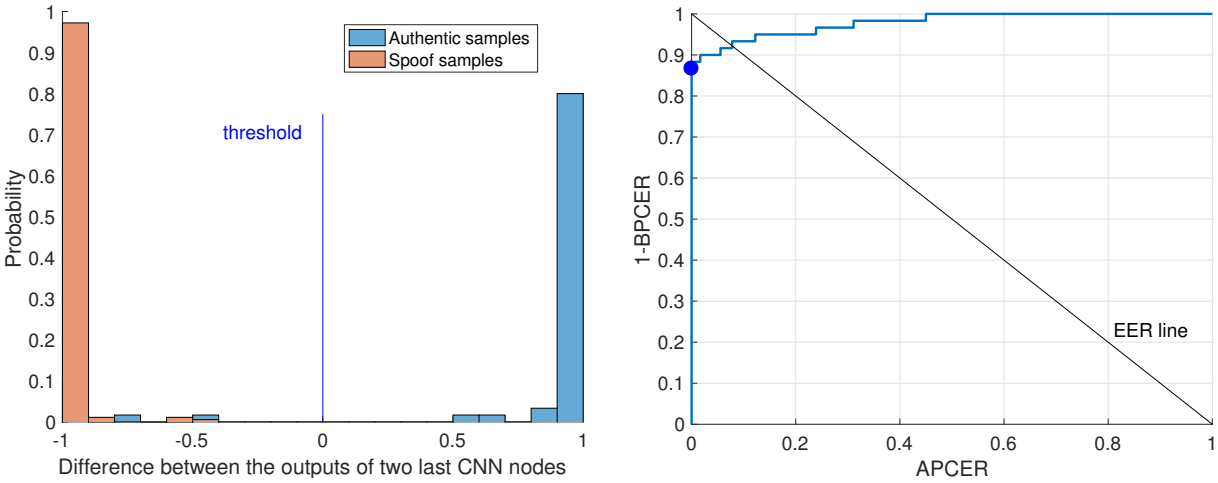


Figure 5.4: Same as Fig. 5.3 except the variant: training on **CASIA**, testing on **CASIA**.

definition-based, mobile-based and video-based attacks) while the CASIA dataset includes additional two kinds of attack (warp-based and cut-based photo attacks). Another source is the fact that data comes from different sensors, which potentially produce samples with different resolutions, color distributions, backgrounds, etc. The VGG architecture

finds very specific features and even when it is tuned to the specific problem, it does not generalize well to be agnostic to specific properties of data acquisition process.

Table 5.3: Performance results obtained with the **cross-dataset** evaluations of **face PAD** and using the overall testing set of each dataset.

Training	Test	APCER (%)	BPCER (%)	HTER (%)	ROC and ePDF
Replay-Attack	CASIA	42.67	51.67	47.16	Fig. 5.5
CASIA	Replay-Attack	89.44	10.0	49.72	Fig. 5.6

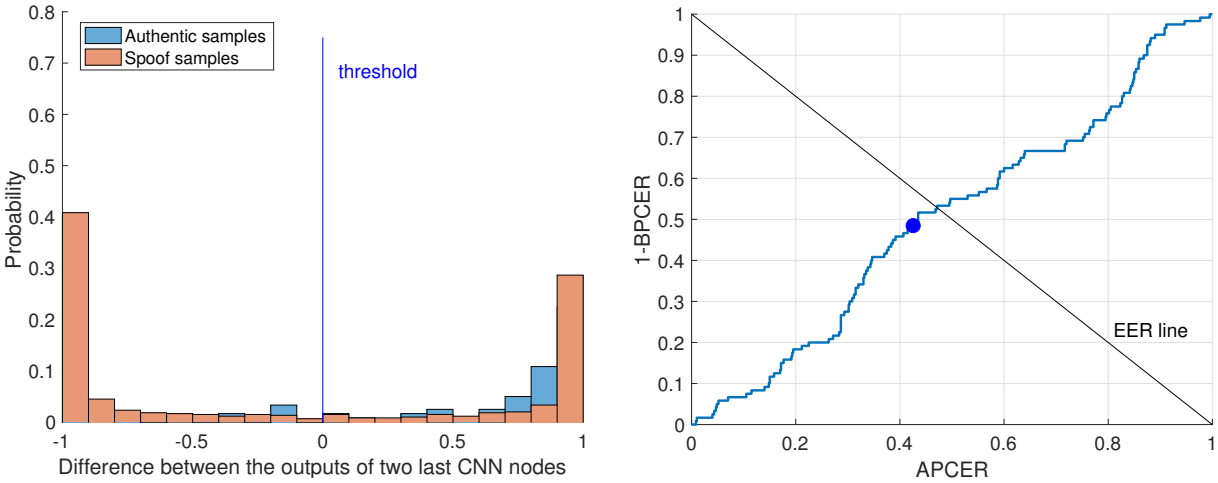


Figure 5.5: Same as Fig. 5.3 except the variant: training on **Replay-Attack**, testing on **CASIA** (cross-dataset testing).

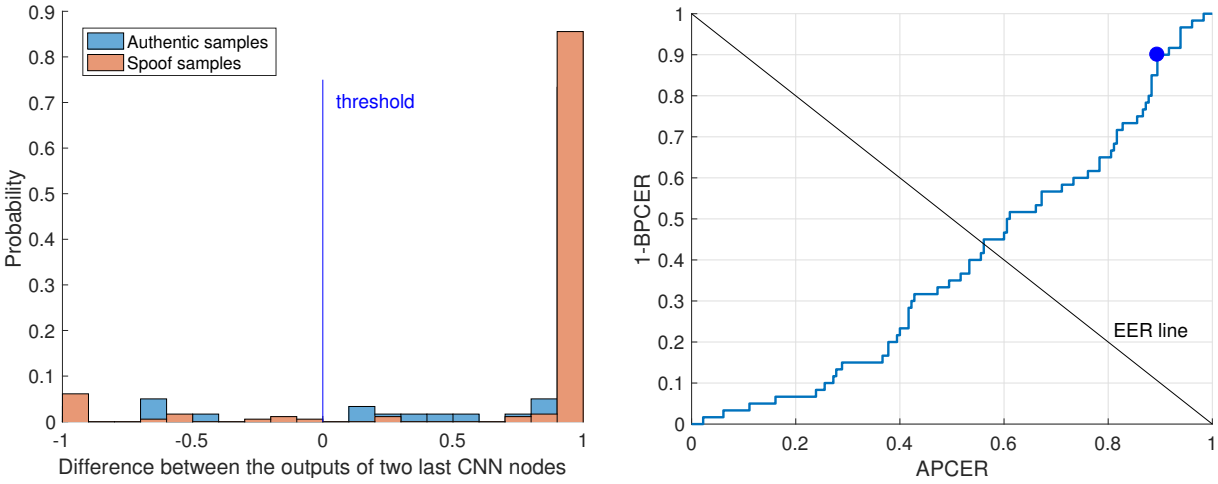


Figure 5.6: Same as Fig. 5.3 except the variant: training on **CASIA**, testing on **Replay-Attack**.

5.5.2 Fingerprints

This section presents how our VGG-based approaches perform in detection of fingerprint attack presentation. As for experiments with face benchmarks, we used the training

subsets (as defined by dataset creators) to make a cross-validation-based training, and separate testing subsets in final performance evaluation. Fingerprint benchmarks are composed of subsets gathering mixed attacks (for instance glue, silicone or gelatin artifacts) and acquired by different sensors (cf. Table 5.1).

Same-sensor results. In this scenario, samples acquired by different sensors are not mixed together. That is, if the classifier is trained with samples acquired by sensor X, only sensor X samples are used in both the validation and final testing. As in previous experiments, 10 statistically independent estimation-validation pairs of non-overlapping subsets were created, and the solution presenting the lowest HTER over ten validations was selected for testing. Table 5.4 as well as Figures 5.7 and 5.8 show the same-sensor testing results averaged over all sensors (used to build a given dataset) and presented for each benchmark separately. These results suggest that the older benchmark (LivDet2009) is relatively difficult since almost 20% of spoofing samples were falsely accepted in a solution that falsely rejects only 3.45 % of authentic examples.

Table 5.4: Performance results obtained in **same-dataset** evaluations of **fingerprint PAD** using a part of testing samples acquired by the same sensor as in the training procedure. Results are averaged over all subsets representing different sensors.

Training	Testing	APCER (%)	BPCER (%)	HTER (%)	ROC and ePDF
LivDet2009	LivDet2009	19.37	3.45	11.4	Fig. 5.7
LivDet2013	LivDet2013	6.8	2.79	4.795	Fig. 5.8

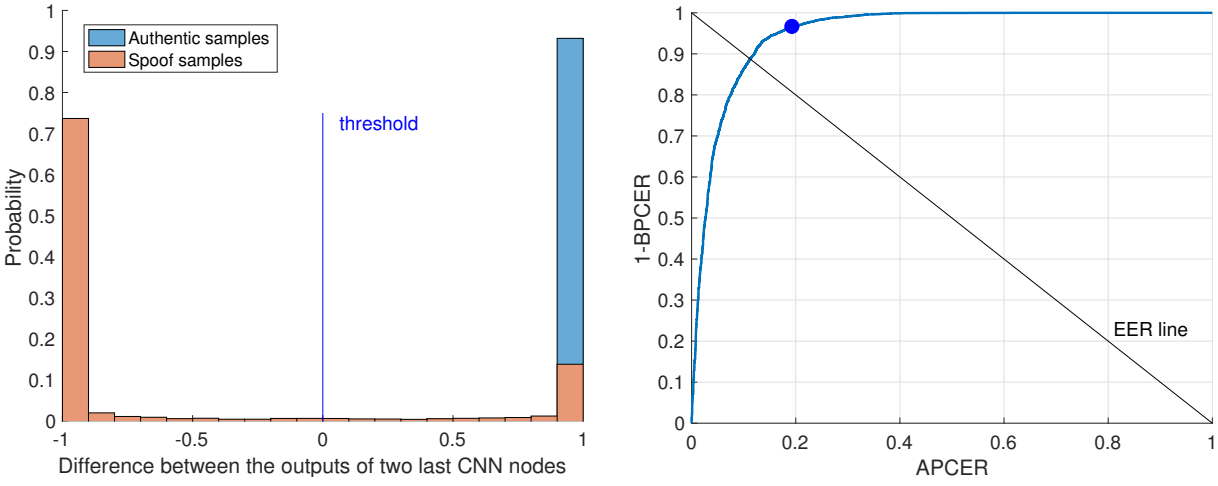


Figure 5.7: **Left:** Empirical distributions of the difference between two CNN output nodes (after softmax) obtained separately for authentic and spoof **fingerprint** samples. **Right:** ROC curve. Variant: training on **LivDet2009**, testing on **LivDet2009**. As in previous plots, the threshold shown in blue color on the left plot and the blue dot on the ROC plot correspond to the approach when the predicted label is determined by the node with the larger output.

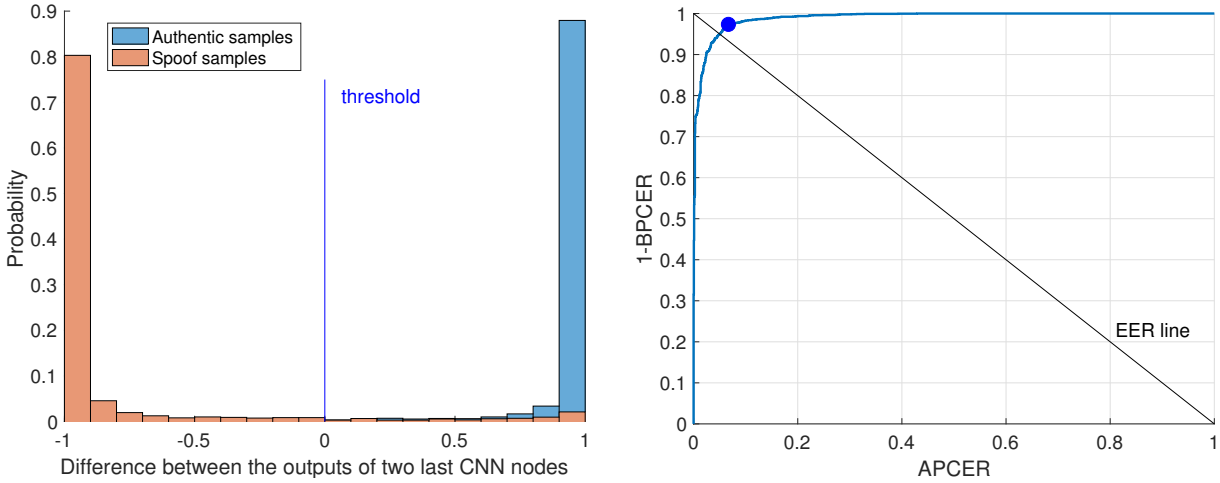


Figure 5.8: Same as Fig. 5.7 except the variant: training on **LivDet2013**, testing on **LivDet2013**.

Cross-sensor results. For cross-sensor analysis, the newer benchmark (LivDet2013) was selected. Each subset (estimation, validation and testing) was divided into two disjoint subsets of samples: a) acquired by ItalData and Swipe sensors, and b) acquired by Biometrika and CrossMatch sensors. Table 5.5 shows that, as with the other modalities, we can observe serious problems with recognition of both artifacts or genuine samples (two first rows of Table 5.5). Figures 5.10 and 5.9, illustrating these results, suggest that a better balance between APCER and BPCER can be found if there is a possibility to adjust the acceptance threshold.

Table 5.5: Performance results obtained in **cross-dataset** evaluations of **fingerprint PAD** using a part of testing samples acquired by different sensor as in the training procedure. All data comes for LivDet2013 fingerprint benchmark. IS = Italdata+Swipe, BC = Biometrika+CrossMatch.

Training	Testing	APCER (%)	BPCER (%)	HTER (%)	ROC and ePDF
IS	BC	24.9	4.01	14.1	Fig. 5.9
BC	IS	2.8	75.6	39.18	Fig. 5.10
IS	IS	3.4	2.37	2.88	Fig. 5.11
BC	BC	2.65	13.1	7.87	Fig. 5.12

For completeness, same-sensor results are also presented on this dataset in two last rows of Table 5.5, and in Figs. 5.12 and 5.11. As expected, a solution based on deep network achieves much better accuracy when the type of sensor in known.

5.5.3 Iris

This last section presents the results of iris presentation attacks detection. Two iris PAD benchmarks were used, as described in Section 5.4), and both same-dataset and cross-dataset experiments were carried out. Each dataset (Warsaw LivDet2015 and AVTS) are already split by their creators into training and testing subsets. We followed this split and

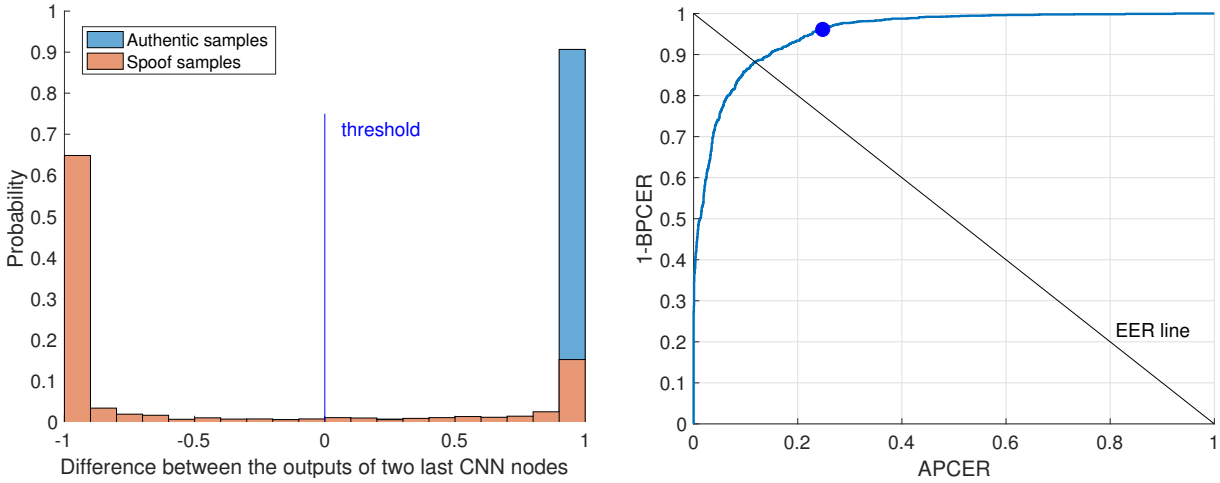


Figure 5.9: Same as Fig. 5.7 except the variant: training on **Italdata+Swipe**, testing on **Biometrika+CrossMatch**.

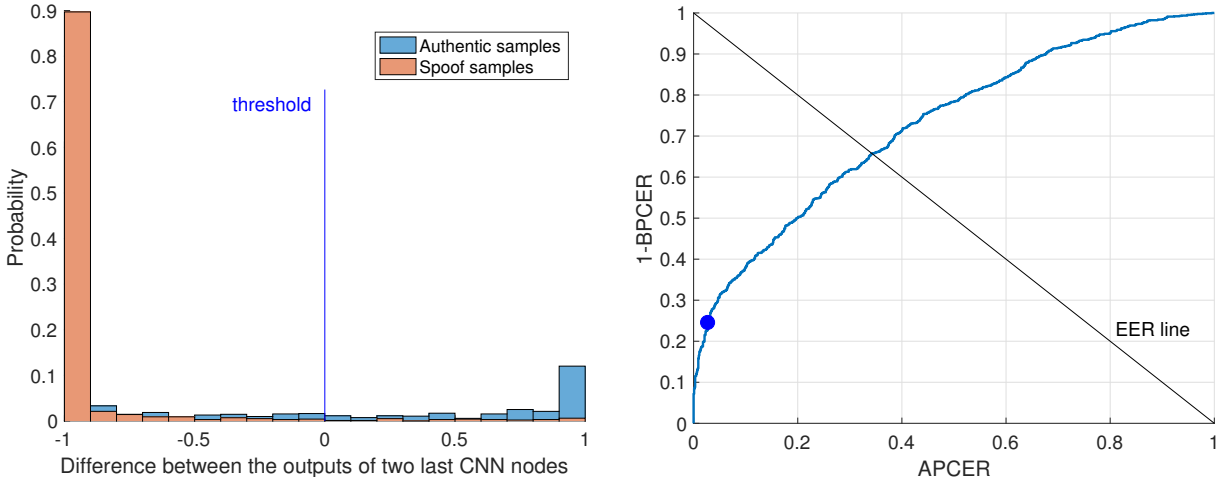


Figure 5.10: Same as Fig. 5.7 except the cross-sensor that training is realized on samples composed **Biometrika+CrossMatch**, testing on **Italdata+Swipe**.

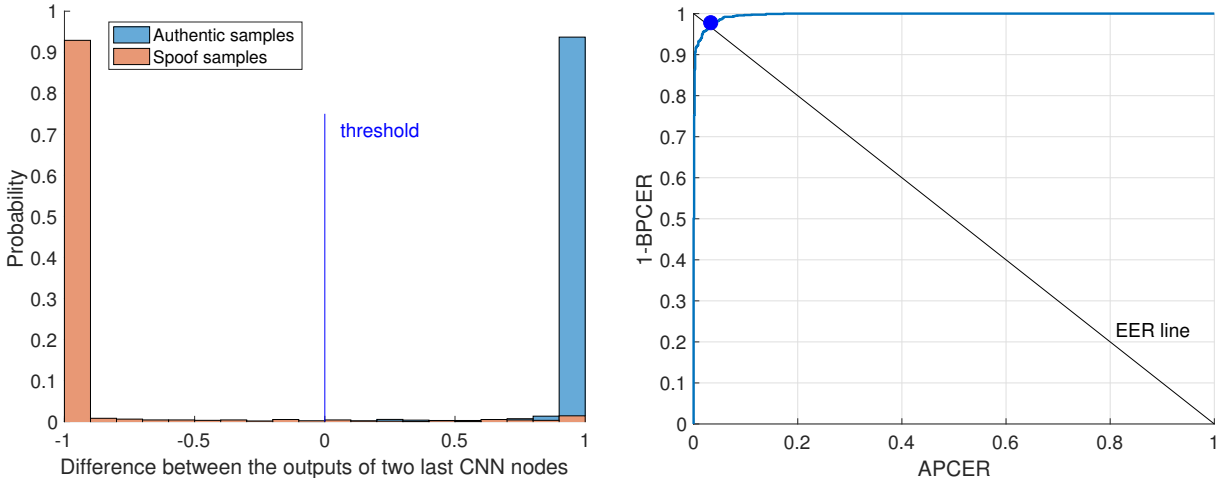


Figure 5.11: Same as Fig. 5.7 except the variant: training on **Italdata+Swipe**, testing on **Italdata+Swipe**.

