GENERALISABLE PRESENTATION ATTACK DETECTION FOR MULTIPLE TYPES OF BIOMETRIC CHARACTERISTICS

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer.nat.)

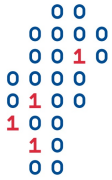
vorgelegte Dissertation von

LÁZARO JANIER GONZÁLEZ SOLER, B.SC.

geboren in Havanna, Kuba

1. Gutachten: Prof. Dr. Ralf Dörner - HOCHSCHULE RHEINMAIN
2. Gutachten: Prof. Dr. Julian Fierrez - AUTONOMOUS UNIVERSITY OF MADRID
3. Gutachten: Prof. Dr. Martin Drahansky - BRNO UNIVERSITY OF TECHNOLOGY

Tag der Einreichung: 05.09.2022
Tag der Prüfung: 14.12.2022



Promotionszentrum
Angewandte Informatik
HAW Hessen

PARTNER:
Frankfurt University of Applied Sciences
Hochschule Darmstadt
Hochschule Fulda
Hochschule RheinMain

Darmstadt, December 2022

SUPERVISORS:
Prof. Dr. Christoph Busch - HOCHSCHULE DARMSTADT
Prof. Dr. Andreas Heinemann - HOCHSCHULE DARMSTADT
Prof. Dr. Marta Gomez-Barrero - HOCHSCHULE ANSBACH

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

ABSTRACT

Biometric systems have experienced a large development in recent years since they are accurate, secure, and in many cases, more user convenient than traditional credential-based access control systems. In spite of their benefits, biometric systems are still vulnerable to attack presentations (APs), which can be easily launched by a fraudulent subject without having a wide expert knowledge. This way, he/she can gain access to several applications, such as bank accounts and smartphone unlocking, where biometric systems are frequently deployed. In order to mitigate such threats and increase the security of biometric systems, the development of reliable Presentation Attack Detection (PAD) algorithms is of utmost importance to the research community.

In the context of PAD, we explore in this Thesis different strategies and methods in order to improve the generalisation capability of PAD schemes. To that end, we propose the definition of a semantic common feature space which successfully discriminates bona fide presentations (BPs)[1] from APs. In essence, this process is seeking for those significant features extracted from known PAI species samples that are observed in unknown PAI species. In addition, we explore several handcrafted techniques in order to build a reliable description of features per biometric characteristic studied. The experimental evaluation shows that a common feature space can be computed through the fusion between generative models and discriminative approaches. Remarkable detection performances for high-security thresholds lead to the construction of a convenient (i.e., low BP rejection rates or Bona fide Presentation Classification Error Rate (BPCER)) and secure (i.e., low AP acceptance rates or Attack Presentation Classification Error Rate (APCER)) PAD subsystem.

Keywords: Biometric systems, presentation attack detection, generalisable feature spaces, semantic common feature spaces.

1 biometric presentation without the goal of interfering with the operation of the biometric system [96]

ZUSAMMENFASSUNG

Die Verwendung biometrischer Systeme hat sich in den letzten Jahren stark verbreitet, da Genauigkeit, Sicherheit und Benutzerfreundlichkeit im Gegensatz zu herkömmlichen Zugangskontrollsystemen deutlich zunahm. Trotz ihrer Vorteile sind biometrische Systeme immer noch anfällig für Angriffe, die von einer betrügerischen Person ohne großes Fachwissen leicht ausgeführt werden können. Auf diese Weise kann er/sie sich Zugang zu verschiedenen Anwendungen verschaffen, z. B. zu Bankkonten und zur Entsperrung von Smartphones, wo biometrische Systeme häufig eingesetzt werden. Um solche Bedrohungen zu entschärfen und die Sicherheit biometrischer Systeme zu erhöhen, ist die Entwicklung zuverlässiger Algorithmen zur Erkennung von Präsentationsangriffen (Presentation Attack Detection, PAD) für die Forschungsgemeinschaft von größter Bedeutung.

Im Zusammenhang mit PAD untersucht diese Arbeit verschiedene Strategien und Methoden, um die Generalisierungsfähigkeit von PAD-Verfahren zu verbessern. Zu diesem Zweck wird ein gemeinsamer semantischen Merkmalsraums definiert, der eine erfolgreiche Unterscheidung zwischen bona fiden Darstellungen und Angriffsdarstellungen ermöglicht. Im Wesentlichen geht es bei diesem Prozess um die Suche nach signifikanten Merkmalen, die aus bekannten Angriffsmustern extrahiert wurden und bei unbekannten Angriffen zu beobachten sind. Darüber hinaus erforscht die Arbeit verschiedene handwerkliche Techniken, um eine zuverlässige Beschreibung der Merkmale für jedes untersuchte biometrische Merkmal zu erstellen. Die experimentelle Auswertung zeigt, dass durch die Fusion zwischen generativen Modellen und diskriminativen Ansätzen ein gemeinsamer Merkmalsraum berechnet werden kann. Bemerkenswerte Erkennungsleistungen für hochsichere Schwellenwerte führen zur Konstruktion eines benutzerfreundlichen (d.h. niedrige Ablehnungsquoten für bona fide Präsentationen) und sicheren (d.h. niedrige Akzeptanzquoten für Angriffspräsentationen) PAD-Subsystems.

PUBLICATIONS

JOURNAL PAPERS

[1] L. J. Gonzalez-Soler, M. Gomez-Barrero, and C. Busch. "On the Generalisation capabilities of Fisher Vector based Face Presentation Attack Detection." In: *IET Biometrics* 10.5 (2021), pp. 480–496.

[2] L. J. Gonzalez-Soler, M. Gomez-Barrero, and C. Busch. "Towards Generalisable Facial Presentation Attack Detection by analysing Facial Regions." In: *Trans. on Biometrics, Behavior, and Identity Science (TBIOM)* (2022). Under Review.

[3] L. J. Gonzalez-Soler, M. Gomez-Barrero, L. Chang, A. Perez-Suarez, and C. Busch. "Fingerprint Presentation Attack Detection Based on Local Features Encoding for Unknown Attacks." In: *IEEE Access* 9 (2021), pp. 5806–5820.

[4] L. J. Gonzalez-Soler, M. Gomez-Barrero, J. Kolberg, L. Chang, A. Perez-Suarez, and C. Busch. "Local Feature Encoding for Unknown Presentation Attack Detection: An Analysis of Different Local Feature Descriptors." In: *IET Biometrics* 10.4 (2021), pp. 374–391.

BOOK CHAPTERS

[1] L. J. Gonzalez-Soler, M. Gomez-Barrero, J. Patino, M Kamble, M. Todisco, and C. Busch. "Fisher Vectors for Biometric Presentation Attack Detection." In: *Handbook of Biometric Antispoofing*. Springer, 2022.

CONFERENCE PAPERS

[1] L. J. Gonzalez-Soler, M. Gomez-Barrero, and C. Busch. "Evaluating the Sensitivity of Face Presentation Attack Detection Techniques to Images of Varying Resolutions." In: *Norwegian Information Security Conf. (NISK)*. 2020.

[2] L. J. Gonzalez-Soler, M. Gomez-Barrero, and C. Busch. "Fisher Vector Encoding of Dense-BSIF Features for Unknown Face Presentation Attack Detection." In: *Proc. Intl. Conf. of the Special Interest Group on Biometrics (BIOSIG)*. LNI. GI, 2020, pp. 1–11.

[3] L. J. Gonzalez-Soler, M. Gomez-Barrero, L. Chang, A. Perez-Suarez, and C. Busch. "On the Impact of Different Fabrication Materials on Fingerprint Presentation Attack Detection." In: *Proc. Int. Conf. on Biometrics (ICB)*. 2019.

[4] L. J. Gonzalez-Soler, J. Patino, M. Gomez-Barrero, M. Todisco, C. Busch, and N. Evans. "Texture-based Presentation Attack Detection for Automatic Speaker Verification." In: *Proc. Intl. Workshop on Information Forensics and Security (WIFS)*. 2020, pp. 1–6.

[5] L. J. Gonzalez-Soler, M. Gomez-Barrero, M. Günther, and C. Busch. "Analysing the feasibility of using Objectosphere for Face Presentation Attack Detection." In: *Norwegian Information Security Conf. (NISK)*. 2021.

[6] L. J. Gonzalez-Soler, M. Gomez-Barrero, M. Kamble, M. Todisco, and C. Busch. "Dual-Stream Temporal Convolutional Neural Network for Voice Presentation Attack Detection." In: *Proc. Int. Workshop on Biometrics and Forensics (IWBF)*. 2022.

[7] L. J. Gonzalez-Soler, K. A. Barhaugen, M. Gomez-Barrero, and C. Busch. "When Facial Recognition Systems become Presentation Attack Detectors." In: *Proc. Intl. Conf. of the Special Interest Group on Biometrics (BIOSIG)*. LNI. GI, 2022, pp. 1–8.

NON-FIRST-AUTHOR PUBLICATIONS

[1] M. Ibsen, L. J. Gonzalez-Soler, C. Rathgeb, P. Drozdowski, M. Gomez-Barrero, and C. Busch. "Differential Anomaly Detection for Facial Images." In: *Proc. Intl. Workshop on Information Forensics and Security (WIFS)*. Best Paper Award. 2021, pp. 1–6.

Science is what we understand well enough
to explain to a computer.
Art is everything else we do.

— knuth:1996 [knuth:1996]

ACKNOWLEDGMENTS

Many thanks to everyone who kept my balance during this challenge. In particular, I would like to thank my supervisors Christoph Busch, Marta Gomez-Barrero, and Andreas Heinemann for their lesson and guidance throughout my research. Special thanks to Marta Gomez-Barrero and Christoph Busch for trusting me and giving me the opportunity to grow professionally with them. An extra thanks to Marta for her daily supervision, for her dedication, for her time to improve my works, thank you very much! In addition, I would also like to thank my colleagues at das/sec for welcoming me in a cooperative and working environment. Thanks to Daniel, Jascha, Christian, Ulrich, Mathias, Marcel, Fabian, Juan, Jannis, Pia, Siri and Torsten for their warm welcome. I also thank my RESPECT colleagues from EURECOM; Max, José, Nick, and Madhu for all the collaborations during the project. Special thanks to my former colleagues at CENATAV, Airel, Chang, Aguado, and Katy, for their friendship and the good times we had together.

Last but not least, I would like to thank my family for their support. To my wife, Dailé, for her special encouragement at times when sadness gripped my heart. Many thanks to Daile for her endless love during these difficult years. Thanks to my parents for encouraging me to continue with this challenge. In short, a thousand thanks to life for giving me this great opportunity.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACER	Average Classification Error Rate. xix, 94, 95, 117
AE	Autoencoder. 13
AP	attack presentation. v, xiv, xvi, xvii, 2–4, 7, 10, 12, 14, 15, 17, 26, 29, 30, 32, 33, 36–39, 49, 54–63, 69, 75, 76, 78–81, 83, 88, 90–94, 96, 99, 101, 102, 105, 108, 113, 115–117
APCER	Attack Presentation Classification Error Rate. v, xvi, xix, 9, 43, 44, 51, 69, 72, 75, 76, 83, 94, 95, 111–113, 121, 122
ASV	Automatic Speaker Verification. 98, 101, 104
AUC	Area Under Curve. xix, 13, 71, 72
BCE	Binary Cross Entropy. 99, 108
BM	Boltzmann Machines. 15
BMM	Bernoulli Mixture Model. 24–26, 31
BoW	Bag of Words. xi, xiii, 23, 24, 26, 29, 31–34, 50, 51
BP	bona fide presentation. v, xiv, xvi, xvii, 3, 4, 10, 12–15, 17, 24, 26, 29, 30, 32, 33, 37–40, 43, 49, 50, 53, 55–57, 60–63, 69, 72, 75, 78–81, 92–94, 99, 101, 102, 105, 108, 112, 113, 115, 116
BPCER	Bona fide Presentation Classification Error Rate. v, xvi, xix, 9, 43, 44, 51, 55, 69, 72, 75, 76, 83, 85, 94, 95, 111–113, 117, 121, 122
BRIEF	Binary Robust Independent Elementary Features. xi, xxii, 18, 22, 23, 36, 51
BSIF	Binarized Statistical Image Features. xi, xii, xv, xvi, xviii, xix, 11, 18, 21, 22, 24, 32, 34, 36, 37, 40, 41, 44, 45, 47, 51, 53–55, 60, 61, 66, 68, 69, 72–78, 80, 81, 83, 85, 95, 98, 103–106, 112, 113, 118
CNN	Convolutional Neural Network. xii, xvii, xix, xxi, 5, 12–14, 51, 57, 58, 68, 73, 97, 99, 100, 104, 108, 110, 111, 113, 118
CQCC	Constant Q Cepstral Coefficients. 12, 98, 101, 103, 104, 106, 107
CQT	Constant-Q Transform. xvii, 12, 98, 100, 103, 104, 106, 107, 110, 112, 113

D-EER Detection Equal Error Rate. xiv, xvii–xix, 9, 13, 14, 33–42, 44–50, 52, 59, 66–77, 80–84, 86–90, 92, 95, 96, 103–113, 116–118, 122

DET Detection Error Trade-off. xiv–xvi, 46, 48, 72, 73, 75, 77, 83, 85

FAST Features from Accelerated Segment Test. xxi, 23

FCL Fully Connected Layer. 57, 60, 89, 99, 100

FMR False Match Rate. 75, 76, 122

FNMR False Non-Match Rate. 75, 76, 122

FR Face Recognition. 54, 55

FV Fisher Vector. xi–xiii, xv, xvii–xix, 24–26, 29, 31–35, 38–40, 49–51, 53, 61, 62, 66, 68, 69, 71–74, 76, 78–80, 92, 95, 97–99, 101, 103–106, 110–116, 118

GAN Generative Adversarial Network. 15, 118

GMM Gaussian Mixture Model. xv, 13–15, 24–26, 31, 34, 51, 54, 77–80, 93, 101, 112

HOG Histogram of Oriented Gradients. xi, xxii, 4, 11, 18–20, 24, 32, 34, 51

IQA Image Quality Assessment. 4, 59, 60

LBP Local Binary Pattern. xi, xxi, 4, 11, 14, 18, 20, 24, 32, 51, 60, 61, 81, 97, 98, 103, 106, 118

LFCC Linear Frequency Cepstral Coefficients. 12, 97, 101, 103, 104, 106, 107

LgA Logical Access. 101–113

LOO leave-one-out. xvi, 69, 72, 75, 84

LPQ Local Phase Quantization. xi, 4, 11, 21, 60, 61, 98, 103, 106

LSTM Long Short-Term Memory. 12, 99, 100

MB-LBP Multi-Scale Block LBP. xi, 20, 98, 103, 106

MCNN Multi-Channel CNN. 14

MFCC mel-frequency cepstral coefficients. 12

oFAST FAST keypoint orientation. 23

ORB	Oriented FAST and Rotated BRIEF. xi, 18, 23, 36, 41, 51
PAD	Presentation Attack Detection. v, xi, xii, xiv, xv, xviii, xix, 3–7, 9–17, 29, 31, 32, 36–38, 40, 44, 46, 50, 51, 53–57, 59–64, 66–69, 72, 75–77, 80–92, 94–101, 103, 104, 107, 109, 111, 113–119, 122
PAI	Presentation Attack Instrument. xiii, xvii, 1–8, 11–15, 18, 32, 38, 40, 43, 44, 49–53, 60, 69, 75, 77, 79, 84, 85, 101, 102, 104, 106, 107, 110, 116, 117, 122
PBX	Private Branch Exchange. 102
PCA	Principal Component Analysis. 25, 26, 67, 78, 116
PhA	Physical Access. 101–113
PHOG	pyramid HOG. 20
PSTN	Public Switched Telephone Network. 102
rBRIEF	rotation aware BRIEF. 23
SIFT	Scale Invariant Feature Transform. xi, xiii, xviii, 18, 19, 22, 24, 29, 31, 32, 34, 37, 40, 41, 45, 47, 48, 50, 51, 53, 116, 118
STD	Standard Deviation. 40, 66, 73, 80, 86–88, 96, 105, 111
STFT	Short-Time Fourier Transform. 12, 21, 97, 98, 100, 103, 104, 106, 107, 110, 113
SURF	Speed-Up Robust Features. xi, 18, 19, 24, 32, 40, 44, 45, 50, 51, 53, 69, 116, 118
SVM	Support Vector Machine. xiii–xv, 13, 26, 30–32, 54, 60, 68, 78, 79, 97–99
t-DCF	minimum normalised tandem Detection Cost Function. 104, 106, 107
TTS	text-to-speech synthesis. 8, 101
UMT	Universal Material Translator. 15
VAE	Variational Autoencoder. 15, 118
VC	Voice Conversion. 8, 97, 101
VLAD	Vector of Locally Aggregated Descriptors. xi, xiii, 26, 29, 31–34, 50, 51
VoIP	Voice-over-IP. 102

1 INTRODUCTION

Biometrics is the science of recognising a subject's identity from physical or behavioural attributes. [98]. Whereas biometric characteristics such as fingerprint, face, and iris are considered biological, gait, signature, and keystrokes are behavioural characteristics. The human voice, in turn, combines both biological and behavioural properties, as the ability to talk needs to be learnt.

Depending on the application context, biometric systems can operate in verification or identification mode [134]. Biometric verification is the process where the input probe is compared to one reference template (i..e, 1:1 comparison) to verify a claimed identity. In contrast, biometric identification compares the input probe to all references (i.e., 1:n comparison) in order to find the biometric identifier associated to the probe [134]. In recent years biometric systems have been steadily evolving. Several studies [98] have shown that the extensive development of biometric systems has increased security and accuracy in many applications such as border controls, financial transaction authentication, and mobile device unlocking. This is due to the fact that biometric characteristics such as fingerprint, face, iris, or voice offer a high discriminative capability (i.e., they are "unique") and cannot be forgotten or shared with other subjects [134].

In spite of their advantages, biometric systems are still vulnerable to different external attacks [167], as shown in Fig. 1.1. In particular, we focus on attacks on the capture device, known as "Attack presentations", which can be easily launched by any subject without having a vast expert knowledge. As a consequence of the wide development experienced by several social networks (e.g., Facebook, LinkedIn, Instagram, or YouTube) a non-authorised subject can learn from a video tutorial and create an artificial copy of our biometric characteristics, denoted as Presentation Attack Instrument (PAI). Thus, she/he could gain access to those unattended applications (e.g., remote authentication for automated payment - pay-by-face [57]) which do not require direct monitoring. Fig. 1.2 shows examples of PAIs which can be easily fabricated by any unauthorised subject to bypass biometric systems. The fabrication of a face PAI can be for instance carried out by downloading a target photo or video from any social media and then replicating them over a printed papersheet. Videos could be also replied directly over the biometric system capture device using an iPad. Furthermore, the free access to large-scale public databases together with the recent advances in the creation of very realistic fake contents or "Deep Fakes" also pose a serious threat to

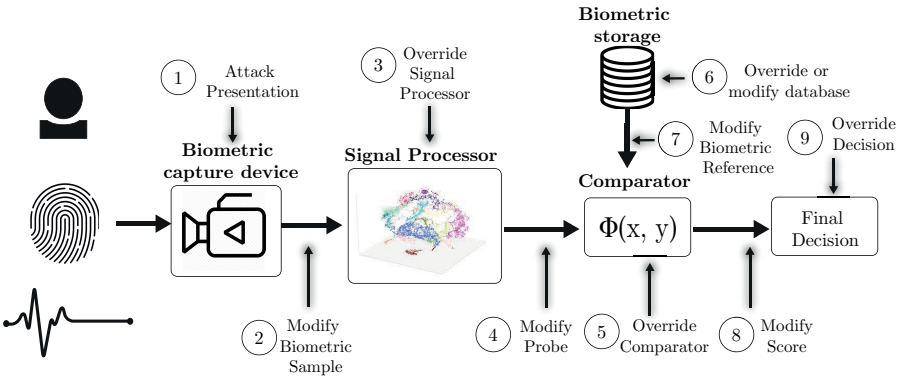


Figure 1.1: Attack points on biometric systems derived from [95]

Figure 1.2: Example of web-collected PAIs commonly launched over the capture device of a face biometric system.

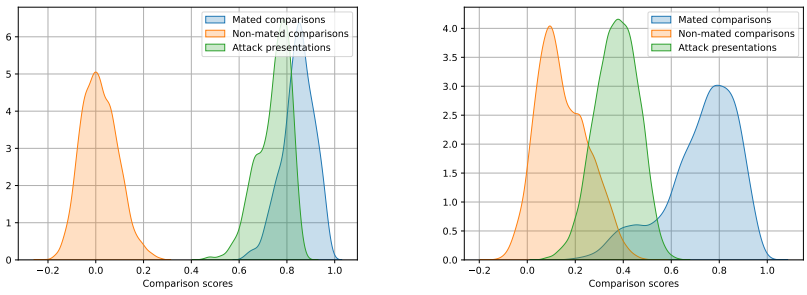
biometric systems [204]. Based on the intention of a malicious attacker, APs can be broadly categorised:

- *Impersonation attacks*: non-authorised subjects that create an artificial copy of the biometric characteristics to look like someone else and thus gain access through biometric systems.
- *Concealing attacks*: subjects that try to hide their own identity, e.g. by using make-up, to avoid detection by a biometric system (e.g. subjects in blacklists).

In our research, we focus on impersonation attacks, as they have proven to be a real threat to the security of current academic and commercial biometric systems. [184]).

1.1 MOTIVATION

The risk posed by PAIs is not only reduced to an academic issue. APs were addressed for the first time in 1998 [220]. Willis and Lee showed how four out of six evaluated biometric systems were vulnerable to PAIs. In 2000, Zwiesele *et al.* [237] conducted a comparative study on biometric identification systems which revealed the high vulnerability



(a) REPLAY-MOBILE database. (b) CSMAD-Mobile database.

Figure 1.3: Score distributions for mated comparisons, non-mated comparisons, and attack presentations.

of these systems to PAIs. Two years later, Matsumoto *et al.* [137] analysed the weakness of eleven commercial fingerprint-based biometric systems to gummy fingerprints. The experimental evaluation reported that 68% to 100% of the PAIs created with cooperative methods were falsely accepted as bona fide presentations (i.e., pristine). In 2009, Japan reported the use of PAIs in one of its airports, and in 2013, a Brazilian doctor used artificial silicone fingerprints to tamper a biometric attendance system at the Sao Paulo hospital [178]. One year later in 2014, a German hacker going by the name "Starbug" demonstrated how he was able to clone a fingerprint of the German defence minister, Ursula von der Leyen, using only publicly available photographs in which her hand was visible [58].

In order to confirm the above statements, we evaluate the risk of APs stemming from the REPLAY-MOBILE [35] and CSMAD-Mobile [166] databases to circumvent the ArcFace scheme [42]. This is one of the best performing open-source biometric recognition algorithm used for face recognition [42]. Fig. 1.3 reports the score distributions for *i*) comparisons between samples from the same subject (i.e., mated comparisons), *ii*) comparisons between samples from different subjects (i.e., non-mated comparisons), and *iii*) comparisons between AP and BP samples of the same subject (i.e., attack presentations). As it can be observed, a high percentage of AP scores (green distribution) overlaps with the score distribution of BP samples (blue distribution), thus confirming the need to equip biometric systems with a Presentation Attack Detection (PAD) module.

In order to prevent those security threats, several PAD techniques have been proposed in the literature [135]. They aim at determining whether a sample stems from a live subject (i. e., this is a BP) or from an artificial replica (i. e., this is an AP). Depending on how those PAD methods are integrated into a biometric system, they can be categorised as hardware- and software-based algorithms [135]. The former seek to spot PAIs by detecting the biological characteristics of the captured subject using a special sensor: electric resistance [45],

temperature [45, 189], and blood pressure [122], among others [46, 47, 62, 113].

In contrast to hardware-based approaches, software-based techniques are more interoperable, as they are not dependent on sensing devices and can therefore be deployed in many more applications than the former. These algorithms assume that properties in BP must be intrinsically opposed to the ones in APs mainly due to capturing properties or the materials used in the fabrication of the artificial replicas or PAIs. In this context, different strategies have emerged in the last decades: textural handcrafted features such as Local Binary Pattern (Local Binary Pattern (LBP)) [157, 221], Histogram of Oriented Gradients (Histogram of Oriented Gradients (HOG)) [2], and Local Phase Quantization (LPQ) [2, 165], frequency domain analysis [102, 126, 175], and Image Quality Assessment (IQA) [61, 62, 103]. More recently, the success of deep learning techniques in several academic and industrial fields has led to the development of more sophisticated PAD approaches which considerably outperform earlier PAD algorithms [33, 67, 175].

In spite of the aforementioned and other efforts, current PAD algorithms struggle to generalise well beyond the PAI species (i.e., attack types) on which they were trained. Specifically, the best performing deep learning-based techniques have reported a high detection performance for identifying PAIs when both PAI species and acquisition conditions are known a priori. However, there are still some issues to be resolved:

- *Poor generalisation capabilities for unknown PAI species:* Current state-of-the-art techniques face difficulties to detect unknown PAI species (i.e., APs created with a particular PAI species different from those in the training set), thereby resulting in a degradation of the detection accuracy.

- *Poor generalisation capabilities across several datasets:* Most state-of-the-art PAD methods show a decreasing detection performance when evaluated over a new database. Since capture devices might age and will eventually be replaced, PAD methods must be able to successfully classify samples acquired with a new capture device. Therefore, generalisation across multiple datasets is of utmost importance.

- *Specialisation on a particular biometric characteristic:* The most sophisticated PAD algorithms have been developed to detect PAIs through a particular type of biometric characteristic (e.g., face, fingerprint, or voice). Therefore, their application across different modalities (i.e., types) is not straightforward and could lead to wide performance deterioration.

- *Large number of hyperparameters:* Deep learning-based approaches are usually based on dense Convolutional Neural Networks (CNNs) having a large number of learnable parameters (exceeding 2.7 million [40]). Such models are not viable in mobile environments with limited resources.

1.2 RESEARCH QUESTIONS

In order to tackle the unresolved issues derived from the motivation, the following research questions are defined for this Thesis:

RQ 1 Keeping in mind that fingerprints consist of ridges and valleys, can the lack of ridge continuity be used to detect the artefacts produced in the fabrication of PAIs?
Is there a close relationship between the lack of ridge continuity and those artefacts?
Can these features aid in successfully detecting unknown PAI species?

RQ 2 Can different colour spaces unveil discriminative features to be capable of successfully detecting facial PAIs?
How can the facial artefacts, produced in the creation of PAIs, be perceived in different colour spaces?

RQ 3 What is the most appropriate facial region to identify PAIs?
Taking into account that the face consists of several regions such as the mouth, eyes, eyebrows, or chin, then how many facial regions are required to correctly identify a PAI.
What is the minimum or the optimal number of facial regions needed to detect PAIs?

RQ 4 Can the image resolution affect face PAD process?

1) Given that several lower, medium and high-resolution capture devices are employed for acquiring face images, how can the facial artefacts be detected in different image resolutions?
2) Keeping in mind that numerous lower, medium and higher resolution capture devices are employed for replay attacks, how can the image resolution of such devices affect or help the detection capability of PAD approaches?
3) How does the combination between replay and capture device resolutions affect the detection capability of PAD approaches?

RQ 5 Can a general framework be built to successfully detect known PAI species and unknown PAI species by generalising across different biometric characteristics?

1.3 THESIS OUTLINE

An overview of the contents covered in this Thesis is organised as follows:

- Chapter 1 introduces general concepts of biometrics and describes how risky attack presentations can be for the security of access controls. As a result, research questions are defined with the main focus on improving the generalisation capability of PAD.
- Chapter 2 summarises state-of-the-art PAD approaches for the three biometric characteristics investigated in this Thesis i. e., fingerprint, face, and voice. PAI species, metrics, evaluation scenario employed through this Thesis are defined.
- Chapter 3 describes the theory on which our research is based. In particular, several handcrafted approaches, as well as our proposed semantic common feature space for improving the PAD generalisation capability, are presented.
- Chapter 4 reports the evaluation of our common feature spaces for fingerprint PAD. In addition, it makes an analysis over PAI species used in the fabrication of PAIs and proposes different generalisable approaches focused on the definition of semantic common feature spaces. To answer the **RQ 1**, we summarise in this Chapter the results in [75, 77, 78].
- Chapter 5 evaluates the the best performing common feature space for facial PAD. The PAD performance of different facial regions is also explored. A comprehensive study about the impact of the image resolution variation for facial PAD is performed. In this Chapter, **RQ 2, 3**, and **4** are answered with the results in [70–73].
- Chapter 6 extends the applicability of the best performing common feature space for voice PAD, thereby answering the **RQ 5** [76, 79, 80]. In this Chapter, an analysis of several 1D-audio-waveforms-to-spectrogram transformations is performed. In addition, we propose a framework which exploits the image representations of spectrograms for voice PAD.
- Chapter 7 concludes the contributions of this Thesis by answering the research questions and highlighting open directions which emerged from our research.

2

RELATED WORK

This Chapter describes the state-of-the-art PAD techniques analysed in our Thesis (see Sect 2.3). Most PAD approaches are focused on improving the generalisation issues described in Chapter 1. In order to enhance the reader understanding, some general PAD concepts are defined below. In addition, metrics employed in the evaluation of PAD mechanisms as well as PAI species used in the fabrication of PAIs are summarised in this Chapter (see Sect. 2.1). Scenarios that are generally employed in the PAD assessment are also described in Sect. 2.2

The main definitions used throughout the Thesis compliant with the standard ISO/IEC 30107-1 [95] are introduced in the following:

- Bona fide presentation: "Interaction of the biometric capture subject and the biometric data capture subsystem in the fashion intended by the policy of the biometric system" [97]. A normal or pristine presentation.
- Attack presentation: "Presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system" [97]. An attack to the capture device to either conceal the own identity or impersonate someone else.
- PAI: "Biometric characteristic or object used in an AP" [95]. For instance, a replayed face photo, a gummy fingerprint, or a replayed speech.
- PAI species: "Class of presentation attack instruments created using a common production method and based on different biometric characteristics" [97]. Tab. 2.1 describes the main PAI species per biometric characteristic employed in the creation of PAIs. A complete overview about the fabrication of these PAI species per biometric characteristic and their impact on the biometric performance can be found in *i*) fingerprint [135], *ii*) face [62], and *iii*) voice [214].

Table 2.1: PAI species used in the fabrication or generation of PAIs.

	PAI species	Description
Fingerprint	Gelatin, Latex Silicone, PlayDoh Wood Glue, EcoFlex	Cooperative: the target subject cooperates in the fabrication of the PAI
	Silgum, Modasil Liquid EcoFlex RTV, Silicone Rubber	Non-cooperative: a latent fingerprint is digitalised, enhanced, and printed over a transparent film
Face	Cut	The face of the attacker is placed behind the hard copies of photos, where eyes have been cut out
	Printed	The attackers place their face behind the hard copies of high-resolution digital photographs
	Video-replay	The attackers replay face videos using tablets or smartphones
	Digital-replay/ Photo-replay	The attackers replay a face image using tablets or smartphones
	Silicone, Transparent Masks Half Mask, Papercraft Mannequin	The attackers create a 3D mask of the target face or use a Mannequin
	Funny Eye Paperglasses Partial Paper	The attackers waer a kind of glasses or the eye region from the target face
	Obfuscation Impersonation Cosmetic	The attackers apply makeup over the face to impersonate someone else or to hide the identity
Voice	Physical Access	The attackers record the target voice using a smartphone or other device.
	Logical Access	PAIs are generated using either a text-to-speech synthesis (TTS) or Voice Conversion (VC) technologies

2.1 STANDARDISED METRICS FOR THE EVALUATION OF PAD MECHANISMS

In order to establish a fair benchmark with the state-of-the-art PAD approaches, we follow the metrics defined in the international standard ISO/IEC 30107-3 for biometric PAD [97]:

- APCER: "Proportion of attack presentations using the same PAI species wrongly classified as bona fide presentations in a specific scenario" [97].
- BPCER: "Proportion of bona fide presentations misclassified as attack presentations in a specific scenario" [97].
- D-EER: "PAD operation point where APCER = BPCER".

Together with the above metrics, we report the BPCERs for several fixed operating thresholds:

- $BPCER_{10}$: BPCER at a fixed operation point APCER = 10%, i. e., 10/100 attack presentations are misclassified.
- $BPCER_{20}$: BPCER at a fixed operation point APCER = 5%, i. e., 5/100 attack presentations are misclassified.
- $BPCER_{100}$: BPCER at a fixed operation point APCER = 1%, i. e., 1/100 attack presentations is misclassified.

2.2 EVALUATION SCENARIOS

We focus on several scenarios commonly employed in the evaluation of PAD algorithms:

- Known PAI species: "scenario where an analysis of all PAI species is performed. In all cases, PAI species for testing are also included in the training set".
- Unknown PAI species: "scenario where PAI species used for testing are not incorporated in the training set". Depending on the biometric characteristic at hand, different protocols are followed in the experiments.
- Cross-database: "scenario where the capture device employed for the acquisition of test samples is different from the one used for capturing the training images. Both datasets contain the same PAI species to ensure that the performance degradation is due to the dataset change and not to the unknown PAI species".
- Cross-session: "scenario where different data collection sessions across different seasons or even years for the same capture device are used for training and testing".

Table 2.2: Summary of relevant studies focused on PAD generalisation per biometric characteristic. The results are reported in terms of D-EER(%).

	Method	Known PAI species	Unknown PAI species	Cross-database
Fingerprint	Rattani *et al.* [169]	-	19.70%	-
	Ding and Ross [44]	-	17.70%	-
	Nogueira *et al.* (VGG) [144]	3.87%	6.30%	30.70%
	Pala and Bhanu (TripleNet) [151]	2.41%	5.86%	15.20%
	Chugh and Jain (FSB-v1) [30]	1.70%	3.50%	18.90%
	Chugh and Jain (FSB-v2) [31]	1.11%	2.93%	17.91%
Face	CSURF + FV [16]	1.70%	-	25.80%
	Textural fusion [17]	2.43%	-	21.30%
	DeepPixelBis [65]	0.42%	5.97%	-
	DTN [128]	-	16.10%	-
	CDCN++ [30]	0.69%	11.95%	18.15%
	DR-UDA [31]	3.63%	-	17.93%
	TTN-S [218]	0.89%	8.00	11.55%
Voice	B01 [76]	9.78%	11.04%	-
	B02 [76]	11.96%	13.53%	-
	OneClassVoice [232]	0.20%	2.19%	-
	RW-Resnet [133]	-	2.98%	-

2.3 Presentation Attack Detection techniques

According to the international standard ISO/IEC-30107-3 [97], PAD aims of determining whether a sample stems from a live subject (i.e., it is a BP) or from an artificial replica (i.e., it is an AP). As mentioned in Chapter 1, PAD techniques can be broadly classified in two categories: Hardware- and Software-based. In the following, we describe PAD algorithms on the basis of the approach on which they are based. In Tab. 2.2 we also report some studies focused on PAD generalisation per biometric characteristic.

2.3.1 *Hardware-based Approaches*

Hardware-based techniques integrate an extra sensor into the capture device to detect the biological characteristics of a human body. Such living characteristics are, for instance, intrinsic properties (e.g., blood pressure [122], skin structure analysis through Optical Coherence Tomography (OCT) [11, 34, 39, 186, 187], electric resistance [45], reflectance [117, 215], or the combination of the two latter through impedance [113]), involuntary signals (e.g., thermal radiation stemming either from fingertips [45] or faces [189]), responses to external stimuli (e.g., motion estimation [115]), or articulatory gestures and oral airflow for speech [216, 230]). In general, those methods have reported a high detection performance to spot particular PAI species. However, the integration of an extra sensor can significantly increase the production cost (e.g., a thermal sensor for an iPhone exceeds EUR

250[1]). Moreover, their accuracy considerably decreases for the detection of unknown PAI species, as the sensing technology employed is designed for particular PAI species [62].

2.3.2 *Software-based Approaches*

Whereas hardware-based techniques are mostly expensive and not user-friendly (e.g., subjects are asked to make pressure in some fingerprint hardware-based mechanisms [135]), software-based approaches detect PAIs by analysing a single image or a set of frames acquired with the same capture device used for recognition purposes. They often provide high security, efficiency, and interoperability, thereby leading to a wide development in the last decade [62, 135, 179]. In the following, we introduce the categories on which those software-based PAD algorithms are based. These are handcrafted-based (Sect. 2.3.2.1), deep learning-based (Sect. 2.3.2.2), anomaly detection-based (Sect. 2.3.2.3), domain adaptation-based (Sect. 2.3.2.4), and generative models-based techniques (Sect. 2.3.2.5).

2.3.2.1 *Handcrafted-based Methods*

Depending on the biometric characteristic, several properties have been explored [135]. Skin distortions [6, 233] and perspiration produced by fingertip pores [43, 161] reported promising results one decade ago. However, for contact-based capturing approaches they depend on the pressure applied by subjects on the capture device surface during the acquisition process.

For facial PAD, numerous properties at different timeslot have been analysed: involuntary gestures such as eye-blinking [99, 114, 152, 156], face and head gestures (e.g., nodding, smiling, looking in different directions) [5, 14, 199]. Despite those and other efforts, these approaches fail to spot PAIs such as printed attacks whose eye region is replaced by the attacker's eyes. Furthermore, video-replay attacks cannot be successfully detected.

To compensate for such weaknesses, several studies have analysed texture properties. Handcrafted-based techniques commonly employ processing tools such as: Fourier Spectrum to describe the global frequency of images [102, 124, 126], Gaussian [235] or Gabor [196] filters to extract a particular frequency information, wavelet multiresolution analysis [1], statistical models to detect image noise [142], and traditional texture descriptors (e. g., LBP [28, 157, 221], HOG [48], BSIF [7], and LPQ [165]).

A common approach for handcrafted speech features is to decompose one-dimensional voice signals into many orthogonal or quasi-orthogonal signals that convert them into a two-dimensional sig-

1 https://amz.run/44Mp

nal [10]. These approaches include various techniques, of which the most popular are the Short-Time Fourier Transform (STFT) and mel-frequency cepstral coefficients (MFCC), which have great relevance in many tasks in audio processing [93, 109].

Numerous handcrafted methods in the literature attempt to capture the artefacts that give away artificial/replayed speech [174]. The Constant Q Cepstral Coefficients (CQCC) [202] is one of the most successful techniques. CQCC are implemented on the basis of the CQT [22], a perceptually-inspired alternative to Fourier-based approaches for time-frequency analysis. As reported in the literature, CQCC generalise across different databases (i.e. ASVspoof 2015 [200], ASVspoof [52], and RedDots replayed database [110]), resulting in performances close to the state-of-the-art in each case.

Alternative methods include those generating high-dimensional magnitude- and phase-based features, which have shown a good ability to discriminate between BPs and APs [26, 188]. Features extracted using linear sub-band processing were also explored, such as the Linear Frequency Cepstral Coefficients (LFCC) [194], which have been shown to detect APs with high accuracy in the recent ASVspoof 2019 [214]. The basic motivation behind sub-band processing is that artefacts of converted speech occur differently in different sub-bands.

2.3.2.2 *Deep learning Methods*

The great performance reported by CNNs on several pattern recognition applications has led to the development of several sophisticated algorithms for PAD. These techniques have reported a high detection performance which outperforms most of the aforementioned handcrafted-based methods. In 2014, Yang *et al.* [226] fine-tuned the ImageNet pretrained CaffeNet [101] and VGG-face [154] models for bi-class classification. Following this idea, Nogueira *et al.* [144] established a benchmark between three CNNs, achieving the best results in the LivDet 2015 competition with an overall accuracy of 95.5%. Pala and Bhanu [151] trained a triple-stream CNN fed with randomly patches extracted from images. Based on the fact that PAIs produces spurious minutiae on a fingerprint image, Chugh *et al.* [30, 31] proposed a framework for independently classifying minutiae-centred local patches extracted from a fingerprint image.

In the context of facial PAD, Xu *et al.* [222] combined Long Short-Term Memory (LSTM) units with CNNs to learn temporal features from face videos. The authors showed that the spatio-temporal features were helpful for facial PAD, thereby resulting in a reduction by half of the error rates reported by handcrafted feature baselines (5.93% vs. 10.00%). Keeping spatio-temporal features in mind, Gan *et al.* [63] proposed a 3D CNN for facial PAD, which, unlike traditional 2D CNNs, extracts the temporal and spatial dimension features from a frame sequence. Atoum *et al.* [9] also combined two-stream CNNs

for extracting local features and depth estimation maps from facial images.

Deep residual learning [85], successfully used for image processing tasks, was adopted for voice PAD. In particular, ResNet has been introduced to avoid vanishing/exploding gradients earlier in deep CNN architectures. This model is successfully used along with an image-like speech spectrogram (e.g., Mel spectrogram) for the detection of voice PAI [25, 205]. Recently, end-to-end approaches, which make use of the raw voice waveform, have also employed residual networks [133]. In this case, the weights of the first 1D convolution layer can be learnable [170] or fixed [195] and forced to have a *sinc* curve. In both cases, performances are comparable with the state-of-the-art [214, 223].

In spite of the advances achieved for handcrafted- and deep learning-based approaches, they still struggle to identify PAIs when, *i*) PAI species employed in the fabrication remain unknown in training (i.e., unknown PAI species) and *ii*) samples in training and testing sets are acquired with different capture devices under different acquisition conditions (i.e., cross-database), thereby leading to a generalisation ability decrease.

2.3.2.3 *Anomaly Detection-based Methods*

In order to overcome the above generalisation shortcomings, several anomaly detection-based PAD methods have been proposed. In 2013, de Freitas Pereira *et al.* [59] already reported poor generalisation capabilities of state-of-the-art face PAD methods to unknown PAI species. In fact, the error rates increased by at least 100% with respect to the evaluation of known PAI species. Motivated by those findings, Arashloo *et al.* [8] experimented over several unknown PAI species scenarios and concluded that anomaly detection approaches trained only on BP data can reach a detection performance comparable to the results attained by binary classification-based techniques. Those results were reported only in terms of the Area Under Curve (AUC), thus lacking a proper quantitative analysis compliant with the ISO/IEC 30107-3 standard on biometric PAD [97]. Following a similar idea, Rattani *et al.* [169] proposed an automatic adaptation of Weibull-calibrated SVMs and evaluated it over the LivDet 2011 database. The experimental assessment showed that D-EERs oscillated between 20 and 30% in the presence of unknown PAI species. On the other hand, Ding and Ross analysed an ensemble of one-class SVMs trained only on BP samples in [44], which lowered the error rates to 10-22% over the same generalisation task.

