KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Security and Assessment of Biometric Authentication: Attacks, Defenses, and Metrics

Tim Van hamme

Supervisors:
Prof. dr. ir. W. Joosen
Dr. ir. D. Preuveneers

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

November 2023

Security and Assessment of Biometric Authentication: Attacks, Defenses, and Metrics

Tim VAN HAMME

Examination committee:
Prof. dr. ir. O. Van der Biest, chair
Prof. dr. ir. W. Joosen, supervisor
Dr. ir. D. Preuveneers, supervisor
Prof. dr. T. Van Cutsem
Prof. dr. J. Davis
Prof. dr. E. Kindt
(KU Leuven & Universiteit Leiden)
Prof. dr. B. Crispo
(University of Trento)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

November 2023

© 2023 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Tim Van hamme, Celestijnenlaan 200A box 2402, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface



For a bystander the PhD journey is often characterized by milestones called 'academic publications' or, in PhD slang, 'papers.' Nevertheless, the PhD trajectory is much richer than its academic output. It serves as an opportunity for personal and professional growth.

My PhD journey is one that felt scary and exciting as we set course for the unknown, with the certainty that we would encounter the boundaries of (our) knowledge. Along this path, I oscillated between profound satisfaction and gratitude and the persistent feeling of being an imposter. Fortunately, I was not alone on this journey, surrounded by remarkable individuals to whom I extend my deepest gratitude.

Being a part of a lab that recognizes the importance of doing the right things with the right people at the right time has been invaluable. Wouter, we owe the vibrant culture and conducive work environment to you. Thinking of you, Wouter, reminds me of the profound quote: "Wisdom cannot be imparted. Wisdom that a wise man attempts to impart always sounds like foolishness to someone else... Knowledge can be communicated, but not wisdom. One can find it, live it, do wonders through it, but one cannot communicate and teach it."

Davy and Enrique, you imparted knowledge about science with unparalleled dedication. Davy, your commitment to your work and team has been a guiding force in my growth. I'm glad to be your colleague. Enrique, your passion for science and technology is inspiring. I've enjoyed the many conversations and learned a lot from you and your keen perspective on life.

The many colleagues and collaborators who left their mark on our work, me as person, and the great culture within the lab. Vera, I truly appreciate your encouraging words and your ability to communicate complex topics. I believe you are a true idealist, both in your work and in your personal life. Make sure to be kind to yourself and take care of that crazy, hard-working, loveable perfectionist that is your partner. Giuseppe, I envy your natural flair. You've become a real man. The world is at your feet, do not let dark thoughts cloud your mind. Sander, thank you for your belief in me; now, believe

in yourself and embrace your remarkable growth. Willem, your optimistic nature is a powerful asset, but ensure your attention is not too divided. Abel, our conversations often present new perspectives; never underestimate your worth. Ilias, your eagerness for knowledge and debate is inspiring, even if occasionally challenging. Victor, your admirable work ethic is a source of inspiration, and witnessing you embrace life is truly fulfilling. Annick, it has been delightful getting to know you. Your unwavering dependability and wealth of experience have garnered my admiration. It's individuals like you who play a pivotal role in sustaining the business, especially during challenging times when everything appears to be unraveling.

To my esteemed jury, I extend my heartfelt appreciation for taking the time to read and engage in insightful discussions about this text. Your valuable feedback and contributions have enriched this work beyond measure. Thank you for your time, expertise, and thoughtful consideration.

Outside work, I'm surrounded by exceptional individuals. My family's unwavering love and respect, coupled with the support of dependable friends, have been my pillars. I am genuinely grateful for my environment. Thank you all!

Abstract

In today's digital age, where our personal identity is increasingly defined and expressed through online platforms, the need for robust authentication methods has never been more critical. As we navigate the interconnected realms of social media, online banking, and email accounts, we stand vulnerable to the rising tide of cyberattacks and data breaches. Our digital identities are, in many ways, the keys to our lives, and safeguarding them is a paramount concern.

The advent of machine learning has ushered in a new era of authentication, promising heightened security and reduced user friction by harnessing the intricacies of appearance, and even user behavior. Yet, as we embrace these innovations, lingering concerns over the reliability of biometrics loom large. To address these concerns this dissertation delves into three core themes: Behavioral Authentication, Novel Attacks, and Security Evaluation Frameworks.

Behavioral Authentication: We focus on reliably extracting identifying patterns from user behavior. Our research in gait recognition explores various learning approaches in challenging open-world scenarios, achieving an equal error rate of approximately 1%. Additionally, we highlight the importance of biometric template protection to secure sensitive biometric data, which we apply to gait authentication. This work is then followed by a comprehensive security evaluation of protected gait authentication systems.

Novel Attacks: In this theme, we delve into innovative attacks on biometric authentication systems. We introduce and assess the first masterkey attack against free-text keystroke dynamics. The concept of a biometric masterkey, which matches a surprisingly large portion of the population, opens new doors for adversaries. Our evaluation shifts the perspective from pondering *'how easy is it to create a fake biometric'* to *'how easy a biometric can be guessed.'* For instance, a skilled adversary can guess typing behavior approximately two times more easily than a 4-digit PIN.

Security Evaluation Frameworks: Our third central theme focuses on proper security evaluation for biometric systems. We propose an evaluation framework to assess

the security of fuzzy commitment-protected biometrics and apply it to our work on gait authentication systems. Our findings reveal that longer effective key lengths are not always better, as adversaries can exploit knowledge of the biometric distribution. Furthermore, we introduce a new evaluation framework to assess biometric security, drawing insights from our attack work. This framework is applied to keystroke and face biometrics, demonstrating that face biometrics are more secure than a 4-digit PIN. We also explore the security implications of adversaries having partial knowledge of the victim's soft biometrics, such as age, sex, and typing speed.

