

Doctoral thesis

Doctoral theses at NTNU, 2021:213

Ali Khodabakhsh

Automated Authentication of Audiovisual Contents: A Biometric Approach

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical
Engineering
Dept. of Information Security and
Communication Technology



Norwegian University of
Science and Technology

Ali Khodabakhsh

Automated Authentication of Audiovisual Contents: A Biometric Approach

Thesis for the Degree of Philosophiae Doctor

Gjøvik, June 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology



Norwegian University of
Science and Technology

NTNU

Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

© Ali Khodabakhsh

ISBN 978-82-326-6166-4 (printed ver.)
ISBN 978-82-326-6101-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2021:213

Printed by NTNU Grafisk senter

To my beloved parents and my sweet loving wife, for their sincere affection, encouragement, and support throughout this journey.

“Su axar, çuxurun tapar.”

“Water flows and carves its own path.”

(Azerbaijani proverb)

Declaration of Authorship

I, Ali Khodabakhsh, hereby declare that this thesis and the work presented in it are entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed:

(Ali Khodabakhsh)

Date: 25/05/2021

Abstract

Following the introduction of image manipulation tools such as Adobe Photoshop in the early 2000s, the public trust in image authenticity dropped and the need for the development and deployment of image authentication techniques became apparent. Recently, we face a similar situation for video content as photo-realistic video manipulation tools like Deepfake are becoming available and within the reach of the general public as well as bad actors. In human to human communication, face and voice modalities play a crucial role, and not surprisingly, the same modalities are most under attack by forgers.

Historically, the task of audiovisual content authentication was the focus of the field of multimedia forensics, with more than 15 years of accumulated literature. Following the increase in the popularity of biometric systems in practice, these systems have also faced similar challenges and felt the need for content authentication. Consequently, the field of presentation attack detection is born to protect biometric systems against fake biometric presentations. Due to the parallel nature of the presentation attack detection problem, defined as protecting a biometric system from presentation attacks, to the audiovisual content authentication problem, defined as protecting the viewer from fake content, the field of biometric presentation attack detection can provide a solid basis for approaching the multimedia authentication problem.

The primary objective of this thesis is to address the audiovisual content authentication problem on the face modality by vulnerability assessment and mitigation of detected vulnerabilities with reliance on biometric and presentation attack detection knowledge. To this end, after producing a taxonomy of existing generation techniques, subjective tests are done to assess the vulnerability of viewers to the most prevalent generation techniques with reliance on data collected from the wild. Following this process, the generation techniques the viewers are most susceptible to were identified. The discovered vulnerabilities are then mitigated individually by the introduction of effective detection techniques that outperform existing solutions. Furthermore, the vulnerability of existing general-purpose detection methods was analyzed and it was discovered that these methods show limited generalization capacity when faced with new generation methods. To mitigate this vulnerability, with reliance on an anomaly extraction approach, a generalizable detection method is introduced and empirically evaluated against the state-of-the-art methods. Additionally, all the datasets that are collected during the course of this thesis work are made publicly available to stimulate further research on this topic.

Acknowledgements

I would like to thank the department of information security and communication technology for funding my research on this exciting topic and providing me with an ideal work environment as well as exceptional opportunities for academic as well as personal growth. I would also like to thank the administration and IT teams for their timely and kind support, my colleagues for providing a friendly and supportive community, and the volunteers for data collection who were kind enough to participate in the subjective experiments. I would like to thank the members of the dissertation committee for taking time out of their busy schedules to review the thesis and providing their invaluable feedback. Lastly, I would like to thank my supervisor Christoph Busch and my co-supervisor Raghavendra Ramachandra who did not hesitate to offer me their selfless and unconditional support and guidance while they provided me with fantastic research equipment as well as academic collaboration and research stay opportunities throughout this period.

Contents

List of Tables	xvi
List of Figures	xix
List of Abbreviations	xxi
I Overview	1
1 Introduction	3
1.1 Motivation and Problem Description	4
1.2 Research Objectives	6
1.3 Research Methodology	9
1.4 List of Included Research Publications	12
1.5 Scope and Outline of the Thesis	13
2 Background and Related Work	15
2.1 Generation Techniques	15
2.1.1 Physical Attacks	15
2.1.2 Digital Attacks	17

2.1.3	Datasets	20
2.2	Detection Techniques	21
2.2.1	Presentation Attack Detection	21
2.2.2	Deepfake Detection	22
3	Summary of Published Articles and Contributions	25
3.1	Viewers' Vulnerabilities	25
3.1.1	Taxonomy	26
3.1.2	Subjective tests	26
3.1.3	Look-alike recognition	27
3.1.4	Inter-frame Forgery Detection	28
3.1.5	Contributions	29
3.2	Generalizability	30
3.2.1	Generalizability of existing methods	30
3.2.2	Generalizable Detector	31
3.2.3	Contributions	33
4	Conclusion and Future Work	35
4.1	Limitations	36
4.2	Future Work	37
4.2.1	Scalable Solutions	37
4.2.2	Diversity in Datasets	38
4.2.3	Robustness Against Adversarial Attacks	38
II	Viewers' Vulnerability	51
5	Article 1: A Taxonomy of Audiovisual Fake Multimedia Content Creation Technology	53

5.1	Abstract	53
5.2	Introduction	53
5.3	Personation Methods	55
5.3.1	Visual	56
5.3.2	Auditory	60
5.3.3	Combinations	62
5.4	Detection Techniques	62
5.5	Discussion	62
5.6	Future work	63
6	Article 2: Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content	67
6.1	Abstract	67
6.2	Introduction	68
6.3	Data and Methodology	69
6.3.1	Dataset	69
6.3.2	Protocol	71
6.3.3	Test Setup	73
6.3.4	Performance Evaluation	73
6.4	Results and Discussion	75
6.5	Conclusion and Future Work	78
6.6	Acknowledgement	80
7	Article 3: Action-Independent Generalized Behavioral Identity Descriptors for Look-alike Recognition in Videos	83
7.1	Abstract	83
7.2	Introduction	84
7.3	Proposed Method	86

7.3.1	Preprocessing	86
7.3.2	The proposed recognition system	87
7.3.3	Look-alike mining	88
7.4	Experiment Setup	89
7.4.1	1000LP Dataset	89
7.4.2	Detector	90
7.5	Results and Discussion	90
7.6	Conclusion	94
8	Article 4: Unit-Selection Based Facial Video Manipulation Detection	99
8.1	Abstract	99
8.2	Introduction	99
8.3	Methodology	102
8.3.1	Morph Cut Dataset	102
8.3.2	Morph-cut Detection	103
8.4	Experiment Setup	104
8.4.1	Morph Cut Dataset Details	104
8.4.2	Proposed Detector	105
8.4.3	Baseline Methods	105
8.5	Results and Discussion	106
8.6	Conclusion	109
III	Generalizability	113
9	Article 5: Fake face detection methods: Can they be generalized?	115
9.1	Abstract	115
9.2	Introduction	116

9.3	Fake Face in the Wild Dataset (FFW)	118
9.4	Fake Face Detection Techniques	120
9.5	Experimental Evaluation	120
9.5.1	Evaluation Metrics	120
9.5.2	Experimental Protocol	121
9.6	Results and Discussion	121
9.6.1	Performance on the Known Fake Face Attacks (TestSet-I)	121
9.6.2	Performance on the Unknown Fake Face Presentations (TestSet-II)	122
9.6.3	Performance on the FaceSwap/SwapMe Dataset (TestSet-III)	125
9.7	Conclusion and Future Work	125
10	Article 6: A Generalizable Deepfake Detector based on Neural Conditional Distribution Modelling	129
10.1	Abstract	129
10.2	Introduction	130
10.3	Methodology	131
10.3.1	Pixel RNN	131
10.3.2	Classification	132
10.3.3	Generalization Performance	132
10.4	Experiment Setup	132
10.5	Results and Discussion	134
10.5.1	Features	134
10.5.2	Known Synthetic Face Detection	135
10.5.3	Unknown Synthetic Face Detection	136
10.6	Conclusion	136
11	Article 7: Unknown Presentation Attack Detection against Rational	

Attackers	141
11.1 Abstract	141
11.2 Introduction	142
11.3 Literature Review	143
11.3.1 Anti Counter Forensics	143
11.3.2 Presentation Attack Detection	144
11.3.3 DeepFakes Detection	145
11.4 Theory	145
11.4.1 Rational Attacker	145
11.4.2 Multiple Attackers	147
11.4.3 Detection Strategy	148
11.4.4 Requirements	149
11.4.5 Generation-based Feature Sets	150
11.4.6 Minimax Objective Function	151
11.5 Proposed Method	151
11.5.1 Pixel-Level Probability Distribution Modelling	151
11.5.2 Dimensionality Reduction	152
11.5.3 Categorical Margin Maximization Loss	153
11.5.4 Unknown Attack Detection	155
11.6 Experiment Setup	156
11.6.1 Datasets	156
11.6.2 Parameters	157
11.6.3 Metrics	159
11.7 Presentation Attack Detection	160
11.7.1 Representation Adequacy	160
11.7.2 One-class classification	163

11.7.3	Detection Performance	163
11.7.4	Detection Cost	169
11.8	Deepfake Detection	169
11.9	Conclusion	172
11.10	Acknowledgment	173

List of Tables

5.1	Summary of personation techniques	63
5.2	Estimated detection difficulty	63
7.1	Performance of proposed methods	91
7.2	Performance compared to existing methods	92
8.1	Video count in proposed dataset	104
8.2	Video parameters	104
8.3	Network architecture	106
8.4	Detection accuracy compared to baseline	107
9.1	FFW dataset statistics	119
9.2	Accuracy of classifiers on testset I	122
9.3	Performance of classifiers on testset I	122
9.4	Performance of classifiers on testset II	123
9.5	CNN EER performances on subcategories	124
9.6	Performance of classifiers on testset III	125
10.1	Detection performance on known attacks compared to baseline	135

10.2	Detection performance on unknown attacks	137
11.1	Representative works on anti counter forensics, presentation attack detection and DeepFakes detection	146
11.2	Detection performance for each of the anomaly measures and their combination on the SiW-M dataset	164
11.3	Performance comparison on the task of known attack detection on the SiW-M dataset	164
11.4	Performance comparison on the task of unknown presentation at- tack detection on the SiW-M dataset	168
11.5	Performance of the detector in few-shot learning scenarios on the SiW-M dataset	170
11.6	Performance of the proposed detection methods for the protocol II task of OULU-NPU dataset	170
11.7	Performance of the proposed detection methods for the task of known attack detection on Deepfake detection task on the Face- Forensics++ dataset	171
11.8	Performance of the proposed detection methods for the task of un- known attack detection on Deepfake detection task on the Face- Forensics++ dataset	172

List of Figures

1.1	Research outline	9
2.1	Examples of generation techniques	16
2.2	Traditional video tampering categories	18
2.3	Video tampering spectrum	18
5.1	Points of vulnerability	56
5.2	Examples of personation techniques	57
6.1	The six categories of fake faces	71
6.2	Subjective test interface	74
6.3	Subjective detection percentages per category	76
6.4	Subjective detection percentages per video	77
6.5	Clue reliability statistics	77
6.6	Effect of familiarization, biometric reference, and knowledge of the target on detection performance	78
6.7	Detection percentage vs age	79
7.1	Look-alike pair examples	84

7.2	Feature extraction pipeline	87
7.3	Proposed network architecture	88
7.4	Subjective face recognition test user interface	89
7.5	DET curve for the proposed methods	91
7.6	t-SNE of embeddings for enrollment utterances	93
7.7	Visualization of facial landmark significance	93
8.1	An example of a morph-cut transition.	101
8.2	Training and evaluation pipelines	105
8.3	DET curve of proposed method in comparison to baseline	107
8.4	Prediction error image examples	108
8.5	Histogram of average prediction error per frame	108
9.1	Examples of fake faces	116
9.2	BRISQUE quality score distribution	119
9.3	Examples from FFW dataset	120
9.4	LBP-SVM score distribution on testsets I and II	124
9.5	Inceptionv3 score distribution on testsets I and II	124
10.1	Training and evaluation pipelines of the proposed method	133
10.2	Classifier architecture	134
10.3	Image log-likelihood histograms for pristine and synthetic data	135
10.4	Examples of log-likelihood matrices	136
11.1	The pipeline of the designed detection mechanism	152
11.2	The architecture of the classifier network	159
11.3	Example frames from BF and each PAS from the SiW-M dataset along with their corresponding log-likelihood matrices	161

11.4	Average and standard deviation of the log-likelihood matrices from the SiW-M dataset	161
11.5	The t-SNE graph on the average log-likelihood matrices for all the data available in the SiW-M dataset	162
11.6	Detection performance according to the starting PCA component .	165
11.7	Detection error trade-off curve for the one-class detector in PAD on the SiW-M dataset	166
11.8	Detection error trade-off curve of the discriminative detector for the known attack detection on PAD task on the SiW-M dataset . .	166
11.9	Average and standard deviation of the log-likelihood matrices in the FaceForencisc++ dataset	171
11.10	Average and standard deviation of the log-likelihood matrices in the OULU-NPU dataset	171

List of Abbreviations

1000LP	100 Look-alike Pairs
AAM	Active Appearance Models
ACER	Average Classification Error Rate
ADAM	Adaptive Moment Estimation
AI	Artificial Intelligence
APCER	Attack Presentation Classification Error Rate
AVTTS	AudioVisual Text-To-Speech
BF	Bona Fide
BFR	Behavioral Face Recognition
BPCER	Bona fide Presentation Classification Error Rate
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
CDNN	Convolutional Deep Neural Network
CF	Counter Forensics
CFG	Computer Generated Face Image
CG	Computer Generated
CGI	Computer Generated Imagery
CI	Confidence Interval
CIA	Central Intelligence Agency
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DARPA	Defense Advanced Research Project Agency
DET	Detection Error Trade-off
DF	DeepFake
DFDC	DeepFake Detection Challenge
DNN	Deep Neural Network
EER	Equal Error Rate
ELU	Exponential Linear Unit

F2F	Face2Face
FACS	Facial Action Coding System
FBI	Federal Bureau of Investigation
FFW	Fake Faces in the Wild
FPS	Frames Per Second
FS	FaceSwap
GAN	Generative Adversarial Network
GPU	Graphical Processing Unit
GMM	Gaussian Mixture Model
HMM	Hidden Markov Models
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
JPEG	Joint Photographic Experts Group
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LOO	Leave-One-Out
LSTM	Long Short Term Memory
MPA	Most Powerful Attack
NIST	National Institute of Standards and Technology
NT	Neural Textures
PAD	Presentation Attack Detection
PAS	Presentation Attack Species
PCA	Principal Component Analysis
PRNU	Photo Response Non-Uniformity
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics
SiW-M	Spoofing in the Wild Multiple Attacks
SSS	Statistical Speech Synthesis
SVM	Support Vector Machine
t-SNE	T-distributed stochastic neighbor embedding
UBM	Universal Background Model

Part I

Overview

Chapter 1

Introduction

The advent of deep learning-based generation techniques in recent years along with the reduction in the cost of computation resulted in the feasibility of low-cost photo-realistic video generation. The introduction of such methods as open-source tools on the internet provided an opportunity for bad actors to weaponize them for personal and political gain. Relying on the fact that the face is the main modality of human communication in daily life, methods that can realistically produce facial videos have an immense potential for abuse. The infamous Deepfake¹ is an example of such tools that was initially shared on Reddit and used for the purpose of generating fake pornography and later for fake news generation. These technologies have started to be seen as a big cyber threat against business, politics, identity, national security, and democracy to an extent that a bill was passed in the US senate² to report at specified intervals on the state of digital content forgery technology and some social media platforms announced that they would remove these content in the wake of the 2020 US elections³. Consequently, it is paramount to address the detection of newly introduced fake content to preserve trust in video content. Historically, the detection of fake content has been the focus of the field of multimedia forensics. However, the research community can benefit greatly from the accumulated knowledge in relevant fields of biometric presentation attack detection and machine learning.

Relying on the strong background of the Norwegian biometrics laboratory in the

¹<https://github.com/deepfakes/faceswap>

²<https://www.congress.gov/bill/116th-congress/senate-bill/2065>

³<https://www.reuters.com/article/us-facebook-deepfake/facebook-to-remove-deepfake-videos-in-run-up-to-2020-u-s-election-idUSKBN1Z60JV>

field of biometrics and presentation attack detection, the research for this thesis was initiated to utilize the knowledge towards the detection of fake audiovisual content. This thesis aims to investigate the vulnerabilities of the viewers and the existing detection techniques, and provides solutions for mitigation of the identified vulnerabilities.

1.1 Motivation and Problem Description

Methods for the generation of realistic image content have existed for decades and the media environment, as well as individuals, have adapted to the presence of these techniques. However, due to the complexity and cost of photo-realistic video generation, a video has been considered a reliable medium and valid evidence by society. Traditionally, realistic facial video manipulation has been challenging and required sophisticated editing tools, complex and time-consuming processes, and domain expertise. Early generation methods required a significant amount of data from a target individual to only modify lip motion. The advent of deep learning and the availability of low-cost computational resources has changed this situation and the quality of synthesized materials that become available. Advancements in data availability and the evolution of deep learning techniques resulted in new methods for automated photo-realistic video synthesis as well as manipulation of facial attributes and facial behavior. Open-source software such as Deepfake and FaceSwap⁴ and even mobile applications such as Reface⁵ have been released facilitating the generation of fake videos without the requirement of experience and expertise.

Many of the introduced generation methods are developed with the innocent purpose of improving realism in movie production and video games by the entertainment industry. These very same technologies, however, have been abused for blackmailing people and producing fake content to spread misinformation and manipulate public opinion. These technologies have shown great potential for causing significant damage to trust in society and fake videos depicting an individual have become a great public concern. According to the visual threat intelligence service Sensity, up until the writing of this thesis, more than 3,000 public figures were targeted using more than 80,000 fake videos⁶. Furthermore, the number of Deepfakes online is roughly doubling every six months showing exponential growth⁷.

Consequently, the detection of fake audiovisual content has received significant

⁴<https://github.com/MarekKowalski/FaceSwap>

⁵<https://reface.ai>

⁶<https://sensity.ai/>

⁷<https://sensity.ai/deepfake-threat-intelligence-a-statistics-snapshot-from-june-2020/>

attention from not only public institutes but also industry and big corporations. Governmental bodies, as well as the news industry, are becoming aware of the potential menace carried by these technologies. There is a growing interest in the detection of this content demonstrated through the increasing number of dedicated workshops in top conferences as well as international projects such as MediFor project⁸ funded by Defense Advanced Research Project Agency (DARPA) and competitions such as Media Forensics Challenge initiated by the National Institute of Standards and Technology (NIST) and the recent Deepfake Detection Challenge⁹ organized by Facebook.

Traditional detection methods developed in the field of media forensics relied on handcrafted methods using features such as in-camera fingerprints and out-camera fingerprints. These methods are highly dependent on the specific recording conditions and scenarios and are not robust against unseen conditions. Making the matters worse, when audiovisual content is shared on the internet, they are often automatically modified by the sharing platform via operations such as compression and resize, as well as meta-data removal, further reducing the effectiveness of methods reliant on acute artifacts. Despite the continuous research and the numerous tools that are developed by the forensics community in the past, the recent changes in the generation techniques and sharing environment challenge the existing forensic methods and demonstrates the need for further investment and development of new and timely detection mechanisms. Even though deep learning and low-cost computational resources provided the grounds for the development of the generation techniques, the same advancements can also provide an opportunity for the development of more effective detection techniques. Large efforts are being directed towards proposing new detection methods as well as improving the existing solution.

In the arms race between the generation and detection technologies, it is crucial to have a clear understanding of the vulnerabilities that exist on the detection side and mitigate them accordingly. In contrast to the presentation attack scenario, where the system to be protected was a biometric capture device of the biometric system, humans have an innate acute ability to detect fake audiovisual content based on the semantic and physical inconsistencies present in the video (55). This ability has historically made photo-realistic generation a difficult problem, especially with regards to the objects that humans are most familiar with such as the human face (56). Therefore, detection methods need to complement the detection abilities of humans rather than replacing them to be able to protect viewers from fake content. Subsequently, an understanding of which generation techniques the viewers are

⁸<https://www.darpa.mil/program/media-forensics>

⁹<https://www.kaggle.com/c/deepfake-detection-challenge>

most susceptible to is crucial. Another factor that limits the vulnerability surface is the technological feasibility and cost of generation techniques. Without a clear picture of which generation methods are widespread and what is possible with the existing technologies, the scope on which a proposed detection method may be effective can be limited, limiting its applicability in real life. Thus, this dissertation work primarily focuses on discovering and addressing the vulnerabilities of viewers and existing general-purpose detection solutions with reliance on a clear understanding of the attack surface.

Video content provides a rich collection of information that can be used for detection, including the visual, behavioral, and auditory modalities as well as the correspondences between them. However, utilizing all the information available simultaneously would require extensive investment in research in multiple directions and thus falls outside the limits of a Ph.D. thesis. Consequently, the strategy taken in this thesis is to prioritize and invest in the most promising modalities for detection. Despite the great developments in the field of speech synthesis, the progress has been much slower compared to the visual modality where dozens of new generation techniques are introduced every year. For example, tools such as Deepfake which are at the center of attention only modify the visual modality. On the other hand, realistic synthesis in the visual modality has been a much more challenging task due to the larger number of details that are required to be perfected. As a result, it can be argued that detection based on the visual modality would be the most promising direction for research as it both poses the major threat and provides the biggest opportunity for detection. As such, in this thesis, similar to the general trend in the community, the visual modality is focused on as the main modality for detection.

1.2 Research Objectives

The research objectives of the thesis are to discover the vulnerabilities of the viewers and existing detection methods and mitigate them. To achieve this, the goal of this thesis can be broken down into the following objectives:

1. A study on possible facial video generation methods with existing technologies needs to be conducted to serve as a basis for vulnerability assessment.
2. The vulnerabilities of the viewers need to be assessed against the most prevalent generation techniques.
3. New detection methods need to be proposed for the discovered vulnerabilities and their performance empirically proved in real-life scenarios.

4. The vulnerabilities of the existing detection methods need to be assessed via rigorous testing in real-life scenarios.
5. The discovered vulnerabilities need to be mitigated by the introduction of new detection methods and their evaluation in realistic scenarios.

Based on these research objectives, the following research questions are formulated:

RQ 1: What methods of photo-realistic facial video generation are feasible with the existing technology and which ones are the most difficult to detect by viewers? (Related chapters: 5, 6)

Following the introduction of Deepfakes, most of the existing literature has a focus on the generative adversarial network (GAN) based facial video generation techniques. However, GAN-based methods are not the only methods of generation, and there is a much wider range of generation methods evident from the various techniques used in the movie industry before the introduction of Deepfakes. Neglecting other generation techniques can result in the development of detection methods that have fundamental weaknesses against them. Thus, it is important to study possible methods of generation to have a complete picture of the threats the viewers may face in reality. Furthermore, as the viewers have a high sensitivity to artifacts on a face image, methods that have not yet reached a sufficient level of realism are easy to recognize as fake and thus do not pose a real threat. For example, earlier subjective tests done on Deepfakes have shown that the videos generated by this method can be detected by viewers with high accuracy if the video quality is sufficient. Consequently, the purpose of this research question is to study the threat environment by providing a comprehensive overview of possible generation methods, and identification of the methods which pose a real threat to the viewers, i.e. are hard to detect by individuals. This could be achieved by an extensive review of possible generation techniques and carefully designed subjective tests and would provide a solid ground for tackling the detection challenge based on empirical evaluation of the threat environment.

RQ 2: How can we detect the generation methods that the viewers are most vulnerable to? (Related chapters: 7, 8)

After the identification of the most effective generation techniques, it is crucial to propose an effective detection mechanism to mitigate the vulnerability. Otherwise, the results of the previous research question would

serve as a mere guide for forgers on the vulnerabilities of the viewers. The purpose of this research question is to measure the performance of the existing solutions for detection and introduce new detection methods that perform well in real-life scenarios. To this end, feature sets that have enough discriminative capacity need to be found and new detection mechanisms need to be introduced. The proposed methods must be evaluated quantitatively in comparison to the existing solutions on datasets that represent the deployment conditions.

RQ 3: Will the existing detection methods satisfy the requirements of detection in real-life scenarios? (Related chapter: 9)

In recent years, there has been a growing interest in general-purpose end-to-end detection methods that provide a unified detection mechanism capable of detecting various generation techniques simultaneously. For the task of fake facial video detection, there has been a growing number of such methods since the introduction of Deepfakes with near-perfect detection rates. Despite the appealing results, the performance of these methods is often only evaluated on specific conditions and thus would not directly signify their performance after deployment. One of the shortcomings of these methods is their tendency to overfit the training conditions. The purpose of this research question is to evaluate the performance of the existing methods in a more realistic scenario where the detector is tasked with the detection of fake videos collected from the wild. As a result, the specific shortcomings of these methods can be discovered, paving the way for research into more effective detection techniques.

RQ 4: How can we address the vulnerabilities of the existing methods and improve the applicability in real-life scenarios? (Related chapters: 10, 11)

As general-purpose classification methods are tasked with optimizing their performance on a specific training scenario, these methods would internally extract discriminative feature sets that perform best for classification and weight them according to their importance for detection to maximize their objective function. However, if in testing conditions there is a mismatch between the discriminative feature sets or their relative importance, their performance can significantly drop. The purpose of this research question is to answer this limitation by the introduction of methods robust to a mismatch between training and test conditions. To this end, we probe into the use of unbiased anomaly representations for one-class and two-class classification for detection. This approach would enable the preservation of a

more complete feature set for detection, as well as the introduction of a more robust importance weighting scheme.

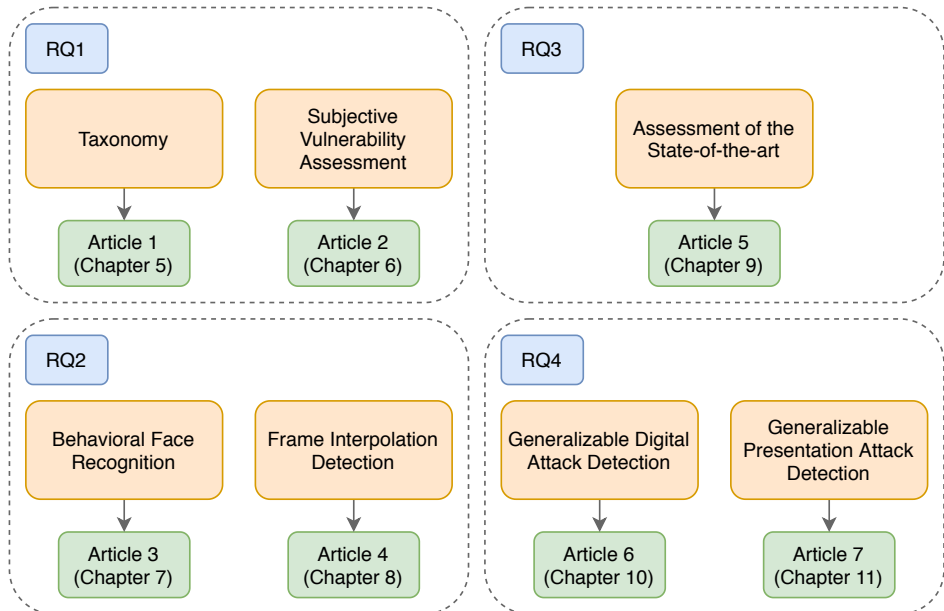


Figure 1.1: Research outline and published articles as per the research questions.

1.3 Research Methodology

Considering the aforementioned research questions as a basis, the following general research methodologies are designed. These methodologies are used throughout the thesis work and target achieving research objectives.

- **Data Collection from the Wild**

There is a lack of datasets in the literature that address the facial video authenticity problem on various generation methods. The existing datasets such as the Deepfake detection challenge (DFDC) dataset (17) often only contain data from one generation technique or are small datasets collected in a controlled environment and lack the variability observed in the real world. Reliance on datasets without sufficient variability to represent real-life conditions for the development of detection techniques may result in detectors with low detection performance in deployment as shown in the results of the DFDC challenge. To avoid such limitations, I relied on the most diverse datasets available and introduced large-scale datasets for cases where

there is no such dataset in the literature. As a result, three datasets were introduced in Chapters 7, 8, and 9 based on data collected from the wild. All these datasets are based on videos collected from YouTube (a popular video-sharing platform), and no unnecessary constraints were enforced during data collection to preserve diversity.

- **Vulnerability Assessment**

Without a clear picture of the existing vulnerabilities, the allocation of research effort can have a lesser impact on real-world applications. Consequently, before investing research effort in the direction of detection of specific generation techniques, a comprehensive vulnerability assessment based on data collected from the wild is done to find the most immediate threats that need to be addressed. Chapters 6 and 9 represent the vulnerability assessment studies. The first study is a subjective vulnerability assessment study in the online form on participants' personal devices to simulate the conditions of real-world encounters. In the second study, the vulnerabilities of a collection of existing detection methods are studied against test conditions corresponding to the diversity that exists in real-world data.

- **Feature-set Selection**

Proper selection of feature sets form the basis of any detection mechanism and is of utmost importance, as a limited feature set may lack enough discriminative information needed for detection. Furthermore, the forgers are actively working towards attacking the feature sets that are commonly used by the viewers and the detectors to maximize their chance of success (11). In chapters 7, 8, and 10, three novel feature sets were introduced based on information that are commonly neglected in a video signal. These features are in order, face behavioral biometric information, frame interpolation prediction errors, and observation log-likelihood for individual pixel intensities. Furthermore, it is shown that by the use of reliable feature sets, the detection complexity is reduced significantly, and it becomes possible to use much simpler detectors for detection. The introduced feature sets increase the barrier for successfully bypassing the detector as the forger is required to invest in attacking additional feature sets which are difficult to model.

- **Detection Algorithms**

After the selection of an appropriate feature set for detection, robust detection algorithms need to be introduced which can utilize them adequately. Deep learning-based algorithms have consistently shown their superiority to traditional handcrafted methods and thus are the machine learning method of

choice throughout this thesis. In Chapter 7 the use of embedding spaces for both features and identities is proposed by utilization of triplet loss objective function and statistical pooling over time. In Chapters 8 10 the discriminative power of the selected features made it possible to perform detection with primitive convolutional neural network (CNN) architectures. In Chapter 11 to simplify the detection network, a principal component analysis (PCA) based dimensionality reduction scheme is introduced followed by a simple deep neural network (DNN) for detection. Furthermore, assuming the rationality of the attacker, to achieve the objective of minimizing the error rate for the most powerful attack, a new loss function is introduced which exaggerates the loss for most powerful attacks and suppresses the loss for easily detectable samples. In all cases, due to the adequacy of the selected feature set, it was possible to utilize smaller detectors compared to other methods in the literature.

- **Performance Metrics**

In classification problems, the accuracy of classification is defined as the percentage of correctly classified samples over all test samples. However, in binary classification, this metric does not capture the whole picture, as the accuracy depends on the decision threshold value. Consequently, the method of choice for reporting the performance of these systems is the receiver operating characteristic (ROC) curves and their derivatives, notably the detection error trade-off (DET) curve. These curves report the missed detection rate and false alarm rates for every threshold value and make it possible to evaluate the missed detection rate of a system at any desired false alarm rate. To represent the performance of a system in a single threshold independent value, the equal-error-rate (EER) measure is used which represents the missed detection and false alarm rates on the point where these two values are equal.

ISO-IEC 30107-3 (27) provides a set of metric definitions with the goal of a unified metric vocabulary and improving the comparability of the proposed methods in the field of presentation attack detection (PAD). Accordingly, the terms attack presentation classification error rate (APCER) and bona fide presentation classification error rate (BPCER) are used to report the missed detection and false alarm rates of a PAD system respectively.

1.4 List of Included Research Publications

The following publications are part of this dissertation:

1. A. Khodabakhsh, C. Busch and R. Ramachandra, "A Taxonomy of Audiovisual Fake Multimedia Content Creation Technology," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, 2018, pp. 372-377.
2. A. Khodabakhsh, R. Ramachandra and C. Busch, "Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content," 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), Berlin, Germany, 2019, pp. 1-6.
3. A. Khodabakhsh and H. Loiselle, "Action-Independent Generalized Behavioral Identity Descriptors for Look-alike Recognition in Videos," 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2020, pp. 151-162.
4. T. Nielsen, A. Khodabakhsh and C. Busch, "Unit-Selection Based Facial Video Manipulation Detection," 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2020, pp. 87-96.
5. A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik and C. Busch, "Fake Face Detection Methods: Can They Be Generalized?," 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2018, pp. 1-11.
6. A. Khodabakhsh and C. Busch, "A Generalizable Deepfake Detector based on Neural Conditional Distribution Modelling," 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2020, pp. 191-198.
7. A. Khodabakhsh, "Unknown Presentation Attack Detection against Rational Attackers," arXiv preprint arXiv:2010.01592, 2020. (Submitted to IET biometrics)

Additionally, during the course of the PhD, a number of other publications were produced which are listed below:

1. A. Khodabakhsh, M. Pedersen, C. Busch, "Subjective Versus Objective Face Image Quality Evaluation For Face Recognition," 2019 International Conference on Biometric Engineering and Applications (ICBEA), Stockholm, Sweden, 2019, pp. 36-42.
2. E. Haasnoot, A. Khodabakhsh, C. Zeinstra, L. Spreeuwiers, R. Veldhuis, "FEERCI: A Package for Fast Non-Parametric Confidence Intervals for Equal Error Rates in Amortized $O(m \log n)$," 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2018, pp. 1-5.
3. A. Khodabakhsh, E. Haasnoot, P. Bours, "Predicted Templates: Learning-curve Based Template Projection for Keystroke Dynamics," 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2018, pp. 1-5.

1.5 Scope and Outline of the Thesis

The main scope of this thesis is to investigate and effectively mitigate the vulnerabilities of the viewers as well as the general-purpose detection methods against various photo-realistic facial video generation techniques with reliance on comprehensive vulnerability assessment and robust deep learning based countermeasures. The vulnerabilities of the viewers were evaluated through subjective tests on the most prevalent generation techniques. The vulnerabilities of the existing detectors were studied through rigorous tests on data from diverse generation techniques. To this end, several datasets were produced and shared publicly using data collected from the wild to address the lack of large public datasets on specific generation techniques. Furthermore, for each discovered vulnerability, an appropriate feature set is introduced and a detection mechanism is proposed for achieving acceptable detection performance in real-life scenarios. With reliance on the distinction between physical attacks (attacks generated before recording by a camera) and digital ones (video editing or computer-generated), the thesis presents various robust machine learning countermeasures for each attack category via the use of spatial and temporal features and proposes the utilization of complementary biometric characteristics, prediction-based anomaly features, and reliable generation artifacts to this end. The intended audience of the thesis is digital forensics and biometric presentation attack detection professionals as well as researchers from the fields of video processing and machine learning.

This thesis is divided into three parts: Part I presents an overview of the thesis, Part II presents the published articles with a focus on viewers' vulnerabilities, and Part III presents the published articles with a focus on detector generalizability. In Part I, the first chapter discusses an introduction to the thesis by describing the motivation and problem description, followed by the research objectives and questions as well as the methodology, list of published articles, and finally the scope of the thesis. Chapter 2 provides a brief background on the subject to introduce the core concepts relating to this study through explaining the related works to this thesis. Chapter 3 provides a detailed summary of each of the research articles included in this thesis and summarizes their contributions. Finally, Chapter 4 concludes this part and provides a perspective for future research directions.

The research articles presented in Part II and III are reformatted versions of the actual publications. Chapter 5 presents a taxonomy of facial video generation techniques while Chapter 6 provides the results of the subjective tests designed based on the taxonomy to evaluate the vulnerabilities of the viewers. Based on the results of these articles, two generation techniques were identified and targeted for the development of detection methods which are presented in Chapters 7 and 8. The proposed detection methods attempt to differentiate people based on their facial behavior and detect inter-frame manipulations with reliance on frame interpolation traces respectively. The weaknesses of the existing general-purpose detectors are investigated in Chapter 9 through performance evaluation on data collected from the wild. Based on the results of this study, efforts were concentrated on the development of a generalizable detection method based on anomaly representations which are presented in Chapter 10 for digital attacks and Chapter 11 for physical attacks.

Chapter 2

Background and Related Work

A summary of the related work to this thesis is provided to give an overview of the state of the matters at the time of the writing. First, the existing generation techniques are summarized and organized into groups followed by a list of relevant datasets in the literature. Afterward, detection techniques are described in relation to generation techniques and the field of study they originated from. This chapter provides the background necessary for understanding the concepts and methods described throughout this document.

2.1 Generation Techniques

Generation techniques can be broadly categorized into physical and digital techniques that take place before being recorded by a sensor (in this case a video camera), and which take place on a recorded video by manipulation of the content or outright synthesis respectively. Each category consists of widely different techniques which are described briefly in this section. Furthermore, a summary of the available datasets from both categories on the modality of the face is also provided. For ease of interpretation, relying on the presentation attack detection terminology, each generation technique is referred to as an attack. Furthermore, specific content presented to the viewer would be called a probe.

2.1.1 Physical Attacks

Physical attacks can broadly fall into one of two subcategories, the attacks that use a human or attacks that use an artificial object. Unfortunately, physical attacks are understudied and the existing literature on physical attacks is limited to the first article in this thesis (36). The use of artificial objects for realistic attacks is infeasible with existing technology due to the complexity of recreating the com-



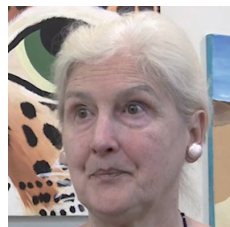
(a) Real-F Mask (48)



(b) Prosthetic Makeup¹



(c) Video Rewrite (12)



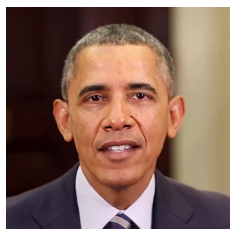
(d) Morph-cut (8)



(e) Video Face Replacement (15)



(f) Face2face (74)



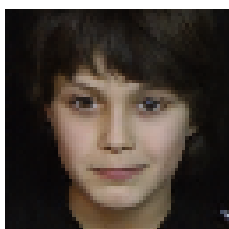
(g) Synthesizing Obama (73)



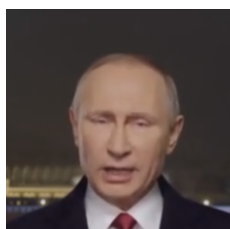
(h) VDub (21)



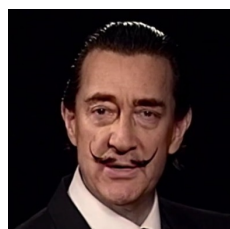
(i) RecycleGAN (7)



(j) Every Smile is Unique (80)



(k) Deep Video Portrait (40)



(l) Dali Lives (45)

Figure 2.1: Examples of generation technologies.

plex facial muscle configuration and movements, and no existing humanoid robot has achieved a convincing behavioral and physiological resemblance to a target individual. Consequently, existing attack methods are limited to the human category, ranging from the use of look-alikes and identical twins to the application of prosthetic makeup and masks. The use of look-alikes for deception has been historically documented in cases such as the impersonation of a general during World War II (14) as well as the use of political decoys by Adolf Hitler, Joseph Stalin, Henry Kissinger, Saddam Hussein, and many others. Due to the difficulty of finding cooperative look-alike actors for a specific individual, the application of various levels of makeup can be used as a substitute to increase the likeliness of an actor to a specific individual (65). For example, 3D masks (Figure 2.1(a)) can be built using soft materials with 3D printers and casting (25) and the cost of creating such masks is getting lower. These methods are often publicly used for political satire by actors such as Tracy Ullman in Tracey Breaks the News series¹ (Figure 2.1(b)). These methods are also commonly used by government agencies such as the FBI and CIA to infiltrate possible terrorist groups. As digital attacks are becoming more cost-effective compared to physical ones, they are more likely to be used in attacks.