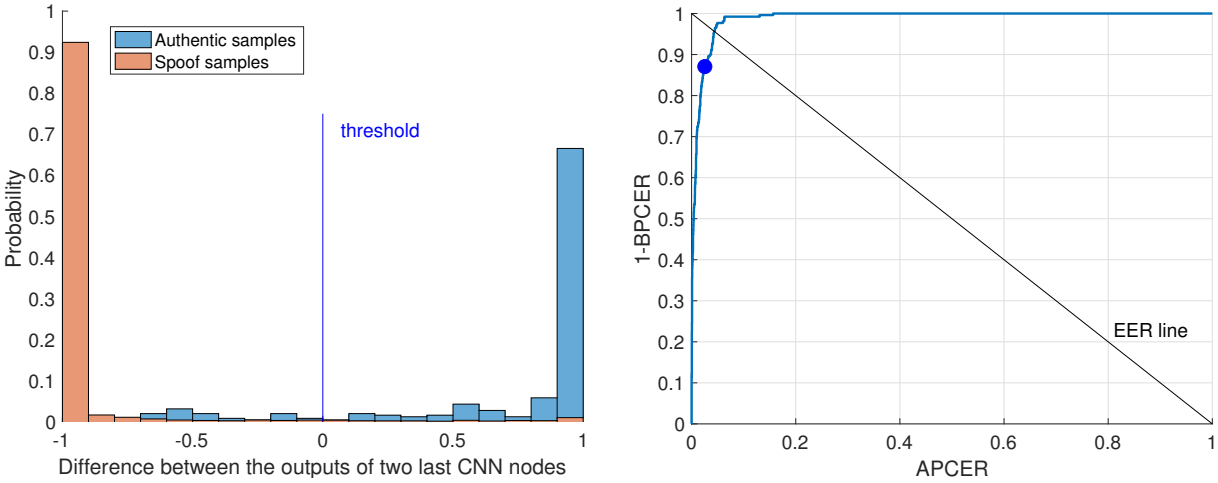


Figure 5.12: Same as Fig. 5.7 except the variant: training on **Biometrika+CrossMatch**, testing on **Biometrika+CrossMatch**.

used the testing subset only in final performance evaluation. The training subset, used in method development, was randomly divided 10 times into estimation and validation disjoint subsets used in cross-validation when training the classifiers.

The average HTER's over 10 splits calculated for validation subsets were approx. 0.0001 and 0.0 for Warsaw and AVTS datasets, respectively. HTER = 0.0 for 5 out of 10 splits of Warsaw training dataset. This means that the VGG-based feature extractor followed by a classification layer trained on our data was perfect on the AVTS dataset, and also it was perfect on half of the splits of the Warsaw benchmark. Since there is no "best split" for either of two datasets, we picked one trained solution presenting perfect performance on the training subsets to evaluate them on the test sets.

Same-dataset results. Table 5.6 presents the testing results obtained in the scenario when both training and testing sets come from the same benchmark. APCER and BPCER refer to classification task, that is each sample belonging to the testing set was classified to one of two classes (authentic or presentation attack) based on posteriori probabilities of class membership estimated by the softmax layer of the trained network. Hence, single APCER and BPCER (point estimators) are presented since this protocol is equivalent to a single acceptance threshold. The results obtained in this scenario are astonishing: the classifiers trained on disjoint subsets of samples originating from the same dataset are either perfect (ATVS benchmark) or close to perfect (a perfect recognition of spoofing samples of Warsaw benchmark with only 0.15% of authentic samples falsely rejected). Figures 5.13 and 5.14 present empirical distributions of the difference between two CNN output nodes and the corresponding ROC curves. The distributions are well separated for both benchmarks, suggesting high performance of the VGG-based solution applied for known spoofing samples.

Cross-dataset results. Table 5.7 shows how catastrophically bad this method may be if tested on **cross-dataset** samples. ATVS and Warsaw samples differ significantly in terms of image properties such as contrast and visibility of iris texture. Especially, all the

Table 5.6: Performance results obtained in **same-dataset** evaluations of **iris PAD** using the overall testing set of each dataset.

Training	Testing	APCER (%)	BPCER (%)	HTER (%)	ROC and ePDF
Warsaw	Warsaw	0.0	0.15	0.075	Fig. 5.13
ATVS	ATVS	0.0	0.0	0.0	Fig. 5.14

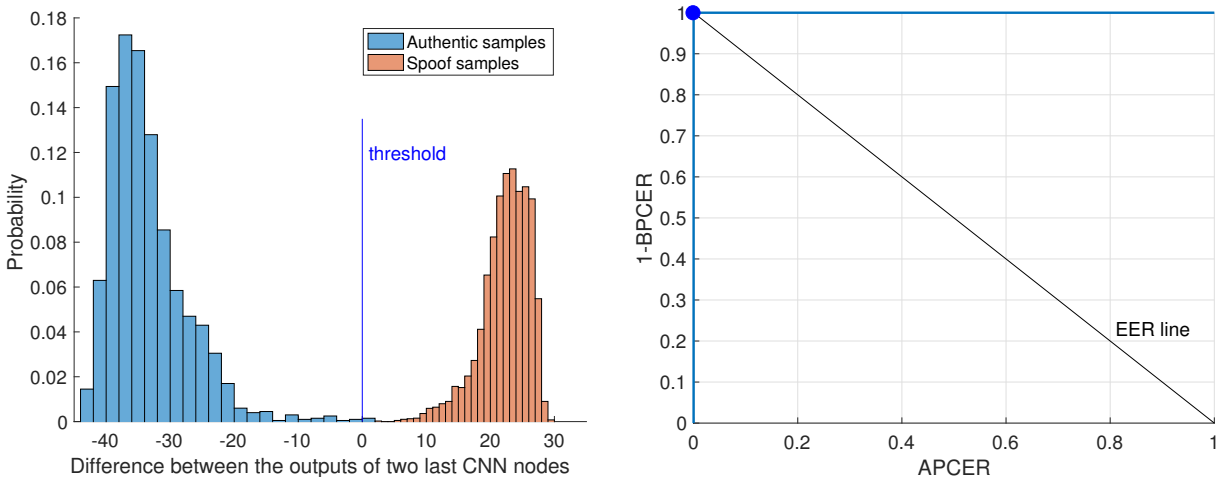


Figure 5.13: **Left:** empirical distributions of the difference between two CNN output nodes (before softmax) obtained separately for authentic and spoof **iris** samples. **Right:** ROC curve. Variant: training on **Warsaw LivDet2015**, testing on **Warsaw LivDet2015**. The threshold shown in blue color on the left plot and the blue dot on the ROC plot correspond to the approach when the predicted label is determined by the node with the larger output.

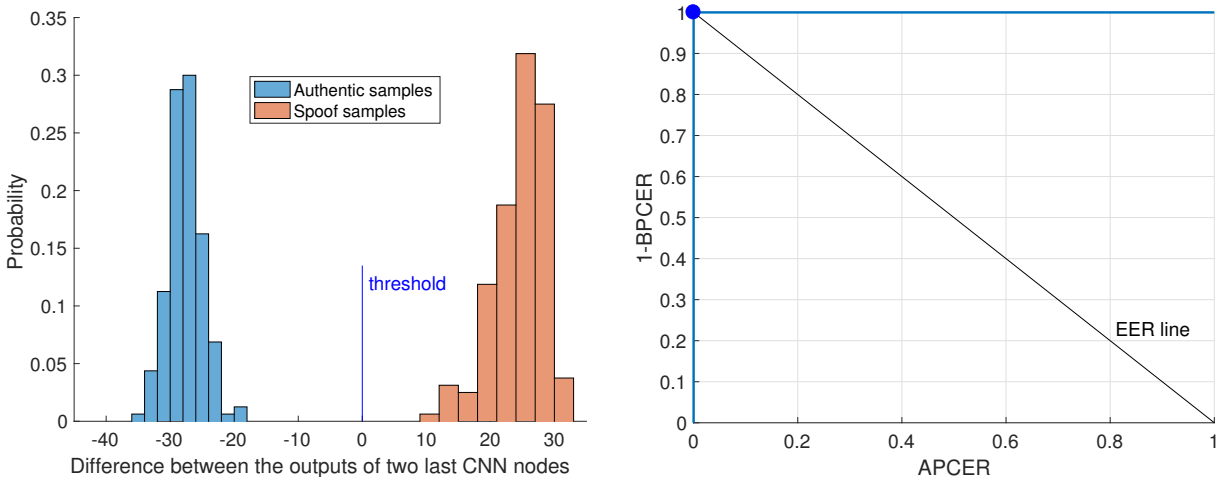


Figure 5.14: Same as Fig. 5.13 except the variant: training on **ATVS**, testing on **ATVS**.

printouts used to produce Warsaw fake samples were able to spoof an example commercial iris recognition system, which is not the case in the ATVS benchmark. Hence, due to non-accidental quality of Warsaw samples, this database seems to be more realistic and more difficult to process than the ATVS. Indeed, training on Warsaw (the "difficult" benchmark) and testing on ATVS (the "easier" benchmark) yields good results. Figure

5.15 presents well separated empirical distributions of the difference between the output nodes of the network obtained for authentic samples and spoofs.

Table 5.7: Performance results obtained in **cross-dataset** evaluations of **iris PAD** using the overall testing set of each dataset.

Training	Testing	APCER (%)	BPCER (%)	HTER (%)	ROC and ePDF
Warsaw	ATVS	0.0	0.625	0.312	Fig. 5.15
ATVS	Warsaw	99.9	0.0	49.99	Fig. 5.16

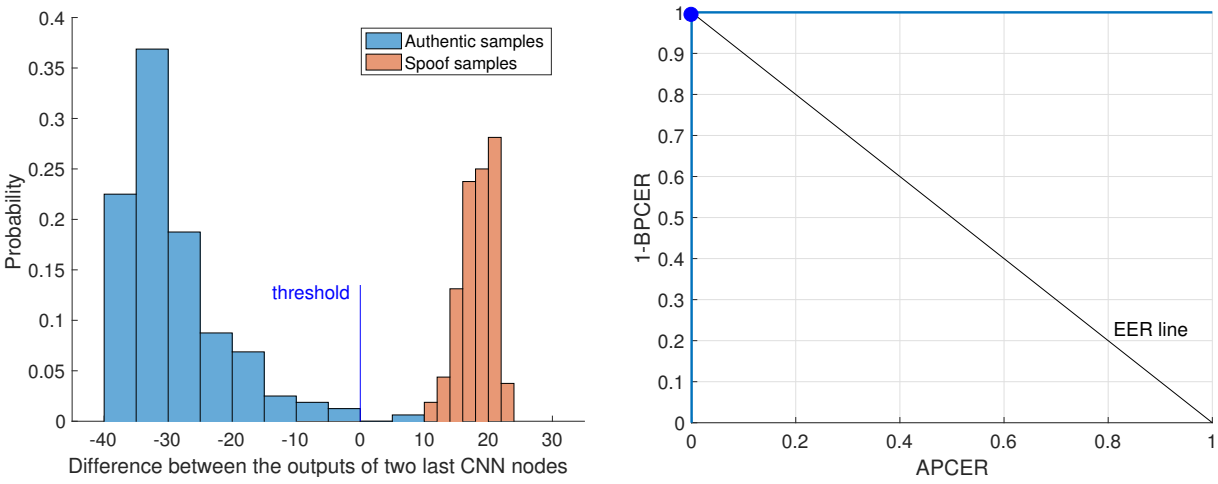


Figure 5.15: Same as Fig. 5.13 except the variant: training on **Warsaw LivDet2015**, testing on **ATVS**.

However, training on ATVS and testing on Warsaw yields almost null abilities to detect spoofs (APCER = 99.9%). This may suggest that exchanging a single layer put on top of the VGG-based feature extraction (trained for a different problem than spoofing detection) is not enough to model various qualities of iris printouts prepared independently by different teams and using different acquisition hardware. Fig. 5.15 confirms that almost all scores obtained for spoofing samples are on the same side of the threshold as for authentic samples. Certainly, if the threshold can be adapted (which is not typically done in the tests), one can find other proportion between APCER and BPCER, for instance a threshold shifted from 0 to -21.9 results in the EER 13.2%.

5.6 Conclusions

In this chapter, we proposed a PAD solution for three modalities widely employed for designing biometric systems (*i.e.*, face, iris and fingerprint) based on VGG network architecture, a deep network architecture originally proposed for object recognition. We showed a methodology to adapt the VGG network to the two-class spoofing classification problem, which was evaluated using six benchmarks available for scientific purposes. The experiments were conducted taking into account the main challenges existing in this research field such as classification across different types of attempted attacks, biometric

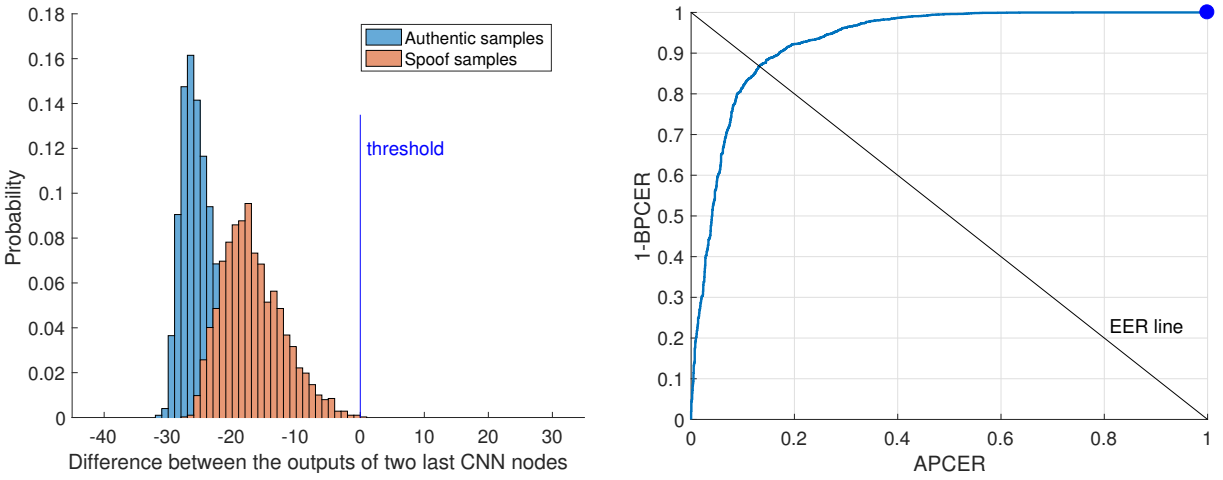


Figure 5.16: Same as Fig. 5.13 except the variant: training on **ATVS**, testing on **Warsaw LivDet2015**.

sensors, and qualities of samples used during attack. In this section, we discuss two main takeaways observed after the analysis presented in this chapter.

The first conclusion is that deep learning is an astonishingly powerful approach to detect image-based presentation attacks in three considered modalities. Note that the final solution is a subtle modification of the VGG network, trained for a different task, not related to presentation attack detection. In the case of iris and fingerprints, the starting network is not even related to the same object recognition task. The results showed that we can use deep learning to detect spoofing attacks in some cases (AVTS iris benchmark) even perfectly. In this simple approach, we have changed only the last layer, connected strictly to the classification task performed by the VGG network. However, one can consider replacing two or all fully connected layers and utilize the output of the convolutional part of the network more efficiently.

The second takeaway comes from the cross-dataset and cross-sensor experiments. These exceptionally poor results seem to be related to the flexibility that characterizes convolutional networks. The flexibility allows them to "decide" which discovered properties of the input data they use in the classification task. But if they are not trained on data that contains a reasonable sampling of the situations present during testing, then they fail terribly, since most of the features no longer correspond to the new data.

This, however, is not a surprising result and simply calls for solutions that take prior knowledge about the modeled phenomenon into account. Apparently the current fascination with deep learning has brought back an old debate: should we use models that are based on our understanding of the problem, which is neither full nor accurate (called feature engineering or "hand-crafted" solutions) or rather flexible models that learn everything from the data (called feature learning or "data-driven" solutions)? It seems that a reasonable mixture of both approaches should present the best reliability. We firmly believe the solution to this problem is in taking the *best of both worlds*.

Acknowledgment

We thank Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) through the DeepEyes project, the São Paulo Research Foundation (FAPESP) through the DéjàVu project (Grant #2015/19222-9), and Microsoft Research for the financial support.

Chapter 6

Leveraging Shape, Reflectance and Albedo from Shading for Face Presentation Attack Detection

"The true work of art is but a shadow of the divine perfection."
—Michelangelo, *Italian sculptor, painter, architect and poet (1475–1564)*

"We can easily forgive a child who is afraid of the dark; the real tragedy of life is when men are afraid of the light."
—Plato, *Greek author & philosopher in Athens (427 BC–347 BC)*

Abstract

Presentation attack detection is a challenging problem in the biometrics that aims at exposing an impostor user that seeks to deceive the authentication process by showing to the acquisition sensor a synthetic sample containing the biometric data of a genuine user previously enrolled in the authentication system. In facial biometrics systems, this kind of attack is performed with a photograph, video, or 3D mask containing the biometric information of a genuine identity. In this paper, we investigate a novel approach to detecting face presentation attacks based on intrinsic properties of the scene such as albedo, depth, and reflectance properties of the facial surfaces, which were recovered through a shape-from-shading (SfS) algorithm. In other to extract meaningful patterns from the different maps obtained with the SfS algorithm, we designed a novel shallow CNN architecture for learning features useful for the presentation attack detection (PAD). We conduct the experimental results considering the intra- and inter-dataset evaluation protocol. The former protocol aims to test a PAD system under a scenario in that the acquisition sensor and the identities are known during the training phase, whereas the latter protocol aims to test a PAD system in a completely unknown scenario, which includes the identities, sensors, and the presentation attack instruments (PAI) used to

generate the attempted attacks. Experimental results showed the effectiveness of the proposed method considering several types of the photo- and video-based presentation attacks, and in the cross-sensor scenario, besides achieving competitive results for the inter-dataset evaluation protocol.

6.1 Introduction

BIOMETRICS is an active research field whose today's challenges go far beyond the making of a high precision system. Nowadays, the security aspects of a biometric system are essential for a successful authentication mechanism due to the vast possibility that an impostor user has for attacking a biometric system. Among different possibilities, presentation attack is the easiest way to deceive such systems since this kind of attack can be performed directly on the acquisition sensor without any previous knowledge of the internal components of the system. It is characterized by the action of presenting a synthetic biometric sample of some valid user to the acquisition sensor in order to authenticate itself as a legitimate user. In this kind of attack, the impostor user does not need any advanced information about internal components of the system, and the minimum requirement for a successfully accomplished presentation attack is a trustworthy copy of the biometric data of a target user such as photographs, digital video, or even a 3D mask [158].

Although several advances have been reported in the literature, face presentation attack problem is still an open problem. According to the IJCB 2017 competition on generalized face presentation attack detection (PAD) in mobile [30], the best algorithm for detecting this kind of attack presented an error detection rate about 10.0%, which means that ten out of every hundred attempted attacks were successfully accomplished, giving to an impostor user unauthorized access to the biometric system. Considering a system with hundreds of thousands of users, this detection rate makes the authentication process unfeasible in practice.

The major problem with the current solutions for PAD is the lack of ability to work in an uncontrolled environment. In fact, the PAD solutions available in the literature present impressive accuracy rates, with near-perfect classification results, for isolated datasets, i.e., in one specific domain. However, when we consider challenging evaluation scenarios, those algorithms present extremely low performance, sometimes becoming worse than random. For this reason, researchers have promoted efforts to report their results considering two evaluation protocols known as intra- and inter-datasets. Intra-dataset evaluation protocol consists of testing a PAD algorithm using data that came from the same source as the training data with training and testing sets being collected using the same acquisition sensor or in the same environment. Inter-dataset protocol consists of testing a PAD algorithm using data from a different source as the training data, which means we have data from a different domain (i.e., different sensor and environment). Such evaluation protocol is more challenging and more suitable for reflecting a real operating scenario.

According to recent results reported in the literature, the Half Total Error Rates (HTER) can increase drastically taking into account these evaluation protocols.

Pinto et al. [189] proposed a PAD method based on analysis of the noise and artifacts left in the synthetic biometric sample during its manufacture such as blurring, printing effect, banding effect among others. Although the authors achieved a low HTER value for the intra-dataset evaluation protocol (2.8%), the HTER of this technique increases significantly, considering the inter-dataset protocol (34.4%). Even in the state-of-the-art techniques, values for error rates are still too high. Boulkenafet et al. [28] reported an HTER of 2.9% and 16.7% considering the intra- and inter-dataset protocols, respectively, which means a degradation of about 4×, in terms of HTER, which is far from an acceptable value in practice.

In this paper, we present a novel approach to distinguishing a synthetic face from real ones, taking into account the optical and physical properties of the scene captured by the acquisition sensor. Our method takes advantage of the depth information, associating it with light properties of the scene to detect an attempted attack, using a classic technique in computer vision known as shape-from-shading (SfS). SfS was firstly proposed by Horn et al. [97] and aims to estimate the shape of an object based on the shade information present in its surface. In contrast with the photometric stereo techniques, SfS techniques require only one image of the object under analysis. Moreover, the estimation of the shape using SfS does not require any additional hardware, which makes possible the application of our technique in devices equipped with only an RGB camera such as smartphones and webcams. Fig. 6.1 illustrates a face surface reconstruction using the Tsai's algorithm [188], which will be described in details in Section 6.3.

(a) RGB image (b) Depth map (c) Reflectance map (d) Albedo map

Figure 6.1: Example of a facial surface reconstruction using an SfS algorithm for presentation attack video frame.

6.1.1 Optical Properties of the Light and Rationale of Proposed Approach

According to the law of refraction [249], the physical mechanism of the light refraction can be characterized in terms of absorption and irradiation of the light incident in a surface. Complementary, the reflection's law governs the reflection of the incident light and states that the incident ray, the reflected ray, and the normal to smooth conducting surfaces (e.g., mirror or polished metal) all lie in the same plane [249]. Basically, the beam of light that affects a flat surface may be absorbed, transmitted, and reflected, and that, the light

reflected can be mathematically understood by Snell and reflection laws, which predict the directions of the light refracted and reflected, respectively, taking into account the refraction index of the material and the roughness of its surface, that is, the smoothness or texture of the surface.

understanding of the physical mechanism of the light refraction, in terms of absorption and irradiation of the light incident on a surface, while the reflection's law governs the reflection of the incident light. According to refraction and reflection laws, the beam of light that affects a flat surface may be absorbed, transmitted, and reflected, and the directions of the reflected and refracted light can be predicted considering the refraction index of the material and the roughness of its surface, i.e., the smoothness or texture of the surface.