In summary, our contributions pave the way toward more reliable biometric authentication by refining recognition methods, deepening our understanding of the threat landscape, and subjecting these systems to rigorous security assessments.

Beknopte samenvatting

In het digitale tijdperk van vandaag, waar onze persoonlijke identiteit steeds meer wordt gedefinieerd en uitgedrukt via online platforms, is de behoefte aan robuuste authenticatiemethoden nog nooit zo kritiek geweest. Terwijl we navigeren door de onderling verbonden domeinen van sociale media, online bankieren en e-mailaccounts, staan we kwetsbaar voor de toenemende golf van cyberaanvallen en datalekken. Onze digitale identiteiten zijn op vele manieren de sleutels tot ons leven, en het beschermen ervan is een belangrijke zorg.

De opkomst van machine learning heeft een nieuw tijdperk van authenticatie ingeluid, met beloftes van verhoogde beveiliging en verminderde gebruikersfrictie door de uiterlijk en zelfs gebruikersgedrag te benutten. Toch blijven er, terwijl we deze innovaties omarmen, zorgen bestaan over de betrouwbaarheid van biometrie.

Om deze zorgen aan te pakken, richt deze dissertatie zich op drie kernthema's: Gedragsauthenticatie, Nieuwe Aanvallen en Beveiligingsevaluatiekaders.

Gedragsauthenticatie: We richten ons op het betrouwbaar extraheren van identificerende patronen uit het gebruikersgedrag. Ons onderzoek naar tredherkenning verkent verschillende leermethoden in uitdagende open-wereldscenario's en bereikt een foutmarge van ongeveer 1%. Daarnaast benadrukken we het belang van biometrischesjabloon bescherming om gevoelige biometrische gegevens te beveiligen, wat we toepassen op tredauthenticatie. Dit werk wordt gevolgd door een uitgebreide beveiligingsevaluatie.

Nieuwe Aanvallen: In dit thema duiken we in innovatieve aanvallen op biometrische authenticatiesystemen. We introduceren en beoordelen de eerste meester-sleutelaanval tegen vrije-tekst toetsaanslagdynamiek. Het concept van een biometrische meester-sleutel, die een verrassend groot deel van de bevolking matcht, opent nieuwe deuren voor aanvallers. Onze evaluatie verschuift het perspectief van *'hoe eenvoudig is het om een nep-biometrie te maken'* naar *'hoe eenvoudig kan een biometrie worden geraden.'* Bijvoorbeeld, een hacker met kennis van biometrische authenticatie kan het typegedrag ongeveer twee keer gemakkelijker raden dan een 4-cijferige pincode.

Beveiligingsevaluatiekaders: Ons derde centrale thema richt zich op een juiste beveiligingsevaluatie voor biometrische systemen. We stellen een evaluatiekader voor om de beveiliging van fuzzy commitment-beschermde biometrieën te beoordelen en passen het toe op ons werk over tredauthenticatie. Onze bevindingen onthullen dat langere effectieve sleutellengtes niet altijd beter zijn, omdat tegenstanders kennis van de biometrische distributie kunnen uitbuiten. Bovendien introduceren we een nieuw evaluatiekader om biometrische beveiliging te beoordelen, geïnspireerd door ons onderzoek naar aanvallen. Dit kader wordt toegepast op toetsaanslag- en gezichtsbiometrieën, waaruit blijkt dat gezichtsbiometrieën veiliger zijn dan een 4-cijferige pincode. We verkennen ook de beveiligingsimplicaties van hackers die gedeeltelijke kennis hebben van de eigenschappen van het slachtoffer, zoals leeftijd, geslacht en typsnelheid.

Samenvattend banen onze bijdragen de weg naar meer betrouwbare biometrische authenticatie door herkenningstechnieken te verfijnen, ons begrip van het dreigingslandschap te verdiepen en deze systemen te onderwerpen aan rigoureuze beveiligingsevaluaties.

Contents

Abstract iii

Beknopte samenvatting v

Contents vii

List of figures xiii

List of tables xvii

List of acronyms xix

Overview of datasets xxi

1 Introduction 1
- 1.1 Challenges of biometric authentication . . . 2
- 1.2 Pillars of the thesis . . . 5
  - 1.2.1 Behavioral authentication . . . 6
  - 1.2.2 Novel attacks . . . 8
  - 1.2.3 Security evaluation . . . 9
- 1.3 Other contributions . . . 11
- 1.4 Dissertation outline . . . 16

2 Background 19
- 2.1 Introduction . . . 19
- 2.2 Biometric system . . . 20
  - 2.2.1 System design . . . 21
  - 2.2.2 The role of ML in biometric authentication . . . 23
  - 2.2.3 Attack surface . . . 23
  - 2.2.4 Evaluation metrics . . . 26
- 2.3 Sensing: System input . . . 27

2.3.1 Sensors . . . 27
2.3.2 Pre-processing . . . 28
2.3.3 Attacks and defenses . . . 29
2.4 Memory: Biometric database . . . 30
2.4.1 Template enrollment . . . 30
2.4.2 Template comparison . . . 31
2.4.3 Attacks and defenses . . . 31
2.5 The brain: extracting and comparing the identifying features . . . 33
2.5.1 Feature extraction . . . 33
2.5.2 Distance functions . . . 34
2.5.3 Learned functions . . . 35
2.5.4 Attacks and defenses . . . 35
2.6 Conclusion . . . 37
2.6.1 Biometric authentication as an open set problem . . . 37
2.6.2 Threats linked to new factors and deep learning . . . 38
2.6.3 Future directions . . . 38