2.1.2 Digital Attacks

Despite the availability of photo-realistic image editing tools, the application of similar techniques for video editing has been too labor-intensive and thus limited to high-budget applications. However, recently, thanks to the availability of higher computational power and the advent of deep learning-based data-driven generation techniques such as generative adversarial networks (GAN), many video-realistic methods have been proposed in the last five years. For the purpose of this study, a digital attack is defined as any digital process that alters content to change the meaning conveyed by a video or outright synthesizes a video with a fabricated meaning. Processes such as the application of imperceptible compression or creating a synthetic copy of an existing video are not considered attacks.

Traditionally, digital attacks were categorized as inter-frame and intra-frame tampering as shown in Figure 2.2 where the former signifies temporal manipulation and the later spatial manipulation. However, as new methods such as complete synthesis and tampering by use of footage from different sources are introduced, a new more representative categorization is needed to cover the whole range of attacks. Tampering can be viewed as a spectrum according to their deviation from the authentic sources as shown in Figure 2.3. Editing and inter-frame tampering can be used to cleverly change the order of frames to change the meaning

¹<https://www.imdb.com/title/tt6941630/>

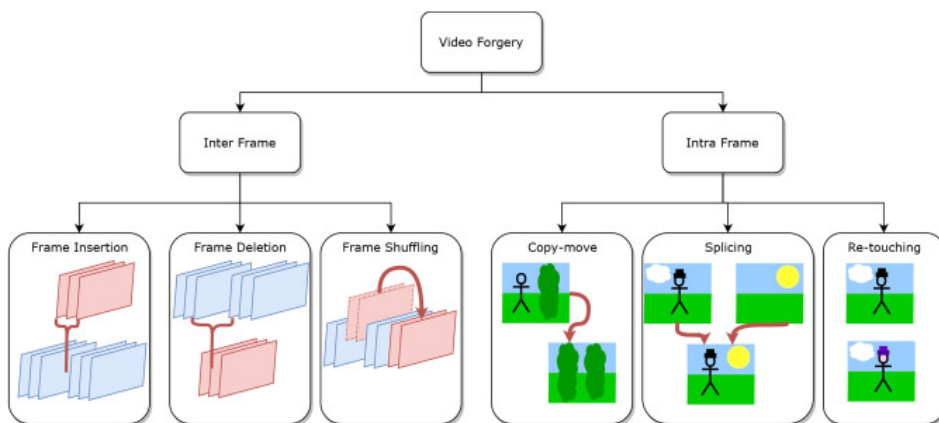


Figure 2.2: The traditional categorization of video tampering methods (31).

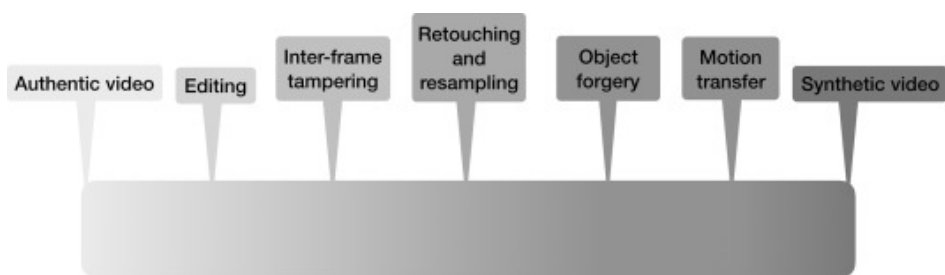


Figure 2.3: The spectrum of digital attacks (31)

conveyed in footage to a desired one by an attacker in a seamless manner. Due to the simplicity of these methods and their higher photo-realism, these methods were explored for facial video synthesis early-on for applications such as audio-visual speech synthesis and video dubbing (53). The first automatic face animation technique (Figure 2.1(c)) was proposed in 1997 (12) where a dataset of visemes was extracted from footage and concatenated by use of morphing to synthesize a new sequence. Ezzat et al. (19) improved this method by proposing the use of a single frame representation of visemes and pre-computed optical flow correspondences to reduce the amount of data needed for synthesis. Recent methods such as (8) provide higher flexibility by enabling the operator to manipulate the video by simply editing the text transcript as well as a higher realism by using intermediate frame mining along with morphing (Figure 2.1(d)). Existing commercial video editing tools such as Adobe Premiere Pro² and Avid media composer³ rely on these methods to provide a user-friendly interface for reordering frames as well as administering invisible transitions to cover the scene cuts, mainly for the application of video summarization.

Retouching and resampling attacks rely on the application of transforms or filters on the pixel intensity values to cover the traces of manipulation as an anti-forensic measure, and often happen after the application of a more severe attack. As an example, color correction methods that exist in tools such as Adobe After Effects can be used to blend videos that are recorded days apart. These methods can also be used to alter the meaning in a video, for example by color histogram adjustment to change the perception of the time of the day. Some compression methods can also be used for covering traces as well as changing the meaning as they do not explicitly conserve the meaning of a video. Compression methods have been shown to have a negative effect on the detection performance of detectors (67). Compression artifact removal (84) and video upscaling methods both in the spatial (70) and temporal (29) domains can further be used to cover the manipulation traces.

Object forgery also known as intra-frame tampering or region tampering refers to the removal and addition of objects in a video with the maintenance of temporal coherence. As shown in Figure 2.2, these attacks are traditionally categorized as copy-move, splicing, and retouching where the data used may come from two different videos. Face swapping is an example of direct application of object forgery on facial videos, where the facial region of the video is replaced with another individual's face from another video. Video Face Replacement (15) is one of the first automatic face swap methods that warps the source face to the target face based on the corresponding 3D geometry (Figure 2.1(e)). A similar system is proposed

²<https://www.adobe.com/products/premiere.html>

³<https://www.avid.com/media-composer>

in (22) where the original facial expression of the target is preserved by non-rigid warping of the source. Object forgery can also be done to replace a part of a scene with a synthetic image in order to reduce the computational and modeling costs of the synthesis method. Deepfakes are an example of such methods as the face region is replaced by a GAN-generated image where the behavior is kept intact and the appearance is altered to another individual's likeness.

Style and motion transfer methods are a category of stronger attacks, where the style or the motions in existing footage are manipulated to match that of another footage. Face2face method (74) (Figure 2.1(f)) and synthesizing Obama (73) (Figure 2.1(g)) are examples of motion transfer where the behavior of the individual in a footage is modified based on behavior from another source and an alternative speech track respectively. For the Face2face method, the texture is further optimized in (76) to improve realism, and the amount of required footage is reduced to a single image in (6). In (21), the authors propose the use of a high-quality 3D face capturing technique for altering the face of an actor to match the mouth movements of a dubber (Figure 2.1(h)). Style transfer methods such as (7) and (78) achieve a similar effect by treating the likelihood of the individual as the style and generating photo-realistic face images from semantic segmentation masks (Figure 2.1(i)).

Synthesis of the whole video in a photo-realistic manner is a challenging task due to the number of details that are needed to be considered as well as the acute sensitivity of humans to synthesis artifacts also expressed by the concept of the uncanny valley (56). Despite this, there has been incremental progress in realistic synthesis in recent years. In (80), the authors generate a photo-realistic smile sequence from a single aligned face image using a series of conditional long short-term memory (LSTM) networks (Figure 2.1(j)). Image-to-image translation has been used in (40) to convert computer graphic rendering of faces to real images (Figure 2.1(k)). In (28) conditional adversarial networks are used to translate facial landmarks into a realistic video. The use of frame prediction is investigated in (50) where CNN, LSTM, and deconvolutional neural networks were used together for the generation. More recently, the Salvador Dalí Museum created a realistic appearance of the painter himself (Figure 2.1(l)) in an exhibition called Dalí Lives using archival footage from interviews (45).

2.1.3 Datasets

There are very few datasets in the literature that can be used for the objective of physical attack detection. In (88), the authors introduce a private dataset of videos collected from 39 twin pairs. The proposed dataset of look-alikes (37) is the only dataset of videos from look-alikes in the literature with 85,000 videos from 1000 look-alike pairs. Detection of makeup attacks ranging from cosmetic

makeup to masks is mainly studied under the topic of presentation attack detection. Datasets that include videos of mask attacks are the private Morpho dataset (18) with 199 mask attacks, 3DMAD dataset (18) with 3D mask attack videos from 17 subjects, and HKBU-MARs dataset (48) with 12 masks recorded with multiple devices under various lighting conditions. A more diverse dataset of attacks is introduced in (49) by the name of SiW-M which not only includes 3D mask attacks but also includes attacks with silicone masks and transparent masks. Furthermore, it includes makeup attacks of both cosmetic and impersonation types as well as obfuscation attacks. This dataset contains 1,630 videos of length 5 to 7 seconds from 13 types of attacks.

The majority of the datasets in the literature with a focus on digital facial manipulations are limited to Deepfakes. DF-TIMIT dataset (42) includes 620 Deepfake videos at sizes 64×64 and 128×128 , Deepfake Detection dataset contains 3,068 Deepfake videos from volunteer actors, and Celeb-DF dataset (47) contains 5,639 Deepfake videos at various resolutions. Recently, two large-scale datasets are introduced, namely, the DFDC dataset (17) consisting of 100,000 Deepfake videos as well as 19,000 pristine videos at $240p$ to $2160p$ resolutions and the DeeperForensics dataset (30) with 10,000 Deepfake videos as well as 50,000 pristine ones. A few datasets include multiple attacks, namely the proposed FFW dataset (39) which includes a set of 150 manipulated videos collected from the wild, and the FaceForensics++ dataset (68) which includes 4000 videos of Deepfakes, CGI, and splicing as well as 1000 pristine videos. The main focus of all these datasets is object forgery and motion transfer attacks. The proposed Morph-Cut dataset (60) is the only dataset in the literature that includes editing attacks.

2.2 Detection Techniques

Detection of physical attacks and digital attacks has traditionally been done in two separate fields of study of facial video presentation attack detection and multimedia forensics. With the advent of Deepfakes, a lot of research effort has been directed towards the detection of Deepfake attacks as its own subfield.

2.2.1 Presentation Attack Detection

Despite the fundamental difference between the task of presentation attack detection with its focus on protecting biometric systems from attacks and the task of protecting viewers from fake content, there is a considerable overlap corresponding to the presentation attacks that are also photo-realistic. The overlap mainly covers makeup attacks and mask attacks which are better studied in the field of presentation attack detection. Passive presentation attack detection methods, also known as software-based methods, try to use the available data in the probe for

making a decision, in contrast to hardware-based and challenge-response methods where additional data is extracted to facilitate detection. These methods rely on physiological signs of life such as eye blinking and facial expression changes as well as texture and deformation features for detection (25). Consequently, these methods can be categorized into static methods and dynamic methods, corresponding to the use of static features such as texture and the use of motion.

The texture-based methods try to learn the facial micro-textures that characterize real faces and have been effectively used for the detection of photo attacks. Local binary patterns (LBP) (13) is a popular texture descriptor followed by different learning algorithms for detection. Translation to a more proper feature space has also been investigated for detection in (87) using Fourier transformation. Another group of methods focus on quality degradation detection as the quality of a generated probe is often lower than the reference due to the imperfections of manufacturing stages during the generation process (20). Other characteristics of the human face and skin such as absorption, reflection, scattering, and refraction have also been used for detection (41). Texture-based methods work best when the resolution of probes is sufficiently high for analysis of the textures, and their performance drops when facing bad illumination conditions and post-processing stages such as compression.

Physiological signs of life can be used as dynamic features. For example, humans blink on average three times per minute and irregular blinking rates can be used for liveliness detection (83). Pulse is another sign of life that can be extracted from video footage using Eulerian video magnification (82) and be used for liveliness detection (9, 64, 61). It is important to mention that while these methods can be effective against the mask and prosthetic makeup attacks, they would fail against a look-alike or light cosmetic makeup attacks as these attacks also contain the liveliness and texture features corresponding to real faces. To distinguish look-alikes and identical twins, distinct facial features such as marks (72), face asymmetry (32), and aging-related features (44) can be used. Biometric systems can also be used for distinguishing look-alikes and identical twins after these systems are fine-tuned specifically on this task (3). The use of unique behavioral features has also been proposed in (88) and (37).

2.2.2 Deepfake Detection

Following concerns over the use of Deepfakes for fake news, hoaxes, and financial fraud, communities of media forensics, biometric anti-spoofing, and data-driven deep learning have joined efforts to address these threats. Consequently, there is a growing interest in Deepfake detection evident from the growing number of workshops, conferences, and competitions dedicated to this topic (77). Most pro-

posed Deepfake detection methods try to detect artifacts that are produced by the GAN pipeline. In (54) the authors rely on the color difference between pristine images and GAN-generated images. The use of convolutional traces of the generative model for detection is proposed in (23). GAN fingerprints caused by specific GAN architecture are further studied as a means of detection in (86). The use of eye color mismatch, missing reflections, and missing details in the eye and teeth regions were proposed in (52). The artifacts that are caused by the misalignment of the 3D head pose and the synthesized region are used in (85). Eye blinking patterns have also been proposed for detection in (33). As GAN-generated images have a fixed resolution compared to the target video, there will be a resolution mismatch between the background and the facial region which was used in (46) for detection.

A number of articles propose the use of general approaches that are not specific to GAN generation artifacts. The neural activation difference of face recognition systems such as VGG-Face (62), OpenFace (5), and FaceNet (71) have been used by (79) for classification. In (57) the authors propose the use of steganalysis features extracted as pixel co-occurrence matrices. Inconsistencies between lip movement and audio speech were also investigated in (42) as a means of detection. Facial expression and head movement correlations of four individuals were extracted based on facial landmarks and modeled in (2) for detection. Steganalysis and mesoscopic features are other general-purpose features that have shown to work well for Deepfake detection in (1) and (89). In order to utilize both spatial and temporal information for detection, 3DCNNs were used in (81) for improved performance on low-quality videos. The temporal inconsistencies are also used in (24) and in (69) via recurrent neural networks, and via optical flow fields in (4). The discriminative power of individual regions of the face was studied in (77). The spatial and spectral features extracted from photo response non-uniformity (PRNU) patterns are used in (66).

Several studies have taken a machine learning-based approach for improving detection performance. The multi-task incremental learning of new types of GAN-generated images was explored in (51) as a measure of improving the performance against new attacks. Attention mechanisms have also been applied to improve the performance of detection systems in (16). The use of general-purpose image classifiers was proposed in (68) which outperformed the specialized methods on the proposed dataset. In (58), it has been shown that multitask learning for both detection and localization of manipulated regions improves the detection performance. Capsule networks have also been shown to perform on par with existing methods with fewer parameters (59). In (10) restricted Boltzmann machine networks are used for detecting manipulated patches in the image.

Chapter 3

Summary of Published Articles and Contributions

The research done towards this PhD is split into two threads with two different views on the problem. One thread is initiated towards the analysis of the vulnerabilities of the viewers and responding to each vulnerability accordingly. A second thread is also initiated with the goal of finding a generalized detection solution following the observation that existing general-purpose detectors suffer from generalizability issues. The following sections discuss the research findings and contributions of the published articles during these studies.

3.1 Viewers' Vulnerabilities

To analyze the vulnerabilities of the viewers to the existing generation techniques, the first step taken was to study the existing generation techniques and form an understanding of what is possible with current technology. Resulting from this study, a taxonomy of possible generation techniques was produced in (36), forming the basis for studying viewers' vulnerabilities. Next, in (38) a collection of videos from the most prevalent generation techniques is collected and subjective tests are designed to see which ones the viewers are most susceptible to. Resulting from this, two methods of video generation to which the viewers were susceptible were identified, namely, unit-selection-based editing and use of look-alikes.

For detection, with a focus on the aforementioned two types of generated videos, two solutions are proposed. To detect unit-selection-based editing which is a form of inter-frame forgery, the traces of frame-interpolation are successfully used for detection in (60). To detect look-alikes, and by extension, videos where the

physiological attributes are not reliable for authentication, a face recognition method based solely on facial behavior is introduced in (37). The four aforementioned studies are summarized below.

3.1.1 Taxonomy

Detection of generated videos needs an understanding of different generation techniques as well as their strengths and weaknesses. The importance of face in human interactions results in the sensitivity of humans to any subtle imperfections in a generated video. As a result, generating realistic facial videos was a major challenge until recently, and historically was limited to impersonations and subtle video editing. Availability of powerful 3D rendering hardware and software resulted in the introduction of CG face videos in the movie industry in the early 2000s and deep learning based synthesis methods made realistic low-cost generation possible in 2017 after the introduction of Deepfakes.

For a video to be perceived as realistic, it should have a realistic representation in all three available modalities, namely appearance, behavior, and speech. Furthermore, these three modalities can be independently generated and combined, and as such, they can be analyzed independently. The source of appearance, behavior, and speech in a generated video can be a living human and alternatively his/her recording, or a computer model and its physical realization. As such, the facial video generation techniques cover a wide range of methods ranging from impersonation to android robots to CG faces to video editing.

3.1.2 Subjective tests

Based on the produced taxonomy, a collection of 24 realistic five-second videos from six of the most prevalent generation techniques was collected from a public video-sharing website. The six techniques correspond to impersonation by look-alikes, impersonation with prosthetic makeup, 3D CGI faces, GAN-generated faces, inter-frame forgery, and partial video editing. These videos were used along with an equal number of pristine videos for a subjective test where the natural encounter in social media is simulated on the web and the opinion of the viewers on whether the videos are “real” or “fake” is asked. In total, 30 people participated in the subjective tests, and their performance was measured on the detection task. The viewers were randomly subjected to a familiarization step and viewed the videos along with a biometric reference pristine video of the target individual half the time to analyze the effect of these parameters.

After the analysis of the results, it became apparent that the participants had a low detection rate on videos generated by two of the six techniques, namely the use of look-alikes and inter-frame forgery. These methods have the smallest foot-

print in the produced video compared to the other four categories. Furthermore, the results show that the availability of a biometric reference and familiarization reduced the number of errors the participant made while knowing the target individual made participants more uncertain without a loss in correct classification for pristine videos. Another interesting observation was that the demographic parameters can have an impact on detection performance as older participants were more confident in their decision which resulted in a higher error rate. This is an example of human bias in decision-making in contrast to the widely discussed bias in AI, and not taking such biases into account can leave certain portions of the society more vulnerable to attacks. Consequently, this shows the importance of studying and taking into account these biases via promoting participant diversity and monitoring its effect on subjective studies.

Another interesting observation is that a literature review (as described in details in Chapters 7 and 9) reveals a lack of attention to the detection of the attacks which the subjective experiments identified as the most effective. This effect is rooted in the fact that the viewer vulnerabilities can have little to no intersection with the detector performance (43) and detector vulnerabilities (26). Nevertheless, the majority of articles in the literature focus on these factors regardless of whether the viewers are already able to notice the visible artifacts in the generated content (43) or outright replicate human's detection ability by relying on visible artifacts for detection (77). As the objective of a detector is protecting the viewers from attacks, the detector needs to complement the human ability in detection, and this can be most effectively achieved by reliance on objective measurement of the subjective vulnerabilities. In other words, subjective tests give a clearer understanding of the attack landscape and guide research towards solutions that address the viewer's vulnerabilities. Consequently, in the subsequent publications, the vulnerabilities identified in the subjective vulnerability assessment study are targeted for mitigation.

3.1.3 Look-alike recognition

Relying on the results of the subjective test study, a next study was initiated on the recognition of look-alikes. Look-alikes and identical twins pose a challenge to both humans and face recognition systems as the physiological likeness of the pairs reduces the efficacy of physiological appearance in recognition. Prior studies have shown the independence of behavioral and physiological attributes, the permanence of behavioral attributes, and the possibility of behavioral face recognition in fixed-phrase and fixed-action scenarios in controlled environments. The behavioral attributes of individuals are available in video footage and can provide a robust alternative to physiological attributes when the latter is not sufficient for face recognition. As such, a study is initiated to probe the possibility of action-

independent face recognition without constraints on the environment as a substitute for physiological face recognition. For this purpose, 1000 look-alike pairs were mined using a state-of-the-art face recognition system along with verification of discrimination difficulty using subjective tests. Based on these 1000 pairs, a dataset of 85,656 utterances was created using videos that were originally collected from YouTube without any specific constraints on the recording environment.

The size of the produced dataset enables the training of deep learning solutions. To single out behavioral information in the videos, the locations of facial landmark positions were extracted and normalized based on the facial pose, size, and average landmark position after normalization. Then a network architecture is designed that consists of a 1D convolutional feature embedding extractor followed by a statistical pooling layer over time dimension and a couple of fully connected layers to map the input landmark position matrices to an identity embedding. The network was trained on a separate set of 4500 non-overlapping identities using triplet loss on euclidean distances and tested on the look-alike pairs. The results of the study show that the proposed method on the proposed dataset can achieve an equal-error-rate of as low as 8% for verification where the state-of-the-art physiological face recognition system had an EER of 30%. As the proposed method relies solely on facial behavior, it would be robust to any physiological attribute manipulation which conserves behavioral information faithfully.

3.1.4 Inter-frame Forgery Detection

Based on the results of the subjective tests, the second category of generated videos where the subjects are susceptible to is unit-selection-based video manipulation. In these methods, the forger uses existing footage of a target individual, and by cutting, reordering, and joining via frame interpolation, he/she can manipulate the actions of the individual and thus the meaning in the footage. Inter-frame forgery detection on facial videos is a largely unexplored field of research despite the ease of generating such content with commercially available tools. Popular video editing tools such as Adobe Premiere Pro and Avid Media Composer contain tools for such manipulation which are accessible through their easy-to-use graphical user interface. For the purpose of the detection of these manipulations, there exists no dataset in the literature. Consequently, a dataset of 1000 videos was introduced based on the automation of the video generation process of Adobe Premiere Pro Morph Cut transition using an internal scripting language called Extendscript. The dataset is based on videos that are originally collected from YouTube and a pre-filtering step based on face bounding-box movement is done to ensure no abrupt jumps exist in the produced footage. The quality of the generated videos was further assured by a manual post-filtering step.

For detection, the frame interpolation traces¹ were focused on. Any interpolation method would have to rely on the information available in the frames before and after to fill in the intermediate frames. Consequently, the interpolation method would generate a smooth transition with lower-than-natural variability in the generated frames. To uncover the amount of natural variability in a frame, another frame interpolation technique can be used to predict the frame based on the frames before and after. The prediction error calculated as the difference between the actual frame and the parallel predicted frame will have traces of over-smoothness in interpolated frames. The results show such differences between the prediction error of interpolated frames and pristine frames. Using the prediction error as a discriminative feature, a simple neural network is used to detect interpolations at frame level with an accuracy of 95%.

3.1.5 Contributions

The contributions of this thread of research can be summarized as follows:

- Providing one of the first taxonomies on the topic of facial video generation techniques, dating before the advent of Deepfakes, with a comprehensive view and covering a wide range of possible digital and physical generation techniques.
- Identification of the most effective generation techniques using the existing examples of the most prevalent generation techniques via subjective tests with a realistic setup, as well as studying the effect of parameters such as familiarity, biometric knowledge, and demographic information on the performance of individuals.
- Collecting the biggest dataset of look-alike pairs in the literature and composing a large-scale video dataset of these pairs big enough for the development of deep learning solutions.
- Introduction of the first general-purpose action independent behavioral face recognition system in the literature which performs well on the videos from the wild relying only on normalized landmark movements.
- Generation of the first facial inter-frame forgery dataset in the literature containing 1000 videos based on the automation of the generation process of the most popular commercial video editing tool. The dataset is also the biggest dataset of inter-frame forgery in the literature.

¹<https://www.youtube.com/watch?v=RA4mAPitn4E>

- Utilization of frame prediction error for a simplified two-step inter-frame forgery detection method with a detection rate of above 95%.
- Making the generated datasets available to the public to stimulate further research.

3.2 Generalizability

One of the major challenges for existing detectors including the ones introduced in the previous section is generalizability. A detector learned on a specific set of generation techniques tends to overfit the features that discriminate against these generation techniques from the pristine data. However, when a new generation technique appears for which the discriminative features are different from the learned features, the performance of the detector drops drastically. The performance can drop further if there is no overlap between the discriminative features of the new technique to the previous ones. To address this issue, a first study was done in (39) to evaluate the performance of the state-of-the-art digital manipulation detection techniques on videos generated by unknown generation techniques. The results of this study showed that existing solutions do not perform well when faced with videos generated with an unknown generation technique. To address this issue, the use of generalizable features learned by a generative model for digital manipulation detection is proposed in (35) with a frame-level unknown manipulation detection accuracy of 95.7%. This work is further extended to detect physical manipulations in Chapter 11 where the rationality of the attacker was also taken into account with reliance on game theory. The proposed method outperformed existing state-of-the-art on the task of presentation attack detection when facing rational attackers. The three aforementioned studies are summarized below.

3.2.1 Generalizability of existing methods

To measure the generalization capacity of digital manipulation detectors, a set of 150 videos were collected from YouTube which would fit the definition of digitally manipulated realistic videos. The videos were generated with various techniques including GAN, CGI, face swap, and manually tuned synthesis by artists. The biggest dataset of digitally manipulated videos at the time of this study was the FaceForensics dataset (67), on which the state-of-the-art detection methods were tested and introduced. To evaluate the performance of the state-of-the-art systems on unknown generation techniques, these systems were trained on the FaceForensics dataset and tested on the collected dataset. The results show that even though these models have a near-perfect detection accuracy on the test set of the original dataset, they fail to detect the videos in the collected dataset with the lowest EER being 27%. These results show that the detection algorithms overfit the

generation techniques available during training. These results were further verified by doing tests on a dataset of face swap images where the EER further worsened to higher than 40%. Furthermore, the results show that, as expected, the performance of the detectors is slightly better on the videos that are generated with a similar generation method to the existing methods in the original dataset.

3.2.2 Generalizable Detector

As explained before, the detection performance of the existing state-of-the-art is reduced drastically when faced with unknown generation techniques. To systematically approach the problem of generalization, a theoretical understanding of the underlying cause is needed. When a discriminative model is trained on videos from a set of generation techniques, the model tries to find the discriminative features that are most useful in the detection of these videos. Furthermore, the model would rank these features according to their usefulness for detection. As the discriminative features that are useful for the detection of unknown generation techniques are unknown, if there is a mismatch between these features and the features that the model relies on the most, the model is expected to perform poorly. In this scenario, anomaly features may prove more useful for the purpose of detection. Anomaly features describe the distribution of pristine data rather than the best discriminative features for detecting any specific set of generation techniques, and consequently, are not biased towards any specific set of generation techniques. Therefore, even though using anomaly features for detection might result in a lower detection rate on known generation techniques compared to the discriminative detectors, they are expected to perform better when faced with unknown generation techniques.

A complete feature set would be a set that fully defines the distribution of pristine data in the pixel representation. If the distribution of pristine data was available, the likelihood of a probe to this distribution would serve as an ideal anomaly feature for detection. However, due to the complexity of this distribution, existing methods for modeling the distribution focus on a segment by segment distribution modeling. One such method is the generative model PixelRNN which models the distribution of individual pixel values conditioned on the pixel values before in raster order. This generative model can be used to measure the likelihood of observing every individual pixel value at each pixel location. Using this model, an image can be converted to a likelihood matrix of the same size, which can be used to find the overall likelihood of observing the image as well as to locate the anomalies in the image at pixel level.

The results of the studies show that the overall likelihood of observing individual images has limited capacity in discriminating pristine and generated images with EERs around 25% as the distributions have a significant overlap. As such, the loc-

ations of the anomalies were investigated further for detection; as visual inspection and t-SNE representation showed that there are differences between pristine and generated frames in this representation, and samples from a specific generation technique tend to cluster together. To use this anomaly representation for detection, two approaches were investigated. The first approach was to directly use them as input to a simple convolutional neural network for training. This approach could perform on par with the state-of-the-art when faced with frames from known generation techniques on FaceForensics++ (68) dataset. However, when faced with samples from unknown generation techniques, the method achieved an average frame-level detection accuracy of 95.7%.

In the second approach, these representations were refined to a compressed representation for video-level detection while conserving the detection-relevant information. Assuming the location of anomalies on the face to be constant, the frame-level representations can be denoised by averaging over time dimension, resulting in a single fixed-length denoised representation per video. In addition, using principal component analysis trained on the training data, a hyper-plane can be defined where the pristine data lies on. After sorting the PCA components according to the explained variability on the training data in descending order, the PCA representation can show in which directions the pristine data has the least variability, and the distance to the PCA hyper-plane can further represent the unexplained variability of a given input. By combining the last elements of the PCA representation with the unexplained variability, a second anomaly feature can be extracted corresponding to the explained energy of the representation along the components where the representation of pristine data shows the least amount of energy. This representation performs equally well compared to the overall image likelihood as an anomaly score, and at the same time is independent of it with a correlation value of 0.15.

In this study, a further parameter was introduced corresponding to the rationality of the attacker when choosing a generation technique. In all existing methods in literature on the performance of detectors when facing multiple generation techniques, it is assumed that the attacker chooses a generation technique at random with equal probability. However, relying on game theory, it can be shown that the attackers tend to use the most powerful attack (MPA) available to them defined as the attack with the highest expected success rate across his options. Consequently, when facing a rational attacker, the average detection rate would not represent the real-life performance of the system, and the performance of the detector when facing the most powerful attack should be used as a measure.

The two attack independent anomaly features explained earlier, when combined, have an unknown digital attack detection MPA EER of 8.2% on FaceForensics++

dataset and 27.1% on physical attack detection on SiW-M (49) dataset. To improve the detection performance against a rational attacker when all the attacks are known attacks, a new loss function named categorical margin maximization loss on an L2 normalized embedding space is introduced to maximize the performance accordingly. This objective function exaggerates the loss for samples that are on the margin between pristine and attack videos and suppresses the loss for samples that are classified correctly. Consequently, most of the training loss would come from generation techniques for which the classification performance is the lowest, resulting in the optimization of performance for MPA while a more balanced performance is achieved across known techniques. The proposed discriminative classifier uses the compressed PCA representation as input and achieves an MPA EER of 0.7% on the digital manipulations and 9.7% on presentation attack detection tasks. The logic-based fusion of the discriminative and the one-class classifier results in a slight improvement of performance in most cases. Finally, thanks to the capacity of the proposed representation in clustering the samples from the same generation technique together, the proposed method achieves a remarkable few-shot learning capacity which allows it to reduce MPA EER by 15.2% with only five samples.

3.2.3 Contributions

The contributions of the second thread of research can be summarized as follows:

- Collection of the first dataset of digitally manipulated videos from the wild and making the datasets public to stimulate further research on the problem of generalizability.
- Experimental demonstration of the limited generalization capacity of state-of-the-art detection methods.
- Introduction of a generalizable detection system that achieves an unknown MPA detection accuracy of 89.3% on widely different digital generation techniques at frame-level along with pixel-level explainability.
- Game theoretical formulation of interactions between the attacker and the detector, and justification of the use of performance on the most powerful attack as a more realistic performance metric for a detection system.
- Introduction of a new loss function named categorical margin maximization loss that optimizes the performance of the detector towards the highest MPA detection rate on known attacks.

- Introduction of a novel detection mechanism for both known and unknown attack detection with few-shot learning capacity based on pixel-level log-likelihood values along with a robust fusion mechanism for the combination of a discriminative and a one-class classifier.

Chapter 4

Conclusion and Future Work

This thesis aimed to provide an understanding of the vulnerability environment concerning facial video authenticity at this point in time and provide solutions for mitigation of the discovered vulnerabilities. To this end, the research objectives and research questions were formulated in Section 1.2 and extensive research work is done to address these questions through publications included in Parts II and III. These seven publications are the main contributions towards the composition of this thesis. The thesis emphasizes the importance of vulnerability assessment and the generalizability of introduced detection methods. The results in this thesis strongly suggest the following general conclusions regarding the main questions to be answered by this research:

- Despite the significant attention directed towards Deepfake detection, several other generation techniques including traditional ones are available to malicious actors for creating photo-realistic facial videos. Comprehensive analysis of the generation techniques and subjective vulnerability assessment can result in the discovery of previously unknown vulnerabilities. Examples of these techniques that were discovered through our research are editing-based methods and the use of physiological similarity. These methods can be very effective as they apply minimal alterations in footage and do not produce significant artifacts. Furthermore, the vulnerability assessment results on the viewers show that there can be demographic differences between the detection ability of parts of society.
- A video contains a rich collection of discriminative information, and while a forger may achieve realism in certain features, passive detection with reliance on other features in a video is feasible. For example, as the main

focus of generation techniques is realism and similarity of the physiological attributes, the behavioral attributes can provide further information that can be used for detection. Attention to feature set selection can result in the identification of features that not only perform well on the detection task but also reduce the detection complexity.

- Classification-based methods tend to overfit the conditions in the training data and their near-perfect performance on limited scenarios do not guarantee their deployability. Due to the ever-growing number of generation techniques, there is an immediate need for detection methods that can perform well against new generation techniques. Anomaly features have shown a better capacity for the detection of unknown attacks and can provide a solid basis for the detection of unknown generation techniques as they do not suffer from the same weakness of overfitting the known generation techniques. Even though the classification-based methods may provide better performance on known generation techniques, considering the rationality of the attacker, the weakness against unknown generation techniques reduces their utility in real-life scenarios.

4.1 Limitations

Following the introduction of deep learning fueled by the availability of high computational capacity of GPUs, the field of machine learning has been experiencing advancements at an incredibly rapid pace. In this context, computer-generated content is not an exception and new generation techniques being introduced on a frequent basis. As a result, it is important to take into account the time frame of 2017 to 2020 as the context in which the research presented in this thesis was done. In this time period, the computer-based generation techniques were in their infancy and few generation techniques could achieve enough realism to bypass human judgment. The first realistic generation techniques with roots in academic research were instigated in 2016 following the introduction of early techniques such as Face2Face (74) and Synthesizing Obama (73). At the end of 2017, Deep-fakes¹ were introduced by amateur developers and gained much attention due to its open-source implementation and public availability and this was shortly followed by the appearance of commercial applications.

Considering the attack landscape, the results of the presented taxonomy and subjective vulnerability assessment studies conducted in late 2018 showed that many of these generation techniques did not pose an imminent threat at the time due to significant visual artifacts visible to the viewers. Thus the focus of the studies in

¹<https://github.com/deepfakes/faceswap>

Part II of the thesis was shifted away from DeepFake detection. However, following the introduction of more effective generation techniques in the subsequent years and the refinement of existing techniques, the quality of these methods soon improved and this led to much higher success in bypassing human judgment, as shown recently by Korshunov and Marcel in (43). To respond to these changes in the attack landscape, the research presented in Part III tries to address vulnerabilities to more recent attacks as well.

Since the submission of the articles included in Part II and III, we have seen further advancements in the quality of the generated content and the introduction of advanced generation techniques such as StyleGAN2 (34), Neural Voice Puppetry (75), and DeepFaceLab (63). The performance of the proposed detection methods against these new attacks is unknown and requires further investigation. The dynamic attack landscape shows the need for constant subjective vulnerability assessment and reevaluation of existing detection methods against new attacks, as well as research into new detection techniques with more emphasis on endurance against the constant changes in the attack landscape. The approaches proposed in Chapters 10 and 11 describe our efforts in this direction.

4.2 Future Work

Based on the research work carried under the scope of this thesis, the following research directions are outlined worth considering for future work.

4.2.1 Scalable Solutions

In video content, there are large amounts of information present for the purpose of detection. Consequently, the lack of sufficient data in each observation is not the limiting factor for the detection performance, and the bottleneck comes from the ability to utilize the information for the task. As a result, the majority of the developed state-of-the-art solutions in the literature rely on extensive computations for achieving state-of-the-art detection performance. In contrast, considering the deployment environment where hundreds of hours of video are uploaded to online services such as YouTube every minute, these solutions would face major scalability issues. This calls for further research in the direction of the development of efficient and scalable solutions with a focus on reducing the computational cost of detection while maintaining competitive performance to the high-cost state-of-the-art. An example of such solutions can be the extraction of compact yet reliable feature sets from video data using low-cost video processing methods. The definition of evaluation criteria which take the number of operations for decision making into account would direct the competition towards efficient detectors.

The challenges arising from the high computational expense of video processing

are further evident from the observation that most groundbreaking advancements in the field come from a limited number of research groups having access to sufficiently large computational resources. In these conditions, replicating the state-of-the-art by itself would be a major logistic challenge for independent researchers, raising the entry barrier to the competition and consequently limiting the speed of progress in research.

4.2.2 Diversity in Datasets

There is a large and ever-growing number of generation techniques that are within the reach of malicious actors for creating fake facial videos. Despite the significant progress made in recent years in the development of solutions with near-perfect detection rates on specific attacks, the results of competitions such as the Deepfake detection challenge shows that these solutions have limited applicability when faced with new attacks. Despite the introduction of large-scale datasets in recent years, these datasets are often limited in terms of the variety of generation methods. Reliance on these datasets for the development of detection solutions enforces bias on the trained detectors. Lack of diversity in a dataset limits the performance evaluation scenarios for unknown attack detection, and the existence of biases towards specific generation techniques makes the optimized detectors ill-fitted for real-life applications. The detection of unknown attacks is still a challenging problem and requires a significant research effort to be resolved. Possible solutions may be reliance on robust feature sets for detection which can be learned by observing several generation techniques during training, as well as further developments on few-shot learning methods.

4.2.3 Robustness Against Adversarial Attacks

Similar to any other forensic scenario, the forgers are actively looking for exploits in the detection mechanisms to increase their chance of success. This dynamic nature calls for the development of flexible detectors that are capable of adapting to changes in observed attacks. For example, the effects of obfuscation on the performance of the detection systems with methods such as obstructing the face, artifact removal, and bad recording conditions are largely understudied. Continuous efforts towards anticipating and monitoring new attack techniques and assessing the vulnerability of detection techniques and content consumers would further increase the awareness of the threat environment and provide grounds for mitigation of the vulnerabilities before the forgers can exploit them.