When a beam of light affects a truly flat surface, each incident ray is reflected at the same angle that we have between the surface normal and such incident ray, but on the opposite side of the surface normal. In contrast, when a beam of light affects rough surfaces, the incident light is reflected in several different directions. An ideal diffuse reflecting surface that reflects the incident light in all directions is said to exhibit a Lambertian reflection. These two processes are known as specular and diffuse reflection, respectively. Although many materials can exhibit both types of reflection, some materials reflect the light more diffuse than specular way (e.g., paper fibers, non-absorbing powder such as plaster, poly-crystalline material such as white marble, among others) [70, 116, 165, 252].

The reflecting power of the material is another interesting physical property that we believe to be useful for the presentation attack detection problem. This property is also known as surface albedo and can be defined as a measure of how much light incident on a surface is reflected without being absorbed. In other words, this property measures the reflectivity of a material and gives an estimate of the level of the diffuse reflection [103,275]. Thus, the objects that appear white reflect most of the incident light, indicating a high albedo, whereas the dark objects absorb most of the incident light, indicating a low albedo.

Finally, the last physical property we investigate in this work is the depth information associated with an object in the scene. Considering the presentation attack instruments known in the literature (e.g., photograph, video replay, mask), we clearly have a significant loss of depth information, excepting the mask-based presentation attacks. In fact, several works published in the literature have successfully investigated features able to characterize the depth information of face regions to point out an attempted attack [66, 67, 84]. Basically, these approaches propose to use depth sensor, such as Microsoft's Kinetic sensor, to find an accurate depth map of the scene.

6.1.2 Contributions and Organization

In light of these remarks, we propose a novel method for detecting face presentation attacks using an SfS technique, which is used for recovering the surface details of the faces captured by the acquisition sensor. Our hypothesis is that the reconstructed surface from the presentation attack samples might contains strong evidence of synthetic patterns in comparison to the RGB images. To the best of our knowledge, this is the first attempt

at using this kind of algorithm for the PAD problem. In summary, the main contributions of this paper are:

- a new method for face presentation attack detection based on intrinsic properties of the surfaces reconstructed with an SfS algorithm;
- a new shallow CNN network able for learning discriminant features from the albedo, reflectance, and depth maps for the PAD problem, which achieved competitive results for the intra- and inter-dataset evaluation protocol;
- an investigation of a new data-driven PAD algorithm, where the depth maps are recovered without any extra hardware, which allows the use of the proposed method in systems equipped with a single RGB sensor;
- an evaluation of the proposed approach considering challenging protocols such as inter-dataset protocol and the cross-sensor scenario; and
- the investigation of using an SfS technique in a challenging problem upon hundreds of thousands of real images.

We organize the remaining of this paper as follows. Section 6.2 presents some relevant related approaches to face presentation attack detection. Section 6.3 describes the proposed method. Section 6.4 presents the datasets and evaluation protocols used in this paper, besides the experimental results and a comparison with methods available in the literature. Finally, Section 6.5 presents the conclusions and possible directions for future work.

6.2 Related Work

Texture analysis is undoubtedly an important and promising line of investigation that made possible many progress in this research field toward finds good PAD algorithms. Back to the First Competition on Counter Measures to 2D Facial Spoofing Attacks [36], the best proposed algorithms [151, 214] explored different texture descriptors, such as Local Binary Patterns (LBP), Gray-Level Co-Occurrence Matrix (GLCM), Histogram of Oriented Gradients (HOG), among others, for detecting printed-based attempted attacks [8].

In order to push the state-of-the-art to past the breaking point, the Second Competition on Counter Measures to 2D Face Spoofing Attacks [42] presented to the community a novel dataset (Replay-Attack dataset) [41] containing three different attacks types, print-, photo-, and video-based attacks, where the winner teams addressed the problem through a feature-level fusion of texture- and motion-based features. The Replay-Attack was quite challenging at that time, making a possible further interesting investigation of other cues for detecting face presentation attacks.

Erdogmus and Marcel [65–67] explored depth information for detecting face presentation attacks by analyzing both color and depth data obtained by Microsoft's Kinect sensor. The authors proposed to use the Local Binary Patterns (LBP) descriptor in both color

and depth images to produce feature vectors, which were used to feed a Linear Discriminant Analysis (LDA) classifier to reveal an attempted attack. Pinto et al. [189, 190, 193] also brought alternatives for detecting face presentation attacks exploiting the residual noise presents in the fake biometric sample left during their recapture and reconstruction such as blurring effect, printout artifacts, Moiré patterns, among others. Similar, Garcia and Queiroz [77] and Wen et al. [260] explored these and other artifacts related to image distortions caused mainly by the recapture process of the original biometric signal.

Another cue that has been an object of investigation in the literature is regarding the reflectance of the objects. Although skin reflectance presents great variation due to different tonalities of human skin [49, 161], researchers have successively used it in several applications [55, 92, 144]. In these cases, however, the reflectance is measured through extra-devices, for instance, thermal infrared imagery and near-infrared imagery. Alternatively, some computational methods for estimating the reflectance map of a scene from RGB images [5, 48, 98] have been proposed in the literature to decompose an RGB image into their reflectance and illumination components [128].

CNN-based techniques also have been considered in the literature. Menotti et al. [162] proposed a framework for optimizing CNN architectures for PAD problem considering different modality, include face biometrics. The authors also proposed a shallow CNN network was applied for detecting iris, fingerprint, and face presentation attacks. Although the authors achieved good results using this technique, this work did not consider more challenging protocols such as inter-dataset protocol and cross-sensor scenario. Atoum et al. [11] combines a patch- and depth-based CNN for face PAD, in which the authors also achieved good results for the intra-dataset evaluation protocol. However, this works also reported their results only in the intra-dataset protocol.

Several works in the literature presented a fine-tuning of existing deep architectures such as AlexNet, VGG, VGG-Face, and GoogleLeNet [142, 180, 192, 269]. However, in general, these architectures achieved near-perfect classification results for the intra-dataset and, at same time, very poor results (close to random) for the inter-dataset protocol. Recently, Rehman et al. [208] proposed a new CNN-based anti-spoofing technique using the VGG-11 architecture, in which the authors reported impressive results for the intra- and inter-dataset scenario. However, a serious methodological failure described by the authors in Sec. 4.2.2 of the original paper [208], made any comparison unfeasible. As mentioned by the authors, part of the testing dataset was used to estimate the threshold τ, which was used for computing the APCER, BCPER, and HTER values. More precisely, considering the inter-dataset protocol, in which we have a training dataset and a testing dataset, the authors used the training partition contained in the test dataset for estimating the threshold τ, which obviously biased the reported results. In contrast to Rehman et al., this paper and the other ones published in the literature use the testing dataset only to report the performance results.

Differently from the previous works in the literature, in this work, we propose a PAD technique that takes advantage of depth, albedo, and reflectance information from RGB-images, without the necessity of any extra-device such as Microsoft's Kinect or an infrared sensor. Instead of using different methods for computing each one of these components, we propose to use a shape-from-shading algorithm, which enables us to estimate these three

representations from a single RGB image. We also propose a new CNN architecture able to work in the intra- and inter-dataset scenario. To the best of our knowledge, our work is the first one to deal with these three schemes simultaneously using a shape-from-shading technique for detecting face presentation attacks.

6.3 Proposed Method

In this section, we present our proposed method for face PAD which is based on intrinsic properties of the surface such as reflectance, albedo, and shape. As previously described, we propose the use of SfS for measuring these properties and use them as input for a Convolutional Neural Network (CNN) method which learns discriminative features for detecting presentation attacks. The advantage of using an SfS method, instead of using an extra-device sensor, is two-fold: (i) a shape-from-shading method gives us an estimation of these three properties at once, and (ii) we came up with a completely data-driven method, which enables our method for use in biometric systems equipped with only an RGB camera such as smart-phones.

The human ability for perceiving the shape of the objects from its shading is one of the most important aspects of the human visual system. This ability is essential for the human understanding of the world under a three-dimensional perspective [201]. Studies have been published showing that human can accurately use shading cues to infer changes in the surface orientation [183, 201, 261]. In computer vision, there are two main classes of methods for estimating the shape from shading: photometric stereo and shape-from-shading methods. An essential difference between them is that photometric stereo methods require two or more images of the same object under different lighting conditions, whereas shape-from-shading methods require only one image of the object to estimate its normal surfaces, making SfS methods very attractive to our problem [240].

We believe that some SfS methods are more appropriate to be applied in our problems than others approaches found in the literature, according to assumptions and restrictions imposed during the formulation of the problem. For instance, methods that add a smoothness constraint to the surface might be inadequate to be used in our problem because such constraint is not contemplated when recovering the shape of faces due to some cavities.

Our work is based on Tsai's approach [188], which does not impose any restriction that could render its use improper for the PAD problem. Fig. 6.2 illustrates our proposed method, which is explained in details next.

6.3.1 Surface Reconstruction: Recovering the Depth, Reflectance and Albedo maps

The Tsai's algorithm [188] uses a linear approximation of reflectance function R to estimate the depth function Z from a single image. The main idea is to apply a discrete approximation for the surface normal using the finite differences method in order to linearize the reflectance function R in terms of Z, and then to solve the linear system through the Jacobi iterative method [220].

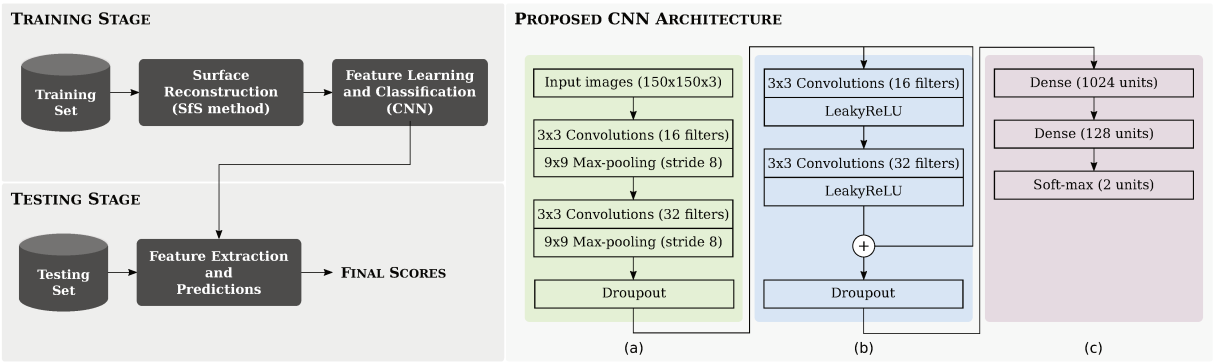


Figure 6.2: Overview of the proposed method for face presentation attack detection. First, we reconstruct the face surfaces by using the SfS method, which produces estimates for the albedo, reflectance, and depth maps. Then, we use a representation learning techniques based on CNN to learn discriminative features from these maps estimated from the training samples, which give us a classification model to decide if a given testing sample is a bona fide presentation or a presentation attack.

Let us, for instance, suppose that a point at position (x, y, z), in camera coordinates, is at a distance z from the image plane and there is a mapping between points in camera coordinates onto the image plane created using the parallel projection (not taking into account any sort of distortion). Assuming that depth information is a function of image plane coordinates $Z = Z_{x,y}$, then the change in depth δz of the point related to the change in image plane coordinate (x, y) can be expressed by using the Taylor series expansion of the function Z about point (x, y) as:

$$\delta z \approx \frac{\partial z}{\partial x}\delta x + \frac{\partial z}{\partial y}\delta y \tag{6.1}$$

The gradient of the surface at point (x, y, z) is the vector $(p, q) = (\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y})$ and, therefore, the normal of a surface patch is related to the gradient by $\mathbf{n} = (p, q, 1)$, since the gradient vector is orthogonal to the level surface $Z_{x,y}$.

Now, suppose a Lambertian surface, which has only diffuse reflectance and the brightness is proportional to the energy of the incident light. In this case, the amount of light energy falling on a surface is proportional to the area of the surface as seen from the light source position, which can be expressed as:

$$\begin{aligned} E_{x,y} &= R(p, q) = \rho I(\mathbf{n} \cdot \mathbf{s}) \\ \Rightarrow R(p, q) &= \rho \frac{(-p, -q, 1)}{\sqrt{1 + p^2 + q^2}} \cdot \frac{(-p_s, -q_s, 1)}{\sqrt{1 + p_s^2 + q_s^2}} \end{aligned} \tag{6.2}$$

where $E_{x,y}$ is the intensity at pixel (x, y), I is the illuminance (or strength of light), $\mathbf{n} = (-p, -q, 1)$ is the surface normal, $\mathbf{s} = (-p_s, -q_s, 1)$ is the light source direction, and ρ is the albedo of the surface.

The SfS method employed in this work uses a discrete approximation for p and q as shown in Equation 6.3 and performs a linear approximation of Equation 6.4 based on

Taylor series expansion considering the first order terms of the function f about a given depth map Z^{n-1}, which give us a linear system of equations (Equation 6.5).

$$\begin{aligned} p &= \frac{\partial z}{\partial x} = Z_{x,y} - Z_{x-1,y} \\ q &= \frac{\partial z}{\partial y} = Z_{x,y} - Z_{x,y-1} \end{aligned} \tag{6.3}$$

$$\begin{aligned} 0 &= f(E_{x,y}, R(\partial z/\partial x, \partial z/\partial y)) \\ 0 &= E_{x,y} - R(Z_{x,y} - Z_{x-1,y}, Z_{x,y} - Z_{x,y-1}) \end{aligned} \tag{6.4}$$

$$\begin{aligned} 0 &= f(Z_{x,y}) \\ &\approx f(Z_{x,y}^{n-1}) + (Z_{x,y} - Z_{x,y}^{n-1}) \frac{d}{dZ_{x,y}} f(Z_{x,y}^{n-1}) \end{aligned} \tag{6.5}$$

When we consider the $Z_{x,y} = Z_{x,y}^{n}$, that is, the depth at n-th iteration, the Equation 6.5 can be rewritten (Equation 6.6) and solved by using the Jacobi iterative method [220], considering an initial estimate of the depth map $Z_{x,y}^{0} = 0$.

$$Z_{x,y}^{n} = Z_{x,y}^{n-1} + \frac{-f(Z_{x,y}^{n-1})}{\frac{df(Z_{x,y}^{n-1})}{dZ_{x,y}}} \tag{6.6}$$

The reflectance and albedo maps also can be obtained directly from the Equation 6.2. After find the depth map $Z_{x,y}^{n}$ at point (x, y), the reflectance map can be computed by using the Equation 6.7, while the albedo map can be found through the Equation 6.8.

$$R(p, q) = \max\left(0, \rho \frac{pp_s + qq_s + 1}{\sqrt{1 + p^2 + q^2}}\right) \tag{6.7}$$

$$\rho_{x,y}^{(n)} = \frac{I_{x,y}}{\mathbf{n}_{x,y}^{(n)} \cdot \mathbf{s}} \tag{6.8}$$

6.3.2 Convolutional Neural Network for Learning Intrinsics Surface Properties

Convolutional Neural Networks (CNNs) [133] is a well-known machine learning technique designed to learn discriminative features from input data and also a mapping function for classification purpose. Their ability to learn an efficient and effective representation space from data has been extensively reported by the scientific community, producing impressive results in many applications such as object recognition [239], video analysis [185], presentation attack detection [30, 162], among others.

Inspired by Menotti et al. [162] and He et al. [95] approaches, our CNN architecture is

composed of a variant of *SpoofNet* followed by one residual block. The original SpoofNet is a shallow CNN architecture composed of two convolutional layers, containing 16 and 64 filters, respectively, with a kernel size of 5×5. Each convolutional layer is followed by a max-pooling layer, with a kernel size of 3×3 and a stride of 2, and by a local normalization layer with a kernel size of 9×9.

In contrast to *SpoofNet*, our new CNN architecture for face presentation attack detection is described in Fig.6.2. Essentially, we propose an ensemble of two shallow CNN architectures (Fig. 6.2(a) and Fig. 6.2(b)) containing an identity shortcut link used to connect the output of these two blocks. Finally, we use two dense layers followed by a soft-max classifier with two units (Fig. 6.2(c)).

6.4 Experimental Results

In this section, we present the experimental results of our proposed approach. Section 6.4.1 describes the datasets used in the experiments, whereas Section 6.4.2 describes the experimental protocols used to validate our approach. Section 6.4.3 shows the experimental setup of the proposed method regarding its parameters, and Sections 6.4.4 and 6.4.5 show the obtained results using the maps obtained with the shape-from-shading algorithm and feature learning process. The remaining sections describe the performance results considering the intra- and inter-dataset evaluation protocol and a comparison among the proposed method and other approaches reported in the literature.

6.4.1 Datasets

We evaluated the proposed method in three datasets freely available in the literature for scientific purpose, which is described in details in the following sections:

Replay-Attack dataset

This dataset contains videos of presentation attacks and bona fide presentations of 50 identities, which were recorded with a webcam with a pixel resolution of 320×240. This dataset provides three types of presentation attacks: print-, mobile- and video- attacks with high definition resolution, which were split into three subsets: the training set with 360 videos; the development set containing 360 videos; and testing set with 480 videos, totaling $1,000$ videos of presentation attacks and 200 videos of bona fide presentation.

CASIA dataset

This dataset comprises 600 videos of presentation attacks and bona fide accesses of 50 identities. The authors recorded both presentation attack and bona fide presentation videos in three different qualities: (i) low-quality videos captured by an old USB camera with 480×640 pixel resolution; (ii) normal-quality videos, which were recorded by a new USB camera with 480×640 pixel resolution; and (iii) high-quality videos captured with a Sony NEX-5 camera with $1,920 \times 1,080$ pixel resolution. The types of presentation attacks contained in this dataset include warped photo attacks, cut photo attacks, photos

and video attacks. Finally, this dataset provides 240 videos for training and 360 videos for testing, totaling 150 videos of bona fide presentations and 450 videos of presentation attacks.

UVAD dataset

This dataset contains bona fide presentation and presentation attack videos of 404 identities, all created at Full HD quality. The videos were recorded in two sections considering different illumination conditions and environments. In total, this dataset provides 16,268 presentation attack videos and 808 bona fide presentation videos, which were recorded through six acquisition sensors of different manufacturers (Sony, Kodak, Olympus, Nikon, Canon, and Panasonic). The video attacks were simulated with seven different display devices, also with HD and Full HD quality. The authors recommend using the videos from Sony, Kodak and Olympus sensors for training, and the videos from Nikon, Canon and Panasonic sensors for testing. This evaluation protocol provides 3,872 for training and 6,416 for testing, totaling 404 bona fide presentation videos and 9,884 presentation attack videos.

6.4.2 Experimental Protocols

The performance of the proposed method is measuring by using two metrics recommended by ISO/IEC 30107-3 [106], Attack Presentation Classification Error Rate (APCER) and the Bona fide Presentation Classification Error Rate (BPCER), wherein the APCER is the proportion of presentation attack incorrectly classified as bona fide presentations and the BPCER is the proportion of bona fide presentations incorrectly classified as presentation attack. Although, the ISO/IEC does not define measures that aggregate these two measures, in this work we additionally use two measures for that, the Equal Error Rate (EER) and Half Total Error Rate (HTER), since the evaluation protocol for some datasets recommends to use them. The EER value is defined by the threshold in that the APCER and BPCER rates are equals, and the HTER is the average of APCER and BPCER measures computed in a threshold τ, which must be defined in a development set.

We evaluated our approach upon the two experimental protocol, the intra- and inter-dataset protocols. In the intra-dataset scenario, we validate the proposed method into each dataset separately, and we follow the official protocols defined in each dataset considered in this work. Therefore, the Replay-Attack dataset is comprised of three subsets: the training set, which was used to fit a classification model; the development set used to find the EER threshold; and the test set, which was used only to report the APCER, BPCER, and HTER values. For the datasets composed of two subsets (CASIA and UVAD), we use the training set to fit a classification model and to find the EER threshold, and the test set to report the final results in terms of APCER, BPCER, and HTER. We also reported the EER value obtained in the test set for the CASIA dataset, as suggested by the dataset's authors. In the inter-dataset scenario, we use one dataset for training the proposed method and a a different dataset to test it .

6.4.3 Experimental Setup

This section describes the parameter configurations and implementation details of the proposed method for reproducibility purposes of the results presented in this paper.

Regarding the shape-from-shading algorithm used in this work, the only parameter required by this algorithm is the light source direction, whose value has been set to coordinate $(0, 0, 1)$. Therefore, we considered that the primary light source is perpendicular to the faces during the acquisition, which is a reasonable choose taking into account the datasets used in our experiments. As the shape-from-shading algorithm works upon images, we subsample the videos to have about 61 frames per video (≈ 2 seconds). Moreover, we apply the shape-from-shading algorithm to each channel of the RGB frames cropped in the face regions, whose locations were provided by the datasets' authors. Finally, we resize the maps found by the algorithm to (150×150), which were used to feed the CNN networks.

We conducted the training process of the CNN networks using 150 epochs and batches of 64. We used the Adadelta solver for minimizing the categorical cross-entropy objective function using a learning rate of 1e$-$2 without the learning decay strategy. Finally, we use an L2 regularization in the soft-max classifier, whose value was configured to 1e$-$4. The seeds were pre-defined in order to obtain reproducibility of our results. Finally, the class decision (bona fide presentation vs. presentation attack) for an input video was taken considering the fusion scores of its 61 frames by computing the median. We use the Keras (version 2.1.3) and Tensorflow (version 1.4.1) frameworks [1] to implement the proposed CNN network and the source code of all proposed methods will be freely available [2].