More recently, Nikisins *et al.* [143] showed how a one-class Gaussian Mixture Model (GMM) can outperform two-class classifiers depending on the PAI species included in the test set. Following the same anomaly detection paradigm, Xiong and AbdAlmageed studied in [221] the detection performance of one-class SVMs and Autoencoders (AEs) in

combination with LBP descriptors. In most of the scenarios tested, the detection rates increased with respect to common bi-class classifiers. Liu *et al.* also analysed in [128] the performance of a Deep Tree Network (DTN) by clustering the PAI species into semantic sub-groups. The experimental evaluation focused on unknown PAI species over the challenging SiW-M database [128], reported a mean D-EER of 16% which is considerably higher than those for known PAI species. Chingovska and Dos Anjos [27] explored the feasibility of client-specific information for facial PAD. Finally, George and Marcel [66] also combined a one-class GMM with a Multi-Channel CNN (MCNN), which fed with face samples acquired at different light spectra (i.e., RGB, thermal, and infrared). Although the experimental evaluation over the SiW-M dataset showed a performance improvement with respect to the DTN technique, its generalisation capability to detect unknown PAI species was still poor (i.e., a D-EER of 12.00%).

Anomaly detection for voice PAD was also explored in a one-class classification framework by Zhang *et al.* [232]. To avoid overfitting to known PAI species, the authors introduced two different margins in the softmax loss function for better modelling the BP speech and isolating the PAIs.

2.3.2.4 *Domain Adaptation-based Methods*

Generally, acquisition conditions such as appearance, illumination, or capture devices vary between datasets. In order to overcome poor cross-database generalisation issues, new PAD approaches have explored Domain Adaptation to transfer the knowledge learned from a source domain to a target domain [64]. By assuming that the relationship between BP and AP face samples on a given subject can be modelled with a linear transformation, Yang *et al.* [227] proposed a subject domain adaptation method to synthesise virtual features. Following this idea, Li *et al.* [125] transformed knowledge learned from a labelled source domain to an unlabelled target domain by minimising the Maximum Mean Discrepancy [130] for facial PAD. De Freitas Pereira [158] proposed a CNN-based method which builds a common feature space from face images, captured on different visual spectra domains, for improving face recognition. To transfer knowledge to the unlabelled target domain, Wang *et al.* [210, 211] proposed an unsupervised domain adaptation with disentangled representation, which builds a feature space shared between the source and target domains. Even if this common feature space appeared to be suitable to overcome cross-database issues, experimental results showed a poor detection performance over known PAI species scenarios (i.e., D-EERs of 3.20%, 6.00%, and 7.20% for CASIA Face Anti-spoofing [236], MSU-MFSD [219], and Rose-Youtu [125] databases, respectively). Other domain adaptation approaches for face PAD can be summarised in [118, 217].

Regarding domain adaptation for fingerprint PAD, Gajawada *et al.* tried to tackle the dependency on the PAIs contained in the training set from a different perspective in [60]. They propose a so-called deep learning-based "Universal Material Translator (UMT)". Given a reduced number (e.g., five) of known AP samples, the UMT generated synthetic PAI samples by embedding the main appearance features of those PAIs with known BP samples. Those synthetic samples were reutilised for training its detector, thereby resulting in an improvement up to 17% over the baseline. Despite these promising results, it should be noted that this approach does require some PAI samples (i.e., five) which should be carefully selected.

Domain adversarial training for voice PAD has been explored in [212]. In this paper, the authors treated the cross-database scenario as a domain-mismatch problem and addressed it using a domain adversarial training framework. The same authors further proposed a dual-adversarial domain adaptation [213] framework to enable fine-grained alignment of APs and BPs separately by using two domain discriminators.

2.3.2.5 *Generative-based Methods*

Nowadays, generative models are the vanguard of unsupervised learning. Techniques such as GMM [138], Boltzmann Machines (BM) [53], Variational Autoencoder (VAE) [108], and Generative Adversarial Network (GAN) [83] have been successfully applied in numerous computer vision [119], speech recognition and generation [87], and natural language [36, 112] tasks. Those algorithms try to capture the inner data probabilistic distribution to generate new similar data [150]. However, to the best of our knowledge, a rather limited number of works has been employed for PAD. Engelsma and Jain [51] fed several GANs with BP samples acquired by a RaspiReader fingerprint capture device. The experimental results for high-security threshold over unknown PAI species showed a detection performance very sensitive to the training set.

In voice PAD, generative models have been purposely used to augment the data in the training phase. Recently, Wang *et al.* [232] have proposed a vocoder replay channel response estimation based on MelGAN [116] and HifiGAN [121] on the ASVspoof 2021 [223], the results of which showed a good generalisation ability.

2.4 SUMMARY

Recently, PAD has been an active research field. In spite of the efforts achieved, current techniques still lack high generalisation capability to detect challenging unknown PAI species under different scenarios. They also decrease their detection performance when are deployed on biometric systems with a capture device different to those employed

for capturing their training samples (i.e., cross-database scenarios). The performance degradation can be observed in Tab. 2.2. Based on this fact, we focus our research on improving the generalisation capabilities of the PAD module under these previous scenarios.

3

SEMANTIC COMMON FEATURE SPACE

Malicious attackers live with us and can launch frequent unknown PAI species against the biometric system's capture device under different conditions to circumvent their security. Therefore, the development of generalisable PAD approaches, which can be successfully employed for several types of biometric characteristics, is of utmost importance for the research community. In this Chapter, we summarise the main theory on which our Thesis is based.

Based on the assumption that unknown PAI species share homogeneous properties with known PAI species and heterogeneous with BPs, we define generalisable common feature spaces which can be successfully combined with discriminative models for PAD through different biometric characteristics, as shown in Fig. 3.1. The proposed algorithms allow the definition of semantic sub-groups constructed from the known PAI species which are observed on unknown PAI species. Thus, the generalisation of the PAD module can be improved. To demonstrate the feasibility of our generalisable approaches, they are then evaluated over three types of biometric characteristics namely fingerprint (see Chapter 4), face (see Chapter 5), and voice (see Chapter 6), which are different to each other and can be captured with a smartphone.

Our generalisable PAD techniques are based on four main steps: *i*) features (Sect. 3.1) are extracted from a regular grid of points (Sect. 3.1.1) along the whole input biometric sample (i.e., fingerprint, face, and voice); *ii*) a generalisable common feature space is built by the definition of semantic sub-groups from the aforementioned features (Sect. 3.2); *iii*) the final descriptor, which represents the biometric sample at hand and emphasises the AP related properties, is subsequently transformed to a new feature space based on the learned semantic common feature space; and *iv*) a BP or AP decision is finally taken by a discriminative model (Sect. 3.3).

3.1 HANDCRAFTED DESCRIPTORS

As mentioned in Chapter. 1, we focus on three biometric characteristics, namely fingerprint, face, and voice, each of which has intrinsic properties that make them different from each other. Whereas fingerprint comprises mainly ridges and valleys [98], facial images are composed of facial aesthetic units [81]. In addition, voice data are generally depicted by time-domain, frequency-domain, or time-frequency-domain representations known as spectrograms [76]. Therefore, the creation

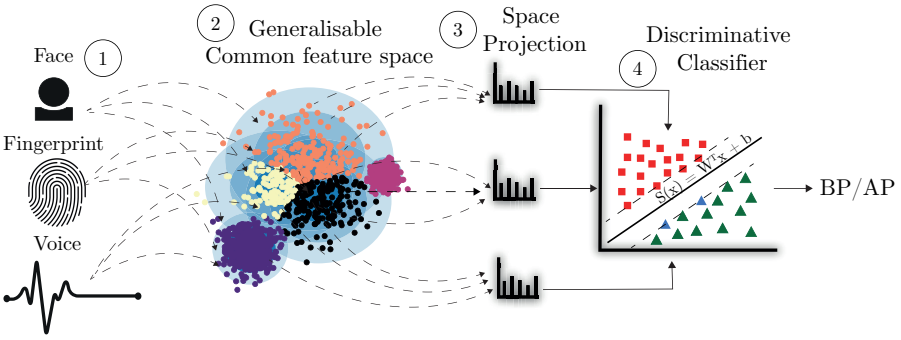


Figure 3.1: General overview of our generalisable common feature space-based approaches.

of a single set of universal features to represent those characteristics would lead to low biometric performance. In this context, we explore in the Thesis several continuous and binary descriptors, which are briefly described in this section. In particular, we have considered *i*) gradient- (SIFT, Speed-Up Robust Features (SURF), HOG), *ii*) intensity difference- (Binary Robust Independent Elementary Features (BRIEF), Oriented FAST and Rotated BRIEF (ORB)), and *iii*) texture-based (LBP, BSIF) features. The reason behind choosing not only continuous but also binary descriptors lies in their higher efficiency at the cost of a small performance loss for other tasks.

3.1.1 *Dense Multi-scale Features*

Since artefacts produced in the fabrication of PAIs might be located in any area of the input image, we follow in our approaches the strategies in [15] for the feature computation. Therefore, local descriptors are densely extracted at fixed points on a regular grid with an uniform spacing (e.g., 3 pixels). In addition, those artefacts might have different sizes. Hence, descriptors are computed over four circular patches with different pixel radii $\sigma = \{4, 6, 8, 10\}$. Thus, each point in the grid is represented by four descriptors, as depicted in Fig. 3.2.

3.1.2 *Scale Invariant Feature Transform*

SIFT [131] is one of the most popular histogram-based descriptors due to its robustness to changes in scale, translation, rotation, and other imaging parameters. In addition, the SIFT descriptor has shown to provide robust recognition capabilities across different affine distortions, changes in 3D viewpoints, addition of noise, and illumination changes. This method involves four stages to generate the set of image features: *i)* scale-space extrema detection, *ii)* keypoint localization, *iii)* orientation assignment, and *iv)* keypoint descriptor. In our investigation, we utilise steps three and four in our implementation, as keypoints are

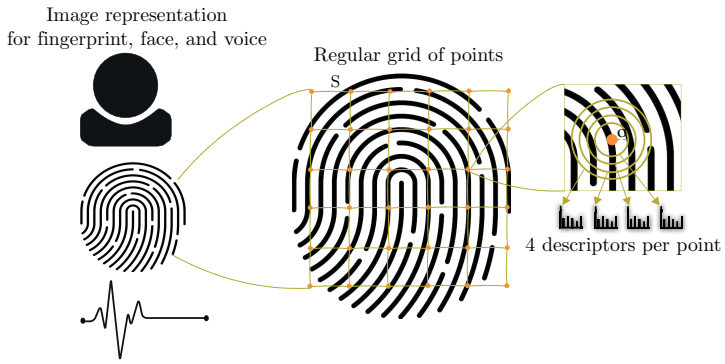


Figure 3.2: Feature extraction over the biometric samples. Dense multi-scale features are computed for face and voice data, as it is shown exemplary for fingerprint.

fixed over a regular grid along with the whole biometric image (see Sect. 3.1.1). To efficiently compute SIFT descriptors, we use the implementation provided in [207], which delivers a speed-up of up to 60x by exploiting the uniform sampling and overlapping between descriptors and using linear interpolation with integral image convolution.

3.1.3 *Speed-Up Robust Features*

SURF [12] is a keypoint-based descriptor, like SIFT, which employs the Haar wavelet transform to approximate the image gradient. In particular, SURF computes the first order Haar wavelet responses in the x and y directions at the orientation assignment step. Similarly to SIFT descriptors, the area around the interest keypoint is subsequently divided into $4\times$ 4 sub-regions, and the Haar wavelet responses are computed and L_2 normalised. The final feature vector is the concatenation of the accumulated wavelet responses in each direction and the summation of its absolute values, thus leading to a 64-dimensional vector per keypoint. In our methodology, we selected the 128-dimensional variant, which also includes the first Haar wavelet responses in diagonal directions.

3.1.4 *Histogram of Oriented Gradients*

HOG [38] is a local image descriptor capturing the intensity gradients and edge directions to describe the shape and appearance of an object within an image. As the previous descriptors, the HOG features are computed over localised cells. Therefore, it is invariant to geometric and photometric transformations. In this particular case, the cells comprise usually 8×8 pixels, and a histogram of edges orientation within that cell is computed. Afterwards, cell blocks of 16×16 pixels are normalised, in order to provide better illumination invariance. In our implementation, we used a multi-scale HOG extension named

pyramid HOG (PHOG), which has reported good results in static facial expression analysis [37] and fingerprint PAD [49]. In this case, the gradient is joined at several pyramid levels, and a histogram is computed for each grid.

3.1.5 *Local Binary Pattern*

LBP [145] is a texture descriptor originally developed for the analysis of two-dimensional texture, which has obtained excellent results in multiple tasks. It is invariant to rotation, illumination, and orientation changes. More specifically, it represents an image with a histogram of uniform patterns corresponding to micro-features in the image. These histograms allow capturing both shape and textural features from an image. In our methodology, a multi-resolution analysis is included, by computing the aforementioned histograms on different window sizes. In more detail, let X be a circular image patch with radii σ and S pixels around the centre. Then, the LBP descriptor is defined as:

$$\mathrm{LBP}_{S,\sigma} = \sum_{i=0}^{S-1} f(g_i - g_c)2^i, \tag{3.1}$$

where g_i with $i = 0 \dots S-1$ are gray intensity values around the center g_c in the image patch. $f(g_i - g_c g_i - g_c)$ is defined as:

$$f(g_i - g_c) = \begin{cases} 1, & g_i - g_c \geq 0 \\ 0, & g_i - g_c < 0 \end{cases} \tag{3.2}$$

In order to capture more information and thereby increase the descriptor distinctiveness, we compute several LBP patterns by combining various radii σ. The LBP histograms are subsequently built from those patterns at different scales by varying the window size and sliding over the whole image. Finally, the computationally efficient implementation provided in [182] is used.

3.1.6 *Multi-Scale Block LBP*

Multi-Scale Block LBP (MB-LBP) [231] encodes the intensities of rectangular regions with the LBP operator, which allows describing several local structures of an image. Whereas the LBP descriptor is defined for each pixel by thresholding the 3×3 neighbourhood pixel values with the centre pixel value, the MB-LBP operator represents each pixel x by comparing the central rectangle average intensity g_x with those of its neighbourhood rectangles. Therefore, it can detect numerous image structures such as lines, edges, spots, flat areas, and corners [231], at different scales and locations. Unlike LBP, the MB-LBP descriptor can

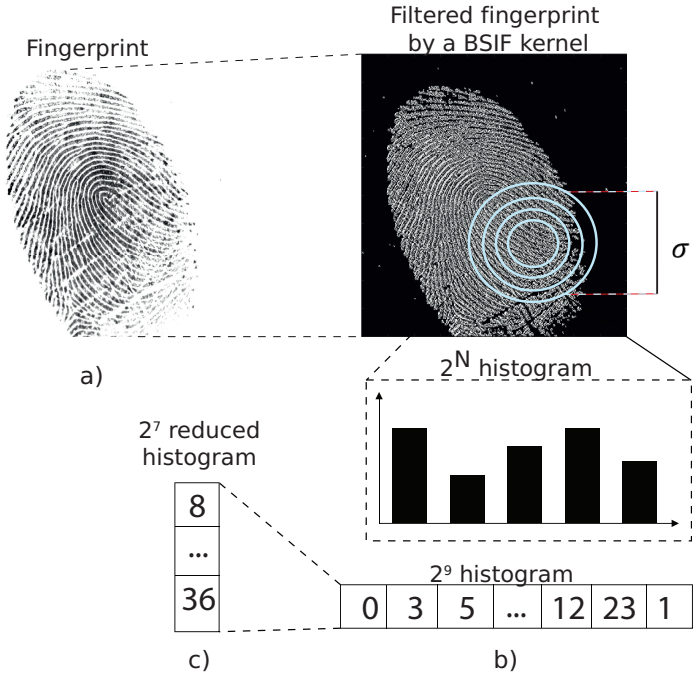


Figure 3.3: BSIF descriptors computed from $N = 9$ filters of size $l = 5$. *a*) fingerprint image, *b*) BSIF histograms, computed densely at fixed points on a regular grid, with a fixed stride S for four local patches with different window size, and *c*) a reduced BSIF histogram.

thus capture large scale structures that may be the dominant features of images, with 256 binary patterns. In our work, we compute the MB-LBP descriptor for several rectangle sizes $R_x = \{3, 5, 7, 9\}$.

3.1.7 *Local Phase Quantization*

LPQ [146] is a texture descriptor designed to deal with blurred images. It represents an image patch of size $l \times l$ centred on a pixel x as a 256-histogram by using the local phase information, extracted by a STFT. Let $F_{u_{i=1..4}}$ be the outputs of the STFT for the pixel x using four bi-dimensional spatial frequency u_0, u_1, u_2 and u_3, the LPQ features for x are defined as a vector whose components are formed by stacking the real and imaginary part of $F_{u_{i=1..4}}$. Subsequently, the vector elements are quantized using a previously defined function and then represented as a integer value in the range $[0 \ldots 255]$. In order to make the LPQ coefficients statistically independent, a decorrelation step based on whitening transform was performed.

3.1.8 *Binarized Statistical Image Features*

BSIF [104] is a local image descriptor computed by binarising the responses of a given image to a set of pre-learned filters to obtain a

statistically meaningful representation of the data. In particular, let X be an image patch of size $l \times l$ and $W = \{W_1, \ldots, W_N\}$ a set of linear filters of the same size as X. Then, we compute binarised responses b_n:

$$b_n = \begin{cases} 1 & \sum_{u,v} W_n(u,v)X(u,v) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

All the filter responses b_n are subsequently stacked to form a bit string $\mathbf{b}$ with size N for each pixel. Subsequently, $\mathbf{b}$ is transformed to a decimal value, and then a 2^N histogram for X is computed. In our work, 60 filter sets with different sizes $l = \{3, 5, 7, 9, 11, 13, 15, 17\}$ and number of filters $N = \{5, 6, 7, 8, 9, 10, 11, 12\}$ were obtained from [104].

Like the SIFT computation, the BSIF histograms are densely extracted over a regular grid with a fixed stride S of 3 pixels, and for each point on the grid, histograms are computed over four circular patches σ, as depicted in Fig. 3.3b). Therefore, each point in the grid is represented by four BSIF histograms. In our implementation, we followed the BSIF reduction strategy described in Sect. 5.1.1 which represents each 2^N BSIF histogram as a 128-component vector (see Fig. 3.3-c).

3.1.9 *Binary Robust Independent Elementary Features*

BRIEF [24] is a binary noise-resistant local descriptor, whose computation time is two orders of magnitude faster than SIFT. This is achieved by exploiting the fact that image patches can be efficiently classified on the basis of a relative small number of pairwise intensity comparisons τ. Thus, the BRIEF binary descriptor represents a smoothed patch like a bit string constructed from a set of binary intensity tests. More specifically, let X be a square smoothed image patch, then a binary test τ can be defined as:

$$\tau(X; x, y) = \begin{cases} 1 & \text{if } X(x) < X(y) \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

where x and y are locations in X, and $X(x)$ is the gray value of X at x. Previous locations are randomly pre-fixed according to a Gaussian distribution around the patch centre. Finally, by using a set of η binary tests, we can obtain a $\eta-$bitstring as follows:

$$f_\eta X = \sum_{i=0}^{\eta-1} \tau(X; x_i, y_i)2^i. \tag{3.5}$$

In our implementation, we select $\eta = 256$, since it has shown a better trade-off between effectiveness and efficiency in many real applications [136].

3.1.10 *Oriented FAST and Rotated BRIEF*

ORB [173] is a binary descriptor built upon BRIEF [24] and Features from Accelerated Segment Test (FAST) [172], which additionally provides rotation invariance. The algorithm starts by detecting FAST points in the image, at different scale pyramid levels, and by adding an effective measure of corner orientation, to conform the final FAST keypoint orientation (oFAST) features. Then, a rotation aware BRIEF (rBRIEF) descriptor is computed and combined with oFAST to obtain the final ORB descriptor.

In more details, rBRIEF first steers the BRIEF descriptor according to the orientation of the keypoints, θ. To that end, rBRIEF discretises θ to increments of $2\pi/30$ (12 degrees), and constructs a lookup table of precomputed BRIEF patterns, thereby obtaining rotation-invariant features in an efficient manner. However, steering BRIEF leads to a loss of variance in the responses, and thus to less discriminative features. In addition, both BRIEF and its steered version show some correlation in the tests. To tackle these issues, ORB runs a greedy search among all possible binary tests to find the ones that have both high variance and means close to 0.5, as well as being uncorrelated.

3.2 COMMON FEATURE SPACE REPRESENTATIONS

3.2.1 *Bag of Words*

This technique was first developed for text categorization tasks, in which a text document is assigned to one or more categories based on its content [129]. To that end, Bag of Words (BoW) represents a text document by a sparse histogram of word occurrence based on a visual vocabulary. Following this idea, Csurka *et al.* [37] adopted and applied this method to represent local features from an image in terms of the so-called "visual words". Our common feature space is built upon this approach.

As proposed in [74], the BoW representation first computes the visual vocabulary as a codebook with K different centroids or visual words by k-means clustering. Then, the BoW is defined as the histogram of the number of image local descriptors assigned to each visual word. Its computation is summarised in Fig. 3.4. An m-level pyramid of spatial histograms is used in order to incorporate spatial relationships between patches. For that purpose, the fingerprint image is partitioned into increasingly fine sub-regions, and the feature descriptors inside each sub-region are assigned to the closest centroid among the K visual words, using a fast version of k-means clustering [50]. Subsequently, the histograms inside each sub-region are computed and transformed into a single and final feature vector by a homogeneous kernel map [208].

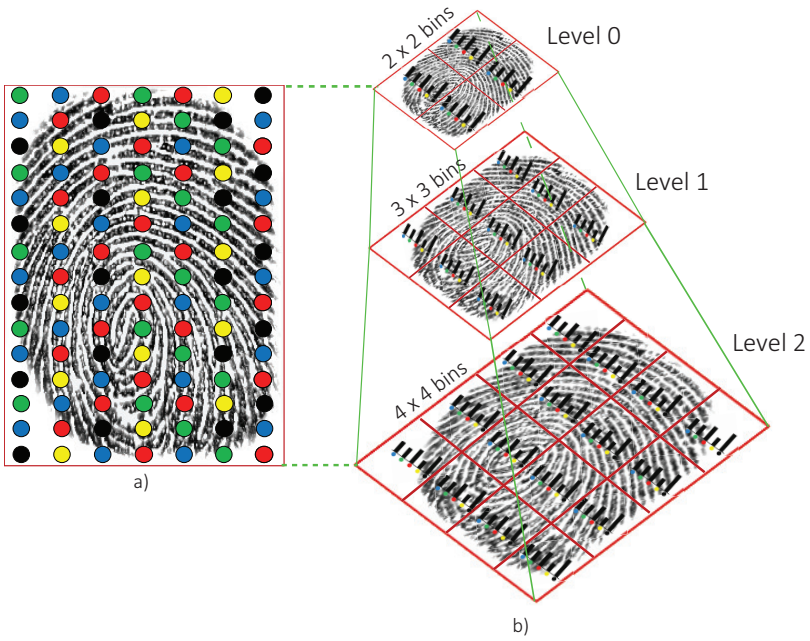


Figure 3.4: Example of pyramid of spatial histograms. a) Quantized features using k-means. b) 3-level pyramid of spatial histograms built from quantized features.

3.2.2 *Fisher Vector*

BoW approaches encode local features using a *hard assignment*, in which a local descriptor is only assigned to one visual word based on a similarity function. In contrast, the Fisher Vector (FV) method derives a kernel from a generative model of the data (e.g., GMM [176] or Bernoulli Mixture Model (BMM) [206]), and describes how the distribution of a set of local descriptors, extracted from unknown PAI species, differs from the known PAI species distribution previously learned by the adopted generative model [176]. The aforementioned generative model can be understood as a *probabilistic* visual vocabulary, thereby allowing a *soft assignment*. Thus, the FV paradigm encodes not only the number of descriptors assigned to each region but also their position in terms of their deviation with respect to the predefined model. Therefore, the final transformed features are more robust to new samples, which may stem from unknown scenarios and thus differ from the samples used for training.

As proposed in [159], we train a GMM model with diagonal covariances from local continuous features (e.g., SIFT, LBP, BSIF, HOG, SURF) extracted on one previous step. In particular, a GMM on K-components, which is represented by their mixture weights (π_k), means (μ_k), and covariance matrices (σ_k), with $k = 1, \ldots, K$, allows discovering semantic sub-groups from known PAI speciess and BP samples, which could successfully enhance the detection of unknown PAI species. In order to build those semantic groups, the local de-

scriptors are first decorrelated using Principal Component Analysis (PCA) [100], hence reducing their size to $d = 64$ components while retaining 95% of the system variance. Then, the FV representation which captures the average statistics first-order and second-order differences between the local features and each semantic sub-groups previously learnt by the GMM is computed [183].

Let $\mathbf{X}$ be a local descriptor of size d and $S_K = \{(\pi_k, \mu_k, \sigma_k) : k = 1 \dots K\}$ a set of K semantic sub-groups learnt by the GMM. The FV representation for $\mathbf{X}$ is defined as the conditional probability:

$$FV_X = P(X|S_K) \tag{3.6}$$
$$= P(X|\mu_k, \sigma_k) \tag{3.7}$$

By applying Bayesian properties, we can rewrite the previous equation as:

$$\phi_k^1 = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{d} \alpha_i(k) \left(\frac{X_i - \mu_k}{\sigma_k} \right), \tag{3.8}$$
$$\phi_k^2 = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{d} \alpha_i(k) \left(\frac{(X_i - \mu_k)^2}{\sigma_k^2} - 1 \right), \tag{3.9}$$

where $\alpha_i(k)$ is the soft assignment weight or the posterior probability of the i-th feature $\mathbf{X}_i$ to the k-th Gaussian [183]. Therefore, the FV representation that defines a fingerprint image is finally obtained by stacking the differences: $\phi = \left[\phi_1^1, \phi_1^2, \dots, \phi_K^1, \phi_K^2\right]$, thereby resulting a $2 \cdot d \cdot K = 2 \cdot 64 \cdot K$ size vector.

On the other hand, for encoding binary features we train a BMM, whose K-components are represented by the mixture weights (π_B^k) and means (μ_B^k), with $k = 1 \dots K$ [206]. Therefore, a closed-form approximation of FV representation is computed as follows:

$$\phi_{\mu_{kd}} = \left(\frac{1}{T} \sum_{t=1}^{T} \gamma_k(x_t) \frac{(-1)^{1-x_t^d}}{\mu_{kd}^{x_t^d}(1-\mu_{kd})^{1-x_t^d}} \right) F_{kd}^{-\frac{1}{2}}, \tag{3.10}$$

where

$$\gamma_k(x_t) = \frac{\pi_k p_k(x_t|\theta|)}{\sum_{k=1}^{K} w_\pi p_k(x_t|\theta|)} \tag{3.11}$$
$$F_{kd} = T\pi_k \left(\frac{\sum_{k=1}^{K} \pi_k \mu_{kd}}{\mu_{kd}^2} + \frac{\sum_{k=1}^{K} \pi_k (1-\mu_{kd})}{(1-\mu_{kd})^2} \right) \tag{3.12}$$

It is worth noting that the FV representation based on BMM approach only takes into account the gradients with respect to μ_{kd}. Therefore, the KD-dimensional FV representation of a fingerprint sample is defined as $\phi_B = \left[\phi_{\mu_{kd}}\right], k = 1, \dots, K$ and $d = 1, \dots, D$

Finally, the FV representation based on BMM yields a compact vector, whose size Kd is the half of FV encoding built upon GMM approach. In addition, BMM, unlike GMM, does not require data decorrelation (i.e., PCA to the extracted local features is not applied).

3.2.3 *Vector of Locally Aggregated Descriptors*

In order to reduce the high-dimension image representation proposed by the FV and BoW approaches, gaining in efficiency and memory usage, we have finally studied the VLAD methodology [100]. This is a simplified non-probabilistic version of FV, which models the data distribution from the accumulative distances between a local descriptor $\mathbf{X}$ and its closest visual word $\mathbf{c}$ in the visual vocabulary. Therefore, as in the BoW approach, a visual vocabulary needs to be computed in the first step with the k-means algorithm.

In particular, a d-dimensional local feature descriptor $\mathbf{X}$ can be represented by a VLAD descriptor $\mathbf{V_X}$ of size Kd as follows:

$$\mathbf{V_X} = \sum_{j=1}^{d} \left(\sum_{X:NN(X)=c_i} X_j - c_{i,j} \right), \tag{3.13}$$

where $\mathbf{X}_j$ and $c_{i,j}$ denote the j-th component of $\mathbf{X}$, and its corresponding closest visual word $\mathbf{c_i}$. In our method, $\mathbf{V_X}$ is subsequently L_2-normalised in order to further improve the classification accuracy. Similar to FV, VLAD also applies PCA over data for their decorrelation.

3.3 DISCRIMINATIVE MODELS

For the final decision, separated linear SVMs are employed to classify the final features extracted with our approaches (see Sect. 3.2). SVMs are popular since they perform well in high-dimensional spaces provided by the above feature representations, avoid over-fitting, and have good generalisation capabilities. According to [89], when the feature's dimensionality is so greater than the number of instances employed for training, a non-linear mapping does not improve the performance. Therefore, the use of a linear kernel would be good enough to achieve a high classification accuracy.

In order to find the optimal hyperplane separating BPs from APs, the optimisation algorithm bounds the loss from below. Therefore, we have trained a linear SVM as follows: The SVM labels the BP samples as +1 and the APs as -1, thereby yielding the corresponding $\mathbf{W}'$ (weights) and $\mathbf{b}'$ (bias) classifier parameters.

Subsequently, given a feature descriptor $\mathbf{x}$ which was previously yielded by a particular encoding approach, the final score $s_\mathbf{x}$, which estimates the class of the sample at hand, is computed as the confidence

of such decision (i.e., the absolute value of the score is the distance to the hyperplane):

$$s_{\mathbf{x}} = \mathbf{W}' \cdot \mathbf{x} + \mathbf{b}' \tag{3.14}$$

4

FINGERPRINT PRESENTATION ATTACK DETECTION

Forensic investigators have used fingerprints for personal identification for many decades, thus reporting a very high biometric performance [98]. A fingerprint image comprises ridges and valleys [98], as shown in Fig. 4.1. Ridges are represented as dark lines whereas valleys are bright lines. Generally, different features may be analysed in a fingerprint sample:

- *Level 1 (Global features)* which consists of dense singular points and the main ridge orientation including the arch, tented arch, left loop, right loop, and whorl.
- *Level 2 (Local features)* which comprise dense minutiae details such as ridge ending and bifurcation.
- *Level 3 (Fine features)* which include concrete details of ridges such as their width, shapes, contours, and strength of sweat pores.

In order to address **RQ 1**, we explore in this Chapter the three common feature spaces together with several handcrafted descriptors described in Chapter 3. Those handcrafted techniques allow the description of fingerprint properties such as lack of ridge's continuity, texture changes, and grey intensity differences, which might differ between a BP and an AP. In order to build a robust and generalisable semantic common feature space, we combine the three approaches (i.e., FV, BoW, and VLAD) with the best performing descriptor (Sect. 4.1); this is named *Space Fusion*. In addition, the best performing descriptors are fused with the best performing common feature space for a sturdy fingerprint representation (Sect. 4.2); this is named *Descriptor Fusion*. In this Chapter, we summarise the results in [72, 75, 77].

4.1 SCORE LEVEL FUSION OF COMMON FEATURE SPACES

In a first approach, we explore the above common feature space (i.e., BoW, VLAD, and FV) in combination with SIFT. Fig. 4.2 shows an overview of the proposed PAD approach. In the first common processing step, SIFT are densely extracted from the whole input image, as indicated in Sect. 3.1.1. Subsequently, the three image representations are applied to transform the local descriptors into a common feature space: *i*) BoW (Sect. 3.2.1), *ii*) FV (Sect. 3.2.2), and *iii*) VLAD

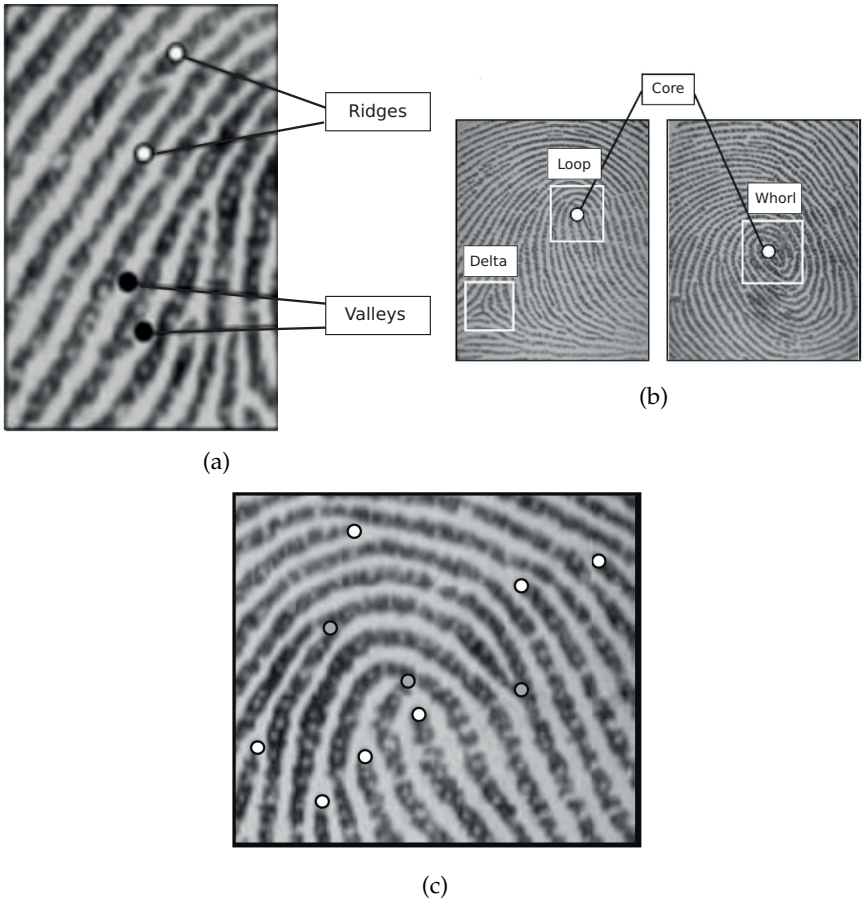


Figure 4.1: *a*) Ridges and Valleys in a fingerprint, *b*) singular regions over the ridge orientation, and *c*) termination and bifurcation minutiae. Images were taken from [98].

(Sect. 3.2.3). Finally, the BP vs. AP decision for a sample at hand is taken by a linear SVM (see Sect. 3.3).

Given that the use of complementary information could improve the detection capabilities of an approach, we also evaluate the fusion between the three proposed representations using a weighted sum method as follows:

$$s_f = \alpha \cdot s_1 + \beta \cdot s_2 + (1 - \alpha - \beta) \cdot s_3, \tag{4.1}$$

where $\alpha + \beta \leq 1$, and s_1, s_2 and s_3 represent the individual scores produced by our three representations. Taking into account that LivDet databases do not include a validation set, the α and β weighted values are computed from each LivDet's training set.

4.2 SCORE LEVEL FUSION OF DIFFERENT DESCRIPTORS

In a second approach, we explore the combination of descriptors described in Sect. 3.1 with the best performing common feature space

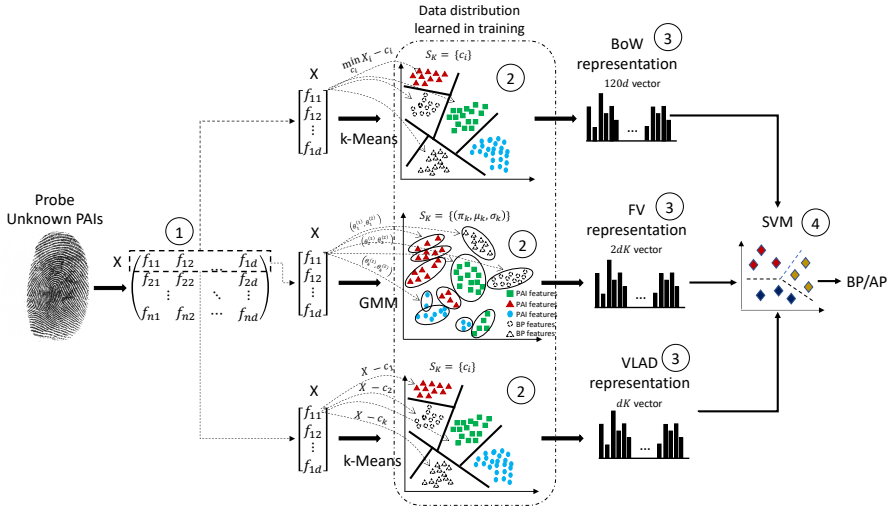


Figure 4.2: Overview of the Space Fusion-based approach. First, SIFT descriptors are densely computed at different scales over the whole input image. These features are subsequently encoded using a previously learned common feature space by means of three different approaches: *a*) BoW, *b*) FV, and *c*) VLAD. The fingerprint descriptor per representation is separately classified using a linear SVM and then combined by a score-level fusion.

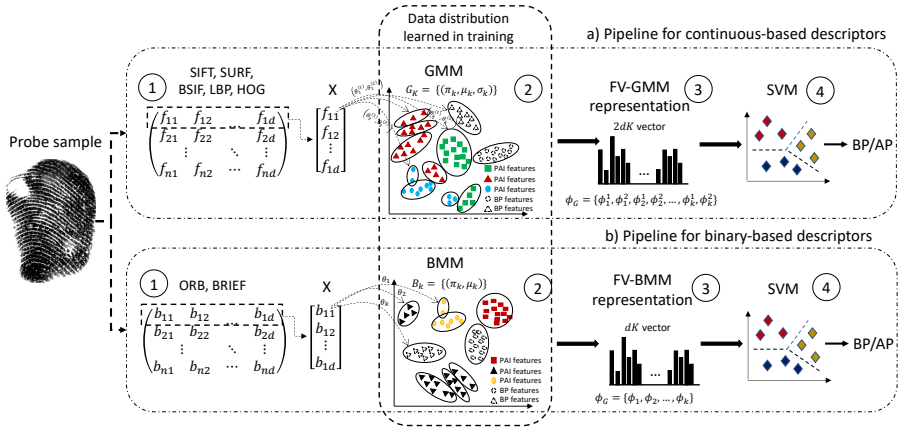


Figure 4.3: Overview of the descriptor fusion-based scheme, which consists of four steps. First, local features are densely computed at different scales. These features are subsequently encoded using a previously learned common feature space. The fingerprint descriptor is classified using a linear SVM. *a*) it refers to the particular pipeline used for continuous-based descriptors, and *b*) it represents the PAD overview for binary-based descriptors. Finally, the SVM outputs for the best performing descriptors are merged by a score-level fusion.

(i.e., FV). Fig. 4.3 shows an overview of the proposed PAD approach, which consists on three main steps: *i)* local features are extracted from a fingerprint sample, both real- and binary-valued (see Sect. 3.1); *ii)* an unsupervised GMM or BMM learns the distribution of the afore-

mentioned decorrelated features, which are subsequently encoded by computing the gradient of the sample log-likelihood with respect to the learned model parameters (i.e., using FV); and *iii)* a BP vs. AP decision is finally performed by a linear SVM (see Sect. 3.3).

In essence, we analyse to which extent different descriptors complement each other to improve the final PAD performance. To that end, the individual descriptor based PAD scores are fused with a weighted sum as follows:

$$s_f = \alpha \cdot s_1 + \beta \cdot s_2 + (1 - \alpha - \beta) \cdot s_3, \tag{4.2}$$

where $\alpha + \beta < 1$, and s_1, s_2 and s_3 represent the individual scores produced by the best three performing descriptors described above. Similar to the approach described in Sect. 4.1, the α and β weighted values are computed from each LivDet's training set.

4.3 EXPERIMENTAL SETUP

In order to perform a fair evaluation of the detection performance of the proposed PAD approaches for different scenarios, we define three main goals:

- Analyse the impact of the key parameter: the visual vocabulary size K (see its definition in Sect. 3.2) on the detection performance of the three common feature spaces (i.e., BoW, VLAD, and FV) and different local descriptors (i.e., SIFT, SURF, LBP, BSIF, and HOG).
- Study the impact of ridge pattern quality on the detection performance of the algorithms.
- Benchmark the detection performance of our approaches with the top state-of-the-art approaches.
- Explore the impact of different materials used in the fabrication of PAIs on the detection performance of our techniques.
- Evaluate realistic and challenging scenarios with unknown PAI species, cross-database, and cross-session settings.

4.3.1 *Databases*

The experiments were conducted on the well-known benchmarks provided by LivDet 2011 [225], LivDet 2013 [68], LivDet 2015 [140], and LivDet 2017 [141]. The Fingerprint Liveness Detection Competition (LivDet) is a biannual challenge aimed at evaluating recent academic and industry research in the field of PAD. A summary of their main

Table 4.1: Databases summary, including the list of PAI species available. The unknown PAI species at testing for LivDet 2015 and LivDet 2017 are highlighted in bold.

DB	# Samples	Capture device	PAI species
LivDet 2011	16,000	Digital 4000B Sagem MSO300	Gelatin, Latex, PlayDoh, Silicone, Wood Glue
		Biometrika FX2000 Italdata ET10	EcoFlex (platinum-catalysed silicone), Gelatine, Latex, Silgum, Wood Glue
LivDet 2013	8,000	Biometrika FX2000 Italdata ET10	EcoFlex (platinum-catalysed silicone), Gelatine, Latex, Modasil, Wood Glue
LivDet 2015	19,431	GreenBit DactyScan26 Biometrika HiScan-PRO Digital Persona U.are.U 5160	EcoFlex (platinum-catalysed silicone), Gelatin, Latex, Wood Glue, **Liquid EcoFlex**, **RTV**
		Crossmatch L Scan Guardian	Body Double, **OOMOO (silicone rubber)**, PlayDoh, EcoFlex, **novel form of gelatine**
LivDet 2017	18,984	Digital Persona U.are.U 5160 GreenBit DactyScan84C Orcanthus Certis2 Image	EcoFlex, Body Double, Wood Glue, **Gelatin**, **Latex**, **Liquid EcoFlex**

features is presented in Tab. 4.1. It should be noted that, unlike the previous databases, LivDet 2015 and LivDet 2017 contain unknown PAI species in the test set, which are not included in the training set. Therefore, both LivDet databases are used to evaluate the proposed PAD subsystems on unknown PAI species scenarios. In our research, the LivDet 2019 and LivDet 2021 databases are not used, as they have not been made public to the research community.