Most existing research relies on the incremental evolution of detection mechanisms based on the empirical evaluation of the proposed methods in specific scenarios, and as a result, the research lacks a theoretical foundation for guidance. In

contrast, researchers in the field of digital forensics have taken great steps towards understanding and formulating the underlying phenomena and developing methods that give a significant advantage to the detection side in an arms race. Consequently, the research can significantly benefit from the adaptation of a similar approach and creating a strong theoretical backbone that would serve as fertile grounds for the development of robust detection systems. For example, reliance on anomaly-based or complete feature sets can increase the range of discriminative information available for detection. Furthermore, a game-theoretic view on the dynamics between the attackers and defenders can provide ways to model the interactions and result in the formulation of more realistic objective functions to improve the robustness of detectors against adversaries.

Bibliography

- [1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *WIFS*, pages 1–7. IEEE, 2018.
- [2] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR Workshops*, June 2019.
- [3] B. Ahmad, M. Usama, J. Lu, W. Xiao, J. Wan, and J. Yang. Deep convolutional neural network using triplet loss to distinguish the identical twins. In *2019 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6, 2019.
- [4] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.
- [5] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. 2016.
- [6] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. Bringing portraits to life. *ACM Trans. Graph.*, 36(6), November 2017.
- [7] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [8] Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. Tools for placing cuts and transitions in interview video. *ACM Trans. Graph.*, 31(4), July 2012.

- [9] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh. Computationally efficient face spoofing detection with motion magnification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 105–110, June 2013.
- [10] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer. Detecting facial re-touching using supervised deep learning. *IEEE Transactions on Information Forensics and Security*, 11(9):1903–1913, 2016.
- [11] Rainer Böhme and Matthias Kirchner. *Counter-Forensics: Attacking Image Forensics*, pages 327–366. Springer New York, New York, NY, 2013.
- [12] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *SIGGRAPH '97*, pages 353–360, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.
- [13] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2012.
- [14] Meyrich Edward Clifton-James. *I was Monty's Double*. Panther Books, 1958.
- [15] Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, Daniel Vlastic, Wojciech Matusik, and Hanspeter Pfister. Video face replacement. *ACM Trans. Graph.*, 30(6):130:1–130:10, December 2011.
- [16] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [17] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset, 2020.
- [18] N. Erdogmus and S. Marcel. Spoofing face recognition with 3d masks. *IEEE Transactions on Information Forensics and Security*, 9(7):1084–1097, July 2014.
- [19] T. Ezzat and T. Poggio. Miketalk: a talking facial display based on morphing visemes. In *Proceedings Computer Animation '98 (Cat. No.98EX169)*, pages 96–102, 1998.

-
- [20] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Transactions on Image Processing*, 23(2):710–724, Feb 2014.
- [21] P. Garrido, L. Valgaerts, H. Sarmadi, I. Steiner, K. Varanasi, P. Pérez, and C. Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Comput. Graph. Forum*, 34(2):193–204, May 2015.
- [22] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [23] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [24] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.
- [25] Javier Hernandez-Ortega, Julian Fierrez, Aythami Morales, and Javier Galbally. Introduction to face presentation attack detection. In *Handbook of Biometric Anti-Spoofing*, pages 187–206. Springer, 2019.
- [26] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3348–3357, January 2021.
- [27] ISO/IEC 30107-3:2017. Information technology - Biometric presentation attack detection - Part 3: Testing and reporting. Standard, International Organization for Standardization, September 2017.
- [28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

- [29] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slo-mo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [30] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deepforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [31] Pamela Johnston and Eyad Elyan. A review of digital video tampering: From simple editing to full synthesis. *Digital Investigation*, 29:67–81, 2019.
- [32] F. Juefei-Xu and M. Savvides. An augmented linear discriminant analysis approach for identifying identical twins with the aid of facial asymmetry features. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–63, 2013.
- [33] T. Jung, S. Kim, and K. Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020.
- [34] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [35] A. Khodabakhsh and C. Busch. A generalizable deepfake detector based on neural conditional distribution modelling. In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2020.
- [36] A. Khodabakhsh, C. Busch, and R. Ramachandra. A taxonomy of audio-visual fake multimedia content creation technology. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 372–377, April 2018.
- [37] A. Khodabakhsh and H. Loisel. Action-independent generalized behavioral identity descriptors for look-alike recognition in videos. In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6, 2020.
- [38] A. Khodabakhsh, R. Ramachandra, and C. Busch. Subjective evaluation of media consumer vulnerability to fake audiovisual content. In *QoMEX*, pages 1–6, 2019.

-
- [39] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch. Fake face detection methods: Can they be generalized? In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–11, 2018.
- [40] Hyeonwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Trans. Graph.*, 37(4), July 2018.
- [41] Youngshin Kim, Jaekeun Na, Seongbeak Yoon, and Juneho Yi. Masked fake face detection using radiance measurements. *J. Opt. Soc. Am. A*, 26(4):760–766, Apr 2009.
- [42] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *CoRR*, abs/1812.08685, 2018.
- [43] Pavel Korshunov and Sébastien Marcel. Deepfake detection: humans vs. machines. *CoRR*, abs/2009.03155, 2020.
- [44] T. Hoang Ngan Le, Keshav Seshadri, Khoa Luu, and Marios Savvides. Facial aging and asymmetry decomposition based approaches to identification of twins. *Pattern Recognition*, 48(12):3843–3856, 2015.
- [45] M Lee. Deepfake salvador dali takes selfies with museum visitors. *The Verge*, 2019.
- [46] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [47] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2020.
- [48] Siqi Liu, Baoyao Yang, Pong C. Yuen, and Guoying Zhao. A 3d mask face anti-spoofing database with real world variations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.
- [49] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

- [50] William Lotter, Gabriel Kreiman, and David D. Cox. Unsupervised learning of visual structure using predictive generative networks. *CoRR*, abs/1511.06380, 2015.
- [51] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva. Incremental learning for the detection and classification of gan-generated images. *arXiv preprint arXiv:1910.01568*, 2019.
- [52] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, 2019.
- [53] Wesley Mattheyses and Werner Verhelst. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66(Supplement C):182–217, 2015.
- [54] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues. *CoRR*, abs/1812.08247, 2018.
- [55] M. Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.
- [56] M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley [from the field]. *IEEE Robotics Automation Magazine*, 19(2):98–100, 2012.
- [57] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019(5):532–1, 2019.
- [58] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *CoRR*, abs/1906.06876, 2019.
- [59] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos, 2019.
- [60] T. Nielsen, A. Khodabakhsh, and C. Busch. Unit-selection based facial video manipulation detection. In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2020.
- [61] Nicklas Overgaard, Ctirad Sousedik, and Christoph Busch. Eulerian video magnification for fingerprint liveness detection. *NISK Journal*, 2014.
- [62] OM Parkhi, A Vedaldi, and A Zisserman. Deep face recognition. pages 1–12. British Machine Vision Association, 2015.

-
- [63] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. Deepfacelab: A simple, flexible and extensible face swapping framework. *CoRR*, abs/2005.05535, 2020.
- [64] K. B. Raja, R. Raghavendra, and C. Busch. Video presentation attack detection in visible spectrum iris recognition using magnified phase information. *IEEE Transactions on Information Forensics and Security*, 10(10):2048–2056, 2015.
- [65] C. Rathgeb, P. Drozdowski, D. Fischer, and C. Busch. Vulnerability assessment and detection of makeup presentation attacks. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2020.
- [66] Christian Rathgeb. Prnu-based detection of facial retouching. *IET Biometrics*, 9:154–164(10), July 2020.
- [67] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR*, abs/1803.09179, 2018.
- [68] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE ICCV*, pages 1–11, 2019.
- [69] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *CVPR Workshops*, June 2019.
- [70] Mehdi S. M. Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [71] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, June 2015.
- [72] N. Srinivas, G. Aggarwal, P. J. Flynn, and R. W. Vorder Bruegge. Analysis of facial marks to distinguish between identical twins. *IEEE Transactions on Information Forensics and Security*, 7(5):1536–1550, 2012.
- [73] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13, July 2017.

- [74] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR'16*, pages 2387–2395, June 2016.
- [75] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 716–731, Cham, 2020. Springer International Publishing.
- [76] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.
- [77] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020.
- [78] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [79] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces, 2020.
- [80] Wei Wang, Xavier Alameda-Pineda, Dan Xu, Pascal Fua, Elisa Ricci, and Nicu Sebe. Every smile is unique: Landmark-guided diverse smile generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [81] Yaohui Wang and Antitza Dantcheva. A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes. In *FG'20, 15th IEEE International Conference on Automatic Face and Gesture Recognition, May 18-22, 2020, Buenos Aires, Argentina.*, Buenos Aires, Argentina, May 2020.
- [82] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph.*, 31(4), July 2012.

- [83] J. Yang, Z. Lei, S. Liao, and S. Z. Li. Face liveness detection with component dependent descriptor. In *2013 International Conference on Biometrics (ICB)*, pages 1–6, 2013.
- [84] Ren Yang, Mai Xu, Zulin Wang, and Tianyi Li. Multi-frame quality enhancement for compressed video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [85] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019.
- [86] Ning Yu, Larry P Davis, and Mario Fritz. Attributing fake images to gans: Analyzing fingerprints in generated images. 2018.
- [87] Dehai Zhang, Da Ding, Jin Li, and Qing Liu. Pca based extracting feature using fast fourier transform for facial expression recognition. In *Transactions on engineering technologies*, pages 413–424. Springer, 2015.
- [88] Li Zhang, KengTeck Ma, Hossein Nejati, Lewis Foo, Terence Sim, and Dong Guo. A talking profile to distinguish identical twins. *Image and Vision Computing*, 2014.
- [89] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE*, pages 1831–1839, 2017.

Part II

Viewers' Vulnerability

Chapter 5

Article 1: A Taxonomy of Audiovisual Fake Multimedia Content Creation Technology

A. Khodabakhsh, C. Busch and R. Ramachandra, "A Taxonomy of Audiovisual Fake Multimedia Content Creation Technology," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, 2018, pp. 372-377.

5.1 Abstract

The spread of fake and misleading multimedia content on social media has become commonplace and is effecting society and its decision procedures negatively in many ways. One special case of exploiting fake content is where the deceiver uses the credibility of a trustworthy source as the means of spreading disinformation. Thanks to advancements in technology, the creation of such content is becoming possible in audiovisual form with limited technical knowledge and at low cost. The potential harm of circulation of these content in media calls for the development of automated detection methods. This paper offers a categorization of such fake content creation technology in an attempt to facilitate further study on generalized countermeasures for their detection.

5.2 Introduction

Consumption of digital media and its impact on decision procedures (e.g. elections) has reached a majority-owned relevance over traditional media (e.g printed

newspapers) in our world of ubiquitous information devices (Smartphone, tablets). Along with that cultural change, we must accept for the consumed content an inherent loss of data authenticity. The lack of proper fact-checking and third-party filtering on these platforms compared to traditional media resulted in the prevalence of misinformation and disinformation on these media (3). The spread of fake content can have a long-lasting impact on individuals opinions even after presentation of factual information (17).

One special case of fake content is where a deceiver uses the identity of another person (e.g. an authority figure) to disseminate false information, taking advantage of his/her credibility. Recent advancements in technology made it possible to create such content in audiovisual form (Fig. 5.2(i)) (13, 26), using commodity devices, and at low cost. A demonstration of existing technologies has been made available online for the purpose of public awareness: <http://futureoffake.news.com>.

These content are of special importance as talking faces are a natural way of communication for humans, and are preferred to other forms of communication. Furthermore, despite considerable progress on detection of fake textual content (23), very little effort has been directed to protect consumers from fake multimedia content. On the other hand, manual detection is very costly and the capacity of authentication can be out-competed by the mass of user-generated content. "Personation" is defined by the *Oxford English Dictionary* as "The action of assuming a character, or of passing oneself off as someone else, esp. for fraudulent purposes"¹. In the context of this study, audiovisual personation can be described as any attempt to assume the identity of another person in a audiovisual form, with intent to deceive. The cases of convincing personations in history have been limited to people with natural similarity (e.g. the actor Clifton James, who resembled General Montgomery in a deception mission in World War II (8)). However, as technology advances, a wide range of virtual and artificial personation techniques are becoming available, and examples of their use can be found in many real-life applications.

For personations to be successful in deception, the created content should be of high quality to pass the multimodal judgment of naive media consumers in naturalness and similarity of speech, appearance, and behavior. As a result, they should be based on a good understanding of the human perception of reality and identity. A notably related concept is the uncanny valley (15) hypothesis. This hypothesis states that after a specific point, the more an artificial entity resembles a human outlook and behavior, the presentation will elicit a more negative emotional response

¹"personation, n." OED Online. Oxford University Press, June 2017. Web. 21 December 2017.

from the observer. Nevertheless, despite the difficulty and expenses of climbing up again from the depth of the uncanny valley, a vast amount of effort has been dedicated to the creation of realistic artificial humans, and many instances of artificial entities have achieved realism in the sense of being indistinguishable from reality to unsuspecting humans. The Hollywood industry with its need for realistic yet low-cost animated scenes stimulated significant innovation in this domain over recent years.

This article proposes categories to group existing technologies for the creation of plausible audiovisual personation content with the goal of providing a comprehensive overview of deception attempts and creating a ground for the development of generalized detectors.

The rest of this paper is organized as follows: Section 5.3 describes the technologies and the motivation behind the development of these technologies. Section 5.4 describes briefly the existing detection methods. Section 5.5 summarizes the study and discusses its implications, and finally section 5.6 will conclude and describe future work.

5.3 Personation Methods

The technology for generation of artificial lifelike human appearance is advancing with the goal of creating the experience of submersion and a greater degree of presence and natural interaction with the artificial entity. The consumer may be aware of the unreality of the entity, however, the apparent realism makes it cognitively possible to have suspension of disbelief. The artificial entities may be digital (e.g. an avatar), or physical (e.g. an android robot). These technologies have applications in communication (e.g. telepresence, customer service, advertisement), training (e.g. education, simulation), health-care (elderly care, physical and psychological therapy), assistance (companionship, museum guides, office robots, software office assistants), entertainment (e.g. cinematography, satire, video games, stage shows), and covert disinformation attacks. Based on the application, the resulting systems can create a passive representation, or be interactive.

For the purpose of this article, the technologies can be categorized by the point of application in the consumption process of the audiovisual content. This is motivated by the difference in technical demands of content generation, and thus detection approaches at each application point. Fig. 5.1 shows the lifetime of an audiovisual content. A video depicting a person is recorded by a camera, and after traveling the network (including storage devices), it is shared by a publisher and displayed to the consumer. Given an audiovisual representation of a person, the points of suspicion are a false presentation at the camera, digital tampering of the

recorded video, or replacement with a computer generated (CG) counterpart.

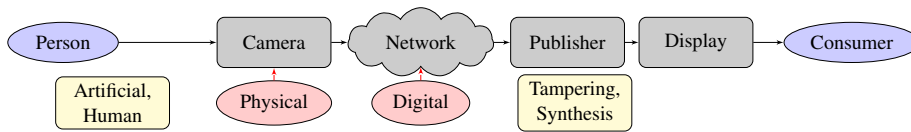


Figure 5.1: Points of vulnerability of video transmission medium to deception attempts.

Based on these points of vulnerability and different modalities of the audiovisual signal, the following categorization is used to cover existing personation technology which will be discussed first for visual and subsequently for audio content:

5.3.1 Visual

A visual personation requires naturalness and similarity to the target person in appearance and behavior. The behavior of the personation can be modeled and applied independently of the appearance, and thus it is described separately.

Physical

A physical visual personation requires a convincing appearance of a person or an object with the resemblance of the target person. This item can be created as an artist's impression, or be created using the scan or cast of the face of a person.

Artificial Artificial visual personation can be described as any physical artifact (i.e. movable dummies and fleshly robots) that can convincingly resemble the target person in appearance and ability to move. Due to the complexity of the human facial muscle configuration and movements, it is not possible to puppeteer the artifact mechanically. Thus the artificial personation devices are usually operated by robots. Such robots are called androids and can have a photorealistic resemblance to the target person thanks to realistic skin and hair like material used in their production. These androids are mainly developed by robotics community for natural human-robot interactions, and have applications ranging from entertainment to education and health-care. Notable examples are animatronics of US presidents at the hall of presidents in the Walt Disney world resort², and Geminoid robots (Fig. 5.2(a)) created by Hiroshi Ishiguro at the Intelligent Robotics Lab at Osaka University (16).

The facial movements are typically modeled by motors acting facial action units on the face. Due to mechanical limitations, these robots have jerky movements and their behavior is easily detectable as unnatural. To avoid these limitations in

²<http://www.popularmechnics.com/technology/robots/a23699/robot-presidents-disney/>

facial motion, some androids use a screen as a face (e.g. Life Imaging Projection System aka L.I.P.S (Fig. 5.2(b)))³. Another notable example is the shape-shifting robot WD-2, which can replicate the face of a person based on the 3D scan of his face. The high cost of building and the unnatural movements limits the application of these androids in personation attempts.



(a) The Geminoid (16) (left) and Hiroshi Ishiguro (right) (b) Life Imaging Projection System³ (14) (c) Natalie Portman (left) and Keira Knightley (right) (d) George W. Bush (left) and Steve Bridges (right) (6)



(e) Ezzat et al. (10) (f) Aimi Eguchi⁴ (g) Keanu Reeves (left) and his digital double (right) et al. (22) (18) (h) Seymour et al. (26) (i) Thies et al. (25) (j) Suwajanakorn et al. (25)

Figure 5.2: Illustration of different visual personation technologies.

Human This category is the oldest personation method that has been used for deception. The cost of personation varies depending on the apparent natural similarity (i.e. biometric twins) of the target person and the personator. In case of lack of sufficient resemblance, the personator can use heavy or prosthetic makeup and masks to change his appearance⁵. The result is often of sufficient similarity to be recognized as the target person. Many applications of this technique exist and are mainly around the entertainment industry, such as “fake shemps”⁶ and impersonators.

An example for identical twins is Leslie H. Gearren⁷ acting for Linda Hamilton in “*Terminator 2: Judgment Day*” as her double. Natural similarity of actors Keira

³<https://news.yale.edu/2001/03/19/heads-will-be-talking-yales-digital-media-arts-center>

⁴<http://newsfeed.time.com/2011/06/24/japanese-scientists-build-a-perfect-and-fake-pop-star/>

⁵<https://www.boredpanda.com/game-of-thrones-make-up-art-transformation-paolo-ballesteros/>

⁶<https://web.archive.org/web/20071115162315/http://en.allexpert.com/q/Horror-Film-2863/Horror-Film-Staff.htm>

⁷<http://www.imdb.com/name/nm0357696/bio>

Knighley for Natalie Portmans character has also been used in “*Star Wars: Episode I – The Phantom Menace*” (Fig. 5.2(c)). Many examples for prosthetic makeup exist in satire (e.g. Steve Bridges as George Bush (Fig. 5.2(d)) (6)). Using humans for personation has been done for political purposes too. The best-documented example of political decoys is personation of Bernard Montgomery by Clifton James (8). This method of personation is surprisingly effective in convincing people. The main advantage of this technique compared to the other methods is complete naturalness of the muscle control of the resulting personation.

The impersonator needs to learn the gestures and mannerism of the target person in order for the personation to be convincing. For such applications, actors are usually the best choice as of their experience in realistic mimicking of behavior. This will provide similarity on top of realism of their movement.

Digital

Using computer algorithms, a video of a talking face can be a digitally modified copy or be completely synthetic. Different technologies evolved for the creation of animated faces based on these two categories for applications such as virtual actors and automated dubbing.

Tampering An authentic video of a person can be manipulated and modified to change the content of the recording. This can be done manually using video editing software (Fig. 5.2(f)) (e.g. splicing and morph cut in Adobe Premiere) or automatically using techniques such as active appearance models (AAM) (10). These changes can require signal processing steps minimally, as of removing a single word manually and morphing the before and after images, or extensively, as for automatic concatenation of visemes in audiovisual text-to-speech (AVTTS).

One of the earliest examples of automatic tampering is Video Rewrite system (5). Since these methods produce the original frames of the recorded video or their morphed copy, the result is generally photorealistic and similar to the target person. However, the realism of dynamics is limited by the amount of variability in the existing footage. The more tampering and morphing happens between incoherent frames, the more temporal artifacts will be visible in the resulting video. This method has been successfully used for AVTTS and achieved high realism scores in subjective tests (Fig. 5.2(e)) (10). The limitation of this method is that it requires a long expressive recording with consistent light and a fixed pose for desirable results. However, the capture and animation process is much simpler and computationally cheaper compared to synthesis and results in higher quality videos.

Synthesis The high computational cost and difficulty in 3D modeling of human facial details and rendering of digital characters, as well as the extreme sensitivity of humans to details of facial texture and motion, makes the generation of synthetic faces hard. However, due to the flexibility these models provide for synthesis in different lighting conditions, from different angles, and with the minimal amount of capture needed compared to tampering techniques, there has been a lot of interest and effort in creating realistic synthetic faces (4). The existing technology has been used to synthesize faces of sufficient realism by the movie industry in the past decade (Fig. 5.2(g)). However, the realism of synthesis is a function of computational costs such as the number of polygons and reflection and shading resolution, making the technology limited to high budget non-realtime applications. Nevertheless, in some cases, it may be possible to reduce computational costs by only synthesizing the face partially and splicing it over some existing footage (4).

The advancements in computational graphics and graphics processing unit capacity slowly bring the possibility of photorealistic 3D rendering to real-time and on personal computers (Fig. 5.2(h)) (22). The capture procedure of faces usually requires the use of multiview face capture systems (9). It has also recently become possible to infer the high-resolution texture of faces using a single low-resolution photo of the face (21). Morph target animation can be used along with facial rigging to animate the face mesh.

These models can present very high photorealism (22) thanks to methods for perfecting the details (e.g. skin reflectance modeling) (9). These synthetic faces have also found applications in AVTTS (10) and robotics (2).

Animation source

The aforementioned physical and digital artificial entities have interfaces for animation (e.g. based on FACS). To answer how to animate these characters using their interface, there are several solutions developed and are described in this section.

Motion capture Motion capture (mocap) technology has advanced tremendously recently, and many markerless mocap systems have been developed with high accuracy (7). This enables the actor or impersonator to control the actions of the virtual or artificial character with ease and accuracy. Based on the resolution of the motion capture device, the movements can be indistinguishable from real movements. These systems have been applied by the movie industry as well as for virtual reality and telepresence applications.

Synthesis In many cases, it is not possible to entirely rely on motion capture for animation of the characters. Examples include video games and autonomous

robots. Early systems were animated using predefined actions that were coded manually (19). Example of these systems are the terminal-analog systems that were early attempts to animate AVTTS characters. There have been attempts to animate characters automatically using models such as hidden Markov models (HMMs). These models can be trained on existing footage of the target person, and used for the synthesis of proper behavior in new situations. Another type of synthesis is the use of text or speech features to animate the character in the video (Fig. 5.2(j)) (25). These systems have applications in AVTTS as well as automated dubbing.

5.3.2 Auditory

Humans rely on dynamics and high-level auditory features for recognizing people, and vocal-tract similarity does not affect the human perception as much as the dynamics of speech. The resulting situation requires realistic virtual and physical artificial beings to have natural sounding personations, as well as having similarity in high-level features.

Physical

Physical methods rely on physical entities for generation of personation speech. These can be broadly categorized into artificial and living.

Artificial A speech personation audio can be generated using biomechanical modeling of human vocal apparatus (11). These systems are hard to develop as the vocal apparatus of humans is not visible and not measurable as easily as faces. The limitations are similar to those of artificial visual personation technologies. The technology has not reached maturity for use in personation.

Human Professional impressionists can successfully imitate the voice of many different people. This ability shows that no alteration to the vocal tract is needed, and impersonation is an ability that can be learned by practice. Impersonations usually mimic the mannerism of the target person and try to adjust their voice dynamics to match that of him/her. The resulting speech is convincingly similar and sounds natural to the human ear. Impersonation is usually used by impressionists for entertainment, however, instances of their use have been recorded for personal and political gains. A notable example is the personation of President Truman's voice on the telephone to persuade foreign leaders to vote in particular ways at the United Nations⁸.

⁸<http://www.trumanlibrary.org/oralhist/wright.htm>

Digital

Speech signal can be manipulated and generated digitally as well. Many different systems have been developed with high naturalness and intelligibility for real-life applications. Similar to digital visual personation techniques, these techniques are also categorizable to tampering and synthesis.

Tampering A synthetic speech can be generated by concatenation of speech samples from a target speaker. The concatenation footprints can be minimal, in the case of removal of a word from an audio, or audible when extensively done (e.g. diphone synthesis). The automated systems generating this kind of synthetic speech are typically called unit-selection speech synthesis systems (12). Due to the use of natural human voice for the generation of the synthetic speech, the resulting audio has very natural human-like sound, resembling the voice of the target speaker. However, due to the collection of each unit from a different context, the high-level features such as style and intonation of speech are often lost. Some of these artifacts can be corrected using post-processing of the pitch and duration of phonemes after synthesis (e.g. using Pitch Synchronous Overlap and Add (PSOLA)). Unit-selection systems are the most used type of synthetic speech and are employed in many real-life applications in our everyday lives.

Synthesis Many technologies for speech synthesis rely on models of speech. These systems include but are not limited to: Statistical speech synthesis (SSS) (29), Articulatory speech synthesis, and voice conversion (1). The most used type of synthetic speech generation systems is SSS. These systems model the distribution of speech features using HMMs in a similar manner to speech recognition systems, and later synthesize speech using parameter generation algorithm. The resulting speech lacks the naturalness of the unit-selection systems, but has more cohesion, is more flexible, and can model the high-level and dynamic features of the speech to some extent. The similarity is also high as the synthesis parameters are generated from the distribution of speech features extracted from genuine speech. The possibility of speaker adaptation on these systems makes them a good candidate for automated personation attempts.

Another type of synthetic speech that requires attention is voice conversion. Given an audio signal from a target speaker, the system can learn a mapping from feature space of the personator to that of the target speaker. This model can later be used to convert the voice of a personator to the target speaker's voice. As of now, these systems lack naturalness in their generated audios but may improve tremendously as the technology advances.

Wavenet (27) represents another interesting type of speech synthesis system that

relies on waveform synthesis rather than feature synthesis. The clear naturalness and resemblance of human speech using waveform synthesis is promising and can pass human judgment.

Animation source

Digital speech synthesis systems usually get a text as an input for generating the output audio. The text may be accompanied by affective information as well. The exception to this is the voice conversion systems that act in a similar manner as the motion capture systems.

5.3.3 Combinations

Multimodal personations require combination of visual and auditory modalities. This is challenging, as humans rely on both visemes and phonemes to understand speech, and thus are extremely sensitive to small disaccords between modalities. The technology of choice for each modality can vary depending on the application of the system. Of course, these techniques can be combined on each modality as well, producing “*hybrid*” personations. This can be done to take advantage of their fusion to reduce the need for extra modeling, avoiding artifacts, or reducing computational costs. Obfuscation may also be employed concurrently to achieve the same goals.

5.4 Detection Techniques

Different detection technologies are being developed stemming from fields of digital video forensics, biometrics, and fake news detection. Presentation attack detection technologies address the detection of physical attempts while tampering detection and computer-generated detection technologies provide solutions for the detection of digital tampering and synthesis attempts respectively (Fig. 5.1). Despite considerable achievements, to date, no generalized method for automated detection of such content has been developed (20, 24, 28).

A different approach in development is to utilize contextual and style-based information as well as relying on external sources of knowledge for verification of veracity of a piece of information (23). Hitherto, the task of personation detection remains mostly a manual endeavor.

5.5 Discussion

In this study, we attempted to categorize all the existing applicable technologies for audiovisual personation. The list of personation technologies can be summarized in Table 5.1. It can be seen that the visual, auditory, and animation factors of a given entity can each one be created by a human, by modifying an existing

record, or by synthesis from a model, and done independently of one another. This classification simplifies the description of any personation technology as well as the formulation of weaknesses and strengths of these methods.

Table 5.1: Summary of the different personation technologies.

		Visual	Auditory	Animation
Physical	Artificial	Androids - Screens	Biomechanical - Loudspeaker	-
	Human	Twins - Prosthetic makeup	Impersonation	Impersonation
Digital	Record-based	Image-based (Tampering)	Unit-selection (Tampering)	Cloning
	Model-based	CG (Synthesis)	SSS (Synthesis)	Autonomous

An estimation of the detection difficulty of personation attempts for viewers is given in Table 5.2. Given the difficulty of detecting record-based models, it can be concluded that major risks of existing technologies are presented by these personations. A lower level of risk arises from modeling of reality by humans and model-based systems.

Table 5.2: An estimate of detection difficulty of personation attempts for humans, along with generation cost approximation. (E: Easy, M: Moderate, H: Hard) Naturalness and similarity are estimated for Visual (V), aniMation (M), and Auditory (A) aspects.

		Naturality			Similarity			Creation Cost	
		V	M	A	V	M	A	Model	Prod.
Physical	Artificial	H	E	E	H	E	E	High	Low
	Human	H	H	H	M	H	H	Mod	Low
Digital	Record-based	H	H	H	H	M	H	Low	Low
	Model-based	M	M	M	H	H	H	Mod	Low

5.6 Future work

In this study, different techniques that are usable for personation are listed and explained. Future work consists of studying the risk assessment of these attacks and applicable detection technologies. Creation of a dataset based on this classification and objective evaluation of the performance of different detectors would be the next step.

References

- [1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *ICASSP'88*, pages 655–658 vol.1, Apr 1988.
- [2] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström. Furhat: A back-projected human-like robot head for multiparty human-machine interaction. In *COST'11*, pages 114–130, Berlin, Heidelberg, 2012. Springer-Verlag.
- [3] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. NBER Working Papers 23089, National Bureau of Economic Research, Inc, 2017.
- [4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99*, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.
- [5] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *SIGGRAPH '97*, pages 353–360, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.
- [6] E. Bumiller. A new set of bush twins appear at annual correspondents' dinner. *The New York Times*, page 1, 2006.
- [7] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, 33(4):43:1–43:10, July 2014.
- [8] M. E. Clifton-James. *I was Monty's Double*. Panther Books, 1958.
- [9] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *SIGGRAPH '00*, pages 145–156, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.
- [10] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *SIGGRAPH '02*, pages 388–398, New York, NY, USA, 2002. ACM.
- [11] K. Fukui, E. Shintaku, M. Honda, and A. Takanishi. Mechanical vocal cord for anthropomorphic talking robot based on human biomechanical structure. *The Japan Society of Mechanical Engineers*, 73(734):112–118, 2007.
- [12] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP'96*, volume 1, pages 373–376 vol. 1, May 1996.

-
- [13] Z. Jin, G. J. Mysore, S. Diverdi, J. Lu, and A. Finkelstein. Voco: Text-based insertion and replacement in audio narration. *ACM Trans. Graph.*, 36(4):96:1–96:13, July 2017.
- [14] G. Lucas and R. McCallum. Star wars: Episode I – the phantom menace, 1999.
- [15] M. Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.
- [16] S. Nishio, H. Ishiguro, and N. Hagita. Geminoid: Teleoperated android of an existing person. In *Humanoid robots: new developments*. InTech, 2007.
- [17] B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, Jun 2010.
- [18] J. Oreck. The matrix reloaded: Unplugged, 2004.
- [19] F. I. Parke. Computer generated animation of faces. In *Proceedings of the ACM Annual Conference - Volume 1*, ACM '72, pages 451–457, New York, NY, USA, 1972. ACM.
- [20] R. Ramachandra and C. Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Comput. Surv.*, 50(1):8:1–8:37, Mar. 2017.
- [21] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li. Photorealistic facial texture inference using deep neural networks. *CoRR*, abs/1612.00523, 2016.
- [22] M. Seymour, C. Evans, and K. Libreri. Meet mike: epic avatars. In *ACM SIGGRAPH 2017 VR Village*, page 12. ACM, 2017.
- [23] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36, Sept. 2017.
- [24] K. Sitara and B. Mehtre. Digital video tampering detection: An overview of passive techniques. *Digital Investigation*, 18(Supplement C):8–22, 2016.
- [25] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13, July 2017.
- [26] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR'16*, pages 2387–2395, June 2016.

- [27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.
- [28] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, 66(Supplement C):130–153, 2015.
- [29] H. Zen, K. Tokuda, and A. W. Black. Review: Statistical parametric speech synthesis. *Speech Commun.*, 51(11):1039–1064, Nov. 2009.

Chapter 6

Article 2: Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content

A. Khodabakhsh, R. Ramachandra and C. Busch, "Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content," 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), Berlin, Germany, 2019, pp. 1-6.

6.1 Abstract

Advancements in computer graphics and artificial intelligence have facilitated the generation of graphics faking human faces and this technology can be misused for personal or political gain. Media consumers are exposed to hours of audiovisual content daily and their vulnerability to fake audiovisual content has not been fully studied and understood; this is in marked contrast to the fact that automated fake content generation techniques are readily accessible to the public. A first step to address this vulnerability is to study the effectiveness of existing methods in passing human judgment. To this end, we examined the performance of 30 participants in the detection of 48 real and fake videos. The fake videos were sourced from six different methods of generation and were collected from a public video sharing website¹, ranging from prosthetic makeup to Deepfakes. Our results show that the participants failed to detect two different types of fake videos. However, participants detection performance improves when they have prior knowledge

¹<https://www.youtube.com/>

about the displayed individual in the form of a biometric reference video (introducing the individual and its behavior) that can be referenced during the test.

6.2 Introduction

The consumption of audiovisual content is on the rise due to the increase in network speed and the richness and appeal of such content compared to traditional forms of media. Further, the consumption of media from free and unreliable sources such as social media channels has increased dramatically in recent years. These two factors combined can cause a massive proliferation of fake news in audiovisual format. This is alarming, because in contrast to text-based fake content detection, audiovisual fake content detection is in its infancy, and only few automatic detection methods are in place with limited applicability (14). An important case of audiovisual content is the case of talking faces, being a usual part of online videos due to it being the most natural way of communication between humans. The generation of videos of fake faces has become possible thanks to advancements in computer graphics and more recently in artificial intelligence. Many methods claim to have video-realism, some of which are available for public use. One recent example of using fake faces is the Xinhua agency's AI presenter².

Humans are shown to be vulnerable to digitally manipulated images (12). In 2012, Farid et al. (4) measured the performance of humans in detecting fake face images generated using computer graphics. Their results show above chance detection accuracy in different resolutions and compression settings. In similar studies (2, 3), authors try to pinpoint contributing factors in detection such as positioning of illumination sources and shadowing, color, and partial occlusion of the face. However, in a more recent study in 2018, Rossler et al. (11) studied the detection performance of humans on fake face images extracted from a specific fake video generation algorithm. Their results show that human detection accuracy can be as low as random guessing after video refinement and compression. This study tries to provide insights into the open question, can people distinguish real videos from fake ones? The results from this study's simulated real-life scenario will shed new light on media consumer vulnerabilities; it will also provide a review of the effectiveness of new and traditional audiovisual fake face generation methods in the current point in time.

The rest of this paper is organized as follows: Section 6.3 describes the experimental methodology and includes details on the dataset, the test protocol, and the test setup. Section 6.4 discusses the results of this study and then Section 6.5 presents our conclusions and proposals for future work.

²<https://www.bbc.com/news/technology-46136504>

6.3 Data and Methodology

In this study, a real (a.k.a bona fide) video is defined as a continuous recording of the target individual without any modification that changes the representation or appearance of that target individual and the content of the utterance. Alternatively, a fake video is anything to the contrary and can be described as either impersonated, manipulated, or synthetic media related to the target individual. The target individual is the natural person whose appearance is used for generation of the fake video.

To reach the objective of this study, a set of videos were required that represents the status of today's technology in fake video generation, and a test setup that simulates real-life video encounters.

6.3.1 Dataset

The scenario in this study is limited to continuous scenes of talking heads. As to study the effect of visual and auditory features rather than the textual content of the videos, only short utterances were considered for this study. A dataset consisting of 48 videos, each five seconds in duration, were manually collected from YouTube. The videos were selected such that they have a size of at least 640×480 pixels, and were manually screened for sufficient lighting and frontal face visibility conditions. The videos are selected such that they do not contain any meaningful uttered sentence, avoiding leakage of information about the real- or fake-ness of the video.

Half of the videos fitted the criteria of "fake", and categorized to six categories based on the technique used to generate them, meanwhile the other 24 represent the "real" video control set. Due to the very limited number of actual fake videos matching the selection criteria, the selected fake material represents an extent of videos that can be used as a fake video. Following the taxonomy introduced in (7), the fake categories are as follows:

1. Physical

- (a) Look-alike: The individual in the video is a look-alike of the target individual. The voice may not match the target individual.
- (b) Prosthetic Makeup: The individual in the video wears prosthetic makeup and impersonates the target individual.

2. Digital

- (a) Computer Graphics Imagery (CGI): The scene has been generated us-

ing CGI. The voice may come from an impersonator or the target individual.

- (b) Interframe forgery (Morph-cut): To alter the spoken audio content, the video has been cut and rejoined in a seamless manner, by using the Adobe Premiere Pro Morph-cut³ video transition.

3. Hybrid

- (a) Face CGI: This technique is similar to the CGI technique, with the difference in that only the face or a part of the face was synthesized and then overlaid on the recorded footage.
- (b) Face GAN: Similar to Face CGI, only the face is replaced. Yet the synthetic face is generated by Generative Adversarial Networks (GAN) using Faceswap⁴ or an alternative open-source application based on the same concept.

The selection process chose the most video-realistic examples encountered from each fake category, fitting the overall criteria of duration and quality. The chosen videos in each category were further filtered for video-realism by three colleagues in our research lab. The selected videos partially overlap with the FFW dataset(8). The sources of the videos guaranteed their status as fake. Facial regions of all the fake videos are depicted in Figure 6.1. For the control set, 24 videos were randomly selected from the VoxCeleb(10) dataset after filtering those with regards to the same duration and quality criteria.

To address the effect of having a biometric reference included in the test, each video in the real and fake categories was paired with a supporting biometric reference from the target individual. The selection criteria for biometric reference videos were the same as for the “real” category and partially selected from VoxCeleb dataset. The participants’ detection performance was first stabilized by using a short mock test that was based on five pairs of video and biometric references that are separate from the experimental/control datasets. The target individuals in the videos were adults who were either celebrities or political personalities of varying in age and gender.