6.4.4 Evaluation of the Proposed CNN Architecture

In this section, we evaluate the CNN network proposed in this work, which was inspired by the SpoofNet [162], a shallow network designed for the presentation attack problem and by the Residual Networks (ResNet) [95]. Here, we show the effectiveness of the proposed SpoofNet-based residual block for detecting presentation attacks by comparing our CNN network with the SpoofNet and ResNet, besides of Xception network [43], a relative new CNN architecture that outperforms Inception V3 on the ImageNet dataset [133]. For both Xception and ResNet networks, we performed a fine-tuning of a pre-trained version trained upon the ImageNet dataset since we do not have much data for training them from scratch. Thus, after loading the pre-trained weights, we remove the top layer and we freeze the remain layers to indicate that such layers will not be trained. Thereafter, we add a fully connected layer with $1,024$ units followed by a soft-max layer with 2 outputs.

Table 6.1 depicts a comparison among these CNN networks for the CASIA dataset using the intra-dataset protocol. Both the SpoofNet and the proposed network outperforms the ResNet and Xception networks. We believe that shallow networks are more suitable for the PAD problem due to nature of the patterns to be learned by the networks, which came from the artifacts added in the synthetic samples such as blurring, banding effect,

[1] https://keras.io and https://www.tensorflow.org

[2] The source code will be public and freely available for scientific purposes on GitHub, upon acceptance of this paper.

Moiré patterns, among others. Noticeably, such patterns can be better understood as low-level features and deeper networks are suitable for learning high-level features such as part of complex objects.

Table 6.1: Performance results (in %) for the CASIA dataset considering the intra-dataset evaluation protocol.

Architecture	Map Type	APCER	BPCER	HTER	Mean HTER
ResNet	Albedo	68.89	68.89	68.89	58.21
	Depth	34.81	48.89	41.85	
	Reflectance	65.56	62.22	63.89	
Xception	Albedo	8.52	78.89	43.70	38.08
	Depth	18.15	55.56	36.85	
	Reflectance	29.63	37.78	33.70	
SpoofNet	Albedo	8.15	14.44	11.30	11.11
	Depth	14.44	11.11	12.78	
	Reflectance	8.52	10.00	**9.26**	
Proposed Method	**Albedo**	6.67	8.89	**7.78**	**8.64**
	Depth	11.11	5.56	**8.33**	
	Reflectance	15.19	4.44	9.81	

6.4.5 How to Feed the Proposed CNN Network with the Different Maps?

Table 6.2: Performance results (in %) for the Replay-Attack and CASIA datasets considering the intra-dataset evaluation protocol.

Dataset	Map Type	HTER
Replay-Attack	Albedo	8.00
	Depth	2.62
	Reflectance	3.87
	Majority Vote	3.38
	Concatenated Maps	**3.12**
CASIA	Albedo	7.78
	Depth	8.33
	Reflectance	9.81
	Majority Vote	5.37
	Concatenated Maps	**2.41**

In this section, we evaluate two strategies to extract meaningful information from the different maps using the proposed CNN network. The experiments presented in this section were performed using the intra-dataset evaluation protocol. The first strategy consists of training a CNN network for each one of the three types of maps available (albedo, reflectance and depth maps), which give us three CNN-based classifiers. After,

the fusion approach based on the majority vote is employed in order to have a final score to decide if a testing sample is a presentation attack or a genuine access. The second approach consists of giving to the network the concatenated maps in order to have an input tensor of $150 \times 150 \times 9$. Table 6.2 shows the obtained results considering these two strategies.

According to obtained results, the concatenated maps outperforms the majority fusion strategy with a relative error reduction of 7.69% for the Replay-Attack dataset and more than 50.0% for the CASIA dataset. Besides of having a significant reduction in the overall time consuming for providing the final decision score, once we need to train only one model, the concatenated maps strategy also facilitate the training of our CNN-based classifier. This is because different maps may behave as a data augmentation approach towards avoiding possible problems regarding over-fitting.

6.4.6 Intra-dataset Evaluation Protocol

In this section, we present the performance results of our approach for the datasets considered in this work. Here, we followed the evaluation protocol defined for each dataset and we also reported the performance results using the metrics suggested by the datasets' authors.

Replay-Attack Dataset

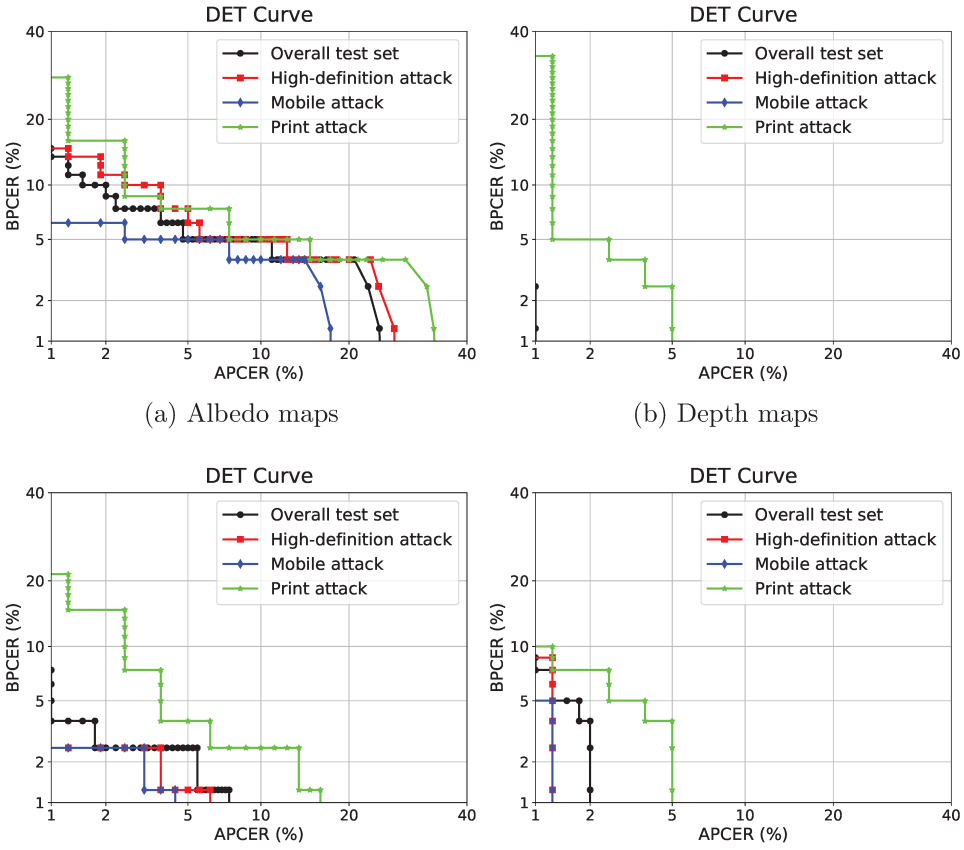
Fig. 6.3 shows the obtained Detection Error Tradeoff (DET) curves for the different maps obtained by the shape-from-shading algorithm and for the three type of presentation attacks contained in this dataset. The aims of this experiment are investigating the discriminability of these maps for detecting the different attack types. The results indicate that mobile-based presentation attacks were the most easily detected by the proposed algorithm. Considering the depth map (Fig. 6.3(b)), the proposed approach achieved an HTER of 2.62% considering the overall test set and perfect BPCER rates for all attack types. Table 6.3 shows the performance results for the network trained using the depth maps.

Table 6.3: Performance results (in %) for the Replay-Attack dataset considering the presentation attacks simulations individually.

Attack Type	APCER	BPCER	HTER
Hight-Def	4.37	0.00	2.19
Mobile	**3.12**	**0.00**	**1.56**
Print	11.25	0.00	5.63
Overall test	5.25	0.00	2.62

CASIA Dataset

Fig. 6.4 illustrates the obtained DET curves considering the different presentation attack simulations. Here, the network trained with the concatenated maps achieved the best



(a) Albedo maps (b) Depth maps

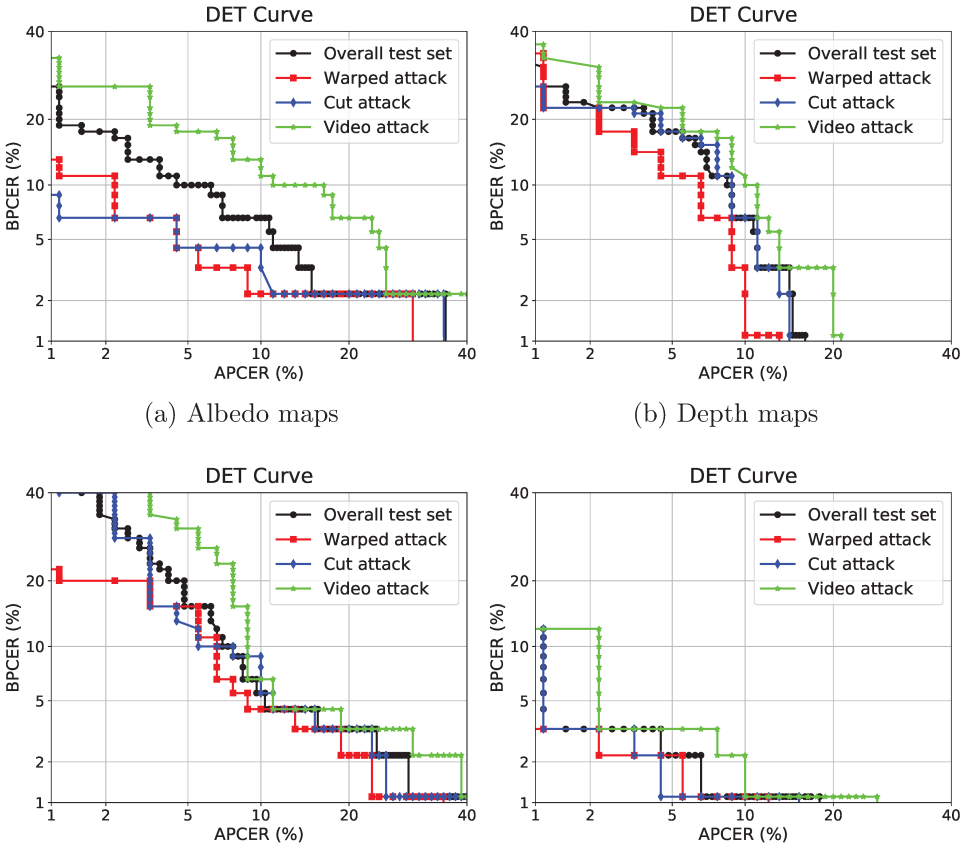
(c) Reflectance maps (d) Concatenated maps

Figure 6.3: Results obtained on Replay-Attack dataset for the three attack types and for the different maps obtained with the shape-from-shading algorithm.

performance results for all categories of attack present in this dataset. Furthermore, the warped- and cut-based attacks were detected easier than video-based attempted attacks. We also notice that network trained with the concatenated maps showed a more robustness to deal with the different types of presentation attack. Table 6.4 shows the error rates for this network, which achieved an HTER of 2.41%. For the warped attack simulations, we achieved an APCER rate of 0.00%, which means the network detected all warped photo attack simulations.

Table 6.4: Performance results (in %) for the CASIA dataset considering the presentation attacks simulations individually.

Attack Type	APCER	BPCER	HTER	EER
Warped photo	**0.00**	**3.33**	**1.67**	**2.22**
Cut photo	1.11	3.33	2.22	3.33
Video	3.33	3.33	3.33	3.33
Overall test	1.48	3.33	2.41	3.33



(a) Albedo maps (b) Depth maps

(c) Reflectance maps (d) Concatenated maps

Figure 6.4: Results obtained on CASIA dataset for the different attack types and for the different maps obtained with the shape-from-shading algorithm.

UVAD Dataset

In this section, we evaluate the proposed method in a challenging scenario with presentation attacks and bona fine presentations, both captured with different sensors, which is named in the literature as a cross-sensor scenario [265]. Table 6.5 shows the obtained results for the different maps, which indicates that network trained using the depth maps is the most discriminative network for detecting the presentation attack in this dataset. Although the HTER of 14.51% obtained in this dataset is higher than previous datasets, this result is the lowest achieved in the literature as shown in Section 6.4.8.

Table 6.5: Performance results (in %) for the UVAD dataset considering the different maps obtained with the shape-from-shading algorithm.

Map Type	APCER	BPCER	HTER
Albedo	24.58	20.00	22.29
Depth	**10.68**	**18.33**	**14.51**
Reflectance	22.07	31.67	26.87
Concatenated Maps	12.40	21.67	17.03

6.4.7 Inter-dataset Evaluation Protocol

Here, we present the obtained results for the inter-dataset evaluation protocol, which is the most challenging evaluation protocol nowadays. The difficulty of this evaluation protocol raises up from the fact that we have a training and testing scenarios quite different in terms of acquisition sensors, light conditions, and environment (e.g, different background).

Table 6.6 shows the obtained results of the proposed method trained with the CASIA dataset and tested upon the other ones. Surprisingly, the proposed method achieved an outstanding performance result for the Replay-attack dataset when we consider only the video-based attempted attack videos for training our CNN network, with an HTER of 9.75%. On the other side, our method achieved an APCER, BPCER, and HTER of 34.81%, 24.44%, and 29.63%, respectively, by using the Replay-Attack dataset for training and the CASIA dataset for testing and considering the reflectance maps. Finally, considering the UVAD dataset for training and the CASIA dataset for testing, the proposed method achieved an APCER, BPCER, and HTER values of 66.67%, 12.22%, and 39.44%, respectively, using the depth maps.

Table 6.6: Results (in %) obtained with the cross-dataset protocol considering both presentation attacks simulations individually and the overall test sets of each dataset.

Training Set	**Testing Sets**					
CASIA	**Replay-Attack**			**UVAD**		
	APCER	**BPCER**	**HTER**	**APCER**	**BPCER**	**HTER**
Video	10.75	8.75	**9.75**	57.93	21.67	39.80
Overall	8.25	51.25	29.75	34.79	36.67	35.73
Warped	65.75	27.50	46.62	36.75	30.00	**33.38**
Cut	92.00	2.50	47.25	58.07	23.33	40.70
	Concatenated Maps			**Depth Maps**		

6.4.8 Comparison with State-of-the-Art Methods

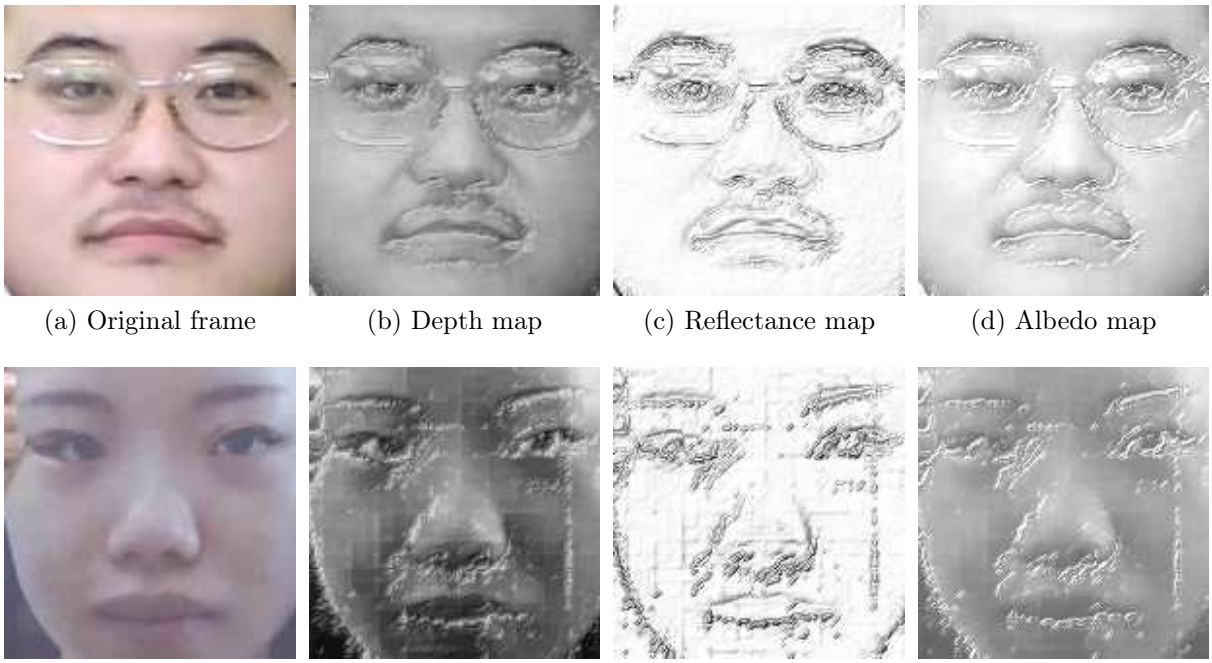
In this section, we compare the proposed method with other method available in the literature. We select the most effective CNN networks designed for the PAD problem, include the networks specifically designed to estimate depth maps from the RGB images without using any kind of extra device [11]. We notice that most effective methods that take into account the intra-dataset evaluation protocol achieved poor performance results in the inter-dataset protocol, as shown in Table 6.7. The proposed method achieved the lowest HTER for the inter-dataset protocol and competitive results for the intra-dataset evaluation protocol, which demonstrates the potential of the proposed method. Considering the complexity of the existing networks, i.e. the number of convolutional layers, the proposed CNN architecture provides a reasonable trade-off between performance and hardware requirement, which can be directly translated into memory consumption and training timing of the network.

Table 6.7: Comparison among the existing CNN-based methods considering the intra- and inter-based evaluation protocols for the datasets considered in this work.

Methods	**Intra-Dataset Protocol**				**Inter-Dataset Protocol**		
	Replay-Attack	**CASIA**		**UVAD**	**Replay-Attack**	**CASIA**	**UVAD**
	HTER	**EER**	**HTER**	**HTER**	**HTER**	**HTER**	**HTER**
Li et al. [142] (Fine-tuned VGG-Face)	4.30	5.20	–	–	–	–	–
Li et al. [142] (DPCNN)	6.10	4.50	–	–	–	–	–
Atoum et al. [11] (Patches and Depth-Based CNNs)	0.72	2.67	2.27	–	–	–	–
Menotti et al. [162] (Architecture Optimization)	0.75	–	–	–	–	–	–
Li et al. [143] (Hybrid CNNs)	1.60	**2.20**	–	–	–	–	–
Pinto et al. [192] (Fine-tuned VGG network)	**0.00**	–	6.67	–	49.72	47.16	–
Yang et al. [269] (Fine-tuned AlexNet)	2.68	–	6.25		41.36	42.04	–
Patel et al. [180] (GoogLeNet + Eye-Blink Detection)	0.50	–	–	12.40	31.60	–	
Proposed Method	3.12	3.33	2.41	14.51	**9.75**	**29.63**	33.38

6.4.9 Visual Assessment

In this section, we show a visual assessment of the albedo, reflectance, and depth maps generated by the Tsai's algorithm, the shape-from-shading algorithm used in this work. Fig. 6.5 depicts these maps computed from a bona fide presentation and from a presentation attack video frame. These examples illustrate how the artifacts affect the reconstruction of the surface, specially, in this example, of the depth and reflectance maps. We believe that the way how the Tsai's algorithm computes the depth might improve the highlighting of the artifacts present in the presentation attack images. As mentioned in the Section 6.3.1, this method performs the estimation of the depth locally, which means that each point (x, y) is reconstructed interdependently. When the algorithm tries to compute the first and second order derivative of outliers (e.g., noise or printing artifact), we came up with a situation where the approximation could not be applied, which produces the white spots in the reconstructed maps. Moreover, the first and second order derivative computations can potentially highlight the printing artifacts, i.e., horizontal and vertical lines. Fig. 6.4.9 shows the details of the reconstructed surface considering a video frame of both classes of the PAD problem, in which we evidence natural pattern for the genuine access (e.g., skin roughness) and synthetic patterns for presentation attack image (e.g., horizontal and vertical lines).



(a) Original frame (b) Depth map (c) Reflectance map (d) Albedo map

(e) Original frame (f) Depth map (g) Reflectance map (h) Albedo map

Figure 6.5: Example of a bona fide presentation video frame (first line) a presentation attack video frame (second line). First column illustrate original frames captured by the acquisition sensor, whereas the other columns show their respective maps.

(a) Reconstructed surface of the nose region of a bona fide presentation

(b) Reconstructed surface of the nose region of a presentation attack

Figure 6.6: Details of the reconstructed surface for the video frames showed in Fig. 6.5 from a genuine access (a) and an attempted attack (b), in which we found strong evidence of a natural (skin roughness) texture pattern and of a synthetic (horizontal and vertical lines) texture pattern for these respective classes.

6.5 Conclusion and Future Work

In this paper, we proposed an algorithm for detecting presentation attacks based on intrinsic properties of the scene such as albedo, reflectance, and depth of the scene. We showed that these properties are useful for detecting different types of presentation attacks with satisfactory results in terms of error rates. We also proposed a novel CNN network specially designed for learning features from these different maps. The ability of CNN networks in learning from data was crucial for our method to achieve the reported results, since the hand-crafting feature engineering of these different maps could take much work effort.