4.4 RESULTS AND DISCUSSION

4.4.1 *Known PAI species*

4.4.1.1 *Effect of the Semantic Sub-groups*

In the first set of experiments, we optimise the algorithms' detection performance in terms of the main key parameter: the visual vocabulary size K. To that end, we focus on the known PAI species scenario, in order to avoid a bias due to other variables. We test the following range of values: $K = \{256, 512, 1024\}$, since $K > 1024$ would yield too long feature vectors, not usable for real-time applications. Tab. 4.2 reports the D-EER values for the adopted K configurations over the space fusion-based scheme (see Fig. 4.2). As it can be observed, the best K values on average are $K = 512$ for FV and $K = 1024$ for VLAD and BoW. In particular, the FV representation reports an D-EER of 2.23%, which is approximately two and three times lower than the ones attained by the remaining encodings (4.88% for VLAD and 6.34% for BoW). This observation, in turn, indicates that FV is able to successfully separate a BP from an AP given a reduced number

Table 4.2: Detection performance, in terms of D-EER(%), of our proposed representations combined with the SIFT descriptor for different K values. The best results per encoding and capture device are highlighted in bold.

DB	Dataset	FV			VLAD			BoW		
		256	512	1024	256	512	1024	256	512	1024
2011	Biometrika	**2.80**	4.10	5.70	8.40	**8.30**	**8.30**	8.10	7.10	**6.40**
	Digital P.	0.70	**0.30**	**0.30**	2.00	1.30	**0.95**	2.20	1.40	**1.30**
	Italdata	3.20	**2.40**	4.50	**9.70**	16.10	13.3	16.20	**7.50**	12.70
	Sagem	1.72	1.60	**1.42**	3.00	**2.65**	**2.65**	6.48	6.53	**5.26**
2013	Biometrika	**0.30**	0.50	0.50	2.50	3.10	**1.80**	3.10	2.50	**1.80**
	Italdata	**0.30**	**0.30**	**0.30**	1.00	**0.80**	**0.80**	4.40	3.90	**3.70**
2015	GreenBit	1.60	**1.30**	1.40	4.80	**3.60**	3.80	4.20	**4.00**	4.40
	Digital P.	7.30	**6.50**	**6.50**	9.70	9.40	**8.90**	16.00	14.70	**13.40**
	Hi_Scan	4.60	**4.30**	4.50	6.50	6.80	**5.80**	11.40	10.50	**9.00**
	Crossmatch	1.06	**1.03**	**1.03**	3.62	3.62	**2.50**	7.03	6.21	**5.40**
	Avg.	2.36	**2.23**	2.62	5.12	5.57	**4.88**	7.91	6.43	**6.34**

of semantic sub-groups built by GMM, in contrast to the VLAD and BoW.

In a second set of experiments, we also optimise the detection performance of the best performing common feature space (i.e., FV) in combination with those local descriptors defined in Sect. 3.1 (see Fig. 4.3). Tab. 4.3 reports the D-EER for several descriptors over different number of semantic sub-groups (i.e., K). As it should be observed, most descriptors report their best D-EER for a small number of semantic sub-groups with the exception of BSIF and HOG, which obtain their optimum performance for $K = 1024$. Consequently with the results reported in Tab. 4.2, the SIFT descriptor yields an D-EER of 2.23% for $K = 512$ which is up to three times lower than the D-EER attained for $K = 1024$.

Table 4.3: Detection performance, in terms of D-EER (%), of the descriptors in combination with FV for different K values. The best results per descriptor are highlighted in bold, and the best D-EER per dataset is underline.

DB	Dataset	SIFT			BSIF			SURF			HOG			LBP			ORB			BRIEF		
		256	**512**	1024	256	512	**1024**	256	**512**	1024	256	512	**1024**	**256**	512	1024	256	**512**	1024	256	**512**	1024
2011	Biometrika	**2.80**	4.10	5.70	**2.40**	2.80	2.60	**4.00**	5.30	4.90	15.40	15.50	**12.80**	7.80	**6.90**	7.70	7.50	**6.20**	**6.20**	11.60	**10.70**	11.70
	Digital P.	0.70	**0.30**	**0.30**	0.60	0.50	**0.40**	1.50	1.20	**1.10**	2.30	1.00	**0.30**	**1.10**	**1.10**	1.60	4.20	**3.40**	3.70	1.30	1.30	**1.20**
	Italdata	3.20	**2.40**	4.50	6.50	6.90	**6.10**	16.40	**13.10**	13.40	**19.70**	20.50	20.00	**15.60**	19.30	20.90	**8.30**	9.90	11.60	16.10	**15.60**	**15.60**
	Sagem	1.72	1.60	**1.42**	1.42	1.23	**1.08**	**0.59**	0.69	0.89	9.92	**7.71**	7.91	**7.81**	8.50	10.80	3.09	3.00	**2.70**	6.29	6.19	**5.11**
2013	Biometrika	**0.30**	0.50	0.50	0.40	0.40	**0.30**	0.90	**0.80**	1.00	**1.90**	2.00	**1.90**	**2.50**	**2.50**	2.70	3.40	2.80	**2.60**	4.40	**3.20**	3.60
	Italdata	**0.30**	**0.30**	**0.30**	0.40	**0.30**	**0.30**	0.70	0.70	**0.60**	2.30	2.20	**2.00**	6.00	**5.20**	6.40	1.90	1.20	**1.10**	5.30	4.00	**2.90**
2015	GreenBit	1.60	**1.30**	1.40	**1.50**	2.50	2.20	3.60	**3.50**	3.90	5.40	**5.30**	5.40	4.80	**4.10**	4.90	2.80	**2.70**	3.10	**5.30**	6.00	6.40
	Digital P.	7.30	**6.50**	**6.50**	**4.10**	4.40	4.30	7.90	**7.00**	7.40	11.70	**11.20**	11.30	12.90	**12.50**	13.00	**16.50**	16.90	17.10	18.20	**16.60**	17.40
	Hi-Scan	4.60	**4.30**	4.50	4.90	4.60	**4.50**	8.50	8.70	**8.60**	8.30	7.10	**6.90**	13.50	13.60	**13.10**	14.80	**14.50**	15.70	9.90	**9.60**	**9.60**
	Crossmatch	1.06	**1.03**	**1.03**	2.47	**2.12**	2.22	8.34	**7.53**	7.62	2.59	**1.88**	2.09	6.34	**6.21**	6.59	**9.40**	**9.40**	10.46	5.41	5.06	**5.03**
	Avg.	2.36	**2.23**	2.62	2.47	2.58	**2.40**	5.24	**4.85**	4.94	7.95	7.44	**7.06**	**7.84**	7.99	8.77	7.19	**7.00**	7.43	8.38	**7.83**	7.85

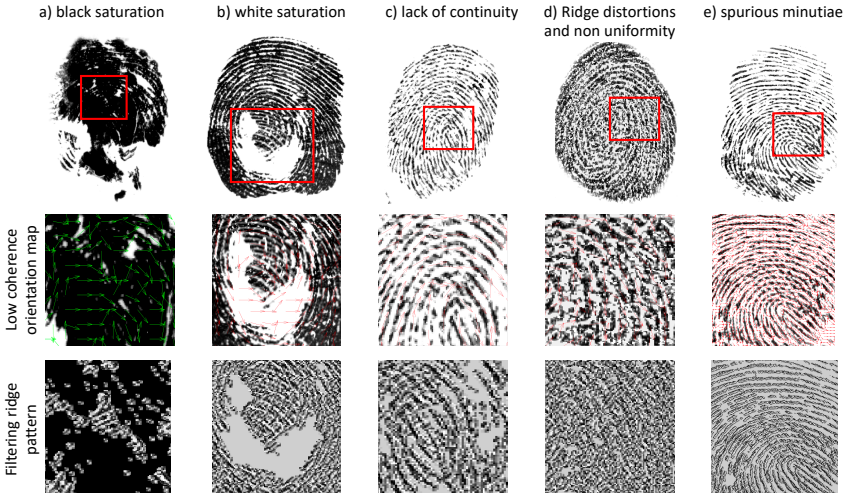


Figure 4.4: Several artefacts over the fingerprint ridge pattern which are frequently found on the AP samples: *a*) higher black saturation, *b*) high white saturation, *c*) lack of continuity on the ridge pattern, *d*) unwanted noises and ridge distortions, and *e*) spurious minutiae produced by earlier artefacts.

4.4.1.2 *Gradient vs. Texture vs. Intensity Differences*

Taking a close look at Tab. 4.3 we observe that the gradient-based descriptors report on average the best detection performance for all databases (i.e., D-EER = 4.17%), followed by texture-based features (i.e., D-EER = 5.12%), and finally, intensity difference-based descriptors (i.e., D-EER = 7.42). By carefully analysing several PAI species from the LivDet databases, we noted that there exist at least five common artefacts which are fully represented by gradient- and texture-based descriptors and hence they could be employed for fingerprint PAD (see Fig. 4.4). Specifically, the gradient computed over fingerprints allows representing their orientation field, hence capturing some ridge pattern characteristics such as black and white saturation on the ridges, lack of continuity, ridge distortions or unwanted noises, non-ridge uniformity, and spurious minutiae, among others which produce a high number of low coherence areas. Consequently, those ridge properties could be also captured by convolving a fingerprint image with a suitable kernel, as shown in Fig. 4.4 third row. As mentioned in Sect. 3.1.8, we employed sixty filter kernels for the BSIF computation. The best performing filter configurations per LivDet dataset are reported in Tab. 4.4. As it should be noted, a texture-based descriptor as BSIF achieves its best detection performance for small-size filter kernels in most cases (i.e., $N \leq 9$): large-size filter kernels can lead to a deterioration of the fingerprint ridge pattern structure, thus removing the aforementioned artefacts. Finally, we can see in Tab. 4.3 that intensity difference-based features analysed by ORB and BRIEF

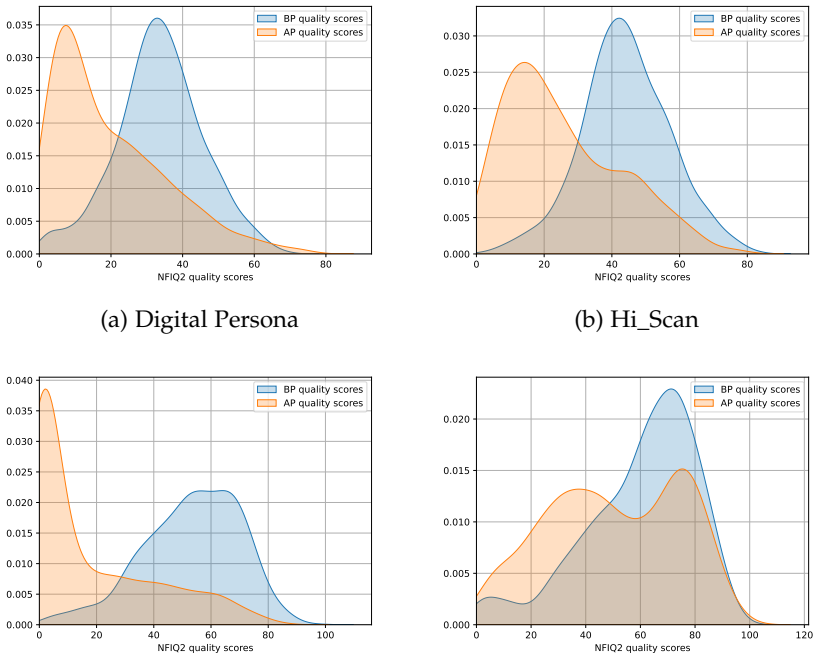
Table 4.4: Best performing BSIF filter configurations per dataset and K value. The best results per dataset are highlighted in bold.

DB	Dataset	BSIF parameters		
		256	512	1024
2011	Biometrika	N **= 5,** l **= 5**	$N = 5, l = 5$	$N = 5, l = 5$
	Digital P.	$N = 11, l = 7$	$N = 11, l = 7$	N **= 11,** l **= 7**
	Italdata	$N = 7, l = 11$	$N = 11, l = 5$	N **= 8,** l **= 3**
	Sagem	$N = 8, l = 7$	$N = 8, l = 7$	N **= 8,** l **= 7**
2013	Biometrika	$N = 6, l = 3$	$N = 6, l = 3$	N **= 6,** l **= 5**
	Italdata	$N = 10, l = 5$	N **= 6,** l **= 9**	N **= 9,** l **= 7**
2015	GreenBit	N **= 7,** l **= 5**	$N = 7, l = 3$	$N = 7, l = 3$
	Digital P.	N **= 6,** l **= 3**	$N = 6, l = 3$	$N = 6, l = 3$
	Hi-Scan	$N = 7, l = 3$	$N = 7, l = 3$	N **= 7,** l **= 3**
	Crossmatch	$N = 8, l = 9$	N **= 12,** l **= 9**	$N = 12, l = 9$
2017	Digital P.	$N = 12, l = 7$	N **= 12,** l **= 7**	$N = 12, l = 7$
	GreenBit	$N = 7, l = 3$	$N = 7, l = 3$	N **= 7,** l **= 3**
	Orcanthus	N **= 6,** l **= 3**	$N = 6, l = 3$	$N = 7, l = 5$

are not suitable to detect an AP attempt, thereby resulting in a poor detection performance.

4.4.1.3 *Effect of the Fingerprint Quality*

We also perceive in Tab. 4.3 that the best performing descriptor (i.e., SIFT) attains a poor detection performance for two out of four datasets in LivDet 2015. In particular, it attains an D-EER of 4.30% and 6.20% for Hi_Scan and Digital Persona, which are respectively three and five times worse than the ones reported by GreenBit and Crossmatch. According to [69], most PAD techniques submitted to LivDet 2015 did not perform well due to the small image size. However, by carefully analysing the fingerprint quality provided by the NFIQ2.0 approach [191] for the entire LivDet 2015 datasets in Fig. 4.5, we found that most BP images in the Digital Persona and Hi_Scan datasets yield a poor NFIQ2.0 quality, in contrast to the ones in GreenBit and Crossmatch. Whereas 8% and 30% of the fingerprints in Digital Persona and Hi_Scan present a good NFIQ2.0 quality greater than 50% (good quality), most BP samples in GreenBit (i.e., 63%) and Crossmatch (i.e., 72%) pose a good NFIQ2.0 quality score. Therefore, both capture devices include some sensor technology which produces a high noise degree on the fingerprint samples, and hence also affects the detection



(a) Digital Persona (b) Hi_Scan

(c) GreenBit (d) Crossmatch

Figure 4.5: NFIQ2.0 quality distribution for the LivDet 20115 datasets.

performance of most state-of-the-art PAD methods [69, 144], even our approach.

The above observation is also confirmed in Fig. 4.6, which reports the detection performance of our best common feature space (i.e., FV) for different fingerprint image quality ranges over the LivDet 2015. As it can be noted, all descriptor categories achieve a detection performance improvement with the BP ridge pattern quality. In particular, gradient-based descriptors yield a mean D-EER of 2.36% for BP images with a NFIQ2.0 quality greater than 40, which outperforms the texture- and intensity difference-based features by a relative 20% and 73%, respectively. These findings in turn confirm the soundness of the gradient-based descriptors to capture the aforementioned ridge pattern artefacts and hence detecting the AP attempts. Consequently with these results, we show in Fig. 4.7 an example of a misclassified BP with a poor NFIQ2.0 quality. As a conclusion, we do confirm that the orientation field, representing the fingerprint ridge pattern, can be successfully employed as a discriminative feature to detect AP attempts whose capture devices do not include a high noise degree over the BP ridge pattern.

4.4.2 *Impact of Different Fabrication Materials*

Now, we study the impact of several PAI species used in the fabrication of PAIs on the PAD performance. In 2019, Chugh and Jain [32]

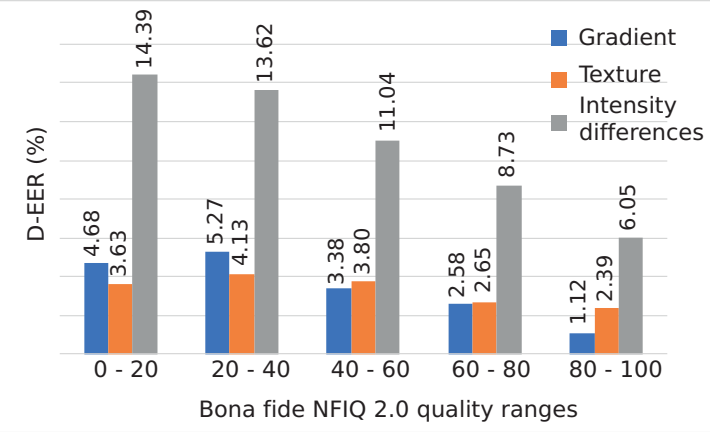


Figure 4.6: D-EER benchmark in terms of NFIQ2.0 quality per descriptor category.

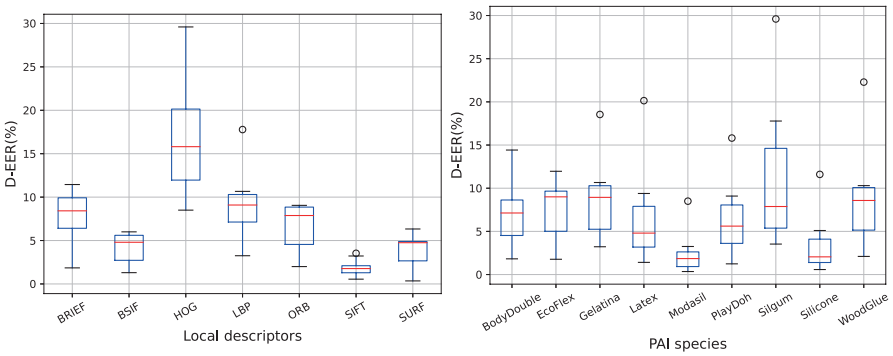
(a) BP with NFIQ2.0 = 29.

(b) AP with NFIQ2.0 = 29.

Figure 4.7: BP and AP samples which report the same NFIQ2.0 quality. *a*) a misclassified BP sample whose ridges include a high noise degree, and *b*) an AP image with a high noise degree.

analysed 12 different PAI species over a database acquired with a single CrossMatch capture device. The authors grouped the features extracted by the Fingerprint Spoof Buster PAD method [31] to derive a training set, comprising only six PAI species. Thus, they achieved a similar detection performance to the algorithm trained with the entire set of PAI species. In our Thesis, we address some main questions remaining unanswered: *i*) to which extent are some PAI species harder to detect? and *ii*) how does this difficulty vary for different feature extractors and for different sensors?

To address the above questions, we evaluate the impact of different PAI species included in the LivDet databases on the proposed descriptors in combination with the FV. In the experimental evaluation, all testing and training images are acquired using the same capture device. In order not to bias the results, the same set of PAI species is



(a) Local feature descriptor.

(b) PAI species.

Figure 4.8: Evaluation of different PAI species on the PAD performance.

used in the fabrication of training and test samples. The results per descriptor and PAI species are reported as boxplots in Fig. 4.8.

Taking a look at Fig. 4.8-a), we observe that SIFT yields the lowest error rates across different PAI species (i.e., a mean D-EER of 1.88%). Its Standard Deviation (STD) (i.e., size of the corresponding box) is also the lowest (0.96%), thereby indicating its higher robustness to different PAI species with respect to the other descriptors considered. Following SIFT, the BSIF-based representation achieves a D-EER of 4.07%, which is slightly higher than the one attained by the SURF (i.e.., D-EER = 3.59%). However, the STD reported by SURF (STD = 2.13%) is higher than the one yielded by BSIF (STD = 1.84). Therefore, we may conclude that the latter is more robust to PAI species variability than the former. In addition, we note that there is no direct relationship between the overall performance of a particular PAD method and its robustness to different PAI species, even if these are known during the training.

Regarding Fig. 4.8-b), it should be seen that the highest variability is yielded by the PAIs fabricated with Silgium (D-EER = 11.57% $\pm$ 9.30%), thereby reflecting its high resemblance with the BP samples. In contrast to Silgum, the PAIs created using Modasil show a very distinct appearance, with no noise in the ridges as in the BP samples. Those are easier to detect by all descriptors, thereby resulting in the lowest error rates (i.e., D-EER = 2.54% $\pm$ 2.80%). To sum up, we can conclude that, as could be expected, some PAI species (e.g., Silgium) are harder to detect than others (e.g., Modasil).

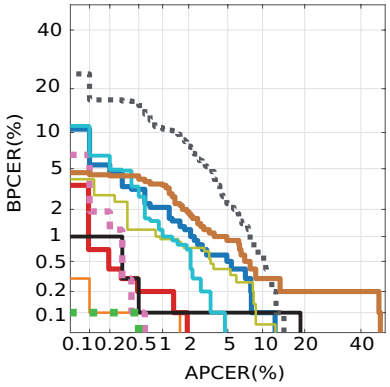
4.4.2.1 *Benchmark with the State Of The Art*

Finally, we establish a benchmark in Tab. 4.5 of the space (i.e., the combination of the three common feature spaces with SIFT, as shown in Fig. 4.2) and the descriptor (i.e., the combination of the three best performing descriptors, SIFT, SURF, and BSIF, with the FV, as depicted in Fig. 4.3) fusions with the current top state-of-the-art. It is worth noting that, in our investigation, we also experimented with

the fusion between the best descriptors per category (i.e., SIFT, BSIF, and ORB). However, the low discriminative power of the Intensity differences-based descriptors led to a clear performance deterioration for unknown PAI species. As it may be observed, both proposed fusions achieve the state-of-the-art approaches for most datasets. In particular, the descriptor fusion reports, on average, remarkable D-EERs of 0.95%, 0.30%, and 1.46% for the three LivDet databases. In contrast to most Deep learning approaches, our descriptor fusion is also able to yield a good detection performance for Digital Persona in LivDet 2015 (i.e., D-EER of 0.10%). As was mentioned, most algorithms submitted to LivDet 2015 did not perform well on Digital Persona due to the small image size [69]. Moreover, most state-of-the-art algorithms decrease their detection performance for those samples acquired by capture devices (e.g., Digital Persona U.are.U 5160 and Biometrika Hi-Scan-PRO) whose acquisition technology produces a high unwanted noise degree on the fingerprint ridge pattern.

Table 4.5: Benchmark in terms of the D-EER(%) with the top state-of-the-art. The best results are highlighted in bold.

DB	Dataset	VGG [144]	TripleNet [151]	FSB-v1 [30]	TinyFCN [153]	FSB-v2 [31]	FLDNet [234]	Space Fusion	Descriptor Fusion
2011	Biometrika	5.20	5.15	2.60	**1.10**	1.24	-	2.40 ($\alpha = 0.7$, $\beta = 0.0$)	1.50 ($\alpha = 0.6$, $\beta = 0.4$)
	Digital P.	3.20	1.85	2.70	1.10	1.61	-	**0.10** ($\alpha = 0.8$, $\beta = 0.0$)	**0.10** ($\alpha = 0.2$, $\beta = 0.8$)
	Italdata	8.00	5.10	3.25	4.75	2.45	-	2.20 ($\alpha = 0.8$, $\beta = 0.0$)	**1.90** ($\alpha = 0.3$, $\beta = 0.7$)
	Sagem	1.70	1.23	1.80	1.56	1.39	-	1.13 ($\alpha = 0.8$, $\beta = 0.2$)	**0.29** ($\alpha = 0.5$, $\beta = 0.0$)
	Avg.	4.52	3.33	2.59	3.12	1.67	-	1.46	**0.95**
2013	Biometrika	1.80	0.65	0.60	0.35	**0.20**	0.36	0.30 ($\alpha = 0.9$, $\beta = 0.0$)	0.30 ($\alpha = 0.0$, $\beta = 0.6$)
	Italdata	0.40	0.50	0.40	0.40	**0.30**	1.35	**0.30** ($\alpha = 0.1$, $\beta = 0.0$)	**0.30** ($\alpha = 0.0$, $\beta = 0.8$)
	Avg.	1.10	0.58	0.50	0.38	**0.25**	0.86	0.30	0.30
2015	GreenBit	4.60	-	2.00	**0.20**	0.68	0.53	1.30 ($\alpha = 1.0$, $\beta = 0.0$)	1.00 ($\alpha = 0.3$, $\beta = 0.5$)
	Digital P.	5.64	-	1.76	3.40	**1.12**	3.61	6.20 ($\alpha = 1.0$, $\beta = 0.0$)	**0.10** ($\alpha = 0.6$, $\beta = 0.1$)
	Hi_Scan	6.28	-	1.08	**0.35**	1.48	2.95	4.30 ($\alpha = 1.0$, $\beta = 0.0$)	3.80 ($\alpha = 0.5$, $\beta = 0.3$)
	Crossmatch	1.90	-	0.81	1.09	**0.64**	1.78	1.03 ($\alpha = 1.0$, $\beta = 0.0$)	0.94 ($\alpha = 0.1$, $\beta = 0.8$)
	Avg.	4.61	-	1.39	1.26	**0.97**	2.22	3.20	1.46



	BPCER10(%)	BPCER20(%)	BPCER100(%)
Biometrika11	0.10	0.50	2.10
Digital11	0.00	0.00	0.10
Italdata11	0.30	0.90	3.50
Sagem11	0.00	0.00	0.20
Biometrika13	0.10	0.10	0.10
Italdata13	0.00	0.00	0.00
Crossmatch15	0.07	0.40	0.90
GreenBit15	0.00	0.00	1.00
Hi-Scan15	0.60	2.30	10.60
Digital15	0.00	0.00	0.00
Avg.	0.12	0.42	1.85

Figure 4.9: Presentation attack detection error trade-off between BPCER over APCER over the known PAI species scenario for the descriptor-fusion-based approach.

4.4.2.2 *In-depth Detection Performance Analysis*

In order to analyse the feasibility of our best fusion approach (i.e., descriptor fusion) for an operational real application, we evaluate in Fig. 4.9 its detection performance in compliance with the ISO/IEC 30107-3 [97]. As it may be observed, the performance varies considerably from the best (i.e., Digital Persona subset from LivDet 2015, with a BPCER = 0.00% for APCER $\geq$ 0.50% or APCER = 0.00% for BPCER $\geq$ 0.05%) to the worst case (i.e., Hi_Scan subset from LivDet 2015, with a BPCER_{100} = 10.60%). Specifically, our algorithm reports on average a BPCER_{100} of 1.48% and 0.05% for LivDet 2011 and LivDet 2013, respectively, which are up to seven times lower than the ones attained by the current state-of-the-art techniques (i.e., a BPCER_{100} of 9.68% for FSB-v1 [30] and 4.05% for FSB-v2 [31] on the LivDet 2011, and a BPCER_{100} of 0.20% for FSB-v1 [30] and 0.05% for FSB-v2 [31] on the LivDet 2013).

Consequently, the proposed fusion is able to achieve a top detection performance for most datasets in LivDet 2015: a mean BPCER_{100} of 0.63% for the Crossmatch, GreenBit, and Digital Persona datasets (i.e., state-of-the-art: 1.61% in FSB-v1 [30] and 0.92% in FSB-v2 [31]). In addition, it should be noted that our method yields its worst detection performance for the Hi_Scan capture device, thereby resulting in a BPCER_{100} of 10.60%. The Hi_Scan dataset includes high-resolution fingerprints with sizes of 1000 $\times$ 1000 pixels where the ROI for most PAIs only covers a 40% of a whole image. In contrast, the ROI for BP samples covers up to 70% of pixels in the images. Since our proposed approach extracts the descriptors from the whole image, we think that a ROI segmentation or a reduction of the points on the regular grid to particular landmarks such minutiae, for the feature extraction, could lead to a detection performance improvement for this type of high-resolution capture device.

Finally, it should be noted that, in general, a good balance between high user convenience or usability (i.e., low BPCER) and high security (i.e., low APCER) can be achieved with the proposed method. In particular, the BPCER ranges between 0.12% and 1.85% for higher security thresholds (i.e., 1.00% $\leq$ APCER $\leq$ 10.00%) confirms the remarkable detection performance of the fusion between gradient and texture-based descriptors for this baseline scenario.

4.4.3 *Unknown PAI species*

One of the main objectives of this Thesis is to deal with scenarios with unknown factors. Therefore, we analyse in detail the detection performance of both fusion representations for unknown PAI species. Two sets of experiments are then performed following three different protocols. In both experiments, all training and test images were acquired by the same capture device.

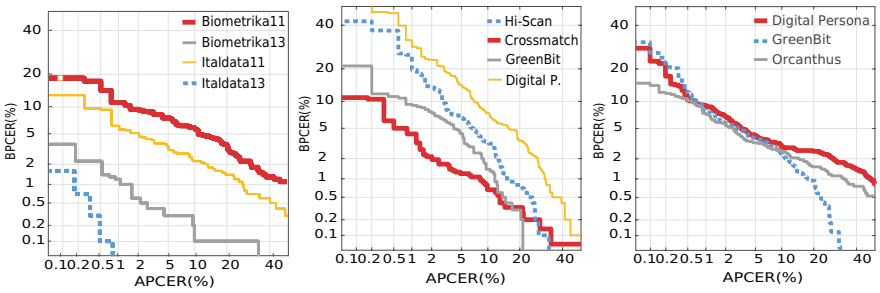
To evaluate the generalisation capability of our PAD algorithms, we select the LivDet 2015 database in which unknown PAI species were used for the fabrication of PAIs in the test set (see Tab. 4.1). In addition, we follow the experimental protocol described in [144] where the LivDet 2011 and 2013 databases are involved. Following this idea, an unknown PAI species evaluation is also carried out over the LivDet 2017. Tab. 4.6 shows the corresponding D-EER values for all subsets.

Focusing first on the LivDet 2011 and LivDet 2013 databases for the challenging scenario, a similar trend to the baseline scenario can be observed for the three selected local descriptors: the gradient-based descriptors achieve on average the best performance for most datasets (average D-EER = 2.08% for SURF), followed by the texture-based descriptor (D-EER = 2.36% for BSIF). In addition, it should be noted that both the three descriptors as well as two fusion algorithms outperform the top state-of-the-art. In particular, the descriptor fusion yields a mean D-EER of 1.00%, which is approximately three and ten times better than the ones attained by the best methods. These results can be also observed for the LivDet 2017 database, reporting on average an D-EER of 3.97% which is better than the one attained by the LivDet 2017 winner [141]. Finally, it is important to highlight that most techniques report a performance deterioration for those test datasets including PAIs fabricated with the PAI species Silgum (see Tab. 4.6, Bio11 and Ita11 rows).

Table 4.6: Detection performance of our fusion representations, in terms of D-EER (%), for several unknown PAI species scenarios.

Protocol	Dataset	PAI species		SIFT	BSIF	SURF	Descriptor Fusion	Space Fusion	FSB-v2 [31]	FLDNet [234]	LivDet 2017 Winner [141]†
		Train	Test								
Proposed by [144]	Bio11	EcoFlex, Gelatine, Latex	Silgum, Woodglue	6.33	3.72	4.61	**1.73**	4.78	4.60	-	-
	Bio13	Modasil, Woodglue	EcoFlex, Gelatine, Latex	1.00	**0.87**	1.31	1.04	1.50	1.30	**0.87**	-
	Ita11	EcoFlex, Gelatine, Latex	Silgum, Woodglue, Other	3.78	4.61	2.00	**1.11**	3.60	5.20	-	-
	Ita13	Modasil, Woodglue	EcoFlex, Gelatine, Latex	0.30	0.23	0.41	**0.13**	0.50	0.60	0.94	-
		Avg.		2.85	2.36	2.08	**1.00**	2.60	2.93	-	-
LivDet 2015	Crossmatch	Body Double, EcoFlex,PlayDoh	Gelatine, OOMOO	1.37	2.56	6.87	2.01	**1.34**	-	2.66	-
	Digital P.	EcoFlex_00-50,Latex, Gelatine, Woodglue	Liquid EcoFlex, RTV	9.40	5.80	10.00	8.45	8.85	-	**3.06**	-
	GreenBit			4.20	5.45	7.80	4.65	4.20	-	**0.46**	-
	Hi-Scan			6.65	8.65	11.60	6.00	6.65	-	**3.38**	-
		Avg.		5.41	5.61	9.07	5.28	5.26	-	**2.39**	-
LivDet 2017*	Digital P.	Woodglue, EcoFlex, Body Double	Gelatine, Latex, Liquid EcoFlex	5.11	5.97	5.03	**4.22** ($\alpha = 0.3$, $\beta = 0.4$)	4.84 ($\alpha = 0.9$, $\beta = 0.1$)	-	-	4.41
	GreenBit			5.64	5.35	5.88	4.06 ($\alpha = 0.3$, $\beta = 0.4$)	5.35 ($\alpha = 0.9$, $\beta = 0.1$)	-	-	**3.56**
	Orcanthus			6.16	4.04	7.00	**3.63** ($\alpha = 0.6$, $\beta = 0.2$)	5.62 ($\alpha = 1.0$, $\beta = 0.0$)	-	-	6.29
		Avg.		5.64	5.12	5.97	**3.97**	5.15	-	-	4.75

† The overall classification errors reported by LivDet 2017 winner in this work are the complement of the overall accuracy achieved in [141]
* The D-EER results were achieved at $K = 512$ for SIFT, and $K = 256$ for BSIF and SURF.



(a) Proposed in [144]. (b) LivDet 2015 [140]. (c) LivDet 2017 [141].

Figure 4.10: DET curves on the unknown PAI species scenario for the best performing fusion algorithm (i.e., Descriptor Fusion).

Regarding the experiments run on LivDet 2015 (see Tab. 4.6, mid row), the trend observed for LivDet 2011 and LivDet 2013 is confirmed: gradient-based descriptors show the best performance, followed by the texture-based one (i.e., 5.41% vs. 5.61%). On the other hand, it may be noted that our fusion method suffers a high-performance deterioration for Digital Persona and Hi_Scan due to their fingerprint quality: 90% and 70% of fingerprint images in those datasets report a NFIQ2.0 quality below 50%, thereby producing an accuracy decrease for most PAD techniques [144, 153, 234]. Therefore, those images are unsuitable to the PAD task and hence for a real fingerprint recognition system.

To conclude the analysis of this scenario, the ISO/IEC-compliant evaluation of the fusion approach over all datasets is presented in Fig. 4.10. As was expected, the performance is worse than that over the known PAI species scenario (see Fig. 4.9). Nevertheless, an average BPCER_{100} of 8.30%, BPCER_{20} of 3.47%, and BPCER_{10} of 1.93% can be achieved, thus still granting a secure and usable system. In addition, it should be noted that the gap in performance at the D-EER for Digital Persona and Hi_Scan in the LivDet 2015 database is here confirmed for all operating points (i.e., a BPCER_{100} of 19.60% for Hi_Scan and a BPCER_{100} of 29.30% for Digital Persona).

4.4.4 *Cross-database and Cross-session Evaluations*

We now evaluate the cross-database scenario, where different capture devices might be used for training and testing at some point in time. This scenario is likely to happen during a long-time deployment, where the fingerprint capture device might age and eventually stop working. Therefore, the fabrication and acquisition of the entire set of earlier known PAI species with the new capture device at hand might not be possible or at least require some time, thereby being not available for high-security applications.

Table 4.7: Performance evaluation in terms of D-EER for cross-session and cross-database scenarios.

(a) Cross-database protocol.

Train - Test	SIFT	BSIF	SURF	Descriptor Fusion	Space Fusion	FSB-v2 [31]	FLDNet [234]
Bio11 - Ital11	11.30	11.65	11.40	**9.65**	11.45	25.35	-
Bio13 - Ital13	1.80	2.95	**0.95**	1.55	1.80	4.30	2.10
Ital11 - Bio11	**2.40**	10.05	7.95	3.70	7.40	25.21	-
Ital13 - Bio13	**0.80**	3.30	1.35	0.90	0.75	3.50	2.90
Avg.	4.08	6.99	5.41	**3.95**	5.35	14.59	-

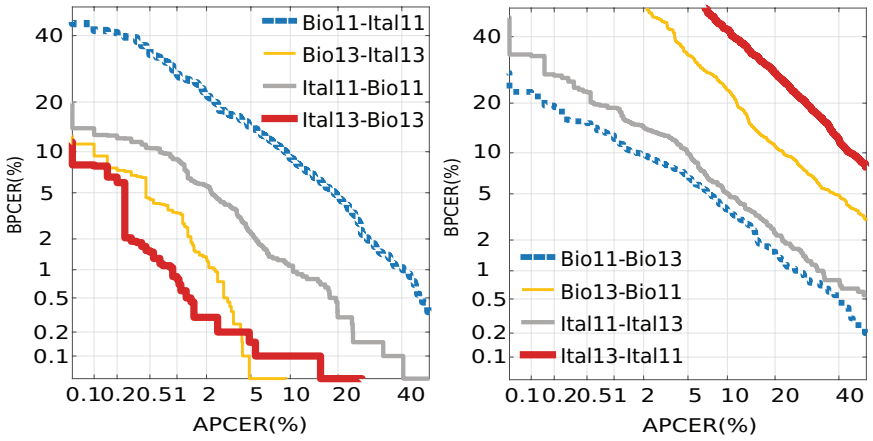
(b) Cross-session protocol.

Train - Test	SIFT	BSIF	SURF	Descriptor Fusion	Space Fusion	FSB-v2 [31]	TripleNet [151]
Bio11 - Bio13	6.80	3.95	**3.90**	5.90	4.00	7.60	14.00
Bio13 - Bio11	**12.70**	16.95	18.55	14.75	13.60	31.16	34.05
Ital11 - Ital13	5.60	**2.95**	9.40	6.95	5.60	6.70	8.30
Ital13 - Ital11	**11.50**	21.95	20.10	24.10	17.50	26.16	44.65
Avg.	**9.15**	11.45	12.99	12.92	10.18	17.91	25.25

a) Biometrika 2011 b) Italdata 2011 c) Biometrika 2013 d) Italdata 2013

Figure 4.11: Appearance behaviour across capture devices for two fingerprints: a) fingerprint sample in Biometrika 2011, b) fingerprint sample in Italdata 2011, c) fingerprint sample in Biometrika 2013, and d) fingerprint sample in Italdata 2013.

In order to study the generalisability of our proposed methods for cross-database, we adopt four training set - test set configurations proposed by [144]. Tab. 4.7-a) reports the corresponding D-EER values. As it may be observed, a gradient-based descriptor (i.e., SIFT) still provides the lowest error rates (D-EER = 4.08%). However, it should be noted that BSIF deploys a similar detection performance for the same pairwise configurations. Even if Italadata 2011 and Biometrika 2011 visually show different texture patterns (see Fig. 4.11), SIFT is able to yield similar D-EER values when these datasets are used for testing (i.e., D-EER of 11.65% for Bio11-Ital11 vs. 10.05% for Ital11-Bio11). Consequently, similar behaviour can be perceived for Biometrika 2013 and Italdata 2013: these look visually more similar, thereby resulting in a low D-EER for all descriptors. Therefore, in order to successfully achieve the interoperability requirement between capture devices, the



(a) Cross-database. (b) Cross-session.

Figure 4.12: DET curves on the cross-database and cross-session scenarios for the best performing fusion representation.

selection of a new sensor must be carefully performed taking into account the five fingerprint ridge pattern properties mentioned in Sect. 4.4.1.2. This could in turn avoid a gap in the detection performance of state-of-the-art techniques.

Regarding the fused representations, we note that they considerably outperform the state-of-the-art by a relative 73% (i.e., D-EER = 3.95% for the descriptor fusion vs. 14.59% for FSB-v2 [31]), thereby showing its generalisation capability for this challenging scenario.

In Fig. 4.12-a), the ISO/IEC-compliant evaluation of the descriptor fusion scheme is depicted. We can first observe the increased detection performance gap between LivDet 2011 and LivDet 2013 for all operating points. In particular, our approach yields a mean $BPCER_{100}$ of 2.40% for datasets in LivDet 2013, which is approximately nine times lower than the one reported for datasets in LivDet 2011 (i.e., $BPCER_{100}$ = 18.18%). Despite this detection performance gap, our fused method is able to achieve a mean $BPCER_{100}$ of 10.14%, thus providing both user convenience and security.

Finally, we evaluate the scenario where different data collection sessions for the same capture device are used for training and testing. To that end, we select two datasets (i.e., Biometrika and Italdata), whose sensors were respectively used for fingerprint acquisition in the LivDet 2011 and LivDet 2013 competitions. Tab. 4.7-b) shows the corresponding D-EER values.

As in most scenarios analysed, a gradient-based descriptor (i.e., SIFT) provides the best detection performance. In particular, SIFT reports a mean D-EER below 10%, which outperforms the remaining descriptors, their fusion (i.e., the descriptor fusion), and the top state-of-the-art techniques. It should be noted that our descriptors and their fusion suffer a detection performance deterioration when the LivDet 2011 dataset is employed for testing. Whereas datasets acquired at

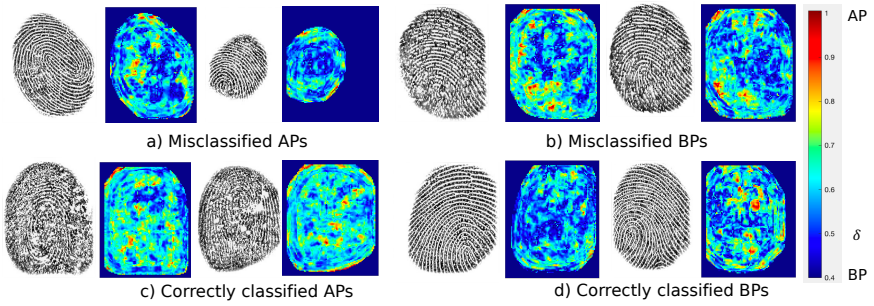


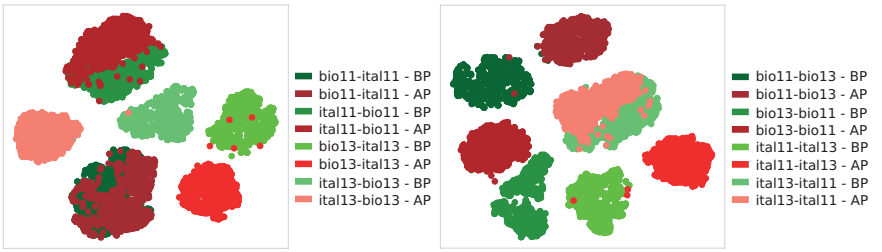
Figure 4.13: Heatmaps with the predicted scores for misclassified and correctly classified samples.

different years with the same capture device look visually similar (see Fig. 4.11), they report a different detection performance. Specifically, the evaluation of LivDet 2011 attains a mean D-EER of 16.95%, in contrast to 5.43% reported by testing LivDet 2013. Similar to the unknown PAI species results, the inclusion of an unknown PAI species as Silgum in LivDet 2011 is one of the issues leading to a detection accuracy decrease of our approach. Finally, the Fig. 4.12-b) confirms the detection performance gap over the LivDet 2011 and 2013 databases (i.e., blue and grey vs. yellow and red): a higher $BPCER_{100}$ over 60% for LivDet 2011 yields a non-usable fingerprint system.