3 Gait template protection using HMM-UBM 41
3.1 Introduction . . . 42
3.2 Related work . . . 43
3.3 Feature extraction . . . 44
3.4 HMM-UBM gait recognition . . . 46
3.4.1 Unprotected system . . . 46
3.4.2 Protected system . . . 48
3.5 Dataset and experimental protocol . . . 50
3.6 Experiments . . . 52
3.7 Discussion and Conclusions . . . 55

4 On the security of biometrics and fuzzy commitment cryptosystems: a study on gait authentication 57
4.1 Introduction . . . 58
4.2 Related work . . . 60
4.2.1 On gait authentication . . . 60
4.2.2 On template protection . . . 62
4.2.3 On fuzzy extraction . . . 62
4.3 Gait authentication systems . . . 63
4.3.1 Remote biometric authentication systems architectures . . . 63
4.3.2 Stream representation . . . 66
4.3.3 Discriminative approach: Neural networks . . . 66
4.3.4 Generative approach: HMM-UBM . . . 67
4.3.5 Intramodal fusion . . . 68
4.4 Security analysis . . . 68
4.4.1 Threat models . . . 68

4.4.2 Attacks against unprotected system . . . 70
4.4.3 Attacks against protected system . . . 71
4.5 Experiments . . . 73
4.5.1 Database . . . 73
4.5.2 Experimental setup . . . 74
4.5.3 Experimental results . . . 75
4.6 Discussion . . . 79
4.6.1 On the security of the unprotected systems . . . 79
4.6.2 On the security of the protected system . . . 82
4.6.3 On generalizability of the security analysis . . . 84
4.7 Conclusion . . . 85

5 Masterkey attacks against free-text keystroke dynamics and security implications of demographic factors 87
5.1 Introduction . . . 88
5.2 Background and related work . . . 89
5.2.1 Biometric authentication pipeline . . . 90
5.2.2 Keystroke biometrics . . . 92
5.2.3 Single-shot and continuous authentication . . . 93
5.2.4 Threat surface of biometric systems . . . 93
5.2.5 Masterkey attacks . . . 94
5.3 Masterkey attack for free-text keystroke dynamics . . . 95
5.3.1 Threat model . . . 95
5.3.2 Attack methodology . . . 96
5.4 Guessing metric evaluation framework . . . 97
5.4.1 Brute force adversaries . . . 97
5.4.2 Guessing metrics for knowledge factors . . . 98
5.4.3 Guessing metrics for biometrics . . . 100
5.4.4 Parameterization of guessing metrics . . . 101
5.5 Experimental setup . . . 102
5.5.1 Keystroke data . . . 102
5.5.2 Keystroke model . . . 103
5.6 Evaluation . . . 105
5.6.1 Masterkeys in feature space . . . 105
5.6.2 Masterkeys in input space . . . 107
5.6.3 Demographics analysis . . . 108
5.7 Discussion . . . 110
5.7.1 Comparison with knowledge-based factors . . . 110
5.7.2 Continuous authentication . . . 111
5.7.3 Limitations . . . 111
5.7.4 Future work . . . 112
5.8 Conclusion . . . 112

6 A novel evaluation framework for biometric security: Assessing guessing difficulty as a metric 115
6.1 Introduction . . . 116
6.2 Related work . . . 118
6.2.1 Masterkey attacks . . . 119
6.2.2 Evaluating the effect of masterkeys . . . 121
6.2.3 Effective key length of biometrics . . . 122
6.2.4 The research gap . . . 123
6.3 The brute-force adversary . . . 124
6.3.1 Non-uniformity . . . 126
6.3.2 Fuzzy matching . . . 126
6.3.3 Uncountable biometrics . . . 128
6.4 Effective size of the biometrics space . . . 130
6.4.1 Upperbound on the guessing difficulty for biometrics . . . 130
6.4.2 Relationship between discrete and continuous biometric templates . . . 134
6.5 Measuring guessing difficulty . . . 136
6.5.1 Guessing metrics for knowledge-based factors . . . 136
6.5.2 Guessing metrics for biometrics . . . 139
6.5.3 Relationship with effective key length . . . 141
6.5.4 Usage in security evaluations . . . 142
6.6 Experimental setup . . . 142
6.6.1 Keystroke dynamics . . . 142
6.6.2 Face recognition . . . 144
6.7 Effect of sample size . . . 145
6.7.1 Impact of statistical hazards . . . 147
6.7.2 The region of stability . . . 149
6.8 Analysis of face and keystroke biometrics . . . 150
6.8.1 Comparing face and keystrokes with knowledge factors . . . 150
6.8.2 Choosing the superior model . . . 151
6.8.3 Dictionary specificity . . . 153
6.9 Discussion . . . 154
6.10 Conclusion . . . 155

7 Conclusion 157
7.1 Summary of the contributions . . . 157
7.2 Charting future paths . . . 161
7.3 Concluding remarks . . . 166

A Additional materials for masterkey attacks against keystroke dynamics 167
A.1 Mathematical symbols . . . 167
A.2 Results of statistical analysis . . . 167

Bibliography **171**

List of publications **199**

List of figures

1.1 The three pillars that reflect the research activities of this PhD. The pillars support the end goal of validity and reliability, security, and privacy for biometric authentication. 5
1.2 Situating research results in the research landscape. 11

2.1 Architecture of a biometric authentication system [43] 21
2.2 Deployment of the building blocks in a biometric pipeline. In the local authentication scenario all building blocks are implemented on the user device. In a remote authentication scenario the sensor is always local, but feature extraction can be executed remotely or locally. The former requires transmitting the raw biometric data to the remote server, while the latter sends only the biometric feature vectors. 22
2.3 The equal error rate is defined by the similarity score at which the FAR and FRR are equal. 26