To eliminate low-level clues that participants might use to identify the fake videos, the following metrics were measured to assure an overlapping distribution between both sets: head size, head pose, image and facial quality. In both real and fake sets the average head size was ≈ 128 pixels, average BRISQUE(9) was $\approx 36\%$, and

³<https://helpx.adobe.com/premiere-pro/using/morph-cut.html>

⁴<https://github.com/deepfakes/faceswap>



Figure 6.1: The faces in the six categories of fake faces. Going left to right, the columns correspond to the following categories in order: Look-alike, Prosthetic Makeup, CGI, Morph-cut, Face CGI, and Face GAN.

average face quality(1) was $\approx 61\%$. The distribution of facial pose in both sets also has a high overlap. The list of videos in the dataset are made available online⁵.

6.3.2 Protocol

The aim of this test is to measure the following:

- Participants (i.e. media consumers) performance in the detection of the most video-realistic fake samples in each category.
- Effect of presence of a biometric reference upon the detection performance.
- Effect of familiarizing the participants with different categories of fake content with a guide on the shortcomings of each fake face generation method on their detection performance.
- Effect of prior knowledge of the target individual on the detection performance.

⁵<http://ali.khodabakhsh.org/fake-faces-for-subjective-testing/>

- Possible correlations between demographic information and subjective detection performance.
- Common clues used by participants.

The test aims to have a measurement corresponding to the real-life scenario and utilizes a web-based interface that participants access through their personal multimedia device (limited to devices with a large display, e.g. laptop or tablet). To make sure the participants can use both modalities, they were given guidelines for screen and audio adjustment.

The experiment sessions were split into five parts. The first part was used to briefly explain the test and also collect participants' demographic information (age, gender, education, and occupation). In addition to this, the existence of any visual deficiency is probed, along with a question regarding the expected expertise of the participant in the task.

The second part consists of a familiarization step, where fake videos are described and a set of videos depicting examples of each category is shown to the participants. To measure the effectiveness of familiarization, the familiarization page is shown before the test in half of the population, and after the test in the other half.

The third step corresponds to a mock test with a fixed order. This step tries to stabilize the performance of the participants and to reduce any inconsistency in their performance caused by the learning process. This step follows the same set of questions as the rest of the test. The answers for these videos were to be discarded in the analysis.

The fourth step is the main part of the test, and was organized as follows: a video is shown to the participant, sometimes along with a biometric reference, and the participant is asked to answer a set of multiple choice questions about the video in question. The questions address the decision of the participant on the video being real or fake and ask if the participant knows of the target individual. Furthermore, the participant is asked about the main clue that led to their decision to be selected from a list of clues, with the option of mentioning additional clues in a comment box. This process is then repeated for the remaining 47 videos. To avoid any effect of ordering in the test, the videos were shown in a randomized order, and the biometric reference video appeared randomly in half of the videos. The participants were also allowed to have a *no answer* choice in the questions, if they were uncertain of their response.

Finally, a feedback page is presented to the participants that provides a visualization of their performance to reward them by increasing their awareness of their

mistakes and vulnerabilities.

6.3.3 Test Setup

The test was implemented using the online survey tool Limesurvey(13). The participants were invited by an email that included a token which limited each participant to a single test trial. The participants were able to stop the test at any point and resume later. The participants were asked to take the test on a large display with adequate brightness at arms distance, and have their audio on, and to be connected to a high-speed internet connection.

The first page of the test included the previously described demographic questions. International Standard Classification of Education (ISCED) 2011 was used to measure the participants' highest completed level of education and Standard Occupational Classification (SOC) System was used to classify their occupation. The participants were asked if they have any deficiencies in their vision, defined as any deficiency that had not been corrected (e.g. by corrective lenses) at the time of the test. They were also asked about their level of expected expertise in the task and given a choice between none, very low, moderate, quite high, and very high. The parameters corresponding to the ordering of videos, biometric reference, and familiarization page was randomly initialized and saved for the analysis step.

The familiarization page is now available online⁶. It contains seven videos illustrating the different fake content generation methods used in the test, along with a description of fake video categories used in this study, and the artifacts they typically create.

A typical test survey page with a biometric reference is shown in figure 6.2. The playback quality of videos were set to "medium"⁷, corresponding to 30fps, 360p videos in VP9 format for video and Opus for audio. The mock test included five videos similar to the actual test, with two real and three fake videos of which two had a biometric reference while three were presented alone.

The mock test was followed by the main test, where the 48 videos were presented in a randomized order, 24 with biometric reference and 24 without biometric reference. The time taken to answer each question set is also recorded.

6.3.4 Performance Evaluation

The performance evaluation metrics used in the experiment are from the ISO/IEC 30107-3 standard (5), they include: Attack Presentation Classification Error Rate

⁶<http://ali.khodabakhsh.org/research/fake-faces-and-fake-face-detection/>

⁷https://developers.google.com/youtube/iframe_api_reference

 Resume later Exit and clear survey

Biometric Reference:


This video is an authentic video of the person in question to be used as a reference:
(Press the button to play and wait until the end of the playback)
 (Playback is known to have issues in Google Chrome browser.)



▶▶

Video in Question:

Please answer the questions about the following video:
(Press the button to play and wait until the end of the playback)
 (Playback is known to have issues in Google Chrome browser.)



▶▶

★ Is the video in question a real recording of **Marisel Hemingway** or is it faked?

Real and Authentic	Uncertain	Fake (Impersonation, Manipulated, Synthetic, etc.)
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

★ Please select an answer to the following questions:

	Yes	Uncertain	No
Have you seen any versions of the video in question before?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Do you know of Marisel Hemingway ?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

★ Please let us know what is the **main clue** helped you with your decision? (Additional and missing ones can be described in the comment box below.)

● Choose one of the following answers

- Prior Knowledge
- Head/Face
- Background/Body
- Movements
- Words
- Audio/Voice
- Movement Synchrony
- Audio/Video Synchrony
- Uncertain
- Other

Please enter your comment here:

[Next](#)

Figure 6.2: The test interface for a sample video with a biometric reference based on the Limesurvey tool.

(APCER) and Bona Fide Presentation Classification Error Rate (BPCER). APCER measures the proportion of fake (i.e. presentation attack) videos incorrectly classified as real (i.e. bona fide), while BPCER measures the proportion of real videos mistaken for fake.

In addition, to evaluate the confidence intervals for detection accuracies, Clopper-Pearson method was used on the binomial distribution of decisions with a 95% confidence interval. For evaluating the significance of difference between distributions, two-tailed student's t-test was used with a significance threshold of 0.05 (specified otherwise). Lastly, Pearson's correlations were reported along with their confidence interval using a Student's t distribution for a transformation of the correlation, and p-values below 0.05 were considered significant(6).

6.4 Results and Discussion

The results presented in this work are based on the participation of volunteers affiliated with our campus, as well as acquaints who were interested in taking the test. During four weeks 30 people participated in the test. 60% of the participants have a master degree, while 23% have a doctorate. 77% of the participants self-identified as male while the remaining self-identified as female. 67% were employed in Computer and Mathematics, while 13% were variously employed in Education, Training, and Library services. The average age was 31.2 with a standard deviation of 7.5. The participants' average time to complete the test was 39 minutes; this corresponds to an average of 37.6 seconds per video and 4.5 minutes for familiarization.

Out of 30 participants, five had vision deficiencies; but their performance was not statistically significantly different from the performance of the rest of the experimental cohort, so their data has been included (p-value of t-test is 0.58, 0.29, and 0.66 for correct, uncertain, and incorrect choices. $n = 5$ for with and $n = 25$ for without vision deficiency.). Out of 30 participants, 18 expected to have moderate expertise and six expected to have very low expertise. The remaining six participants were equally distributed between *quite high* and *none*. The participants were of different nationalities, with 93% from Eurasia. The participants, when asked, did not mention any mismatch in presentation or low-level patterns useful for distinguishing between real and fake videos.

The participants had a below 30% BPCER and APCER in detecting real and fake videos respectively, except for the look-alike and Morph-cut categories. No statistical significance (with a 95% confidence) was observed between the other categories of the fake and average time taken to answer each question per category. Figure 6.3 shows the percentage of correct, uncertain, and incorrect identification of real

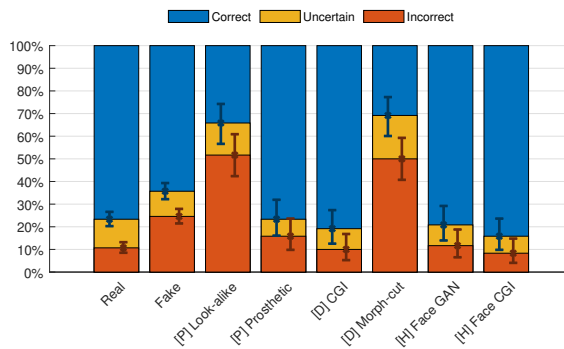


Figure 6.3: Overall choice percentages with 95% confidence intervals of the participants in the real and fake categories, along with each subcategory of fake videos. The letter before the fake category names correspond to the classification of fake videos ([P] Physical, [D] Digital, and [H] Hybrid) (7).

and fake videos, along with the performance in each fake category separately.

Figure 6.4 shows the percentages of correct, uncertain, and incorrect identification for every single video sorted by detection accuracy, along with their corresponding category. The look-alike and morph-cut videos are gathered around the left-hand side, while the other four categories are distributed in-between the real videos. A close inspection of outliers in each category shows these videos having special lighting conditions. For example, the most misclassified sample of prosthetic makeup is the face on the fourth row, second column, in Figure 6.1. The most misclassified example of CGI and Face GAN are the faces at row one column three and column six respectively. It is also interesting to observe that the percentage of uncertain answers per video never exceeds 25% even when the percentage of incorrect reaches above 50%. This implies that the participants were on average, confident of their decision in all the videos. The three videos that were classified correctly 100% of the time were of well-known political personalities (presidents of the united states) and are shown in row one column two, row two column five, and row four column six in Figure 6.1.

The most common main clues used is shown in Figure 6.5. The difference in usage shows participants relied mostly on Head/Face compared to other clues. It is also interesting to see that the distribution of clues is different in fake and real videos. In addition, some clues resulted in different performance across classes. For example, when participants mentioned movements as the main clue, BPCER was 95% while APCER was only 49%. No statistical significance (with a 95% confidence) was observed between the accuracy of detection given a specific clue

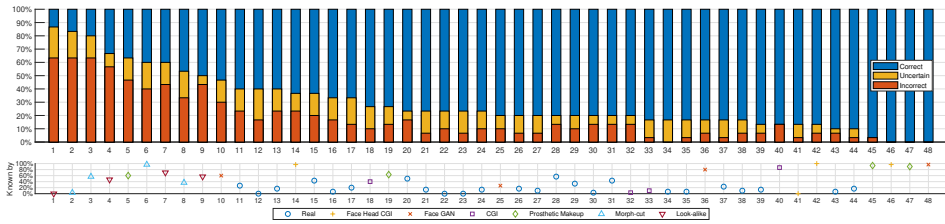


Figure 6.4: The percentage of correct, uncertain, and incorrect choices per video, sorted by the percentage of correct from low to high. The category of videos is shown in the plot below with colored markers along with the percentage of population that knew the subject in the video. Look-alike and Morph-cut samples are concentrated in the left side of the graph.

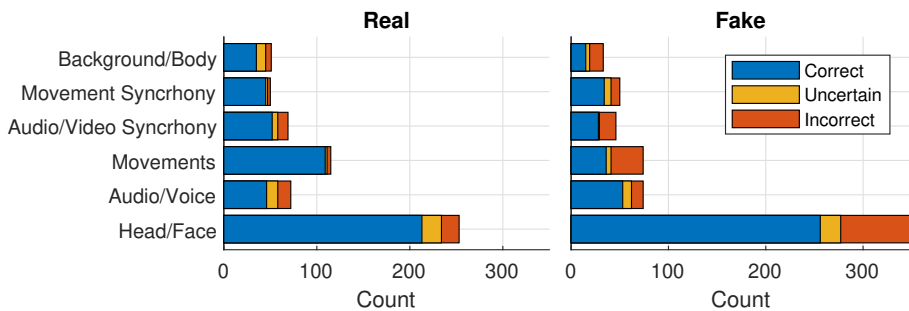


Figure 6.5: The number of correct, uncertain, and incorrect choices, given the most common main clue selected by the participants. The difference in distribution and accuracy of clues in real and fake categories are visible.

due to small sample size. To measure clue diversity per participant compared to the clue diversity in the whole group, the clue entropy is calculated. The average participant entropy was measured to be 2.36 while the total entropy was 3.12, showing that participants tended to focus on a smaller set of clues in comparison to the population.

The presence of familiarization was accompanied by a shift in the distribution of incorrect percentage towards lower values (t-test $p = 0.07$, $n = 12$ for with and $n = 18$ for without familiarization) and reduced the inter-participant variability for incorrect and uncertain responses as shown in Figure 6.6(a). Yet this reduction only caused an insignificant increase in uncertain and correct responses (t-test $p = 0.90$ and 0.60 respectively at aforementioned sample sizes). This shows that the provided familiarization oriented their decisions, yet was not effective in increasing their overall accuracy. As shown in Figures 6.6(b) and 6.6(c), Having a biometric reference shifted the distribution of incorrect percentages to lower values

(t-test $p = 0.05$, $n = 30$ for both conditions), while knowing the target individual mostly shifted the distribution of uncertain percentages in the same direction (t-test $p < 0.01$, $n = 30$ for both conditions). The distribution of correct percentages was shifted towards higher values when the target individual was known (t-test $p < 0.01$).

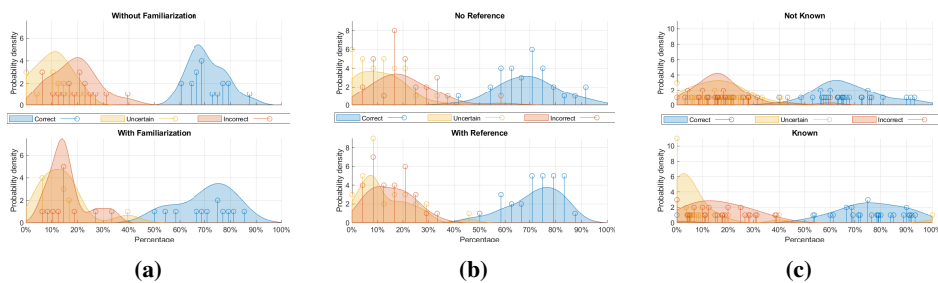


Figure 6.6: The probability density estimates of the percentage of correct, uncertain, and incorrect choices for each participant in the with and without (a) familiarization, (b) biometric reference, and (c) knowledge of the target individual scenarios along with the original distribution. A shift towards lower values is observable in the incorrect distribution in (a), while in there is (b) a shift towards lower values in the incorrect and uncertain distributions. In (c) the uncertain distribution shifted significantly towards lower values.

The following were observed between the demographic information and the performance of individual participants: Due to the small population size no significant correlation was observed comparing the level of education and gender to performance. Level of expected expertise in the task had a positive trend in comparison to the number of correct responses, yet the 95% confidence intervals for these values were overlapping. A moderate positive correlation was observed between the age and the number of incorrect answers ($p = 0.07$), simultaneously a moderate negative correlation existed between the age and the number of uncertain ($p = 0.02$), canceling the overall effect on the number of correct, as shown in Figure 6.7.

6.5 Conclusion and Future Work

We evaluated the performance of 30 participants in distinguishing fake videos from real ones using a web-based platform. 48 pair of videos were collected from an online video sharing website, 24 of which could fit the definition of fake and were generated using six different methods ranging from prosthetic makeup to Deep-fakes.

The results suggest the vulnerability of participants to the traditional methods more than the new methods, specifically to look-alikes and interframe forgery. This aligns well with the long history of use of look-alikes as fake faces, especially as

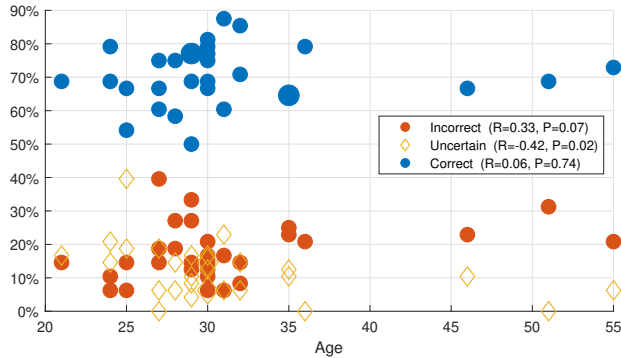


Figure 6.7: Scatter plot of percentage of correct, uncertain, and incorrect choices for each participant vs their age. Overlapping points were represented by the area of the marker. A positive correlation exists between age and incorrect percentage, while a negative correlation is observed between age and uncertain.

political decoys. Interframe forgery, on the other hand, has a limited footprint as it only affects a part of the video. The footprint is further covered using the morph-cut technique for smoothing the transition in jump cuts. Yet both these techniques are expensive in practice, due to the difficulty of finding look-alike impersonators, and of finding long videos depicting consistent scenes of the target person to be used in the morph-cut setting.

It can also be concluded that the selected fake videos from CGI, Face CGI, Face GAN, and Prosthetic Makeup techniques had not yet reached convincing video-realism. The results also suggest special lighting setups to be effective in resulting in more errors in the population, obfuscating the artifacts caused by the generation method.

The existence of a biometric reference reduces the number of errors, while knowing of the target individual reduces the uncertainty, contributing to a higher number of correct classification. The presented familiarization was not effective in increasing the accuracy of participants, yet it caused a lower number of incorrect choices which was in turn compensated with a higher number of uncertain ones. Furthermore, it is observed that individuals rely on a small set of clues for their decision, and the main clue supporting the participants' decision is in the head/face area.

Many parameters did not yet provide any statistically significant difference due to the small number of participants and videos per category. This will be investigated in more detail in our future work. The selected techniques also had a big difference in ease of detection, limiting the effect of conditions such as familiarization and biometric reference in detection accuracy. Furthermore, the population was not

representative of the general population, limiting the implications of the findings.

This study shows the performance of suspecting audience in the specific task of differentiating real and fake videos. However, in a real-life scenario, the audience is not actively judging every and each video for them being real or fake, and the major source of trust comes from the publisher of the video. The future work for this study consists of a test design that measures the subjective performance in an unsuspecting manner, on a wider and more diverse population. Furthermore, the auditory and visual aspects of the signal will be studied separately. Effects of lighting conditions and popularity of the subjects will also be studied in further detail.

6.6 Acknowledgement

We thank the Department of Information Security and Communication Technology, NTNU for funding, Erwin Haasnoot and Adam Szekeres for guidance on the test setup, and blind reviewers and Dr. Carl Stuart Leichter for invaluable comments during the review process.

References

- [1] J. Chen, Y. Deng, G. Bai, and G. Su. Face image quality assessment based on learning to rank. *IEEE Signal Processing Letters*, 22(1):90–94, Jan 2015.
- [2] S. Fan, T.-T. Ng, J. S. Herberg, B. L. Koenig, and S. Xin. Real or fake?: Human judgments about photographs and computer-generated images of faces. In *SIGGRAPH Asia 2012 Technical Briefs*, SA '12, pages 17:1–17:4, New York, NY, USA, 2012. ACM.
- [3] S. Fan, R. Wang, T.-T. Ng, C. Y.-C. Tan, J. S. Herberg, and B. L. Koenig. Human perception of visual realism for photo and computer-generated face images. *ACM Trans. Appl. Percept.*, 11(2):7:1–7:21, July 2014.
- [4] H. Farid and M. J. Bravo. Perceptual discrimination of computer generated and photographic faces. *Digital Investigation*, 8(3):226–235, 2012.
- [5] ISO/IEC 30107-3:2017. Information technology - Biometric presentation attack detection - Part 3: Testing and reporting. Standard, International Organization for Standardization, Sept. 2017.
- [6] M. G. KENDALL. *The advanced theory of statistics*. Number 4th Ed. Macmillan, 1979.
- [7] A. Khodabakhsh, C. Busch, and R. Ramachandra. A taxonomy of audiovisual fake multimedia content creation technology. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 372–377, April 2018.
- [8] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch. Fake face detection methods: Can they be generalized? In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–11, 2018.
- [9] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, Dec 2012.
- [10] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Proc. Interspeech 2017*, pages 2616–2620, 2017.
- [11] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR*, abs/1803.09179, 2018.

- [12] V. Schetinger, M. M. Oliveira, R. da Silva, and T. J. Carvalho. Humans are easily fooled by digital images. *Computers & Graphics*, 68(Supplement C):142–151, 2017.
- [13] C. Schmitz et al. Limesurvey: An open source survey tool. *LimeSurvey Project Hamburg, Germany*. URL <http://www.limesurvey.org>, 2012.
- [14] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36, Sept. 2017.

Chapter 7

Article 3: Action-Independent Generalized Behavioral Identity Descriptors for Look-alike Recognition in Videos

A. Khodabakhsh and H. Loisel, "Action-Independent Generalized Behavioral Identity Descriptors for Look-alike Recognition in Videos," 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2020, pp. 151-162.

7.1 Abstract

There is a long history of exploitation of the visual similarity of look-alikes for fraud and deception. The visual similarity along with the application of physical and digital cosmetics greatly challenges the recognition ability of average humans. Face recognition systems are not an exception in this regard and are vulnerable to such similarities. In contrast to physiological face recognition, behavioral face recognition is often overlooked due to the outstanding success of the former. However, the behavior of a person can provide an additional source of discriminative information with regards to the identity of individuals when physiological attributes are not reliable. In this study, we propose a novel biometric recognition system based only on facial behavior for the differentiation of look-alikes in unconstrained recording conditions. To this end, we organized a dataset of 85,656 utterances from 1000 look-alike pairs based on videos collected from the wild, large enough

for the development of deep learning solutions. Our selection criteria assert that for these collected videos, both state-of-the-art biometric systems and human judgment fail in recognition. Furthermore, to utilize the advantage of large-scale data, we introduce a novel action-independent biometric recognition system that was trained using triplet-loss to create generalized behavioral identity embeddings. We achieve look-alike recognition equal-error-rate of 7.93% with sole reliance on the behavior descriptors extracted from facial landmark movements. The proposed method can have applications in face recognition as well as presentation attack detection and Deepfake detection.



Figure 7.1: Examples of look-alike identity pairs in the proposed 1000 look-alike pairs (1000LP) dataset. Each column shows one pair of look-alikes. The identities in the proposed dataset are a subset of the identities in the VGGFace2 (4) dataset.

7.2 Introduction

Distinguishing visually similar individuals, be it identical twins or look-alikes with physical make-up or plastic surgery, has been challenging for both humans and face recognition algorithms (16). In the context of video communication, this vulnerability is further exacerbated as other means of identity verification are often not available. Moreover, the use of look-alikes and make-up for fraud has an advantage over digital manipulation methods as they don't produce any digital footprint in the received signal to be used for detection. Furthermore, despite the rise of advanced digital video manipulation methods such as Deepfakes, subjective tests show higher susceptibility of viewers to fake videos containing look-alikes rather than digitally manipulated videos (11). Fortunately, a video signal contains additional clues on the identity of the person in the form of facial behavior (3, 15).

Among existing methods for behavioral face recognition (BFR), the vast majority of studies focus on fixed-phrase authentication or specific emotional responses. Chen et al. (6) propose use of dense optical flow vector distance for identification in a fixed-phrase scenario. In (5) Cetingul et al. experiment with dense motion features, lip contour motion features, and lip shape features with a hidden-Markov-model (HMM) classifier. Zafeiriou and Pantic (28) use principal component analysis (PCA) followed by linear discriminant analysis (LDA) on dense facial deformation features in spontaneous smile for biometric recognition. Wang and

Liew (25) show that behavioral lip biometrics based on temporal shape descriptors and motion vector representation outperforms physiological lip biometrics based on texture descriptors. Gavrilescu (9) proposes a multi-state neural network on individual facial expressions extracted in the form of facial action coding system (FACS). More recently, Iengo et al. (10) use neural networks on dynamic facial features to achieve a fixed-phrase recognition rate of 98.2% and Taskirar et al. (24) use statistical properties of facial distances during different phases of smile facial expression for face recognition.

A number of publications have attempted to address unconstrained BFR. Matta and Dugelay (18) propose using rigid head displacements along with GMM and Bayesian classifiers for person recognition. Ye and Sim (26) use locally similar facial deformation patterns for identification through the calculation of local deformation profile similarity. In (22), Shreve et al. quantify the type and intensity as well as the temporal dynamics of action units (AU) via calculating histogram distances and dynamic time warping (DTW) distance. Yuan et al. (27) propose the usage of active shape models on lip contour along with gaussian mixture models (GMM) for authentication in smartphone applications.

BFR has also been used in multi-modal biometric recognition as well as presentation attack detection (PAD). Notably, Zhao and Pietikainen (30) introduce local binary patterns (LBP) on three orthogonal planes and volume LBPs and thus incorporates immediate neighborhood frames of the video for face recognition. Kim et al. (13) use long short-term memory (LSTM) cells on top of convolutional neural networks (CNN) to capture smile facial dynamics. More recently, Pan and Deravi (19) use support vector machine (SVM) on AU histogram features for presentation attack detection. Finally, Agrawal et al. (1) model facial expressions of four individuals using facial landmarks and SVM to detect Deepfakes.

To distinguish look-alikes from each other many image-based methods have been proposed. Klare et al. (14) provide a taxonomy of facial features and analyze the discriminative power of these features for identical twin identification. The only video-based solution is proposed by Zhang et al. (29), where they extracted six types of face motion from the talking profile of identical twins and use the similarity of aligned motion sequences for classification by an SVM model. To the best of the authors' knowledge, there exists no publicly available video dataset of look-alikes in the literature. The only related video dataset in the literature is the private dataset by Zhang et al. (29) collected from 39 pairs of twins at the Mojiang International Twins Festival. There also exists a couple of related datasets containing solely images. Lamba et al. (16) collected the only dataset on look-alikes consisting of 500 images from 50 celebrities and their look-alikes. Phillips et al. (20) collected a dataset of 435 twins consisting of 24050 images.

All aforementioned publications rely on small data collected in controlled environments, and few of them address emotion- and utterance-independent detection with limited success, and as such, among all publications regarding this topic, none have addressed the unconstrained BFR in real-world scenarios. In this study, we introduce a general-purpose action-independent identity descriptor extractor based on facial behavior for distinguishing look-alikes. To this end, we also provide the first large-scale look-alike video dataset named “1000 look-alike pairs (1000LP)” which consists of approximately 23,000 real-world videos collected from a public video-sharing platform¹, for which both humans and state-of-the-art recognition systems fail at differentiation². Among the aforementioned literature, the approach in this article is in the same line of research as is taken by Zhang et al. (29) and Agrawal et al. (1). The rest of this article is organized as follows: in Section 7.3 the proposed method is described, while Section 7.4 includes the details of the collected dataset as well as the experiment setup. The results of the experiments are discussed in Section 7.5 and the article is concluded in Section 7.6.

7.3 Proposed Method

The physiological likeliness of two individuals due to natural similarity or application of physical or digital makeup may lead to false-positives in face recognition. In these cases, the facial behavior can be a source of complementary information for face recognition. Facial behavior contains identifiable information and has a significant role in person identification by humans (3, 15). In our proposed method, after face detection and facial landmark extraction in each frame, we train a convolutional deep neural network (CDNN) which maps the sequence of normalized landmark positions in the video to a vector in a generalized behavior space in an end-to-end manner. This approach enables the recognition of persons that are previously unseen by the detector by simply calculating the distance between behavior-vectors extracted from a pair of videos. Furthermore, as the network only sees the landmarks, it is guaranteed to be void of influence by the physiological likeliness of the individuals. Furthermore, landmarks are not as sensitive to disturbances and quality-related issues as other features such as optical and motion fields are and can be extracted with higher confidence.

7.3.1 Preprocessing

We use the open-source facial behavior analysis toolkit OpenFace (2) to extract the landmark positions from videos. The toolkit provides face detection as well as pose estimation and 3D landmark positions for each frame in the video. For

¹<http://www.youtube.com>

²The dataset is publicly available for download at <http://ali.khodabakhsh.org/research/1000lp/>

landmark positions to be independent of the camera position and head rotation angle, we use the pose estimation information to rotate the 3D landmarks in 3D space to achieve a frontal pose of zero degrees roll, yaw, and pitch. Further on, the landmark positions in each video are scaled to match a fixed scale used over the whole dataset. The scaling is done such that the inner eye corner landmarks would be on average 0.5 units apart. Finally, the landmarks are individually normalized using their mean and standard deviation across the whole training dataset. The aim of the aforementioned normalization steps is to convert the landmark positions to rotation-independent displacements from the average position. Even though the pose information can also contain additional behavioral identity information, they were left out due to their dependence of the estimated pose to the camera angle and position. Figure 7.2 visualizes the preprocessing pipeline.

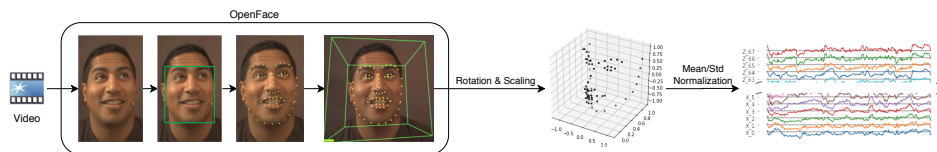


Figure 7.2: Feature extraction pipeline.

7.3.2 The proposed recognition system

To extract identity-sensitive yet action-independent information from the time series of landmark movements, it is fruitful to rely on the distribution statistics of the landmark deviations. However, due to the noisy nature of the estimated 3D landmark positions extracted from 2D videos in the pre-processing step, a refinement step proves necessary. However, the refinement criteria are ambiguous as the correct landmark position is not available. Furthermore, the movements are correlated to a large extent and contain redundancies. Motivated by the recent success of x-vectors (23) in the field of speaker recognition, we propose the network architecture shown in Figure 7.3 for end-to-end learning of the appropriate refinement for the best identification performance before statistical pooling. In this architecture, four 1D-convolutional layers are applied to the input time series. By using max-pooling layers across time, the receptive field of the final layer of the stack can be increased. Following the convolutional layers, a linear mapping is learned to map the output of the last convolutional layer to the feature-embedding space. After calculation of the mean and standard deviation of the feature-embeddings across time, the resulting fixed-length vector is then used for generating identity embeddings by two fully-connected layers. Instead of using class labels for training the network, we use triplet loss (21) to enable better generalization capacity for unseen identities. Furthermore, batch normalization is used after the input layer, the stat-

istical pooling layer, and between the output of neurons and activation functions to reduce the learning time of the network. No activation is used on the output of the feature-embedding mapping layer and the final layer to enable the network to utilize the full embedding space.

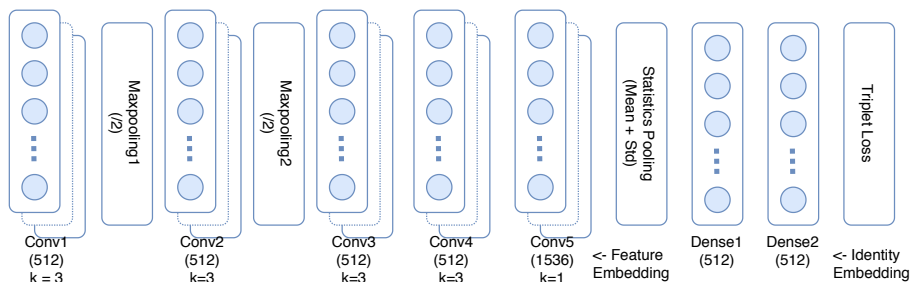


Figure 7.3: Proposed network architecture.

The Euclidean distance between identity embeddings can directly be used as a biometric dissimilarity metric. In the case of multiple enrollment samples from multiple identities, it is also possible to use the proposed system as a preprocessing step, and train a softmax layer for classification directly on extracted identity embeddings.

7.3.3 Look-alike mining

The VoxCeleb2 dataset (7) contains over 1 million utterances from more than 6,000 celebrities collected from YouTube. The identities in this dataset are a subset of identities in the VGGFace2 (4) dataset. To mine for Look-alike identities, we used the ArcFace (8) face recognition system to compare the average embeddings for each identity in the VGGFace2 dataset that appears in VoxCeleb2 dataset as well. After sorting the scores of the resulting 36M comparison pairs, the top 2,000 pairs with the highest similarity score are selected for a subjective face recognition test. Among the top pairs, there exist pairs of identical twins as well.

In the subjective face recognition test, for each look-alike pair of identities, four images are selected from each identity from the VGGFace2 dataset and shown to participants. The task for the participants was to check whether the two sets of images correspond to the same identity or two different people. The user interface is shown in Figure 7.4. Due to the large number of comparisons, the test was done by 20 participants, each labeling 200 pairs such that each pair is labeled by two people. From the resulting comparisons, the pairs that were labeled as the same people by at least one participant were selected as look-alikes and formed the 1000 look-alike pairs (1000LP) dataset. Figure 7.1 shows examples of the resultant look-alike pairs. To assure the reliability of the selected look-alike pairs,

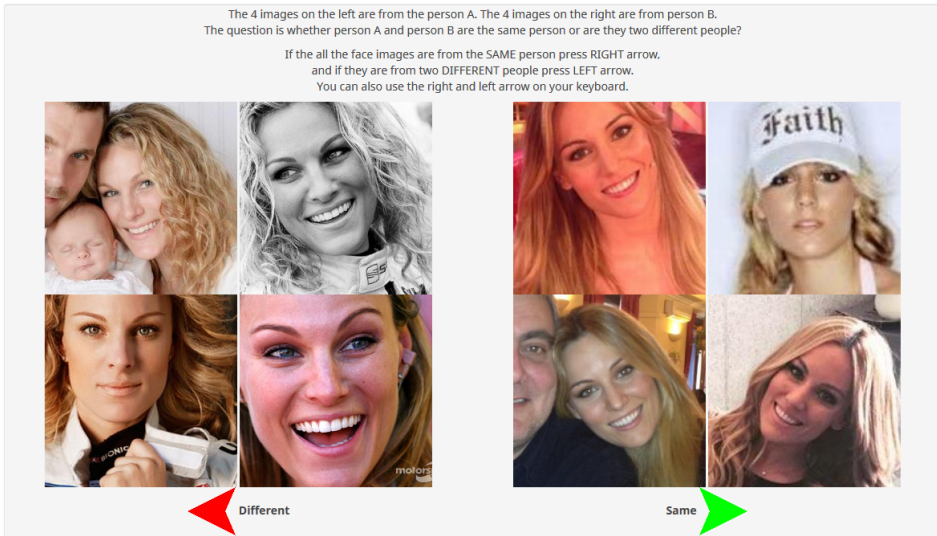


Figure 7.4: Subjective face recognition test user interface.

the equal-error-rate (EER) is calculated for the resulting look-alike pairs using the ArcFace network, resulting in an unacceptably high EER of 30.32%.

7.4 Experiment Setup

The selected 1000 pairs of look-alikes consist of 1634 unique identities. The remaining 4500 identities in VoxCeleb2 are available for training the network. The rest of this section describes the details of the organized test dataset and the parameters used for training.

7.4.1 1000LP Dataset

The utterances available in the VoxCeleb2 dataset are in the format of cropped faces sized 224×224 pixels at 25 frames per second in AVC1 format. There is a total of 1,128,246 utterances which originate from 150,480 YouTube videos. After filtering out all utterances with a length of less than 8 seconds and discarding all utterances for which face landmark detection failed, a total of 253,361 utterances remained for training and 85,656 utterances for testing. The median length of the remaining utterances is 10.7 seconds. From the 4500 train identities, 15% of them were held for validation purposes, and the remaining were used for training. For the test identities, one-third of videos (28,368 utterances) were separated for enrollment, and the remaining videos were used for testing. Resulting from this, 127,332 test trials were created, out of which 57,288 are client trials and

70,044 are impostor trials³. Special care is taken in the selection of the enrollment and test utterances such that if an utterance from a YouTube video is used in the enrollment, no utterances from the same video remains in the test trials. Thus, the performance is assured to correspond to the cross-video performance in real-life use.

7.4.2 Detector

The network parameters are shown in Figure 7.3. The breadth of the network along with the dimension of the final embedding is set to 512, with only the exception of expanded feature embedding dimensions of three times the breadth. The total number of trainable parameters in the network was 5.3M. A kernel size of 3 is used in the convolutional stack while max-pooling is done with a stride of two, resulting in a receptive field of 23 frames (roughly one second) before statistical pooling. The normalized input had a dimension of 204 corresponding to 3D coordinates for the 68 landmarks. The model was trained using TensorFlow⁴ with a batch size of 256 and the learning rate was manually adjusted towards minimizing validation loss. Semi-hard triplet loss on L2 distance of L2 normalized network outputs was used and the model was trained for 10 epochs. The hyper-parameters are selected according to the highest network performance on validation data.

7.5 Results and Discussion

The verification and identification performance of the proposed method for Euclidean similarity as well as softmax probabilities are reported in Table 7.1. The Euclidean similarity scoring performs better in identification mode than softmax probabilities and achieves an identification accuracy of 79.84% on video level. This is remarkable considering the large number of identities enrolled in the system (1634). Despite the high identification accuracy, the EER of the Euclidean similarity measure is 13.04%. Softmax probabilities, however, achieve a much better EER of 7.93% in verification mode. This discrepancy shows that softmax probabilities perform better in separating score distributions of client and impostor trials, but fails to preserve the ranking order of similarities. The detection error tradeoff (DET) curve is shown in Figure 7.5 visualizing the fact.

In order to be able to interpret the performance of the proposed method, it is compared to the reported results for existing BFR methods in the literature in Table 7.2. It is important to emphasize that all previous methods have only been tested on videos with controlled and semi-controlled recording environments. Among

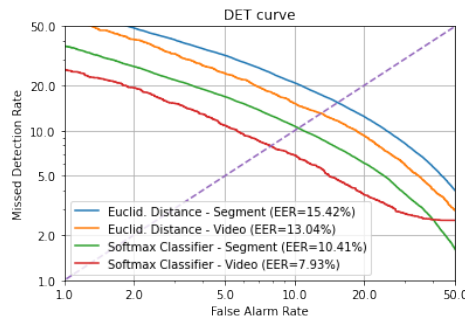
³The dataset is publicly available for download at <http://ali.khodabakhsh.org/research/10001p/>

⁴<https://www.tensorflow.org/>

		Verification	Identification	
		EER (%)	Top-1 (%)	Top-5 (%)
Euclidean	Segment (~ 10 sec)	15.42	60.57	77.83
Distance	Video (~ 4 seg)	13.04	79.84	92.61
Softmax	Segment (~ 10 sec)	10.08	65.47	81.00
Classifier	Video (~ 4 seg)	7.93	73.87	86.33

Table 7.1: The performance of the proposed methods.

Figure 7.5: Detection error tradeoff (DET) curve for the proposed methods.



the methods that operate on non-predetermined motion, the proposed method has the lowest EER and a comparable recognition rate despite the number of enrolled identities being orders of magnitude larger.