4.4.5 *Visualisation of the FV Representation*

We show in Fig. 4.13 the visualisation of the scores predicted by our best approach (i.e., descriptor fusion-based pipeline) for misclassified and correctly classified samples taken from the LivDet 2015. It should be noted that both the heatmaps for misclassified BPs and correctly classified APs contain a high number of low coherence areas or unwanted noise, in contrast to the ones yielded for wrongly classified APs and correctly classified BPs. Those areas of low coherence are produced by the capture devices in the sample acquisition. In addition, we may observe that our proposed method fails for those PAIs having a perfectly defined ridge pattern, as depicted in Fig 4.13-a). As it was mentioned above, local analysis for particular landmarks such as minutiae could lead to an improvement for these challenging cases.

Finally, a t-SNE visualisation in Fig. 4.14 for the cross-database and cross-session scenarios shows the capability of the FV representation to separate an AP from a BP. We can observe that feature spaces for AP samples appear to be, at most cases, closer with each other than with those BP attempts. Even for those testing capture devices such as Biometrika 2011 (see Fig. 4.14-b)), which contains PAI species unknown in the Biometrika 2013 training set, we can note that our approach was able to find a set of semantic sub-groups from known



(a) Cross-database. (b) Cross-session.

Figure 4.14: t-SNE visualisation of the FV common feature space for the cross-database and cross-session scenarios.

samples to successfully fit those unknown PAI species. This, in turn, confirms the aforementioned hypothesis in Sect. 4.

4.4.6 *Summary*

To summarise the findings on the fingerprint PAD, we can highlight the following takeaway messages:

- Among the three common feature space (i.e., BoW, FV, and VLAD), the best detection performance is obtained for FV, and the worst for BoW.
- Gradient-based descriptors (i.e., SIFT and SURF) successfully represent low coherence areas produced by several fingerprint ridge pattern artefacts such as black saturation, white saturation, lack of continuity, unwanted noises and ridge distortions, thereby resulting in the best detection performance in most scenarios.
- A NFIQ2.0 evaluation over the LivDet 2015 database showed that different analysed descriptors improved their detection performance as the ridge pattern of BP fingerprints enhanced. Therefore, the NFIQ2.0 can be employed as an secure indicator to obtain a reliable PAD module: an D-EER $< 2.58\%$ is reported when the BP fingerprint quality is greater than 60 (i.e., NFIQ2.0 > 60).
- The proposed PAD methods, based on local image descriptors and common features spaces, are able to outperform the state-of-the-art techniques not only in the baseline scenario (i.e., both the PAIs and the acquisition devices are known a priori) but also in more realistic and challenging scenarios (i.e., unknown PAI species, cross-session, and cross-database). In particular, the D-EER is reduced by up to four times for the cross-database (i.e., D-EER = 3.95% vs. 14.59%).
- Further, a fusion at score's level between three common feature spaces deployed a performance improvement in most cases,

thereby resulting in a $BPCER_{100}$ in the range of 1.98% - 17% in the presence of unknown PAI species. This, in turn, confirmed that the hard quantisation computed by VLAD and BoW can be used as additional information to enhance the soft quantisation built by the FV approach. We think that for those non-improvement cases, proper tuning of the fusion parameters could enhance its detection performance.

- In addition, the ISO/IEC compliant evaluation revealed that the best performing fusion approach provides an usable system (i.e., low BPCER) even for a high security (APCER = 1%) operating point: it achieves an average $BPCER_{100}$ = 1.85% for the baseline scenario; $BPCER_{100} < 1.20\%$ for LivDet 2013, $BPCER_{100} < 16\%$ LivDet 2015 and $BPCER_{100} < 9\%$ LivDet 2017 for the unknown PAI species scenario; and $BPCER_{100} < 11\%$ for the cross-database evaluation.

- Texture-based features also yield the best detection performance right after gradient-based descriptors. In particular, BSIF achieves its best performance for small filter sizes (i.e., $N < 9$), which capture most of aforementioned artefacts. Given that BSIF depends on a set of filters previously learnt from thirteen natural images, we think that the use of filters trained for the particular fingerprint PAD task or extracted from intermediate CNN layers could unveil other ridge artefacts, hence improving the BSIF performance.

- Even if the Intensity differences- or binary-based descriptors (i.e., ORB and BRIEF) offer a lower computational load, their performance is not competitive against their continuous counterparts for fingerprint PAD purposes (i.e., SIFT, SURF, BSIF, HOG, and LBP).

- The fusion of gradient- and texture-based information considerably improves the detection performance of the single descriptors, even in the scenarios where textural features alone achieve considerably higher error rates (e.g., cross-database).

- The semantic sub-groups learned by the GMM allow modelling most aforementioned artefacts produced in the creation of PAIs. A better artefact description by the semantic sub-groups depends on that the input features follow a Gaussian distribution. In order to remove this GMM constrain and hence improve the FV representation, new deep generative models, which have shown to be more powerful for learning data distribution, could be evaluated.

- Whereas deep learning-based fingerprint PAD approaches require large databases for optimising thousands of parameters,

our proposal attained a high detection performance by tuning a small number of them (K, α, and β) from a small dataset.

- Further, most state-of-the-art techniques yield a poor detection performance over Digital Persona in LivDet 2015 due to its small image size, our fusion method reports a remarkable D-EER of 0.10%.

- Since Hi_Scan contains images with a high size of 1000×1000 pixels where the ROI area for PAIs only covers a 40% of the whole image, our best fusion-based representations is unable to report a reliable detection performance, thereby resulting in a BPCER_{100} of 10.60%. A reduction of the points on the regular grid to specific landmarks such as minutiae, for the feature extraction, or a ROI segmentation could improve its error rates.

5 FACE PRESENTATION ATTACK DETECTION

In this Chapter, we evaluate the feasibility of using FV for face PAD (Sect. 5.1). Whereas gradient-based descriptors such as SIFT and SURF showed to be an appropriate choice for fingerprint samples (see Chapter 4), in which minutiae can be regarded as landmarks within the image, we anticipate that for facial images the textural information is more relevant than the geometric information related to facial landmarks. Therefore, we combine the FV representation with a compact version of BSIF, extracted from local patches of the facial image (Sect. 5.1.1). We also extend the assumption that unknown PAI species share more texture, shape and appearance features with known PAI species than with those BP samples. Hence, the FV representation would allow tackling the aforementioned issues on PAD generalisation to unknown PAI species. Furthermore, we analyse in this Chapter the PAD performance of several facial regions such as the mouth, nose, and eyes (Sect. 5.2) and study the sensitivity of PAD algorithms to images of varying resolutions (Sect. 5.3). In general, we summarise the results in [71–73] and answer the **RQ 2**, **RQ 3**, and **RQ 4**.

5.1 APPLICATION OF FV FOR FACE PAD

5.1.1 *Improved Local BSIF*

Usually, PAIs include details (e.g. acute edges around the eyes in CASIA cut attacks [236]) which can be successfully detected by the quantization of filtered features (see Fig. 5.2). Therefore, we directly explore the combination of the BSIF features with the best generalisable common feature space (i.e., FV). As mentioned in Sect. 3.1.8, BSIF [104] is a local image descriptor computed by binarising the responses of a given image to a set of pre-learned filters to obtain a statistically meaningful representation of the data. However, the BSIF features are transformed into a high-dimensional vector as the number of filters N increases, not suitable for a combination with the FV representation. The histograms also become sparse vectors as the number of linear filters N increases as they are densely extracted following the strategy described in Sect. 3.1.1. Hence, we computed the number of zero and non-zero components per number of filters over the CASIA Face Anti-Spoofing database [236] in Fig. 5.3-a) and noted that the number of non-zero components remains under 223 in all cases, having an average value of 128. We will thus represent each 2^N BSIF histogram as a 128-component vector by summing the

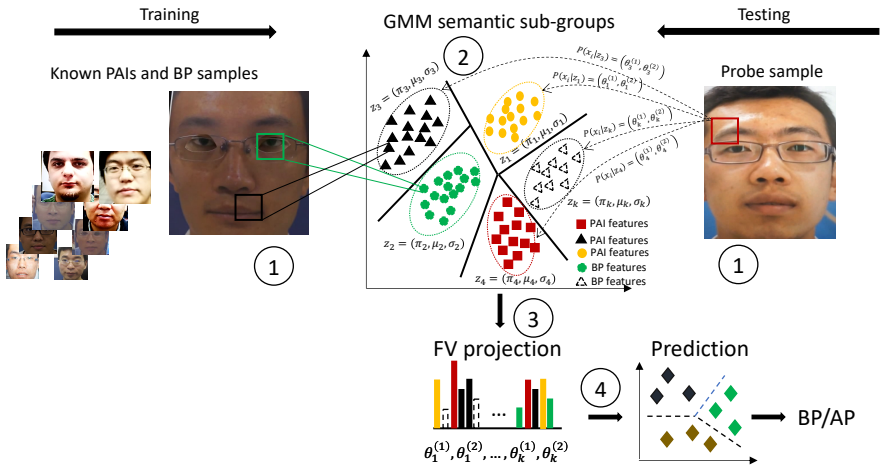


Figure 5.1: Face PAD approach overview which comprises three steps: *a*) local BSIF features are densely extracted per RGB channel, *b*) the feature distribution (i.e., semantic sub-groups) is subsequently learned by training an unsupervised GMM, *c*) the loglikelihood among the BSIF components and the parameters of the semantic sub-groups from the facial feature vector are computed; and *d*) the face representation is classified using a linear SVM.

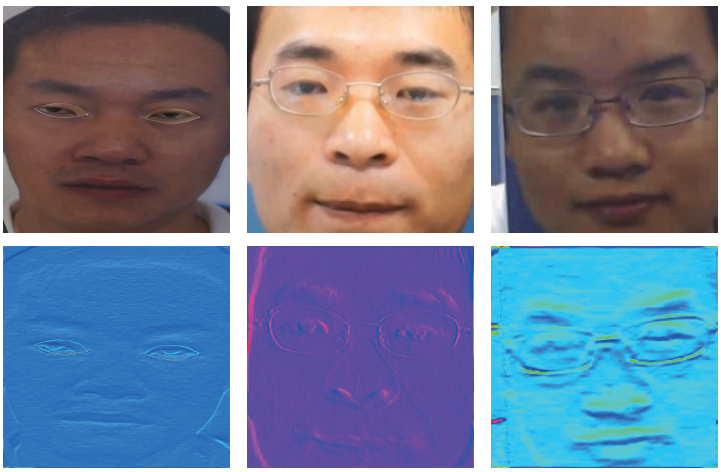
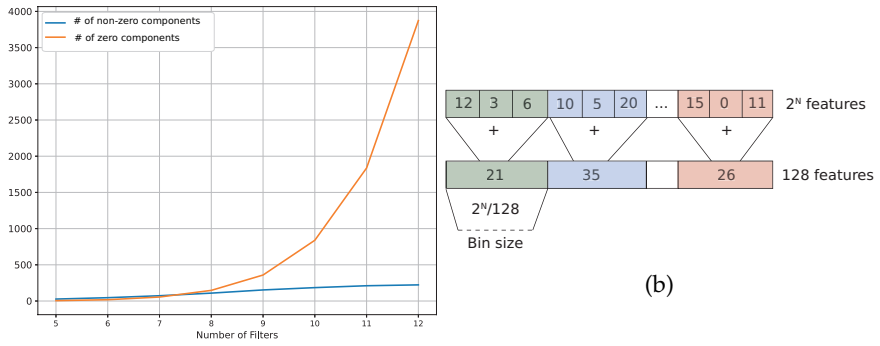


Figure 5.2: Visualisation of the artefacts on three PAI species after convolving the face image with a particular BSIF filter.

elements for each sequential $2^N/128$ sub-set in the original histogram (see an example in Fig 5.3-b). This representation reduces the storage requirements down to 12.5% for N = 10 or 3.1% for N = 12 while leading to a high PAD performance (see Sect. 5.5).

5.2 FACIAL REGION ANALYSIS FOR PAD

A peculiarity of most PAD approaches in Face Recognition (FR) systems is that they detect AP attempts through the analysis of the full



(a) (b)

Figure 5.3: *a*) Average number of zero and non-zero components of dense BSIF histograms for different numbers of filters N, and *b*) a reduction example where a local BSIF histogram of size $2^N = 512$ is represented as a 128-component vector.

face region, thus ignoring facial occlusions by different accessories, as shown in Fig. 5.4. In particular, the use of accessories such as masks to prevent respiratory infection, glasses, or traditional clothes have resulted in a detection performance deterioration of most PAD algorithms. In fact, those PAD methods [23, 105, 181, 228], which have demonstrated the advantage of local face patches in defending against a variety of PAI species, might drop their performance in detecting BPs when pristine local patches contain some of the aforementioned accessories. Thus, these approaches might also fail to correctly separate an occlusion from an intentional AP attempt. Up to now, few approaches have evaluated the impact of some aforementioned occlusions for PAD. Fang *et al.* [55] evaluated the effect of masked attacks on the detection performance of seven state-of-the-art PAD schemes, showing that real masks pose a serious threat for operational FR systems: PAD methods assessed wrongly classified up to 48.25% masked BP samples as APs. Subsequently, Fang *et al.* defined in [54] partial attack labels and use them for training two state-of-the-art architectures (i.e., DeepPixelBis [65] and MixFaceNets [20]). The experimental results demonstrated a detection performance improvement for masked attacks with respect to [55]: relative BPCER enhancements of 5.16% and 85.28% are achieved by DeepPixelBis and MixFaceNets, respectively. Despite the detection performance improvements attained, there exist still an uncountable number of accessories which might drop the accuracy of the PAD subsystems.

To fill this gap in the literature, we abstract from the fact that the input facial samples could contain some of the mentioned occlusions and present a comprehensive analysis of the feasibility of several facial regions for PAD. These facial regions could, in turn, be used in the presence of these occlusions without scarifying the final detection performance.

Figure 5.4: Examples of web-collected facial images occluded by different accessories such as masks, glasses, hands, paper, and tattoos.

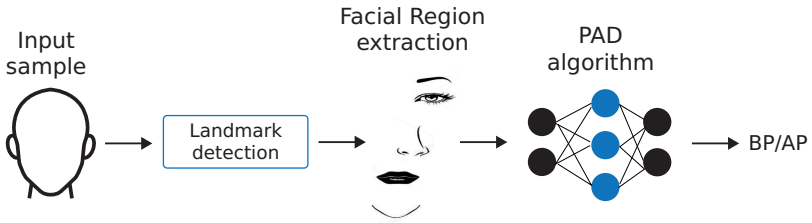


Figure 5.5: Proposed framework to conduct the facial region study and evaluate PAD approaches.

5.2.1 *Proposed Framework*

In our study, we explore the feasibility of using 14 facial regions for PAD purposes: both eyes, both eyebrows, central face, chin, jaw, left eye, right eye, left eyebrow, right eyebrow, mouth, nose, left face, right face regions. Fig. 5.5 shows the framework proposed to conduct our analysis, which is based on two main steps: *i*) the facial region is detected and extracted (see Sect. 5.2.2), and *ii*) the face region is the input to a PAD approach (see Sect. 5.2.3) for BP vs. AP decision.

Table 5.1: Definition of facial regions by landmarks.

	Region	Enclosing landmarks
1	Full Face	The entire face region
2	Left Face	[8, 16, 24, 27]
3	Right Face	[0, 8, 19, 27]
4	Central Face	[0, 16, 19, 24, 30]
5	Jaw	[1, 8, 15, 28]
6	Both Eyebrows	[17, 19, 24, 26]
7	Both Eyes	[0, 16, 17, 26, 28]
8	Left Eyebrow	[22, 24, 26]
9	Right Eyebrow	[17, 19, 21]
10	Left Eye	[16, 26, 27, 28]
11	Right Eye	[0, 17, 27, 28]
12	Mouth	[48, 50, 52, 54, 57]
13	Nose	[27, 31, 33, 35]
14	Chin	[5, 8, 11, 57]

5.2.2 *Facial Regions Extraction*

For facial region detection and extraction, we consider the open-source toolbox dlib [106] which extracts 68 landmarks per face. Based on such landmarks, we define 14 different facial regions in Tab. 5.1. For a comprehensive analysis, we divide these regions into two groups: single (i.e., mouth, nose, chin, left eye, right eye, left eyebrow, and right eyebrow) and composite (i.e., both eyes, both eyebrows, central face, jaw, left face, and right face, full face). Fig. 5.6 shows an example of those landmarks together with some facial regions.

5.2.3 *PAD Methods*

Five state-of-the-art CNN approaches are independently evaluated: *i*) AlexNet [120] which outperformed all traditional machine learning and computer vision approaches in the ImageNet challenge [120], *ii*) DenseNet [91] comprising 121 layers, *iii*) ResNet [86] version with 101 layers, *iv*) MobileNetV2 [177] proposed mainly for mobile applications, and *v*) a recent lightweight CNN named MNasNet [198].

In our implementation, the last Fully Connected Layer (FCL) for all deep learning architectures studied is modified to a single neuron with a sigmoid activation for the BP vs. AP binary decision. We train the algorithms using the Adam optimiser [107] and use the ImageNet pre-trained weights to initialise the networks. A learning rate of 10^{-4}

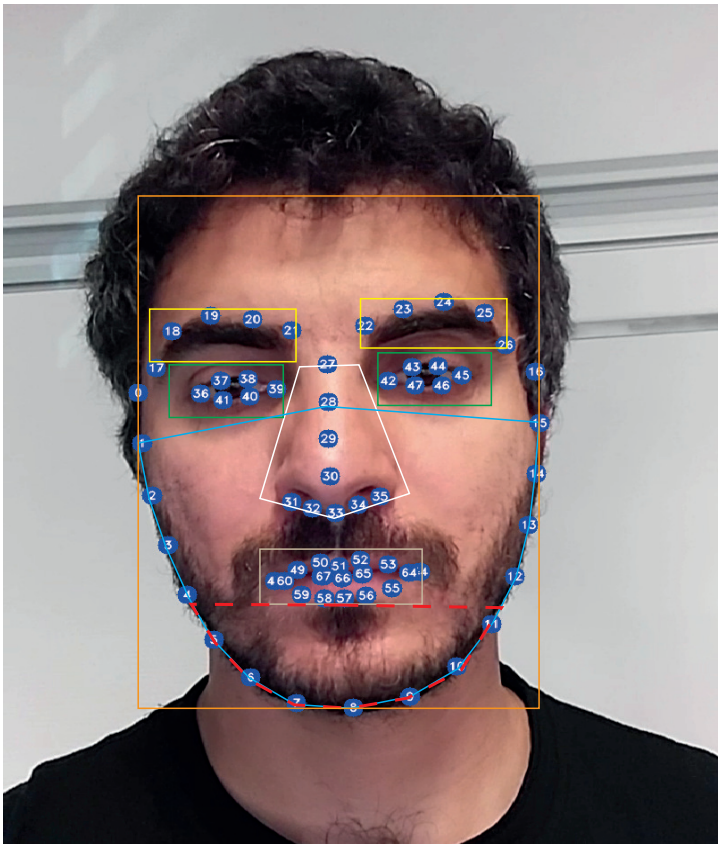


Figure 5.6: Examples of some facial regions (i.e., mouth, nose, left and right eyes, left and right eyebrows, chin, jaw, and full face).

with a weight decay parameter of 10^{-6} was used. The framework was implemented in PyTorch [155] and the CNNs are trained on the Nvidia GPU Tesla M10 with 16 GB DRAM.

5.2.4 *Facial Region Utility*

We define a new metric named *Facial Region Utility* which combines the correlation between facial regions and the detection performance of algorithms when they are trained using a particular facial region and evaluated on another one. This metric reports a value in the range $[0, \ldots, 1]$ which indicates the usefulness of a particular region for training to spot an AP based on the other region in a probe image. Formally, the *Facial Region Utility* for a probe facial region R_P with respect to a trained region R_T is defined as follows:

$$U(R_T, R_P) = \frac{|C(R_T, R_P)| + (1 - P(R_T, R_P))}{2}, \tag{5.1}$$

where $C(R_T, R_P)$ is the Pearson correlation coefficient between R_T and R_P. $C(R_T, R_P)$ reports a value in the range $[-1, 1]$ indicating how correlated the features of R_P are with those of R_T. Since the

Table 5.2: Benchmark of state-of-the-art approaches in terms of classification accuracy (%) under the Quality Test and Overall Test protocol in [236].

method	low	normal	high	overall
LBP+SVM [162]	**83**	78	**90**	80
Network A [171]	**84**	**91**	79	80
Network B [171]	**86**	**93**	80	81
Network C [171]	**94**	**94**	82	87
ShallowCNN [163]	**93**	**92**	84	88

direction of the Pearson correlation between R_T and R_P does not lead to any improvement, we apply the absolute value over the coefficient. $P(R_T, R_P)$ represents the normalised D-EER when R_P is evaluated using an algorithm trained over the region R_T. Utility values close to 1 state that R_T can be employed for training whilst R_P can be successfully used for detecting an AP in the probe image. To normalise the D-EER values to the range $[0, 1]$, we employ the traditional Min-Max normalisation [180]:

$$normalised_{\text{D-EER}} = \frac{\text{D-EER} - \mathbf{min}_{\text{D-EER}}}{\mathbf{max}_{\text{D-EER}} - \mathbf{min}_{\text{D-EER}}}, \tag{5.2}$$

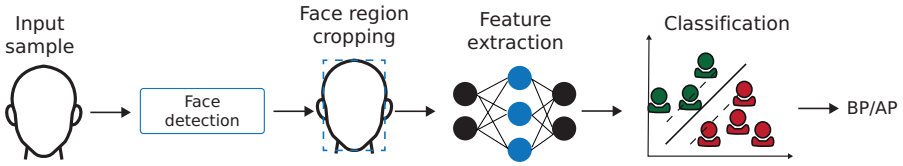
where $\mathbf{min}_{\text{D-EER}}$ and $\mathbf{max}_{\text{D-EER}}$ are, respectively, the minimum and maximum values of the set of D-EERs computed by the evaluated PAD methods (see Sect. 5.2.3) on different training and testing configurations of the facial regions.

To make the equation 5.1 clear to readers, we show the boundary cases. Let **A** be a PAD algorithm, for the best case, we assume that the performance of **A** on two face regions (i.e., $P_{\mathbf{A}}(R_T, R_P)$) would result in a D-EER = 0.0, and R_T and R_P are highly correlated (i.e., $C(R_T, R_P) = 1$). Therefore, the *Facial Region Utility* between R_T and R_P would achieve the highest value (i.e., $U(R_T, R_P) = 1$). On the contrary, for the worst case, $P_{\mathbf{A}}(R_T, R_P) = 1,0$ and $C(R_T, R_P) = 0$, leading to a $U(R_T, R_P) = 0$.

5.3 SENSITIVITY TO IMAGES OF VARYING RESOLUTIONS

In a third approach, we study the sensitivity of facial PAD techniques to images of varying resolutions. Nowadays, most PAD algorithms have reported a performance degradation when they are trained with face images of varying resolutions, as shown in Tab. 5.2. Up to now, very few works have addressed these issues. In 2013, Galbally *et al.* [61] evaluated the potential of general IQA as a protection tool against

Figure 5.7: Proposed framework to analyze the effect of images with varying resolutions.



PAIs and showed that a face sample acquired in an attack attempt has different quality than a BP image. Following that idea, Bhogal *et al.* [13] also explored six non-reference IQA metrics to detect APs on iris, fingerprint, and facial characteristics. As a result, the authors found that the best quality measure and classification setting highly depends on the target database, thereby recommending its optimisation for each particular application. More recently, Agarwal *et al.* [3] showed how several image transformations such as gamma correction, log transform, and brightness control can help a non-authorised subject to circumvent a PAD algorithm. In addition, the authors demonstrated that such image transformations decrease the detection performance of handcrafted- and deep learning-based approaches.

In spite of those valuable efforts, one main question remains unanswered: could the utilisation of images with varying resolutions affect the detection performance of any PAD method? In other words, to which extent are PAD approaches sensitive to training sets containing images of varying quality? And how could this difficulty affect the PAD generalisation capabilities?

5.3.1 *Proposed Framework*

In order to address the above questions, we analyse different PAD algorithms following the three-step overview depicted in Fig. 5.7: *i*) to remove non-useful information, faces are first detected with the Tensorflow[1] Face Detection method, *ii*) a global texture descriptor is then extracted per image, and *iii*) a BP or AP decision is finally taken by a linear SVM. To extract global features from images, we select three well-known texture descriptors: LBP, LPQ, and BSIF, and five deep learning approaches (i.e., MobileNet [88], MobileNetV2 [177], InceptionV3 [190], Xception [29], and DenseNet121 [91]). All of these descriptors have been widely employed for face PAD [8, 31, 143] and are summarised in Tab. 5.3. In our implementation, we use the ImageNet [41] pre-trained deep learning models, and the final descriptor is computed from the last layer after removing the FCLs.

1 https://github.com/yeephycho/tensorflow-face-detection

Table 5.3: Summary of the descriptors used in the image resolution analysis.

Handcrafted descriptors			Deep learning descriptors	
descriptor	**parameters**	**length**	**architecture**	**length**
LBP [145]	R = {1, 2, 3} P = {8, 16, 24}	59	MobileNet [88] MobileNetV2 [177]	$7 \times 7 \times 1024 = 50176$ $7 \times 7 \times 1280 = 62720$
LPQ [146]	R = {3, 5} $\alpha = 1$	256	InceptionV3 [190] Xception [29]	$5 \times 5 \times 2048 = 51200$ $7 \times 7 \times 2048 = 100352$
BSIF [104]	N = {5, 6, 7, 8, 9, 10, 11, 12} L = {3, 5, 7, 9, 11, 13, 15, 17}	2^N	DenseNet121 [91]	$7 \times 7 \times 1024 = 50176$

5.4 EXPERIMENTAL SETUP

The experimental evaluation aims to address the following goals:

- Analyse the impact of different BSIF filter configurations on the PAD performance of our common feature space FV.
- Study the detection performance for different colour spaces (i.e., RGB, HSV, and YC_bC_r).
- Analyse the feasibility of using different facial regions for PAD.
- Study the effect of using image with varying resolutions in the detection performance of PAD subsystems.
- Evaluate the correlation and detection performance of facial regions as well as their utility for being used on real applications where some face parts might be occluded
- Benchmark the detection performance of the FV approach with the top state-of-the-art for known PAI species, unknown PAI species, and cross-database.
- Establish a benchmark of state-of-the-art using our proposed facial region analysis for a real application where subjects wore masks to prevent respiratory infections.

5.4.1 *Databases*

In order to reach our goals, the experimental evaluation was conducted over five well-established databases, which are summarised in Tab. 5.4:

- **CASIA-FASD** [236] contains 600 short videos of BPs and APs stemming from 50 different subjects, and acquired under different conditions. The dataset comprises three PAI species: *i*) warped photo attacks or printed attacks, cut photo attacks, and video-replay attacks.

- **REPLAY-ATTACK (RA)** [28] consists of 1200 short videos (around 10 seconds in mov format) of both BPs and APs of 50 different subjects, acquired with a 320 × 240 low resolution webcam of a 13-inch MacBook Laptop. The video samples were recorded under two different conditions: *i*) controlled, with an uniform background and artificial lighting, and *ii*) adverse, with natural illumination and non-uniform background. In addition, this database comprises three PAI species: printed attacks, photo-replay attacks, and video-replay attacks.

- **REPLAY-MOBILE (RM)** [35] comprises 1190 video clips of printed, photo-replay, and video-replay attacks of 40 subjects under different lighting conditions. Those videos were recorded with two smartphone capture devices: an iPad Mini2 and a LG-G4 smartphone, thereby allowing the evaluation of PAD approaches for the mobile scenario.

- **MSU-MFSD** [219] contains 440 video clips of photo-replay and video-replay attacks of 35 subjects. Those PAI species were acquired with two camera types: MacBook Air 13-inch and front-camera in the Google Nexus 5 smartphone. The MSU-MFSD database comprises two particular scenarios: *i*) a mobile phone is used to capture both bona fide presentations and presentation attacks, simulating the application of mobile phone unlock, and *ii*) the printed photos used for attacks are generated with a state-of-the-art colour printer on larger sized paper.

- **SiW-M** [128] consists of 968 videos of 13 PAI species including challenging attacks such as silicone masks, obfuscation, and cosmetic makeup, among others. 660 BP videos from 493 subjects are also included in the dataset. Those subjects are diverse in ethnicity and age, and the videos were collected in 3 sessions: *i*) a room environment where the subjects were recorded with few variations such as pose, lighting and expression; *ii*) a different and so larger room where the subjects were recorded with lighting and expression variations; and *iii*) a mobile phone mode where the subjects are moving while the phone camera is recording. Extreme pose angles and lighting conditions are also introduced. As mentioned in Chapter 1, the impersonation attacks are the focus of our Thesis. However, we evaluate on this database the feasibility of our proposed common feature space (i.e., FV) for concealing attacks such the obfuscation.

- **Collaborative Real Mask Attack (CRMA)** [55] consists of 423 BP videos and 12690 attacks of 47 subjects. The videos were acquired with three different high-definition capture devices on realistic scenarios. The PAI species are *i*) both unmasked (BM0) and masked (BM1) bona fide presentations, *ii*) printed and video

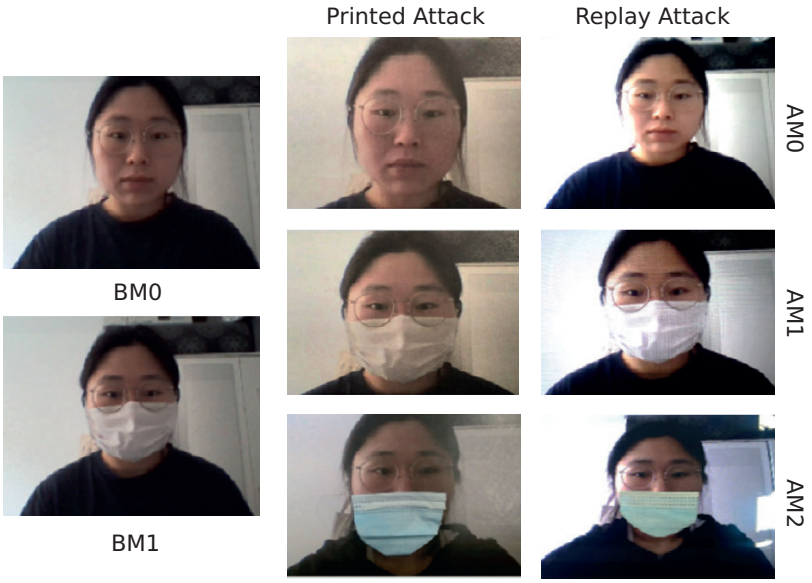


Figure 5.8: Example of BPs and APs in the CRMA database taken from [55].

replay attacks from subjects not wearing a mask (AM0), *iii*) printed and video replay attacks from subjects wearing a mask (AM1), and *iv*) partial attack where the unmasked printed/replayed faces are covered with real masks (AM2). The CRMA is challenging due to different face masks, multiple capture devices, and several capture distances. An example of different BP and AP samples is depicted in Fig. 5.8.

- **OULU-NPU** [19] consists of 4950 high-resolution short video sequences of BP and AP attempts stemming from 55 subjects. The BP samples were acquired in three different sessions under different illumination conditions and background scenes. The PAI species are printed and video-replay attacks which were recorded using the frontal cameras of six mobile phones. This database defines four different protocols as follows:
  - Protocol 1 focuses on the generalisation ability of PAD techniques across different environment conditions (i.e., illumination and background scenes).
  - Protocol 2 is designed to evaluate the PAD generalisation ability when the tested PAI species remain unknown from the training set.
  - Protocol 3 analyses the capture device interoperability following a Leave One Camera Out (LOCO) protocol, where samples recorded by five smartphones are used for training whilst videos captured by the sixth mobile device are used in the evaluation.
  - Protocol 4 is the most challenging scenario, as it combines all described protocols. In particular, the generalisation

ability of PAD approaches across previously unknown illumination conditions, background scenes, PAI species, and capture devices are simultaneously evaluated.

Since most databases contain videos, we select a random frame per video to conduct our experiments.

Table 5.4: A summary of databases considered in our experiments for facial characteristics.

DB	#Samples	Capture device	Capture conditions	PAI species
CASIA-FASD	600	Low-quality USB camera Normal-quality USB camera High-quality Sony NEX-5 camera	Natural scenes	Printed attacks, Cut photo, Video replay
REPLAY-ATTACK	1,200	Low-quality 13-inch MacBook webcam	Controlled, adverse scenes	Printed, Photo replay, Video replay
REPLAY-MOBILE	1,190	High-quality iPad Mini 2 High-quality LG G4	Controlled, adverse direct sunlight, lateral sunlight, diffuse and complex backgrounds	Printed, Photo replay, Video replay
MSU-MFSD	440	Low-quality 13-inch MacBook webcam Low-Quality Google Nexus 5 camera	Natural scenes	Printed, Video replay
SiW-M	968	High-quality Logitech C920 webcam High-quality Canon EOS T6	Controlled adverse scenes	Printed, Video replay, Half mask, Silicone mask, Transparent, Papercraft, Mannequin, Obfuscation, Impersonation, Cosmetic, Funny Eye, Paper Glasses, Partial Paper
CRMA	13,133	iPad Pro, Galaxy Tab S6, Surface Pro-6	Realistic scenes	Printed and Video replay of subjects wearing masks
OULU-NPU	4,950	Samsung Galaxy S6 edge, HTC Desire EYE, MEIZU X5, ASUS Zenfone Selfie, Sony XPERIA C5 Ultra Dual, OPPO N3	Controlled and adverse scenes	Printed and Video replay

5.5 RESULTS AND DISCUSSION

5.5.1 *known PAI species*

5.5.1.1 *Effect of the Semantic Sub-groups*

Following the experimental evaluation in Chapter. 4, we first need to find the optimal configuration of our proposed common feature space in terms of the key parameters: the filter size l, the number of BSIF filters N, and the number of semantic sub-groups K. Following the overall protocol provided by the datasets [28, 35, 219, 236], we compute the D-EER for each of sixty filter configurations (i.e., one error rate for each filter set employed by our improved BSIF) and report in Tab. 5.5 the mean and STD for each fixed K value. As it may be observed, the detection performance of our method increases with K: a D-EER of 0.45% on average is achieved for $K = 1024$, hence this K value will be considered for the remaining experiments. In addition, we may observe that the STD is below 1.0% in all datasets, hence indicating that a statistically meaningful representation of face data can be obtained using different BSIF filters, regardless of the values chosen for N and l.

It should be noted that there is a high difference between the error rates attained for CASIA and the ones achieved for the remaining datasets. Specifically, D-EERs for CASIA are up to 20 times greater than the ones reported for other databases. We think that this divergence is mainly given by the image resolution variation employed for training and testing our approach. Whereas the REPLAY-ATTACK, REPLAY-MOBILE, and MSU-MSFD databases consist of images acquired with fixed low or high-resolution capture devices respectively, face images in CASIA were obtained with a mix of low-, medium-, and high-resolution capture devices. Therefore, we analyse in Sect. 5.5.6 the impact of using image of varying resolutions on the PAD performance. As a result, we anticipate that our approach is, like most PAD algorithms, affected by the image quality varying.

5.5.1.2 *Colour Space Analysis*

According to Boulkenafet *et al.* [18], the RGB colour space has limited discriminative power for face PAD due to the high correlation between the three colour components. In contrast, HSV and YC_bC_r are based on the separation of the luminance and chrominance components, thereby providing additional information for learning more discriminative features. Based on that observation, we evaluate in Tab. 5.6 the detection performance of the proposed FV approach for the three aforementioned colour spaces. As it can be observed and contrary to the conclusions drawn in [18], RGB appears to be the colour space including the most discriminative features for facial PAD, thereby

Table 5.5: Detection performance, in terms of D-EER (%), of our proposed common feature space for different K values. The best result is highlighted in bold.

DB \ K	256	512	1024
CASIA-FASD	2.02 ± 0.93	1.95 ± 0.79	1.79 ± 0.82
REPLAY-ATTACK	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
REPLAY-MOBILE	0.01 ± 0.03	0.00 ± 0.00	0.00 ± 0.00
MSU-MFSD	0.02 ± 0.09	0.01 ± 0.08	0.01 ± 0.08
Avg.	0.51	0.49	**0.45**

resulting, on average, in a D-EER of 0.45%. However, taking a closer look, we can observe that the three colour spaces report similar error rates in three out of four datasets (i.e., REPLAY-ATTACK, REPLAY-MOBILE, and MSU-MFSD): mean D-EERs of 0.003%, 0.03%, and 0.19% are achieved by RGB, HSV, and YC_bC_r respectively.

In order to validate the detection performance achieved by our proposed method using current colour spaces, we select the non-parametric Mann-Whitney test [132] with a 95% of confidence to verify the statistical significance of the sixty error rates reported by different colour spaces. To that end, we define the null hypothesis H_0 and alternative hypothesis H_1 as:

- H_0: two colour spaces provide the same discriminative information for face PAD.
- H_1: two colour spaces do not provide the same discriminative information for face PAD.

Then, an all-against-all comparison per dataset is performed. As a result of this test, we do confirm that the RGB only provides the most discriminative information for one out of four databases: error rates attained by the RGB claim to be statistically higher than the ones reported by the other colour spaces for the CASIA database. In contrast, for the three remaining databases (i.e., REPLAY-ATTACK, REPLAY-MOBILE, and MSU-MFSD), Mann-Whitney results state that the three colour spaces include the same discriminative information, thereby confirming their similar detection performances reported in Tab. 5.6. The reason for this difference with respect to [18] is that we carried out a feature decorrelation with PCA before finding the semantic sub-groups, thereby leading to the detection of similar features for the three colour spaces.

Table 5.6: Detection performance in terms of D-EER (%) of the FV for the best performing $K = 1024$.

DB \ Colour	RGB	HSV	YC_bC_r
CASIA-FASD	1.79 ± 0.82	2.35 ± 1.07	2.20 ± 1.05
REPLAY-ATTACK	0.00 ± 0.00	0.02 ± 0.05	0.10 ± 0.26
REPLAY-MOBILE	0.00 ± 0.00	0.00 ± 0.00	0.12 ± 0.26
MSU-MFSD	0.01 ± 0.08	0.08 ± 0.35	0.35 ± 0.81
Avg.	**0.45**	0.79	0.69

Table 5.7: Benchmark with state-of-the-art in terms of D-EER (%) for the known PAI species scenario using $K = 1024$ on RGB. The best results per database are highlighted in bold.

Method	CASIA	RA	MSU	RM
BSIF-SVM [164]	10.21	-		-
MBSIF-TOP [7]	7.20	-		-
CSURF + FV [16]	2.80	0.10	2.20	-
Texture fusion [18]	4.60	1.20	1.50	-
Depth CNNs [9]	2.67	0.72	0.35 ± 0.19	-
ResNet-15-3D [84]	2.22	0.25	-	-
FaceSpoofBuster [21]	3.88	5.50	-	-
shallowCNN-LE [163]	4.00	3.70	8.41	-
DR-UDA [211]	3.30	1.30‡	6.30	-
SPMT + SSD [185]	0.04	0.03	-	-
DeepPixBiS [65]	-	-	-	**0.00**
WeightedAvg. [56]	-	1.43	-	9.95
HR-CNN [139]	**0.02**	-	0.04	-
Our Method (FV)	1.79 ± 0.82	**0.00 ± 0.00**	**0.01 ± 0.08**	**0.00 ± 0.00**
Best D-EER†	0.37	0.00	0.00	0.00
$BPCER_{100}$	0.00	0.00	0.00	0.00

† The best D-EER as well as the $BPCER_{100}$ per dataset are attained for $N = 10$ filters of size $l = 9$.
‡ Half Total Error Rates (HTER) reported in [211]

5.5.1.3 *Benchmark with the State of the Art*

Finally, we benchmark in Tab. 5.7 our FV with the top state-of-the-art PAD techniques for the best performing colour space and K value (i.e., RGB and $K = 1024$). On the one hand, it can be observed that a baseline implementation based on BSIF and SVMs in [7], where extracted features have not been transformed with the FV technique,

reports a D-EER of 10.21%, in contrast to the best D-EER achieved in this work for CASIA (i.e., 0.37%).

On the other hand, it may be also observed that the FV representation does not produce a reliable detection performance: it depends on a good feature extractor for the specific data domain. Specifically, the combination of FV and SURF descriptors, which has shown a remarkable detection performance for fingerprint PAD in Chapter. 4, achieves D-EERs of 2.80% and 0.10% for CASIA and REPLAY-ATTACK respectively, which are still far away from the ones reported in this work. Therefore, we can conclude that the use of the improved BSIF descriptors presents a clear advantage for facial PAD with respect to gradient-based features such as SURF.

In addition, the texture fusion approach in [18] can be also outperformed by a relative 96% and 100% respectively, depending on the testing database. Among the next five deep learning-based techniques, the lowest D-EER reported are 2.22% and 0.25%, which are also twelve and twenty-five times worse than our best results for CASIA and REPLAY-ATTACK databases, respectively. In contrast, the last two approaches analysed [139, 185] outperform our technique by one order of magnitude for the CASIA database. However, our best result is three-time better than the ones reported in [185] for REPLAY-ATTACK (i.e., 0.00% vs 0.03%). It should be noted that the authors of those works admit that their PAD approaches are time-consuming methods. Very low computational cost is an additional advantage of our approach, which needs about 0.7 seconds per single attempt transaction (i.e., image load, feature extraction, and AP decision), thereby making it suitable for real-time applications. Finally, we can observe that for a high security threshold (i.e., APCER = 1.0%), our proposed method reports a remarkable BPCER of 0.0% for all databases: only one in 100 AP attempts are accepted while zero BPs are rejected by our algorithm when PAI species and capture devices employed in the PAI acquisition are known a priori.