3.1 Template protection approach incorporating quantization metadata to the Fuzzy Commitment 49

4.1 Biometric Authentication Architectures, including hardware security features usage, SFE-protected/unprotected templates-based authentication, and the attack points considered in the security analysis. 64
4.2 The ANN architectures for the TCN and GRU. 66
4.3 The FRR of the system and the imposter rejection rate (IRR) of the template estimation attack based on linear regressors, where the leaked system is the HMM with 4 states and 32 mixtures, with the target systems on the x-axis. The dotted line indicates the mean, while the full line represents the median. 77
4.4 Complexity of biometric hacker attacks on mixed ANN and HMM ensembles with binary template size $n = 127$ 80

4.5 Degrees of freedom of binomial fit of impostor attempts Hamming distance, maximum effective key length for FRR $\leq 5\%$, and required effective key length for a biometric hacker hash preimage attack complexity bounded above 6 bits for HMM ensembles and binary template length 127. 81
4.6 Histogram of Hamming distances of impostor attempts in validation sets and its corresponding binomial and β-binomial fits, for $n = 255$, and the ensemble of $E = 8$ neural networks. 82

5.1 Typical biometric pipeline. The four main attacks target are the sensor (1), the interfaces (2,4,6), the processing blocks (3,7), and the storage (5). 90
5.2 Trade-off between false accepts and false rejects when selecting threshold t. 91
5.3 Comparison of guessing difficulty between two uniform probability distributions of 10^3 and 10^4 users, and the probability distribution of 4-digit PINs [35]. 98
5.4 DET curve (FMR in function of FNMR) of the keystroke biometrics models. 104
5.5 Guessing metrics expressed as effective key lengths on the y-axis, the number of users in the evaluation set (on a logarithmic scale) on the x-axis. In 5.5a the adversary knows the enrolled users, while in 5.5b users are unknown. 105
5.6 Change in α-guesswork (G) with increasing success rate α for the three attack methodologies. *diff* indicates knowledge of the enrolled users. . 108

6.1 Example distributions of biometrics on the surface of a sphere. 6.1a shows an ideal uniform distribution. 6.1b demonstrates the non-uniformity. 6.1c illustrates that the biometrics do not populate the entire surface of the sphere but only a subspace known as the manifold. 120
6.2 Illustration of the primary challenges when computing guessing difficulty of secret data: non-uniformity, fuzzy matching, and uncountability. For randomly chosen secrets, every codeword is equally probable, and each codeword matches only with itself. Human-chosen secrets introduce non-uniformity. Error correction allows one codeword to match multiple codewords. For an ideal biometric, the secrets are no longer discrete but continuous feature vectors, i.e., the codepoints are distributed uniformly on an n-ball. Matching happens by computing the angular distance. When dealing with real biometrics, all challenges come together as the data is non-uniform, the events are uncountable, and the matching is fuzzy. 125

6.3 An optimal guessing sequence of 4 guesses that exhaust all 4-bit codewords, where codewords match if their Hamming distance is 1 or less. The initial success rate of a guess is $\frac{5}{16}$ which drops to $\frac{3}{16}$ for the third and fourth guesses. 127
6.4 The biometric pipeline. The problem space and feature space are clearly indicated, as well as the attack points where an adversary can inject masterkeys. 129
6.5 Comparison of guessing difficulty between two uniform probability distributions of 10^3 and 10^4, and the probability distribution of real-world 4-digit PINs [35]. 138
6.6 The relationship between the effective key length (EKL) found with Algorithm 2 and the α-work-factor μ_α. These metrics are computed for a hypothetical biometric where the users are sampled uniformly on the 15-ball and the threshold is set to $\frac{7}{15}$ 141
6.7 DET curve of the keystroke and face biometric systems. 145
6.8 Wolf Attack Guessing Effectiveness: The y-axis displays guessing metrics represented as effective key lengths, while the x-axis shows the number of users in the evaluation set (logarithmic scale). This illustrates the adversary's performance executing a wolf attack. 146
6.9 Wolf Sample Transferability: Adversarial guessing effectiveness when using wolf samples from an attack dataset to impersonate users in a target dataset. The x-axis shows the unique number of users in the attack dataset on a logarithmic scale, while the y-axis represents guessing metrics expressed as effective key lengths. The target dataset contains 50,000 unique users, and its size is fixed. 147
6.10 Change in α-guesswork (G) with increasing success rate α for the four biometric authentication systems (2 keystrokes and 2 face). 153

7.1 [Repeated picture Figure 1.1] The three pillars that reflect the research activities of this PhD. The pillars support the end goal of valid and reliable, secure, and private biometric authentication. 158

List of tables

3.1 Comparing the EER of related work on accelerometer based gait authentication schemes. 43
3.2 Performance in terms of EER$\%$ of 4 states, 32 Gaussian mixtures per state HMM-UBM system using different sets of features for the different sensors (Left, Center, and Right). 52
3.3 Performance in terms of EER$\%$ using the Center sensor of HMM-UBM configurations with S states and M Gaussian mixtures per state for constant $C = S \cdot M = 128$ using the $\mathrm{RPVH} \cup \mathrm{GD}$ feature set. 53
3.4 Performance in terms of EER$\%$ using the Center sensor of HMM-UBM configurations with $S = 4$ states and different M Gaussian mixtures per state, using the $\mathrm{RPVH} \cup \mathrm{GD}$ feature set. 53
3.5 Unprotected systems incorporated into the protected systems. D stands for the number of eigenfeatures computed for each state, and N is the number of unprotected systems with these parameters incorporated into the protected system. 54
3.6 Protected systems percentual performance figures for different code parameters. 54