The experimental results corroborated the effectiveness of our CNN networks trained using these different maps. Particularly, the network trained with the depth maps and with the concatenated maps presented a more robustness for detecting presentation attacks taken into account the inter-dataset evaluation protocol. For the intra-dataset evaluation protocol, the depth map achieved the best performance results for the UVAD and Replay-Attack datasets, whereas the concatenated maps achieved the best performance results for the CASIA dataset. We believe there could be some complementarity between these maps, which would allow our CNN network to learn good features and deal with this complex dataset that contains several kinds of photo and video presentation attacks.

Unquestionably, the inter-dataset evaluation protocol was the hardest scenario for the proposed method, even considering the cross-sensor scenario, in which we achieved better results than the state-of-the-art, as confirmed through the results obtained for the UVAD dataset. We believe our work could help the community to have a better understanding about this challenging problem, since the proposed method was able to spot strong evidences of presentation attacks considering the photographs- and video-based attempted attacks in the reconstructed surface of the faces.

Future research efforts include the investigation of alternative approaches to combining the albedo, reflectance, and depth maps toward extracting complementary patterns. This is useful for detecting presentation attacks, as well as the investigation of new approaches to recovering the surface properties from shading by taking into account a reflectance model more suitable for our problem, such as Bidirectional reflectance distribution function (BRDF). The study of methods for finding the light source detection that operate in

a real scenario (not with synthetic images) could also be a promising investigation path toward improving the facial surface reconstruction.

Chapter 7

Conclusions and Future Work

"Take delight in the Lord, and he will give you the desires of your heart."
—Bible, *Psalm 37:4*

"In questions of science, the authority of a thousand is not worth the humble reasoning of a single individual."
—Galileo Galilei, *Italian astronomer, mathematician, physicist, philosopher and professor (1564–1642)*

"Science knows only one commandment – contribute to science."
—Bertolt Brecht, *German theatre practitioner, playwright, and poet (1898–1956)*

7.1 Final Remarks

IN this thesis, we proposed a set of algorithms and methodologies for detecting presentation attacks in biometric systems based on face, iris, and fingerprint traits. The hypotheses presented in **Chapter 1** were investigated throughout the other chapters, where we also showed the specific contributions achieved in each work and a discussion regarding the visual assessments of the main approaches presented in this thesis.

The results presented in **Chapters 2 and 3** corroborate the **Hypothesis 1** in which we showed strong evidences of how artifacts affect frequency components of the Fourier spectrum. The Fourier analysis was a powerful tool to recognize strong evidences of these artifacts and our proposed method for collecting such evidences presented results superior to or competitive with traditional image analysis approaches, such as texture and motion analyses. Moreover, the modeling of artifacts collected over time described in **Chapter 3** demonstrated to be a more effective approach than the visual rhythm-based techniques, presented in **Chapter 2**.

Regarding the **Hypothesis 2**, we showed that supervised feature learning techniques based on Convolutional Neural Networks (CNN) were able to learn meaningful telltales

of presentation attacks for different modalities and attack types. In **Chapter 4**, we presented performance results of two methodologies for building convolutional networks, architecture optimization and filter optimization. The architecture optimization methodology was used to investigate a set of shallow CNN architectures, i.e., number of convolutional layers ranging from one to three, whose weights were randomly defined through a uniform distribution. On the other hand, the filter optimization was used to investigate the behavior of these shallow CNNs when operating with filter weights defined through the back-propagation mechanism. The results achieved in this work suggest us that the interplay between these two approaches is a promising strategy for deploying CNN architecture for the PAD problem, which led us to build a promising convolutional network, the *SpoofNet*, which has inspired several other studies in the literature. In **Chapter 5**, we designed a methodology to use deep neural networks to distinguish presentation attacks from bona fide presentations, in different modalities. Our methodology empowered the networks to deal with the small datasets available to our problem and learn useful features for detecting attempted attacks in the intra-dataset scenario. Additionally, we showed, in **Chapter 5**, the limitations of deep architectures for operating in the inter-dataset scenario. The weakness of deep neural networks for modeling the phenomenon related to the problem from the original image space sharpened our understanding of how to build flexible convolutional neural networks taking into account other properties inferred from the original data.

The findings achieved in **Chapters 2 and 3**, in which we presented two feature engineering solutions (named "hand-crafted" solutions), and in **Chapters 4 and 5**, in which we presented different approaches based on representation learning (named "data-driven" solutions), have turned our attention for possibilities of how to harmoniously mix these two approaches in an attempt of taking the best of both approaches.

Lastly, in **Chapter 6**, we presented a novel approach to face PAD based on intrinsic properties of face surfaces reconstructed via shape-from-shading technique (i.e., albedo, reflectance and depth information). Furthermore, we designed a novel shallow CNN architecture suitable for learning meaningful telltales of presentation attacks from the estimations of these properties. This solution presented outstanding results considering the cross-sensor and inter-dataset scenarios, corroborating the **Hypothesis 4**. In this work, we presented a new perspective to deal with the problem by looking more closely into the surface properties of the skin surface, for the bona fide presentations, and synthetic materials such as printed photographs and device screens, for the presentation attacks. The results achieved in this work corroborate the **Hypothesis 3**, which give us new perspectives of dealing with the PAD problem since this proposed method reaches the best performance results even considering challenging evaluation protocol and scenarios.

7.2 Directions for Future Work

We believe that the guidelines presented in this thesis have contributed to research community by providing a better understanding of how artifacts affect biometric samples and by proposing new methods for capturing patterns and nuances related to these artifacts.

Directions for future research include investigations of methods that better quantify different artifacts added during the manufacture and recapture of synthetic samples. The design of methods capable of quantifying the artifacts separately can also yield improvements to detect attempted attacks, since such artifacts may change in magnitude over time, mainly due to the interaction between the acquisition sensor and the presentation attack instruments.

During our investigations, we noticed that we might have cases where artifacts can potentially cancel each other. For instance, the blurring effect caused by sensor defocus can decrease the amount of flicking or banding effects. Therefore, we believe that the design of algorithms to quantify these artifacts individually and over time has much potential to achieve good results, even in cross-sensor and data-set scenarios.

On the other hand, the proposed method based on the intrinsic properties of facial surfaces can be improved by using modeling that better describes facial skin spectrophotometry, such as the bidirectional reflectance distribution function [57]. In addition, the investigation of methods capable of accurately estimating the light source direction can also improve the detection rates in the cross-sensor and inter-dataset scenarios in which we have a test environment (e.g., illumination conditions) different from the training environment.

Finally, we believe that the combination of methods and algorithms introduced in this thesis, toward exploring their possible complementarity, is a promising investigation line to build robust detectors. In this direction, the visual rhythm introduced in **Chapter 2** could be used to feed the CNN networks proposed in **Chapters 4, 5, and 6** toward building robust PAD solutions. Another promising investigation path to improve the algorithms and methodologies presented in this thesis is the adoption of patch-based analysis [11], mainly CNN-based methods that usually require a greater amount of data during the training phase.

7.3 Other Applications to Algorithms Presented in this Thesis

In this section, we present some possible applications in digital forensics that could take advantage of the algorithms presented in this thesis along with techniques developed for solving other problem that could be explored toward improving our algorithms.

7.3.1 Detection of (Illegal) Copyrighted Video Recapture

The recapture detection problem of digital media in general can also benefit from algorithms and techniques developed in this research. This is a topic of interest of the video forensics research field, which is growing active, since recapture is often an indicator of tampering activities [163, 215].

Visentini-Scarzanella et al. [256] proposed a method to detect the recapture of videos based on deformations of the objects in recaptured images. The authors used lens radial distortion models [7] to extract features that are used in a threshold-based classification

scheme. Bestagini et al. [23] proposed a method to detect recaptured images, based on the analysis of ghosting artifacts, which are added to the images during the recapture process. The authors proposed a ghosting artifact detector based on the analysis of peaks at high-frequencies, in the frequency domain. The authors evaluated the proposed approach in a dataset composed of 18 videos (nine originals and nine recaptured), in which the authors reported an accuracy of 91%.

In recent works, Thongkamwitoon et al. [247, 248] proposed an approach to detect recaptured images based on learned edge blurriness using the K-singular value decomposition (K-SVD) approach [1] to learn dictionaries using edge-based features, besides presenting an anti-forensic technique to the our previous work [193], which was designed to detect video-based face spoofing attacks based on cues such as Moiré patterns and aliasing artifacts.

In conclusion, the methods for recapture detection hereby presented aim at modeling artifacts added to the signal during the recapture process. This line of investigation matches with our works presented in **Chapters 2 and 3**, and further investigations regarding the applicability of algorithms and techniques presented in this thesis could be performed, aiming at possible contributions in this research field as well.

7.3.2 Image Tampering Detection

The image manipulation detection problem is another topic in digital forensic research field that can also benefit from the developed algorithm presented in **Chapter 6**, or vice-versa. The analysis of intrinsic properties of light is a great source of inspiration for several algorithms and techniques for detecting image tampering, especially, splicing and copy-move manipulation [54, 212, 213]. Riess et al. [212] presented a method for detecting image manipulation, namely *intrinsic contour estimation*, that use the reflectance map computed by the algorithms proposed in [78, 231] to normalize the images under analysis and so exposing the manipulated candidate regions. Carvalho et al. [54] and Riess [213] proposed algorithms for detecting image tampering based on intrinsic properties of light present in a digital photography under analysis considering the dichromatic reflectance model [230], which are more suitable for computing intrinsic properties of human skin and, therefore, the algorithm presented in **Chapter 6** could be improved by using this reflectance model.

7.4 Publications During this Doctoral Research

The results obtained during the doctoral period were published in important scientific communication vehicles dedicated to academic and scientific communities related to digital forensics, biometrics and image analysis. We divided the publications into related and non-related with the thesis subject:

Publications Related with the Thesis Subject

1. *Using visual rhythms for detecting video-based facial spoof attacks*

We presented this work in **Chapter 2**, which was published in a **peer-reviewed Journal**, with an **impact factor of 5.824**;

Reference

Pinto, A., W. Robson Schwartz, H. Pedrini, and A. de Rezende Rocha. Using visual rhythms for detecting video-based facial spoof attacks. *IEEE Transactions on Information Forensics and Security (T-IFS)*, 10(5):1025–1038, May 2015

2. *Face Spoofing Detection Through Visual Codebooks of Spectral Temporal Cubes*
We discussed about this work in **Chapter 3** and it was also published in a **peer-reviewed Journal**, with an **impact factor of 5.071**;

Reference

Pinto, A., H. Pedrini, W. Robson Schwartz, and A. Rocha. Face spoofing detection through visual codebooks of spectral temporal cubes. *IEEE Transactions on Image Processing (T-IP)*, 24(12):4726–4740, Dec 2015

3. *Deep Representations for Iris, Face, and Fingerprint Spoofing Detection*
We presented this work in **Chapter 4** and it was published in a **peer-reviewed Journal**, with an **impact factor of 5.824**. Especially, my personal contributions for this work include: (1) implementation and execution of experiments involving face and fingerprint modalities; (2) contributions with ideas toward improving the performance results of the method during the experiments, based on expertise acquired in our previous works; and (3) article writing and execution of extra experiments during the revisions;

Reference

D. Menotti, G. Chiachia, **Pinto, A.**, W. Robson Schwartz, H. Pedrini, A. Xavier Falcao, and A. Rocha. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security (T-IFS)*, 10(4):864–879, April 2015

4. *Counteracting Presentation Attacks in Face Fingerprint and Iris Recognition*
We presented this work in **Chapter 5**, which was published in as a **Book Chapter**. Especially, my personal contributions for this work include : (1) ideas in attempt of improving the performance results of the method during the experiments, based on expertise acquired in our previous works; (2) article writing and text revision. The experiments were performed by the undergraduate students Michael Krumdick and Benedict Becker. This work was developed during my one-year doctoral internship at the University of Notre Dame, USA;

Reference

Allan Pinto, Helio Pedrini, Michael Krumdick, Benedict Becker, Adam Czajka, Kevin W. Bowyer, and Anderson Rocha. *Deep Learning in Biometrics*, chapter Counteracting Presentation Attacks in Face Fingerprint and Iris Recognition, page 49. CRC Press, 2018

5. *Leveraging Shape, Reflectance and Albedo from Shading for Face Presentation Attack Detection*
   We presented this work in **Chapter 6** and it was **submitted in a peer-reviewed Journal**, with an **impact factor of 7.982**;

   Reference

   Allan Pinto, Siome Goldenstein, Alexandre Ferreira, Tiago Carvalho, Helio Pedrini, and Anderson Rocha. Leveraging shape, reflectance and albedo from shading for face presentation attack detection. *IEEE Transactions on Neural Networks and Learning Systems (T-NNLS)*, pages 1–11, 2018 (submitted)

6. *Ensemble of Multi-View Learning Classifiers for Cross-Domain Iris Presentation Attack Detection*
   This work does not appear in this thesis, but was partially inspired by the work discussed in **Chapter 4** and it was published in a **peer-reviewed Journal**, with an **impact factor of 5.824**. Especially, my personal contributions for this work include: (1) implementation, execution of experiments, and suggestions of experimental protocols involving the meta-classification step; (2) Adjusts of SpoofNet architecture (**Chapter 4**) to learn artifacts related to iris presentation attack detection problem from Binarized statistical image feature (BSIF) maps [114], instead of raw data; (3) contributions with ideas toward improving the performance results of the method during the experiments, based on expertise acquired in our previous works; and (4) writing and revision of the article. This work was partially developed during my one-year doctoral internship at the University of Notre Dame, USA;

   Reference

   A. Kuehlkamp, **A. Pinto**, A. Rocha, K. W. Bowyer, and A. Czajka. Ensemble of multi-view learning classifiers for cross-domain iris presentation attack detection. *IEEE Transactions on Information Forensics and Security (T-IFS)*, pages 1–13, 2018 (To appear)

Publications Non-Related with the Thesis Subject

7. *Provenance filtering for multimedia phylogeny*
   This work does not appear in this thesis and it was published in a **peer-reviewed Intl. Conference**. This work was developed during my one-year doctoral internship at the University of Notre Dame, USA;

Reference

A. Pinto, D. Moreira, A. Bharati, J. Brogan, K. Bowyer, P. Flynn, W. Scheirer, and A. Rocha. Provenance filtering for multimedia phylogeny. In *IEEE International Conference on Image Processing (ICIP)*, pages 1502–1506, Sept 2017

8. *Spotting the difference: Context retrieval and analysis for improved forgery detection and localization*
   This work does not appear in this thesis and it was published in a **peer-reviewed Intl. Conference**. This work was developed during my one-year doctoral internship at the University of Notre Dame, USA;

Reference

J. Brogan, P. Bestagini, A. Bharati, **A. Pinto**, D. Moreira, K. Bowyer, P. Flynn, A. Rocha, and W. Scheirer. Spotting the difference: Context retrieval and analysis for improved forgery detection and localization. In *IEEE International Conference on Image Processing (ICIP)*, pages 4078–4082, Sept 2017

9. *U-Phylogeny: Undirected provenance graph construction in the wild*
   This work does not appear in this thesis and it was published in a **peer-reviewed Intl. Conference**. This work was developed during my one-year doctoral internship at the University of Notre Dame, USA;

Reference

A. Bharati, D. Moreira, **A. Pinto**, J. Brogan, K. Bowyer, P. Flynn, W. Scheirer, and A. Rocha. U-phylogeny: Undirected provenance graph construction in the wild. In *IEEE International Conference on Image Processing (ICIP)*, pages 1517–1521, Sept 2017

10. *Image Provenance Analysis at Scale*
    This work does not appear in this thesis and it was published in a **peer-reviewed Journal**, with an **impact factor of 5.071**. This work was partially developed during my one-year doctoral internship at the University of Notre Dame, USA;

Reference

D. Moreira, A. Bharati, J. Brogan, **A. Pinto**, M. Parowski, K. W. Bowyer, P. J. Flynn, A. Rocha, and W. J. Scheirer. Image provenance analysis at scale. *IEEE Transactions on Image Processing (T-IP)*, 27(12):6109–6123, Dec 2018

Publications by Type of the Scientific Communication Vehicle

7.5 Source Code Available Along with this Thesis

The dataset and source codes of the algorithms and methodologies presented in this thesis are freely and publicly available on GitHub [1] [2] and Reasoning for Complex Data (RECOD) laboratory [3] websites. Please do not hesitate to contact me (allansp84@gmail.com) if you have any trouble or questions about using these codes.

Table 7.1: Source code developed during this thesis and freely available for reproducibility purposes.

Technique	**Source**
UVAD dataset (**Chapter 2**)	https://tinyurl.com/ho3ozhx
Visual Rhythms (**Chapter 2**)	https://github.com/allansp84/visualrhythm-antispoofing
Spectral Cubes (**Chapter 3**)	https://github.com/allansp84/spectralcubes
SpoofNet (**Chapter 4**)	https://github.com/allansp84/simple-hp

† These links were visited on November 6th, 2018.

[1] http://repo.recod.ic.unicamp.br/public/projects
[2] https://github.com/allansp84
[3] http://recod.ic.unicamp.br

Bibliography

[1] M. Aharon, M. Elad, and A. Bruckstein. k -svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing (TSP)*, 54(11):4311–4322, November 2006.

[2] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481, 2004.

[3] Zahid Akhtar, Giorgio Fumera, Gian Luca Marcialis, and Fabio Roli. Robustness evaluation of biometric systems under spoof attacks. In *International Conference on Image Analysis and Processing*, pages 159–168, 2011.

[4] Zahid Akhtar, Giorgio Fumera, Gian Luca Marcialis, and Fabio Roli. Evaluation of serial and parallel multibiometric systems under spoofing attacks. In *IEEE International Conference on Biometrics: Theory Applications and Systems*, pages 283 –288, September 2012.

[5] Nizar Almoussa. Variational retinex and shadow removal. Technical report, University of California, Department of Mathematics, 2009.

[6] Jonathan Alon, V. Athitsos, Quan Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1685–1699, Sept 2009.

[7] L. Alvarez, L. Gómez, and J.R. Sendra. An algebraic approach to lens distortion by line rectification. *Journal of Mathematical Imaging and Vision (JMIV)*, 35(1):36–50, September 2009.

[8] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: A public database and a baseline. In *IEEE International Joint Conference on Biometrics*, pages 1–7, October 2011.

[9] André Anjos, Laurent El-Shafey, Roy Wallace, Manuel Günther, Christopher McCool, and Sébastien Marcel. Bob: A free signal processing and machine learning toolbox for researchers. In *ACM Conference on Multimedia Systems*, pages 1449–1452, New York, NY, USA, 2012. ACM.

[10] C. Arthur. iPhone 5S fingerprint sensor hacked by germany's chaos computer club. http://tinyurl.com/pkz59rg, September 2013. Accessed: 2016/02/20.

[11] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu. Face anti-spoofing using patch and depth-based cnns. In *IEEE International Joint Conference on Biometrics*, pages 319–328, October 2017.

[12] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A Araujo. Bossa: Extended bow formalism for image classification. In *IEEE International Conference on Image Processing*, pages 2909–2912, September 2011.

[13] Wei Bao, Hong Li, Nan Li, and Wei Jiang. A liveness detection method for face

recognition based on optical flow field. In *International Conference on Image Analysis and Signal Processing*, pages 233–236, April 2009.

[14] Andy Beach. *Real world video compression.* Peachpit Press, Berkeley, CA, USA, first edition, 2008.

[15] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2546–2554, 2011.

[16] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research (JMLR)*, 13:281–305, 2012.

[17] J. Bergstra, D. Tamins, and N. Pinto. Hyperparameter optimization for convolutional vision architecture, 2013.

[18] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

[19] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU math expression compiler. In *The Python for Scientific Computing Conference (SciPy)*, 2010.

[20] James Bergstra, Dan Yamins, and David D. Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *International Conference on Machine Learning*, 2013.

[21] James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 2546–2554. Curran Associates, Inc., 2011.

[22] P. Bestagini, M. Visentini-Scarzanella, M. Tagliasacchi, P. L. Dragotti, and S. Tubaro. Video recapture detection based on ghosting artifact analysis. In *IEEE International Conference on Image Processing*, pages 4457–4461, September 2013.

[23] P. Bestagini, M. Visentini-Scarzanella, M. Tagliasacchi, P. L. Dragotti, and S. Tubaro. Video recapture detection based on ghosting artifact analysis. In *IEEE International Conference on Image Processing*, pages 4457–4461, September 2013.

[24] S. Bharadwaj, T.I. Dhamecha, M. Vatsa, and R. Singh. Computationally efficient face spoofing detection with motion magnification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 105–110, June 2013.

[25] B. Biggio, Z. Akhtar, G. Fumera, G.L. Marcialis, and F. Roli. Security evaluation of biometric authentication systems under real spoofing attacks. *IET Biometrics*, 1(1):11 –24, March 2012.

[26] B. Biggio, Z. Akthar, G. Fumera, G.L. Marcialis, and F. Roli. Robustness of multimodal biometric verification systems under realistic spoofing attacks. In *IEEE International Conference on Biometrics: Theory Applications and Systems*, pages 1–6, October 2011.

[27] Z. Boulkenafet, J. Komulainen, Xiaoyi Feng, and A. Hadid. Scale space texture analysis for face anti-spoofing. In *IAPR International Conference on Biometrics*, pages 1–6, June 2016.

[28] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face anti-spoofing based on color texture analysis. In *IEEE International Conference on Image Processing*, pages 2636–2640, September 2015.

[29] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, August 2016.

[30] Zinelabidine Boulkenafet, Jukka Komulainen, Zahid Akhtar, and Abdenour Hadid et al. A competition on generalized software-based face presentation attack detection in mobile scenarios. In *IEEE International Joint Conference on Biometrics*, pages 688–696, October 2017.

[31] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2559–2566, June 2010.

[32] K.W. Bowyer and J.S. Doyle. Cosmetic contact lenses and iris recognition spoofing. *Computer*, 47(5):96–98, 2014.