5.5.2 *Unknown PAI species*

As it has been mentioned in this Thesis, one of the main goals is to address the detection of unknown PAI species. In particular, we tackle the challenging scenario where PAI species remain unknown in the training set of PAD techniques. To that end, two sets of experiments are carried out over the five selected databases following the LOO protocol described in [8]: one PAI species is evaluated while the remaining PAI species are included in the training set.

5.5.2.1 *Generalisation across Traditional unknown PAI species*

In the first set of experiments, we evaluate the feasibility of our FV method to detect unknown PAI species over traditional PAI species

(i.e., printed attacks, cut photo attacks, and photo and video replay attacks). The corresponding results are reported in Tab. 5.8. It should be noted that error rates for each particular unknown PAI species in CASIA are multiplied by a factor of 2.17% on average with respect to the corresponding D-EERs reported in Tab. 5.7 (i.e., 3.88% vs. 1.79%). In contrast, theD-EERs for the remaining datasets are comparable with their corresponding error rates for the known PAI species scenario. These observations confirm the sensitivity of our approach to training with datasets having images of varying resolutions.

Table 5.8: Benchmark with the state-of-the-art in terms of the AUC (%) for $K = 1024$ and RGB over traditional unknown PAI species. The best results per PAI species are highlighted in bold.

	CASIA			REPLAY-ATTACK			MSU-MFSD			REPLAY-MOBILE		
	Cut	Warped	Video	Digital	Printed	Video	Printed	HR Video	Mobile Video	Digital	Printed	Video
OC-SVM_RGB+BSIF [8]	60.70	95.90	70.70	88.10	73.70	84.30	64.80	87.40	74.70	-	-	-
NN+LBP [221]	88.40	79.90	94.20	95.20	78.90	99.80	50.60	99.90	93.50	-	-	-
DTN [128]	97.30	97.50	90.00	99.90	99.60	99.90	81.60	99.90	97.50	-	-	-
CDCN [229]	99.90	99.80	98.48	99.43	99.92	**100**	70.82	**100**	99.99	-	-	-
TTN-S [218]	**100**	**100**	99.57	**100**	**100**	**100**	87.06	**100**	94.50	-	-	-
our FV (AUC)	99.6	97.9	**99.9**	**100**	**100**	**100**	**99.32**	**100**	**100**	**100**	**100**	**100**
our FV (D-EER)*	3.33	6.67	2.22	0.00	0.00	0.00	1.96	0.00	0.00	0.00	0.00	0.00
our FV (mean D-EER)	4.11 ± 1.99	6.15 ± 2.42	1.37 ± 1.60	0.00 ± 0.00	1.35 ± 1.73	0.00 ± 0.00	6.64 ± 4.62	0.00 ± 0.00	0.11 ± 0.44	0.00 ± 0.00	0.34 ± 0.63	0.02 ± 0.12

* The D-EER and AUC values per dataset are reported for $N = 10$ filters of size $l = 9$.

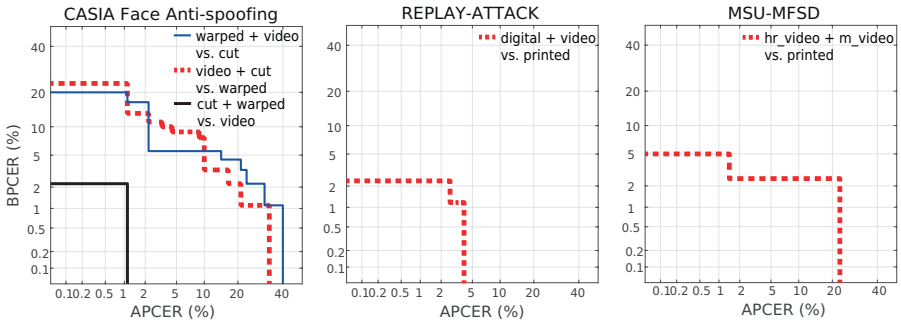


Figure 5.9: Traditional unknown PAI species DET curves over the LOO protocol for the CASIA, REPLAY-ATTACK, and MSU-MFSD databases. The REPLAY-MOBILE database reports a remarkable BPCER = 0.0% for any APCER. unknown PAI species such as digital and video for REPLAY-ATTACK and m_video and hr_video for MSU-MFSD attain a BPCER = 0.0% for any APCER, hence their corresponding curves are not shown.

Regarding the MSU-MFSD database, our method suffers a performance deterioration for printed attacks. Whereas both HR video (i.e., hr_video in Fig 5.9) and mobile video (i.e., m_video in Fig 5.9) attacks report on average a D-EER of 0.0%, printed attacks attain an average error rate of 6.64%, hence indicating that BSIF texture features of the latter are as close to BP semantic sub-groups as the semantic sub-groups defined from video replay attacks. This, in turn, states that the detection performance over unknown PAI species depends on a reliable known PAI species selection for training. Due to the lack of a proper quantitative analysis of the top state-of-the-art methods in compliance with the ISO/IEC 30107-3 standard on biometric PAD [97], we establish a benchmark in terms of AUC. In spite of the previous shortcomings, we can note that for a fixed filter configuration (i.e., $N = 10$ filters of size $l = 9$ pixels), our approach achieves current state-of-the-art results for all datasets, thereby resulting in an AUC close to 100%.

Finally, a high detection performance of our method can be perceived in Fig. 5.9: a BPCER in the range of 0.0% - 17% for any APCER $\geq$ 1.0% confirms the soundness of the common feature space defined by FV to separate an unknown PAI species from a BP attempt.

5.5.2.2 *Generalisation across Challenging unknown PAI species*

In the second set of experiments, we evaluate challenging unknown PAI species such as 3D Masks (i.e., Silicone masks, Transparent masks, and Mannequin Head) and Makeup (obfuscation, impersonation, and cosmetic) in the SiW-M database following the LOO protocol: twelve PAI species are employed for training and the remaining thirteenth PAI species is used for testing. It is worth pointing out that there is no overlap between training and test subjects. Tab. 5.9 reports the D-EER

for $N = 10$ BSIF filters of size $l = 9$ and the best BSIF performing filter configurations. The corresponding DET curves for the latter are depicted in Fig. 5.10. As it may be observed, our best filter performing-based FV representation reports an improvement with respect to the results attained by state-of-the-art methods, thereby yielding a D-EER of 11.44% and a STD of 8.73%. We can also note that this approach attains the top state-of-the-art error rates for the challenging Mask attacks (i.e., D-EER of 9.33%), even though some prior techniques [127, 229] employ additional information such as depth and temporal cues to detect those 3D Mask attacks. In addition, it should be noted that the FV algorithm reports a detection performance deterioration for the BSIF filter setting adopted from the known PAI species evaluation (i.e., $N = 10$ filters of size $l = 9$), thereby resulting in a mean D-EER of 15.86%. Despite the accuracy degradation, this representation is still able to achieve the state-of-the-art schemes, thereby showing its soundness for this scenario. In order to enhance the BSIF computation and remove the dependency on the current 60 filter configurations, we plan as future work to perform the BSIF quantisation over the filters learned by intermediate CNN layers.

Table 5.9: Benchmark with the state-of-the-art for challenging unknown PAI species on RGB for $K = 1024$ in terms of D-EER (%). The best results per PAI species are highlighted in bold.

Methods	Replay	Printed	Mask Attacks					Makeup Attacks			Partial Attacks			Average
			Half	Silicone	Trans.	Papercraft	Manneq.	Obfusc.	Imperson.	Cosmetic	Funny Eye	Paper Glasses	Partial Paper	
Auxiliary [127]	14.00	4.30	11.60	12.40	24.60	7.80	10.00	72.10	10.00	**9.40**	21.40	18.60	4.00	16.95 ± 17.72
DTN [128]	10.00	**2.10**	14.40	18.60	26.50	5.70	9.60	50.20	10.10	13.20	19.80	20.50	8.80	16.12 ± 12.23
DeepPixBis [65]	11.68	7.94	7.22	15.04	21.30	3.78	4.52	26.49	1.23	14.89	23.28	18.90	4.82	12.39 ± 8.25
MCCNN [66]	12.82	12.94	11.33	13.70	**13.47**	0.56	5.60	**22.17**	0.59	15.14	**14.40**	23.93	9.82	12.04 ± 6.92
CDCN++ [229]	**9.20**	5.60	**4.20**	**11.10**	19.30	5.90	5.00	43.50	**0.00**	14.00	23.30	**14.30**	**0.00**	11.95 ± 11.79
FV Method (Optimum)	10.28	7.70	7.98	18.42	17.87	**0.00**	**2.40**	27.93	**0.00**	16.78	17.84	18.22	3.27	**11.44 ± 8.73**
Optimum BSIF filters	$N = 11$	$N = 5$	$T = 7$	$N = 8$	$N = 10$	$N = 5$	$N = 6$	$N = 6$	$N = 6$	$N = 9$	$N = 9$	$N = 11$	$N = 11$	-
	$l = 7$	$l = 3$	$l = 15$	$l = 5$	$l = 7$	$l = 11$	$l = 11$	$l = 11$	$l = 13$	$l = 13$	$l = 13$	$l = 7$	$l = 13$	-
FV Method (Fixed)*	12.49	11.76	14.20	22.94	23.20	5.61	7.19	34.57	1.58	22.07	23.71	23.26	3.65	15.86 ± 9.89

* D-EERs per dataset are reported for $N = 10$ filters of size $l = 9$.

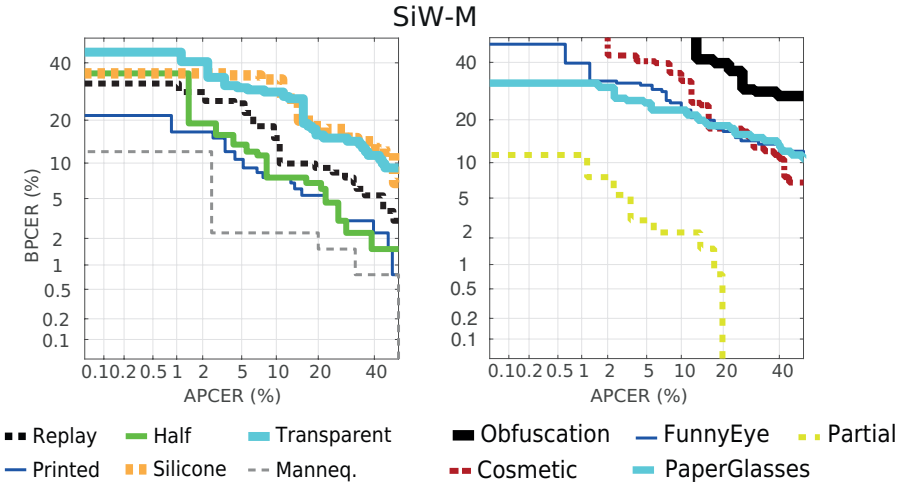


Figure 5.10: Challenging unknown PAI species DET curves over the LOO protocol for the SiW-M database on the best BSIF filter configuration. DET curves for Papercraft and Impersonation attacks are not shown since they report a BPCER = 0.0% for any APCER.

Taking a closer look at Tab. 5.9, we can also see that all PAD techniques report a poor detection performance for the obfuscation attacks included in the SiW-M database: D-EERs are for none of the approaches lower than 22%, which indicates that this is the most challenging PAI species. This is due to the fact that the makeup applied over the faces is subtle and hence looks like real human faces. Given that the majority of subjects in the obfuscation set are not in the BP dataset, a proper evaluation reporting the impact of those cosmetic beautifications on a real face recognition system cannot be carried out [168]. The main question to address the threat of a given PAI is whether it is able to change the appearance of the subject enough to lead to a False-Non-Match. However, other studies about the impact of similar obfuscated images on real deep face recognition systems have reported a high biometric performance: a reliable False Non-Match Rate (FNMR) of 7.80% at a False Match Rate (FMR) of 0.1% for ArcFace [42] and a remarkable FNMR of 1.60% at a FMR = 0.01% for a new ArcFace variant [184] indicate the low dangerousness of those PAI species for facial biometric systems. Based on these observations, we think that those attacks should not be taken into account for PAD training since they can negatively impact the detection of another PAI species (e.g., Transparent Masks, see Tab. 5.9). In other words, we think that by excluding obfuscation attacks from the training set, we could significantly improve the detection performance of current PAD techniques.

Finally, we observe in Fig. 5.10 that our approach reports an average BPCER of 21.53% for the challenging mask attacks: only one in 100 AP attempts are accepted while at most 22 in 100 BPs are rejected by our PAD system. In addition, it should be noted that the proposed method

Table 5.10: Benchmark with the state-of-the-art in terms of the D-EER (%) for the cross-database scenarios over the best BSIF filter configuration. The best results per PAI species are highlighted in bold.

Train	MSU		RA		CASIA		Avg.
Test	CASIA	RA	MSU	CASIA	MSU	RA	
Colour Texture [17]	46.00	33.90	34.10	37.70	24.40	30.30	34.40
Texture fusion [18]	29.20	16.20	21.40	31.20	19.90	**9.90**	21.30
DupGAN [90]	27.10	35.40	36.20	46.50	33.40	42.4	36.83
KSA [125]	**9.10**	33.30	34.90	**12.30**	15.10	39.30	24.00
ADA [209]	17.70	5.10	30.50	41.50	9.30	17.50	20.27
DR-UDA [211]	16.80	**3.00**	29.00	34.20	**9.00**	15.60	**17.93**
FV (Optimum)	24.67	6.57	**11.67**	29.33	12.86	24.36	18.24
Optimum BSIF filters	$N = 7$ $l = 7$	$N = 7$ $l = 3$	$N = 12$ $l = 11$	$N = 10$ $l = 9$	$N = 8$ $l = 13$	$N = 6$ $l = 3$	-
FV (Fixed)*	31.56	25.79	24.29	29.33	33.10	35.71	29.97

* D-EERs per dataset are reported for $N = 10$ filters of size $l = 9$.

achieves a remarkable BPCER of 0.0% for any APCER over these types of impersonation attacks included in SiW-M. This PAI species, unlike the obfuscation attacks in SiW-M, are more challenging for real deep face recognition systems, as they have reported a significant biometric performance deterioration (i.e., FNMR = 47.80% @ FMR = 0.01% [184]).

5.5.3 *Cross-database Evaluation*

Similar to capture devices employed for fingerprint acquisition, most face capture devices age and stop working, hence will be replaced by new sensors for which we have no AP samples for training the PAD systems. Therefore, it is of utmost importance that our PAD methods are robust to those situations. To that end, we select three databases (i.e., CASIA Face Anti-Spoofing, MSU-MFSD, and REPLAY-ATTACK) and establish in Tab. 5.10 a benchmark of our proposed representation with the current state-of-the-art techniques for each training-test configuration. It should be noted that our approach is able to achieve current state-of-the-art results, thereby yielding a D-EER of 18.24% on average for the best BSIF filter configuration. In addition, the best performing deep learning-based scheme for cross-database (i.e., DR-UDA [211]) reports, on average, a D-EER of 17.93%, which is up to twice lower than the worst result reported for this scenario (i.e., 36.83% for DupGAN [90]). In order to improve generalisation cross-database, this method, like DupGAN [90], KSA [125], and ADA [209], is fully based on domain adaptation, which transfers the knowledge learned from a source domain to a target domain. In spite of the results attained for this scenario, the DR-UDA algorithm is unable

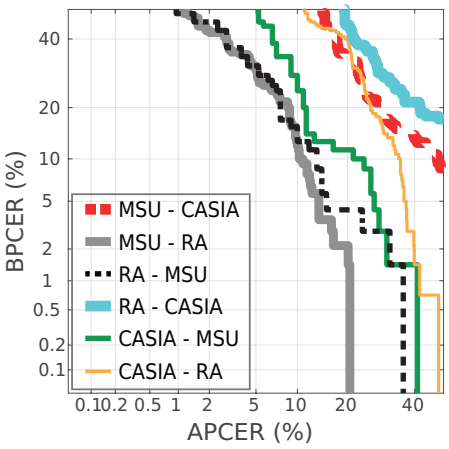


Figure 5.11: DET curves for the cross-database scenario for the best BSIF filter configurations.

to achieve reliable error rates for known PAI species (i.e., D-EERs of 3.20%, 6.00%, and 7.20% for CASIA, MSU-MFSD, and Rose-Youtu databases, respectively).

Consequently with the results reported in Tab. 5.9 for a fixed BSIF filter configuration (i.e., $N = 10$ filters of size $l = 9$) our proposed method decreases its detection performance up to 40%, thereby resulting in a D-EER of 29.97%. This, in turn, states the need of removing the dependency to current BSIF filters in order to keep stable the performance of our algorithm for different PAD scenarios.

On the other hand, it may be observed that our proposed method trained with images of varying resolutions in CASIA performs well for high-resolution face images (i.e., 12.86% for MSU). In contrast, it reports a detection performance decrease of up to 47% for face samples stemming from low-resolution capture devices (i.e., 24.36% for REPLAY-ATTACK). However, by training our approach with high-resolution images in MSU-MFSD, a D-EER of 6.57% can be yielded for those low-quality face images in REPLAY-ATTACK, thereby indicating the need for future studies about the impact of external factors such as image resolution and acquisition conditions over this challenging scenario. Furthermore, unlike current PAD techniques in the literature, a reliable D-EER of 11.67% for high-resolution face images can be attained by tuning our proposed PAD algorithm with low-quality images in REPLAY-ATTACK. These results confirm that PAIs in MSU and REPLAY-ATTACK contain similar artefacts which can be successfully represented by the semantic sub-groups learned by the GMM.

Finally, it should be noted in Fig 5.11 that the proposed algorithm reports a detection performance deterioration for high-security thresholds: a poor average BPCER_{100} of 73.90% confirms the need for new interoperable PAD schemes in order to improve their generalisation capabilities for this scenario without losing accuracy for the remaining scenarios.

5.5.4 *Computational Complexity*

In order to report the computation complexity of our proposed method, we focus on the four main steps carried out to analyse whether a sample at hand is either a BP or AP: *i*) the extraction of compact BSIF histograms, *ii*) PCA projection of such histograms, *iii*) the loglikelihood computation through FV, and *iv*) SVM decision.

5.5.4.1 *Compact BSIF Histograms*

Let $I(x,y)$[2] $\in \mathbb{R}^{3\times M\times M}$ be an input RGB image and W a set of N filters of size $l \times l$. As it was mentioned, in the first step our algorithm extracts compact BSIF histograms from several points sampled over a regular grid on $I(x,y)$. To that end, the facial image $I(x,y)$ is first convolved per RGB channel using each of N filters in W and then quantised by the Eq. 3.3, thereby resulting in $3 \times M \times M \times N \times l \times l \approx M^2 \times l^2 \times N$ operations. Consequently, P histograms are densely extracted per point over the regular grid at different window sizes r, thereby leading to many operations similar to the ones reported by the earlier convolution (i.e., $P \times M^2 \times r^2$). Since both P, N, r, and l take fixed finite values in the ranges: $P = 4$, $N = \{5,\ldots,12\}$, $r = \{4,6,8,10\}$, and l = {5, 7, 9, 11, 13, 15, 17}, the number of operations for the extraction of spatial BSIF histograms is asymptotically bounded by M^2 operations. Hence, its computational complexity is $O(M^2)$.

5.5.4.2 *PCA Projection*

In this second step, the T BSIF histograms of size $d' = 128$ previously extracted for $I(x,y)$ are projected to a low dimensional space $d < d'$ (i.e., $d = 64$) using the PCA base vectors obtained in training time. To that end a matrix multiplication is performed. This has a computational complexity $O(T \times d' \times d)$. By assuming that $T \leq M^2$ and $d' = 128$ and $d = 64$ take fixed finite values, the final computational complexity of the PCA projection is asymptotically bounded by $O(d \times d' \times M^2) \approx O(M^2)$.

5.5.4.3 *The FV Computation*

As it was mentioned, the FV computes the loglikelihood between $T \leq M^2$ descriptors of size d extracted from $I(x,y)$ and the GMM parameters learned in the training: $G_K = \{(\pi_k, \mu_k, \sigma_k) : k = 1\ldots K\}$. To that end, the soft assignment weight (or posterior probability) of

2 We focus on the computational complexity over a square RGB image of size $M \times M$ for a better understanding of the readers. The same reasoning can be applied to rectangular images.

the $i-th$ features x_i is computed for each semantic sub-group (k) as follows [176]:

$$\alpha_i(k) = \frac{\exp\left[-\frac{1}{2}(x_i-\mu_k)'\sigma_k^{-1}(x_i-\mu_k)\right]}{\sum_{j=1}^{K}\exp\left[-\frac{1}{2}(x_i-\mu_j)'\sigma_k^{-1}(x_i-\mu_j)\right]}, \tag{5.3}$$

As it should be noted, the operators involved in the posterior probability computation are the GMM parameters $\{\mu_k, \sigma_k) : k = 1 \dots K\}$ and the descriptor x_i. Whereas, the numerator in the fraction can be computed in $O(d)$, its denominator carries out $K \times d$ operations. Therefore, the computational complexity for $x_i \in \mathbb{R}^d$ is bounded by the maximum between d and $K \times d$, thereby resulting in $O(K \times d)$. Given that this operation is computed for each $x_{i=1,\dots,T}$, the final computational complexity for the soft assignment weight computation is $O(T \times K \times d) \approx O(M^2)$.

Now, the computational complexity for the Eq. 3.11 and 3.12 must be computed. As it should be observed, there are several additions, subtraction, and multiplication operations between vectors which can be sequentially performed. Keeping in mind that the Eq. 3.11 and 3.12 are computed for each $x_{i=1,\dots,T}$ and semantic-sub group $k \in \{1, \dots, K\}$, their computational complexity is bounded by $O(T \times K \times d) \approx O(M^2)$.

5.5.4.4 *SVM Classification*

In the last step, the BP or AP decision is taken by a linear SVM. To that end, the algorithm computes the Eq. 3.14 between the input FV representation $\mathbf{x} \in \mathbb{R}^{1\times 2Kd}$ and the SVM parameters (i.e., $\mathbf{W}' \in \mathbb{R}^{2Kd\times 1}$ and $\mathbf{b}' \in \mathbb{R}$). This has a computational complexity $O(2 \times K \times d) \approx O(K \times d)$

Finally, following the $big-O$ properties, the final computational complexity of our proposed method is asymptotically bounded by the maximum between the computational complexities of the intermediate steps: $O(M^2)$. In a nutshell, the number of operations carried out by our algorithm for the worst case is $O(M^2)$ (i.e., linear to the number of image pixels).

5.5.5 *Visualisation of the FV Representation*

Finally, a t-SNE visualisation in Fig. 5.12 of BP and AP samples in the CASIA database confirms the aforementioned hypothesis, which state that the PAIs share more texture, shape, and appearance features with known PAI species than with those BP samples. Whereas the FV representations of APs (blue, red and yellow) are separated of the BPs (green spots), they are close to each other. However, we can also observe that some PAI species such as warped (yellow) and cut photo

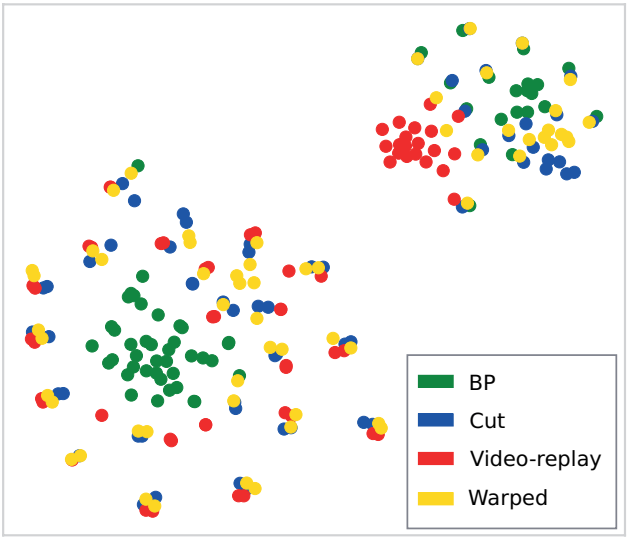


Figure 5.12: t-SNE visualisation for BP vs. AP samples in the CASIA database.

attacks (blue), still overlap with BP samples, thus indicating that the data distribution learned by a GMM model using the BSIF features needs to be improved in order to get a better detection performance.

5.5.6 *Analysis of the Impact of Image Resolution Variation*

As the above results show, our FV common feature space suffers a performance degradation when images of varying resolutions are employed either for training or testing. Following the pipeline in Sect. 5.3.1, we evaluate the sensitivity of using images of varying resolutions on PAD performance. To that end, we select the CASIA database which includes images with the desired quality properties.

5.5.6.1 *Known PAI species: Analysis of a Single Image Resolution*

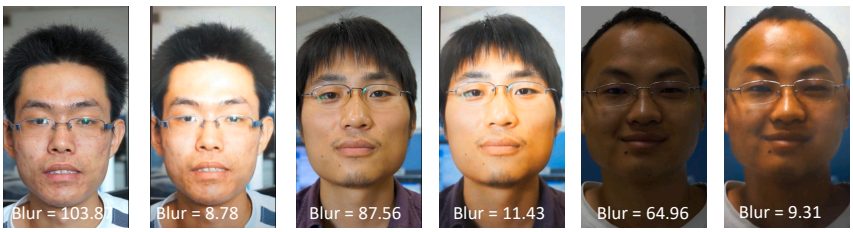
The first set of experiments evaluates the PAD performance under three resolution settings. The corresponding D-EER values are reported in Tab. 5.11. We can first note that for each particular PAI species the error rates attained depend on the image resolution being evaluated. Specifically, D-EERs of 18.01 $\pm$ 1.67, 16.82 $\pm$ 2.67, and 15.65 $\pm$ 6.08 are reported on average for cut, warped, and video-replay attacks, respectively, hence indicating that PAD approaches widely depend on both the PAI species used (different average D-EER) and the resolution settings employed to acquire them (up to 6% STD in the D-EER).

In addition, we observe that the error rates for each PAI species follow different trends depending on the resolution of the images used. Whereas the cut and warped photo attacks achieve their best detection performance on average across all descriptors for face im-

Table 5.11: Average D-EER (%) values under the known PAI species protocol.

Attack	Cut			Warped			Video		
	L	M	H	L	M	H	L	M	H
LBP	20.00	**16.67**	17.78	17.78	18.89	**10.00**	18.89	21.11	**4.44**
LPQ	20.00	**15.00**	20.00	**13.33**	21.67	16.67	22.50	20.00	**1.67**
BSIF	**18.61**	23.56	20.56	16.83	26.03	**12.19**	17.28	20.86	**2.17**
MobileNet	**23.33**	26.67	**23.33**	**16.67**	33.33	36.67	23.33	13.33	**10.00**
MobileNetV2	16.67	**13.33**	**13.33**	**10.00**	16.67	**10.00**	23.33	**6.67**	16.67
InceptionV3	**6.67**	20.00	16.67	16.67	**13.33**	16.67	33.33	33.33	**16.67**
Xception	**16.67**	**16.67**	26.67	**10.00**	16.67	26.67	20.00	**10.00**	16.67
DenseNet121	**10.00**	**10.00**	20.00	10.00	**6.67**	10.00	13.33	**3.33**	6.67
Avg.	**16.49**	17.74	19.79	**13.91**	19.16	17.39	21.50	16.08	**9.37**

Figure 5.13: BP and AP samples with their corresponding blurriness values.



ages acquired with low-resolution capture devices (i.e., D-EERs of 16.49% and 13.91 for cut and warped, respectively), the video-replay shows its best D-EER for high-resolution images (i.e., 9.37%). However, taking a closer look we can note that the handcrafted descriptors LBP and BSIF do perform better with high-quality images not only for video-replay but also for warped attacks. In contrast, the deep learning-based techniques achieve their best detection performance for low to medium quality images for cut and warped attacks, and also for medium quality images for video-replay attacks. To shed some light on these differences, we investigated some intrinsic image properties and confirmed that the screen projection of the video-replay attacks on a high-resolution capture device unveils several blurriness and sharpness artefacts, which are successfully detected by all PAD techniques. In Fig. 5.13 we show some BP and AP samples with their blurriness values, which were computed as the variation of the Laplacian [160].

Table 5.12: Average D-EER values under the known PAI species protocol.

	PAI species \ Quality	Single			Multiple			
		L	M	H	L ∪ M	L ∪ H	M ∪ H	L ∪ M ∪ H
Single	**cut**	**17.94**	21.71	19.64	21.82	24.27	27.16	24.22
	warped	16.02	24.03	**12.41**	23.49	19.74	24.92	24.60
	video	17.49	19.64	**3.00**	19.85	14.64	14.67	18.43
Multiple	**cut ∪ video**	19.31	25.06	**16.47**	22.57	23.80	28.19	25.44
	cut ∪ warped	**16.37**	21.19	16.95	21.47	23.62	28.40	26.61
	warped ∪ video	19.43	23.45	**12.73**	22.52	19.42	24.31	24.08
	cut ∪ warped ∪ video	16.25	24.83	**15.11**	22.15	23.01	26.48	25.74

Table 5.13: D-EER (%) values for single and multiple attack-resolution settings.

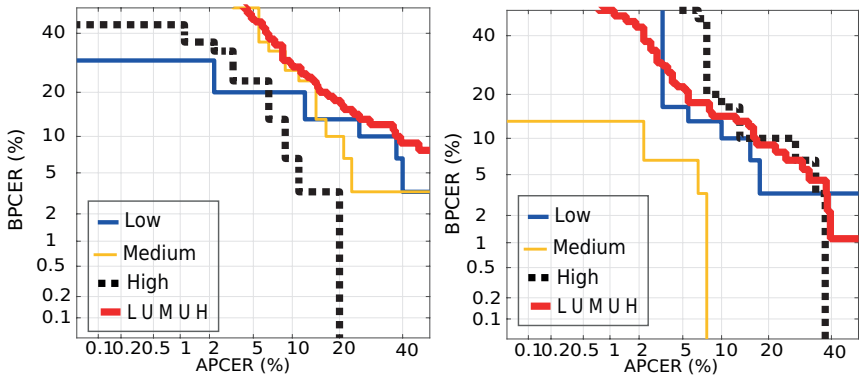
Attack \ Quality	Single	Multiple
Single	16.88 ± 6.17	21.48 ± 4.06
Multiple	18.93 ± 3.99	24.24 ± 2.45

5.5.6.2 *Known PAI species: Analysis of Images of Mixed Resolutions*

We present in Tab. 5.12 a joint evaluation of the proposed descriptors over several PAI species combinations, which were simultaneously acquired with different resolution by their respective capture devices. We can observe that the D-EER values for multiple resolution images are up to seven times higher than the ones achieved for the best single image quality (e.g., 3.00% for high image quality vs. 19.85% for L ∪ M). In contrast, similar D-EER values are reported when several PAI species are employed for training over face images acquired by a single capture device (e.g., 17.94% for Cut photo attacks vs. 16.37% for cut ∪ warped). Thus we can highlight how the utilisation of images of varying resolutions leads to a high PAD performance deterioration across different PAI species.

Finally, it should be also noted that the PAD approaches can be circumvented by launching Cut photo and Warped photo attack samples which were recorded with medium and high-resolution capture devices: a high mean D-EER of 28.40% is attained for that configuration.

Following those observations, we tried to determine which of these two image properties (i.e., PAI species or image resolution) produces the strongest PAD performance deterioration. To that end, we compute in Tab. 5.13 the average for each single and multiple combination depicted in Tab. 5.12. On the one hand, for a configuration where either several PAI species or images of varying resolutions are em-



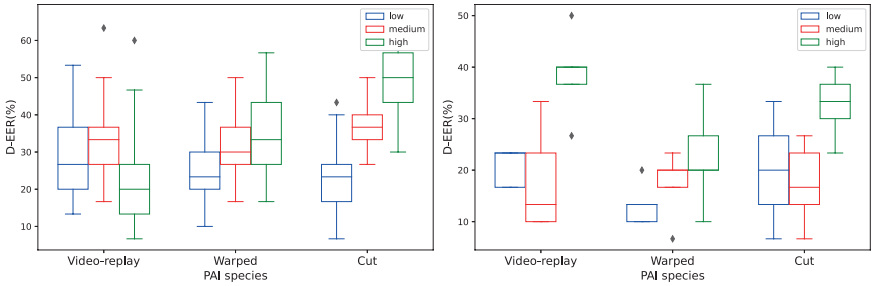
(a) BSIF descriptor. (b) DenseNet descriptor.

Figure 5.14: DET curves for the best handcrafted and deep learning-based approach over the known PAI species scenario. For the BSIF computation, we use $N = 10$ filters of size $l = 13$.

ployed, a high-performance decrease can be seen. In particular, a mean D-EER of 21.48% is reported when PAD approaches are trained with a single PAI species which was acquired under several resolution settings. This number is in turn worse than the one achieved when the PAD methods are trained with multiple PAI species with a single image resolution (18.93%). On the other hand, as could be expected, a high mean D-EER of 24.24% is attained for the worst case where the PAD approaches are optimised utilising several PAI species which were acquired with varying-resolution capture devices. We thus conclude that the utilisation of images with varying quality produces the greatest PAD performance deterioration. In addition, we confirm the PAD performance decrease reported by state-of-the-art algorithms in Tab. 5.2 and answer the question launched in Sect. 5.3: PAD techniques based both on handcrafted and deep learning features are sensitive to images with varying resolutions, which lead to a high accuracy decrease in the detection of APs.

5.5.6.3 *Known PAI species: A Deeper PAD Performance Analysis*

Finally, we show in Fig. 5.14 the DET curves for the best handcrafted and deep learning approaches over the Quality test and Overall test protocols from [236]. As it can be observed, the joint training for the analysed PAI species, which were acquired with three varying-resolution capture devices (i.e., thick red line), yields on average a high BPCER of 67.23% for any APCER $\leq$ 1%. This is in turn higher than the ones attained for every single resolution. In addition, the BSIF descriptor reports its best BPCER value for high-resolution images. In contrast, the deep learning approach achieves its best BPCER for images of medium quality. In this context, we can conclude that either a down-sampling or up-sampling step performed by the deep learning-based descriptors for fitting the size of a given image into the input



(a) Handcrafted descriptors. (b) Deep learning descriptors.

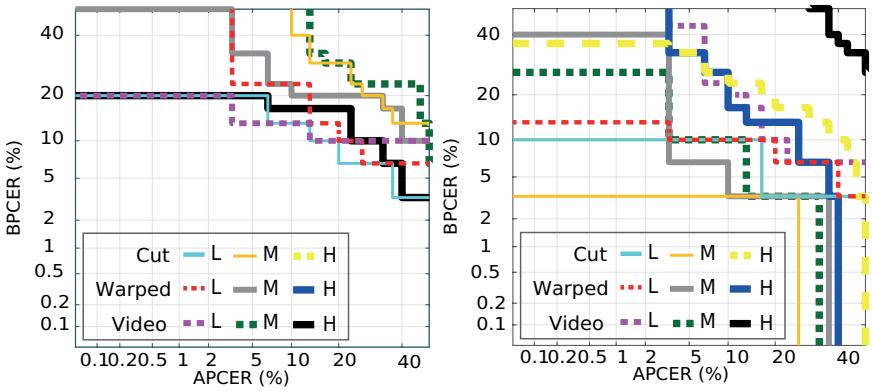
Figure 5.15: Handcrafted vs. Deep Learning performance on the detection of unknown PAI species.

layer can lead to an information loss for low and high-quality input images. In particular, for a database, such as CASIA whose cropped face images pose an average size of 180 $\times$ 157 and 644 $\times$ 545 pixels for low and high-resolution settings, respectively, an up-sampling and down-sampling procedure to fit the image size to 224 $\times$ 224 pixels (input layer size) can approximately affect on average a 65% of pixels of a given face image. This could in turn remove several artefacts produced in the creation of PAIs. In contrast, this re-sizing procedure affects only 37% of the pixels of medium resolution images, thereby leading to higher detection performance.

5.5.6.4 *Unknown PAI species: In-depth Generalisation Capability Analysis*

Once the image resolution issues have been evaluated, we selected the worst-case scenario from the previous experiment (i.e., several PAI species acquired under numerous image quality conditions are employed for training) and assessed the generalisation capability of the PAD approaches in Fig. 5.15. To that end, we follow the LOO protocol in [8].

We can observe in Fig. 5.15 a high D-EER variance for each unknown PAI species evaluated, which confirms the high impact degree of the image resolution on the PAD generalisation capabilities. On the other hand, a mean D-EER of 24.57% $\pm$ 8.64 indicates that the handcrafted PAD approaches perform better for low-resolution samples stemming from unknown PAI species. Depending on the given PAI species, the detection performance of deep learning approaches attained for low-resolution unknown PAI species outperforms the one reported for medium-resolution samples (e.g., 13.33% $\pm$ 4.08 vs. 17.33 $\pm$ 6.41 for warped photo attacks). In addition, the error rates yielded for high-resolution images lead to a considerable detection performance deterioration, thereby resulting in a mean D-EER of 34.14%. We can therefore conclude that the resolution variation is not the only external factor which affects the detection performance of PAD methods. Other acquisition conditions such as the distance between the PAI



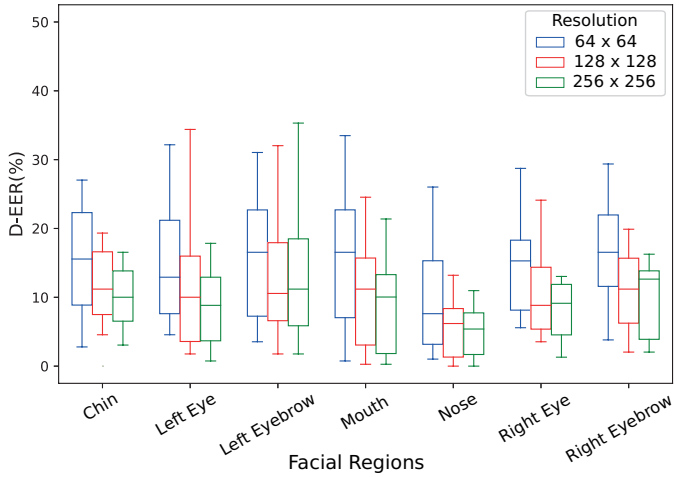
(a) BSIF descriptor.

(b) DenseNet descriptor.

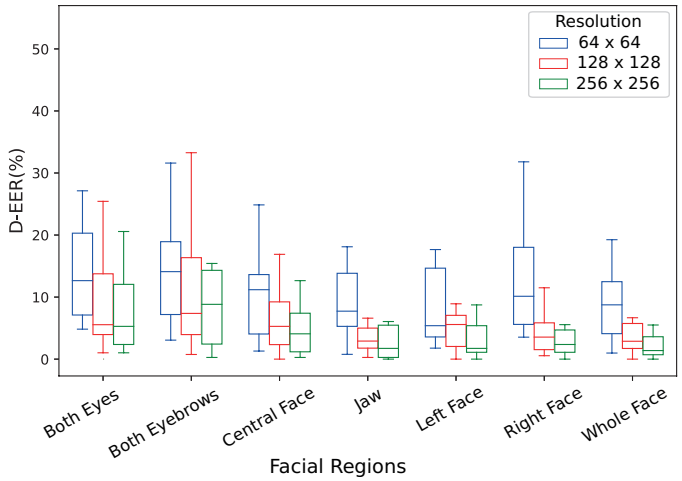
Figure 5.16: DET curves for the best handcrafted and deep learning-based approaches over the unknown PAI species scenario. For BSIF computation, we use $N = 10$ filters of size $l = 13$.

and capture device, which differs for different samples on the CASIA database, can also produce an accuracy decrease of PAD algorithms.

To conclude our analysis, we show in Fig. 5.16 the DET curves for the best handcrafted and deep learning approaches over the unknown PAI species scenario. First, we can note that there is a correlation between the error rates represented in Fig. 5.15 and the detection performance attained by a particular handcrafted and deep learning descriptor: the BSIF descriptor shows its best detection performance for a low-resolution setting, thereby resulting in a BPCER_{20} of 16.67% for the entire set of PAIs. Similarly, the DenseNet descriptor for medium-resolution configuration attains on average a BPCER_{20} of 6.67% which outperforms the BPCER values achieved for the remaining resolution settings (i.e., a BPCER_{20} of 13.44% and 35.55% for low and high resolution samples, respectively). These results confirm that the deep learning approaches report an accuracy decrease when the images size at hand is not close to the size of their input layer. In addition, they reveal that the PAD methods highly depend on the resolution of the capture device and hence should be carefully optimised for each particular application.



(a) Impact on the single regions.



(b) Impact on the composite regions.

Figure 5.17: Impact of image transformation on different facial regions.

5.5.7 *Facial Regions Analysis*

5.5.7.1 *Known PAI species*

EFFECTS OF IMAGE TRANSFORMATION FOR PAD Since the size of facial regions can vary across images, we also investigate the effect of image resolution over facial regions for PAD. To that end, we compute the D-EER per facial region and algorithm defined in Sect. 5.2.3 over three databases: CASIA, RM, and RA. Fig. 5.17 reports the boxplots per facial region over three resolutions i.e., 64×64, 128×128, 256×256: greater resolution configurations might result in a performance deterioration due to pixel value interpolation for the smallest regions. We note that the D-EER improves with the image resolution, thus yielding the best detection performance for an image size of 256×256 pixels. We also observe that those regions having a large image size (e.g., full face, right face, left face, and jaw) report a low STD for

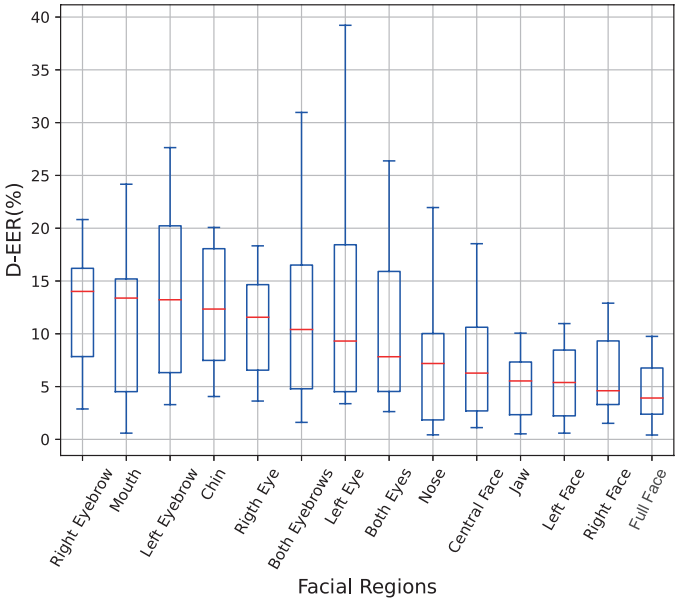


Figure 5.18: Best performing facial regions for known PAI species.

an image resize greater or equal than 128×128 pixels (see red and green boxes in Fig. 5.17-b): the mean STD is approximately 6.99. In contrast, their STD increases when a small size of 64×64 is used (see blue boxes in Fig. 5.17-b): the mean STD is approximately 10.25.