4.1 Mean authentication performance in the test fold of the different systems in terms of EER$(\%)$, HTER$(\%)$ for the validation set EER threshold, and FAR$(\%)$ when the FRR in the validation set is at most 5%. 75
4.2 Security evaluation of protected ensembles for different binary template sizes n. 79

5.1 Guessing metrics computed on $10,000$ users for several combinations of adversary and attack. The results are compared with the entropy of a zero-effort attacker as well as traditional knowledge factors such as passwords [33] and PINs [35] 106

5.2 Estimates of the guessing metrics for relevant demographic splits. We only provide groups with more than 1000 unique users, which excludes several *native languages*. Values with an asterisk (*) are not significantly different, i.e., p-value ≥ 0.01. 114

6.1 Overview of the factors impacting the evaluation of guessing difficulty and the approach used to evaluate them. 126
6.2 The 1^{st} and 99^{th} percentiles of computing p_1 for 1.000 different samples taken uniformly from the 15-ball. The sample size is varied from 100 to 50.000. The angular distance is $\frac{7}{15}$, which gives an FMR of $\approx 1.67\%$. 148
6.3 The target dataset's fraction of non-impersonated individuals is measured during the transfer of wolves from an attack dataset. This transfer involves exploring different attack dataset sizes and the adversary's utilization of various numbers of wolves in the guessing sequence. 149
6.4 The guessing metrics were calculated on a complete dataset consisting of 110,000 users, as well as on random subsamples of a specific size drawn 100 times. The results include the 1^{st} and 99^{th} percentiles for the subsamples. 150
6.5 Guessing difficulty for keystroke dynamics and face recognition. Two different attack techniques are reported, the wolf attack, which provides an upper bound, and the masterkey attack, which provides a lower bound. The results are compared with the entropy of traditional knowledge factors such as passwords [33] and PINs [35] 152
6.6 Estimates of guessing difficulty for keystroke dynamics when the adversary has knowledge of soft-biometric characteristics. We only provide groups with more than 2.000 unique users, which excludes several *native languages*. 156

A.1 An overview of the mathematical symbols used and the section where they were introduced. 168
A.2 p-values for the ANOVA tests. In the case of the native language, the p-values are obtained with an Alexander-Govern test. 169

List of acronyms

BER	Bit Error Rate
BCH	A type of ECC named after Bose–Chaudhuri–Hocquenghem
CMA-ES	Covariance Matrix Adaptation Evolution Strategy
CNN	Convolutional Neural Network
DET	Detection Error Trade-off
ECC	Error Correcting Code
EER	Equal Error Rate
EKL	Effective Key Length
FAR	False Acceptance Rate. Sometimes also called FMR
FIDO	Fast Identity Online
FMR	False Match Rate. Sometimes also called FAR
FN	False Negative
FNMR	False Non-Match Rate. Sometimes also called FRR
FP	False Positive
FRR	False Rejection Rate. Sometimes also called FNMR
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Unit
HL	Hold Latency
HMM	Hidden Markov Model
HTER	Half Total Error Rate
IL	Interkey Latency
IMU	Inertial Measurement Unit
LSTM	Long Short-Term Memory
ML	Machine Learning
PL	Press Latency
PLP	Perceptual Linear Prediction

RNN	Recurrent Neural Network
RL	Release Latency
SVM	Support Vector Machine
OSVM	One-class Support Vector Machine
TN	True Negative
TP	True Positive
UBM	Universal Background Model

Overview of datasets

OU-ISIR gait database, inertial sensor dataset

The research conducted in Chapters 3 and 4 leverages the OU-ISIR [168] dataset published by the Institute of Scientific and Industrial Research (ISIR), Osaka University (OU). The OU-ISIR dataset comprises two subsets optimized for different variation factors. The first subset maximizes the number of subjects, consisting of level walk data from 744 individuals (389 males, 355 females) aged 2 to 78 years. Motion trajectory constraint and signal autocorrelation were employed to automatically extract two different level-walk sequences for each subject using the center IMUZ. In the second subset, emphasis was placed on variations in ground slope, sensor location, and sensor type. Two sequences for level walk, up-slope walk, and down-slope walk were manually extracted for each subject and each sensor, resulting in a sub-subset from 3 IMUZ sensors for 495 subjects and a sub-subset from smartphones for 408 subjects. The data extraction process was synchronized with simultaneously captured videos.

Dhakal keystroke dataset

The Dhakal keystroke dataset [70] is used in Chapters 5 and 6. The Dhakal keystroke dataset captures keystrokes from physical keyboards, recording the keycode, downstroke (press) timing, and upstroke (release) timing. The dataset is noteworthy for containing 15 sentences typed by over 168,000 users, making it the most extensive free text-typing dataset. Participants engaged with a web form to input their typing behavior. They were tasked with memorizing a typical sentence and subsequently typing it out. The sentences, randomly chosen from a corpus of 1525 common English sentences, all contained at least three words.

C-MS-Celeb

The C-MS-Celeb [116] dataset is a cleaned version of the MS-Celeb-1M dataset, comprising 6,464,018 images of 94,682 individuals. The original MS-Celeb-1M dataset, created by Microsoft Research, comprises 10 million face images from the internet. This dataset deemed the world's largest publicly available face recognition dataset, includes images of nearly 100,000 individuals. The dataset prominently includes American and British celebrities. The term "celebrity" extends beyond Hollywood, encompassing individuals who maintain an online presence for professional reasons, such as journalists, artists, musicians, activists, policymakers, writers, and academics.