[33] I. R. Buhan and P. H. Hartel. The state of the art in abuse of biometrics. Technical Report TR-CTIT-05-41, Centre for Telematics and Information Technology University of Twente, Enschede, September 2005.

[34] P. Campisi. *Security and Privacy in Biometrics*, chapter Security and Privacy in Biometrics: Towards a Holistic Approach, pages 1–23. Springer, London, 2013.

[35] E.H. Carazzai. Paranaguá Harbor employees used silicone fingers to circumnvent biometric system in Paraná. http://tinyurl.com/hkoj2jg, February 2014. Accessed: 2016/02/20.

[36] M.M. Chakka, A. Anjos, S. Marcel, Tronci, and et al. Competition on Counter Measures to 2-D Facial Spoofing Attacks. In *IEEE International Joint Conference on Biometrics*, pages 1–6, 2011.

[37] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[38] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[39] Mo Chen, Jessica Fridrich, Jan Lukáš, and Miroslav Goljan. Imaging sensor noise as digital x-ray for revealing forgeries. In Teddy Furon, François Cayre, Gwenaël Doërr, and Patrick Bas, editors, *Information Hiding*, volume 4567 of *Lecture Notes in Computer Science*, pages 342–358. Springer Berlin Heidelberg, 2007.

[40] Giovani Chiachia. Hyperparameter Optimization made Simple, 2014.

[41] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *International Conference of the Biometrics Special Interest Group*, pages 1 –7, September 2012.

[42] I. Chingovska, J. Yang, Z. Lei, and D. et al. Yi. The 2nd Competition on Counter Measures to 2D Face Spoofing Attacks. In *IAPR International Conference on Biometrics*, pages 1–6, June 2013.

[43] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE International Conference on Computer Vision and Pattern Recognition*,

July 2017.

[44] Seong Soo Chun, Hyeokman Kim, Kim Jung-Rim, Sangwook Oh, and Sanghoon Sull. Fast text caption localization on video using visual rhythm. In Shi-Kuo Chang, Zen Chen, and Suh-Yin Lee, editors, *Recent Advances in Visual Information Systems*, pages 259–268, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[45] M.-G. Chung, J. Lee, H. Kim, S. M.-H. Song, and W.-M. Kim. Automatic Video Segmentation based on Spatio-Temporal Features. *Korea Telecom (KT)*, 1(4):4–14, 1999.

[46] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural Networks*, 20(1), 2012.

[47] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber. Deep big simple neural nets for handwritten digit recognition. *Neural Computation*, 22(12):3207–3220, 2010.

[48] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics*, 1(1):7–24, January 1982.

[49] Catherine C. Cooksey, David W. Allen, and Benjamin K. Tsai. Reference data set of human skin reflectance. *Journal of Research of the National Institute of Standards and Technology*, 122:1–5, 2017.

[50] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[51] A. Czajka. Database of iris printouts and its application: Development of liveness detection method for iris recognition. In *Int. Conference on Methods and Models in Automation and Robotics (ICMMAR)*, pages 28–33, August 2013.

[52] A. Czajka. Pupil dynamics for iris liveness detection. *IEEE Transactions on Information Forensics and Security*, 10(4):726–735, April 2015.

[53] Adam Czajka. Iris liveness detection by modeling dynamic pupil features. In Kevin W. Bowyer and Mark J. Burge, editors, *Handbook of Iris Recognition*, pages 439–467. Springer London, London, 2016.

[54] T. J. d. Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. d. R. Rocha. Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security*, 8(7):1182–1194, July 2013.

[55] Congxia Dai, Yunfei Zheng, and Xin Li. Pedestrian detection and tracking in infrared imagery using shape and appearance. *Computer Vision and Image Understanding*, 106(2):288–299, 2007. Special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum.

[56] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, June 2005.

[57] Kristin J. Dana, Bram van Ginneken, Shree K. Nayar, and Jan J. Koenderink. Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics*, 18(1):1–34, January 1999.

[58] J. Daugman. Iris recognition and anti-spoofing countermeasures. In *International Biometrics Conference*, 2004.

[59] J.G. Daugman. High confidence visual recognition of persons by a test of statistical

independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1148–1161, 1993.

[60] J.G. Daugman. Recognizing persons by their iris patterns. In Anil Jain, Ruud Bolle, and Sharath Pankanti, editors, *Biometrics: Personal Identification in Networked Society*, pages 103–121. Kluwer Academic Publishers, 1999.

[61] John G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, 2(7):1160–1169, Jul 1985.

[62] M. De Marsico, M. Nappi, D. Riccio, and J. Dugelay. Moving face spoofing detection via 3d projective invariants. In *IAPR International Conference on Biometrics*, pages 73 –78, April 2012.

[63] J. Doyle and K.W. Bowyer. Notre dame image dataset for contact lens detection in iris recognition, 2014. last access on June 2014.

[64] J.S. Doyle, K.W. Bowyer, and P.J. Flynn. Variation in accuracy of textured contact lens detection based on sensor and lens pattern. In *IEEE International Conference on Biometrics: Theory Applications and Systems*, pages 1–7, 2013.

[65] N. Erdogmus and S. Marcel. Spoofing 2D face recognition systems with 3D masks. In *International Conference of the Biometrics Special Interest Group*, pages 1–8, 2013.

[66] N. Erdogmus and S. Marcel. Spoofing in 2D face recognition with 3D masks and anti-spoofing with kinect. In *IEEE International Conference on Biometrics: Theory Applications and Systems*, pages 1–6, September 2013.

[67] N. Erdogmus and S. Marcel. Spoofing face recognition with 3d masks. *IEEE Transactions on Information Forensics and Security*, 9(7):1084–1097, July 2014.

[68] J. Fierrez-Aguilar, J. Ortega-garcia, D. Torre-toledano, and J. Gonzalez-rodriguez. Biosec baseline corpus: A multimodal biometric database. *Pattern Recognition*, 40:1389–1392, 2007.

[69] Tiago Freitas Pereira, Jukka Komulainen, Andre Anjos, Jose De Martino, Abdenour Hadid, Matti Pietikainen, and Sebastien Marcel. Face liveness detection using dynamic texture. *EURASIP Journal on Image and Video Processing*, 2014(1):2, 2014.

[70] Michael P. Fuller and Peter R. Griffiths. Diffuse reflectance measurements by infrared fourier transform spectrometry. *Analytical Chemistry*, 50(13):1906–1910, 1978.

[71] J. Galbally, F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia. Fingerprint liveness detection based on quality measures. In *Int. Conference on Biometrics, Identity and Security (BIdS)*, pages 1–8, September 2009.

[72] J. Galbally, F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia. A high performance fingerprint liveness detection method based on quality related features. *Future Generation Computer Systems (FGCS)*, 28(1):311–321, 2012.

[73] J. Galbally, S. Marcel, and J. Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014.

[74] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Transactions*

on Image Processing, 23(2):710–724, February 2014.

[75] J. Galbally, J. Ortiz-Lopez, J. Fierrez, and J. Ortega-Garcia. Iris liveness detection based on quality related features. In *IAPR International Conference on Biometrics*, pages 271–276, 2012.

[76] Javier Galbally, Julian Fierrez, and Javier Ortega-garcia. Vulnerabilities in biometric systems: Attacks and recent advances in liveness detection. *Database*, 1(3):1–8, 2007. available at http://atvs.ii.uam.es/files/2007_SWB_VulnerabilitiesRecentAdvances_Galbally.pdf.

[77] D.C. Garcia and R.L. de Queiroz. Face-spoofing 2D-detection based on moiré-pattern analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):778–786, April 2015.

[78] Peter Vincent Gehler, Carsten Rother, Martin Kiefel, Lumin Zhang, and Bernhard Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 765–773, USA, 2011. Curran Associates Inc.

[79] Wilson S. Geisler and Duane G. Albrecht. Cortical neurons: Isolation of contrast gain control. *Vision Research*, 32(8):1409–1410, 1992.

[80] Jan C. Gemert, Jan-Mark Geusebroek, Cor J. Veenman, and Arnold W. Smeulders. Kernel codebooks for scene categorization. In *European Conference on Computer Vision*, pages 696–709. Springer-Verlag, 2008.

[81] L. Ghiani, A. Hadid, G.L. Marcialis, and F. Roli. Fingerprint liveness detection using binarized statistical image features. In *IEEE International Conference on Biometrics: Theory Applications and Systems*, pages 1–6, September 2013.

[82] L. Ghiani, G.L. Marcialis, and F. Roli. Fingerprint liveness detection by local phase quantization. In *International Conference on Pattern Recognition*, pages 537–540, November 2012.

[83] L. Ghiani, D. Yambay, V. Mura, S. Tocco, G.L. Marcialis, F. Roli, and S. Schuckcrs. LivDet 2013 – fingerprint liveness detection competition. In *IAPR International Conference on Biometrics*, pages 1–6, 2013.

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

[85] C. Gottschlich. Convolution comparison pattern: An efficient local image descriptor for fingerprint liveness detection. *PLoS ONE*, 11(2):12, February 2016.

[86] D. Gragnaniello, G. Poggi, C. Sansone, and L. Verdoliva. Fingerprint liveness detection based on weber local image descriptor. In *IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, pages 46–50, September 2013.

[87] D. Gragnaniello, G. Poggi, C. Sansone, and L. Verdoliva. An investigation of local descriptors for biometric spoofing detection. *IEEE Transactions on Information Forensics and Security*, 10(4):849–863, April 2015.

[88] D. Gragnaniello, G. Poggi, C. Sansone, and L. Verdoliva. Local contrast phase descriptor for fingerprint liveness detection. *Pattern Recognition*, 48(4):1050–1058, 2015.

[89] S. J. F. Guimaraes, M. Couprie, N. J. Leite, and A. A. Araujo. A Method for Cut Detection Based on Visual Rhythm. In *Conference on Graphics, Patterns and Images*, pages 297–304, 2001.

[90] Manuel Günther, Dennis Haufe, and Rolf P. Würtz. Face recognition with disparity corrected gabor phase differences. In *International Conference on Artificial Neural Networks and Machine Learning*, pages 411–418, 2012.

[91] P. Gupta, S. Behera, M. Vatsa, and R. Singh. On iris spoofing using print attack. In *International Conference on Pattern Recognition*, pages 1681–1686, August 2014.

[92] Ju Han and B. Bhanu. Human activity recognition in thermal infrared imagery. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 17–17, June 2005.

[93] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, November 1973.

[94] A. Hayter. *Probability and Statistics for Engineers and Scientists*. Cengage Learning, 4th edition, 2012.

[95] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 770–778, June 2016.

[96] A. Hernandez-Vela, M.A. Bautista, X. Perez-Sala, V. Ponce, X. Baro, O. Pujol, C. Angulo, and S. Escalera. BoVDW: Bag-of-Visual-and-Depth-Words for gesture recognition. In *International Conference on Pattern Recognition*, pages 449–452, November 2012.

[97] B. K.P. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA, 1970.

[98] Berthold K. P. Horn and Robert W. Sjoberg. Calculating the reflectance map. *Applied Optics*, 18(11):1770–1779, June 1979.

[99] B.K.P. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. Technical report, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, 1970.

[100] Agnar Hoskuldsson. PLS Regression Methods. *Journal of Chemometrics*, 2(3):211–228, 1988.

[101] Agnar Hoskuldsson. Pls regression methods. *Journal of Chemometrics*, 2(3):211–228, 1988.

[102] X. Huang, C. Ti, Q.-Z. Hou, A. Tokuta, and R. Yang. An experimental study of pupil constriction for liveness detection. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 252–258, 2013.

[103] Tomonori Hyodo. *Radiation physics: proceedings of the International Symposium on Radiation Physics*, chapter Backscattering of Gamma Rays, pages 110–118. NBS special publication. U.S. Dept. of Commerce, National Bureau of Standards, 1977.

[104] Aapo Hyvrinen, Jarmo Hurri, and Patrick O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[105] ISO/IEC. DIS (Draft International Standard) 30107–3, Information technology – Biometric presentation attack detection – Part 3: Testing and reporting, 2016.

[106] ISO/IEC 30107-3:2017. *Information technology – Biometric presentation attack detection – Part 3: Testing and reporting*, 2017.

[107] A. K. Jain and A. Ross. *Handbook of Biometrics*, chapter Introduction to Biometrics, pages 1–22. Springer, 2008.

[108] A. K. Jain and A. Ross. *Handbook of Biometrics*, chapter Introduction to biometrics, pages 1–22. Springer, 2008.

[109] A.K. Jain, A.A. Ross, and K. Nandakumar. *Introduction to Biometrics*, chapter Introduction, pages 1–49. Springer US, Boston, MA, 2011.

[110] Anil K. Jain, Patrick Flynn, and Arun A. Ross. *Handbook of Biometrics*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[111] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the Best Multi-Stage Architecture for Object Recognition? In *IEEE International Conference on Computer Vision*, pages 2146–2153, 2009.

[112] Xiaofei Jia, Xin Yang, Yali Zang, Ning Zhang, Ruwei Dai, Jie Tian, and Jianmin Zhao. Multi-scale block local ternary patterns for fingerprints vitality detection. In *IAPR International Conference on Biometrics*, pages 1–6, 2013.

[113] M. Kanematsu, H. Takano, and K. Nakamura. Highly reliable liveness detection method for iris recognition. In *SICE Annual Conference (SICE)*, pages 361–364, 2007.

[114] J. Kannala and E. Rahtu. BSIF: Binarized statistical image features. In *International Conference on Pattern Recognition*, pages 1363–1366, Nov 2012.

[115] T. Kathikeyan and B. Sabarigiri. Countermeasures against iris spoofing and liveness detection using electroencephalogram (eeg). In *Int. Conference on Computing, Communication and Applications (ICCA)*, pages 1–5, 2012.

[116] M. Kazayawoko, J. J. Balatinecz, and R. T. Woodhams. Diffuse reflectance fourier transform infrared spectra of wood fibers treated with maleated polypropylenes. *Journal of Applied Polymer Science*, 66(6):1163–1173, 1997.

[117] G. Kim, S. Eum, J.K. Suhr, D.I. Kim, K.R. Park, and J. Kim. Face liveness detection based on texture and frequency analyses. In *IAPR International Conference on Biometrics*, pages 67–72, April 2012.

[118] W. Kim, S. Suh, and J.-J. Han. Face liveness detection from a single image via diffusion speed model. *IEEE Transactions on Image Processing*, 24(8):2456–2465, August 2015.

[119] A. Klaeser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *Proceedings of the British Machine Vision Conference*, pages 99.1–99.10. BMVA Press, 2008. doi:10.5244/C.22.99.

[120] N. Kohli, D. Yadav, M. Vatsa, and R. Singh. Revisiting iris recognition with color cosmetic contact lenses. In *IAPR International Conference on Biometrics*, pages 1–7, 2013.

[121] M. Kohn and M. Clynes. Color Dynamics of the Pupil. *Annals of the New York Academy of Sciences*, 156(2):931–950, 1969.

[122] I. Kokkinos and A. Yuille. Scale invariance without scale selection. In *IEEE Inter-*

national Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008.

[123] K. Kollreider, H. Fronthaler, and J. Bigun. Non-intrusive liveness detection by face images. *Image and Vision Computing (JIVC)*, 27(3):233–244, 2009.

[124] J. Komulainen, A. Hadid, M. Pietikainen, A. Anjos, and S. Marcel. Complementary countermeasures for detecting scenic face spoofing attacks. In *IAPR International Conference on Biometrics*, pages 1–7, June 2013.

[125] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection using dynamic texture. In *ACCV International Workshops*, pages 146–157, 2013.

[126] Neslihan Kose and Jean-Luc Dugelay. Countermeasure for the protection of face recognition systems against mask attacks. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–6, 2013.

[127] Neslihan Kose and Jean-Luc Dugelay. On the vulnerability of face recognition systems to spoofing mask attacks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2357–2361, 2013.

[128] Neslihan Kose and Jean-Luc Dugelay. Reflectance analysis based countermeasure technique to detect face mask attacks. In *International Conference on Digital Signal Processing*, pages 1–6, 2013.

[129] A. Krizhevsky. cuda-convnet: High-performance c++/cuda implementation of convolutional neural networks, 2012.

[130] A. Krizhevsky, I. Sutskever, and G. E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[131] A. Krizhevsky, I. Sutskever, and G.E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[132] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[133] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, volume 1 of *NIPS'12*, pages 1097–1105, USA, 2012. Curran Associates Inc.

[134] Supawan Kumpituck, Dongju Li, Hiroaki Kunieda, and Tsuyoshi Isshiki. Fingerprint spoof detection using wavelet based local binary pattern. In *International Conference on Graphic and Image Processing (ICGIP)*, volume 10225, pages 102251C–102251C–8, 2017.

[135] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[136] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[137] E. Lee, K. Park, and J. Kim. *Advances in Biometrics (AB)*, chapter Fake Iris Detection by Using Purkinje Image, pages 397–403. Springer, Berlin, Heidelberg,

2005.

[138] Tai Sing Lee. Image representation using 2d gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):959–971, October 1996.

[139] Ting-Wei Lee, Gwo-Hwa Ju, Heng-Sung Liu, and Yu-Shan Wu. Liveness detection using frequency entropy of image sequences. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2367–2370, 2013.

[140] Jiang-Wei Li. Eye blink detection based on multiple gabor response waves. In *International Conference on Machine Learning and Cybernetics*, volume 5, pages 2852 –2856, July 2008.

[141] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K. Jain. Live face detection based on the analysis of fourier spectra. In *Biometric Technology for Human Identification*, volume 5404, pages 296–303. Proc. SPIE, 2004.

[142] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, December 2016.

[143] L. Li, Z. Xia, L. Li, X. Jiang, X. Feng, and F. Roli. Face anti-spoofing via hybrid convolutional neural network. In *International Conference on the Frontiers and Advances in Data Science (FADS)*, pages 120–124, October 2017.

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

[145] Chengjun Liu and H. Wechsler. A shape- and texture-based enhanced fisher classifier for face recognition. *IEEE Transactions on Image Processing*, 10(4):598–608, April 2001.

[146] Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *IEEE International Conference on Computer Vision*, pages 2486–2493, November 2011.

[147] F. Lourenço and D. Pires. Video shows Samu's medical using silicone fingers, in Ferraz. http://tinyurl.com/akzcrgw, mar. 2013. Accessed: 2016/02/20.

[148] Lovish, A. Nigam, B. Kumar, and P. Gupta. Robust contact lens detection using local phase quantization and binary gabor pattern. In G. Azzopardi and N. Petkov, editors, *Computer Analysis of Images and Patterns*, pages 702–714. Springer International Publishing, 2015.

[149] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[150] J. Lukäs, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, June 2006.

[151] J. Määttä, A. Hadid, and M. Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *IEEE International Joint Conference on Biometrics*, pages 1–7, October 2011.

[152] J. Maatta, A. Hadid, and M. Pietikäinen. Face spoofing detection from single images using texture and local shape analysis. *IET Biometrics*, 1(1):3–10, March 2012.

[153] Davide Maltoni, Dario Maio, Anil K. Jain, and Salil Prabhakar. *Handbook of Fingerprint Recognition*. Springer Publishing Company, Incorporated, 2nd edition, 2009.

[154] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar. Detecting silicone mask-based presentation attack via deep dictionary learning. *IEEE Transactions on Information Forensics and Security*, 12(7):1713–1723, July 2017.

[155] E. Marasco and C. Sansone. Combining perspiration- and morphology-based static features for fingerprint liveness detection. *Pattern Recognition Letters*, 33(9):1148–1156, 2012.

[156] E. Marasco, P. Wild, and B. Cukic. Robust and interoperable fingerprint spoof detection via convolutional neural networks. In *IEEE Symposium on Technologies for Homeland Security (HST)*, pages 1–6, May 2016.

[157] Emanuela Marasco, Peter Johnson, Carlo Sansone, and Stephanie Schuckers. Increase the security of multibiometric systems by incorporating a spoofing detection algorithm in the fusion mechanism. In *International Conference on Multiple Classifier Systems*, pages 309–318, 2011.

[158] Sbastien Marcel, Mark S. Nixon, and Stan Z. Li. *Handbook of Biometric Anti-Spoofing: Trusted Biometrics Under Spoofing Attacks*. Springer Publishing Company, Incorporated, 2014.

[159] Gian Luca Marcialis, Aaron Lewicke, Bozhao Tan, Pietro Coli, Dominic Grimberg, Alberto Congiu, Alessandra Tidu, Fabio Roli, and Stephanie A. C. Schuckers. Livdet 2009 – first international fingerprint liveness detection competition. In Pasquale Foggia, Carlo Sansone, and Mario Vento, editors, *International Conference on Image Analysis and Processing*, volume 5716 of *Lecture Notes in Computer Science*, pages 12–23. Springer, 2009.

[160] G.L. Marcialis, F. Roli, and A. Tidu. Analysis of fingerprint pores for vitality detection. In *International Conference on Pattern Recognition*, pages 1289–1292, August 2010.

[161] Stephen R. Marschner, Stephen H. Westin, Eric P. F. Lafortune, Kenneth E. Torrance, and Donald P. Greenberg. Reflectance measurements of human skin. Technical Report PCG-99-2, Program of Computer Graphics, Cornell University, 1999.

[162] D. Menotti, G. Chiachia, A. Pinto, W.R. Schwartz, H. Pedrini, A.X. Falcao, and A. Rocha. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security*, 10(4):864–879, April 2015.

[163] Simone Milani, Marco Fontani, Paolo Bestagini, Mauro Barni, Alessandro Piva, Marco Tagliasacchi, and Stefano Tubaro. An overview on video forensics. *APSIPA Transactions on Signal and Information Processing*, 1:e2, 2012.