Following the above observations, we also see that the pixel value estimation for the smallest facial regions (i.e., left and right eyes, left and right eyebrows, both eyebrows, both eyes, mouth, nose, and chin) during the resize significantly affects the algorithm's detection performance, thus resulting in a high STD in the ranges $[6.72, \ldots, 13.62]$. These resolution results confirm the findings observed in Sect. 5.5.6: the up-sampling or down-sampling step performed by the deep learning approaches to adjust the size of a given image in the input layer leads to an information loss of artefacts for the smallest or largest sizes, respectively.

Based on the fact that most facial regions report on average their best detection performance for a resize configuration of 256×256 pixels, we select it for further experiments.

DETECTION PERFORMANCE OF FACIAL REGIONS In the second set of experiments, we evaluate the PAD performance for each facial region over CASIA, RM, and RA databases following their corresponding known PAI species protocols. Similar to the above experiment, we compute the D-EER per facial region and algorithm defined in Sect. 5.2.3 and report their detection performance as boxplots in Fig. 5.18. As it may be noted, the training and evaluation of selected approaches using the full face attain on average the best D-EER: a median D-EER of 3.92% (indicated by the central blue mark in the boxplots) outperforms the remaining facial regions. Regarding com-

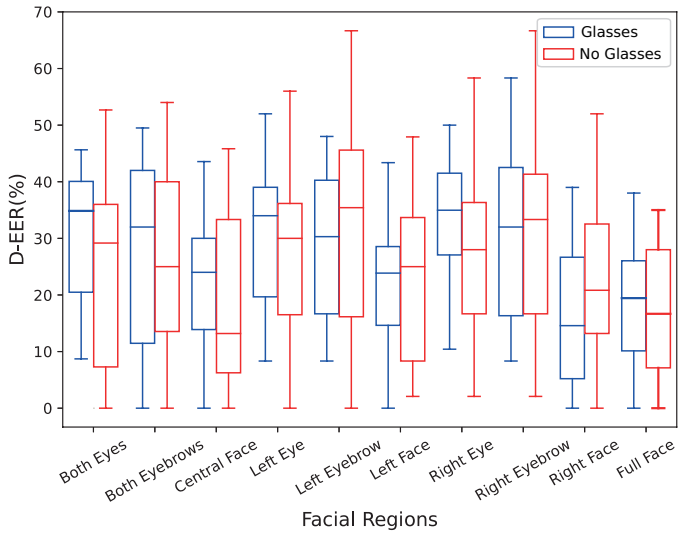


Figure 5.19: Detection performance for images containing glasses (blue boxes) and no glasses (red boxes).

posite regions, we also observe that they report the best performances e.g., right face (median D-EER = 4.61%), left face (median D-EER = 5.38%), jaw (median D-EER = 5.53%), and central face (median D-EER = 6.28%). Furthermore, the error rates of these composite regions tend to their median values, thus resulting in a low STD with respect to the mean values: their STD is in the ranges $[5.88, \ldots, 7.97]$. Among the single regions, the nose achieves the best detection performance, yielding a median D-EER of $7.19\% \pm 7.06$. In fact, this outperforms the performance attained by both eyes (median D-EER = 7.83%). Whereas the 75% of D-EER values for the nose region are below its median, only 25% of error rates for both eyes are below its median, hence indicating that the nose is more suitable for PAD than both eyes. Since the nose is a flat region composed mostly of skin, we think that any variation in quality, colour, or texture can lead to an improvement in the detection of APs.

We observe that the worst regions are right and left eyebrows and mouth which report median D-EERs above 13% and STD values in the ranges $[10.28, \ldots, 12.08]$. We can also see that the union between both regions (i.e., both eyebrows) improves their individual errors by three percentage points (i.e., 10.41% for both eyebrows vs. 14.01% for the right eyebrow). This is because the region comprising both eyebrows includes a flat skin in between which allows algorithms to detect APs. Similar behaviour can be also perceived in the results achieved for both eyes.

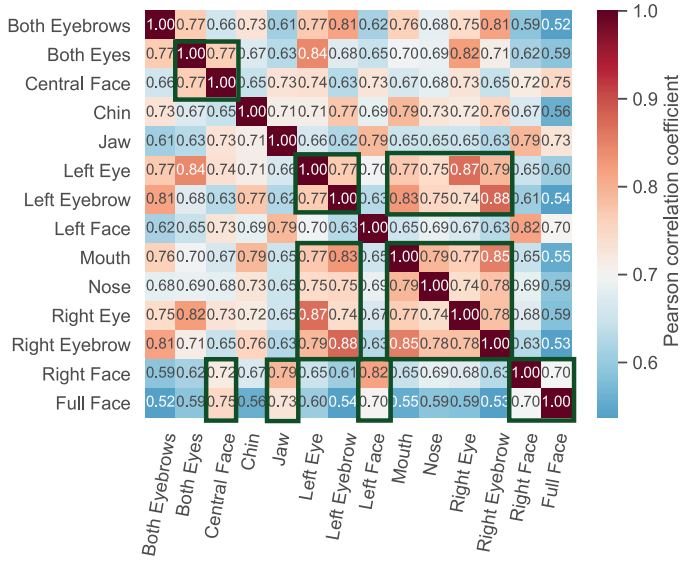
IMPACT OF WEARING GLASSES ON PAD Taking a closer look at Fig. 5.18, we note that most regions around the eyes (i.e., left and right eyes and left and right eyebrows) report a high performance deterioration, thus yielding STD values in the ranges $[10, \ldots, 12]$. Based on

this observation, we investigate the effect of wearing glasses in those regions that might contain such accessories. To that end, we follow the same experimental evaluation used in Sect. 5.5.7.1 and split the training and evaluation sets from the CASIA, RM and RA databases into two balanced sets each containing faces with glasses and faces without glasses. We then show in Fig. 5.19 the boxplots representing D-EERs achieved by the proposed methods per the facial region over the above databases. We perceive that *i*) wearing glasses affects the detection performance of approaches evaluated when trained using either the full face or the central face, *ii*) right and left faces are not affected by wearing glasses, thus yielding a better detection performance when faces contain glasses, *iii*) wearing glasses impact the PAD performance for both left and right eyes along their fusion (i.e., both eyes), and *iv*) whereas the performance for left and right eyebrows is not highly affected by wearing glasses, the fusion region (i.e., both eyebrows) is. The latter is due to the accuracy of the region extraction algorithm: it includes part of the glasses in the final images. These findings complement the study conducted in [149]: wearing glasses also has a negative impact on iris segmentation and thus on iris recognition.

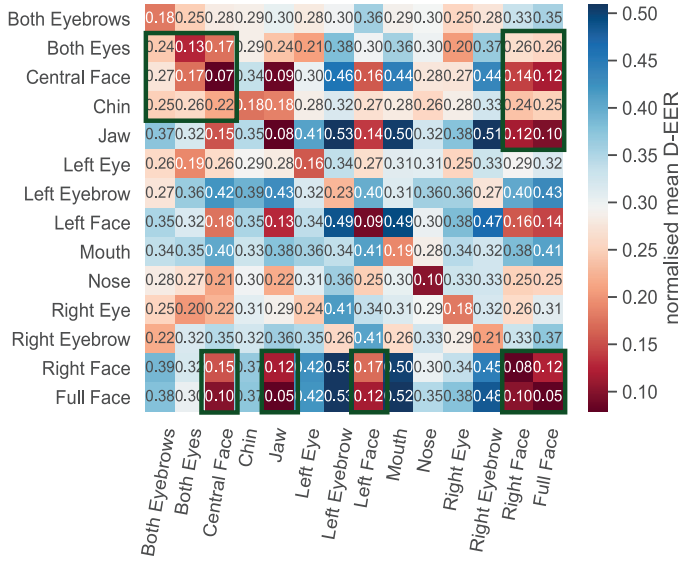
CORRELATION, CROSS-DETECTION PERFORMANCE, AND UTILITY

We explore now the correlation between facial regions. For this purpose, we first train the PAD approaches for each facial region over the CASIA, RM, and RA databases. On the evaluation sets, we extract the latent vectors from the last FCL before the final decision layer and average them. Finally, we illustrate in Fig. 5.20-a the average Pearson correlation coefficient between facial regions. It should be noted that the features representing the facial region combination share at least 50% of their characteristics with each other. We highlight with a green rectangle the facial regions that are highly correlated with each other. Specifically, the latent vectors of facial regions such as the left and right eyes, left and right eyebrows, mouth and nose report as expected a high Pearson correlation ranging from 0.74 to 1.00. As expected, we also see that the right and left regions of the faces share 82% of their characteristics with each other. Therefore, they can be interchangeably used for PAD. Finally, the full face is highly correlated with the central part of the face, followed by the jaw and the left and right regions of the face.

Following the above idea, we also compute the detection performance between facial regions (as it is illustrated in Fig. 5.20-b). In this experiment, we train the architectures using one facial region (depicted by the rows) and evaluate the remaining regions (shown by the columns) over three databases (i.e., CASIA, RM, and RA). Then, the mean D-EER between facial region combinations is reported in Fig. 5.20-b. It is important to point out that error rates are normalised following the eq. 5.2. It should be observed that the evaluation of



(a) Absolute correlation values between facial regions.



(b) Detection performance normalised between facial regions

Figure 5.20: Correlation and detection performance between facial regions. The green rectangles highlight some examples of facial region configurations which report high correlations and detection performances.

facial regions such as the jaw, central face, and left and right regions achieves the best detection when the algorithms are trained using the full face. In fact, the jaw yields the same D-EER to the one attained by the full face (i.e., D-EER = 0.05). Subsequently, the central face and left and right face regions depict similar detection performance (i.e., 0.10 vs. 0.12). As a consequence of these results, we perceive that the full face can be used to spot an AP attempt in the probe image when PAD algorithms are trained on either jaw, left face, right face, or central face regions.

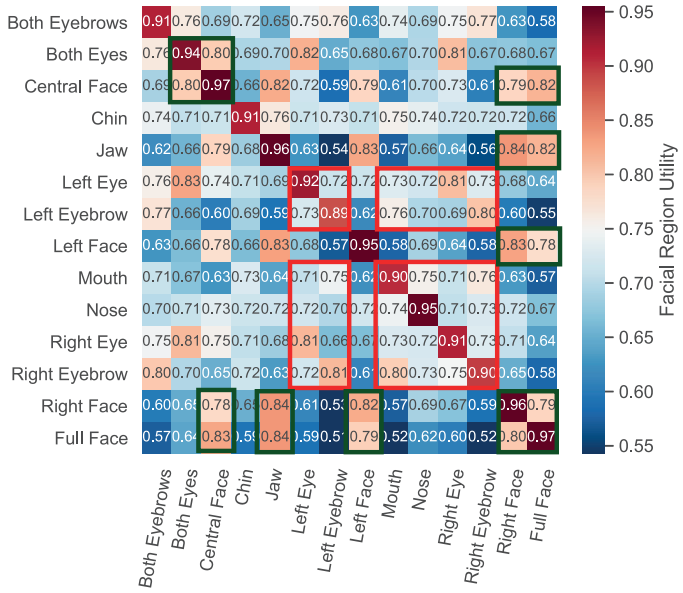


Figure 5.21: *Facial Region Utility* computed from the correlation and detection performance matrices. The green rectangles highlight the combinations of facial regions with a high utility. The red rectangles state those examples of facial region combinations whose correlation and detection performance values show a contrary trend in Fig. 5.20.

Based on Fig. 5.20, we compute in Fig. 5.21 using the eq. 5.1 the *Facial Region Utility*. As it was mentioned in Sect. 5.2.4, this metric indicates the usefulness of a particular region for training to spot an AP based on the other region in a probe image. As it can be observed, the same region used simultaneously for training and testing reports the best utility (i.e., diagonal values). We also note that facial regions such as the full face, left and right faces, central face, and jaw can be used the create a reliable train and test configuration as they report high *Facial Region Utility* values. In particular, the training of a PAD approach over the full face (i.e, red rectangle at the bottom) allows the successful evaluation of regions such as jaw ($U(\cdot) = 0.84$), central face ($U(\cdot) = 0.83$), right face ($U(\cdot) = 0.80$), and left face ($U(\cdot) = 0.79$). It should be noted that the *Facial Region Utility* highly depends both on the correlation and the algorithm's detection performance. We highlight with red rectangles those train-test facial regions which drop their *Facial Region Utility* due to contrary trends depicted in Fig. 5.20. Whereas mouth, nose, right and left eyes and left and right eyebrows pose a high correlation with each other (see Fig. 5.20-a), the detection performance between them decreases considerably (see Fig. 5.20-b). Therefore, they are not suitable for a PAD train-test configuration.

Table 5.14: Benchmark of the state-of-the-art algorithms trained on the full face and evaluated on the regions with the best *Facial Region Utility* in terms of D-EER (%) using the OULU-NPU database.

P	Approach	Facial Regions				
		Full Face	Jaw	Central Face	Right Face	Left Face
1	FV [72]	8.19 ± 1.19	7.59 ± 0.80	13.71 ± 2.39	8.54 ± 1.05	12.06 ± 1.07
	DeepPixelBis [65]	4.17	6.67	6.67	4.48	10.83
	CDCN [229]	4.48	10.83	20.94	15.83	16.68
2	FV [72]	8.30 ± 1.75	7.23 ± 1.37	11.71 ± 2.59	9.53 ± 2.20	10.74 ± 2.56
	DeepPixelBis [65]	2.78	3.83	7.25	8.01	10.52
	CDCN [229]	3.96	11.11	18.89	16.94	17.78
3	FV [72]	8.29 ± 6.75	8.27 ± 6.54	12.59 ± 6.42	10.22 ± 6.50	14.92 ± 8.41
	DeepPixelBis [65]	1.25 ± 1.23	4.10 ± 2.80	6.33 ± 5.49	5.53 ± 5.47	7.62 ± 5.08
	CDCN [229]	1.88 ± 0.93	13.40 ± 3.38	22.78 ± 3.75	13.47 ± 5.30	14.27 ± 5.41
4	FV [72]	24.86 ± 8.47	19.88 ± 10.17	27.66 ± 7.62	20.56 ± 13.12	28.53 ± 6.19
	DeepPixelBis [65]	10.42 ± 10.51	12.71 ± 5.83	16.67 ± 5.74	16.04 ± 13.66	19.58 ± 11.64
	CDCN [229]	13.54 ± 4.21	22.29 ± 7.56	28.33 ± 6.83	22.71 ± 14.88	22.71 ± 12.31

5.5.7.2 *Analysis of the Facial Region Utility on Challenging Scenarios*

In order to verify the usefulness of the *Facial Region Utility*, we select several state-of-the-art PAD techniques and train them using the full face. Then, we evaluate the best utility regions (i.e., jaw, central face, right face, and left face) in Tab. 5.14 over the challenging protocols in the OULU-NPU database. It is important to highlight that the results depicted in Tab. 5.14 might differ from the ones yielded by their corresponding papers. We train and assess these algorithms using a random video frame in contrast to the original pipelines which use all video frames to make the final decision. We can see that the D-EERs improve with the utility of facial regions independently of the evaluated protocol. Specifically, the best detection performance is yielded by the full face, followed by the jaw. According to the *Facial Region Utility*, the central face is the third best region to spot an AP attempt in a probe image after the full face and jaw. However, we may note that this region reports a detection performance decrease with respect to the results achieved by right and left faces. This behaviour mostly happens due to the sensitivity of this region to the use of glasses (see Fig. 5.19). In addition, we observe that the right face outperforms the left face in all experiments. This is mainly due to variables such as the asymmetry of the face and the artificial light positions used in the BP and AP acquisition. The latter causes most of the characteristics separating a BP from an AP to be detected in the right region of the face (see Fig. 5.22).

On the other hand, we observe that the detection performance attained by our semantic common feature space (i.e., FV) shows for the jaw an improvement regarding the remaining regions. Unlike deep learning approaches evaluated, this algorithm derives a kernel from

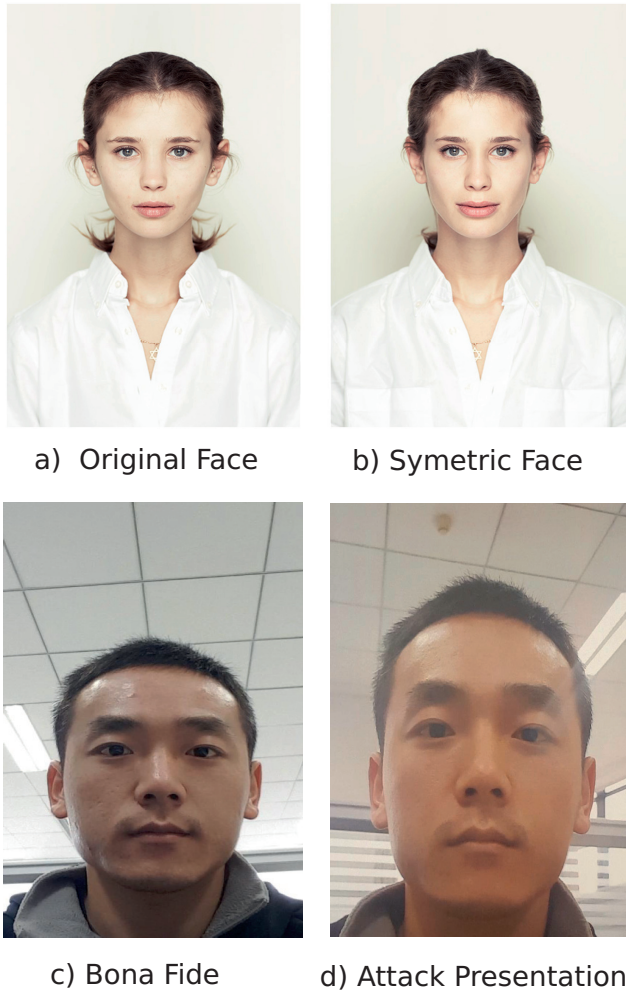


Figure 5.22: Some images that show why the detection performance between left and right faces is different. *a*) and *b*) represent the visual differences between a perfect symmetrical face (i.e., *b*) and its original face (i.e, *a*) [123]. *c* and *d* are examples of BP and AP in OULU-NPU whose artificial light configurations differ with each other.

the parameters learned by a generative model (i.e., GMM [176]) to characterise how the distribution of a set of unknown local descriptors differs from the distribution of known features. Therefore, this does not require the probe image to be similar in terms of shape to the trained samples. It can be observed that deep learning schemes suffer from a detection performance deterioration when the object in the probe image (e.g., the jaw) is different from the one used for training (i.e., the full face). We think that deep learning solutions focused on local patches could improve the above limitation.

Table 5.15: The detection performance of the DeepPixelBis algorithm on the CRMA database. The PAD decision threshold employed in the APCER, BPCER, and ACER computation is the one yielded at a $BPCER_{10}$ on only unmasked data in the development set.

Approach	BPCER (%)		APCER (%) - Print			APCER (%) - Replay			ACER (%)
	BM0	BM1	AM0	AM1	AM1	AM0	AM1	AM1	
DeepPixelBis (Full Face) [65]	63.16	64.04	0.00	0.00	0.00	0.00	0.64	0.00	29.47
$DeepPixelBis_{RW}$ (Full Face) [54]	35.09	41.23	0.00	0.10	1.17	0.19	1.95	0.58	18.58
$DeepPixelBis_{PAL}$ (Full Face) [54]	42.11	51.75	0.00	0.00	0.00	0.00	0.44	0.00	23.35
$DeepPixelBis_{RW-PAL}$ (Full Face) [54]	26.32	29.82	0.00	0.19	1.17	0.00	1.32	0.29	14.81
DeepPixelBis (Central Face)	7.02	15.79	3.12	2.83	7.02	8.48	6.84	9.06	**9.51**

5.5.7.3 *Benchmark with the State Of The Art*

In order to validate the usefulness of facial regions for a real application, we select the best performing algorithm in Tab. 5.14 (i.e., DeepPixelBis) and establish a benchmark with the state-of-the-art techniques in Tab. 5.15 over the CRMA database. To conduct a realistic analysis where the behaviour of the PAD on masked data is still unknown, we follow the experimental setup in [54] and report the APCER and BPCER values by using the threshold $BPCER_{10}$ that is computed on only unmasked data in the development. In this experiment, the algorithms are trained on the full faces and evaluated either on the full face (i.e., the four first rows) or the central face (i.e., the last row). In addition, we compute the ACER due to the lack of a proper evaluation of the state-of-the-art compliant with the ISO/IEC 30107-3 [97] for biometric PAD. Taking a closer look at Tab. 5.15, we note that the evaluation of the algorithms using the full face leads to a significant detection performance deterioration. In particular, the BPCER values for BM0 and BM1 are considerably high (i.e., first row, BPCER $\geq$ 63.16%), thereby confirming our initial hypothesis: PAD algorithms misclassify BPs as an intentional AP when subjects wear some accessories e.g., masks. In fact, the Regional Weight (RW) and Partial Attack Label (PAL) methodologies proposed in [54] to mitigate masked attacks build a secure (APCER $\leq$ 1.95%) but not convenient (BPCER $\geq$ 26.32) PAD subsystem. In contrast, we see that the detection of a facial mask and then the evaluation of the central region result in an overall improvement of the detection performance: an ACER of 9.51% which outperforms the DeepPixelBis [65] and $DeepPixelBis_{RW-PAL}$ [54] method by relative improvements of 67.73% and 17.98%, respectively, allows the building of a secure and convenient system. Finally, it is worth noting that the subjects only wear glasses in 16% of the images in CRMA database unlike OULU-NPU, whose subjects wear masks in 50% of the images. Therefore, these results are not fully biased by this type of accessory.

5.6 SUMMARY

To summarise the findings on the face PAD, we can highlight the following takeaway messages:

- The FV common feature space is able to keep a high generalisation capability for facial images. By combining the FV with a compact BSIF we can obtain a $BPCER_{100}$ in a range of 0.0% to 17% for traditional unknown PAI species such as printed, photo- and video-replay, and cut photo attacks. These results outperformed the top state-of-the-art and confirmed that our PAD approach can yield a secure and convenient system under that challenging scenario.
- A remarkable performance is also reported for more challenging attacks: a mean D-EER of 11.44% showed the FV soundness in the detection of unknown PAI species. In particular, the algorithm was able to yield an APCER of 26.09% for a type of obfuscation attacks, which is up to four times better than the ones reported by current state-of-the-art PAD techniques.
- The experimental results over the common feature space indicated the statistical advantage of RGB with respect to other colour spaces for datasets having images of varying resolutions, thereby resulting in a minimum average D-EER of 0.45% for known PAI species detection.
- Regarding the facial region analysis, we showed that the composite regions achieved the best detection performances among regions. In particular, the full face yielded a median D-EER of 3.92%, followed by the right face (median D-EER = 4.61%), left face (median D-EER = 5.38%), jaw (median D-EER = 5.53%), and central face (median D-EER = 6.28%).
- Further, the experimental results unveiled the existence of a correlation between left and right regions of the face as well as both eyes and eyebrows in terms of PAD performance.
- The proposed *Facial Region Utility* metric indicated those regions capable of being used to improve the performance reported by the full face when the subjects use common accessories.
- A particular use case where individuals wore masks to prevent respiratory infections showed the feasibility of using the central face over the full face in the evaluation: an ACER of 9.51% which outperforms the state-of-the-art methods by a relative improvement up to 67.73%, allows the building of a secure and convenient PAD module. In addition, we noted that the BPCER values yielded by the state-of-the-art were decreased down to 7.02% (BM0) and 15.79% (BM1) for pristine subjects.

- Wearing glasses affects the detection performance of algorithms when either the full face, eyes, or central face is used to detect AP attempts
- Increasing the size of facial regions impacts the detection performance of the analysed algorithms: 256 × 256 pixels reported the best results for all regions. Furthermore, we see that the pixel value estimation for the smallest facial regions such as left and right eyes, left and right eyebrows, both eyebrows, both eyes, mouth, nose, and chin during the resize considerably affects the algorithm's detection performance, thus resulting in high STD values.
- The above observations were also confirmed on the evaluation of the impact of images of varying quality for the facial PAD. Images of varying resolutions produce a high PAD performance decrease, which can be even greater than the use of numerous PAI species.
- Deep learning-based descriptors reported the worst PAD performance deterioration for face images whose sizes are widely different from the input layer size.
- Video replay attacks screened on a high-resolution capture device unveiled several blurriness and sharpness properties, which can be successfully detected by PAD techniques.
- The training of PAD methods with several PAI species which are acquired with varying-resolution capture devices appears to be the worst case for the face PAD task, thereby resulting in a D-EER of 24.24% and a joint $BPCER_{100}$ of 67.23%. This, in turn, confirmed that the image resolution is a requirement which must carefully be taken into account to build a secure and reliable face PAD module.

6 VOICE PRESENTATION ATTACK DETECTION

In this Chapter, we evaluate the feasibility of using FV for voice PAD. To the best of our knowledge, very few works have explored texture-based analysis (Sect. 6.1) for voice PAD. Alegre *et al.* [4] proposed an algorithm based on the combination of LBP and one-class classifiers. Even if the proposed technique reported a poor detection performance for some unknown PAI species such as Voice Conversion (VC), that study showed the generalisation capability of the proposed texture-based representation for voice PAD. Motivated by that fact, we explore in this Chapter several image processing texture descriptors in combination with a SVM (Sect. 6.2), which have been successfully employed for fingerprint [135] and face [62] PAD. In order to improve the generalisation capability of the analysed texture descriptors, we utilise the FV representation (Sect. 6.3). In addition, we establish a benchmark of our generalisable common feature space (i.e., FV) with a new deep learning-based approach namely Dual-Stream Temporal CNN (see Sect. 6.4). This Chapter summarises the results in [76, 79] and answers the **RQ 5**.

6.1 1D AUDIO WAVEFORMS TO 2D SPECTROGRAMS

The visualisation of audio/speech signals is key to many audio analysis tasks, usually involving: *i*) time-domain, *ii*) frequency-domain, or *iii*) time-frequency-domain representations known as spectrograms, which show the signal amplitude over time at a set of discrete frequencies. Many time-frequency representations have been proposed, each with different characteristics. Keeping in mind that an audio signal can be represented as an image, as shown in Fig. 6.1, in this Thesis we focus on the following four time-frequency representations:

- The Short-Time Fourier Transform (STFT) [147] is a time-frequency decomposition based upon the application of Fourier analysis to short segments or windows of the audio signal. As such, it is effectively a filter bank where the bandwidth of each filter is constant and is related to the window function. The STFT is implemented on a 30ms window with a 15ms shift and a 1024-point Fourier transform.
- The Linear Frequency Cepstral Coefficients (LFCC) [193] coefficients are computed from the STFT by applying the discrete cosine transform (DCT) [147]. Generally, only lower-order coeffi-

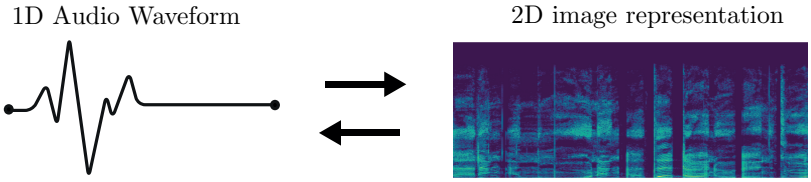


Figure 6.1: A speech sample together with its texture image representation.

cients are retained since they represent the vocal tract configuration.

- The Constant-Q Transform (CQT) [22] is a perceptually motivated approach to time-frequency analysis. In contrast to Fourier-based approaches, the bin frequencies of the filterbank are geometrically distributed. Compared to the STFT, the CQT has a greater frequency resolution for lower frequencies and a greater temporal resolution for higher frequencies. The CQT is applied with a maximum frequency of $F_{max} = F_{NYQ}$, where F_{NYQ} is the Nyquist frequency of 8kHz. The minimum frequency is set to $F_{min} = F_{max}/2^9 \simeq 15$Hz (9 being the number of octaves). The number of bins per octave is set to 96. These parameters result in a time shift of 8.5ms. For both STFT and CQT spectrogram-to-image representations, we perform a min-max normalisation and 8-bit quantisation on the log-magnitude spectrum.
- The Constant Q Cepstral Coefficients (CQCC) [192, 202] stem from the application of cepstral processing to CQT representations. CQCCs offer a time-frequency resolution more closely related to that of human perception. These features were designed specifically for PAD but have also shown to be beneficial for Automatic Speaker Verification (ASV) and utterance verification [201].

6.2 TEXTURE DESCRIPTORS

In order to analyse the texture features extracted from the time-frequency image representations described above, we study some well-known texture descriptors, exposed in Sect. 3.1, in combination with SVMs: LBP [145], MB-LBP [231], BSIF [104], and LPQ [146]. It worth noting that these descriptors are computed over the full image. Therefore, an image is only represented by one feature vector.

6.3 APPLICATION OF FV

In a second approach, we evaluate the feasibility of using our common feature space (i.e., FV) in combination with the best performing texture descriptor and image representation described in Sect. 6.2 and 6.1,

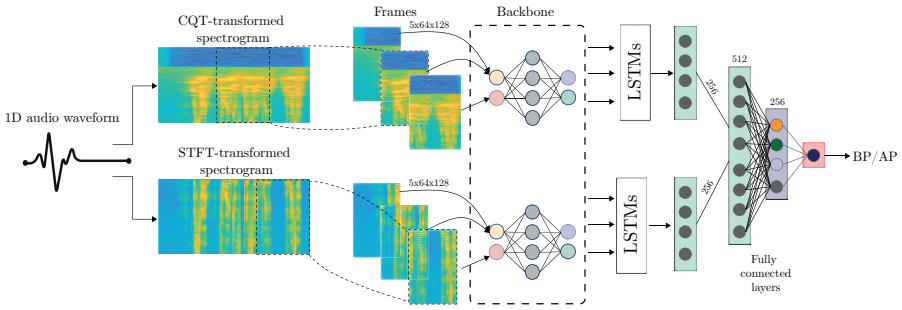


Figure 6.2: General overview of our dual-stream temporal CNN approach.

respectively. In essence, we follow the same pipeline described in Chapter 3: *i*) the input 1D audio waveform is transformed to the corresponding image representation, *ii*) the best performing descriptor in Sect. 6.2 is computed over the whole image and *iii*) projected by the FV. Finally, a linear SVM takes the final decision.

6.4 DUAL-STREAM TEMPORAL CNN

Finally, we propose a dual-stream temporal CNN which leverages temporal information of latent embeddings extracted from the two best image representations in Sect. 6.1 to enhance the generalisation capabilities. Fig. 6.2 shows a general overview of our dual-stream temporal CNN which takes advantage of the temporal latent representation of the input spectrograms for voice PAD. In essence, the input 1D audio waveforms are firstly transformed into 2D images using the image representations described in Sect. 6.1. This transformation will lead to an individual stream in our approach. The new images are then split into several frames and represented by an intermediate latent vector stemming from a traditional CNN (e.g., DenseNet, declared as Backbone in Fig. 6.2). In order to include voice temporal information in the network optimisation, the per-frame latent representation is further processed by a series of LSTM layers, whose output layer is concatenated with the one provided by the other stream. Finally, these final concatenated features are fed to a FCL, which, in turn, inputs a single unit layer for the BP vs. AP decision.

To optimise the network, we use the Binary Cross Entropy (BCE) loss, which is generally employed for binary classification tasks [82]. BCE $\mathcal{L}(\cdot)$ is computed as:

$$\mathcal{L}(x) = y \cdot \log p(x) + (1 - y) \cdot \log(1 - p(x)), \tag{6.1}$$

where $p(x)$ is the predicted probability and y is the true label for the input x. We assign $y = 1$ for BPs and $y = 0$ for APs.

Table 6.1: General architecture of our dual-stream temporal CNN.

Layers	CQT-stream	STFT-stream
Input	$5 \times 64 \times 128$	$5 \times 64 \times 128$
Backbone latent space	5×512	5×512
LSTM (4 layers)	1×256	1×256
Concatenation	1×512	
FCL	1×256	
Sigmoid	1×1	

6.4.1 *Network Architecture*

As mentioned, our dual-stream temporal CNN comprises two streams: one optimised for the CQT representation and the other one for the STFT representation. In our experiments, we first split the input spectrograms into 5 continuous frames, each of which has 64 $\times$128 pixels. A latent representation of 512 features per frame is computed using a given backbone (e.g., DenseNet [91], ResNet [86], or MobileNetv2 [177]). To exploit temporal information of speech images, the above latent representations are fed into 4 hidden LSTM layers each consisting of 256 neurons. The LSTM outputs of each stream are concatenated into a 512 vector, which in turn is further processed by a 256 FCL. Finally, a FCL of a single unit with sigmoid activation is added to produce the binary classification. A summary of the main architecture is shown in Tab. 6.1.

In our implementation, we trained the network from scratch using the Adam optimiser [107]. A learning rate of 1×10^{-4} with a weight decay parameter of 1×10^{-6} was used. The framework was implemented in PyTorch [155] and trained on the Nvidia GPU Tesla M10 with 16 GB DRAM.

6.5 EXPERIMENTAL SETUP

The experimental evaluation has a threefold goal: *i*) evaluate the detection performance of our proposed method over challenging scenarios, *ii*) analyse the effect of unbalanced data over the generalisation capabilities, and *iii*) establish a benchmark with the state-of-the-art PAD techniques.

- Analyse the detection performance of the texture descriptors for the baseline scenario (i.e., known PAI species and unknown PAI species).

- Benchmark the detection performance of our two proposed algorithms for known PAI species and unknown PAI species.
- Study the effect of unbalanced data over the generalisation capabilities.
- Evaluate the detection performance for cross-database.

In order to establish a fair benchmark, we adopt two PAD baseline approaches from the ASVspoof 2019 challenge [214], which use GMM back-end classifier with either CQCC (*B*01) or LFCC (*B*02) features. It should be noted that whereas the baseline approaches employed a bi-cluster GMM model for the BP or AP classification, our analysed FV technique uses it as a generative model to fit the BP and AP data distribution.

6.5.1 *Databases*

The experimental evaluation was conducted over the freely available databases ASVspoof 2019 [203] and 2021 [224] whose characteristics are summarised in Tab. 6.2:

- ASVspoof 2019 database consists of two assessment scenarios: Logical Access (LgA) and Physical Access (PhA)[1]. Both LgA and PhA databases are partitioned into three disjoint datasets: training, development, and evaluation. Whereas the PAIs in the training and development datasets were built with the same algorithms and capture conditions (i.e., it is the known PAI species scenario), PAIs for the evaluation dataset were generated with different techniques and capture conditions (i.e., it is theunknown PAI species scenario). The LgA partition contains PAI samples which were generated using 17 different TTS and VC technologies: six were designated for known PAI species assessment (i.e., A01-A06) and 11 for unknown PAI species (i.e., A07-019 with exception of A16 = A04 and A19 = A06 and hence both two attacks are in the training set). In order to analyse and improve the ASV reliability in different acoustic environments and replay setups, the training and development data for the PhA scenario is created under 27 different acoustic and 9 replay configurations. The replay settings comprise 3 attacker-to-talker (i.e., A, B, C) recording distances and 3 loudspeaker quality (i.e., A, B, C). The evaluation dataset is generated in the same manner as training and development data but with different random acoustic and replay configurations. [203].

1 To avoid confusion with PA (presentation attack), we have named the two partitions of the ASVspoof 2019 database LgA and PhA

Table 6.2: A summary of ASVspoof databases.

	Partition	Dataset	#BP	#AP	PAI species
2019	LgA	training	2580	22800	A01, A03, A03, A04, A05, A06
		development	2548	22296	
		evaluation	7355	63882	A07, A8, A09, A10, A11, A12, A13, A14, A15, A16, A17, A18, A19
	PhA	training	5400	48600	AA, AB, AC, BA, BB, BC, CA, CB, CC
		development	5400	24300	
		evaluation	18090	116640	
2021	LgA	evaluation	18452	163114	C1, C2, C3, C4, C5, C6, C7

- ASVspoof 2021 database includes an extra assessment scenario (i.e., DeepFake) along with the LgA and PhA scenarios. Following the protocol in [224], we use the LgA partition for the cross-database evaluation. More specifically, the proposed algorithms are trained over the LgA partition in the ASVspoof 2019 and evaluated over the same partition in the ASVspoof 2021. ASVspoof 2021 is more challenging than ASVspoof 2019 as it includes new trials per speaker. In contrast to the LgA partition in ASVspoof 2019, this set in ASVspoof 2021 contains BPs and PAIs transmitted over a variety of telephony systems including Voice-over-IP (VoIP) and a Public Switched Telephone Network (PSTN). The data transmission across telephony systems introduces nuisance variability usually expected in several real applications. Both BP and AP samples were treated with one of seven distinct codecs as a result of transmission (i.e., C1 - C7). C1 replicates the LgA scenario in the ASVspoof 2019. C2 and C4-C7 correspond to the transmission across an Asterisk Private Branch Exchange (PBX) system using one of five different codecs operating at either 8 kHz or 16 kHz bandwidths. C3 relates to the transmission over a PSTN system starting from a mobile smartphone and ending at a SIP endpoint hosted on a professional VoIP system.

Table 6.3: Benchmark in terms of D-EER(%) of the texture descriptors for the best parameter configuration per speech-to-image domain transformation for known PAI species.

	CQCC		**LFCC**		**STFT**		**CQT**	
	LgA	PhA	LgA	PhA	LgA	PhA	LgA	PhA
Best BSIF	$N = 6$	$N = 8$	$N = 9$	$N = 11$	$N = 12$	$N = 9$	$N = 9$	$N = 9$
Parameters	$l = 17$	$l = 3$	$l = 5$	$l = 7$	$l = 15$	$l = 5$	$l = 17$	$l = 13$
LBP	32.72	16.39	17.39	28.42	10.12	15.50	9.77	7.65
LPQ	20.72	13.80	19.20	43.59	12.66	15.04	9.54	6.05
BSIF	**18.53**	**13.11**	**14.30**	**18.08**	**0.86**	**11.30**	**2.11**	**4.54**
MB-LBP	19.68	22.30	16.52	24.28	1.10	12.01	3.12	4.98
avg	22.91	16.40	16.85	28.59	6.19	13.46	6.14	5.80

6.6 RESULTS AND DISCUSSION

6.6.1 *Known PAI species*

6.6.1.1 *Texture Analysis*

In order to analyse the detection performance, two sets of experiments were carried out. In the first experiment set, we optimise the detection performance of our texture descriptors in terms of the D-EER for different parameter configurations. Table 6.3 shows the D-EER for the best parameter setting over the development set in the LgA and PhA scenarios. As it may be observed, among all speech-to-image domain transformations, the CQT reports the best detection performance, thereby resulting a mean D-EER of 6.14% and 5.80% for the LgA and PhA scenarios, respectively. In addition, among the texture descriptors, the BSIF unveils the best texture features for the audio PAD task: D-EERs of 0.86% and 4.54% are attained for the LgA with STFT and PhA with CQT, respectively, thereby showing its suitability for the audio PAD task.

In a second set of experiments, we evaluate our FV approach for the best texture descriptor (i.e., BSIF) for known PAI species detection. Tab. 6.4 shows the FV detection performance for the best number of Gaussian clusters per speech-to-image domain transformation. We can observe that STFT achieves the same detection performance for the LgA scenario for the pooled database (i.e., all PAI species) as the one reported by the single BSIF descriptor: a D-EER of 0.86%, which is approximately three times lower than the one attained by the FV encoding with the images in the CQT domain. Alternatively, a different detection performance is reported for the PhA scenario

Table 6.4: Benchmark of our FV method and BSIF for known PAI species.

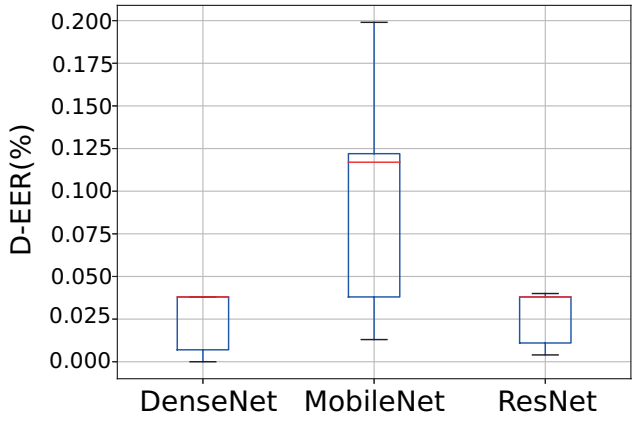
		CQCC		**LFCC**		**STFT**		**CQT**	
		$K = 512$		$K = 512$		$K = 256$		$K = 512$	
	PAI	D-EER	t-DCF	D-EER	t-DCF	D-EER	t-DCF	D-EER	t-DCF
LgA	A01	9.69	0.2973	5.70	0.1749	**0.38**	**0.0101**	1.15	0.0454
	A02	3.99	0.1202	11.22	0.3195	**0.62**	**0.0201**	0.92	0.0294
	A03	11.40	0.3503	6.59	0.2065	**0.74**	**0.0225**	1.19	0.0388
	A04	8.76	0.2650	19.34	0.5258	**0.84**	**0.0282**	2.68	0.0804
	A05	7.14	0.2260	11.40	0.3374	**1.50**	**0.0516**	2.46	0.0715
	A06	17.77	0.5044	12.73	0.5850	**0.82**	**0.0278**	4.09	0.1331
	pooled	10.28	0.3060	13.27	0.3660	**0.86**	**0.0294**	2.55	0.0748
		$K = 256$		$K = 128$		$K = 512$		$K = 256$	
PhA	AA	16.26	0.4332	39.64	0.9045	21.32	0.5416	**8.11**	**0.2038**
	AB	7.65	0.2184	27.26	0.6930	14.28	0.3763	**2.43**	**0.0702**
	AC	6.23	0.1743	22.34	0.5767	9.75	0.2530	**2.30**	**0.0671**
	BA	15.09	0.3859	33.26	0.8107	9.54	0.2334	**5.19**	**0.1291**
	BB	6.53	0.1814	23.05	0.5933	6.19	0.1568	**1.45**	**0.0373**
	BC	5.45	0.1502	19.03	0.5069	4.64	0.1141	**1.16**	**0.0352**
	CA	15.56	0.4072	32.51	0.7987	9.19	0.2299	**5.25**	**0.1365**
	CB	6.42	0.1719	22.73	0.6009	6.20	0.1566	**1.21**	**0.0341**
	CC	5.11	0.1374	19.23	0.5136	4.19	0.1105	**0.96**	**0.0275**
	pooled	9.94	0.2675	27.34	0.6784	11.05	0.2700	**3.68**	**0.0976**

where a D-EER of 3.68% for the CQT outperforms the one attained by STFT (i.e., 11.05%). Finally, it may be noted that the minimum normalised tandem Detection Cost Function (t-DCF)[2] values for both the STFT on LgA and CQT on PhA are respectively below 0.05 and 0.14, hence indicating that the FV provides a high security against PAIs to the ASV systems.