Chapter 1

Introduction

"If you cannot measure it. You cannot improve it."

— Lord Kelvin

In today's digital age, a substantial portion of our identity is expressed and stored online through various platforms and services that we engage with daily. Imagine the wealth of personal data shared on social media, the financial transactions conducted online, or the conversations held through e-mail and chat. This digital transformation has exposed us to unprecedented risks from malicious actors seeking to exploit and misuse our (personal) information. As a result, safeguarding our online presence with reliable, user-friendly, and secure authentication methods that can accurately verify identity is imperative.

Despite the intricate and multifaceted nature of identity, most existing authentication methods still rely on knowledge-based factors, such as passwords or PINs, or possession-based factors, such as tokens or cards. However, these methods suffer from numerous security limitations, making them vulnerable to issues like guessing, phishing, or theft. Furthermore, they often prove to be inconvenient, as exemplified by the existence of password managers – specialized software created to manage the labyrinthine assortment of user credentials. Hence, there is a pressing need for alternative or complementary authentication methods that offer a seamless and secure user experience.

As humans, we do not generally rely on knowledge and ownership as the identifying characteristics of our peers. Instead, we rely on appearance and even behavior. Advances in machine learning (ML) provide computer systems with similar recognition

capabilities. Moreover, ML is very adept at identifying patterns in all sorts of data, such as images, keystroke timings, accelerometer data, and more. These new capabilities enable digital systems to seamlessly and securely verify identity. Appearance based biometrics leverages unique biological characteristics like facial features, providing a nearly frictionless user experience. Taking it one step further, behavioral authentication consistently verifies the user's identity rather than only once by monitoring their behavior for anomalous patterns, e.g., gait (human locomotion), typing behavior, and other system interactions. As a result, the authentication process becomes transparent to the user.

1.1 Challenges of biometric authentication

Biometric authentication promises a seamless and secure authentication experience, but it is not without its own unique set of challenges.

Uncertainty in the Authentication Process Biometric authentication introduces a level of uncertainty into the authentication process. Unlike traditional passwords, with only one correct answer, biometric data is inherently variable. Two biometric acquisitions are never 100% identical. Consequently, the biometric comparison process must strike a delicate balance. On one hand, it must tolerate minor deviations from the original data to prevent legitimate users from being wrongfully denied access. On the other hand, it must be sufficiently rigorous to prevent impostors from gaining unauthorized access. A dependable authentication system must minimize such errors, which can be achieved by selecting a biometric modality that meets the following criteria:

- **Uniqueness:** Each individual's biometric data should be distinct and distinguishable from that of others.
- **Permanence:** Biometric data should exhibit stability and consistency over an extended period.

Once a unique and permanent modality is identified (e.g., iris recognition), the primary challenge is developing methods to distinguish individuals based on their biometric characteristics reliably. Throughout this dissertation, we refer to this ability as "accurate biometric recognition." This capability is then leveraged to learn a representation of specific individuals. Once the concept of a specific individual is learned, it can be used for identity verification (authentication) or recognizing individuals within a group (identification). This dissertation focuses on the former.

Countering Adversarial Efforts As a security control, authentication must actively defend against malicious attempts by adversaries to impersonate others. Security threats range from counterfeit biometrics, such as using a photo to deceive a face authentication system [108, 215], to sophisticated attacks that target the machine learning component. These attacks aim to manipulate the training process [27, 28, 89, 138] or introduce changes to the input that serve as 'optical illusions' for the system [30, 220, 50, 137]. Consequently, relying on biometric authentication demands ongoing vigilance to address emerging threats.

Protecting Stored Biometric Reference Templates It is imperative to protect stored biometric reference templates from a privacy and security perspective. From a privacy standpoint, these templates should be unlinkable [34, 109], preventing authentication providers from colluding to determine which users they have in common. Given the inherent privacy-sensitive nature of biometric data, it must be safeguarded to prevent function creep and minimize damage in the event of a database breach. Protecting stored biometric data is crucial for security, as leaked biometric data can be exploited for presentation attacks. Therefore, biometric templates should be stored in an irreversible, secure form. For added security, revocable templates allow individuals to replace compromised templates with new ones, mitigating the long-lasting consequences of a security breach.

Addressing Non-Uniformity in Biometric Data Biometric data is not uniformly distributed [238, 71, 63], which can negatively impact the security of biometric authentication systems [36, 171, 259]. Currently, metrics assessing biometric authentication systems inadequately address the impact of non-uniformity.

Demographic Bias in Biometric Recognition Models There have been notable findings revealing demographic biases in biometric recognition models [94, 40]. These biases can amplify pre-existing human biases in law enforcement and have led to partial bans on face recognition technology in law enforcement [182, 141].

In conclusion, advancing biometric authentication involves more than just reducing errors by extracting identifying features from biometric data. The system must anticipate and defend against adversaries aiming to impersonate others by manipulating input data. Understanding the robustness of a biometric authentication system requires defining a clear threat model and evaluating an adversary's effectiveness within that model. Such an evaluation extends beyond measuring the ML system's ability to distinguish individuals. Lastly, biometric authentication systems store reference data of individuals, which can be vulnerable to leaks or misuse by service providers. As such, this dissertation explores the reliability of biometric authentication through three key dimensions:

1. **Accurately extracting identifying patterns from (behavioral) biometrics.** The quest for effective authentication systems that seamlessly blend usability, security, and deployability has led to the rise of several new authentication schemes [34]. Facial and fingerprint authentication gained popularity because they hit a sweet spot in the usability-security tension field for smartphone unlock. However, using behavior over appearance can provide an even smoother and more secure authentication experience for several reasons: Firstly, behavior can be collected in the background without interrupting the user. Secondly, behavior allows for continuous identity verification instead of verifying identity at a single moment, making it an effective defense against account hijacking. Despite the many benefits of behavior, they remain relatively unexplored due to lower recognition capabilities than appearance. The primary difficulty with behavior is that it produces less consistent signals influenced considerably by the context. Moreover, when privacy-preserving techniques are applied, the performance might drop further.