The t-distributed stochastic neighbor embeddings (t-SNE) (17) for enrollment utterances for a subset of identities is visualized in Figure 7.6. It is visible that the enrollment utterances of test set identities form concentrated clusters with few outliers. This signifies that the learned embedding space is able to generalize well across unseen identities, and the failure cases probably correspond to the outliers. Figure 7.7 shows landmark significance for a selected set of filters in the first convolutional layer of the network. The significance is measured in terms of the norm of the 3×3 matrix corresponding to multiplicative weights in x , y , and z coordinates of each landmark in frames $t - 1$, t , and $t + 1$. These heatmaps show the reliance of the network on meaningful facial actions such as eyebrow movements, upper lip movements, and movements in the corners of the mouth.

The results of this study show the power of large data in improving the performance and generalizability of BFR systems. Even though this system is trained on 4500 identities, the number of training identities is still much smaller compared to physiological face recognition systems, and there is room for further improvement.

Table 7.2: The performance of the proposed method in contrast to the reported results for existing methods.

Ref.	Subj. #	Environment	Motion	Feature	Classifier	Perf.	Metric
(6)	28	Controlled	Fixed-Phrase Speech	Motion Flow Fields	PCA + LDA	~87%	Recog. Rate
(5)	50	Controlled	Fixed-Phrase Speech	Grid-based Motion Contour-based Motion	LDA + Bayes	5.2% 12.0%	EER
(28)	22	Controlled	Spontaneous Smile	Lip Shape	PCA + LDA	10.4%	EER
(25)	40	Controlled	Fixed-Phrase Speech	Motion Fields	HMM-UBM	2.5%	EER
(9)	64	Controlled	Induced Emotion	Lip Shape Deformation Lip Texture Deformation	MLP	1.92% 8.53%	EER
(10)	20	Controlled	Fixed-Phrase Speech	Facial Action Units	DNN	91.7%	Pertision
(24)	400	Controlled	Spontaneous Smile	Facial Landmarks	DNN	0.64%	EER
(18)	9	Controlled	Unconstrained Speech	Facial Landmark Distances	Euclid. Dist.	31.20%	EER
(26)	97	Controlled	Unconstrained Emotion	Facial Feature Displacement	GMM	19.1%	EER
(22)	96	Ambiguous	Unconstrained Emotion	Local Deformation Patterns	Similarity	18.86%	EER
(27)	20	Ambiguous	Unconstrained Speech	Facial Action Units	Hist. Sim. DTW	~62%	Recog. Rate
(1)	Clinton Sanders Trump Warren	Ambiguous	Unconstrained Speech	Lip Contour	GMM	96.2%	Recog. Rate
Proposed	1634	Unconstrained	Unconstrained Speech	Facial Action Units	SVM	75% 95% 77% 95%	TPR @ 10% FPR
Proposed	1634	Unconstrained	Unconstrained Speech	Facial Landmarks	CDNN	7.93% 79.84% 93.12%	EER Recog. Rate TPR @ 10% FPR

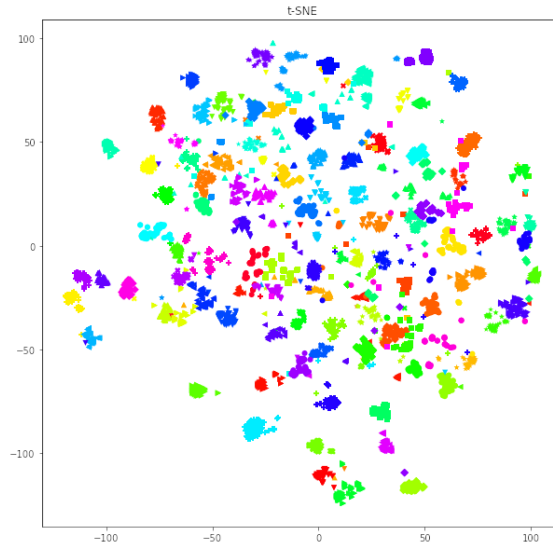


Figure 7.6: t-distributed stochastic neighbor embedding for enrollment utterances. For aesthetic reasons, only the identities with more than 50 enrollment utterances are visualized. Different colors and shapes signify different identities.

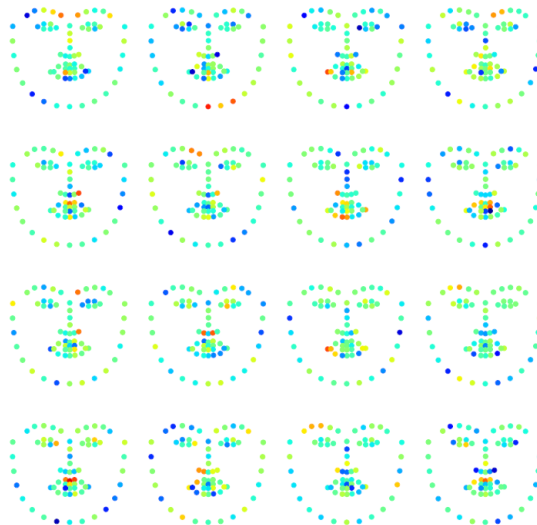


Figure 7.7: Facial landmark significance visualization for selected filters in conv1. The significance is measured as the norm of the 3×3 matrix corresponding to x , y , and z coordinates of the landmark in frames $t - 1$, t , and $t + 1$.

7.6 Conclusion

In this article, we proposed a novel general-purpose action-independent behavioral identity embedding extraction network with acceptable performance for real-life applications. The network benefits from a large number of training samples and identities and proves capable of extracting descriptive embeddings for unseen identities in unconstrained conditions. We also respond to the lack of publicly available large-scale datasets for look-alike detection, as well as publicly available behavioral face recognition systems by releasing the 1000 look-alike pairs (1000LP) dataset and the code for the proposed method.

The proposed method provides a complementary source of identity information that can be used alongside physiological face recognition systems to make them robust against look-alikes, as well as presentation attacks that try to mimic the physiological likeliness. The proposed method is robust to physical and digital spatial signal manipulations as it relies solely on the temporal behavior of the individual in question. Due to the permanence of behavioral face biometrics (3) and its robustness to manipulations and quality degradation, these methods have already found their way into the detection of Deepfakes (1) and can provide a robust alternative to existing narrowly applicable detection methods (12).

References

- [1] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In *CVPR Workshops*, June 2019.
- [2] T. Baltrušaitis, P. Robinson, and L. Morency. Openface: An open source facial behavior analysis toolkit. In *WACV*, pages 1–10, 2016.
- [3] L. Benedikt, D. Cosker, P. L. Rosin, and D. Marshall. Assessing the uniqueness and permanence of facial actions for use in biometric applications. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(3):449–460, May 2010.
- [4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, pages 67–74, 2018.
- [5] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp. Discriminative analysis of lip motion features for speaker identification and speech-reading. *IEEE TIPS*, 2006.
- [6] L.-F. Chen, H. Y. M. Liao, and J.-C. Lin. Person identification using facial motion. In *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, volume 2, pages 677–680 vol.2, Oct 2001.
- [7] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *CoRR*, abs/1806.05622, 2018.
- [8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4685–4694, 2019.
- [9] M. Gavrilescu. Study on using individual differences in facial expressions for a face recognition system immune to spoofing attacks. *IET Biometrics*, 5(3):236–242, 2016.
- [10] D. Iengo, M. Nappi, S. Ricciardi, and D. Vanore. Dynamic facial features for inherently safer face recognition. In *ICIP*, pages 2611–2615, 2019.
- [11] A. Khodabakhsh, R. Ramachandra, and C. Busch. Subjective evaluation of media consumer vulnerability to fake audiovisual content. In *QoMEX*, pages 1–6, 2019.
- [12] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch. Fake face detection methods: Can they be generalized? In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–11, 2018.

- [13] S. T. Kim, D. H. Kim, and Y. M. Ro. Facial dynamic modelling using long short-term memory network: Analysis and application to face authentication. In *BTAS*, 2016.
- [14] B. Klare, A. A. Paulino, and A. K. Jain. Analysis of facial features in identical twins. In *IJCB*, pages 1–8, 2011.
- [15] B. Knight and A. Johnston. The role of movement in face recognition. *Visual Cognition*, 4(3):265–273, 1997.
- [16] H. Lamba, A. Sarkar, M. Vatsa, R. Singh, and A. Noore. Face recognition for look-alikes: A preliminary study. In *IJCB*, pages 1–6, 2011.
- [17] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [18] F. Matta and J. Dugelay. A behavioural approach to person recognition. In *2006 IEEE International Conference on Multimedia and Expo*, pages 1461–1464, 2006.
- [19] S. Pan and F. Deravi. Facial action units for presentation attack detection. In *EST*, pages 62–67, 2017.
- [20] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, and M. Pruitt. Distinguishing identical twins by face recognition. In *Face and Gesture 2011*, pages 185–192, 2011.
- [21] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, June 2015.
- [22] M. Shreve, E. A. Bernal, Q. Li, J. Kumar, and R. Bala. A study on the discriminability of faces from spontaneous facial expressions. In *ICIP*, pages 1674–1678, 2016.
- [23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, pages 5329–5333, 2018.
- [24] M. Taskirar, M. Killioglu, N. Kahraman, and C. E. Erdem. Face recognition using dynamic features extracted from smile videos. In *INISTA*, pages 1–6, 2019.
- [25] S.-L. Wang and A. W.-C. Liew. Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power. *Pattern Recognition*, 2012.

- [26] N. Ye and T. Sim. Towards general motion-based face recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2598–2605, June 2010.
- [27] Y. Yuan, J. Zhao, W. Xi, C. Qian, X. Zhang, and Z. Wang. Salm: Smartphone-based identity authentication using lip motion characteristics. In *SMART-COMP*, 2017.
- [28] S. Zafeiriou and M. Pantic. Facial behaviometrics: The case of facial deformation in spontaneous smile/laughter. In *CVPR 2011 WORKSHOPS*, pages 13–19, June 2011.
- [29] L. Zhang, K. Ma, H. Nejati, L. Foo, T. Sim, and D. Guo. A talking profile to distinguish identical twins. *Image and Vision Computing*, 2014.
- [30] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE TPAMI*, 29(6):915–928, 2007.

Chapter 8

Article 4: Unit-Selection Based Facial Video Manipulation Detection

T. Nielsen, A. Khodabakhsh and C. Busch, "Unit-Selection Based Facial Video Manipulation Detection," 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2020, pp. 87-96.

8.1 Abstract

Advancements in video synthesis technology have caused major concerns over the authenticity of audio-visual content. A video manipulation method that is often overlooked is inter-frame forgery, in which segments (or units) of an original video are reordered and rejoined while cut-points are covered with transition effects. Subjective tests have shown the susceptibility of viewers in mistaking such content as authentic. In order to support research on the detection of such manipulations, we introduce a large-scale dataset of 1000 morph-cut videos that were generated by automation of the popular video editing software Adobe Premiere Pro. Furthermore, we propose a novel differential detection pipeline and achieve an outstanding frame-level detection accuracy of 95%.

8.2 Introduction

Following the evolution of artificial intelligence and the rapid increase in the computational capacity of computers in recent decades, many novel video manipulation techniques have been introduced and became feasible. Despite the original

intention of the developers of these techniques, many of them have the potential of being misused by malicious actors to spread disinformation for political and financial aims. Following the significant media attention to this problem after the introduction of Deepfakes, many research groups attempt to address the vulnerability (19). However, among video manipulation techniques, vulnerability to unit-selection based methods have been overlooked. Unlike Deepfakes and similar generation methods for which synthesis still requires a significant amount of expert knowledge and computational capacity, unit-selection based video manipulation can be flexibly done by commercial software such as Adobe Premiere Pro through their easy to use graphical user interface. Furthermore, subjective tests have shown unit-selection based manipulations to be more difficult to detect for humans than intra-frame manipulations (12). The use of seamless cut-point transitions is commonplace in media for shortening and summarizing the highlights of videos and they go unnoticed more often than not¹.

Due to the less computational cost and the higher video-realism of unit-selection based generation methods, these methods have been explored for synthesis early-on for applications like audio-visual synthesis and video dubbing (14). Even though concatenative generation methods require long videos with constrained recording conditions to be seamless, thanks to searchable public archives of videos, there exists enough footage from interviews on celebrities and political figures for these methods to be feasible. The first automatic technique for face-animation was proposed by Bregler et al. in 1997 (6). They create a database of visemes² from existing footage and, given an input text, they retrieve the visemes and concatenate them using morphing to synthesize a new sentence. More recently, Berthouzoz et al. (5) introduced an editing tool to place visible cuts and seamless transitions in interview videos based on text transcript, which was further developed into the morph-cut transition in Adobe Premiere Pro³ as a replacement for B-roll⁴ and jump-cut transitions⁵ for video summarization. Mattheyses and Verhelst (14) and Johnston and Elyan (11) provide an overview of existing unit-selection based manipulation methods. Among the existing datasets, the biggest that includes inter-frame forgery is VTD 2016 (1) which is comprised of 33 videos, 6 of which contain inter-frame forgery. Johnston and Elyan (11) provide a review of existing video tampering datasets.

¹<https://metro.co.uk/2018/12/13/viewers-baffled-child-appears-t-eleport-tv-interview-8244024/>

²Visemes denote the shape of the mouth when pronouncing specific phonemes. Visemes and phonemes do not share a one-to-one correspondence.

³<https://www.adobe.com/products/premiere.html>

⁴In B-roll transition, a supplemental footage is intercut with the main shot to cover the cuts.

⁵In jump-cut transition, the cut is kept as it is, causing an abrupt jump in the resulting footage.

In the context of facial video manipulation, a substantial amount of research is oriented towards intra-frame facial video manipulation detection (19). However, there exists a gap in knowledge with regards to detection of unit-selection based facial video manipulation, and to the best of our knowledge, there are no dataset and no proposed detection method that explicitly address this vulnerability. Nonetheless, Among the proposed methods for the detection of intra-frame manipulations, some utilize inter-frame information for detection to a limited extent. The authors in (10) and (16) exploit the inter-frame dependencies to detect frame-by-frame manipulations via a convolutional long short-term memory (LSTM) network and a recurrent neural network respectively. Amerini et al. (2) use estimation of the optical flow field as input to a convolutional neural network (CNN) for the detection of inter-frame inconsistencies.

To reduce the visibility of concatenation points in inter-frame forgery, simple gradual transitions such as interpolation, warping, and morphing, as well as more advanced methods such as face-specific warping (9) and intermediate frame mining (5) can be used. Examples of advanced transitions that are already available in video editing software are Adobe Premiere Pro Morph-cut (Figure 8.1) and Avid⁶ Fluid Morph. Despite the core algorithms of these transitions being trade secrets, the name of these transitions implies the use of morphing in some form. Consequently, single-image face morphing detection algorithms that are developed in the context of biometric presentation attack detection become relevant for detection. Scherhag et al. (17) provide a recent survey of existing morphing attack detection methods. Asaad and Jassim (3) used the responses of uniform local binary pattern (LBP) extractors on the image to build a Vietoris-Rips complex for detection. Wandzik et al. (20) use high-level features of pretrained face recognition networks as input for a linear SVM classifier.

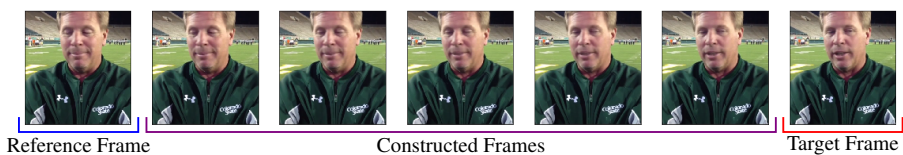


Figure 8.1: An example of a morph-cut transition.

Another set of relevant detection methods can be adopted from general-purpose inter-frame forgery detection, namely frame-insertion and frame-deletion detection methods. Siatara and Mehtre (18) provide an overview of the existing inter-frame forgery detection methods. Notably, Chao et al. (7) detect manipulated

⁶<https://www.avid.com/>

videos by using the consistency in the total optical flow values in the X and Y directions. More recently, Bakas and Naskar (4) used 3D convolutional neural networks with a special difference layer to detect out of place frames in the video sequence.

In this work, we introduce a large-scale dataset of videos containing morph-cut transitions based on videos collected from the wild.⁷ To the best of our knowledge, the Morph Cut dataset is the first of its kind and enables the training of deep learning solutions for the detection task. Furthermore, we introduce a robust neural detection pipeline, capable of detecting the morph-cut position at the frame level in a video. The rest of this article is organized as follows: The dataset and the proposed detector are introduced in Section 8.3. The experiment setup is explained in Section 8.4 and the results are discussed in Section 8.5. Finally, the paper is concluded in Section 8.6.

8.3 Methodology

Due to the lack of datasets containing a sufficiently large number of unit-selection based manipulation in the literature, we decided to generate a dataset and provide it publicly to stimulate further research in inter-frame forgery detection. In this section, we summarize the construction process of the new Morph Cut dataset along with the description of our proposed method for detecting the inter-frame forgeries.

8.3.1 Morph Cut Dataset

The development of deep learning-based detectors requires large-scale datasets. Consequently, as the manual generation of datasets of such scale is impractical, the generation process needs to be automated. Adobe Premiere Pro is a well-known popular video editing application that features a seamless morph-cut transition for cut-point concatenation. Furthermore, Adobe Systems provide the scripting language named Extendscript which can be used for automation of repetitive tasks in video editing. As such, Adobe Premiere Pro morph-cut transition is the perfect candidate to be used for the generation of the dataset. To achieve a seamless transition, the frames before and after transition need to be similar with regards to the background as well as the general body posture.

To ensure the quality of the generated data, we relied on a much larger video dataset consisting of interview videos as the basis for video selection. Thereafter, based on the movements of face bounding-box after face detection in the videos and the

⁷The instructions on how to download the Morph Cut dataset are available at <http://ali.khodabakhsh.org/research/morphcut/>

structural similarity of the frames to one another, the videos were ranked and the most suitable videos were selected for the application of morph-cut. Subsequently, the transition is applied to the videos at random points during the interview and the resulting manipulated videos were manually investigated for videos with visible artifacts to be discarded.

8.3.2 Morph-cut Detection

The unit-selection based video synthesis requires smooth transitions at the cut-points to cover the abrupt changes between the frame before and after. As such, it is safe to assume the existence of frame interpolation during the transition in one form or another. During frame interpolation, the content of the new frame in-between is generated based on the information available in the frame before and after. In contrast, pristine frames contain a natural variability that is not completely explainable based on the information in the frame before and after. Let us consider the frame in the middle to be consisting of two factors, p for the redundant information that is inferable from the frame before and after, and u for the unpredictable natural variability. A good frame interpolation would be able to infer p accurately, however, inference of u is an ill-defined problem. If during the design and training of a frame interpolation method, no mechanism is considered for ignoring u , the objective function would force the interpolation method to generate an average u which minimizes the penalty, yet never occurs in the pristine data. This phenomenon often results in synthetic samples described as over-smooth.

Considering any two frame interpolation methods with the aforementioned characteristics, we hypothesize that the predicted intermediate frames would show more similarity to each other than to the pristine data. The rationale behind this is that the p factor would exist in both pristine and synthetic frames, yet the u factor would only properly occur in pristine data while the frame interpolation methods each would generate an over-smooth average u . Thus it is reasonable for the difference between the natural u and the average u to be greater than the difference between two average u s generated by the two synthesis methods. To use this behavior for interpolation detection, for each frame, the interpolated parallel can be generated from the frame before and after with any other good interpolation method that fits the aforementioned description. Next, the prediction error can be measured as the difference between the interpolated frame and the observed one. Consequently, this difference can be used for distinguishing pristine frames from interpolated ones by using a distance measure. Alternatively, this prediction error *image* can be fed to a classifier which specializes in the detection of interpolated frames for better performance.

8.4 Experiment Setup

We provide the large-scale Morph Cut dataset for the task of unit-selection based facial video manipulation detection training and testing on which we empirically verify the detection hypothesis. Furthermore, in our benchmark we perform the detection task with four applicable detection methods from the literature. The details of the dataset along with the experiment setup is explained in the following.

8.4.1 Morph Cut Dataset Details

The VoxCeleb2 (15) dataset is used as a basis for video selection, which contains a collection of interview videos from celebrities hosted on the video-sharing platform YouTube. The videos are ranked based on the face bounding-box movements, and on the suitable videos, uniform random sampling is applied to select candidate points for morph-cut. Next, the candidates with high structural similarity index (21) are selected and two morph-cut transitions are automatically added to each video using Extendscript. The Morph Cut dataset contains 1,000 videos with an average duration of 2.75 seconds. This dataset adds up to $\sim 83,000$ frames with $\sim 27,500$ morphed frames and a ratio of 33% morphed frames to pristine ones. The videos are split three sets corresponding to training, validation, and the test data according to numbers in Table 8.1. The video parameters are summarized in Table 8.2. The videos are accompanied by frame-level labels corresponding to whether each frame is morphed or pristine. All reported results are based on frame-level classification performance between the morphed frames and the pristine ones.

Table 8.1: The number of videos in each set of the constructed Morph Cut dataset.

Set	Count
Train	700
Dev	150
Test	150

Table 8.2: The parameters used to create each video in the constructed Morph Cut dataset.

Video parameters
MPEG-4 (Base Media / Version 2)
480p (854 × 480)
30 FPS (Frames-Per-Second)
AVC (NTSC)

8.4.2 Proposed Detector

For the detector’s reference frame-interpolation method, the pre-trained CyclicGen (13) convolutional neural network is used. For a given pair of frames, this network produces a high-quality intermediate interpolated frame. Using this network, for each frame in a video, a corresponding interpolated frame is synthesized based on the frame before and after, and the prediction error is calculated in terms of a difference image. The resulting prediction error *images* on cropped face regions are then converted to gray-scale and fed to a simple convolutional neural network for frame-level classification. The input to the network is augmented with the *context* prediction error images of two frames before and after, resulting in an input shape of $64 \times 64 \times 5$. The training and evaluation pipeline is visualized in Figure 8.2 and the classifier network architecture is summarized in Table 8.3.

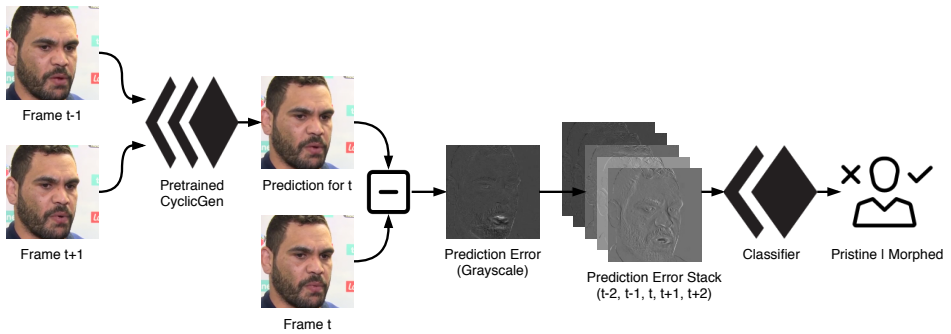


Figure 8.2: The training and evaluation pipeline in the proposed method.

8.4.3 Baseline Methods

For baseline methods to be used in our benchmark, we relied on recently published and reproducible detection methods for face-morph detection (3), time-aware Deepfake detection (10), inter-frame forgery detection (4), and general purpose image classification (8). Among the four methods, (10) and (4) utilize temporal information while (3) and (8) rely only on static face images. All methods provide frame-level decision.

The first method is based on topological data analysis for image tampering detection described in the paper of the same name (3). This method was originally created to detect morphing attacks on face images by extracting features from the texture of the image itself, making the method sensitive to image tampering through the degradation of the image. For this method, we first extract the cropped faces from each frame in the dataset and construct a 1-skeleton of the fullrips simplicial complex for each face image, which is then fed into an SVM classifier to

Table 8.3: The network architecture of the classifier. The network contains 1.6M trainable parameters.

Layer	Output Shape	Parameters
Conv2D	(62, 62, 128)	Kernel=(3,3)
MaxPooling2D	(31, 31, 128)	Pool=(2,2)
Conv2D	(29, 29, 128)	Kernel=(3,3)
MaxPooling2D	(14, 14, 128)	Pool=(2,2)
Conv2D	(12, 12, 256)	Kernel=(3,3)
MaxPooling2D	(6, 6, 256)	Pool=(2,2)
Conv2D	(4, 4, 512)	Kernel=(3,3)
MaxPooling2D	(2, 2, 512)	Pool=(2,2)
Flatten	(2048)	
Dense	(512)	
DropOut	(512)	
Dense	(2)	

attempt and recognize the morphed faces against the pristine ones.

The second method relies on recurrent neural networks for Deepfake detection (10). The cropped face images are used as input to the network and all parameters are kept the same as described in the paper except we are training with fewer epochs. The third method relies on 3D convolutional neural networks for the detection of inter-frame forgery as described in (4). Finally, due to the outstanding performance of the Xception-Net (8) for Deepfake detection task, the pre-trained network is fine-tuned on the task of morph-cut detection on individual images.

8.5 Results and Discussion

Table 8.4 summarizes the detection accuracy of the proposed method in comparison to the baseline methods. The proposed method achieves the highest detection accuracy of 95.1% on the test set, followed surprisingly by the fine-tuned XceptionNet at 77.0%. The other three baseline methods show limited success in the detection of morph-cut frames. The detection-error-tradeoff (DET) curve for the top 3 best-performing methods is shown in Figure 8.3. In this figure, APCER stands for attack presentation classification error rate and BPCER stand for bona fide presentation classification error rate, which correspond to the missed detection and the false alarm rate of a biometric presentation attack detection system respectively following the ISO/IEC 30107 standard terminology⁸. The proposed method achieves an acceptable detection equal-error-rate (EER) of 4.95%.

⁸<https://www.iso.org/obp/ui/#iso:std:iso-iec:30107:-3:ed-1:v1:en>

Table 8.4: The detection accuracy of the proposed method in comparison to the baseline methods. The results show the frame-level performance.

Method	Test Accuracy
Topological Data Analysis (3)	50.2%
Deepfake Video Detection (10)	59.0%
Inter-Frame Forgery C3D (4)	67.4%
Fine-tuned XceptionNet (8)	77.0%
Proposed Method	95.1%

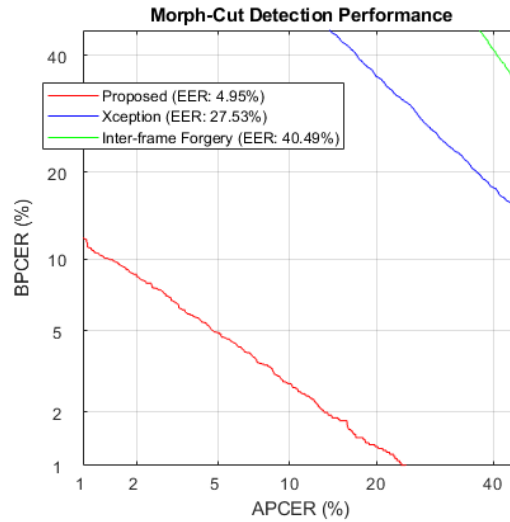


Figure 8.3: The DET curve for the frame-level detection performance of the proposed method, the fine-tuned Xception-Net(8), and the inter-frame forgery detection method(4). The equal-error-rate (EER) value for the aforementioned methods is shown in the figure legend.

Examples of the prediction errors which are used as input to the classifier in the proposed method are visualized in Figure 8.4. Natural variations are clearly visible in prediction errors in pristine frames, while these variations are not observed in the morphed (interpolated) ones. Figure 8.5 shows the probability density distribution of average prediction error per frame over pristine and morphed frames. The morphed frame average prediction error distribution is shifted towards zero compared to the pristine distribution, confirming the hypothesis proposed in Section 8.3.2. The clear distinction between the pristine and morphed frame prediction errors visualized in Figure 8.4 and 8.5 show the effectiveness of prediction error *images* in isolating useful features for morphed face detection.

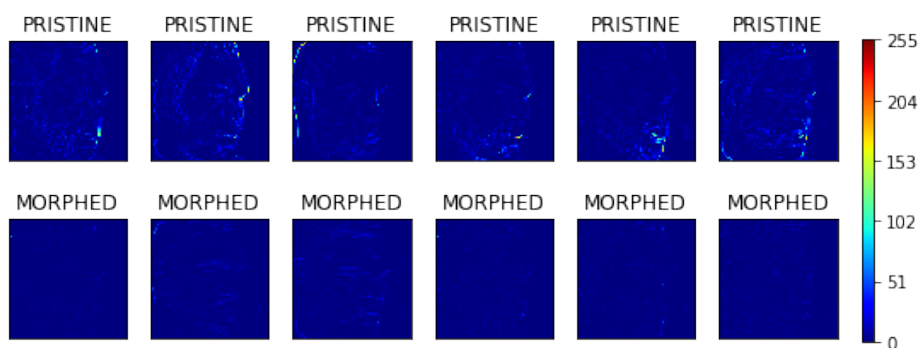


Figure 8.4: Example of prediction error *images* of cropped faces in a six-frame sequence of pristine frames (top) and morph-cut frames (bottom) in a video. The images visualize the absolute gray value difference per pixel between the interpolation output and the actual frame.

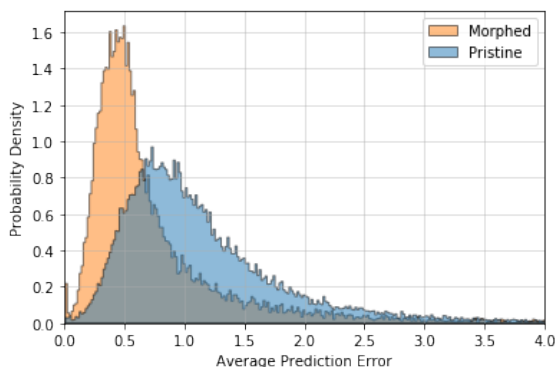


Figure 8.5: The probability density distribution of average prediction error per frame for pristine and morphed frames across the dataset.

8.6 Conclusion

In this article, we addressed the problem of unit-selection based facial video manipulation by providing the first large-scale dataset of videos manipulated by popular video-editing software. Furthermore, we proposed a detection method that relies on frame-interpolation prediction-errors as discriminative features for the detection of morphed frames. The proposed method outperforms the baseline methods by a wide margin. The high frame-level performance of the proposed method shows its capacity in reliably detecting unit-selection based video manipulation and confirms the detection hypothesis that synthetic frames demonstrate higher similarity to each other than to pristine ones.

References

- [1] O. I. Al-Sanjary, A. A. Ahmed, and G. Sulong. Development of a video tampering dataset for forensic investigation. *Forensic Science International*, 266:565–572, 2016.
- [2] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo. Deepfake video detection through optical flow based cnn. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.
- [3] A. Asaad and S. Jassim. Topological data analysis for image tampering detection. In *International Workshop on Digital Watermarking*, pages 136–146. Springer, 2017.
- [4] J. Bakas and R. Naskar. A digital forensic technique for inter-frame video forgery detection based on 3d cnn. In *International Conference on Information Systems Security*, pages 304–317. Springer, 2018.
- [5] F. Berthouzoz, W. Li, and M. Agrawala. Tools for placing cuts and transitions in interview video. *ACM Trans. Graph.*, 31(4), July 2012.
- [6] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *SIGGRAPH '97*, pages 353–360, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.
- [7] J. Chao, X. Jiang, and T. Sun. A novel video inter-frame forgery model detection scheme based on optical flow consistency. In *IWDW, IWDW'12*, page 267–281, Berlin, Heidelberg, 2012. Springer-Verlag.
- [8] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.
- [9] K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlasic, W. Matusik, and H. Pfister. Video face replacement. *ACM Trans. Graph.*, 30(6):130:1–130:10, Dec. 2011.
- [10] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.
- [11] P. Johnston and E. Elyan. A review of digital video tampering: From simple editing to full synthesis. *Digital Investigation*, 29:67–81, 2019.

-
- [12] A. Khodabakhsh, R. Ramachandra, and C. Busch. Subjective evaluation of media consumer vulnerability to fake audiovisual content. In *QoMEX*, pages 1–6, 2019.
- [13] Y.-L. Liu, Y.-T. Liao, Y.-Y. Lin, and Y.-Y. Chuang. Deep video frame interpolation using cyclic frame generation. In *Proceedings of the 33rd Conference on Artificial Intelligence (AAAI)*, 2019.
- [14] W. Mattheyses and W. Verhelst. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66(Supplement C):182–217, 2015.
- [15] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Proc. Interspeech 2017*, pages 2616–2620, 2017.
- [16] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *CVPR Workshops*, June 2019.
- [17] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7:23012–23026, 2019.
- [18] K. Sitara and B. Mehtre. Digital video tampering detection: An overview of passive techniques. *Digital Investigation*, 18(Supplement C):8–22, 2016.
- [19] L. Verdoliva. Media forensics and deepfakes: an overview. *arXiv preprint arXiv:2001.06564*, 2020.
- [20] L. Wandzik, G. Kaeding, and R. V. Garcia. Morphing detection using a general- purpose face recognition system. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1012–1016, 2018.
- [21] Z. Wang and A. C. Bovik. Mean squared error: Lot it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

Part III

Generalizability

Chapter 9

Article 5: Fake face detection methods: Can they be generalized?

A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik and C. Busch, "Fake Face Detection Methods: Can They Be Generalized?," 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2018, pp. 1-11.

9.1 Abstract

With advancements in technology, it is now possible to create representations of human faces in a seamless manner for fake media, leveraging the large-scale availability of videos. These fake faces can be used to conduct personation attacks on the targeted subjects. Availability of open source software and a variety of commercial applications provides an opportunity to generate fake videos of a particular target subject in a number of ways. In this article, we evaluate the generalizability of the fake face detection methods through a series of studies to benchmark the detection accuracy. To this extent, we have collected a new database of more than 53,000 images, from 150 videos, originating from multiple sources of digitally generated fakes including Computer Graphics Image (CGI) generation and many tampering based approaches. In addition, we have also included images (with more than 3,200) from the predominantly used Swap-Face application that is commonly available on smart-phones. Extensive experiments are carried out using both texture-based handcrafted detection methods and deep learning based

detection methods to find the suitability of detection methods. Through the set of evaluation, we attempt to answer if the current fake face detection methods can be generalizable.

9.2 Introduction



Figure 9.1: Examples of different fake faces in contrast to the bona fide presentation.

Face biometrics are widely deployed in various applications as it ensures reliable and convenient verification of a data subject. The dominant application of face recognition is for logical or physical access control to for instance restricted security areas. Implicitly the human visual system applies face recognition to determine, which data subject is the communication partner be it in a face to face conversation or be it in consuming messages while observing a media stream (e.g. news channel). With recent advances in deep learning, it is now possible to seamlessly generate manipulated images/videos in real-time using technologies like image morphing, Snap-Chat, Computer Generated Face Image (CGFI), Generative Adversarial Networks (GAN) and Face2Face (14). These technologies enable an attacker to manipulate the face image either by swapping it with another face or by pixel-wise manipulation to generate a new face image/video. It is well demonstrated in the literature that face recognition techniques fail drastically in detecting generated fake faces (9). Further fake face samples can also be shared by intention with the social media, in order to spread the fake news associated with the target subject. The challenge is not only posed to the biometric systems but also to the general media perception on social media. Thus it is of paramount importance to detect faked face representations to reduce the vulnerability of biometrics systems and to reduce the impact of manipulated social media content.

Traditional biometric systems have addressed this problem of detecting the fake faces using Presentation Attack Detection (PAD) schemes (1, 10). PAD schemes

¹Pinscreen: <http://www.pinscreen.com/>

²<https://www.fakeapp.org/>

³“We the people”: <http://www.macinnesscott.com/vr-art-x>

in the earlier works have investigated and provided remedial measures focused on both attacks with low-cost artefacts (e.g. print, display, and wrap) and high-cost artefacts (like silicon masks). Another kind of attacks based on face morphing takes face images of two different data subjects to generate a new morphed face image which can practically match both the subjects (9). Yet another and recently created method of generating a faked face image/video was presented in (14) that can be used to introduce a personation attack on the target subject. The personation attack can be constructed by the re-enactment process, transferring the facial expressions from the source actor to a target actor, resulting in the manipulated images/video. This generated facial sample through such procedures is referred to as the fake face (5) (11). The generated content shows high sample quality of images/videos, which is difficult to detect even for trained forensic examiners (11). There are recent additions to generate fake face images that include the use of GAN, CGI, Face2face, and others which are highly realistic. The reliable detection of such fake face images is challenging due to the process of re-enactment. This results in infinitesimal variation in the face images that challenges the conventional forensics methods based on extracting edge discontinuities and texture information in spotting manipulated images.

To the best of our knowledge, there exists only one work that has attempted to detect fake faces, which were using only one type of fakes, generated by Face2Face application (11). In their work (11), pre-trained deep Convolutional Neural Network (CNN) based approaches are evaluated on the newly constructed fake face image database. The results reported in (11) show good detection performance of the pre-trained Xception CNN that can be attributed to the fact that both fake face generation and detection are carried out on the training and testing subset of one particular dataset (FaceForensics). While this is an important first step, we need to anticipate that with the evolution of computer vision technologies, fake faces can also be generated using alternative and newer methods. Thus, it is necessary to provide an insight into the generalization of the methods that are used to detect the fake faces to measure the reliability.

In this work, we present a comprehensive and exploratory study on the generalizability of different fake face detection methods based on both recent deep learning methods and conventional texture descriptor based methods. To this extent of studying generalizability, we present a new database created using diverse methodologies for generating fake faces. Further, we also propose the protocols to effectively evaluate the generalizability of both texture based and deep learning based methods. The main contributions of this paper in fake face detection are:

- A new database which we hereafter refer as *Fake Face in the Wild (FFW)* database with more than 53,000 images (from 150 videos) assembled from public sources (YouTube) is introduced. This database shows the largest diversity of different fake face generation methods provided so far.
- In the view of limited public databases available for this key research area, the newly created database will be made available for the public along with the publication of this paper.
- Comprehensive evaluation of 6 different algorithms that include various kinds of deep learning methods such as AlexNet (6), VGG19 (12), ResNet50 (3), Xception (2), GoogLeNet/Inceptionv3 (13), and texture based methods based on Local Binary Patterns (LBP) with Support Vector Machine (SVM).
- Extensive experiments providing insights on the generalization of the algorithms for unseen fake faces are presented. Specifically, fake faces generated using three different methods such as CGI, FakeApp, face swap, etc are considered.