[164] J.C. Monteiro, A.F. Sequeira, H.P. Oliveira, and J.S. Cardoso. Robust iris localisation in challenging scenarios. In *Communications in Computer and Information Science (CCIS)*. Springer-Verlag, 2004.

[165] Kamal Moradi, Christophe Depecker, and Jacques Corset. Diffuse reflectance infrared spectroscopy: Experimental study of nonabsorbing materials and comparison with theories. *Applied Spectroscopy*, 48(12):1491–1497, December 1994.

[166] R. F. Nogueira, R. de Alencar Lotufo, and R. Campos Machado. Fingerprint liveness

detection using convolutional neural networks. *IEEE Transactions on Information Forensics and Security*, 11(6):1206–1213, June 2016.

[167] L. O'Gorman. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE*, 91(12):2021–2040, 2003.

[168] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, July 2002.

[169] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[170] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59, 1996.

[171] V. Ojansivu and J. Heikkilä. *Image and Signal Processing (ISP)*, chapter Blur Insensitive Texture Classification Using Local Phase Quantization, pages 236–243. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[172] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[173] J. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *IEEE International Conference on Computer Vision*, 2014.

[174] A. Pacut and A. Czajka. Aliveness detection for iris biometrics. In *IEEE Int. Carnahan Conferences Security Technology (ICCST)*, pages 122–129, October 2006.

[175] G. Pan, Z. Wu, and L. Sun. *Recent Advances in Face Recognition*, chapter Liveness detection for face recognition, pages 235–252. InTech, 2008.

[176] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *IEEE International Conference on Computer Vision*, pages 1–8, October 2007.

[177] Gang Pan, Lin Sun, Zhaohui Wu, and Yueming Wang. Monocular camera-based face liveness detection by combining eyeblink and scene context. *Telecommunication Systems*, 47:215–225, 2011.

[178] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, number 3 in 1, page 6, 2015.

[179] K. Patel, H. Han, A.K. Jain, and G. Ott. Live face video vs. spoof face video: Use of moiré; patterns to detect replay video attacks. In *IAPR International Conference on Biometrics*, pages 98–105, May 2015.

[180] Keyurkumar Patel, Hu Han, and Anil K. Jain. Cross-database face antispoofing with robust feature representation. In Zhisheng You, Jie Zhou, Yunhong Wang, Zhenan Sun, Shiguang Shan, Weishi Zheng, Jianjiang Feng, and Qijun Zhao, editors, *Biometric Recognition*, pages 611–619, Cham, 2016. Springer International Publishing.

[181] B. Peixoto, C. Michelassi, and A. Rocha. Face liveness detection under bad illumination conditions. In *IEEE International Conference on Image Processing*, pages 3557–3560, September 2011.

[182] Otávio A.B. Penatti, Fernanda B. Silva, Eduardo Valle, Valerie Gouet-Brunet, and Ricardo da S. Torres. Visual word spatial arrangement for image retrieval and classification. *Pattern Recognition*, 47(2):705–720, 2014.

[183] Alex P. Pentland. Finding the illuminant direction. *Journal of the Optical Society of America*, 72(4):448–455, April 1982.

[184] T.F. Pereira, A. Anjos, J.M. de Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *IAPR International Conference on Biometrics*, pages 1–8, 2013.

[185] Mauricio Perez, Sandra Avila, Daniel Moreira, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, 230:279–293, 2017.

[186] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, pages 143–156. Springer-Verlag, 2010.

[187] Q. T. Phan, D. T. Dang-Nguyen, G. Boato, and F. G. B. De Natale. Face spoofing detection using ldp-top. In *IEEE International Conference on Image Processing*, pages 404–408, Sept 2016.

[188] Tsai Ping-Sing and Mubarak Shah. Shape from shading using linear approximation. *Image and Vision Computing*, 12(8):487–498, 1994.

[189] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha. Face spoofing detection through visual codebooks of spectral temporal cubes. *IEEE Transactions on Image Processing*, 24(12):4726–4740, December 2015.

[190] A. Pinto, W. Robson Schwartz, H. Pedrini, and A. Rocha. Using visual rhythms for detecting video-based facial spoof attacks. *IEEE Transactions on Information Forensics and Security*, 10(5):1025–1038, May 2015.

[191] Allan Pinto. A countermeasure method for video-based face spoofing attacks. Master's thesis, University of Campinas, October 2013.

[192] Allan Pinto, Helio Pedrini, Michael Krumdick, Benedict Becker, Adam Czajka, Kevin W. Bowyer, and Anderson Rocha. *Deep Learning in Biometrics*, chapter Counteracting Presentation Attacks in Face Fingerprint and Iris Recognition, page 49. CRC Press, 2018.

[193] Allan Pinto, Helio Pedrini, William Robson Schwartz, and Anderson Rocha. Video-based face spoofing detection through visual rhythm analysis. In *Conference on Graphics, Patterns and Images*, pages 221–228, August 2012.

[194] N. Pinto and D. D. Cox. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 8–15, 2011.

[195] N. Pinto, D. Doukhan, J. J. DiCarlo, and D. D. Cox. A high-throughput screening approach to discovering good forms of biologically-inspired visual representation. *PLoS ONE*, 5(11):e1000579, 2009.

[196] N. Pinto, D. Doukhan, J.J. DiCarlo, and D.D. Cox. A high-throughput screening approach to discovering good forms of biologically-inspired visual representation. *PLoS ONE*, 5(11):e1000579, 2009.

[197] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10):10762–10774, May 2010.

[198] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[199] R. Raghavendra and C. Busch. Robust scheme for iris presentation attack detection using multiscale binarized statistical image features. *IEEE Transactions on Information Forensics and Security*, 10(4):703–715, April 2015.

[200] K.B. Raja, R. Raghavendra, and C. Busch. Video presentation attack detection in visible spectrum iris recognition using magnified phase information. *IEEE Transactions on Information Forensics and Security*, 10(10):2048–2056, October 2015.

[201] V. S. Ramachandran. Perception of shape from shading. *Nature*, 331:163–166, January 1988.

[202] N. K. Ratha, J. H. Connell, and R. M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal*, 40(3):614–634, 2001.

[203] Nalini K. Ratha, Jonathan H. Connell, and Ruud M. Bolle. An analysis of minutiae matching strength. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 223–228, 2001.

[204] C. Rathgeb and A. Uhl. Attacking iris recognition: An efficient hill-climbing technique. In *IEEE/IAPR International Conference on Pattern Recognition (ICPR)*, pages 1217–1220, 2010.

[205] C. Rathgeb and A. Uhl. Statistical attack against iris-biometric fuzzy commitment schemes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 23–30, 2011.

[206] A. Rattani and A. Ross. Automatic adaptation of fingerprint liveness detector to new spoof materials. In *IEEE International Joint Conference on Biometrics*, pages 1–8, Sept 2014.

[207] A. Rattani, W.J. Scheirer, and A. Ross. Open set fingerprint spoof detection across novel fabrication materials. *IEEE Transactions on Information Forensics and Security*, 10(11):2447–2460, November 2015.

[208] Yasar Abbas Ur Rehman, Lai Man Po, and Mengyang Liu. Livenet: Improving features generalization for face liveness detection using convolution neural networks. *Expert Systems with Applications*, 108:159–169, 2018.

[209] W. Ren, S. Singh, M. Singh, and Y.S. Zhu. State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recognition*, 42(2):267–282, 2009. Learning Semantics from Multimedia Content.

[210] Daily Mail Reporter. Face off: Man arrested after boarding plane as an old man - only to land as youthful refugee. http://tinyurl.com/33l9laz, November 2010. Accessed: 2016/02/20.

[211] Daily Mail Reporter. The white robber who carried out six raids disguised as a black man (and very nearly got away with it). http://tinyurl.com/2cvuq59, December 2010. Accessed: 2016/02/20.

[212] Christian Riess and Elli Angelopoulou. Scene illumination as an indicator of image

manipulation. In Rainer Böhme, Philip W. L. Fong, and Reihaneh Safavi-Naini, editors, *Information Hiding*, pages 66–80, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[213] Christian Riess, Sven Pfaller, and Elli Angelopoulou. Reflectance normalization in illumination-based image manipulation detection. In Vittorio Murino, Enrico Puppo, Diego Sona, Marco Cristani, and Carlo Sansone, editors, *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, pages 3–10, Cham, 2015. Springer International Publishing.

[214] W. Robson Schwartz, A. Rocha, and H. Pedrini. Face spoofing detection through partial least squares and low-level descriptors. In *IEEE International Joint Conference on Biometrics*, pages 1–8, October 2011.

[215] Anderson Rocha, Walter Scheirer, Terrance Boult, and Siome Goldenstein. Vision of the unseen: Current trends and challenges in digital image and video forensics. *ACM Computing Surveys*, 43(4):26:1–26:42, October 2011.

[216] Anderson Rocha, Walter Scheirer, Terrance Boult, and Siome Goldenstein. Vision of the unseen: Current trends and challenges in digital image and video forensics. *ACM Comput. Surv.*, 43(4):26:1–26:42, October 2011.

[217] Arun A. Ross, Karthik Nandakumar, and Anil K. Jain. Score level fusion. In *Handbook of Multibiometrics*, volume 6 of *International Series on Biometrics*, pages 91–142. Springer US, 2006.

[218] M. Rousson, T. Brox, and R. Deriche. Active unsupervised texture segmentation on a diffusion based feature space. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 699–704, June 2003.

[219] V. Ruiz-Albacete, P. Tome-Gonzalez, F. Alonso-Fernandez, J. Galbally, J. Fierrez, and J. Ortega-Garcia. Direct attacks using fake images in iris verification. In *First European Workshop on Biometrics and Identity Management (BioID)*, volume 5372 of *Lecture Notes in Computer Science*, pages 181–190. Springer, 2008.

[220] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 2003.

[221] Andrew M. Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y. Ng. On Random Weights and Unsupervised Feature Learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[222] W. Scheirer, A. Rocha, A. Sapkota, and T. Boult. Towards open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, July 2013.

[223] W.J. Scheirer, A. Rocha, A. Sapkota, and T.E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, July 2013.

[224] S. Schuckers, K.W. Bowyer, A.C., and D. Yambay. LivDet 2013 - liveness detection iris competition, 2013.

[225] W.R. Schwartz, A. Kembhavi, D. Harwood, and L.S. Davis. Human Detection Using Partial Least Squares Analysis. In *IEEE International Conference on Computer Vision*, 2009.

[226] A.F. Sequeira, J. Murari, and J.S. Cardoso. Iris liveness detection methods in mo-

bile applications. In *Int. Conference on Computer Vision Theory and Applications (VISAPP)*, volume 3, pages 22–33, January 2014.

[227] A.F. Sequeira, J. Murari, and J.S. Cardoso. Iris liveness detection methods in the mobile biometrics scenario. In *Int. Joint Conference on Neural Network (IJCNN)*, pages 3002–3008, July 2014.

[228] A.F. Sequeira, J. Murari, and J.S. Cardoso. MobBIO a multimodal database captured with a handheld device. In *Int. Conference on Computer Vision Theory and Applications (VISAPP)*, pages 133–139, 2014.

[229] A.F. Sequeira, H.P. Oliveira, J.C. Monteiro, J.P. Monteiro, and J.S. Cardoso. MobILive 2014 - mobile iris liveness detection competition. In *IEEE International Joint Conference on Biometrics*, pages 1–6, September 2014.

[230] Steven A. Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985.

[231] L. Shen and C. Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 697–704, June 2011.

[232] Pedro Silva, Eduardo Luz, Rafael Baeta, Helio Pedrini, Alexandre Xavier Falcao, and David Menotti. An approach to iris contact lens detection based on deep image representations. In *Conference on Graphics, Patterns and Images*, pages 157–164. IEEE, 2015.

[233] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv e-prints*, September 2014.

[234] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[235] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, volume 2, pages 1470–1477, 2003.

[236] Steven W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Pub., San Diego, CA, USA, 1997.

[237] Lin Sun, WaiBin Huang, and MingHui Wu. TIR/VIS Correlation for Liveness Detection in Face Recognition. In *Computer Analysis of Images and Patterns*, pages 114–121, 2011.

[238] Z. Sun, H. Zhang, T. Tan, and J. Wang. Iris image classification based on hierarchical visual codebook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1120–1133, 2014.

[239] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–9, June 2015.

[240] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[241] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, June 2014.

[242] Y. Taigman, M. Yang, M.'A. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[243] B. Tan and S. Schuckers. Spoofing protection for fingerprint scanner by fusing ridge signal and valley noise. *Pattern Recognition*, 43(8):2845 – 2857, 2010.

[244] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, 19(6):1635–1650, June 2010.

[245] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *European Conference on Computer Vision*, pages 504–517, 2010.

[246] S. Tariyal, A. Majumdar, R. Singh, and M. Vatsa. Deep dictionary learning. *IEEE Access*, 4:10096–10109, 2016.

[247] T. Thongkamwitoon, H. Muammar, and P. L. Dragotti. Robust image recapture detection using a k-svd learning approach to train dictionaries of edge profiles. In *IEEE International Conference on Image Processing*, pages 5317–5321, October 2014.

[248] T. Thongkamwitoon, H. Muammar, and P. L. Dragotti. An image recapture detection algorithm based on learning dictionaries of edge profiles. *IEEE Transactions on Information Forensics and Security*, 10(5):953–968, May 2015.

[249] P.A. Tipler and G. Mosca. *Physics for Scientists and Engineers*. Physics for Scientists and Engineers. W. H. Freeman, 2007.

[250] Harold Tipton. *Information Security Management Handbook*. CRC Press, Inc., Boca Raton, FL, USA, 5 edition, 2003.

[251] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, May 2010.

[252] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society of America*, 57(9):1105–1114, September 1967.

[253] R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, and F. Roli. Fusion of multiple clues for photo-attack detection in face recognition systems. In *IEEE International Joint Conference on Biometrics*, pages 1–6, October 2011.

[254] D.A.R. Vigo, F. Shahbaz Khan, J. van de Weijer, and T. Gevers. The impact of color on bag-of-words based object recognition. In *International Conference on Pattern Recognition*, pages 1549–1553, 2010.

[255] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2001.

[256] M. Visentini-Scarzanella and P.L. Dragotti. Modelling radial distortion chains for video recapture detection. In *IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, pages 412–417, September 2013.

[257] Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, and Keying Ye. *Probability & Statistics for Engineers and Scientists*. Pearson Education, Upper Saddle

River, 8th edition, 2007.

[258] Pete Wardern. The SDK for Jetpac's iOS Deep Belief image recognition framework, 2014.

[259] Z. Wei, X. Qiu, Z. Sun, and T. Tan. Counterfeit iris detection based on texture analysis. In *International Conference on Pattern Recognition*, pages 1–4, 2008.

[260] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, April 2015.

[261] P.H. Winston and B. Horn. *The psychology of computer vision*. McGraw-Hill computer science series. McGraw-Hill, 1975.

[262] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.

[263] Cui Xu, Ying Zheng, and Zengfu Wang. Eye states detection by boosting local binary pattern histogram features. In *IEEE International Conference on Image Processing*, pages 1480–1483, October 2008.

[264] D. Yadav, N. Kohli, J.S. Doyle, R. Singh, M. Vatsa, and K.W. Bowyer. Unraveling the effect of textured contact lenses on iris recognition. *IEEE Transactions on Information Forensics and Security*, 9(5):851–862, 2014.

[265] D. Yambay, B. Becker, N. Kohli, D. Yadav, A. Czajka, K. W. Bowyer, S. Schuckers, R. Singh, M. Vatsa, A. Noore, D. Gragnaniello, C. Sansone, L. Verdoliva, L. He, Y. Ru, H. Li, N. Liu, Z. Sun, and T. Tan. Livdet iris 2017 - iris liveness detection competition 2017. In *IEEE International Joint Conference on Biometrics*, pages 733–741, Oct 2017.

[266] D. Yambay, L. Ghiani, P. Denti, G.L. Marcialis, F. Roli, and S. Schuckers. Livdet 2011 – fingerprint liveness detection competition. In *IAPR International Conference on Biometrics*, pages 208–215, 2012.

[267] Junjie Yan, Zhiwei Zhang, Zhen Lei, Dong Yi, and Stan Z. Li. Face liveness detection by exploring multiple scenic clues. In *International Conference on Control Automation Robotics and Vision*, pages 188–193, Dec 2012.

[268] J. Yang, Z. Lei, D. Yi, and S.Z. Li. Person-specific face antispoofing with subject domain adaptation. *IEEE Transactions on Information Forensics and Security*, 10(4):797–809, April 2015.

[269] Jianwei Yang, Zhen Lei, and Stan Z. Li. Learn convolutional neural network for face anti-spoofing. *CoRR*, abs/1408.5601, 2014.

[270] Dong Yi, Zhen Lei, Zhiwei Zhang, and StanZ. Li. Face anti-spoofing: Multi-spectral approach. In Sébastien Marcel, Mark S. Nixon, and Stan Z. Li, editors, *Handbook of Biometric Anti-Spoofing*, Advances in Computer Vision and Pattern Recognition, pages 83–102. Springer London, 2014.

[271] MatthewD. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2014.

[272] L. Zhang, Z. Zhou, and H. Li. Binary gabor pattern: An efficient and robust descriptor for texture classification. In *IEEE International Conference on Image Processing*, pages 81–84, September 2012.

[273] Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition. In *IEEE International Conference on Computer Vision*, volume 1, pages 786–791, 2005.

[274] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and S.Z. Li. A face antispoofing database with diverse attacks. In *IAPR International Conference on Biometrics*, pages 26–31, April 2012.

[275] Q. Zheng and R. Chellappa. Estimation of illuminant direction, albedo, and shape from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):680–702, July 1991.

Appendix A

Ethics Committee Approval

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE COMPUTAÇÃO

Termo de Consentimento Livre e Esclarecido (TCLE)

Campinas, November 21, 2018

Captura de Dados para Avaliação de Segurança em Sistemas Biométricos de Faces

Pesquisador Responsável: Prof. Dr. Anderson de Rezende Rocha

Diretor do Instituto: Prof. Dr. Hans Kurt Edmund Liesenberg

Termo de Consentimento Livre e Esclarecido (TCLE)

Responsável: Prof. Dr. Anderson de Rezende Rocha

Objetivo da pesquisa: A presente pesquisa tem como objetivo a construção de uma base de dados contendo vídeos da face de pessoas adultas para uso científico. Esses dados serão utilizados no desenvolvimento e na avaliação da eficácia de novas metodologias que visam detectar procedimentos que tentam burlar o mecanismo de controle de acesso a um sistema de computação, realizado por um sistema de biometria de face. Os dados também poderão ser utilizados no desenvolvimento de novos métodos para reconhecimento e detecção de face de humanos por computador.

Justificativa: Embora existam algumas bases de dados destinadas à avaliação de metodologias que visam burlar sistemas biométricos faciais, como a Print-Attack[1], do *Instituto de Pesquisa IDIAP*, e a o *benchmark* NUAA[2], da *Universidade de Aeronáutica e Astronáutica Nanjing*, os seus proprietários não permitem a divulgação, feita por terceiros, de dados gerados durante experimentos científicos que utilizam tais bases, o que dificulta a geração e a disseminação de novos conhecimentos adquiridos em pesquisas científicas. Adicionalmente, a existência de bases de dados de vídeos disponíveis para o estudo de possíveis ataques a sistemas de biometria de faces são ainda mais escassos e não temos conhecimento de uma base livre para esse fim.

Procedimentos: Os participantes serão filmados por uma câmera digital, por um período de aproximadamente 2 minutos, em três locais dentro da universidade, com diferentes condições de iluminação e pose.

Para que a filmagem seja realizada de maneira organizada, diminuindo, ao máximo, o desconforto ou riscos aos participantes, uma fila circular será organizada com a finalidade de capturar três vídeos de cada participante, com aproximadamente 40 segundos de duração para cada vídeo.

Os locais escolhidos serão internos e externos ao prédio do Instituto de Computação (IC/Unicamp), de modo a permitir a aquisição de vídeos com diferentes condições de iluminação e pose. Tais locais serão arejados, livres de circulação de veículos automotivos, distantes de construções em andamento ou eventos de outra natureza que possam colocar em risco a segurança e a integridade física dos participantes.

As filmagens dos participantes poderão ocorrer durante o dia ou à noite, sendo que, no período diurno, os participantes não ficarão expostos diretamente aos raios solares durante a realização da pesquisa.

[1] A. Anjos and S. Marcel, "Counter-Measures to Photo Attacks in Face Recognition: A Public Database and A Baseline," in *International Joint Conference on Biometrics (IJCB'11)*, 2011.

[2] X.Tan, Y.Li, J.Liu, and L.Jiang. "Face Liveness Detection from A Single Image with Sparse Low Rank Bilinear Discriminative Model," In *Proceedings of 11th European Conference on Computer Vision (ECCV'10)*, 2010.

Desconfortos e riscos: Durante a realização da pesquisa, poderá haver um certo cansaço ou fadiga dos participantes, visto que eles ficarão em uma fila aguardando a sua vez de participar na coleta em andamento. No entanto, como a duração dos vídeos é curta para cada participante, problemas dessa natureza não deverão trazer desconfortos mais sérios aos participantes. Não há riscos previsíveis.

Benefícios: O participante contribuirá com informações úteis para o desenvolvimento e a avaliação de novos métodos que visam melhorar a segurança dos sistemas de computação, como por exemplo, sistemas que possuem mecanismos de controle de acesso por biometria facial.