2. **Accounting for novel attack techniques.** The adversarial context of authentication implies that adversaries will aim to breach the system. Consequently, we have to understand the threat surface of the system as a whole. Moreover, an improved understanding of novel attacks against the pattern-matching approach that drives biometric authentication informs defensive actions and improves understanding of the limitations of biometric authentication. Currently, the biometric community is mostly concerned with presentation attacks which consist of presenting forged biometric data generated by learning the appearance or behavior of the victim or a very weak zero-effort attacker that uses their own biometric data and hopes for an accidental mismatch. There is a real need to go beyond these two types of adversaries and consider more novel attacks targeting the machine learning component. Furthermore, a well-executed evaluation of new security threats within a clearly defined threat model can be repurposed as a security metric. The security metric, in turn, can be used to describe the system's validity.

3. **Effective security measures for biometric authentication.** A valid and reliable authentication system requires a minimal security level to be offered. To adequately describe whether a biometric authentication system upholds these requirements, we need meaningful measures for the strength of an authentication system that facilitate comparisons with alternative authentication solutions. Such comparison is complicated as biometric authentication schemes allow a tolerance when matching biometrics, while traditional authentication factors, such as passwords, require an exact match. The fuzzy matching approach implies that mismatches can occur, which can be leveraged by an adversary to impersonate their victim more easily. Currently, the probability of a mismatch is evaluated

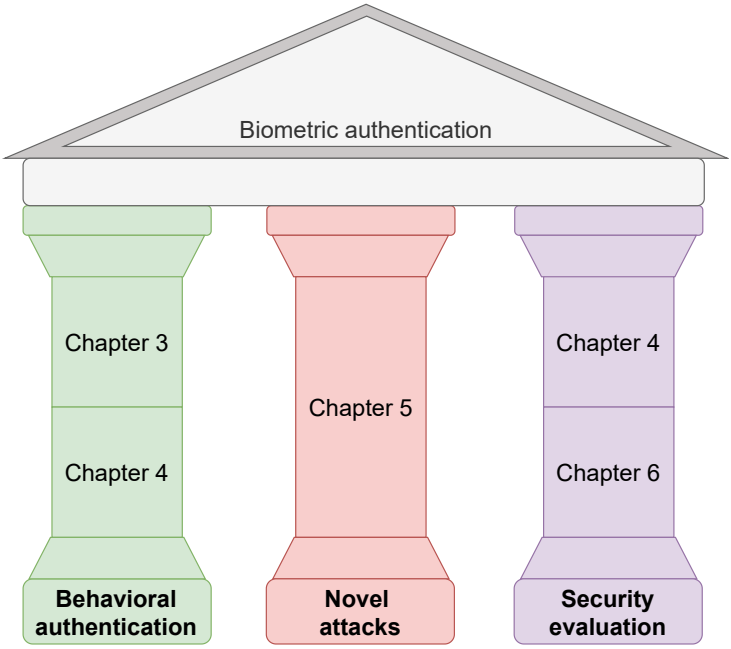


Figure 1.1: The three pillars that reflect the research activities of this PhD. The pillars support the end goal of validity and reliability, security, and privacy for biometric authentication.

without an adversary present and is known as the False Acceptance Rate (FAR). Effective measures should account for adversarial manipulation.

1.2 Pillars of the thesis

The dissertation seeks to contribute to the advancement of biometric authentication systems. We identified three verticals (cf. Figure 1.1) through which we research biometrics. In the first vertical, we research the problem of extracting the identifying features from behavior to offer a seamless authentication experience that respects the security and privacy-sensitive nature of the biometric data. The second vertical explores the attack surface of the biometric authentication system. We further narrow this vertical down by researching how an adversary can exploit common biometric patterns and the fuzziness of biometric matching to gain an advantage. The third vertical researches the metrics used to express the validity of the biometric authentication system, i.e., the security level offered by it. Currently, the most common security metric to describe

the potency of a biometric modality is the false acceptance rate, which is only relevant in the presence of a very weak and unmotivated adversary [30]. Armed with our knowledge of behavioral authentication and its threat surface, we face the challenge of defining meaningful security metrics and allow comparing biometric authentication systems with knowledge-based factors.

In overview, this disseration is organized around three main themes that act as the supporting pillars for trustworthy biometrics. The core themes are itemized below and depicted in Figure 1.1:

- Behavioral authentication
- Novel attack vectors
- Security evaluation

1.2.1 Behavioral authentication

Behavioral authentication provides a frictionless user experience and continuously verifies identity. The user experience of behavioral authentication is seamless because the authentication process leverages the users' natural interactions with the system. When an authentication decision is required, the necessary data has already been collected, and thus, the user must not be interrupted for proof of identity, such as a password. Moreover, the real-time collection of behavioral data allows a different approach to authentication: Instead of only authenticating the user at the beginning of a session, the user's identity is continuously verified. Continuous authentication provides an authentication process that is more secure against session hijacking[1].

Despite its advantages, behavioral authentication poses obstacles that must be addressed. First and foremost, behavior varies depending on the context. Therefore, behavior is often less distinctive than 'strong' biometrics such as fingerprints, iris, and face. Secondly, behavior changes over time, e.g., when a user becomes more skilled in typing on a keyboard, their typing behavior changes. Lastly, the required behavior might have yet to be exhibited, and acquiring it would require significant user friction. For example, imagine a gait authentication system, but the user has been sitting for the last hour.