9.3 Fake Face in the Wild Dataset (FFW)

This section presents the details of the newly constructed database. To simulate the performance of fake face detection methods in the wild, a set of videos from a public video sharing website (YouTube) is collected. This dataset is collected with the special focus on digitally created contents, generated with recently developed technologies. These videos include a wide array of fake images generated through CGI, GANs, manual and automatic image tampering techniques, and their combinations, due to the widespread use of these methodologies. CGI is considered in this work due to the wide availability and the ease of creation of high-quality fake face images that include images of variable sizes. *The key motivation in creating this database can be attributed to non-available public databases for either devising detection methods or the study of generalizability.* This work, therefore, facilitates further research by making the dataset publicly available along with the paper.⁴

Table 9.1 shows a summary of the videos in the FFW dataset. The dataset is created using videos of variable duration ranging from 2 seconds that corresponds to 60 frames up-to 74 seconds that corresponds to more than 2,000 frames. The videos are carefully selected to have a resolution of at least 480p and above and are manually checked for assuring the quality to avoid images with visible artifacts,

⁴Download information available at <http://ali.khodabakhsh.org/ffw/>

face poses, degraded illumination on faces and resolution. The constructed dataset consists of 150 videos, of which 85 videos broadly pertain to face images manipulated via image tampering (e.g., splicing, replacing, etc) and 65 corresponds to the use of CGI. The database thus consists of 53,000 images. In order to have bona fide samples for the evaluation, we have employed publicly available face forensic database (11) resulting in a total of 78,500 bona fide samples from 150 videos.

Table 9.1: Fake Face in the Wild Dataset (FFW) broad statistics. CGI faces were generated using several different graphics engines. Face (FakeApp) were generated in multiple resolutions and with different settings. Face (Other) category includes Face replacement, part of face splicing, and partial CGI faces, some of which were done manually, others automatically (see Figure 9.3 for examples).

Category	Type	# of videos
CGI	Full	50
	Head	22
Tampering	Face (FakeApp)	50
	Face (Other)	28
Total		150

To evaluate the performance on the newly created database, the quality measures are taken into consideration by processing the database through the same compression algorithm such that the quality of both fake and bona fide samples are consistent. This further avoids misleading detection error rates that for instance can be attributed to compression artefacts and bias the detection methods. Figure 9.2 shows the distribution of the average BRISQUE quality assessment (7) measured for FFW database indicating high overlap of the distribution justifying the similar quality. A sample set of images from the FFW dataset can also be seen in Figure 9.3.

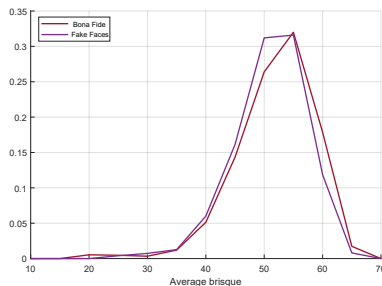


Figure 9.2: Distribution of BRISQUE quality scores for the Fake Faces in the Wild (FFW) dataset.

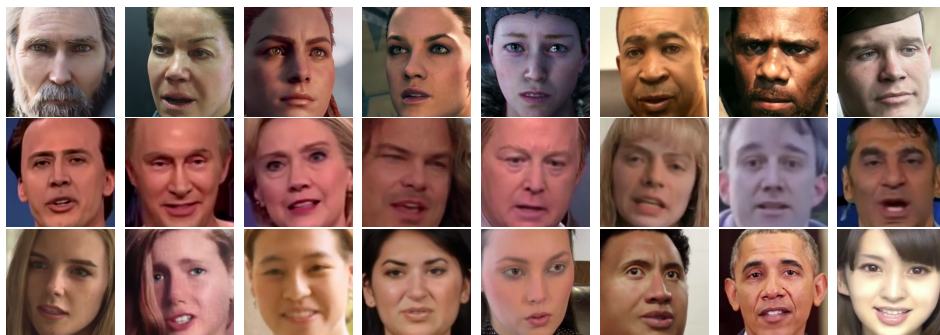


Figure 9.3: Examples from Fake Faces in the Wild (FFW) dataset. Top row: CGI full scene. Middle row: Deepfakes. Bottom row from left to right: Head CGI x2, Face replacement x2, Face CGI x2, Part of face splicing x2.

9.4 Fake Face Detection Techniques

With the goal of detection of a wide range of forged/CG/tampered audiovisual content, many methods originating from image forensics and biometrics presentation attack detection can be adapted. In this perspective, widely used texture-based method - Local Binary Patterns (LBP) and a set of CNN based systems are considered. The selection of CNN architectures AlexNet (6), VGG19 (12), ResNet50 (3), Xception (2), and GoogLeNet/Inceptionv3 (13) is based on the recent works demonstrating very high performance for various tasks. The parameters are optimized when possible on the training data and the details of parameter tuning is presented in 9.5.2.

9.5 Experimental Evaluation

This section presents the experimental evaluation of the FFW dataset. The experiment protocols are designed in accordance with protocols advised in (11). We present the evaluation of detecting known attacks followed by detecting unknown attacks.

9.5.1 Evaluation Metrics

We present the detection error rates in terms of Equal Error Rates (EER) to provide performance in the lines of earlier work. We further supplement the results using the ISO/IEC 30107-3 (4) with Attack Presentation Classification Error Rate (AP-CER) and Bona fide Presentation Classification Error Rate (BPCER) as described in (4).

9.5.2 Experimental Protocol

To effectively evaluate the fake detection methods, we divide the whole database to have three different disjoint partitions such as training set, development set, and testing set. The training set is adopted from the FaceForensics database (11) that has 7,040 bona fide and 7,040 fake face samples. The training set is used to fine tune the pre-trained deep CNN networks. To effectively fine-tune the networks and avoid overfitting, we employ 5 different types of data augmentation on each of the training images that includes translation and reflection. The learning rates of the last layer are boosted such that weights of the earlier layer are not affected and the weights of the last layer are adapted for the new training data. Thus, we have used the *weight learning rate factor* as 10 and *bias learning rate factor* as 20. For the texture based Local Binary Patterns (LBP) (8), the histogram is extracted using (8,1) neighborhoods with a block size of 40 pixels. The training dataset is used to train the SVM classifier.

The development dataset is comprised of 1,500 bona fide and 1,500 fake face samples that are taken from the validation set of FaceForensics database (11). This dataset is used to fix the operating thresholds such as Equal Error Rates (EER). The testing dataset consists of three specific kinds: (1) *To evaluate known artefacts - TestSet-I* - Test set corresponds to test set of FaceForensics database (11) that comprised of 1,500 bona fide and 1,500 fake face samples. This dataset is particularly used to understand the detection performances of known attacks. (2) *To evaluate unknown artefacts - TestSet-II* - The test set in this case consists of a newly constructed FFW dataset. In order to be inline with known attacks, this set is comprised of 1,500 bona fide and 1,500 fake face samples. (3) *To evaluate unknown artefacts - TestSet-III* - This test set comprises of 1,776 bona fide samples and 1,576 fake faces generated using FaceSwap and SwapMe application proposed by (15).

While *TestSet-I* focuses on measuring the performance of the detection algorithms, *TestSet-II* and *TestSet-III* are used to measure the generalizability of the detection techniques. It has to be noted that none of these sets (*TestSet-II* and *TestSet-III*) are used either for training, fine-tuning or validation process.

9.6 Results and Discussion

The detailed results and the obtained performance are provided in this section.

9.6.1 Performance on the Known Fake Face Attacks (TestSet-I)

The performance of texture- and CNN-based methods on known attacks (TestSet-I) are summarized in Table 9.2 and Table 9.3. Following are the main observations:

- CNN-based methods perform well and except for AlexNet, provide a detection accuracy of over 98%. In contrast, LBP features classified with SVM have the accuracy of 96% on the test data.
- In the benchmark of the CNN networks, the Inception network gives the best performance by a large margin.
- The low error rates in accord with a low EER error confirm the stability of the selected threshold point for decision. However, deviation from the selected operating point towards lower BPCER and higher APCER is visible in the results, suggesting slight inaccuracy in EER threshold estimation.

Table 9.2: The accuracy of texture- and CNN-based classifiers on the TestSet I dataset along with their confidence interval (CI).

		Accuracy \pm CI
Texture-based	LBP	96.33% \pm 0.69%
CNN-based	AlexNet	95.83% \pm 0.73%
	VGG19	98.30% \pm 0.47%
	ResNet	98.43% \pm 0.45%
	Xception	98.70% \pm 0.41%
	Inception	99.60% \pm 0.23%

Table 9.3: Performance of the systems on known fake faces from TestSet I. The threshold is computed on the development database.

	APCER	BPCER	EER
LBP	3.80% \pm 0.99%	2.87% \pm 0.86%	3.33%
AlexNet	7.80% \pm 1.38%	1.73% \pm 0.67%	3.73%
VGG19	2.47% \pm 0.80%	0.47% \pm 0.35%	1.40%
ResNet	2.27% \pm 0.77%	0.47% \pm 0.35%	1.40%
Xception	2.47% \pm 0.80%	0.13% \pm 0.19%	1.07%
Inception	0.67% \pm 0.42%	0.47% \pm 0.35%	0.53%

9.6.2 Performance on the Unknown Fake Face Presentations (TestSet-II)

Following the good performance of all neural network solutions along with the LBP features, the generalizability of the learned classifiers are examined on the collected dataset of matching size as shown in Table 9.4 and the observations are:

- The performance of all systems in terms of APCER errors drops significantly, rendering the systems ineffective, classifying most images as bona fide.

- A closer look at the EER values for these systems shows much better than random performance of CNN-based models on the Unknown dataset.
- It can be concluded that the performance of the CNN-based systems is very poor because of the low performance at the selected operating point.

Table 9.4: Performance of the systems on unknown attacks from TestSet II. The threshold is computed on the development database.

	APCER	BPCER	EER
LBP	89.00% \pm 1.62%	2.87% \pm 0.86%	48.73%
AlexNet	91.47% \pm 1.44%	1.73% \pm 0.67%	32.13%
VGG19	90.73% \pm 1.50%	0.47% \pm 0.35%	29.40%
ResNet	89.53% \pm 1.58%	0.47% \pm 0.35%	30.33%
Xception	93.20% \pm 1.30%	0.13% \pm 0.19%	26.87%
Inception	91.93% \pm 1.41%	0.47% \pm 0.35%	27.47%

To illustrate this further, the score histogram of the known and unknown attacks are presented in Figures 9.4 and 9.5 for LBP-SVM and Inception networks respectively. The dotted vertical line indicates the threshold computed on the development database that corresponds to the EER. Figure 9.4 shows the inability of the system in distinguishing unknown attacks by a significant overlap between the bona fide distribution and the distribution of scores from the unknown attacks. However, a close look into Figure 9.5 shows that even though the network is capable of discriminating between unknown attacks and the bona fide to some extent, the weak placement of the decision boundary causes the network to fail. *By setting the threshold of the system to the EER point on the known attacks, even though the system shows optimal performance for the known attacks, it also becomes vulnerable to new types of attacks, where the separability may be less.*

Performance on each Sub-Type of Attacks

To have a closer look at the capability of CNNs in generalization, EERs for each type is calculated separately and reported in Table 9.5.

- From these results, it is visible that the networks perform better in detecting CGI compared to contents generated by FakeApp, or other techniques.
- These results indicate that even though the networks were not trained to detect CGI specifically, they are still somewhat effective for detecting of CGI videos.

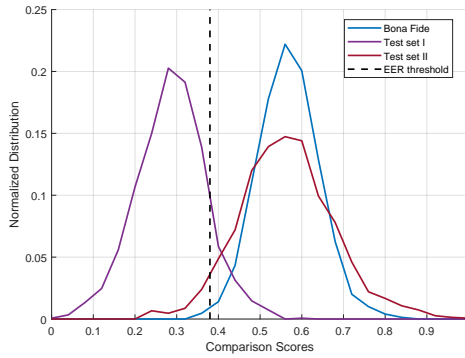


Figure 9.4: LBP-SVM system comparison score distribution on TestSets I and II.

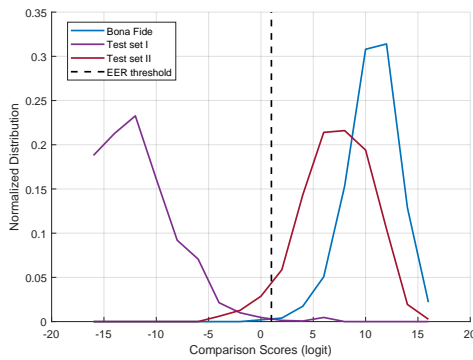


Figure 9.5: Inceptionv3 system comparison score distribution on TestSets I and II.

Table 9.5: CNN performances in terms of EER on subcategories, corresponding to Table 9.1.

	Full CGI	Image Manipulation	
		FakeApp	Other
AlexNet	32.60%	28.80%	34.37%
VGG19	28.00%	31.20%	28.60%
ResNet	28.80%	28.37%	34.40%
Xception	23.60%	25.20%	31.20%
Inception	23.40%	27.40%	31.40%

9.6.3 Performance on the FaceSwap/SwapMe Dataset (TestSet-III)

To investigate the transferability of the generalization ability of the networks on the unknown data of a widely different type, experiments were done on a filtered subset of the FaceSwap/SwapMe dataset as shown in Table 9.6.

- The APCER and EER scores present a further drop in performance.
- These results indicate the lack of transferability of the learned classifiers to the general face forgery classification cases.

Table 9.6: Performance of the systems on FaceSwap/SwapMe dataset from TestSet III. The threshold is computed on the development database.

	APCER	BPCER	EER
LBP	90.16% \pm 1.50%	3.43% \pm 0.86%	46.06%
AlexNet	94.04% \pm 1.19%	5.01% \pm 1.04%	43.02%
VGG19	97.27% \pm 0.82%	2.31% \pm 0.71%	44.93%
ResNet	89.40% \pm 1.55%	8.22% \pm 1.30%	43.79%
Xception	93.15% \pm 1.27%	3.43% \pm 0.86%	40.99%
Inception	71.64% \pm 2.27%	22.58% \pm 1.98%	46.39%

9.7 Conclusion and Future Work

The advancement of image manipulation and image generation techniques have now provided the ability to create seamless and convincing fake face images. The challenging nature of data both for visual perception and algorithmic detection is provided in recent works. The key problem that was not considered up until now is the evaluation of generalizability on existing fake face detection techniques. In order to answer the question of generalizability, in this work, we have created a new database which we refer to as Fake Face in the Wild (FFW) dataset containing 53,000 images from 150 videos that are publicly available. The key observation from this work throws light on deficiencies of detection algorithms when unknown data is presented. This observation holds for both texture descriptors and deep-learning methods, which yet cannot meet the challenge of detecting fake faces. This analysis further emphasizes the importance of validation of detectors across multiple datasets. Proposed detectors that lack such validation can show misleadingly high performances while having limited applicability, and provide little contribution to the ongoing research. As such, advancements in fake face detection technology call for the incorporation of proper cross-dataset validation in all future research as a requirement for publication.

The future work in the direction of fake face detection will involve the development of systematical methods for answering the generalization problem, and employment of multi-modal cues from fake face data.

References

- [1] S. Bhattacharjee and S. Marcel. What you can't see can help you - extended-range imaging for 3d-mask presentation attack detection. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7, Sept 2017.
- [2] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.
- [3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [4] ISO/IEC 30107-3:2017. Information technology - Biometric presentation attack detection - Part 3: Testing and reporting. Standard, International Organization for Standardization, Sept. 2017.
- [5] A. Khodabakhsh, C. Busch, and R. Ramachandra. A taxonomy of audiovisual fake multimedia content creation technology. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 372–377, April 2018.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
- [7] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, Dec 2012.
- [8] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, July 2002.
- [9] R. Raghavendra, K. B. Raja, and C. Busch. Detecting morphed face images. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2016.
- [10] R. Ramachandra and C. Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Comput. Surv.*, 50(1):8:1–8:37, Mar. 2017.
- [11] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR*, abs/1803.09179, 2018.

- [12] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.
- [14] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR'16*, pages 2387–2395, June 2016.
- [15] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. *IEEE*, pages 1831–1839, 2017.

Chapter 10

Article 6: A Generalizable Deepfake Detector based on Neural Conditional Distribution Modelling

A. Khodabakhsh and C. Busch, "A Generalizable Deepfake Detector based on Neural Conditional Distribution Modelling," 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2020, pp. 191-198.

10.1 Abstract

Photo- and video-realistic generation techniques have become a reality following the advent of deep neural networks. Consequently, there are immense concerns regarding the difficulty in differentiating what content is real from what is synthetic. An example of video-realistic generation techniques is the infamous Deepfakes, which exploit the main modality by which humans identify each other. Deepfakes are a category of synthetic face generation methods and are commonly based on generative adversarial networks. In this article, we propose a novel two-step synthetic face image detection method in which general-purpose features are extracted in a first step, trivializing the task of detecting synthetic images. The anomaly detector predicts the conditional probabilities for observing every individual pixel in the image and is trained on pristine data only. The extracted anomaly features demonstrate true generalization capacity across widely different unknown syn-

thesis methods while showing a minimal loss in performance with regard to the detection of known synthetic samples.

10.2 Introduction

Advancements in the computational capacity of modern graphical processing units (GPUs) in the past decades allowed the realization of deep neural network models. Deep learning, among other contributions, provided solutions for the synthesis of photo- and video-realistic content, challenging the existing manipulation detection methods in video forensics. An especial case of such synthetic signals is “Deep-fakes”, which are typically generated by generative adversarial networks (GANs). Deepfakes in combination with obfuscation in various forms have shown to be effective at fooling human subjects (13).

The research community has responded to this threat by developing various detection methods. Yu et al. in (20) made use of unique GAN fingerprints for the detection of fake images generated by these models. RNNs have been used for temporal-aware detection of Deepfakes by Guera et al. in (6). The spectrum domain is used by Zhang et al. (21) for the detection of GAN generated images.

Most of the existing detection methods are, however, complex and have narrow applicability as they are trained to detect specific types of synthetic signals and fail to generalize (7). Few publications try to address the detection of synthetic samples from unknown generation models. In (15), Stehouwer et al. used attention mechanisms and achieved remarkable performance over various generation techniques. Nataraj et al. (11) used pixel co-occurrence matrices for generalized detection across different GAN architectures. In (10), Marra et al. utilized multi-task learning incrementally for detecting synthetic images coming from unknown GAN models. Zhou et al. (22) proposed a two-stream classification network architecture based on steganalysis features. Afchar et al. (1) utilized mesoscopic features along with shallow networks gaining robustness against unknown synthetic images. Rossler et al. (13) evaluated different detection systems on a large dataset of diverse synthetic samples and achieved the best performance with a pretrained XceptionNet neural network. For an extensive review on the related literature, please refer to (19).

Despite major progress in the detection of synthetic face images, the generalization problem across widely different generation techniques remains a major issue. In this article, we propose a novel general-purpose feature. The subsequent trivialization enables a simple detector to reliably detect unknown attacks from widely different generation techniques. The proposed method achieves this by suppressing the content of the input signal while faithfully conserving the detection-relevant

information. The rest of this article is organized as follows: Section 10.3 explains the proposed two-step method along with the rationale behind it. Section 10.4 explains the experimental setup used for showcasing the performance of the method, and Section 10.5 discusses the findings of the article. Finally, Section 10.6 concludes the article.

10.3 Methodology

Synthetic images contain artefacts that can be used for detection and can act like fingerprints for identification of their generation process. These traces, however, are often minuscule and can be severely obscured by the actual content of the images to the extent of becoming imperceptible to the eyes of the viewer as well as the automated detection systems. We hypothesize that in the synthetic face detection task, the actual content of images acts as a strong noise, and removing them would unveil these traces and greatly simplify the task of synthetic face detection. However, this approach requires knowledge of the actual content of the image for reference.

In the absence of a reference to be subtracted from the image, the likelihood of the image to an accurate probability distribution of pristine face images would serve as a suitable proxy. To make the accurate modeling of the probability distribution over the face image space practical, the image can be broken down into smaller segments, and the probability distribution over individual segments of the image conditioned on the previous segments can be modeled.

10.3.1 Pixel RNN

The probability distribution of intensity values in each pixel conditioned on pixels before (in raster order) in pristine images can be modeled with a PixelRNN model (18). In this model, for each pixel i , the probability distribution (in the form of a Logistic mixture model) of observing the current value given all previous pixel values is learned by a recurrent or a masked convolutional neural network. This network would then be able to predict the probability distribution of pixel values for each pixel location conditioned on the pixel values before it. This probability distribution can then be used to measure the likelihood of observing a specific pixel value in location x_i given all pixel values before it ($\log(p(x_i|x_{<i}))$). By repeating this operation over all the pixels in an input image, one can calculate a likelihood matrix with the same size as the input image. Consequently, the probability of observing the input image can be calculated as $\log(p(x)) = \sum_{i=0}^n \log(p(x_i|x_{<i}))$. For the purpose of this study, an improved variant of PixelRNN named Pixel-CNN++ (14) is used.

10.3.2 Classification

The probability of the input image is a feature that can spot anomalies and can directly be used for classification. However, the conditional probability matrix corresponding to the log-likelihood of observing every single pixel intensity can serve as a better feature for classification as it contains additional information with respect to the location of anomalies and the anomaly strength at each location. For achieving a higher detection rate, one can use the model trained in the previous step as an anomaly feature extractor, or in more precise terms a universal background model (UBM). The term UBM signifies that the model is universally used regardless of the synthetic method in question in the detection task. Furthermore, it signifies that the model is a background preprocessing step which postpones the classification task to a second step. Consequently, a classifier can be trained on the output of the UBM model which is in the form of a conditional probability matrix in a supervised manner. Ideally, as the complexity of the detection problem is substantially reduced following the feature extraction step, a simple classifier should be sufficient for detection of synthetic faces. In this study, we use a very simple and small neural network for classification.

10.3.3 Generalization Performance

To measure the generalization capacity of a model, a common practice is to split the generation techniques to known and unknown methods. Next, the model is trained on synthetic data from the known methods and tested on the data from the unknown methods. To show the generalization capacity of our proposed method, we follow the same convention and do generalization tests in a leave-one-out (LOO) manner. For each generation method, we consider all other methods to be known and measure the detection performance on the single unknown method. The overall generalization performance is then measured by aggregating them over all the leave-one-out runs.

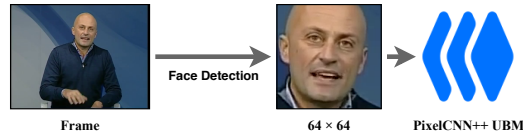
10.4 Experiment Setup

For the purpose of this study, the FaceForensics dataset (13) is selected as a large dataset containing four manipulation techniques, namely Deepfakes¹, Face2Face (17), Faceswap², and Neural Textures (16). This dataset contains 1000 pristine videos along with 1000 from each manipulation technique, each split into three sets of training (with 700 videos), development (with 150 videos), and test (with 150 videos). The videos are collected from YouTube and have a minimum quality of 480p (VGA). The videos are provided in three different quality levels to simu-

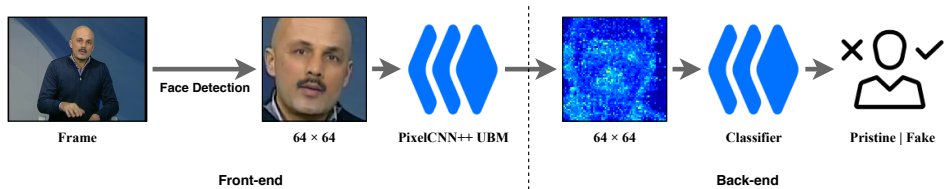
¹<https://github.com/deepfakes/faceswap>

²<https://github.com/MarekKowalski/FaceSwap/>

late the conditions of video processing in social networks. For extraction of face images from the videos, the Dlib toolkit (8) is used, and the detected face images are resized to 64×64 . As the focus of this study is generalizability across the four completely different generation techniques, we limit the experiments to uncompressed data. Subsequently, the models are trained on individual cropped face images from frames as shown in Figure 10.1, and the detection performance is evaluated in terms of the frame-level detection accuracy.



(a) The pipeline for the training of the anomaly detection system. The model is trained on pristine face images only.



(b) The training and evaluation pipeline of the classifier. The pre-trained anomaly detection model is used as an anomaly feature extractor.

Figure 10.1: The training and evaluation pipelines of the proposed method. UBM stands for universal background model and represents the probability distribution based anomaly extraction system.

The UBM model used for experiments is the Tensorflow implementation of PixelCNN++ (14). The default architecture, consisting of three blocks with five ResNet layers and 160 filters in each layer is used. A single model with 94 million parameters is trained for five epochs on natural images only from the training set, with a learning rate of 0.0001 on a single GPU in an end-to-end manner.

As the complexity of the detection problem is reduced in the anomaly feature extraction step to an extent that the synthesis artifacts are visible in its output (see Figure 10.4), a very simple classifier based on LeNet-5 (9) is used for detection of synthetic faces from known and unknown generation methods. The modified architecture summarized in Figure 10.2 is small enough to be trained on a CPU and has less than one million parameters. For each experiment, one classifier is trained on the available training data for 25 epochs with a learning rate of 0.001. The activation function used is the ReLU function, and to improve the convergence speed, batch normalization is used between the output of the layers and the activation function. The overall detection pipeline is shown in Figure 10.1(b).

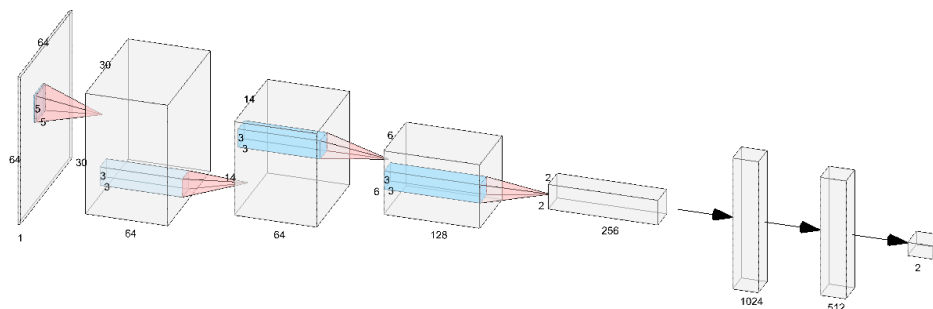


Figure 10.2: The diagram of the classifier architecture. Each convolution is followed by a 2x2 maxpooling layer and ReLU activation. The network has a total of 933,442 parameters.

10.5 Results and Discussion

In this section, we first discuss the characteristics of the anomaly extraction method and then summarize the performance of the method on both known and unknown attack detection scenarios.

10.5.1 Features

Figure 10.3 shows the histogram of log-likelihoods for images in the validation data for pristine images as well as the synthetic images. The log-likelihood values for the pristine images are higher than the synthetic images, however, there is a significant overlap between the distributions. Deepfakes show higher log-likelihood values compared to the other synthesis methods. These results show the discrimination power of the observation probability of the images for synthetic face image detection. However, the image probability distributions have significant overlap, and cannot be relied on as a high-performance detection score.

To achieve a better performance, we can rely on the pixel log-likelihood *images* extracted by the UBM model as anomaly features. Figure 10.4 visualizes examples of these *images* from the pristine data as well as the four generation techniques. In this figure, a drastic difference is observable between the pristine images and the synthetic images. The traces of the synthesis process are visible as low likelihood points in yellow and red on the image. Furthermore, each generation method shows a unique footprint in all examples. The Deepfakes have artifacts in the shape of the spliced synthetic face area over the background image. The Face2Face technique results in low likelihood pixel values on the edges of the 3D facial features such as nose and jawline. FaceSwap technique results in low likelihood areas around the eyes and the mouth. Lastly, NeuralTextures inhibits individual low-likelihood

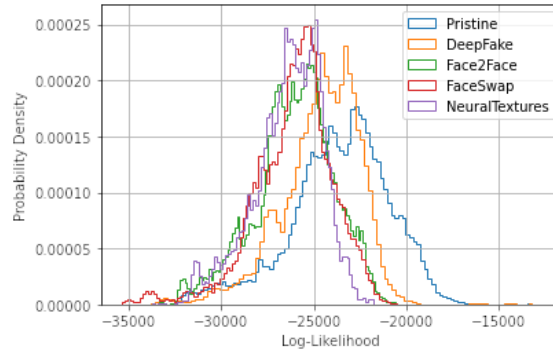


Figure 10.3: The image log-likelihood probability for pristine images and synthetic images in the development data.

pixels on the nose and eye regions.

10.5.2 Known Synthetic Face Detection

To measure the discriminative power of the likelihood images, we used the simple classifier explained in the previous section for synthetic face detection on each individual method. The results are reported in Table 10.1 along with the performance of the baseline methods from (13). The proposed method performs on par with the baseline methods despite having a smaller input image size and a much smaller number of parameters. These results confirm that the log-likelihood images conserve the information valuable for detection faithfully while reducing the detection complexity by removing the unhelpful information.

Table 10.1: The performance of the proposed method in terms of detection accuracy in known synthetic face image detection scenario in comparison with existing methods adapted from (13). (DF: DeepFakes, F2F: Face2Face, FS:FaceSwap, NT:NeuralTextures)

	Input Size	DF [%]	F2F [%]	FS [%]	NT [%]
Steg. Features+SVM(5)	128×128	99.03	99.13	98.27	99.88
Cozzolino et al.(4)	128×128	98.83	98.56	98.89	99.88
Bayar and Stamm(2)	128×128	99.28	98.79	98.98	98.78
Rahmouniet al.(12)	100×100	98.03	98.96	98.94	96.06
MesoNet(1)	256×256	98.41	97.96	96.07	97.05
XceptionNet(3)	299×299	99.59	99.61	99.14	99.36
Proposed Method	64×64	99.30	98.25	99.11	98.46

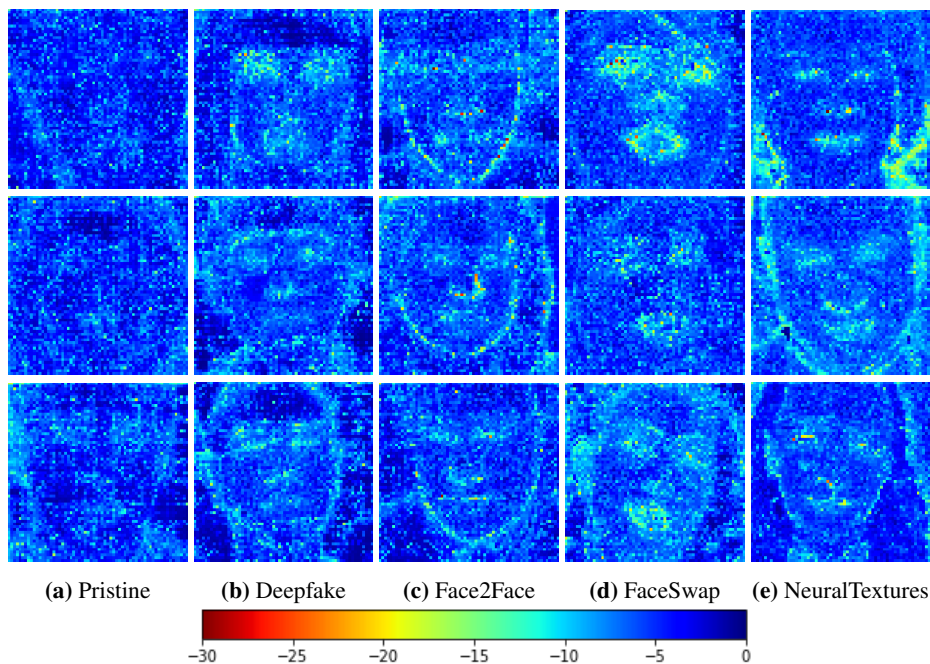


Figure 10.4: Examples of the log-likelihood output matrix of the universal background model on pristine and synthetic face images. The name of the generation method is mentioned below each column. As shown in the color bar, red signifies low log-likelihood probability, while blue signifies high.

10.5.3 Unknown Synthetic Face Detection

The performance of the proposed method in the unknown synthetic face detection scenario is summarized in Table 10.2. The proposed method shows an acceptable detection rates for all four synthesis methods while showing above 96% on three out of four in LOO generalization experiments. The performance of Face2Face method gets slight improvement over the known case due to the larger training data available in the LOO scenario.

10.6 Conclusion

In this article, we introduced a truly generalizable synthetic face image detection method which achieves an outstanding average detection accuracy of 95.73% on unknown synthetic methods. The synthetic methods are from widely different synthesis mechanisms ranging from Deepfakes from generative adversarial networks to FaceSwap. The proposed method consists of a preprocessing step where the content of the image is suppressed, and the anomaly locations and anomaly

Table 10.2: The performance of the proposed method on unknown synthetic samples in terms of detection accuracy. For each method, the system is trained on the other three synthesis data and did not observe a single sample of the method in question during training. The average detection accuracy is also reported. (DF: DeepFakes, F2F: Face2Face, FS:FaceSwap, NT:NeuralTextures)

	DF [%]	F2F [%]	FS [%]	NT [%]	Avg [%]
LOO Detection Accuracy	89.26	98.41	96.80	98.44	95.73

strengths are extracted. The classification is then done by a simple classifier. The anomaly extraction step is trained on natural images only and preserves the detection-relevant information faithfully in the form of observation log-likelihood probability. The detectors' success provides new hopes for addressing the generalization problem over widely different generation processes.

References

- [1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *WIFS*, pages 1–7. IEEE, 2018.
- [2] B. Bayar and M. C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *ACM IH&MMSec*, pages 5–10, 2016.
- [3] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.
- [4] D. Cozzolino, G. Poggi, and L. Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *ACM IH&MMSec*, pages 159–164, 2017.
- [5] J. Fridrich and J. Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012.
- [6] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.
- [7] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch. Fake face detection methods: Can they be generalized? In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–11, 2018.
- [8] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.
- [9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [10] F. Marra, C. Saltori, G. Boato, and L. Verdoliva. Incremental learning for the detection and classification of gan-generated images. *arXiv preprint arXiv:1910.01568*, 2019.
- [11] L. Nataraj, T. M. Mohammed, B. Manjunath, S. Chandrasekaran, A. Flenner, J. H. Bappy, and A. K. Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019(5):532–1, 2019.

-
- [12] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *WIFS*, pages 1–6. IEEE, 2017.
- [13] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE ICCV*, pages 1–11, 2019.
- [14] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.
- [15] J. Stehouwer, H. Dang, F. Liu, X. Liu, and A. Jain. On the detection of digital face manipulation. *arXiv preprint arXiv:1910.01717*, 2019.
- [16] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.
- [17] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR’16*, pages 2387–2395, June 2016.
- [18] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *ICML - Volume 48*, page 1747–1756. JMLR.org, 2016.
- [19] L. Verdoliva. Media forensics and deepfakes: an overview. *arXiv preprint arXiv:2001.06564*, 2020.
- [20] N. Yu, L. P. Davis, and M. Fritz. Attributing fake images to gans: Analyzing fingerprints in generated images. 2018.
- [21] X. Zhang, S. Karaman, and S.-F. Chang. Detecting and simulating artifacts in gan fake images. *arXiv preprint arXiv:1907.06515*, 2019.
- [22] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, pages 1831–1839, 2017.

Chapter 11

Article 7: Unknown Presentation Attack Detection against Rational Attackers

A. Khodabakhsh, Z. Akhtar, "Unknown Presentation Attack Detection against Rational Attackers," arXiv preprint arXiv:2010.01592, 2020. (Submitted to IET biometrics)

11.1 Abstract

Despite the impressive progress in the field of presentation attack detection and multimedia forensics over the last decade, these systems are still vulnerable to attacks in real-life settings. Some of the challenges for existing solutions are the detection of unknown attacks, the ability to perform in adversarial settings, few-shot learning, and explainability. In this study, these limitations are approached by reliance on a game-theoretic view for modeling the interactions between the attacker and the detector. Consequently, a new optimization criterion is proposed and a set of requirements are defined for improving the performance of these systems in real-life settings. Furthermore, a novel detection technique is proposed using generator-based feature sets that are not biased towards any specific attack species. To further optimize the performance on known attacks, a new loss function coined categorical margin maximization loss (C-marmax) is proposed which gradually improves the performance against the most powerful attack. The proposed approach provides a more balanced performance across known and unknown attacks and achieves state-of-the-art performance in known and unknown attack detection

cases against rational attackers. Lastly, the few-shot learning potential of the proposed approach is studied as well as its ability to provide pixel-level explainability.

11.2 Introduction

Over the last decades, there have been major break-throughs in the fields of manufacturing, computing, and communication, resulting in cost reduction as well as higher availability of manufacturing and synthesis processes to the public. Among the beneficiaries of these advancements are the attackers to biometric and forensic systems, who take advantage of these methods to devise new and more powerful attacks. Relying on the fact that face is the main modality of human communication in daily life, and the ever-growing interest in the use of face biometrics in real-life applications, methods that can realistically produce facial videos have an immense potential for abuse. The infamous Deepfake tool¹ is such an example that has repeatedly been used for the purpose of fake news generation to such an extent that a bill was passed in the US senate to report at specified intervals on the state of digital content forgery technology.

Consequently, biometric and forensic systems face new challenges every day as they have to become secure against a wider range of attacks happening at a higher frequency. Making the matters worse, the existing detection solutions are often designed against a specific attack (or set of attacks) in controlled environments and lack the capacity to face the challenges of real-life deployment. This is evident from the results of the recent Deepfake detection challenge² organized by Facebook where the best performing algorithm had a detection rate of only 65% when faced with unknown generation techniques. As such, addressing vulnerabilities of existing solutions and introduction of methods to mitigate these vulnerabilities is of utmost importance for deployment of these systems in practice.

One aspect of challenges for deployment that is rarely studied is the selection process of a rational attacker. It is expected for an attacker with an ever-growing menu of options for attacking to behave rationally and choose the most powerful attack available to him to maximize the chance of infiltration. Furthermore, as the defender does not have knowledge or access to massive amounts of data for all possible attacks available to attackers, his detector would probably be tasked with the detection of unknown attacks or attacks from which only a few training examples are available. Additionally, lack of explainability limits the use of a system in high-stake applications where explainability increases its utility when operated by a human supervisor.

¹<https://github.com/deepfakes/faceswap>

²<https://www.kaggle.com/c/deepfake-detection-challenge>

In this article, to address these challenges, a game-theoretic approach is considered for the formulation of the interactions between the attacker and the detector. Resulting from this, an optimization criterion is formulated and a set of requirements are defined for designing the detector accordingly. To tackle the problem of unknown attack detection and few-shot learning, the use of unbiased compressed feature sets is proposed, and for targeting the optimal performance, a new loss function is defined faithful to the formulated optimization criteria. Finally, the explainability of the proposed method is demonstrated with a few examples. The rest of this article is organized as follows: In section 11.3, the related literature is reviewed and a theoretic basis for the proposed approach is established in section 11.4. Afterward, the proposed method is introduced in section 11.5 and the case study experiment setup is explained in section 11.6. Finally, the results of the experiments are reported and analyzed in section 11.7 and 11.8 and the article is concluded in section 11.9.