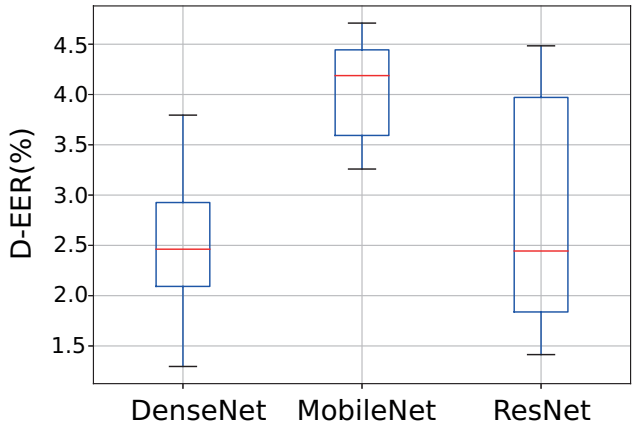
6.6.1.2 *Reliability of the Spectrogram Fusion*

Now, we evaluate the feasibility of fusing the best two speech-image representations (i.e., STFT and CQT) on the detection performance of our dual-stream temporal CNN. To that end, we select three different

2 t-DCF [111] is the primary metric used for the ASVspoof 2019 challenge (https://www.asvspoof.org/), which evaluates the performance between the proposed PAD approaches and an ASV system.



(a) LgA evaluation



(b) PhA evaluation

Figure 6.3: Detection performance per backbone for known PAI species.

backbones which have reported remarkable results in several pattern recognition tasks [92]: DenseNet with 121 layers [91], ResNet with 34 layers [86], and MobileNet version 2 [177]. As it can be seen in Tab. 6.2, the AP samples represent 90% of the whole dataset. Therefore, we select randomly for this experiment the same number of AP samples as BPs at 5 different iterations. Then, we train our proposed method for each random subset and report the D-EER for known PAI species scenarios in Fig. 6.3. As it can be observed, our approach achieves a mean D-EER lower than 0.10% and 4.04% for LgA and PhA, respectively. Whereas ResNet attains the best mean D-EER of 0.03% $\pm$ 0.02 for LgA, DenseNet reports the best mean D-EER of 2.51% $\pm$ 0.93 for PhA. The latter outperforms our common feature space (i.e., FV) combined with BSIF by a relative 31.79%. We also note that the minimum D-EER is yielded by DenseNet for LgA (i.e., D-EER = 0.00%) and PhA (i.e., D-EER = 1.30%). Finally, we observe that low STDs ranging 0.02-0.16 and 0.60-1.34 for LgA and PhA, respectively, indicate that the random selection of APs does not considerably impact the algorithm's detection performance.

Table 6.5: Benchmark in terms of D-EER(%) of the texture descriptors for unknown PAI species detection.

	CQCC		LFCC		STFT		CQT	
Method	LgA	PhA	LgA	LgA	LgA	LgA	LgA	LgA
LBP	24.61	17.87	28.30	25.88	22.44	16.26	17.43	7.16
LPQ	25.10	18.83	24.10	29.77	20.06	16.94	16.89	5.65
BSIF	**20.35**	**14.29**	**18.36**	**19.74**	15.33	**11.29**	**14.73**	**4.94**
MB-LBP	24.84	28.27	20.11	25.20	**10.69**	13.26	15.48	6.13
avg	23.73	19.82	22.72	25.15	17.27	14.44	16.13	5.97

6.6.2 *Unknown PAI species*

6.6.2.1 *Texture Analysis*

As mentioned in Sect. 6.5, one goal of this work is to analyse traditional texture descriptors for unknown PAI species detection. To that end, we select the evaluation dataset and assess the detection performance for the adopted texture descriptors by setting up the same parameters reported for the known PAI species experiment.

The corresponding results are presented in Tab. 6.5. We can note that the BSIF descriptor attains again the best detection performance for most speech-to-image transformations: a D-EER of 4.94% for PhA, which is close to the one reported by the known PAI species scenario (i.e., 3.68%). In addition, the MB-LBP outperforms the remaining descriptors for the STFT-LgA scenario, achieving a D-EER of 10.69%.

Based on this fact, we also evaluated the combination between BSIF and FV for each particular PAI species for the LgA and PhA scenarios in Tab. 6.6 and established a benchmark against the baselines *B*01 and *B*02. As it can be observed, the CQT achieves the best detection performance for the entire set of LgA-PAIs (i.e., a D-EER of 6.83%). In addition, this outperforms the adopted baselines for the most challenging PAIs for LgA scenario under the ASVSpoof 2019 database: D-EERs of 7.91% and 1.94% are respectively achieved for A10 and A13, which are two and five times lower than the ones reported by the baselines. Moreover, their corresponding t-DCF values are better than the ones attained by the baselines.

Consequently, for the PhA scenario our CQT-based FV approach attains for the pooled a D-EER of 3.66%, which is three times lower than the one yielded by the baselines (i.e., a D-EER of 11.04% for *B*01 and a D-EER of 13.54% for *B*02). Furthermore, we outperform the baselines for most PAI species: a D-EER in the range 1.10-7.64%

Table 6.6: Benchmark with the state of the art (*B*01 and *B*02) of our FV method and BSIF for unknown PAI species.

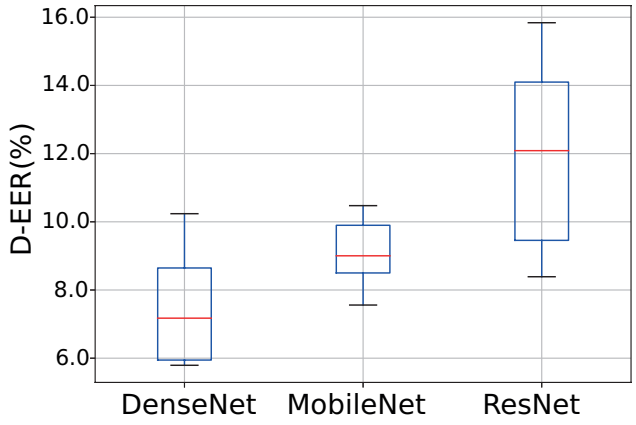
		CQCC		LFCC		STFT		CQT		B01		B02	
		$K = 512$		$K = 512$		$K = 256$		$K = 512$		$K = 2$		$K = 2$	
	PAI	D-EER	t-DCF	D-EER	t-DCF	D-EER	t-DCF	D-EER	t-DCF	D-EER	t-DCF	D-EER	t-DCF
LgA	A07	7.80	0.2445	17.78	0.5147	0.31	0.0100	4.45	0.1376	**0.00**	**0.0000**	12.86	0.3263
	A08	6.62	0.1958	1.75	0.0466	1.25	0.0356	4.60	0.1431	**0.04**	**0.0007**	0.37	0.0086
	A09	3.23	0.0944	1.16	0.0351	0.30	0.0089	0.89	0.0270	0.14	0.0060	**0.00**	**0.0000**
	A10	9.29	0.2784	17.35	0.5111	10.81	0.3140	**7.91**	**0.2438**	15.16	0.4149	18.97	0.5089
	A11	2.18	0.0685	9.32	0.2688	1.40	0.0430	3.41	0.1032	**0.08**	**0.0020**	0.12	0.0027
	A12	7.08	0.2212	17.21	0.4900	5.84	0.1689	5.08	0.1560	**4.74**	**0.1160**	4.92	0.1197
	A13	8.30	0.2648	23.91	0.6964	5.01	0.1415	**1.94**	**0.0634**	26.15	0.6729	9.57	0.2519
	A14	8.83	0.2686	8.74	0.2585	3.15	0.0971	2.05	0.0638	10.85	0.2629	**1.22**	**0.0314**
	A15	4.56	0.1415	5.13	0.1517	4.32	0.1345	4.48	0.1377	**1.26**	**0.0344**	2.22	0.0607
	A16	7.54	0.2363	15.22	0.4259	0.74	0.0234	2.04	0.0669	**0.00**	**0.0000**	6.31	0.1419
	A17	34.39	0.9115	21.22	0.5745	31.20	0.8643	16.62	0.4766	19.62	0.9820	**7.71**	**0.4050**
	A18	36.08	0.9536	32.19	0.8853	5.94	**0.1793**	10.28	0.3080	3.81	0.2818	**3.58**	0.2387
	A19	26.94	0.7321	23.08	0.6628	4.62	0.1388	11.49	0.3454	**0.04**	**0.0014**	13.94	0.4635
	pooled	14.62	0.3691	16.79	0.3837	7.76	**0.1881**	**6.83**	0.1926	9.57	0.2366	8.09	0.2116
		$K = 256$		$K = 128$		$K = 512$		$K = 256$		$K = 2$		$K = 2$	
PhA	AA	23.14	0.5396	37.94	0.8827	19.45	0.4954	**7.64**	**0.1985**	25.28	0.4975	32.48	0.7359
	AB	17.27	0.4162	26.89	0.7049	14.44	0.3714	**2.05**	**0.0548**	6.16	0.1751	4.40	0.1295
	AC	11.28	0.2859	22.55	0.5913	10.70	0.2772	2.17	0.0594	**2.13**	**0.0529**	3.95	0.1121
	BA	19.74	0.4967	31.24	0.7644	10.78	0.2785	**5.17**	**0.1360**	21.87	0.4658	24.59	0.6011
	BB	14.40	0.3679	23.11	0.6132	7.06	0.1877	**1.22**	**0.0341**	5.26	0.1483	4.29	0.1252
	BC	9.66	0.2517	19.32	0.5157	5.42	0.1458	**1.23**	**0.0336**	1.61	0.0433	3.20	0.0888
	CA	18.64	0.4709	28.35	0.7084	9.88	0.2568	**5.46**	**0.1408**	21.10	0.5025	21.63	0.5524
	CB	13.08	0.3358	22.40	0.5926	6.27	0.1692	**1.22**	**0.0335**	4.70	0.1360	3.92	0.1194
	CC	9.05	0.2383	18.83	0.5087	5.22	0.1382	**1.10**	**0.0308**	1.79	0.0461	3.06	0.0895
	pooled	15.68	0.3837	26.20	0.6649	11.21	0.2815	**3.66**	**0.0946**	11.04	0.2454	13.54	0.3017

together with a t-DCF between 0.03-0.20% unveils a reliable and secure generalisation capability for this scenario.

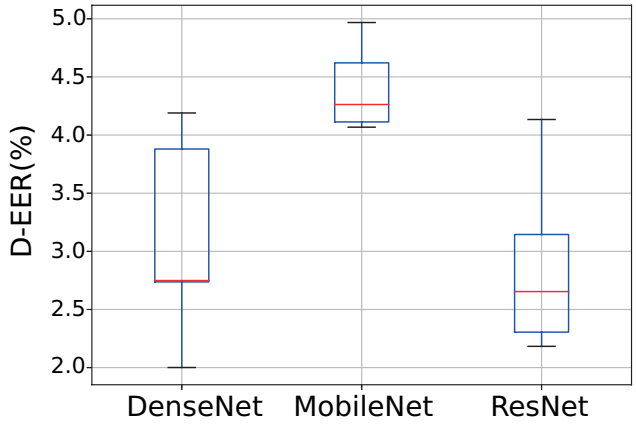
6.6.2.2 *Spectrogram Fusion*

Now, we compute the detection performance of our dual-stream combined with the three studied backbones (i.e., DenseNet, MobileNet, and ResNet) in Fig. 6.4. We observe that the mean D-EER is multiplied by a factor of 82 for DenseNet, 93 for MobileNet, and 457 for ResNet in comparison with the D-EERs reported for the known PAI species evaluation. Specifically, the best performing backbone (i.e., DenseNet) achieves a mean D-EER of 7.56% $\pm$ 1.89, resulting in a minimum D-EER of 5.79%.

In contrast to the results reported for LgA, the detection performance per backbone for PhA is similar to those yielded for the known PAI species. Mean D-EERs ranging 2.88%-4.41% confirm the above observation over the same experiment: the features for unknown samples stemming from the PhA partition follow a similar distribution to the ones for the spectrograms in the training set. Hence, we strongly think that algorithms for voice PAD should be able to achieve similar results for known PAI species and unknown PAI species over the PhA partition.



(a) LgA evaluation



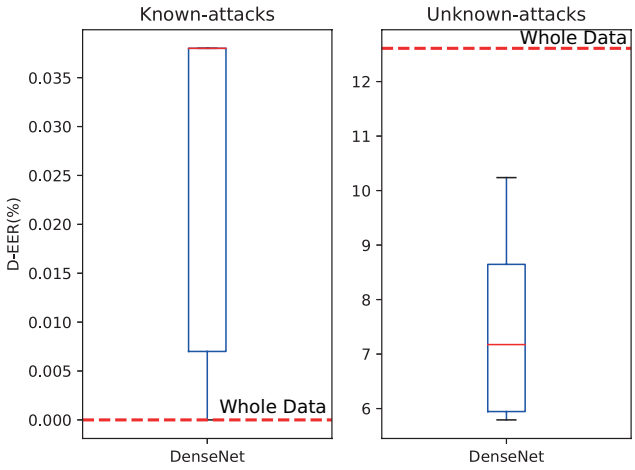
(b) PhA evaluation

Figure 6.4: Detection performance per backbone for unknown PAI species.

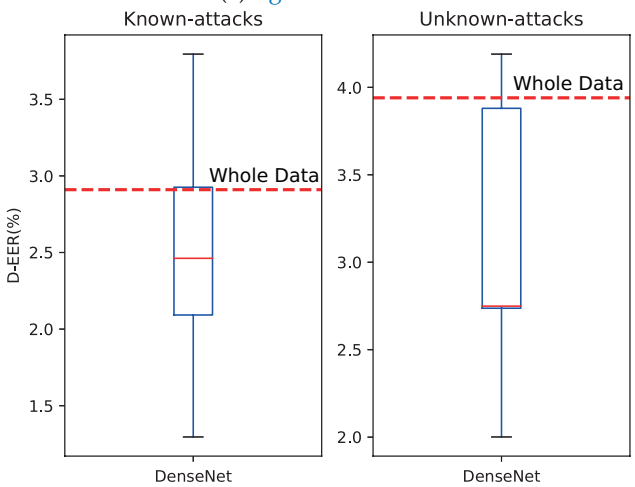
6.6.2.3 *Impact of Unbalanced Dataset over unknown PAI species*

On the other hand, we evaluate to what extent the detection performance of our CNN-based method is affected when trained with the entire database. To that end, we selected the best performing backbone (i.e., DenseNet). In order to avoid bias in classifier training, we optimise the BCE loss in our approach by setting up weights per category (i.e., 0.90 for BPs and 0.10 for APs). Fig. 6.5 shows a benchmark of our proposed algorithm when it is trained with unbalanced (i.e., the entire dataset, red dashed line) and balanced (i.e., a random selection of AP samples) databases. We can see that the training with the entire database yields similar results to those attained by the random selection of samples (see *a*) and *b*), column 1). Even, it achieves a D-EER of 0.00% for LgA (see *a*), column 1) which is lower than the mean value reported by training with a balanced database. This is because the features computed for the evaluation set follow the same distribution as those of the training set.

In contrast to the results reported for known PAI species, we can observe that training with an unbalanced database considerably in-



(a) LgA evaluation



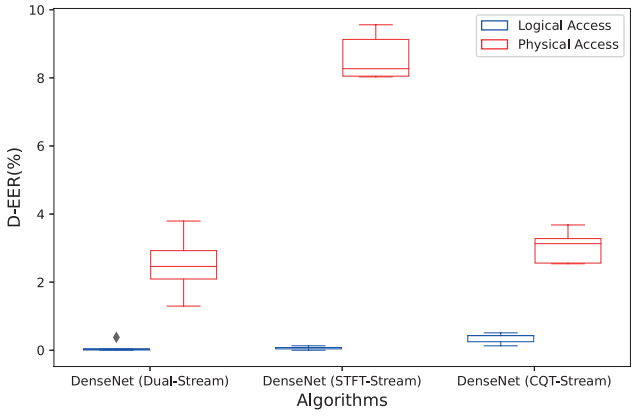
(b) PhA evaluation.

Figure 6.5: Benchmark of our proposed method trained with data random selection and the entire data (dashed red line).

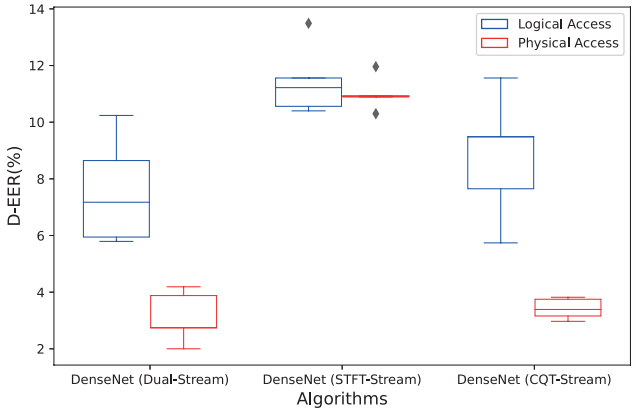
creases the D-EER compared to training using the same number of samples per category for unknown PAI species. In particular, a D-EER of 12.61%, which is approximately twice higher than the one reported by the mean of the data random selection (i.e., 7.56%), is achieved for LgA. Subsequently, we can also note a decrease in the detection performance of our proposed approach when trained with the unbalanced database: a D-EER of 3.94% for an unbalanced database vs. a mean D-EER of 3.11% for a balanced database confirms the impact of training with an unbalanced database in the unknown PAI species detection. We think that future studies focused on the PAD generalisation should consider the issues of unbalanced databases.

6.6.2.4 *Ablation Study*

Finally, we conduct an ablation study of the dual-stream with respect to each stream separately. Fig. 6.6 establishes a benchmark between the



(a) Known PAI species evaluation



(b) Unknown PAI species evaluation.

Figure 6.6: Performance benchmark of the dual-stream with respect to each stream separately.

proposed dual-stream with respect to each single scheme (i.e., STFT or CQT). As it can be observed, the dual-stream scheme is capable of outperforming both singular streams for the known PAI species and unknown PAI species scenarios. In particular, for the challenging unknown PAI species scenario, the fusion algorithm achieves mean D-EERs of 7.56% and 3.11% for LgA and PhA, respectively, which improve the single pipelines by up to a relative 33.97% and 71.73%. These results do confirm that the fusion between the best performing spectrogram representations (i.e., STFT and CQT) improves the detection performance of every single pipeline, especially for the unknown PAI species scenario. These two spectrogram representations contain different frequency information that complements each other to improve the final decision.

6.6.3 *Cross-database Evaluation*

We also assess the ability of our dual-stream temporal CNN to spot PAIs across different databases and compare it with our FV represen-

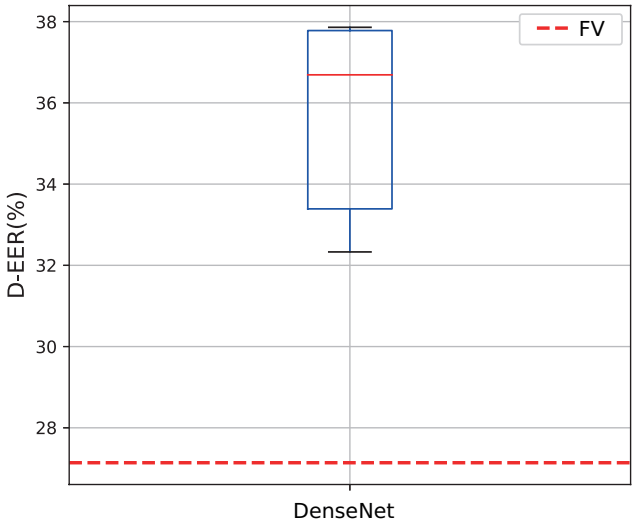
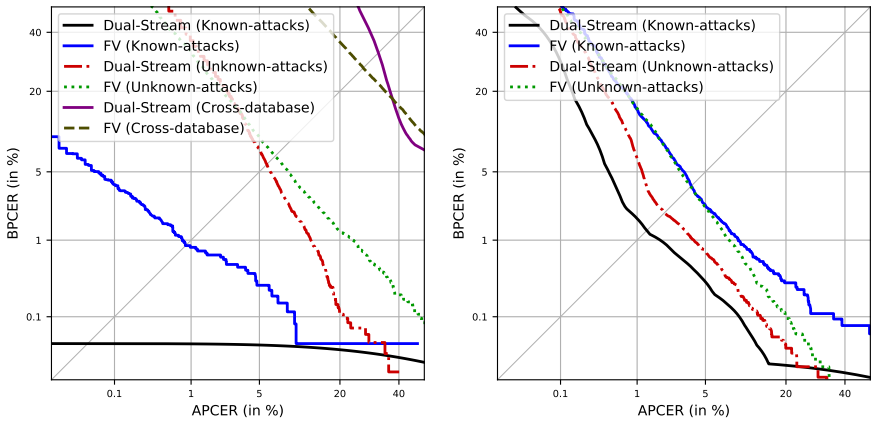


Figure 6.7: Cross-database evaluation for the best performing backbone (i.e., DenseNet).

tation. To that end, we follow the cross-database protocol defined in Sect. 6.5 and compute in Fig. 6.7 the D-EER for the best performing backbone (i.e., DenseNet) over the models trained over the five random sets mentioned in Sect. 6.6.1.2. We observe that the CNN-based algorithm achieves a mean D-EER of 35.61% with a STD of 2.58% which is worse than the one attained by our common feature space FV (i.e., 27.14%). Depending on the selection of the training set, a minimum D-EER of 32.33% is yielded, which shows that the selection of training samples is a challenge for PAD generalisation and should be taken into account in future research. In addition, these results confirm the need of enhancing the generalisation capability of neural networks. A considerable improvement of our results for this challenging scenario would be the combination of our dual-stream temporal CNN with those backbones which are developed for instance for domain adaptation [213]. Furthermore, the latest CNN families of EfficientNet architectures [197] could enhance the final decision of our framework.

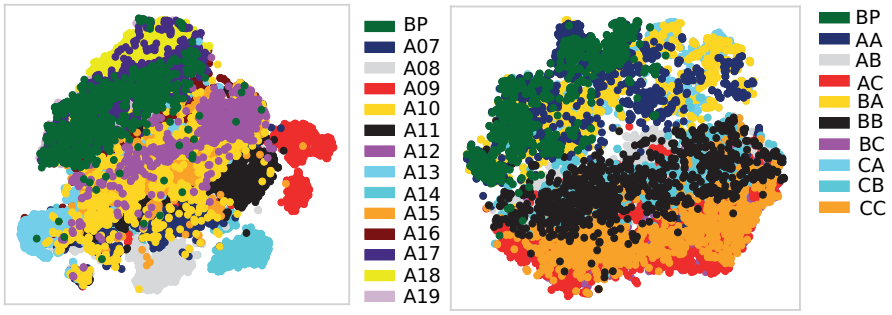
6.6.4 *In-depth Performance Analysis*

We establish a benchmark in Fig. 6.8 between both proposed algorithms for LgA and PhA. As it can be seen, the dual-stream approach considerably outperforms the FV for known PAI species and unknown PAI species in most scenarios. In particular, for PhA, a BPCER $\leq$ 1.78% and a BPCER $\leq$ 6.59% at an APCER $\geq$ 1.0% for known PAI species and unknown PAI species, respectively confirm the soundness of our learnable features for operating over this challenging scenario. For LgA, the FV technique reports, for a high-security threshold (i.e.,



(a) LgA evaluation (b) PhA evaluation

Figure 6.8: Benchmark for known PAI species, unknown PAI species, and cross-database. Diagonal light-gray lines represent the D-EER (%).



(a) LgA common feature space. (b) PhA common feature space.

Figure 6.9: t-SNE visualization of common feature spaces learned by the FV-based approach for the CQT transformation.

APCER = 1.0%), a BPCER = 32.83% and BPCER = 82.72% for unknown PAI species and cross-database, respectively. These are better than the ones attained by the dual-stream method. Consistent with the results shown in Fig. 6.7, the cross-database performance computed by our both algorithms suffers a significant decrease for high security thresholds, thus indicating the need for further research on these scenarios.

6.6.5 *Visualisation of the FV Representation*

Finally, a t-SNE visualisation in Fig. 6.9 shows that most PAI species share more homogeneous features with each other than with those BPs. Thus, this confirms our hypothesis mentioned in Chapter 1. In spite of the results, we note the overlap of some PAI species such as A17, A18, AA, BA, and CA with the BP features. This indicates that the data distribution learned by a GMM using the BSIF features needs to

be improved in order to get a better generalisable FV common feature space.

6.6.6 *Summary*

To summarise the findings on the voice PAD, we can highlight the following takeaway messages:

- Among four image representations of speech audio, CQT and STFT unveiled textural differences between BPs and APs, thereby resulting in the best detection performance across several traditional texture descriptors.
- In particular, BSIF representation reported the best detection performance for known PAI species and unknown PAI species scenarios.
- The combination of BSIF with our common feature space (i.e., FV) yielded D-EERs in the range 0.86%-6.83% for known PAI species and unknown PAI species scenarios, thus reporting a performance improvement with respect to the use of BSIF alone.
- The fusion of the two best spectrograms (i.e., CQT and STFT) through our dual-stream temporal CNN method outperformed the FV representation for most scenarios. More specifically, a BPCER $\leq$ 1.78% and a BPCER $\leq$ 6.59% at an APCER $\geq$ 1.0% for known PAI species and unknown PAI species, respectively confirm the soundness of our learnable features for operating over this challenging scenario.
- Further, this fusion reports D-EERs which outperform each spectrogram-based pipeline by up to a relative 33.97% and 71.73% for LgA and PhA, respectively. This does confirm that the complementary information between STFT and CQT leads to an improvement in the detection of unknown PAI species.
- In contrast to the results for PhA, the FV confirmed its soundness for the challenging unknown PAI species and cross-database scenarios over LgA. This outperformed the dual-stream approach for higher security thresholds (i.e., APCER $\leq$ 1.0%).
- In spite of the results, the FV representation decreases its performance for this challenging scenarios over LgA: a BPCER = 32.83% and BPCER = 82.72% for unknown PAI species and cross-database, respectively indicate the need for further research on these scenarios.
- The results in Sect. 6.6.2.3 confirmed the need of further research about the selection of PAI species to train PAD algorithms. The

training over an unbalanced dataset considerably decreases the generalisation capabilities of the PAD module.

- Finally, the results also confirmed the feasibility of using our common feature space FV for voice PAD.

7

CONCLUSIONS AND FUTURE DIRECTIONS

Nowadays, numerous investigations carried out over a number of years have shown that APs launched over the capture device to impersonate someone else are a real threat to the security of biometric systems. To prevent these threats, different PAD algorithms have been developed in the last decades. Those techniques have reported a remarkable detection performance to spot AP attempts whose PAI species are known a priori. However, they drop their accuracy when unknown PAI species are employed. In addition, those methods are specifically proposed for a particular biometric characteristic, hence their application on a different biometric characteristic leads to high-performance degradation.

This Thesis investigated and proposed new algorithms mainly focused on improving the generalisation capabilities in challenging scenarios where either PAI species or capture devices remain unknown. Specifically, we explored the use of different techniques which propose the definition of a semantic common feature space summarising those features of BPs and APs that persist in different unknown PAI species. Among three different common feature spaces studied, the FV representation appeared to be capable of improving the generalisation capabilities across different biometric characteristics such as fingerprint, face, and voice. In essence, this algorithm characterises how the distribution of a set of local descriptors, extracted from unknown PAI species, differs from the distribution of known APs and BPs, which was previously learned by a generative model. Therefore, the final transformed features are more robust to new samples, which may stem from unknown scenarios and thus differ from the samples used for training.

The FV representation was successfully evaluated over three different types of biometric characteristics (i.e., fingerprint, face, and voice) in combination with several handcrafted descriptors. The experimental results reported in terms of the metrics defined in the international ISO/IEC 30107-3 for biometric PAD [97] showed a remarkable detection performance over both unknown PAI species and cross-database scenarios for these three biometric modalities. It is worth noting that the FV common feature space was assessed in the international Fingerprint Liveness Detection (LivDet) competition 2019, resulting in the best overall accuracy among participants (i.e., overall accuracy of 96.17% [148]).

In general, the contributions of our Thesis per research question are:

> ***RQ 1**: Keeping in mind that fingerprints consist of ridges and valleys, can the lack of ridge's continuity be used to detect the artefacts produced in the fabrication of PAIs?*
> *Is there a close relationship between the lack of ridge continuity and those artefacts?*
> *Can these features aid in successfully detecting unknown PAI species?*

The experimental results in Sect. 4.3 showed that gradient-based descriptors (i.e., SIFT and SURF) successfully represent low coherence areas produced by several fingerprint ridge pattern artefacts such as black saturation, white saturation, lack of continuity, unwanted noises, and ridge distortions, thereby resulting in the best detection performance in most scenarios. In addition, a NFIQ2.0 evaluation over the LivDet 2015 database depicted that the different analysed descriptors improved their detection performance as the ridge pattern of BP fingerprints enhanced: an D-EER $< 2.58\%$ is reported when the BP fingerprint quality is greater than 60 (i.e., NFIQ2.0 > 60). Therefore, the lack of ridge's continuity can be used as a suitable indicator to separate a BP from an AP. Finally, the combination between the SIFT and the FV representation reported a remarkable detection performance in challenging scenarios such as unknown PAI species, cross-database, and cross-session.

> ***RQ 2**: Can different colour spaces unveil discriminative features to be capable of successfully detecting facial PAIs?*
> *How can the facial artefacts, produced in the creation of PAIs, be perceived in different colour spaces?*

Boulkenafet *et al.* reported in [17] that the RGB colour space has limited discriminative power for face PAD due to the high correlation between the three colour components. In contrast, HSV and YC_bC_r are based on the separation of the luminance and chrominance components, thereby providing additional information for learning more discriminative features. However, the experimental results in this Thesis indicated the statistical advantage of RGB with respect to other colour spaces for datasets having images of varying resolutions, thereby resulting in a minimum average D-EER of 0.45% for known PAI species detection. Contrary to the conclusions drawn in [18], we observed that the three explored colour spaces reported similar error rates in three out of four datasets (i.e., REPLAY-ATTACK, REPLAY-MOBILE, and MSU-MFSD): mean D-EERs of 0.003%, 0.03%, and 0.19% are achieved by RGB, HSV, and YC_bC_r respectively. The main reason for this difference with respect to [18] is that we carried out a feature decorrelation with PCA before finding the semantic sub-groups, thereby leading to the detection of similar features for the three colour spaces.

***RQ 3**: What is the most appropriate facial region to identify PAIs?*
Taking into account that the face consists of several regions such as the mouth, eyes, eyebrows, or chin, then how many facial regions are required to correctly identify a PAI.
What is the minimum or the optimal number of facial regions needed to detect PAIs?

The experimental results in Sect. 5.5.7 showed that the composite regions achieved the best detection performances among regions. In particular, the full face yielded a D-EER of 3.92%, followed by right face (D-EER = 4.61%), left face (D-EER = 5.38%), the jaw (D-EER = 5.53%), and central face (D-EER = 6.28%). However, the use of the nose region is capable of achieving error rates comparable to those reported by the composite regions. In addition, we proposed a *Facial Region Utility* metric which indicated the usefulness of a particular region for training to spot an AP based on the other region in a probe image. A practical application where individuals wore masks to prevent respiratory infections showed the feasibility of using the central face over the full face in the evaluation: an ACER of 9.51% which outperforms the state-of-the-art methods by a relative improvement up to 67.73%, allows the building of a secure and convenient PAD module. In addition, we noted that the BPCER values yielded by the state-of-the-art were decreased down to 7.02% (BM0) and 15.79% (BM1) for pristine subjects.

***RQ 4**: Can the image resolution affect face PAD process?*

1) *Given that several lower, medium, and high resolution capture devices are employed for acquiring face images, how can the facial artefacts be detected in different image resolutions?*
2) *Keeping in mind that numerous lower, medium, and higher resolution capture devices are employed for replay attacks, how can the image resolution of such devices affect or help the detection capability of PAD approaches?*
3) *How does the combination between replay and capture device resolutions affect the detection capability of PAD approaches?*

In Sect. 5.5.6, we explored the impact of using image of varying resolutions to detect APs. We noted that the utilisation of images of varying quality for the facial PAD produces a high-performance decrease, which can be even greater than the use of numerous PAI species; *ii*) the current deep learning-based descriptors report the worst PAD performance deterioration for face images whose sizes are widely different from the input layer size; *iii*) Video-replay attacks,

screened on a high-resolution capture device, include several blurriness and sharpness properties, which can be successfully detected by PAD techniques; and *iv*) training PAD methods with several PAI species which are acquired with varying-resolution capture devices appears to be the worst case for the face PAD task, thereby resulting in a D-EER of 24.24% and a joint $BPCER_{100}$ of 67.23%. Therefore, we can confirm that the image resolution is a requirement which must carefully be taken into account in order to build a secure and reliable face PAD module.

> ***RQ 5**: Can a general framework be built to successfully detect known PAI species and unknown PAI species by generalising across different biometric characteristics*

The main finding of this Thesis was the use of semantic common feature spaces to improve PAD generalisation capabilities. In particular, we reported that the FV representation is able to define semantic feature sub-groups from known samples which are then found in unknown PAI species. The experimental evaluation conducted over three different types of biometric characteristics (i.e., fingerprint, face, and voice) confirmed the soundness of FV to detect both known PAI species and unknown PAI species. Therefore, this representation based on the combination of generative and discriminative models can be successfully employed to build a general framework for biometric PAD.

7.1 FUTURE WORK

Based on the findings of this Thesis, some future directions emerge:

- As it was shown through this Thesis, the FV representation is able to define a semantic common feature space that allows improving the generalisation capability on different biometric characteristics. This approach was combined with traditional handcrafted descriptors such as SIFT, SURF, BSIF, and LBP, among others. Therefore, its combination with deep features learnt from powerful CNNs is expected to enhance our results.
- Further, the FV computation with more powerful generative models such as VAEs and GANs could improve results for the challenging cross-database scenario. A major advance would be the combination of the FV representation with the above deep generative models in an end-to-end scheme. Thus, we would have a robust approach capable of building a secure biometric system against unknown PAI species.
- Given that the FV representation reported a remarkable generalisation ability over different types of biometric characteristics (i.e.,

fingerprint, face, and voice), we strongly think that transforming its features into a multimodal common feature space would be a viable way forward to overcome the current state-of-the-art limitations.

- Since our research is mostly focused on capture devices that acquire the images under visible spectrum, the properties of near-infrared images might be analysed for improving PAD.
- Finally, the feasibility of our PAD subsystems should be evaluated for the emerged contactless fingerprint attacks.

GLOSSARY

APCER	Attack Presentation Classification Error Rate. "proportion of attack presentations using the same PAI species wrongly classified as bona fide presentations in a specific scenario" [97]. xvi, xix, xx, 9, 43, 44, 51, 69, 72, 75, 76, 83, 94, 95, 111–113, 121, 122
attack presentation	"presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system" [97]. An attack to the capture device to either conceal the own identity or impersonate someone else. xiii, 1, 3, 6, 7, 9, 121
biometric characteristic	"biological and behavioural characteristic of an individual from which distinguishing, repeatable biometric features can be extracted for the purpose of biometric recognition" [97]. v, xviii, 1, 2, 4–7, 9–11, 17, 115, 118, 123
bona fide presentation	"interaction of the biometric capture subject and the biometric data capture subsystem in the fashion intended by the policy of the biometric system" [97]. A normal or pristine presentation. 3, 7, 9, 121
BPCER	Bona fide Presentation Classification Error Rate. "proportion of bona fide presentations missclassified as attack presentations in a specific scenario" [97]. xvi, xix, xx, 9, 43, 44, 51, 55, 69, 72, 75, 76, 83, 85, 94, 95, 111–113, 117, 121, 122
$BPCER_{100}$	BPCER at a fixed operation point APCER = 1%, i. e., 1/100 attack presentations is misclassified. 9, 43, 46, 48, 49, 51, 52, 68, 77, 95, 96, 118
$BPCER_{10}$	BPCER at a fixed operation point APCER = 10%, i. e., 10/100 attack presentations are misclassified. xix, 9, 46, 94
$BPCER_{20}$	BPCER at a fixed operation point APCER = 5%, i. e., 5/100 attack presentations are misclassified. 9, 46, 85

cross-database — "scenario where the capture device employed for the acquisition of test samples is different from the one used for capturing the training images. Both datasets contain the same PAI species to ensure that the performance degradation is due to the dataset change and not to the unknown PAI species". xi, xii, xv–xix, 9, 10, 13–16, 32, 46–51, 61, 76, 77, 101, 110–113, 115, 116, 118

cross-session — "scenario where different data collection sessions across different seasons or even years for the same capture device are used for training and testing". xi, xv, xviii, 9, 32, 46–50, 116

D-EER — Detection Equal Error Rate. PAD operation point where APCER = BPCER. xvii–xix, xxi, 9, 13, 14, 33–42, 44–50, 52, 59, 66–77, 80–84, 86–90, 92, 95, 96, 103–113, 116–118

FMR — False Match Rate. "proportion of the completed biometric non-mated comparison trials that result in a false match" [94]. xxi, 75, 76

FNMR — False Non-Match Rate. "proportion of the completed biometric mated comparison trials that result in a false non-match" [94]. xxi, 75, 76

known PAI species — "scenario where an analysis of all PAI species is performed. In all cases, PAI species for testing are also included in the training set". v, xi, xii, xiv, xvi–xix, 5, 9, 10, 13, 14, 17, 24, 33, 43, 46, 53, 61, 66, 68, 70, 72, 73, 77, 79–83, 86, 87, 95, 100, 101, 103–108, 110–113, 116, 118

PAD — Presentation Attack Detection."automated determination of a presentation attack" [95]. v, xi, xii, xiv, xv, xviii, xix, xxii, 3–7, 9–17, 29, 31, 32, 36–38, 40, 44, 46, 50, 51, 53–57, 59–64, 66–69, 72, 75–77, 80–92, 94–101, 103, 104, 107, 109, 111, 113–119, 122

PAI — Presentation Attack Instrument. "biometric characteristic or object used in a presentation attack" [95]. For instance, a replayed face photo, a gummy fingerprint, or a replayed speech.. xvii, xxii, 1–8, 11–15, 18, 32, 38, 40, 43, 44, 49–53, 60, 69, 75, 77, 79, 84, 85, 101, 102, 104, 106, 107, 110, 116, 117

PAI species — "class of presentation attack instruments created using a common production method and based on different biometric characteristics" [97]. xiv, xv, xvii–xix, 4, 6–11, 13, 14, 33, 36, 38–40, 44, 45, 49, 54, 55, 61–65, 69, 71, 72, 74–76, 79, 80, 82–84, 96, 102, 103, 106, 112, 113, 115, 117, 118, 121–123

unknown — "not seen in training". 9, 13, 24, 44, 63, 64, 69, 93, 94, 107, 115, 122

unknown PAI species — "scenario where PAI species used for testing are not incorporated in the training set". v, xi, xii, xiv, xvi–xix, 4, 5, 9–11, 13–15, 17, 24, 32, 33, 41, 44–46, 49–51, 53, 61, 69–72, 74, 75, 84, 85, 95, 97, 100, 101, 106–113, 115, 116, 118

BIBLIOGRAPHY

[1] A. Abhyankar and S. Schuckers. "Fingerprint liveness detection using local ridge frequencies and multiresolution texture analysis techniques." In: *Proc. Intl. Conf. on Image Processing*. IEEE. 2006, pp. 321–324.

[2] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore. "Face presentation attack with latex masks in multispectral videos." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 81–89.

[3] A. Agarwal, A. Sehwag, R. Singh, and M. Vatsa. "Deceiving face presentation attack detection via image transforms." In: *Proc. Intl. Conf. on Multimedia Big Data (BigMM)*. IEEE. 2019, pp. 373–382.

[4] F. Alegre, A. Amehraye, and N. Evans. "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns." In: *Proc. Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*. 2013, pp. 1–8.

[5] A. Ali, F. Deravi, and S. Hoque. "Liveness detection using gaze collinearity." In: *Proc. Intl. Conf. on Emerging Security Technologies*. 2012, pp. 62–65.

[6] A. Antonelli, R. Cappelli, D. Maio, and D. Maltoni. "A new approach to fake finger detection based on skin distortion." In: *Advances in Biometrics*. Springer, 2005, pp. 221–228.

[7] S. R. Arashloo, J. Kittler, and W. Christmas. "Face spoofing detection based on multiple descriptor fusion using multiscale dynamic binarized statistical image features." In: *IEEE Trans. on Information Forensics and Security* 10.11 (2015), pp. 2396–2407.

[8] S. R. Arashloo, J. Kittler, and W. Christmas. "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol." In: *IEEE Access* 5 (2017), pp. 13868–13882.

[9] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu. "Face anti-spoofing using patch and depth-based CNNs." In: *Proc. Intl. Joint Conf. on Biometrics (IJCB)*. 2017, pp. 319–328.

[10] *Audio Processing Systems*. J. Wiley and L. Sons, 2008. Chap. 4, pp. 97–113.

[11] E. Auksorius and A. Boccara. "Fingerprint imaging from the inside of a finger with full-field optical coherence tomography." In: *Biomedical Optics Express* 6.11 (2015).

[12] H. Bay, T. Tuytelaars, and L. Van Gool. "Surf: Speeded up robust features." In: *Proc. Intl. European Conf. on Computer Vision (ECCV)*. 2006, pp. 404–417.