Métodos alternativos: Não há métodos alternativos para a obtenção das informações desejadas, pois a única forma para obtenção de vídeos da face dos participantes é por meio de filmagem com uma câmera digital.

Acompanhamento e assistência: Não se aplica.

Esclarecimentos: Em caso de dúvidas, o participante poderá entrar em contato, a qualquer momento, com Prof. Dr. Anderson de Rezende Rocha, pelo telefone (19) 3521-5854 ou pelo e-mail anderson.rocha@ic.unicamp.br, responsável por essa pesquisa.

Possibilidade de inclusão em grupo controle ou placebo: Não se aplica.

Possibilidade de desistência: É assegurado ao participante o direito de poder abandonar a pesquisa a qualquer momento e sem aviso prévio. O participante que assim o fizer não sofrerá penalidade alguma, constrangimento ou represália de qualquer natureza.

Critérios de Inclusão e Exclusão dos Participantes: Os participantes serão convidados a participar da pesquisa por meio da divulgação de comunicados que serão afixados nos quadros de avisos do Instituto de Computação da Unicamp. Adicionalmente, esses comunicados serão enviados por e-mail para os alunos de graduação, pós-graduação e para os docentes do mesmo instituto. Dado o caráter voluntário da pesquisa, caso o número de participantes não seja suficiente para a realização da mesma, comunicados serão enviados às listas de e-mails dos alunos de graduação e pós-graduação dos outros institutos pertencentes à Unicamp. Para participar da pesquisa, o candidato deverá atender aos seguintes critérios: (1) ser maior de 18 anos; e (2) residir na Região Metropolitana de Campinas. Ressalta-se que a participação é voluntária e o interessado pode desistir sem prejuízos às partes em qualquer momento sem necessidade de justificativa prévia. Em caso de desistência de um candidato, o candidato subsequente será chamado sem maiores prejuízos à pesquisa ou às partes envolvidas.

Sigilo de dados confidenciais: Não se aplica.

Despesas decorrentes da participação: Os participantes não terão nenhuma despesa com a pesquisa, dessa forma, não está previsto ressarcimento decorrente da participação na pesquisa.

Riscos previsíveis: A pesquisa não envolve riscos previsíveis, visto que o participante será apenas filmado com um câmera digital comum, procedimento simples que não oferece risco ou dano à integridade física, psíquica, moral ou de outra natureza.

Ao participante: O participante receberá uma cópia deste *Termo de Consentimento Livre e Esclarecido (TCLE)* para posteriores consultas ou esclarecimentos de dúvidas, contendo a assinatura do responsável pela pesquisa na última página deste termo, bem como a sua rubrica em todas as páginas.

Autorização para uso de imagem: O participante autoriza a utilização de sua imagem apenas para o uso em pesquisas científicas, incluindo a confecção e divulgação de trabalhos científicos em todo o território brasileiro e no exterior. É assegurado ao participante que seus dados capturados não sofrerão nenhuma manipulação que possa lhe causar eventuais constrangimentos ou danos morais e psíquicos. As imagens dos participantes serão armazenadas em dispositivos acessíveis apenas pelos pesquisadores envolvidos na pesquisa. Caso outros pesquisadores ou grupos de pesquisa tenham interesse em utilizar esses dados em seus trabalhos, os mesmos terão que assinar um termo de responsabilidade sobre a utilização e proteção de tais dados. Ao concordar com este *Termo de Consentimento Livre e Esclarecido (TCLE)*, o participante estará cedendo permanentemente os direitos de sua imagem para os fins descritos neste parágrafo.

Sobre o responsável: O participante poderá entrar em contato, a qualquer momento, com o responsável pela pesquisa, Prof. Dr. Anderson de Rezende Rocha, por meio dos seguintes canais de comunicação:

Prof. Dr. Anderson de Rezende Rocha

Endereço profissional: Av. Albert Einstein, 1251, Sala 40,
Cidade Universitária, Campinas-SP, 13083-852

Telefone: (19) 3521-5854

e-mail: anderson.rocha@ic.unicamp.br

Comitê de ética em pesquisa: Caso o participante tenha alguma denúncia ou reclamações referentes aos aspectos éticos da presente pesquisa, este deve entrar em contato com o Comitê de Ética em Pesquisa (CEP) da Faculdade de Ciências Médicas (FCM) da Universidade Estadual de Campinas (UNICAMP), por e-mail, telefone ou pessoalmente, no seguinte endereço:

Comitê de Ética em Pesquisa (CEP/ FCM/ UNICAMP)

Endereço: Rua Tessália Vieira de Camargo, 126
Cidade Universitária, Campinas-SP, 13083-887

Telefone: (19) 3521-8936 ou (19) 3521-7187

e-mail: cep@fcm.unicamp.br

Sendo assim, eu, ...,
declaro ter lido as cláusulas acima e concordo em participar da pesquisa referente a este *Termo de Consentimento Livre e Esclarecido (TCLE)*, que está sob a responsabilidade do Prof. Dr. Anderson de Rezende Rocha, pesquisador e professor do Instituto de Computação (IC) da Universidade Estadual de Campinas (UNICAMP).

Participante

Responsável: Prof. Dr. Anderson de Rezende Rocha

Appendix B

Convolutional Network Operations

Our networks use classic convolutional operations that can be viewed as linear and non-linear image processing operations. When stacked, these operations essentially extract higher level representations, named *multiband images*, whose pixel attributes are concatenated into high-dimensional feature vectors for later pattern recognition.[1]

Assuming $\hat{I} = (D_I, \mathbf{I})$ as a multiband image, where $D_I \subset Z^2$ is the image domain and $\mathbf{I}(p) = \{I_1(p), I_2(p), \ldots, I_m(p)\}$ is the attribute vector of a m-band pixel $p = (x_p, y_p) \in D_I$, the aforementioned operations can be described as follows.

Filter Bank Convolution

Let $\mathcal{A}(p)$ be a squared region centered at p of size $L_\mathcal{A} \times L_\mathcal{A}$, such that $\mathcal{A} \subset D_I$ and $q \in \mathcal{A}(p)$ iff $\max(|x_q - x_p|, |y_q - y_p|) \leq (L_\mathcal{A} - 1)/2$. Additionally, let $\Phi = (\mathcal{A}, W)$ be a filter with weights $W(q)$ associated with pixels $q \in \mathcal{A}(p)$. In the case of multiband filters, filter weights can be represented as vectors $\mathbf{W}_i(q) = \{w_{i,1}(q), w_{i,2}(q), \ldots, w_{i,m}(q)\}$ for each filter i of the bank, and a multiband filter bank $\Phi = \{\Phi_1, \Phi_2, \ldots, \Phi_n\}$ is a set of filters $\Phi_i = (\mathcal{A}, \mathbf{W}_i)$, $i = \{1, 2, \ldots, n\}$.

The convolution between an input image $\hat{I}$ and a filter Φ_i produces a band i of the filtered image $\hat{J} = (D_J, \mathbf{J})$, where $D_J \subset D_I$ and $\mathbf{J} = (J_1, J_2, \ldots, J_n)$, such that for each $p \in D_J$,

$$J_i(p) = \sum_{\forall q \in \mathcal{A}(p)} \mathbf{I}(q) \cdot \mathbf{W}_i(q). \tag{B.1}$$

Rectified Linear Activation

Filter activation in this work is performed by rectified linear units (RELUs) of the type present in many state-of-the-art convolutional architectures [130, 194] and is defined as

$$J_i(p) = \max(J_i(p), 0). \tag{B.2}$$

[1] This appendix describes convolutional networks from an image processing perspective, therefore the use of terms like image *domain*, image *band*, *etc.*

Spatial Pooling

Spatial pooling is an operation of paramount importance in the literature of convolutional networks [135] that aims at bringing translational invariance to the features by aggregating activations from the same filter in a given region.

Let $\mathcal{B}(p)$ be a pooling region of size $L_\mathcal{B} \times L_\mathcal{B}$ centered at pixel p and $D_K = D_J/s$ be a regular subsampling of every s pixels $p \in D_J$. We call s the *stride* of the pooling operation. Given that $D_J \subset Z^2$, if $s = 2$, $|D_K| = |D_J|/4$, for example. The pooling operation resulting in the image $\hat{K} = (D_K, \mathbf{K})$ is defined as

$$K_i(p) = \sqrt[\alpha]{\sum_{\forall q \in \mathcal{B}(p)} J_i(q)^\alpha}, \tag{B.3}$$

where $p \in D_K$ are pixels in the new image, $i = \{1, 2, \ldots, n\}$ are the image bands, and α is a hyperparameter that controls the sensitivity of the operation. In other words, our pooling operation is the L_α-norm of values in $\mathcal{B}(p)$. The stride s and the size of the pooling neighborhood defined by $L_\mathcal{B}$ are other hyperparameters of the operation.

Divisive Normalization

The last operation considered in this work is divisive normalization, a mechanism widely used in top-performing convolutional networks [130, 194] that is based on gain control mechanisms found in cortical neurons [79].

This operation is also defined within a squared region $\mathcal{C}(p)$ of size $L_\mathcal{C} \times L_\mathcal{C}$ centered at pixel p such that

$$O_i(p) = \frac{K_i(p)}{\sqrt{\sum_{j=1}^{n} \sum_{\forall q \in \mathcal{C}(p)} K_j(q)^2}} \tag{B.4}$$

for each pixel $p \in D_O \subset D_K$ of the resulting image $\hat{O} = (D_O, \mathbf{O})$. Divisive normalization promotes competition among pooled filter bands such that high responses will prevail even more over low ones, further strengthening the robustness of the output representation $\mathbf{O}$.

Appendix C

Copyright Permissions

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

Taylor and Francis Group LLC Books LICENSE TERMS AND CONDITIONS

Jul 23, 2018

This is a License Agreement between Allan Pinto ("You") and Taylor and Francis Group LLC Books ("Taylor and Francis Group LLC Books") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Taylor and Francis Group LLC Books, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	4394800794229
License date	Jul 20, 2018
Licensed content publisher	Taylor and Francis Group LLC Books
Licensed content title	Deep Learning in Biometrics
Licensed content date	Mar 12, 2018
Type of Use	Thesis/Dissertation
Requestor type	Academic institution
Format	Print, Electronic
Portion	chapter/article
Number of pages in chapter/article	50
The requesting person/organization is:	Allan da Silva Pinto/University of Campinas
Title or numeric reference of the portion(s)	Chapter 11
Title of the article or chapter the portion is from	Counteracting Presentation Attacks in Face, Fingerprint, and Iris Recognition
Editor of portion(s)	N/A
Author of portion(s)	Allan Pinto, Helio Pedrini, Michael Krumdick, Benedict Becker, Adam Czajka, Kevin W. Bowyer, and Anderson Rocha
Volume of serial or monograph.	N/A
Page range of the portion	11/245-295
Publication date of portion	2018
Rights for	Main product
Duration of use	Life of current edition
Creation of copies for the	no

disabled	
With minor editing privileges	yes
For distribution to	Worldwide
In the following language(s)	Original language of publication
With incidental promotional use	no
The lifetime unit quantity of new product	Up to 2,000,000
Title	Analysis of Intrinsic and Extrinsic Properties of Biometric Samples for Presentation Attack Detection
Instructor name	Prof. Dr. Anderson de Rezende Rocha
Institution name	Institute of Computing, University of Campinas
Expected presentation date	Sep 2018
Billing Type	Invoice
Billing Address	Allan Pinto Av. Albert Einstein, 1251 Cidade Universitária Zeferino Vaz Campinas, Brazil 13180230 Attn: Allan Pinto
Total (may include CCC user fee)	0.00 USD
Terms and Conditions	

TERMS AND CONDITIONS

The following terms are individual to this publisher:

Taylor and Francis Group and Informa healthcare are division of Informa plc. Permission will be void if material exceeds 10% of all the total pages in your publication and over 20% of the original publication. This includes permission granted by Informa plc and all of its subsidaries.

Other Terms and Conditions:

Please make sure the appropriate source is credited. Insert information as appropriate. Each copy containing our material must bear a credit line in the following format: Copyright (Insert © Year) From (Insert Title) by (Insert Author/Editor Name). Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc This permission does not cover any third party copyrighted work which may appear in the material requested.

STANDARD TERMS AND CONDITIONS

1. Description of Service; Defined Terms. This Republication License enables the User to obtain licenses for republication of one or more copyrighted works as described in detail on the relevant Order Confirmation (the “Work(s)”). Copyright Clearance Center, Inc. (“CCC”) grants licenses through the Service on behalf of the rightsholder identified on the Order Confirmation (the “Rightsholder”). “Republication”, as used herein, generally means the inclusion of a Work, in whole or in part, in a new work or works, also as described on the

Order Confirmation. “User”, as used herein, means the person or entity making such republication.
2. The terms set forth in the relevant Order Confirmation, and any terms set by the Rightsholder with respect to a particular Work, govern the terms of use of Works in connection with the Service. By using the Service, the person transacting for a republication license on behalf of the User represents and warrants that he/she/it (a) has been duly authorized by the User to accept, and hereby does accept, all such terms and conditions on behalf of User, and (b) shall inform User of all such terms and conditions. In the event such person is a “freelancer” or other third party independent of User and CCC, such party shall be deemed jointly a “User” for purposes of these terms and conditions. In any event, User shall be deemed to have accepted and agreed to all such terms and conditions if User republishes the Work in any fashion.
3. Scope of License; Limitations and Obligations.
3.1 All Works and all rights therein, including copyright rights, remain the sole and exclusive property of the Rightsholder. The license created by the exchange of an Order Confirmation (and/or any invoice) and payment by User of the full amount set forth on that document includes only those rights expressly set forth in the Order Confirmation and in these terms and conditions, and conveys no other rights in the Work(s) to User. All rights not expressly granted are hereby reserved.
3.2 General Payment Terms: You may pay by credit card or through an account with us payable at the end of the month. If you and we agree that you may establish a standing account with CCC, then the following terms apply: Remit Payment to: Copyright Clearance Center, 29118 Network Place, Chicago, IL 60673-1291. Payments Due: Invoices are payable upon their delivery to you (or upon our notice to you that they are available to you for downloading). After 30 days, outstanding amounts will be subject to a service charge of 1-1/2% per month or, if less, the maximum rate allowed by applicable law. Unless otherwise specifically set forth in the Order Confirmation or in a separate written agreement signed by CCC, invoices are due and payable on “net 30” terms. While User may exercise the rights licensed immediately upon issuance of the Order Confirmation, the license is automatically revoked and is null and void, as if it had never been issued, if complete payment for the license is not received on a timely basis either from User directly or through a payment agent, such as a credit card company.
3.3 Unless otherwise provided in the Order Confirmation, any grant of rights to User (i) is “one-time” (including the editions and product family specified in the license), (ii) is non-exclusive and non-transferable and (iii) is subject to any and all limitations and restrictions (such as, but not limited to, limitations on duration of use or circulation) included in the Order Confirmation or invoice and/or in these terms and conditions. Upon completion of the licensed use, User shall either secure a new permission for further use of the Work(s) or immediately cease any new use of the Work(s) and shall render inaccessible (such as by deleting or by removing or severing links or other locators) any further copies of the Work (except for copies printed on paper in accordance with this license and still in User's stock at the end of such period).
3.4 In the event that the material for which a republication license is sought includes third party materials (such as photographs, illustrations, graphs, inserts and similar materials) which are identified in such material as having been used by permission, User is responsible for identifying, and seeking separate licenses (under this Service or otherwise) for, any of

such third party materials; without a separate license, such third party materials may not be used.
3.5 Use of proper copyright notice for a Work is required as a condition of any license granted under the Service. Unless otherwise provided in the Order Confirmation, a proper copyright notice will read substantially as follows: “Republished with permission of [Rightsholder’s name], from [Work's title, author, volume, edition number and year of copyright]; permission conveyed through Copyright Clearance Center, Inc. ” Such notice must be provided in a reasonably legible font size and must be placed either immediately adjacent to the Work as used (for example, as part of a by-line or footnote but not as a separate electronic link) or in the place where substantially all other credits or notices for the new work containing the republished Work are located. Failure to include the required notice results in loss to the Rightsholder and CCC, and the User shall be liable to pay liquidated damages for each such failure equal to twice the use fee specified in the Order Confirmation, in addition to the use fee itself and any other fees and charges specified.
3.6 User may only make alterations to the Work if and as expressly set forth in the Order Confirmation. No Work may be used in any way that is defamatory, violates the rights of third parties (including such third parties' rights of copyright, privacy, publicity, or other tangible or intangible property), or is otherwise illegal, sexually explicit or obscene. In addition, User may not conjoin a Work with any other material that may result in damage to the reputation of the Rightsholder. User agrees to inform CCC if it becomes aware of any infringement of any rights in a Work and to cooperate with any reasonable request of CCC or the Rightsholder in connection therewith.
4. Indemnity. User hereby indemnifies and agrees to defend the Rightsholder and CCC, and their respective employees and directors, against all claims, liability, damages, costs and expenses, including legal fees and expenses, arising out of any use of a Work beyond the scope of the rights granted herein, or any use of a Work which has been altered in any unauthorized way by User, including claims of defamation or infringement of rights of copyright, publicity, privacy or other tangible or intangible property.
5. Limitation of Liability. UNDER NO CIRCUMSTANCES WILL CCC OR THE RIGHTSHOLDER BE LIABLE FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL OR INCIDENTAL DAMAGES (INCLUDING WITHOUT LIMITATION DAMAGES FOR LOSS OF BUSINESS PROFITS OR INFORMATION, OR FOR BUSINESS INTERRUPTION) ARISING OUT OF THE USE OR INABILITY TO USE A WORK, EVEN IF ONE OF THEM HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In any event, the total liability of the Rightsholder and CCC (including their respective employees and directors) shall not exceed the total amount actually paid by User for this license. User assumes full liability for the actions and omissions of its principals, employees, agents, affiliates, successors and assigns.
6. Limited Warranties. THE WORK(S) AND RIGHT(S) ARE PROVIDED “AS IS”. CCC HAS THE RIGHT TO GRANT TO USER THE RIGHTS GRANTED IN THE ORDER CONFIRMATION DOCUMENT. CCC AND THE RIGHTSHOLDER DISCLAIM ALL OTHER WARRANTIES RELATING TO THE WORK(S) AND RIGHT(S), EITHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. ADDITIONAL RIGHTS MAY BE REQUIRED TO USE ILLUSTRATIONS, GRAPHS, PHOTOGRAPHS, ABSTRACTS, INSERTS OR OTHER PORTIONS OF THE WORK (AS OPPOSED TO THE ENTIRE WORK) IN A MANNER CONTEMPLATED

BY USER; USER UNDERSTANDS AND AGREES THAT NEITHER CCC NOR THE RIGHTSHOLDER MAY HAVE SUCH ADDITIONAL RIGHTS TO GRANT.

7. Effect of Breach. Any failure by User to pay any amount when due, or any use by User of a Work beyond the scope of the license set forth in the Order Confirmation and/or these terms and conditions, shall be a material breach of the license created by the Order Confirmation and these terms and conditions. Any breach not cured within 30 days of written notice thereof shall result in immediate termination of such license without further notice. Any unauthorized (but licensable) use of a Work that is terminated immediately upon notice thereof may be liquidated by payment of the Rightsholder's ordinary license price therefor; any unauthorized (and unlicensable) use that is not terminated immediately for any reason (including, for example, because materials containing the Work cannot reasonably be recalled) will be subject to all remedies available at law or in equity, but in no event to a payment of less than three times the Rightsholder's ordinary license price for the most closely analogous licensable use plus Rightsholder's and/or CCC's costs and expenses incurred in collecting such payment.

8. **Miscellaneous.**

8.1 User acknowledges that CCC may, from time to time, make changes or additions to the Service or to these terms and conditions, and CCC reserves the right to send notice to the User by electronic mail or otherwise for the purposes of notifying User of such changes or additions; provided that any such changes or additions shall not apply to permissions already secured and paid for.

8.2 Use of User-related information collected through the Service is governed by CCC's privacy policy, available online here:
http://www.copyright.com/content/cc3/en/tools/footer/privacypolicy.html.

8.3 The licensing transaction described in the Order Confirmation is personal to User. Therefore, User may not assign or transfer to any other person (whether a natural person or an organization of any kind) the license created by the Order Confirmation and these terms and conditions or any rights granted hereunder; provided, however, that User may assign such license in its entirety on written notice to CCC in the event of a transfer of all or substantially all of User's rights in the new material which includes the Work(s) licensed under this Service.

8.4 No amendment or waiver of any terms is binding unless set forth in writing and signed by the parties. The Rightsholder and CCC hereby object to any terms contained in any writing prepared by the User or its principals, employees, agents or affiliates and purporting to govern or otherwise relate to the licensing transaction described in the Order Confirmation, which terms are in any way inconsistent with any terms set forth in the Order Confirmation and/or in these terms and conditions or CCC's standard operating procedures, whether such writing is prepared prior to, simultaneously with or subsequent to the Order Confirmation, and whether such writing appears on a copy of the Order Confirmation or in a separate instrument.

8.5 The licensing transaction described in the Order Confirmation document shall be governed by and construed under the law of the State of New York, USA, without regard to the principles thereof of conflicts of law. Any case, controversy, suit, action, or proceeding arising out of, in connection with, or related to such licensing transaction shall be brought, at CCC's sole discretion, in any federal or state court located in the County of New York, State of New York, USA, or in any federal or state court whose geographical jurisdiction covers the location of the Rightsholder set forth in the Order Confirmation. The parties expressly

submit to the personal jurisdiction and venue of each such federal or state court.If you have any comments or questions about the Service or Copyright Clearance Center, please contact us at 978-750-8400 or send an e-mail to info@copyright.com.
v 1.1