11.3 Literature Review

Considering the task of forgery detection or presentation attack detection on the face modality, there exist three relevant threads of research. First is the field of multimedia forensics, and more specifically, anti-counter forensics (CF). This thread of research takes an adversarial view on the problem and tries to optimize the performance of the detection system facing an adversary who is actively working towards undermining the performance of the detector. Second is the field of presentation attack detection (PAD) in which the objective of the detector is to secure a biometric system against attacks from different presentation attack species (PAS). Lastly, the newly established thread of Deepfake detection is considered that was initiated to address new phenomenon of availability of automated open-source photo-realistic digital video manipulation techniques on the internet. In this article, the terminology proposed for the field of presentation attack detection is relied on. Consequently, the act of forgery is called attacking the detector and generation techniques used by the forger are called attack species.

11.3.1 Anti Counter Forensics

The majority of solutions in the literature are designed neglecting the fact that an attacker works actively to undermine the performance of the detection system (11). To address the vulnerability to CF attacks, many anti-CF techniques have been developed, with a focus on detecting the traces left by CF techniques. Anti-forensic techniques often target a specific CF technique, and as a result, an obvious problem occurs when the attacker anticipates the use of the anti-CF technique and adjust accordingly. In turn, the defender would need to resort to the introduction of a new detection system to detect the anti-CF attacks, resulting in a never-ending iterative

loop with unforeseeable outcomes (70). A possible solution to this problem is to design techniques that are intrinsically more resistant to CF attempts (15)(57). For example, in (27) the authors proposed the use of second-order statistics derived from co-occurrence matrices and show robustness against CF attacks. Zhang et al. (84) used a reduced feature set based on assumptions on the attacker's data manipulation strategy. A combination of one-class and two-class classifiers is proposed in (10). Another interesting approach is the randomization of the feature selection process (17). In (5) and (6), the authors propose the reuse of the original feature space for the detection of CF attacks by retraining for the task of double JPEG compression detection. The third group of solutions rely on game theory to model the interactions between the detector and the attacker and improve the performance of the detector at the final equilibrium (7) (69)(8). All aforementioned methods address the case where attacker has a limited choice of CF attacks and do not consider selection process of attacks in optimization of the detector.

11.3.2 Presentation Attack Detection

Similar to anti-CF techniques, the existing PAD research can be categorized into three branches: (1) PAD systems that address specific PASs, (2) PAD systems that increase or optimize the feature set to detect a higher variety of attacks, and finally, (3) PAD systems that rely on game theory to model the interactions between attacker and defender and optimize the PAD performance accordingly.

The early PAD methods addressed PAD for specific PAS, examples of which are methods relying on features such as blinking, head movement, and textures (56, 40, 58). Different Features have been used (80), (71), (59),(13), such as 2D Fourier spectrum (78), local binary patterns (LBP) (26). Authors in (83), (44) and (62) presented central difference convolutional networks, layer-by-layer progressive compact space generation, and style transfer techniques, respectively. Many PAD methods rely on an augmented feature set using additional hardware. Examples include 3D depth camera (77), multi-spectral camera (85), and microphones (18). However, these techniques require the addition of often expensive hardware to the pipeline, which may not be feasible in all applications. A few studies tried to use generalizable feature sets for PAD. In (78), the authors propose the use of image distortion analysis. The use of 25 general image quality features for PAD is investigated in (32). In (42), a regression function is learned to map the image quality assessment scores. The use of pixel-level supervision for improving features is investigated in (47) and regional self-supervision in (28). A limited number of studies tried to address the generalizability of PAD systems (63) (9) using a one-class classification approach (4) (55), deep metric learning model (61), and zero-shot (48). To the best of our knowledge, no game-theoretic approach is proposed to model interactions between attacker and defender.

11.3.3 DeepFakes Detection

Several approaches have been proposed for detecting DeepFakes, such as lack of asymmetry in computer-generated imagery(24), spatio-temporal deformations of a 3D face model (25), use of periodic blood flow (20), generation flaws (51), blinking (45), and blood flow (19), face warping artifacts (46), use of face landmark locations (82), head pose consistencies (81), mesoscopic features (1), architecture-specific GAN fingerprints (41) (52) (49), convolutional neural networks (CNNs) (64), attention mechanism (23) and capsule networks(54), long short-term memory (LSTM) networks (35), recurrent CNNs (65), and optical fields (3). However, such detectors tend to overfit to the known attacks and show limited generalizability (37). The problem of generalization has been studied in a few articles using, e.g., auto-encoder in (21) and (30), incremental learning in (50), pre-processing artifacts in (79), transferability of the network in (76), time dimension with attention mechanism in (31). Other works are (43), (22), (2) and (74). Most studies have a heavy focus on DNNs/GAN generated artifacts and do not consider other types of manipulations. Also, none of the aforementioned studies take into account the rationality of the attacker nor the case in which the attacker has multiple choices of attack species. A summary of the most representative works in anti counter forensics, presentation attack detection and deepfakes detection is presented in Table 11.1.

11.4 Theory

In this section, I introduce the definition of a rational attacker and formulate such an attacker's pay-off equation and decision-making process. Furthermore, I discuss the detection strategy facing such an attacker and define the requirements for a PAD system accordingly. Lastly, I justify the use of one-class detection techniques based on generative models for unknown attack detection.

11.4.1 Rational Attacker

In most existing literature the selection process of the attackers for which attack species to use is neglected and assumed to be that of random selection, resulting in the proposed detectors having fundamental weaknesses. A rational attacker is defined as an attacker who, knowing the pay-offs to his possible choices, selects the one with the highest pay-off. From a game-theoretic perspective, the interactions between an attacker x and the defender can be modeled by a sequential asymmetric game in which the defender chooses a detector after which the attacker administers their attack of choice. An attacker would have to choose among a set of attack species A_x which represents all his options. The pay-off u_i for the attacker for an

Table 11.1: Representative works on anti counter forensics, presentation attack detection and DeepFakes detection. AUC = Area Under the Curve; ER = Error Rate; EER = Equal Error Rate; HTER = Half Total Error Rate; Acc = Accuracy.

References	Approach	Database	Performance	Year
Anti Counter Forensics				
De Rosa <i>et al.</i> (27)	Second-order statistics derived from co-occurrence matrices	UCID.v2	AUC = 90%	2015
Barni <i>et al.</i> (6)	Higher-order features both in spatial and frequency domain + SVM	RAISE	AUC = 98%	2017
Chen <i>et al.</i> (17)	Randomisation of the feature space + SVM	RAISE-2k	ER = 90%	2019
Presentation Attack Detection				
Boulkenafet <i>et al.</i> (13)	Speeded-Up Robust Features and Fisher Vector Encoding + Softmax classifier with a cross-entropy loss function	Replay-Attack, CASIA, MSU	EER = 0.1%, 2.8%, 2.2%	2017
Wang <i>et al.</i> (77)	CNNs-based texture features + depth information from Kinect + SVM	In-House	HTER = 1.2%	2017
Liu <i>et al.</i> (48)	Zero-shot Deep Tree Network partitioning spoof samples into semantic sub-groups in an unsupervised fashion	CASIA, Replay-Attack, MSU	AUC = 90%, 99.9%, 81.6%	2019
Li <i>et al.</i> (44)	Deep-learning system (i.e., CompactNet) for learning a compact space tailored for face PAD	Replay-Attack, OULU-NPU, HKBU-MARs V1	HTER = 0.7%, 6.0%, 14.8%	2020
DeepFakes Detection				
Li <i>et al.</i> (45)	Eye-blinking analysis using Long-term Recurrent Convolutional Networks	CEW, Generated DeepFakes	AUC = 99%	2018
Nguyen <i>et al.</i> (54)	CNNs-based capsule network	FaceForensics	Acc = 83.33%	2018
Du <i>et al.</i> (30)	Locality-Aware AutoEncoder	FaceSwap	Acc = 68.06%	2020

attack $a_i \in A_x$ can be formulated as:

$$\begin{aligned}
 u_i &= r(1 - p_i) - c_f p_i - c_i \\
 &= r - p_i(r + c_f) - c_i \\
 &\cong -p_i(r + c_f) - c_i
 \end{aligned} \tag{11.1}$$

where $r > 0$ is the reward for a successful attack, p_i is the probability of detection (detection rate) for the attack species a_i , $c_f > 0$ is the cost of failure for the attacker, and $c_i > 0$ is the cost of the attack. To account for the budget of the attacker, I assume the budget allows all attack species that are in A_x , and any attack that requires a higher budget is excluded from A_x .

The attacker can, with the help of trial and error as well as consultation from the experience of other attackers, have an accurate estimate of p_i for $a_i \in A_x$. The attacker's goal is to choose an attack species that maximizes the pay-off function if the highest pay-off is higher than the pay-off of not attacking the system. As $r + c_f$ is constant for every individual attacker, the optimization corresponds to the selection of an attack species with the lowest weighted sum depending on p_i and c_i . In practice, it is fruitful for the defender to take c_i into account, and low-cost attack species are expected to occur more frequently than the high-cost ones. However, because measuring c_i for individual attack species falls outside the scope of this study, I assume the worst-case scenario in which the cost of all possible attack species are assumed zero, enabling all attackers to use more effective attacks regardless of the cost of the attack, as long as their budget allows the attack to be included in A_x . Consequently, the pay-off formula boils down to $u_i \cong -p_i$, and the choice of the attacker would be the attack with the lowest p_i , referred to as the *most powerful attack* (MPA). The values for p_i s depends solely on the choice of the detector by the defender.

11.4.2 Multiple Attackers

A detection system faces not only one attacker but different attackers with different sets of A_x . Gathering statistics about the availability of attack species to the attackers would provide further knowledge about the probability of observing a specific MPA during the detection scenario. However, as such statistics are often not available for individual attackers, a conservative approach would be to construct a union set of all possible attack species for groups of attackers A_{X_k} and assume all attack species in A_{X_k} are available to all attackers from category k . By doing so, the PAD scenario is further simplified as the distinction between individual attackers collapses and all attackers in each category become identical.

For example, using the budget as a categorizing factor, the attackers can be categorized to low-budget and high-budget and the attack set for low-budget attackers A_{X_l}

and high-budget attackers A_{X_h} can be constructed. Next, using the probability of an attacker belonging to each category $p(X_k)$ and the performance of the detector D on the MPA from that category $perf(A_{X_k}|D)$, the expected overall performance of the system can be estimated as $\sum_k p(X_k) \times perf(A_{X_k}|D)$. Other examples of categorizing factors are expertise, time-budget, and access to unknown attacks or anti-forensic attacks. As the categorization of the attackers and calculation of the probability of attackers belonging to each category falls outside the scope of this study, I assume a single category A_X for all attackers. From here on, I use the term *attacker* to refer to the hypothetical attacker that can administer all attacks in A_X .

11.4.3 Detection Strategy

For deciding the best detection strategy, the accurate estimate of detection rate for individual attack species by the attacker can be interpreted as equivalent to having full knowledge over the detection performance over all $a_i \in A_X$. Due to the sequential nature of the game, the defender needs to choose p_i s for individual attack species before the attacker decides which attack to choose. Subsequently, the rational attacker will choose the MPA which has the lowest detection rate depending on the defender's choice of detector.

Let us assume the set A denotes all possible attack species. In A , two attack species are considered different if they have different manufacturing/generation process, including generation parameters such as manufacturer expertise, quality, and obfuscation. From the perspective of an attack detection system, an attack species can be categorized into one of three subsets: (1) Known attack species (A_k) to which detector is exposed in training process and its performance optimized, (2) Unknown attack species (A_u) to which detector is not exposed to and its performance is unknown, and (3) Anti-forensic attack species (A_a) signifying the attack species that are designed with knowledge over the weaknesses of the detector in mind and render the detector useless. These three subsets cover the whole set A . It is important to mention that these subsets can be expanded as new attacks are invented (become possible) and added to A .

To the extent of the knowledge available to the defender, A_k constitutes the set of all possible attack species, all while the attacker may be able to administer attacks falling outside A_k . The defender can know the detection rate for attack species in A_k and optimize them accordingly, however, he cannot know the detection rate for attack species in A_u . The best the defender can do in this case is to make an educated guess of what the minimum detection rate can be for attack species in A_u . To achieve this, every individual attack species in A_k can be left out as an imaginary unknown attack species during training, and the minimum detection

rate across all leave-one-out (LOO) trials can be used as a rough estimate of the detection rate across MPA in A_u .

The pay-off for the defender can be formulated as

$$v_i = -c_d - c_m(1 - p_i), \quad (11.2)$$

where c_d is a constant cost of detection, c_m is the constant cost of missed detection, and p_i is the probability of detection of attack a_i which matches the definition of p_i for the attacker. Knowing that the attacker will choose MPA, i.e. the attack species with the lowest p_i , the defender's best strategy would be to maximize the minimum p_i across both A_k and A_u to maximize v_i . There is a further objective of reducing the detection cost such that c_d is not prohibitively large, i.e. $c_d \ll c_m(1 - p_i)$. The defender needs to choose to maximize p_i either for $a_i \in A_k$ or $a_i \in A_u$, while limiting c_d according to the application dependant c_m . As mentioned in Section 11.4.2, it is also possible to categorize the attackers to the ones with access to attack species from A_u and the ones without, and define an objective function that takes into account the minimum detection rate over both A_k and A_u . Yet, as the defender does not possess any knowledge over A_u , it logically follows that he does not have any knowledge about the probability of the attackers being able to use attacks that belong to A_u either, and would need to resort to an educated guess of the probability instead. In this study, I try to maximize the detection rate for MPA from A_k and A_u independently, corresponding to the cases where $A_X \subset A_k$ and $\exists a_i \in A_X, a_i \in A_u$ respectively, and propose a fusion scheme that can be used to combine the resulting detectors without a significant loss of performance in either case.

11.4.4 Requirements

Following the aforementioned explanations, it is evident that the common approach towards improving the average detection performance across known attacks is not viable when the detectors are deployed and face rational attackers. Consequently, a more sophisticated approach is needed to be taken based on these analyses where the performance of a system is optimized considering the MPAs, unknown attacks, and adversarial attackers. To this end, the following set of requirements can be defined as guidance for the development of a robust detection system:

- It should have an optimal minimum detection rate across known attack species.
- It should have an acceptable minimum expected detection rate across unknown attack species.

- It should be able to learn to detect an unknown attack species optimally once it becomes known by a few examples.
- The cost of detection should not outweigh the cost of miss-detection.
- It should be robust against adversarial attacks.

The first two requirements can be directly justified according to the formulation of the problem provided in Sections 11.4.3. The third requirement follows directly from the first two for the case when an unknown attack species becomes known. In this case, the newly known attack species qualifies for a known attack species and should follow the first requirement, even though there might exist only a limited number of available examples from it. Consequently, the detector should be able to learn to increase the detection rate of the previously unknown attack species to match that of known ones.

There are certain solutions in the literature that attempt to address the last requirement (17), however, to the best of our knowledge, there exists no method to prove the robustness mathematically, and empirical proofs would be limited to the specific anti-CF attacks that are considered. Consequently, for a detector to achieve robustness against adversarial attacks, it needs to survive the test of time. As such, fulfilling this requirement falls outside the scope of this study.

11.4.5 Generation-based Feature Sets

It is common practice to rely on discriminative models for the detection of attacks. However, the objective of a discriminative model requires it to focus on the discriminative features between bona fide (BF) and known attack species. Consequently, these models do not learn discriminative features that are not directly useful for the detection of the presented known attacks. As such, these models often fail to infer information on unknown attacks where the discriminative feature set is different from the learned ones. In contrast, the objective of a generative model trained on BF data requires it to model all variability in the BF data to the capacity of the model, and because of this, does not over-represent some features while under-representing the others. Using feature sets extracted by a generative model, a detector is expected to be more robust to unknown attack species as it has access to more informative feature sets (60), only limited by the capacity of the generator in learning the feature set corresponding to BF data (53). Namely, GANs have shown to be more effective for open-set recognition (29). Hence, generative models can be used for anomaly extraction more effectively in unknown attack detection scenarios. Even though the features extracted using the generative model are not optimized for detection and might not outperform the discriminative

features used by a discriminative model on known attacks, it can be demonstrated that they would generalize better on unknown attack species as they have no bias regarding what the attack should look like (29)(53)(60).

11.4.6 Minimax Objective Function

Considering the known attack detection scenario, another limitation of most existing discriminative detectors is the reliance on the average loss for optimizing the parameters. However, as argued in Section 11.4.3, the performance of a detector against a rational attacker is not determined by the average detection rate, but the detection rate on the MPA. Accordingly, optimizing the average detection rate does not necessarily translate to the optimization of the detection rate against the MPA all while posing challenges for the detection of the under-represented attack species. In response to this limitation, objective functions that rely on minimizing the maximum loss (or maximizing the minimum gain) are proposed as a reliable alternative, for which the GAN loss (34) is a famous example.

11.5 Proposed Method

According to the requirements defined in Section 11.4.4, two separate detection methods are proposed for both scenarios of known and unknown attack detection. Furthermore, a fusion mechanism is introduced to combine the decision of the two detectors for a unified solution with few-shot learning capabilities. Both proposed methods rely on pixel-level generator-based anomaly features and its compact representation extracted to achieve better performance across unknown attack species. For the purpose of known attack detection, a new loss function is introduced which follows the defined objective of maximizing the minimum detection rate. For the purpose of unknown attack detection, I construct a generator-based one-class detector that relies on attack-unspecific anomaly-sensitive information extracted from the detection pipeline.

11.5.1 Pixel-Level Probability Distribution Modelling

A distribution model for BF images can provide an ideal model for presentation attack detection, as it would be a generative model that contains the complete feature-set and can also provide a single detection score in the form of the likelihood of an observation to the BF distribution. However, due to the complexity of the distribution of BF images, the large amounts of data needed to train such distribution properly, and finally the curse of dimensionality, it is deemed impractical. However, by breaking down the problem into modeling segments of an image rather than the whole image, there exist practical solutions.

PixelRNN (75) is a generative model that models the pixel intensity value probabil-

ity distribution conditioned on previous pixel values in raster order. This approach can be used to calculate log-likelihood values for observing individual pixels in an image, and once these values are aggregated, they can be used to estimate the log-likelihood of observing the input image as a whole. The pixel-level log-likelihood values can further be used for the localization of low-likelihood pixels (anomalies) in the input. In the proposed approach, the aggregated log-likelihood value is used as the first anomaly measurement for the one-class classifier, and a dimensionality reduction scheme is proposed for simplification of the description of the localization information for extracting the second anomaly measurement which are also used for training the proposed discriminative detector for the known attack detection (Fig. 11.1).

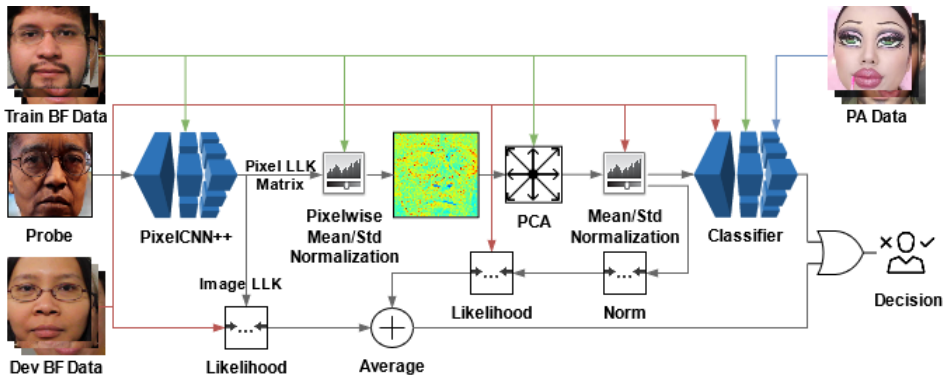


Figure 11.1: The pipeline of the designed detection mechanism for both the discriminative classifier and the generator-based one-class detector. The red, green, and blue arrows signify the use of the data in the training of the models pointed to. The gray arrows correspond to the flow of the probe data.

11.5.2 Dimensionality Reduction

The pixel-level log-likelihood values provide valuable information about the severity of the anomalies at each location in the image. However, dealing with features the same size as the input video proves challenging, especially when the amount of training data is limited. To tackle this problem, the following dimensionality reduction scheme is proposed: As the location of anomalies is expected to remain roughly constant in a video, one can average the pixel-level log-likelihood values across the cropped face frames across the whole input video. This step will serve two purposes, firstly it collapses the data in the time dimension, and secondly, it reduces the noise in the frame-level representations. Next, I use a principal component analysis (PCA) transformation learned on BF data to reduce the dimensionality further (Fig. 11.1).

PCA transformation extracts the directions where the variability of the BF data is most explained. It can also be used to extract the directions in which the input data shows little variability. The components for which the BF data shows little variability fits well with the definition of anomaly features, and they are a good representation of the similarities between the BF samples. Additionally, the unexplained variability of input after transformation to the PCA space can provide further anomaly clues. This unexplained variability can be measured as the distance between the input and its projection on the PCA hyper-plane. Thus, I augment the PCA transformed features with the measurement of unexplained variability. The resulting compact representation manages to conserve the discriminative information in the input video effectively while reducing the dimensionality further by a factor of ≈ 1000 .

The amount of shift across the PCA dimensions where BF samples show little variability, along with the unexplained variance measurement can directly be used for one-class detection. To reduce it to a single score, the energy of the input across these dimensions can be measured by calculating the norm of the signal across them. However, as the unexplained variability is on a different scale compared to the PCA transformation values, a normalization step is required. Normalization can be done by making the distribution of the BF samples across these dimensions zero-mean unit-variance.

11.5.3 Categorical Margin Maximization Loss

As the performance of a system in deployment is measured according to its performance for the MPA, a new minimax loss function needs to be introduced that optimizes the detector towards achieving the highest MPA detection rate possible. In this approach, motivated by the success of the triplet loss (67), I introduce categorical margin maximization loss (C-marmax) that weighs attacks exponentially according to the difficulty of classification, and thus focuses on reducing the loss from the most difficult samples (MPAs) at each batch during the training. Using C-marmax, the network transforms the aforementioned compact representations to embeddings on a unit hyper-sphere where the distance between the BF data and attacks are maximized while the distance between attacks from the same species, as well as BF samples to each other, is minimized. In this loss, the distances between attacks from one species to other species are ignored as we don't have any information about the similarity or dissimilarity between distribution across any two attack species. Hence the detector is *categorical* as it only considers distances between observations from different categories (i.e. BF vs attacks) for calculating the loss value. Finally, to exaggerate the loss from samples belonging to the MPA and suppress the loss from other attack species, the loss attributed to the anchors is exaggerated according to the distances such that the network pays more attention

to marginal anchors to fulfill the objective of maximizing the minimum detection rate.

In attack detection scenarios, there are a few classes, and it is possible to rely on the distance to the center of distribution in a batch rather than the distance between individual samples. To this end, in each batch, I compute the location of the center of distribution for each attack species as well as BF data on the unit hyper-sphere, and according to the label of the inputs, I use these centers to measure the distance of the anchor to the positive distribution d_p and the negative distribution d_n . To achieve the maximum margin possible between the distribution of BF samples and PA samples in the embedding space, a fixed margin is not defined. Instead, the ratio $\frac{d_p}{d_n}$ is used for the maximum d_p and minimum d_n in a batch from each class, requiring the numerator to be minimized to zero, while the denominator is maximized to the maximum possible value of 2 on the unit hyper-sphere. To avoid the loss value to become infinity when d_n is zero, the ratio is modified to $\frac{d_p}{d_p+d_n}$ which is equivalent to $\frac{d_p}{d_n}$ when $d_p \ll d_n$. Furthermore, to exaggerate the loss for marginal observations (where d_p is high) in comparison to non-marginal observations (where d_p is low), exponentiation is used, and the resulting formula becomes $(\frac{d_p}{d_p+d_n})^g$.

As the defined loss does not maximize the distance between centers of distributions directly, to assure that the center of distributions are far from each other, the minimum distance between two centers are floored at $\sqrt{2}$ corresponding to 90 degrees on the unit hyper-sphere, with a second loss term. The final loss function is summarized as follows:

$$\begin{aligned} loss_m &= \left(\frac{\max\{d(a, C_p)\}}{(\max\{d(a, C_p)\} + \min\{d(a, C_n)\})} \right)^g \\ loss_c &= \max\{\min\{\sqrt{2} - d(C_p, C_n)\}, 0\} \\ loss &= loss_m + 0.1 \times loss_c \end{aligned} \tag{11.3}$$

where, d stands for euclidean distance, a signifies the anchor, C_p is the center of the positive class, C_n is the center of the negative class, g is the exaggeration factor, $loss_m$ is the margin loss, and $loss_c$ is the center loss. During decision making, the euclidean distance to the center of BF distribution can be used for scoring. This distance can further be converted to an attack detection probability value by division by 2.

In comparison to the triplet-loss, the proposed modifications result in a tunable exaggeration of the loss in misclassified samples and suppression of the loss in the correctly classified ones and relax the need for a fixed margin constant. Having no constant margin, the network can continue training even after a specific margin

is achieved between the classes until the maximum margin on the hyper-sphere is reached. Furthermore, by using the center of the distribution instead of the distance between individual anchors, the loss becomes less stochastic, allowing faster convergence. To the same effect, the categorical nature of the proposed loss relaxes the untrue assumption that all attacks come from the same distribution regardless of their corresponding attack species.

11.5.4 Unknown Attack Detection

As argued in Section 11.4.5, a discriminative model may overfit to certain discriminative features that correspond to the bias in known attack species used in training. This also holds true for the presented C-marmax loss, as even though it tries to achieve a balanced attack detection performance across known attack species, it may exclude discriminative features that may be important for the detection of unknown attack species. As such, to detect unknown attacks, a one-class detector is proposed which does not have a bias towards any specific attack species, or in other words, for it all attacks are unknown. As explained in Section 11.5.1, the log-likelihood value of observing an image serves as a good general-purpose anomaly detection measure. However, this metric does not include the other important discriminative feature available in the pixel-level log-likelihood data, namely the location information. As explained in Section 11.5.2, the location relevant anomalies can be represented by the components in a PCA transform trained on BF where the BF data show the least variability. Furthermore, this representation can be augmented by the unexplained variance in the form of the distance of an observation to the PCA hyper-plane. Finally, the energy of the signal across the resulting representation after normalization can be used as an anomaly score. Following these steps, a second location-sensitive anomaly measure is derived. Assuming a Gaussian distribution for BF scores for both anomaly measures, using the BF score distribution, one can calculate the likelihood of an observation belonging to this distribution as the final probability score. For the final score of the one-class detection scheme, I simply average the two resulting likelihood scores from the log-likelihood measure and the PCA-based measure (Fig. 11.1).

To fuse the probability scores from the discriminative detector and the one-class detector when they are employed together, I use the following logic: If the discriminative detector decides that a sample is an attack, it most certainly is one. However, if the discriminative detector decides that the sample is a BF, the defender cannot be sure that the sample is a BF as it might come from an unknown attack. So the one-class detector is to be consulted for a decision. This two step decision logic can be interpreted as using an *OR* gate on the decision of the discriminative and the one-class detector decisions. However, as both systems provide a probab-

ity scores rather than a decision, considering that $A \vee B = A + B - AB = \overline{\overline{A} \times \overline{B}}$, the following fusion formula is proposed that mirrors the logic level decision making:

$$\begin{aligned} p_{PA}(x|D, O) &= 1 - p_{BF}(x|D) \times p_{BF}(x|O) \\ &= 1 - (1 - p_{PA}(x|D)) \times (1 - p_{PA}(x|O)) \end{aligned} \quad (11.4)$$

where p_{PA} corresponds to the probability of belonging to the attack category, p_{BF} corresponds to the probability of belonging to the BF category, and O and D correspond to one-class and discriminative detector models.

11.6 Experiment Setup

For measuring the effectiveness of the proposed method, its application on both tasks of presentation attack detection and Deepfake detection are considered. In this section, a description of the datasets used is provided, followed by the parameters used in training. Lastly, the measures used for evaluation of the method are described.

11.6.1 Datasets

To show the performance of the proposed method for presentation attack detection, the SiW-M dataset³ (48) is selected due to its large collection of presentation attack species. Similarly, the FaceForensics++ dataset⁴ (64) is chosen for the task of Deepfake detection as it contains the widest choice of species between the available datasets.

SiW-M

This dataset consists of 660 BF videos from 493 subjects from diverse ethnicity and age. Furthermore, it includes 966 PA videos from 13 different PAS collected under various environmental conditions, extreme face pose angles and lighting conditions. The videos are around six seconds in length. This dataset is specifically designed for the evaluation of generalization performance across unknown PAS. The attack species in this dataset are categorized into replay, print, mask, makeup, and partial attacks. The PAS available in this dataset are form a diverse set of attacks including print and display attacks as well as transparent masks and impersonation makeup. This dataset also includes PAS corresponding to partial attacks.

For training the models, 530 randomly chosen BF videos are used, while 65 randomly chosen BF videos were kept for development purposes, leaving 65 videos

³<http://cvlab.cse.msu.edu/siw-m-spoof-in-the-wild-with-multiple-attacks-database.html>

⁴<https://github.com/ondyari/FaceForensics>

for testing. For training the classifier in the unknown case, a LOO setup is used and for each attack species, all the videos from other attack species are used for training, along with the training and development BF data. For few-shot learning, an additional randomly chosen one or five videos from the targeted attack species are included in the training, while in the known case 50% of the videos are included.

FaceForensics++

FaceForensics++ dataset contains four PAS corresponding to Deepfakes⁵, Face2Face (73), Faceswap⁶, and Neural Textures (72). The dataset contains 1,000 BF videos and 1,000 videos from each PAS, each split into three sets, reserving 72% for training, 14% for validation and allocating 14% for evaluation. The videos are collected from YouTube and after manipulation, recompressed in three video qualities for evaluation of performance under various compression levels. For the purpose of analyzing performance over unknown attacks, only the non-compressed version of the data is used. Similar to the SiW-M dataset, both known and LOO unknown attack detection experiments are considered.

11.6.2 Parameters

The proposed method has a number of parameters corresponding to face detection, the pixel-level log-likelihood extraction model, the PCA model, and finally the classifier. In this study, the videos are considered as a set of frame images. The face region is extracted in each frame after face detection using the Dlib toolkit (38), and the cropped faces are resized to 128×128 .

The overall pipeline of the proposed detection mechanism is visualized in Fig. 11.1 along with information about where the training data, development data, and known attack data is used. The input image is first processed by the PixelCNN++ model trained using the training data, resulting in an aggregated observation log-likelihood and pixel-wise log-likelihood matrices. The aggregated observation log-likelihood is compared to the distribution of BF values learned from development data to acquire the first generator-based anomaly measure. The pixel-wise log-likelihood matrices are further normalized to zero-mean unit-variance using the distribution of pixel values in the training data before applying the PCA transform. The PCA transform is learned using the training data, and the PCA transformed representation is augmented with the unexplained variance measure and normalized to zero-mean unit-variance across all dimensions using the development data. Then after sorting the components based on the explained variance of training data in descending order, the last components are used for calculating the norm. This

⁵<https://github.com/deepfakes/faceswap>

⁶<https://github.com/MarekKowalski/FaceSwap/>

value is then compared to the distribution of BF scores learned on development data for calculating the second generator-based anomaly measure. The first and second probability scores are combined by averaging, resulting in a single one-class classification score. The augmented and normalized PCA representations are then passed to the discriminative classifier trained on BF data from training and development set along with attack data from known attacks.

PixelCNN++

For pixel-level log-likelihood matrix extraction, a PixelCNN++⁷ (66) model is trained on the resized cropped face images extracted from the BF training data. The model consists of three hierarchies with five ResNet layers in each, with 160 filters with a receptive field of 3×3 in each layer, resulting in 95 million parameters. Concatenated ELU (68) is used for activation and pixel intensity values are modeled using 10 logistic distributions. For regularization, dropout with a probability of 50% is used. The model is trained with a batch size of one and the ADAM (39) optimizer with a learning rate of 10^{-5} is used for 500 epochs on a single randomly chosen frame per training video in each epoch.

The log-likelihood matrix is then generated by concatenating the pixel log-likelihood values for each of the 10 logistic distributions for each color channel, resulting in a matrix of size $128 \times 128 \times 30$. For calculating the log-likelihood of observing the video, the likelihood of observing each individual frame is calculated using the weighted sum for the individual logistic distributions across the whole cropped face image. These values are then averaged across time to measure the average log-likelihood of the observed input video to be used for one-class detection. For extracting location-sensitive features, after averaging the pixel-level log-likelihood matrix values across the whole input video, at each pixel location, the distribution of log-likelihoods are normalized such that the BF training data has a distribution of zero-mean unit-variance, resulting in a matrix of size $128 \times 128 \times 30$ per video.

Principal Component Analysis

In the next step, these matrices are extracted from the BF training data to train a PCA model with sorted components according to the explained variance across these components in descending order. Unexplained variance is measured by calculating the euclidean distance between each input and its projection on the PCA hyper-plane and added to the end of the PCA representation. The PCA representation is normalized to have zero-mean unit-variance for BF data from the validation set. For one-class detection, to measure the energy of the input video across the last 10% of the PCA representation, the norm after normalization is used. Using

⁷<https://github.com/openai/pixel-cnn>

the distribution of the norm values across the validation data, a single Gaussian model is trained for calculating the likelihood of a given input to the BF distribution. The same approach is taken for the video log-likelihood values collected directly from the output of the PixelCNN++ model. These two likelihood values are averaged to calculate the final score of the generator-based one-class detector.

Classifier

The PCA representation is also used for the training of the discriminative classifier using the aforementioned loss function. A DNN model with four hidden layers, each with 512 ReLU activated units is trained for mapping its input to the L2 normalized embedding space of six dimensions (Fig. 11.2). Due to the limited amount of training data available for training the classifier, dropout regularization with a rate of 50% is used on the output of each hidden layer, along with L2 regularization with a factor of 10^{-6} . Oversampling is done by using random segments of the training videos and their vertically flipped copies while testing is done on the whole test videos. The training data is balanced by repetition to have 50% BF samples and $\frac{50\%}{\#PAS}$. The loss function only has one tunable parameter g , which was set to two to achieve fast conversion. Training is done with a batch size of 128 for 100 epochs with a fixed learning rate of 10^{-3} using the ADAM optimizer. Finally, the detection probability score is calculated by measuring the Euclidean distance of the embedding to the average of the validation data embeddings divided by two. The fusion between the probability score calculated by the generator-based one-class detector and the discriminative detector is done using the formula in Section 11.5.4.

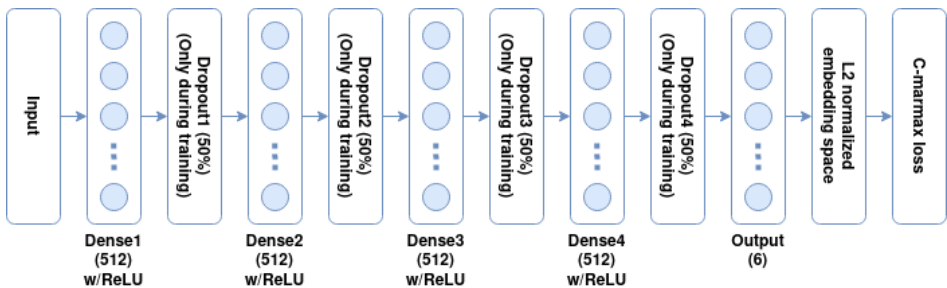


Figure 11.2: The architecture of the classifier network.

11.6.3 Metrics

To evaluate the performance of the proposed system, the threshold less equal-error-rate (EER) metric is used. EER measures the error rate when the missed detection percentage is equal to the false alarm percentage. For evaluation of performance

across all attack species, the EER value for the MPA is chosen by measuring the maximum EER across all species following the arguments represented in Section 11.4.3. Furthermore, the detection error trade-off (DET) curve is used for showing the missed detection rate for each false alarm value. Missed detection corresponds to the bona fide presentation classification error rate (BPCER) and false alarm corresponds to attack presentation classification error rate (APCER) in ISO/IEC 30107 terminology⁸. In the experimental result tables, we report are $ACER@APCER=5\%$. The $BPCER@APCER = 5\%$ can be calculated as $BPCER = (ACER \times 2) - 5\%$.

11.7 Presentation Attack Detection

In this section, the adequacy of the proposed generator-based anomaly representations is first explained. Later, the performance of the proposed method based on these representations is evaluated and compared to the existing solutions in both known and unknown attack detection scenarios. Lastly, the few-shot learning capacity of the proposed method is investigated and the computational cost of the pipeline is reported.

11.7.1 Representation Adequacy

Fig. 11.3 shows examples of the log-likelihood matrices extracted by the Pixel-CNN++ model for sample frames from BF data as well as each attack species. It can be seen that BF data shows few single anomaly pixels corresponding to the natural variations in the BF frame as well as anomalies around the location of the glasses. However, each attack species shows its own pattern of anomalies corresponding to the locations where it is observed. For example for the obfuscation makeup attack, the anomalies correspond to where the eyebrow and beard lines are drawn, for the mannequin attack they correspond to the skin regions, for the paper mask to the fold locations, and for the replay attack to the overexposed regions of the face. These examples show the capacity of the representation to provide explainability at pixel-level.

To further analyze the unique patterns from each attack species, the average log-likelihood matrix for each species is presented in Fig. 11.4. The average and standard deviation of log-likelihood values for training BF data are shown in the first column. From these two images, it can be seen that most of the natural variability in the training data corresponds to the eye and the nasal dorsum as well as the background, while the periocular region of the face contains a lower natural log-likelihood. After normalization of the average log-likelihood matrices for test data using these two matrices, it can be seen that the test BF data matches

⁸<https://www.iso.org/obp/ui/iso>

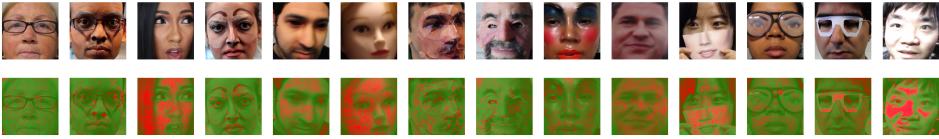


Figure 11.3: Example frames from BF and each PAS from the SiW-M dataset along with their corresponding log-likelihood matrices below them. Red pixels show the location of anomalies from the perspective of the PixelCNN++ model. From left to right: BF, Cosmetic Makeup, Impersonation Makeup, Obfuscation Makeup, Half Mask, Mannequin, Paper Mask, Silicone Mask, Transparent Mask, Print, Paper Cut, Funny Eye, Paper Glasses, and Replay.