[13] A. Bhogal, D. Söllinger, P. Trung, and A. Uhl. "Non-reference image quality assessment for biometric presentation attack detection." In: *Proc. Intl. Workshop on Biometrics and Forensics (IWBF)*. IEEE. 2017, pp. 1–6.

[14] J. Bigun, H. Fronthale, and K. Kollreider. "Assuring liveness in biometric identity authentication by real-time face tracking." In: *Proc. Intl. Conf. on Computational Intelligence for Homeland Security and Personal Safety (CIHSPS)*. 2004, pp. 104–111.

[15] A. Bosch, A. Zisserman, and X. Munoz. "Image Classification using Random Forests and Ferns." In: *Proc. Int. Conf. on Computer Vision (ICCV)*. 2007.

[16] Z. Boulkenafet, J. Komulainen, and A. Hadid. "Face antispoofing using speeded-up robust features and fisher vector encoding." In: *IEEE Signal Processing Letters* 24.2 (2016), pp. 141–145.

[17] Z. Boulkenafet, J. Komulainen, and A. Hadid. "Face spoofing detection using colour texture analysis." In: *IEEE Trans. on Information Forensics and Security* 11.8 (2016), pp. 1818–1830.

[18] Z. Boulkenafet, J. Komulainen, and A. Hadid. "On the generalization of color texture-based face anti-spoofing." In: *Image and Vision Computing* 77 (2018), pp. 1–9.

[19] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. "OULU-NPU: A mobile face presentation attack database with real-world variations." In: 2017.

[20] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper. "Mixfacenets: Extremely efficient face recognition networks." In: *Proc. Intl. Joint Conf. on Biometrics (IJCB)*. IEEE. 2021, pp. 1–8.

[21] R. Bresan, A. Pinto, A. Rocha, C. Beluzo, and T. Carvalho. "FaceSpoof Buster: a Presentation Attack Detector Based on Intrinsic Image Properties and Deep Learning." In: *arXiv preprint arXiv:1902.02845* (2019).

[22] J. Brown. "Calculation of a constant Q spectral transform." In: *Journal of the Acoustical Society of America* 89.1 (), pp. 425–434.

[23] R. Cai, H. Li, S. Wang, C. Chen, and A. Kot. "DRL-FAS: A novel framework based on deep reinforcement learning for face anti-spoofing." In: *IEEE Trans. on Information Forensics and Security (TIFS)* 16 (2020), pp. 937–951.

[24] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. "Brief: Binary robust independent elementary features." In: *Proc. Intl. European Conf. on Computer Vision*. Springer. 2010, pp. 778–792.

[25] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury. "Generalization of Audio Deepfake Detection." In: *Proc. Odyssey*. 2020, pp. 132–137.

[26] X. Cheng, M. Xu, and T. F. Zheng. "Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019." In: *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC)*. 2019, pp. 540–545.

[27] I. Chingovska and A. Dos Anjos. "On the use of client identity information for face antispoofing." In: *IEEE Trans. on Information Forensics and Security* 10.4 (2015), pp. 787–796.

[28] I. Chingovska, A. Anjos, and S. Marcel. "On the Effectiveness of Local Binary Patterns in Face Anti-spoofing." In: 2012.

[29] F. Chollet. "Xception: Deep learning with depthwise separable convolutions." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition*. 2017, pp. 1251–1258.

[30] T. Chugh, K. Cao, and A. K. Jain. "Fingerprint spoof detection using minutiae-based local patches." In: *Proc. Intl. Joint Conf. on Biometrics (IJCB)*. 2017, pp. 581–589.

[31] T. Chugh, K. Cao, and A. K. Jain. "Fingerprint Spoof Buster: Use of Minutiae-Centered Patches." In: *IEEE Trans. on Information Forensics and Security* 13.9 (2018), pp. 2190–2202.

[32] T. Chugh and A. K. Jain. "Fingerprint presentation attack detection: Generalization and efficiency." In: *Proc. Intl. Conf. on Biometrics (ICB)*. IEEE. 2019, pp. 1–8.

[33] T. Chugh and A. K. Jain. "Fingerprint spoof detector generalization." In: *IEEE Trans. on Information Forensics and Security (TIFS)* 16 (2020), pp. 42–55.

[34] T. Chugh and A. Jain. "OCT Fingerprints: Resilience to Presentation Attacks." In: *arXiv preprint arXiv:1908.00102* (2019).

[35] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel. "The REPLAY-MOBILE Face Presentation-Attack Database." In: *Proc. Intl. Conf. on Biometrics Special Interests Group (BIOSIG)*. 2016.

[36] R. Cotterell and J. Eisner. "A deep generative model of vowel formant typology." In: *arXiv preprint arXiv:1807.02745* (2018).

[37] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. "Visual categorization with bags of keypoints." In: *Proc. Intl. Workshop on Statistical Learning in Computer Vision (ECCV)*. 2004, pp. 1–22.

[38] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection." In: *2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, pp. 886–893. ISBN: 0-7695-2372-2.

[39] L. Darlow, J. Connan, and A. Singh. "Performance Analysis of a Hybrid Fingerprint Extracted from Optical Coherence Tomography Fingertip Scans." In: *2016 Intl. Conf. on Biometrics (ICB)*. 2016.

[40] D. Deb and A. K. Jain. "Look locally infer globally: a generalizable face anti-spoofing approach." In: *IEEE Trans. on Information Forensics and Security (TIFS)* 16 (2020), pp. 1143–1157.

[41] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[42] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. "Arcface: Additive angular margin loss for deep face recognition." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition*. 2019, pp. 4690–4699.

[43] R. Derakhshani, S. Schuckers, L. Hornak, and L. O'Gorman. "Determination of vitality from a non-invasive biomedical measurement for use in fingerprint scanners." In: *Pattern recognition* 36.2 (2003), pp. 383–396.

[44] Y. Ding and A. Ross. "An ensemble of one-class SVMs for fingerprint spoof detection across different fabrication materials." In: *Proc. Intl. Workshop on Information Forensics and Security (WIFS)*. IEEE. 2016, pp. 1–6.

[45] M. Drahansky. "Experiments with skin resistance and temperature for liveness detection." In: *Intl. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE. 2008, pp. 1075–1079.

[46] M. Drahansky and D. Lodrova. "Liveness detection for biometric systems based on papillary lines." In: *Intl. Conf. on Information Security and Assurance*. IEEE. 2008, pp. 439–444.

[47] M. Drahansky, R. Notzel, and W. Funk. "Liveness detection based on fine movements of the fingertip surface." In: *Proc. Information Assurance Workshop*. IEEE. 2006, pp. 42–47.

[48] R. K. Dubey, J. Goh, and V. LL. Thing. "Fingerprint Liveness Detection From Single Image Using Low-Level Features and Shape Analysis." In: *IEEE Trans. on Information Forensics and Security* 11.7 (2016), pp. 1461–1475.

[49] R. K. Dubey, J. Goh, and V. Thing. "Fingerprint Liveness Detection From Single Image Using Low-Level Features and Shape Analysis." In: *IEEE Trans. on Information Forensics and Security* 11.7 (2016), pp. 1461–1475.

[50] C. Elkan. "Using the triangle inequality to accelerate k-means." In: *Proc. Int. Conf. on Machine Learning (ICML)*. 2003, pp. 147–153.

[51] J. Engelsma and A. K. Jain. "Generalizing fingerprint spoof detector: Learning a one-class classifier." In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2019, pp. 1–8.

[52] Serife Kucur Ergünay, Elie Khoury, Alexandros Lazaridis, and Sébastien Marcel. "On the vulnerability of speaker verification to realistic voice spoofing." In: *Proc. Intl. Conf. on Biometrics Theory, Applications and Systems (BTAS)*. 2015, pp. 1–6.

[53] S. E. Fahlman, G. E. Hinton, and T. S. Sejnowski. "Massively parallel architectures for Al: NETL, Thistle, and Boltzmann machines." In: *Proc. National Conf. on Artificial Intelligence (AAAI)*. 1983.

[54] M. Fang, F. Boutros, A. Kuijper, and N. Damer. "Partial Attack Supervision and Regional Weighted Inference for Masked Face Presentation Attack Detection." In: *Proc. Intl Conf. on Automatic Face and Gesture Recognition (FG)*. IEEE. 2021, pp. 1–8.

[55] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper. "Real masks and spoof faces: On the masked face presentation attack detection." In: *Pattern recognition* 123 (2022), p. 108398.

[56] S. Fatemifar, M. Awais, S. Arashloo, and J. Kittler. "Combining multiple one-class classifiers for anomaly based face spoofing attack detection." In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2019.

[57] Finextra. *Worldline takes pay-by-face system on the road*. https://www.finextra.com/newsarticle/35009/worldline-takes-pay-by-face-system-on-the-road/. Last accessed: January 31, 2023. 2020.

[58] Forbes. *Hacker Clones Fingerprint From Politician's Photograph*. https://www.forbes.com/sites/paulmonckton/2014/12/30/hacker-clones-fingerprint-from-photograph/. Last accessed: January 31, 2023. 2014.

[59] T. de Freitas Pereira, A. Anjos, J. De Martino, and S. Marcel. "Can face anti-spoofing countermeasures work in a real world scenario?" In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2013, pp. 1–8.

[60] R. Gajawada, A. Popli, T. Chugh, A. Namboodiri, and A. Jain. "Universal material translator: Towards spoof fingerprint generalization." In: *Proc. Intl. Conf. on Biometrics (ICB)*. IEEE. 2019, pp. 1–8.

[61] J. Galbally, S. Marcel, and J. Fierrez. "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition." In: *IEEE Trans. on Image Processing* 23.2 (2013), pp. 710–724.

[62] J. Galbally, S. Marcel, and J. Fierrez. "Biometric Antispoofing Methods: A Survey in Face Recognition." In: *IEEE Access* 2 (2014), pp. 1530–1552. ISSN: 2169-3536.

[63] J. Gan, S. Li, Z. Zhai, and C. Liu. "3D convolutional neural network based on face anti-spoofing." In: *Proc. Intl. Conf. on Multimedia and Image Processing (ICMIP)*. 2017, pp. 1–5.

[64] Y. Ganin and V. Lempitsky. "Unsupervised domain adaptation by backpropagation." In: *Proc. Intl. Conf. on Machine Learning*. 2015, pp. 1180–1189.

[65] A. George and S. Marcel. "Deep pixel-wise binary supervision for face presentation attack detection." In: *Proc. Intl. Conf. on Biometrics (ICB)*. IEEE. 2019, pp. 1–8.

[66] A. George and S. Marcel. "Learning One Class Representations for Face Presentation Attack Detection using Multi-channel Convolutional Neural Networks." In: *IEEE Trans. on Information Forensics and Security* 16 (2020), pp. 361–375.

[67] A. George and S. Marcel. "On the effectiveness of vision transformers for zero-shot face anti-spoofing." In: *Proc. Intl. Joint Conf. on Biometrics (IJCB)*. IEEE. 2021, pp. 1–8.

[68] L. Ghiani, D. Yambay, V. Mura, S. Tocco, G. L. Marcialis, et al. "LivDet 2013 Fingerprint Liveness Detection Competition 2013." In: *Proc. Intl. Conf. on Biometrics (ICB)*. IEEE. 2013, pp. 1–6.

[69] L. Ghiani, D. A. Yambay, V. Mura, G. L. Marcialis, F. Roli, and S. Schuckers. "Review of the Fingerprint Liveness Detection (LivDet) Competition Series: 2009 to 2015." In: *Image and Vision Computing* 58 (2017), pp. 110–128.

[70] L. J. Gonzalez-Soler, M. Gomez-Barrero, and C. Busch. "Evaluating the Sensitivity of Face Presentation Attack Detection Techniques to Images of Varying Resolutions." In: *Norwegian Information Security Conf. (NISK)*. 2020.

[71] L. J. Gonzalez-Soler, M. Gomez-Barrero, and C. Busch. "Fisher Vector Encoding of Dense-BSIF Features for Unknown Face Presentation Attack Detection." In: *Proc. Intl. Conf. of the Special Interest Group on Biometrics (BIOSIG 2020)*. LNI. GI, 2020, pp. 1–11.

[72] L. J. Gonzalez-Soler, M. Gomez-Barrero, and C. Busch. "On the Generalisation capabilities of Fisher Vector based Face Presentation Attack Detection." In: *IET Biometrics* 10.5 (2021), pp. 480–496.

[73] L. J. Gonzalez-Soler, M. Gomez-Barrero, and C. Busch. "Towards Generalisable Facial Presentation Attack Detection by analysing Facial Regions." In: *Trans. on Biometrics, Behavior, and Identity Science (TBIOM)* (2022).

[74] L. J. Gonzalez-Soler, L. Chang, J. Hernández-Palancar, A. Pérez-Suárez, and M. Gomez-Barrero. "Fingerprint presentation attack detection method based on a bag-of-words approach." In: *Proc. Iberoamerican Congress on Pattern Recognition (CIARP)*. Springer. 2017, pp. 263–271.

[75] L. J. Gonzalez-Soler, M. Gomez-Barrero, L. Chang, A. Perez-Suarez, and C. Busch. "On the Impact of Different Fabrication Materials on Fingerprint Presentation Attack Detection." In: *Proc. Int. Conf. on Biometrics (ICB)*. 2019.

[76] L. J. Gonzalez-Soler, J. Patino, M. Gomez-Barrero, M. Todisco, C. Busch, and N. Evans. "Texture-based Presentation Attack Detection for Automatic Speaker Verification." In: *Proc. Intl. Workshop on Information Forensics and Security (WIFS)*. 2020, pp. 1–6.

[77] L. J. Gonzalez-Soler, M. Gomez-Barrero, L. Chang, A. Perez-Suarez, and C. Busch. "Fingerprint Presentation Attack Detection Based on Local Features Encoding for Unknown Attacks." In: *IEEE Access* 9 (2021), pp. 5806–5820. DOI: 10.1109/ACCESS.2020.3048756.

[78] L. J. Gonzalez-Soler, M. Gomez-Barrero, J. Kolberg, L. Chang, A. Perez-Suarez, and C. Busch. "Local Feature Encoding for Unknown Presentation Attack Detection: An Analysis of Different Local Feature Descriptors." In: *IET Biometrics* 10.4 (2021), pp. 374–391.

[79] L. J. Gonzalez-Soler, M. Gomez-Barrero, M. Kamble, M. Todisco, and C. Busch. "Dual-Stream Temporal Convolutional Neural Network for Voice Presentation Attack Detection." In: *Proc. Int. Workshop on Biometrics and Forensics (IWBF)*. 2022.

[80] L. J. Gonzalez-Soler, M. Gomez-Barrero, J. Patino, M Kamble, M. Todisco, and C. Busch. "Fisher Vectors for Biometric Presentation Attack Detection." In: *Handbook of Biometric Antispoofing*. Springer, 2022.

[81] M. Gonzalez-Ulloa. "Restoration of the face covering by means of selected skin in regional aesthetic units." In: *British Journal of Plastic Surgery* 9 (1956), pp. 212–221.

[82] I. Goodfellow, Y. Bengio, and A Courville. *Deep learning*. MIT press, 2016.

[83] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets." In: *Advances in Neural Information Processing Systems*. 2014, pp. 2672–2680.

[84] J. Guo, X. Zhu, J. Xiao, Z. Lei, G. Wan, and S. Z. Li. "Improving Face Anti-Spoofing by 3D Virtual Synthesis." In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2019.

[85] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[86] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[87] G. E. Hinton et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97.

[88] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." In: *arXiv preprint arXiv:1704.04861* (2017).

[89] Ch. Hsu, Ch. Chang, Ch. Lin, et al. *A practical guide to support vector classification*. Tech. rep. Taipei, Taiwan: National Taiwan University, 2003.

[90] L. Hu, M. Kan, S. Shan, and X. Chen. "Duplex generative adversarial network for unsupervised domain adaptation." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition*. 2018, pp. 1498–1507.

[91] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely connected convolutional networks." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4700–4708.

[92] K. Huang, A. Hussain, W. Wang, and R. Zhang. *Deep learning: fundamentals, theory and applications*. Vol. 2. 2019.

[93] X. Huang and L. Deng. *An Overview of Modern Speech Recognition*. 2010, pp. 339–366.

[94] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*. International Organization for Standardization. 2006.

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

[96] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 2382-37:2022 Information Technology - Vocabulary - Part 37: Biometrics*. International Organization for Standardization. 2017.

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

[98] A. Jain, P. Flynn, and A. Ross. *Handbook of Biometrics*. Springer, 2007.

[99] H. K. Jee, S. U. Jung, and J. H. Yoo. "Liveness detection for embedded face recognition system." In: *Proc. Intl. Journal of Biological and Medical Sciences* 1.4 (2006), pp. 235–238.

[100] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid. "Aggregating local image descriptors into compact codes." In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)* 34.9 (2012), pp. 1704–1716.

[101] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. "Caffe: Convolutional architecture for fast feature embedding." In: *Proc. Intl. Conf. on Multimedia*. 2014, pp. 675–678.

[102] C. Jin, H. Kim, and S. Elliott. "Liveness detection of fingerprint based on band-selective Fourier spectrum." In: *Proc. Intl. Conf. on Information Security and Cryptology (ICISC)*. Springer, 2007, pp. 168–179.

[103] C. Jin, S. Li, H. Kim, and E. Park. "Fingerprint liveness detection based on multiple image quality features." In: *Proc. Int. Conf. on Information Security Applications (ISA)*. Springer, 2011, pp. 281–291.

[104] J. Kannala and E. Rahtu. "BSIF: Binarized statistical image features." In: *Proc. Intl. Conf. on Pattern Recognition (ICPR)*. 2012, pp. 1363–1366.

[105] A. Kantarcı, H. Dertli, and H. Ekenel. "Shuffled patch-wise supervision for presentation attack detection." In: *Proc. Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*. 2021, pp. 1–5.

[106] D. E. King. "Dlib-ml: A machine learning toolkit." In: *The Journal of Machine Learning Research* 10 (2009), pp. 1755–1758.

[107] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization." In: *arXiv preprint arXiv:1412.6980* (2014).

[108] D. P. Kingma and M. Welling. "Auto-encoding variational bayes." In: *arXiv preprint arXiv:1312.6114* (2013).

[109] T. Kinnunen and H. Li. "An overview of text-independent speaker recognition: From features to supervectors." In: *Speech Communication* 52.1 (2010), pp. 12–40.

[110] T. Kinnunen et al. "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research." In: *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 5395–5399.

[111] T. Kinnunen et al. "Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals." In: *IEEE Trans. on Audio, Speech, and Language Processing* (2020).

[112] D. Klein and C. D. Manning. "Fast exact inference with a factored model for natural language parsing." In: *Advances in Neural Information Processing Systems*. 2003, pp. 3–10.

[113] J. Kolberg, D. Gläsner, R. Breithaupt, M. Gomez-Barrero, J. Reinhold, A. von Twickel, and C. Busch. "On the Effectiveness of Impedance-Based Fingerprint Presentation Attack Detection." In: *Sensors* 21.17 (2021).

[114] K. Kollreider, Fronthaler, and J. Bigun. "Verifying liveness by multiple experts in face biometrics." In: *Proc. Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*. 2008, pp. 1–6.

[115] K. Kollreider, H. Fronthaler, and J. Bigun. "Non-intrusive liveness detection by face images." In: *Image and Vision Computing* 27.3 (2009), pp. 233–244.

[116] J. Kong, J. Kim, and J. Bae. "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis." In: *ArXiv* (2020).

[117] N. Kose and J. Dugelay. "Reflectance analysis based countermeasure technique to detect face mask attacks." In: *Proc. Intl. Conf. on Digital Signal Processing (DSP)*. 2013, pp. 1–6.

[118] K. Kotwal, S.Bhattacharjee, P. Abbet, Z. Mostaani, W. Wei, X. Wenkang, Z. Yaxi, and S. Marcel. "Domain-Specific Adaptation of CNN for Detecting Face Presentation Attacks in NIR." In: *IEEE Trans. on Biometrics, Behavior, and Identity Science* (2022).

[119] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in Neural Information Processing Systems*. 2012, pp. 1097–1105.

[120] A. Krizhevsky, I. Sutskever, and G. Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in Neural Information Processing Systems (NeurIPS)* 25 (2012).

[121] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville. "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis." In: (2019). arXiv: 1910.06711 [eess.AS].

[122] P. D. Lapsley, J. A. Lee, Pare D. Ferrin, and N. Hoffman. *Anti-fraud biometric scanner that accurately detects blood flow*. US Patent 5,737,439. 1998.

[123] Laura Stampler. *Here's What Faces Would Look Like If They Were Perfectly Symmetrical*. https://time.com/2848303/heres-what-faces-would-look-like-if-they-were-perfectly-symmetrical/. Last accessed: January 31, 2023. 2014.

[124] H. Lee, H. Maeng, and Y. Bae. "Fake finger detection using the fractional Fourier transform." In: *Biometric ID Management and Multimodal Communication*. Springer, 2009, pp. 318–324.

[125] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot. "Unsupervised domain adaptation for face anti-spoofing." In: *IEEE Trans. on Information Forensics and Security* 13.7 (2018), pp. 1794–1809.

[126] J. Li, Y. Wang, T. Tan, and A. K. Jain. "Live face detection based on the analysis of fourier spectra." In: *Biometric technology for human identification*. Vol. 5404. International Society for Optics and Photonics. 2004, pp. 296–303.

[127] Y. Liu, A. Jourabloo, and X. Liu. "Learning deep models for face anti-spoofing: Binary or auxiliary supervision." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 389–398.

[128] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. "Deep Tree Learning for Zero-shot Face Anti-Spoofing." In: *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4680–4689.

[129] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. "Text classification using string kernels." In: *Journal of Machine Learning Research* 2.Feb (2002), pp. 419–444.

[130] M. Long, H. Zhu, J. Wang, and M. I. Jordan. "Unsupervised domain adaptation with residual transfer networks." In: *Advances in Neural Information Processing Systems*. 2016, pp. 136–144.

[131] G. D. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints." In: *Intl. Journal of Computer Vision* 60 (2 2004), pp. 91–110. ISSN: 0920-5691.

[132] R. Lowry. *Concepts and applications of inferential statistics*. Last accessed: January 31, 2023. 2014. URL: http://vassarstats.net/textbook/.

[133] Y. Ma, Z. Ren, and S. Xu. "RW-Resnet: A Novel Speech Anti-Spoofing Model Using Raw Waveform." In: *arXiv preprint arXiv:2108.05684* (2021).

[134] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. 1st. Springer-Verlag, 2009.

[135] E. Marasco and A. Ross. "A Survey on Antispoofing Schemes for Fingerprint Recognition Systems." In: *ACM Computing Surveys (CSUR)* 47.2 (2015), p. 28.

[136] Y. Martínez-Díaz, L. Chang, N. Hernández, H. Méndez-Vázquez, and L. H. Sucar. "Efficient video face recognition by using fisher vector encoding of binary features." In: *Proc. Intl. Conf. on Pattern Recognition (ICPR)*. 2016, pp. 1436–1441.

[137] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino. "Impact of artificial gummy fingers on fingerprint systems." In: *Electronic Imaging 2002*. International Society for Optics and Photonics. 2002, pp. 275–289.

[138] G. J. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[139] U. Muhammad and A. Hadid. "Face Anti-spoofing using Hybrid Residual Learning Framework." In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2019.

[140] V. Mura, L. Ghiani, G. L. Marcialis, F. Roli, D. A. Yambay, and S. Schuckers. "LivDet 2015 Fingerprint Liveness Detection Competition 2015." In: *Proc. Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE. 2015, pp. 1–6.

[141] V. Mura, G. Orrù, R. Casula, A. Sibiriu, G. Loi, P. Tuveri, L. Ghiani, and G. L. Marcialis. "LivDet 2017 fingerprint liveness detection competition 2017." In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2018, pp. 297–302.

[142] H. P. Nguyen, A. Delahaies, F. Retraint, and F. Morain-Nicolier. "Face presentation attack detection based on a statistical model of image noise." In: *IEEE Access* 7 (2019), pp. 175429–175442.

[143] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel. "On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing." In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2018, pp. 75–81.

[144] R. F. Nogueira, R. de Alencar Lotufo, and R. C. Machado. "Fingerprint Liveness Detection Using Convolutional Neural Networks." In: *IEEE Trans. on Information Forensics and Security* 11.6 (2016), pp. 1206–1213.

[145] T. Ojala, M. Pietikainen, and T. Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)* 24.7 (2002), pp. 971–987. ISSN: 0162-8828.

[146] V. Ojansivu and J. Heikkilä. "Blur insensitive texture classification using local phase quantization." In: *Proc. Intl. Conf. on Image and Signal Processing (ISP)*. 2008, pp. 236–243.

[147] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-time Signal Processing (2Nd Ed.)* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999. ISBN: 0-13-754920-2.

[148] G. Orrù, R. Casula, P. Tuveri, C. Bazzoni, G. Dessalvi, M. Micheletto, L. Ghiani, and G. L. Marcialis. "Livdet in action-fingerprint liveness detection competition 2019." In: *Proc. Intl. Conf. on Biometrics (ICB)*. IEEE. 2019, pp. 1–6.

[149] D. Osorio-Roig, P. Drozdowski, C. Rathgeb, A. Morales-González, E. Garea-Llano, and C. Busch. "Iris Recognition in Visible Wavelength: Impact and automated Detection of Glasses." In: *Proc. Intl. Workshop on Ubiquitous implicit BIOmetrics and health signals monitoring for person-centric applications (UBIO)*. IEEE, 2018, pp. 542–546.

[150] A. Oussidi and A. Elhassouny. "Deep generative models: Survey." In: *Proc. Intl. Conf. on Intelligent Systems and Computer Vision (ISCV)*. 2018, pp. 1–8.

[151] F. Pala and B. Bhanu. "Deep Triplet Embedding Representations for Liveness Detection." In: *Deep Learning for Biometrics*. 2017, pp. 287–307.

[152] G. Pan, L. Sun, Z. Wu, and S. Lao. "Eyeblink-based anti-spoofing in face recognition from a generic webcamera." In: *Proc. Intl. Conf. on Computer Vision*. 2007, pp. 1–8.

[153] E. Park, X. Cui, T. H. Nguyen, and H. Kim. "Presentation attack detection using a tiny fully convolutional network." In: *IEEE Trans. on Information Forensics and Security* 14.11 (2019), pp. 3016–3025.

[154] O. Parkhi, A. Vedaldi, and A. Zisserman. "Deep face recognition." In: *Proc. British Machine Vision Conf. (BMVC)*. British Machine Vision Association, 2015.

[155] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. "Automatic differentiation in pytorch." In: *Proc. NIPS-W*. 2017.

[156] K. Patel, H. Han, and A. K. Jain. "Cross-database face anti-spoofing with robust feature representation." In: *Proc. Chinese Conf. on Biometric Recognition*. 2016, pp. 611–619.

[157] F. Peng, L. Qin, and M. Long. "Face presentation attack detection based on chromatic Co-occurrence of local binary pattern and ensemble learning." In: *Journal of Visual Communication and Image Representation* 66 (2020), p. 102746.

[158] T. De Freitas Pereira. *Learning How To Recognize Faces In Heterogeneous Environments*. Tech. rep. EPFL, 2019.

[159] F. Perronnin, J. Sánchez, and T. Mensink. "Improving the fisher kernel for large-scale image classification." In: *Proc. Intl. European Conf. on Computer Vision (ECCV)*. 2010, pp. 143–156.

[160] S. Pertuz, D. Puig, and M. A. Garcia. "Analysis of focus measure operators for shape-from-focus." In: *Pattern Recognition* 46.5 (2013), pp. 1415–1432.

[161] R. Plesh, K. Bahmani, G. Jang, D. Yambay, K. Brownlee, T. Swyka, P. Johnson, A. Ross, and S. Schuckers. "Fingerprint Presentation Attack Detection utilizing Time-Series, Color Fingerprint Captures." In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2019, pp. 1–8.

[162] L. Po, Y. Li, F. Yuan, and L. Feng. "Face liveness detection using shearlet-based feature descriptors." In: *Electronic Imaging* 25.4 (2016).

[163] X. Qu, J. Dong, and S. Niu. "shallowCNN-LE: A shallow CNN with Laplacian Embedding for face anti-spoofing." In: *Intl. Conf. on Automatic Face & Gesture Recognition*. 2019, pp. 1–8.

[164] R. Raghavendra and C. Busch. "Presentation attack detection algorithm for face and iris biometrics." In: *Proc. European Signal Processing Conf. (EUSIPCO)*. 2014, pp. 1387–1391.

[165] R. Raghavendra, S. Venkatesh, K. Raja, P. Wasnik, M. Stokkenes, and C. Busch. "Fusion of multi-scale local phase quantization features for face presentation attack detection." In: *Proc. Intl. Conf. on Information Fusion (FUSION)*. 2018, pp. 2107–2112.

[166] R. Ramachandra, S. Venkatesh, K. Raja, S. Bhattacharjee, P. Wasnik, S. Marcel, and C. Busch. "Custom silicone face masks: Vulnerability of commercial face recognition systems & presentation attack detection." In: *Proc. Intl. Workshop on Biometrics and Forensics (IWBF)*. 2019, pp. 1–6.

[167] N. K. Ratha, J. H. Connell, and R. M. Bolle. "Enhancing security and privacy in biometrics-based authentication systems." In: *IBM systems Journal* 40.3 (2001), pp. 614–634.

[168] C. Rathgeb, P. Drozdowski, and C. Busch. "Makeup presentation attacks: Review and detection performance benchmark." In: *IEEE Access* 8 (2020), pp. 224958–224973.

[169] Ajita Rattani, Walter J Scheirer, and Arun Ross. "Open set fingerprint spoof detection across novel fabrication materials." In: *IEEE Trans. on Information Forensics and Security* 10.11 (2015), pp. 2447–2460.

[170] M. Ravanelli and Y. Bengio. "Speaker Recognition from Raw Waveform with SincNet." In: *Proc. IEEE Spoken Language Technology Workshop (SLT)*. 2018, pp. 1021–1028.

[171] Y. Rehman, L. Po, and M. Liu. "Deep Learning for face anti-spoofing: an end-to-end approach." In: *Proc. Intl. Conf. on Signal Processing: Algorithms, Architectures, Arrangements, and Applications*. 2017, pp. 195–200.

[172] E. Rosten and T. Drummond. "Machine learning for high-speed corner detection." In: *Proc. European Conf. on Computer Vision (ECCV)*. 2006, pp. 430–443.

[173] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. "ORB: An efficient alternative to SIFT or SURF." In: *Proc. Intl. Conf. on Computer Vision (ICCV)*. 2011, pp. 2564–2571.

[174] M. Sahidullah, T. Kinnunen, and C. Hanilçi. "A comparison of features for synthetic speech detection." In: *Proc. Interspeech*. 2015, pp. 2087–2091.

[175] Md. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, and K. Lee. "Introduction to voice presentation attack detection and recent advances." In: *Handbook of Biometric Anti-Spoofing*. Springer, 2019, pp. 321–361.

[176] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. "Image classification with the fisher vector: Theory and practice." In: *Proc. Intl. Journal on Computer VIsion* 105.3 (2013), pp. 222–245.

[177] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 4510–4520.

[178] S. Schuckers. "Presentations and attacks, and spoofs, oh my." In: *Image and Vision Computing* 55 (2016), pp. 26–30.

[179] A. F. Sequeira, T. Gonçalves, W. Silva, J. R. Pinto, and J. S. Cardoso. "An exploratory study of interpretability for face presentation attack detection." In: *IET Biometrics* 10.4 (2021), pp. 441–455.

[180] L. Al Shalabi, Z. Shaaban, and B. Kasasbeh. "Data mining: A preprocessing engine." In: *Journal of Computer Science* 2.9 (2006), pp. 735–739.

[181] T. Shen, Y. Huang, and Z. Tong. "Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing." In: *Proc. Intl. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 0–0.

[182] C. Silva, T. Bouwmans, and C. Frelicot. "An eXtended Center-Symmetric Local Binary Pattern for Background Modeling and Subtraction in Videos." In: *Proc. Intl. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*. Berlin, Germany, 2015.

[183] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. "Fisher Vector Faces in the Wild." In: *British Machine Vision Conf. (BMVC)*. Vol. 2. 3. 2013, p. 4.

[184] M. Singh, M. Chawla, R. Singh, M. Vatsa, and R. Chellappa. "Disguised faces in the wild 2019." In: *Proc. Intl. Conf. on Computer Vision Workshops*. 2019, pp. 0–0.

[185] X. Song, X. Zhao, L. Fang, and T. Lin. "Discriminative representation combinations for accurate face spoofing detection." In: *Pattern Recognition* 85 (2019), pp. 220–231.

[186] C. Sousedik, R. Breithaupt, and C. Busch. "Volumetric fingerprint data analysis using Optical Coherence Tomography." In: *Proc. Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*. IEEE Computer Society, 2013, pp. 1–6.

[187] C. Sousedik and C. Busch. "Quality of Fingerprint Scans captured using Optical Coherence Tomography." In: *Proc. Intl. Joint Conf. on Biometrics (IJCB)*. IEEE Computer Society, 2014.

[188] K. Srinivas, R. K. Das, and H. A. Patil. "Combining Phase-based Features for Replay Spoof Detection System." In: *Proc. Intl. Symposium on Chinese Spoken Language Processing (ISCSLP)*. 2018, pp. 151–155.

[189] L. Sun, W. Huang, and M. Wu. "TIR/VIS correlation for liveness detection in face recognition." In: *Intl. Conf. on Computer Analysis of Images and Patterns*. Springer. 2011, pp. 114–121.

[190] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the inception architecture for computer vision." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition*. 2016, pp. 2818–2826.

[191] E. Tabassi and NIST. *Development of NFIQ 2.0*. https://www.nist.gov/services-resources/software/development-nfiq-20. 2015 (accessed August 24, 2020).

[192] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco. "An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification." In: *Proc. Odyssey*. 2020.

[193] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco. "Spoofing Attack Detection using the Non-linear Fusion of Sub-band Classifiers." In: *Proc. Interspeech*. 2020.

[194] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco. "Spoofing attack detection using the non-linear fusion of sub-band classifiers." In: *Proc. Interspeech*. 2020.

[195] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher. "End-to-end anti-spoofing with RawNet2." In: *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 6369–6373.

[196] B. Tan and S. Schuckers. "New approach for liveness detection in fingerprint scanners based on valley noise analysis." In: *Journal of Electronic Imaging* 17.1 (2008), pp. 011009–011009.

[197] M. Tan and Q. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In: *Proc. Intl. Conf. on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[198] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. Le. "Mnasnet: Platform-aware neural architecture search for mobile." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2820–2828.

[199] Santosh Tirunagari, Norman Poh, David Windridge, Aamo Iorliam, Nik Suki, and Anthony TS Ho. "Detection of face spoofing using visual dynamics." In: *IEEE Trans. on Information Forensics and Security* 10.4 (2015), pp. 762–777.

[200] M. Todisco, H. Delgado, and N. Evans. "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients." In: *Proc. Odyssey volume=25*. 2016, pp. 249–252.

[201] M. Todisco, H. Delgado, and N. Evans. "Articulation rate filtering of CQCC features for automatic speaker verification." In: *Proc. Interspeech*. 2016.

[202] M. Todisco, H. Delgado, and N. Evans. "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification." In: *Computer Speech & Lang.* 45 (2017), pp. 516–535.

[203] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee. "Asvspoof 2019: Future horizons in spoofed and fake audio detection." In: *Proc. Interspeech* (2019).

[204] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. "Deepfakes and beyond: A survey of face manipulation and fake detection." In: *Information Fusion* 64 (2020), pp. 131–148.

[205] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva. "STC Antispoofing Systems for the ASVspoof2021 Challenge." In: *Proc. of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*. 2021, pp. 61–67.

[206] Y. Uchida, S. Sakazawa, and S. Satoh. "Image retrieval with fisher vectors of binary features." In: *ITE Trans. on Media Technology and Applications* 4.4 (2016), pp. 326–336.

[207] A. Vedaldi and B. Fulkerson. *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. http://www.vlfeat.org/. 2008.

[208] A. Vedaldi and A. Zisserman. "Efficient Additive Kernels via Explicit Feature Maps." In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)* 34.3 (2011).

[209] G. Wang, H. Han, S. Shan, and X. Chen. "Improving Cross-database Face Presentation Attack Detection via Adversarial Domain Adaptation." In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2019.

[210] G. Wang, H. Han, S. Shan, and X. Chen. "Cross-domain Face Presentation Attack Detection via Multi-domain Disentangled Representation Learning." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition*. 2020, pp. 6678–6687.

[211] G. Wang, H. Han, S. Shan, and X. Chen. "Unsupervised Adversarial Domain Adaptation for Cross-Domain Face Presentation Attack Detection." In: *IEEE Trans. on Information Forensics and Security* 16 (2020), pp. 56–69.

[212] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu. "Cross-Domain Replay Spoofing Attack Detection Using Domain Adversarial Training." In: *Proc. Interspeech*. 2019, pp. 2938–2942.

[213] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu. "Dual-Adversarial Domain Adaptation for Generalized Replay Attack Detection." In: *Proc. Interspeech 2020*. 2020, pp. 1086–1090. DOI: 10.21437/Interspeech.2020-1255.

[214] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, and K. Lee. "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech." In: *Computer Speech & Language* 64 (2020), pp. 101–114.

[215] Y. Wang, X. Hao, Y. Hou, and C. Guo. "A new multispectral method for face liveness detection." In: *Proc. Asian Conf. on Pattern Recognition*. 2013, pp. 922–926.

[216] Y. Wang, W. Cai, T. Gu, W. Shao, Y. Li, and Y. Yu. "Secure Your Voice: An Oral Airflow-Based Continuous Liveness Detection for Voice Assistants." In: *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.4 (2019).

[217] Z. Wang, Z. Wang, Z. Yu, W. Deng, J. Li, S. Li, and Z. Wang. "Domain Generalization via Shuffled Style Assembly for Face Anti-Spoofing." In: *arXiv preprint arXiv:2203.05340* (2022).

[218] Z. Wang, Q. Wang, W. Deng, and G. Guo. "Learning multi-granularity temporal characteristics for face anti-spoofing." In: *IEEE Trans. on Information Forensics and Security (TIFS)* (2022).

[219] D. Wen, H. Han, and A. K. Jain. "Face spoof detection with image distortion analysis." In: *IEEE Trans. on Information Forensics and Security* 10.4 (2015), pp. 746–761.

[220] D. Willis and M. Lee. "Six biometric devices point the finger at security." In: *Computers & Security* 5.17 (1998), pp. 410–411.

[221] F. Xiong and W. AbdAlmageed. "Unknown presentation attack detection with face rgb images." In: *Proc. Intl. Conf. on Biometrics Theory, Applications and Systems (BTAS)*. 2018, pp. 1–9.

[222] Z. Xu, S. Li, and W. Deng. "Learning temporal features using LSTM-CNN architecture for face anti-spoofing." In: *Proc. Asian Conf. on Pattern Recognition (ACPR)*. 2015, pp. 141–145.

[223] J. Yamagishi et al. "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection." In: *Proc. of the Automatic Speaker Verification and Spoofing Countermeasures Challenge (ASVSpoof2021)*. 2021, pp. 47–54.

[224] J. Yamagishi, X. Wang, M. Todisco, Md. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, et al. "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection." In: *arXiv preprint arXiv:2109.00537* (2021).

[225] D. Yambay, L. Ghiani, P. Denti, G. L. Marcialis, F. Roli, and S. Schuckers. "LivDet 2011 Fingerprint Liveness Detection Competition 2011." In: *Proc. Intl. Conf. on Biometrics (ICB)*. IEEE. 2012, pp. 208–215.

[226] J. Yang, Z. Lei, and S. Li. "Learn convolutional neural network for face anti-spoofing." In: *arXiv preprint arXiv:1408.5601* (2014).

[227] J. Yang, Z. Lei, D. Yi, and S. Z. Li. "Person-specific face antispoofing with subject domain adaptation." In: *IEEE Trans. on Information Forensics and Security* 10.4 (2015), pp. 797–809.

[228] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu. "Face anti-spoofing: Model matters, so does data." In: *Proc. Intl. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 3507–3516.

[229] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao. "Searching central difference convolutional networks for face anti-spoofing." In: *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5295–5305.

[230] L. Zhang and J. Yang. "A Continuous Liveness Detection for Voice Authentication on Smart Devices." In: *arXiv preprint arXiv:2106.00859* (2021).

[231] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li. "Face detection based on multi-block lbp representation." In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2007, pp. 11–18.

[232] Y. Zhang, J. Fei, and Z. Duan. "One-Class Learning Towards Synthetic Voice Spoofing Detection." In: *IEEE Signal Processing Letters* 28 (2021), pp. 937–941.

[233] Y. Zhang, J. Tian, X. Chen, X. Yang, and P. Shi. "Fake finger detection based on thin-plate spline distortion model." In: *Advances in Biometrics*. Springer, 2007, pp. 742–749.

[234] Y. Zhang, S. Pan, X. Zhan, Z. Li, M. Gao, and C. Gao. "FLDNet: Light Dense CNN for Fingerprint Liveness Detection." In: *IEEE Access* 8 (2020), pp. 84141–84152.

[235] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. "A face antispoofing database with diverse attacks." In: *Proc. intl. Conf. on Biometrics (ICB)*. IEEE. 2012, pp. 26–31.

[236] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z Li. "A face antispoofing database with diverse attacks." In: *Proc. Intl. Conf. on Biometrics (ICB)*. 2012, pp. 26–31.

[237] A. Zwiesele, A. Munde, C. Busch, and H. Daum. "BioIS Study - Comparative Study of Biometric Identification Systems." In: *34th Annual 2000 IEEE Intl. Carnahan Conf. on Security Technology (CCST)*. IEEE Computer Society, 2000, pp. 60–63.

DECLARATION

Hiermit erkläre ich, dass ich die vorliegende Dissertation zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer.nat.) mit dem Titel

Generalisable Presentation Attack Detection For Multiple Types of Biometric Characteristics

selbständig und ohne fremde Hilfe und nur mit den mir zur Verfügung gestellten Hilfsmitteln verfasst habe. Alle wörtlich oder sinngemäß aus veröffentlichten Werken übernommenen Textpassagen und alle auf mündlichen Informationen beruhenden Aussagen sind als solche gekennzeichnet. Die Grundsätze der guten wissenschaftlichen Praxis wurden beachtet. Eine Promotion habe ich noch nicht angestrebt.

Darmstadt, 14 December 2022

Lázaro Janier González Soler

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".