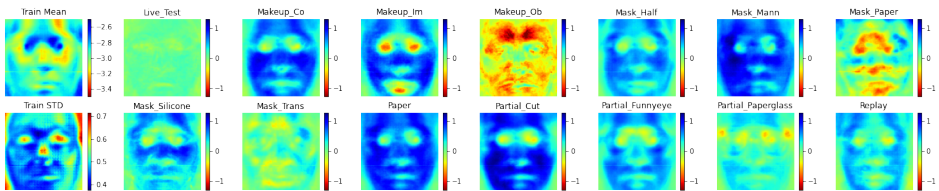


Figure 11.4: Average and standard deviation of the log-likelihood matrices over training data in the first column, along with the average log-likelihood matrices for test BF data and each individual PAS from the SiW-M dataset in the same order as in Fig. 11.3.

the training BF data average, while each attack species show a different pattern for low likelihood and high likelihood regions. Attacks with unusually high likelihood over the skin region are cosmetic makeup, impersonation makeup, half mask, mannequin, silicone mask, print, and partial cut attacks. This effect can be interpreted as the over-smoothness of skin texture in these attacks. Attacks with unusually low likelihood over the skin are obfuscation makeup, paper mask and to some extent transparent mask, which can, in turn, be interpreted as severe anomalies in the skin texture. As expected, partial attacks show anomalies in the region of the image where the attack is applied to.

Fig. 11.5 shows the t-SNE embeddings (36) of the normalized average pixel log-likelihood matrices from each video. From this figure, it is evident that the representation manages to cluster attacks from the same species together with few exceptions. Furthermore, it shows a good separability between BF data and presentation attack data, while the training BF data distribution overlaps with the test BF data. These are remarkable characteristics for the features generated by the proposed anomaly extraction which was trained in an unsupervised manner on only BF examples. This separation is however not perfect, as a cluster of BF samples are located inside the attack distribution with high overlap with partial funny eye and partial paper glass attacks. In addition, clusters of presentation attacks exist inside

the BF distribution. The majority of these samples are from transparent mask, obfuscation makeup, and partial paper glass attacks. By looking at Fig. 11.4 it can be seen that all these attacks have a shared characteristic where the average log-likelihood matrix has lower values on the skin region in contrast to other attacks where the skin region shows higher log-likelihood values.

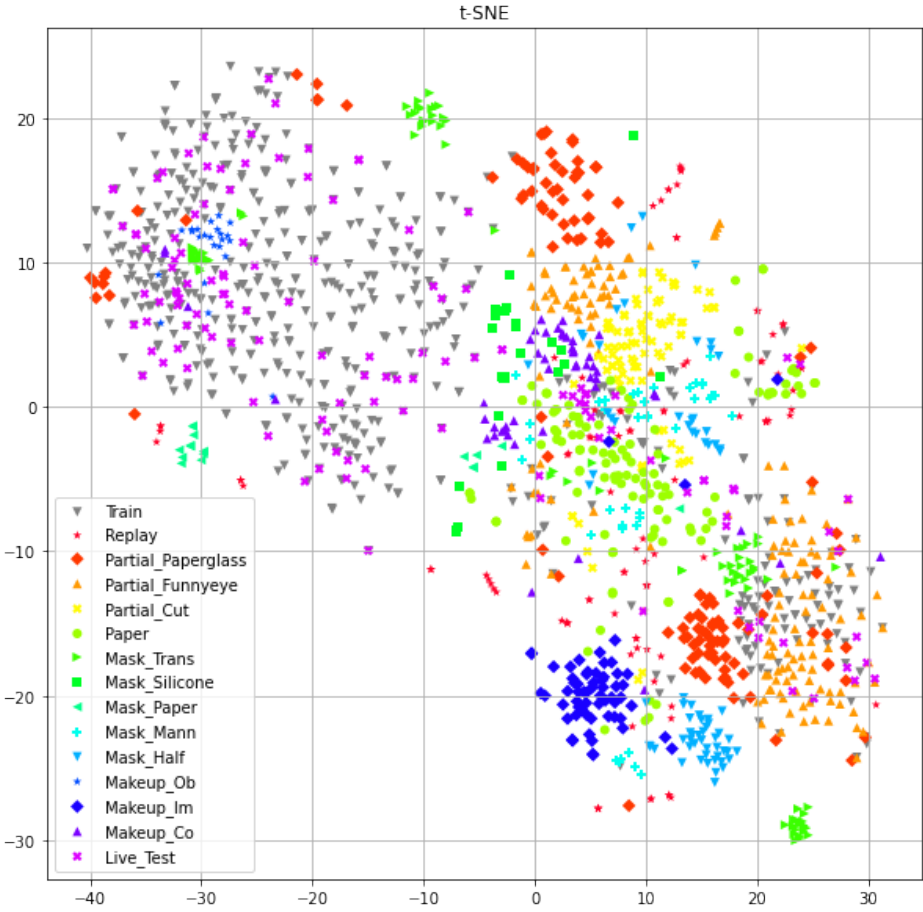


Figure 11.5: The t-SNE graph on the average log-likelihood matrices for all the data available in the SiW-M dataset. Each point represents a video, and each attack species is visualized with a different shape and color. The training BF data is shown with gray dots while the test BF data is shown with pink pluses. A clear separation is visible between BF data and attack data.

11.7.2 One-class classification

The performance of both anomaly measures in the proposed one-class classification scheme, along with the combined one-class detection score for each species is presented in Table 11.2. Even though the EER values for the detection of individual attacks, with the exception of impersonation makeup attack, are far from acceptable, these anomaly measures show a balanced performance across all attack species. For all attacks, it can be seen that the fusion of these two anomaly measures successfully reduces the EER close to the smaller value of the two, and subsequently the MPA EER is reduced by 10%. The method performs significantly better on impersonation makeup attacks compared to the other attacks while transparent mask and paper glasses attacks are the most challenging for the system.

To see the effect of the number of PCA components in the detection rate, Fig. 11.6 shows the average as well as the maximum EER over all species after filtering out the first n components from the PCA representation. It can be seen that, as hypothesized, the last PCA dimensions contain a significant amount of attack-unspecific discriminative information. The correlation between the aggregated log-likelihood measure and the anomaly norm measures is 0.15 signifying the complementing potential of these measures on each other. The combination scores reflect the complementary nature of these measures and results in a detector with an MPA attack detection EER of 27.1%. The DET curve for the resulting one-class detector is shown in Fig. 11.7 for all attack species. This plot reveals three clusters of curves corresponding to transparent mask, silicone mask, partial paper glass, paper mask, partial funny eye, obfuscation makeup, and cosmetic makeup attacks with above 20% EERs, partial paper cut, replay, half mask, print, and mannequin attacks with EERs between 10% and 20%, and finally impersonation makeup with less than 5% EER. The attacks with higher than 20% EER reflect the overlaps observed in Fig. 11.5. The attacks with an EER below 20% show similarities in their average log-likelihood images in Fig. 11.4 while the other attacks each have their individual dissimilar patterns.

11.7.3 Detection Performance

In the following, the detection performance in terms of MPA EER is presented and analyzed for the detection of known attacks, unknown attacks, and few-shot learning.

Known attacks

The performance of the proposed methods in comparison to the existing detection methods which are applied to the SiW-M dataset is reported in Table 11.3. It can be seen that even though the proposed method is outperformed on most individual

Table 11.2: Detection performance for each of the anomaly measures and their combination on the SiW-M dataset.

Method	Metric [%]	Replay		Print		Mask						Makeup			Partial			MPA
		Half	Silicone	Trans.	Paper	Mann.	Obf.	Imp.	Cosm.	FunnyEye	P.Glasses	P.Cut	MPA					
Agg. Log-likelihood	EER	22.08	22.23	19.57	23.99	37.69	29.86	19.09	24.41	16.53	25.88	24.00	37.08	22.23	37.69			
	ACER	23.01	13.93	13.38	27.19	50.57	26.43	13.11	17.23	13.88	26.48	22.95	35.33	13.51	50.57			
Anomaly	EER	13.96	17.64	16.67	29.97	23.86	23.13	20.34	37.85	1.52	22.39	33.54	27.02	18.97	37.85			
	ACER	17.70	27.56	14.14	31.73	33.90	23.40	15.38	31.62	4.79	36.33	31.29	30.03	15.03	36.33			
Combination	EER	15.23	12.47	14.46	25.84	27.08	23.89	11.55	22.99	3.15	23.12	24.18	26.66	15.13	27.08			
	ACER	13.16	10.90	11.11	23.40	30.87	18.09	9.32	17.98	4.79	19.67	16.14	23.97	11.24	30.87			

Table 11.3: Performance comparison between proposed detection method and existing methods on the task of known attack detection on the SiW-M dataset.

Method	Metric [%]	Replay		Print		Mask						Makeup			Partial			MPA
		Half	Silicone	Trans.	Paper	Mann.	Obf.	Imp.	Cosm.	FunnyEye	P.Glasses	P.Cut	MPA					
Auxiliary (47)	EER	4.7	0.0	1.6	10.5	4.6	10.0	6.4	12.7	0.0	19.6	7.2	7.5	0.0	19.6			
	ACER	5.1	5.0	5.0	10.2	5.0	9.8	6.3	19.6	5.0	26.5	5.5	5.2	5.0	26.5			
LLIG (28)	EER	3.5	3.1	0.1	9.9	1.4	0.0	4.3	6.4	2.0	15.4	0.5	1.6	1.7	15.4			
	ACER	3.5	3.1	1.9	5.7	2.1	1.9	4.2	7.2	2.5	22.5	1.9	2.2	1.9	22.5			
One-class	EER	15.7	9.6	12.4	28.7	27.7	22.5	10.3	18.2	3.9	22.9	22.6	26.2	17.6	28.7			
	ACER	13.2	10.9	11.1	23.4	30.9	18.1	9.3	18.0	4.8	19.7	16.1	24.0	11.2	30.9			
C-marmax	EER	9.7	5.6	1.5	6.6	4.5	3.0	3.8	8.0	3.1	7.8	5.7	7.7	6.5	9.7			
	ACER	10.6	5.6	4.3	6.6	8.0	8.6	6.3	13.3	4.7	9.3	6.3	9.2	6.5	13.3			
Fusion	EER	6.1	7.3	3.6	4.5	4.5	3.8	4.8	8.0	1.5	7.8	4.3	8.5	3.4	8.5			
	ACER	9.1	7.1	7.3	11.9	10.2	11.6	7.0	11.7	4.7	12.3	10.8	10.7	7.3	12.3			

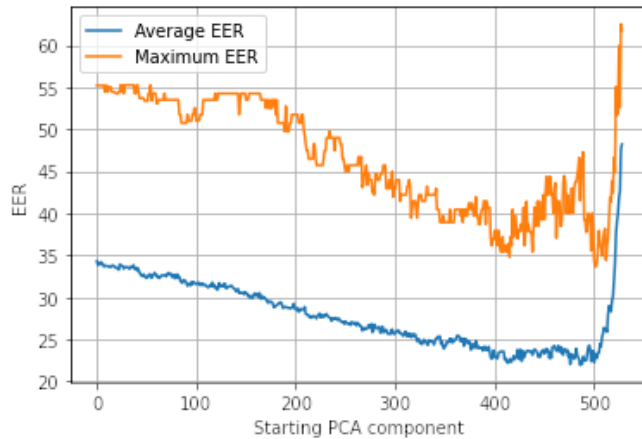


Figure 11.6: Detection performance according to the starting PCA component before calculation of the energy.

attacks, the focus of the loss function on the MPA resulted in a lower EER on the difficult attack species, namely cosmetic makeup. As a result, the proposed discriminative detector achieves 9.7% EER on the MPA, reducing the MPA EER by 37% compared to the best existing detector. The proposed fusion mechanism further reduces the MPA EER to 8.5%. The DET curve for the proposed discriminative detector is shown in Fig. 11.8. It is worth noting that the clusters visible in Fig. 11.7 are merged together and the curves follow a similar course, representing a more balanced detection performance. Furthermore, the curve for impersonation makeup is almost identical to the one-class classification curve, showing that the proposed C-marx loss successfully avoided optimization of performance on this attack which was the easiest to detect using its input. A similar pattern is observable in Table 11.3 where print and impersonation makeup attacks achieved the smallest boost in performance after the application of the discriminative classifier. Due to the small number of test samples, the DET curve shows abrupt changes, showing that more data is needed for a more precise measurement of EERs.

Unknown Attacks

The results for the proposed method along with the performance of existing detectors in unknown attack conditions are presented in Table 11.4. It can be seen that, as expected, the one-class detector performs better than all discriminative detection methods in terms of MPA EER, including the proposed method. However, it is worth mentioning that the discriminative detectors gain an advantage over certain PASs where there is a similarity of the discriminative features between the

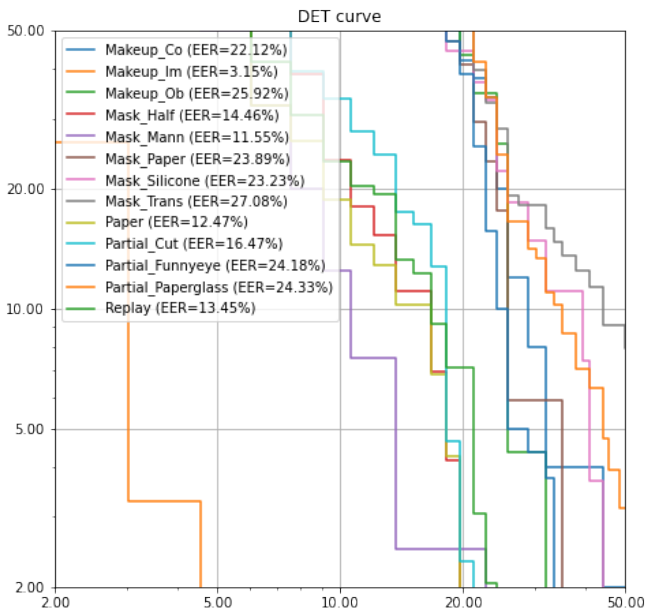


Figure 11.7: Detection error trade-off curve for the one-class detector in PAD on the SiW-M dataset.

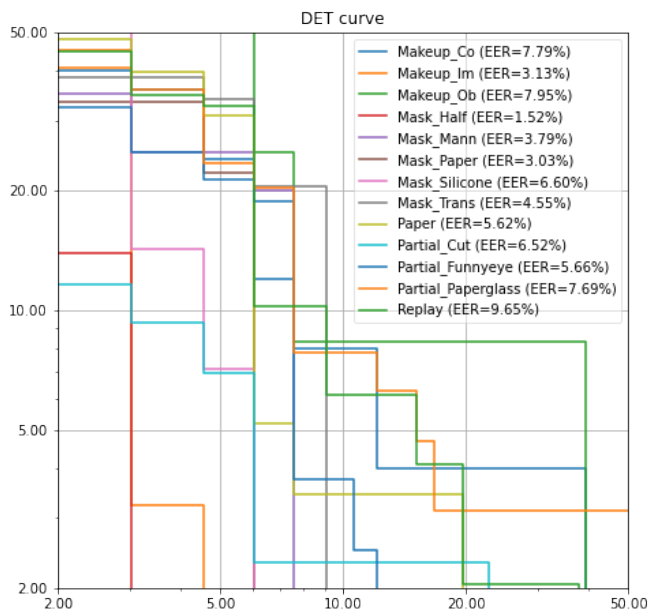


Figure 11.8: Detection error trade-off curve of the discriminative detector for the known attack detection on PAD task on the SiW-M dataset.

unknown PAS and the known ones used in training. This distinction is visible in cases where a significant difference exists between the one-class classifier performance compared to the discriminative classifier such as in the case of silicone mask and mannequin attacks. A close observation of Fig. 11.5 reveals that samples from these two attacks are not clustered together in the anomaly feature space. The proposed fusion method managed to cap the EER for partial paper glasses and partial funny eye attack where the proposed discriminative detection method fails while not hindering the performance on cases where the discriminative detection method performs well. Considering the existing solutions, it can be seen that there only exists one approach that has a better than chance detection rate for MPA, namely LLIG (28). This is concerning as it shows that all other existing methods would be ineffective against a rational attacker, and in the case of (47), would actually increase the efficacy of attacks. The proposed method achieves an MPA EER of 27.8% and outperforms all baseline methods.

OULU-NPU dataset

In Table 11.6, we report results of proposed framework as well as previously presented strategies in the literature on OULU-NPU dataset (14). Fig. 11.10 shows the average and standard deviation of log-likelihood matrices for training data along with the average matrices for test data. The OULU-NPU face presentation attack detection database is composed of 4950 real access and attack videos. The videos were captured utilizing the front cameras of six mobile devices in three sessions with different background scenes and illumination conditions. There are two attacks, i.e., print and video-replay, which were generated via two printers and two display devices. In this study, we adopted OULU-NPU Protocol II because it presents the unknown attack detection scenario, namely, the effect of attack variation is assessed by introducing previously unseen print and video-replay attacks in the test set. We can observe in Table 11.6 that the performance obtained using proposed framework is better than prior methods. For instance, the presented method with c-marmax achieved 2.4% EER, whereas the scheme proposed in (33) obtained 6.0% EER. It is also worth noticing that proposed system with one-class classifier did not perform well, but the proposed system with fusion scheme could avoid a major loss and attain notable accuracy.

Few-shot learning

In Table 11.5, the performance of the proposed method is presented on the task of few-shot learning when having one or five examples, and compared to unknown and known cases. It can be seen that by observation of even one example from an unknown PAS, the performance of the system improves, and the MPA EER is reduced by 45% from 33.5% to 18.3% by observation of five examples. As such,

Table 11.4: Performance comparison of the proposed methods and the existing methods in the literature on the task of unknown presentation attack detection on the SiW-M dataset.

Method	Metric [%]	Replay	Print	Mask				Makeup			Partial			MPA	
				Half	Silicone	Trans.	Paper	Mann.	Obf.	Imp.	Cosm.	Funny	Eye		P.
SVM+LBP (14)	EER	20.8	18.6	36.3	21.4	37.2	7.5	14.1	51.2	19.8	16.1	34.4	33.0	7.9	51.2
	ACER	20.6	18.4	31.3	21.4	45.5	11.6	13.8	59.3	23.9	16.7	35.9	39.2	11.7	59.3
Auxiliary (47)	EER	14.0	4.3	11.6	12.4	24.6	7.8	10.0	72.3	10.1	9.4	21.4	18.6	4.0	72.3
	ACER	16.8	6.9	19.3	14.9	52.1	8.0	12.8	55.8	13.7	11.7	49.0	40.5	5.3	55.8
DTN (48)	EER	10.0	2.1	14.4	18.6	26.5	5.7	9.6	50.2	10.1	13.2	19.8	20.5	8.8	50.2
	ACER	9.8	6.0	15.0	18.7	36.0	4.5	7.7	48.1	11.4	14.2	19.3	19.8	8.5	48.1
CDC (83)	EER	9.2	5.6	4.2	11.1	19.3	5.9	5.0	43.5	0.0	14.0	23.3	14.3	0.0	43.5
	ACER	10.8	7.3	9.1	10.3	18.8	3.5	5.6	42.1	0.8	14.0	24.0	17.6	1.9	42.1
LLIG (28)	EER	6.8	11.2	2.8	6.3	28.5	0.4	3.3	17.8	3.9	11.7	21.6	13.5	3.6	28.5
	ACER	7.4	19.5	3.2	7.7	33.3	5.2	3.3	22.5	5.9	11.7	21.7	14.1	6.4	33.3
One-class	EER	15.2	12.5	14.5	25.8	27.1	23.9	11.6	23.0	3.2	23.1	24.2	26.7	15.1	27.1
	ACER	13.2	10.9	11.1	23.4	30.9	18.1	9.3	18.0	4.8	19.7	16.1	24.0	11.2	30.9
C-marmax	EER	12.2	10.5	27.5	8.2	22.7	13.4	4.3	17.0	0.8	11.3	33.5	33.2	11.1	33.5
	ACER	13.2	8.6	21.0	11.3	16.5	6.7	6.3	11.2	4.0	14.4	41.1	44.4	9.7	44.4
Fusion	EER	10.4	6.5	20.3	10.9	24.6	3.0	3.5	23.0	1.5	12.8	25.1	27.8	7.1	27.8
	ACER	10.9	7.9	13.4	15.8	22.5	6.7	4.8	18.7	4.8	18.9	19.2	25.5	5.9	25.5

the proposed system shows the capacity of significantly reducing the EER after the presentation of a few examples of a new PAS. It can also be seen that, specifically in the case of impersonation makeup attack, the observation of new samples does not reduce the EER. This can be explained by the fact that the EER is already low in the zero-shot case, and the proposed C-marmax loss does not reward further improvement in the EER of this attack as it does not improve the overall MPA EER.

11.7.4 Detection Cost

Due to the big size of the PixelCNN++ model, the extraction of each individual pixel log-likelihood matrix for each frame is the bottleneck and takes roughly 75 milliseconds in our setup. Considering the average length of six seconds for the 24 FPS videos in the dataset, processing each video takes 9 seconds, corresponding to $\times 1.5$ real-time speed. This may account for a prohibitively high detection cost in certain applications such as smartphone-based detection or social media monitoring. However, according to Eq. 11.2 the proposed method can find applications where the cost of a missed detection is high, such as border control and authenticity verification in journalism.

11.8 Deepfake Detection

Fig. 11.9 shows the average and standard deviation of log-likelihood matrices for training data along with the average matrices for test data. It can be seen that most variations in the data are from the background, forehead, and cheeks, while the eye and mouth regions had little variability with a low log-likelihood average. The BF test data average matches that of training BF data. However, there are distinct patterns corresponding to each attack species. In the case of Deepfakes and NeuralTextures, there is a high log-likelihood region on the lower half of the face, corresponding to the possible over-smoothness of the texture. In the case of Deepfakes, there is a low-likelihood region around the eyebrows and the chin line which corresponds to the locations where the artifacts that are the characteristic of Deepfakes often occur. For the Face2Face technique, the pattern corresponds to points with low log-likelihood around the nose and chin line, while for the FaceSwap technique, the pattern corresponds to the eyes, nose, and mouth regions.

Table 11.7 shows the performance of the one-class detector and the proposed discriminative detector as well as their fusion. It can be seen that the one-class detector managed to achieve acceptable MPA EER of 8.21% while the discriminative detector achieved near-perfect video level detection. The Fusion did not degrade the performance of the discriminative detector significantly. It is important to men-

Table 11.5: Performance of the detector in few-shot learning scenarios on the SiW-M dataset.

Method	Metric [%]	Replay	Print	Mask				Makeup			Partial			MPA	
				Half	Silicone	Trans.	Paper	Mann.	Obf.	Imp.	Cosm.	FunnyEye	P.Glasses		P.Cut
Zero shot	EER	12.2	10.5	27.5	8.2	22.7	13.4	4.3	17.0	0.8	11.3	33.5	33.2	11.1	33.5
	ACER	13.2	8.6	21.0	11.3	16.5	6.7	6.3	11.2	4.0	14.4	41.1	44.4	9.7	44.4
One shot	EER	14.8	10.5	22.6	7.6	18.3	6.2	4.3	15.9	0.8	14.7	30.2	26.4	13.3	30.2
	ACER	15.5	11.7	24.8	12.9	18.0	6.2	4.8	18.2	2.5	19.0	42.7	46.0	10.5	46.0
Five shot	EER	16.4	10.0	6.0	9.8	15.4	3.0	11.0	15.9	0.9	18.3	17.0	15.0	6.1	18.3
	ACER	16.3	8.8	9.8	9.1	12.1	7.2	8.9	11.1	2.7	12.4	19.2	23.3	14.5	23.3
Known	EER	9.7	5.6	1.5	6.6	4.5	3.0	3.8	8.0	3.1	7.8	5.7	7.7	6.5	9.7
	ACER	10.6	5.6	4.3	6.6	8.0	8.6	6.3	13.3	4.7	9.3	6.3	9.2	6.5	13.3

Table 11.6: Performance of the proposed detection methods for the protocol II task of OULU-NPU dataset.

Metric [%]	Gradient (12)	Auxiliary (47)	DeepPixBiS (33)	TSCNN-ResNet (16)	LLIG (28)	One-class	C-marmax	Fusion
EER	0.9	-	-	2.0	-	26.6	2.4	3.3
ACER	2.5	2.7	6.0	4.9	3.4	52.4	3.1	3.1

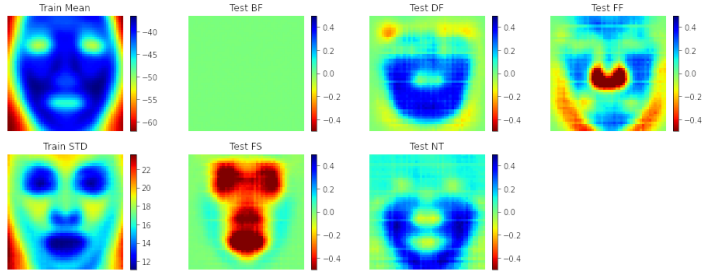


Figure 11.9: Average and standard deviation of the log-likelihood matrices over training data in the first column, along with the average log-likelihood matrices for test BF and each individual attack species in the FaceForencisc++ dataset in the following order: Deepfakes, Face2Face, FaceSwap, NeuralTextures.

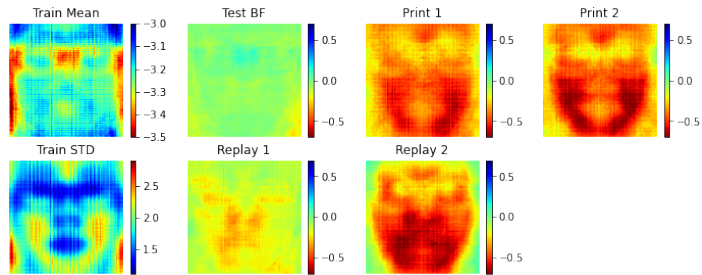


Figure 11.10: Average and standard deviation of the log-likelihood matrices over training data in the first column, along with the average log-likelihood matrices for test BF and each individual attack species in the OULU-NPU dataset in the following order: Printer 1, Printer 2, Replay 1, Replay 2.

Table 11.7: Performance of the proposed detection methods for the task of known attack detection on Deepfake detection task on the FaceForensics++ dataset.

Method	Metric [%]	DeepFake	Face2Face	FaceSwap	NTexture	MPA
One-class	EER	6.43	8.21	2.14	2.14	8.21
	ACER	5.00	8.21	3.21	3.21	8.21
C-marmax	EER	0.00	0.71	0.00	0.36	0.71
	ACER	2.50	2.50	2.50	2.50	2.50
Fusion	EER	0.71	0.36	0.00	0.71	0.71
	ACER	2.50	2.50	2.50	2.50	2.50

tion that known attack detection on the raw subset of the dataset is a solved problem with near-perfect frame-level detection rates reported in the baseline (64).

Table 11.8 reports the detection performance on the LOO unknown attack detection scenario. The low EERs of the discriminative detector shows that there are mutually discriminative features across the known and unknown attacks, especially for Face2Face and NeuralTexture methods. However, the face swap method shows less similarity to other methods and this results in a increase in the EER of the discriminative detector compared to the one-class one. Furthermore, the fusion mechanism managed to lower the MPA EER significantly, and an MPA EER of 2.5% is achieved for the unknown attack detection. Due to the easiness of spotting digital manipulation traces in raw videos, the overall performances in terms of MPA are much lower than for PAD experiments.

Table 11.8: Performance of the proposed detection methods for the task of unknown attack detection on Deepfake detection task on the FaceForensics++ dataset.

Method	Metric [%]	DeepFake	Face2Face	FaceSwap	NTexture	MPA
One-class	EER	6.43	8.21	2.14	2.14	8.21
	ACER	5.00	8.21	3.21	3.21	8.21
C-marmax	EER	5.36	1.07	5.71	1.79	5.71
	ACER	5.00	2.86	5.71	2.86	5.71
Fusion	EER	2.50	1.43	2.50	1.43	2.50
	ACER	3.57	2.50	4.29	2.50	4.29

11.9 Conclusion

The choice of the attack by a rational attacker can have a significant negative impact on the performance of the detection systems in real-life scenarios. In response, after relying on game theory to build a theoretic basis and formulating the interactions between the attacker and the defender, a new detection method is proposed to optimize the performance against attacks from such attackers. Experiments on the tasks of presentation attack detection and Deepfake detection show effectiveness of proposed method in improving detection rate on most powerful attacks both in known attack cases and when the detector faces unknown attacks. Furthermore, the proposed feature set is capable of enabling few-shot learning and explainability at pixel-level. The proposed method shows generalizability across widely different types of attacks ranging from Deepfakes and replay attacks to 3D masks and makeup attacks and is able to show where the artifacts commonly occur for each specific attack species. Also, unsupervised anomaly detection method used is able to produce representations that cluster attacks from the same species together and separate BF samples from attacks in an unsupervised manner with limited training

data in unconstrained recording conditions.

However, this method has two specific short-comings. First, the extraction of the anomaly representations is computationally expensive and thus the system cannot be deployed in applications where processing an input video should be done faster than in real-time such as automated content monitoring on social media. Secondly, despite the proposed method outperforming the state-of-the-art in the task of presentation attack detection, its expected 27.8% performance against the most powerful unknown attack is still far from acceptable for real-life applications, showing the need for further research in this direction. However, the availability of more training data from a more diverse set of attacks may alleviate this limitation.

11.10 Acknowledgment

This research work was funded by the Department of Information Security and Communication Technology at the Norwegian University of Science and Technology.

References

- [1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.
- [2] Z. Akhtar, M. R. Mouree, and D. Dasgupta. Utility of deep learning features for facial attributes manipulation detection. In *IEEE Int'l Conf. on Humanized Computing and Communication with Artificial Intelligence*, pages 55–60, 2020.
- [3] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.
- [4] S. R. Arashloo, J. Kittler, and W. Christmas. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE Access*, 5:13868–13882, 2017.
- [5] M. Barni, Z. Chen, and B. Tondi. Adversary-aware, data-driven detection of double jpeg compression: How to make counter-forensics harder. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2016.
- [6] M. Barni, E. Nowroozi, and B. Tondi. Higher-order, adversary-aware, double jpeg-detection via selected training on attacked samples. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 281–285, 2017.
- [7] M. Barni, M. C. Stamm, and B. Tondi. Adversarial multimedia forensics: Overview and challenges ahead. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 962–966, 2018.
- [8] M. Barni and B. Tondi. The source identification game: An information-theoretic perspective. *IEEE Transactions on Information Forensics and Security*, 8(3):450–463, 2013.
- [9] S. Bhattacharjee, A. Mohammadi, A. Anjos, and S. Marcel. *Recent Advances in Face Presentation Attack Detection*, pages 207–228. Springer, 2019.
- [10] B. Biggio, I. Corona, Z.-M. He, P. P. K. Chan, G. Giacinto, D. S. Yeung, and F. Roli. One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time. In F. Schwenker, F. Roli, and J. Kittler, editors, *Multiple Classifier Systems*, pages 168–180, Cham, 2015. Springer International Publishing.

-
- [11] R. Böhme and M. Kirchner. *Counter-Forensics: Attacking Image Forensics*, pages 327–366. Springer New York, New York, NY, 2013.
- [12] Z. Boulkenafet and et al. A competition on generalized software-based face presentation attack detection in mobile scenarios. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 688–696, 2017.
- [13] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 2017.
- [14] 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 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 612–618, 2017.
- [15] C. Chen, Y. Q. Shi, and W. Su. A machine learning based scheme for double jpeg compression detection. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, 2008.
- [16] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li. Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Transactions on Information Forensics and Security*, 15:578–593, 2020.
- [17] Z. Chen, B. Tondi, X. Li, R. Ni, Y. Zhao, and M. Barni. Secure detection of image manipulation by means of random feature selection. *IEEE Transactions on Information Forensics and Security*, 14(9):2454–2469, 2019.
- [18] G. Chetty. Biometric liveness checking using multimodal fuzzy fusion. In *International Conference on Fuzzy Systems*, pages 1–8, 2010.
- [19] U. A. Ciftci, I. Demir, and L. Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.
- [20] V. Conotter, E. Bodnari, G. Boato, and H. Farid. Physiologically-based detection of computer generated faces in video. In *ICIP*, pages 248–252, 2014.
- [21] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *CoRR*, abs/1812.02510, 2018.
- [22] D. Cozzolino Giovanni Poggi Luisa Verdoliva. Extracting camera-based fingerprints for video forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

- [23] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [24] D. Dang-Nguyen, G. Boato, and F. G. B. De Natale. Discrimination between computer generated and natural human faces based on asymmetry information. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 1234–1238, 2012.
- [25] D. Dang-Nguyen, G. Boato, and F. G. B. De Natale. 3d-model-based video analysis for computer generated faces identification. *IEEE Transactions on Information Forensics and Security*, 10(8):1752–1763, 2015.
- [26] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. Lbp - top based countermeasure against face spoofing attacks. In J.-I. Park and J. Kim, editors, *Computer Vision - ACCV 2012 Workshops*, pages 121–132, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
- [27] A. De Rosa, M. Fontani, M. Massai, A. Piva, and M. Barni. Second-order statistics analysis to cope with contrast enhancement counter-forensics. *IEEE Signal Processing Letters*, 22(8):1132–1136, 2015.
- [28] D. Deb and A. K. Jain. Look locally infer globally: A generalizable face anti-spoofing approach, 2020.
- [29] L. Ditria, B. J. Meyer, and T. Drummond. Opegan: Open set generative adversarial networks. In *Asian Conference on Computer Vision*, 2020.
- [30] M. Du, S. Pentyala, Y. Li, and X. Hu. Towards generalizable deepfake detection with locality-aware autoencoder, 2020.
- [31] T. Fernando, C. Fookes, S. Denman, and S. Sridharan. Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks, 2019.
- [32] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Transactions on Image Processing*, 23(2):710–724, 2014.
- [33] A. George and S. Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

-
- [34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. G. et al., editor, *Adv. in Neural Info. Proc. Sys.*, volume 27, pages 2672–2680, 2014.
- [35] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *IEEE AVSS*, pages 1–6, 2018.
- [36] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS’02, page 857–864, Cambridge, MA, USA, 2002. MIT Press.
- [37] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch. Fake face detection methods: Can they be generalized? In *2018 International Conference of the Biometrics Special Interest Group*, pages 1–6, 2018.
- [38] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.
- [39] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [40] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun. Real-time face detection and motion analysis with application in “liveness” assessment. *IEEE Transactions on Information Forensics and Security*, 2(3):548–558, 2007.
- [41] H. Li, B. Li, S. Tan, and J. Huang. Detection of deep network generated images using disparities in color components. *CoRR*, abs/1808.07276, 2018.
- [42] H. Li, S. Wang, and A. C. Kot. Face spoofing detection with image quality regression. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2016.
- [43] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [44] L. Li and et al. Compactnet: learning a compact space for face presentation attack detection. *Neurocomputing*, 409:191–207, 2020.
- [45] Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.

- [46] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [47] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [48] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [49] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. Do gans leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511, 2019.
- [50] F. Marra, C. Saltori, G. Boato, and L. Verdoliva. Incremental learning for the detection and classification of gan-generated images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2019.
- [51] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, 2019.
- [52] S. McCloskey and M. Albright. Detecting gan-generated imagery using saturation cues. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4584–4588, 2019.
- [53] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li. Open set learning with counterfactual images. In *ECCV*, pages 613–628, 2018.
- [54] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, 2019.
- [55] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel. On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In *2018 International Conference on Biometrics (ICB)*, pages 75–81, 2018.
- [56] G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcam. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

-
- [57] X. Pan, X. Zhang, and S. Lyu. Exposing image forgery with blind noise estimation. *MM&Sec '11*, page 15–20, New York, NY, USA, 2011. Association for Computing Machinery.
- [58] K. Patel, H. Han, and A. K. Jain. Cross-database face antispoofing with robust feature representation. In Z. You, J. Zhou, Y. Wang, Z. Sun, S. Shan, W. Zheng, J. Feng, and Q. Zhao, editors, *Biometric Recognition*, pages 611–619, Cham, 2016. Springer International Publishing.
- [59] K. Patel, H. Han, and A. K. Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283, 2016.
- [60] P. Perera, V. I. Morariu, R. Jain, V. Manjunatha, C. Wigington, V. Ordonez, and V. M. Patel. Generative-discriminative feature representations for open-set recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11814–11823, 2020.
- [61] D. Perez-Cabo, D. Jimenez-Cabello, A. Costa-Pazo, and R. J. Lopez-Sastre. Deep anomaly detection for generalized face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [62] I. A. L. R., L. T. Menon, M. C. O. P. N., A. L. Koerich, and A. S. B. J. au2. Style transfer applied to face liveness detection with user-centered models, 2019.
- [63] R. Ramachandra and C. Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Comput. Surv.*, 50(1):8:1–8:37, Mar. 2017.
- [64] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [65] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *IEEE/CVF CVPR Workshops*, June 2019.
- [66] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *CoRR*, abs/1701.05517, 2017.

- [67] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, June 2015.
- [68] W. Shang, K. Sohn, D. Almeida, and H. Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16*, page 2217–2225. JMLR.org, 2016.
- [69] M. C. Stamm, W. S. Lin, and K. J. R. Liu. Forensics vs. anti-forensics: A decision and game theoretic framework. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1749–1752, 2012.
- [70] M. C. Stamm, M. Wu, and K. J. R. Liu. Information forensics: An overview of the first decade. *IEEE Access*, 1:167–200, 2013.
- [71] X. Tan, Y. Li, J. Liu, and L. Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, pages 504–517, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.
- [72] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), July 2019.
- [73] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.
- [74] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020.
- [75] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *CoRR*, abs/1601.06759, 2016.
- [76] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. Cnn-generated images are surprisingly easy to spot... for now, 2020.
- [77] Y. Wang, F. Nian, T. Li, Z. Meng, and K. Wang. Robust face anti-spoofing with depth information. *Journal of Visual Communication and Image Representation*, 49:332–337, 2017.
- [78] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE TIFS*, 10(4):746–761, 2015.

-
- [79] X. Xuan, B. Peng, W. Wang, and J. Dong. On the generalization of gan image forensics. In Z. Sun, R. He, J. Feng, S. Shan, and Z. Guo, editors, *Biometric Recognition*, pages 134–141, Cham, 2019. Springer International Publishing.
- [80] J. Yang, Z. Lei, S. Liao, and S. Z. Li. Face liveness detection with component dependent descriptor. In *Int’l Conf. on Biometrics (ICB)*, pages 1–6, 2013.
- [81] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE ICASSP*, pages 8261–8265, 2019.
- [82] X. Yang, Y. Li, H. Qi, and S. Lyu. Exposing gan-synthesized faces using landmark locations. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec’19*, page 113–118, New York, NY, USA, 2019. Association for Computing Machinery.
- [83] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [84] F. Zhang, P. P. K. Chan, B. Biggio, D. S. Yeung, and F. Roli. Adversarial feature selection against evasion attacks. *IEEE Transactions on Cybernetics*, 46(3):766–777, 2016.
- [85] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li. Face liveness detection by learning multispectral reflectance distributions. In *Face and Gesture 2011*, pages 436–441, 2011.

ISBN 978-82-326-6166-4 (printed ver.)
ISBN 978-82-326-6101-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)



NTNU

Norwegian University of
Science and Technology