In this dissertation, we aim to push the boundaries of behavioral authentication. Our ambitious goal was to advance the state-of-the-art of IMU-based[2] gait recognition from

[1] Protection against session hijacking depends on the biometric modality. For example keystrokes are continuously used to interact with a personal computer. Gait authentication is less suited as a defense against session hijacking as it requires to continuously walk during the session.

[2] IMU = inertial measurement unit. An IMU measures the acceleration, angular velocity, and direction w.r.t. the earth's magnetic field.

5% to 1% Equal Error Rate (EER). EER is the working point at which False Acceptance Rate (FAR) equals False Rejection Rate (FRR). We have reached this requirement in our applied machine learning research (Chapters 3 and 4). Later, we also developed free-text keystroke authentication systems for the purpose of researching novel attacks, but did not specifically contribute a new approach to keystroke recognition; instead, we replicated prior work[3].

One of the primary challenges to overcome to achieve accurate gait recognition is finding the most informative representation of the time-series data. Moreover, we defined a data representation that simplified learning the identifying characteristics (Chapter 3).

A second challenge is understanding the learning problem and determining a learning approach. Biometric recognition is an open-world problem because not all individuals are known at training time. Furthermore, there are typically only a limited number of data points per individual. Therefore, new concepts (individuals) must be learned based on a few samples. We introduced two distinct approaches to the gait recognition field: generative approaches from voice recognition [124] (Chapter 3) - eigenfeatures extracted from Hidden Markov Models with Universal Background Models (HMM-UBM) - and discriminative approaches used for face recognition [213, 66] (Chapter 4)- deep metric learning.

To further improve recognition performance, we train several systems and fuse their output (Chapter 3 and 4). Every system learns a similar set of identifying characteristics, yet, there are slight differences. These differences are due to machine learning optimizing to a local optimum rather than a global one. Therefore, even systems with the exact same approach and trainable parameters trained on the same training set will differ and provide value when used in conjunction with their siblings. Of course, because the amount of information in biometric data is limited, the marginal return of adding more systems diminishes. In Chapter 4, we describe how to find the optimal working point.

The last challenge for all biometric systems is processing and storing sensitive personal data. A correct working system will store a summary of the biometric signal - the biometric template - in a database and compare it with new biometric samples during operation. To ensure secure and private storage and processing of this sensitive data, biometric template schemes [166] have been proposed. These schemes ensure that biometric templates are revocable, unlinkable, and confidential. From the beginning, we have incorporated such a scheme in our gait authentication work (Chapters 3 and 4). Furthermore, Chapter 4 demonstrates that a leaked biometric

[3]We chose to investigate gait as a biometric modality due to its potential for innovation in a niche area of authentication technology. This decision was influenced by industry collaborations and a desire to contribute to the advancement of biometric authentication. However, for later research we switched to keystroke dynamics as there are bigger datasets, and it exhibits the same interesting characteristics as gait.

template contains sufficient information to reconstruct biometric templates extracted by entirely different systems accurately. Thereby, we further demonstrate the importance of biometric template protection. We show a correlation between unprotected templates and further motivate the use of template protection.

Gait as a niche biometric modality was an excellent starting point for the PhD trajectory. The subject allowed us to contribute to the field while expanding our expertise. However, it is a niche biometric for some inherent limitations w.r.t. permanence, i.e., an injury, change of footwear, or surface impacts human locomotion. For the same reason, the contributions have been achieved in lab conditions without deployment. An actual system must account, among others, for energy constraints, computational resource restrictions, and sensor deviations.

In conclusion, our contributions can be summarized as follows:

- We contribute several approaches to IMU-based gait recognition that improve the state-of-the-art and achieve close to 1% EER in a challenging benchmark dataset. (Chapters 3 and 4)
- We further improve recognition performance by using mono-modal inter-model fusion. (Chapters 3 and 4)
- We demonstrate the importance of biometric template protection by devising a biometric template reconstruction attack that leverages leaked biometric templates. (Chapter 4)
- We apply template protection and thoroughly evaluate its effect on security. (Chapters 3 and 4)

1.2.2 Novel attacks

Machine learning presents a novel attack surface, teeming with an abundance of diverse attack methodologies. For example, academic researchers have shown that ML is vulnerable to adversarial input [229, 26], poisoning [29, 96], membership inference [222], model inversion [83] and model stealing [235]. Despite the enormous interest of the research community, most of these new attacks have primarily been explored in an academic context and have seen limited real-world [16] (ab)use. The limited use of novel attacks can, among others, be attributed to the lack of incentive for the cybercriminal ecosystem to adopt them as attack vectors. According to Apruzzese et al. [16], the primary reason for this lack of incentive is the limited applicability of these novel attacks because they target an isolated ML component rather than a complete system. This situation is regrettable, as understanding how the attack methodologies are incorporated into attack vectors informs practical defensive actions.

In our work, we design and systematically evaluate a novel attack against the biometric authentication system. Moreover, we explore the concept of a masterkey attack against free-text keystroke dynamics. A masterkey is a key that fits multiple locks. The idea of a biometric masterkey is interesting for an adversary as they can leverage such a key to impersonate a surprisingly large part of the population even when they have no knowledge of the biometric traits of their victim, and the biometric recognition model has a low false acceptance rate. To execute a masterkey attack, an adversary requires the ability to spoof data, which is challenging in the presence of security controls such as spoofing detection and specialized secure hardware. Therefore, we specifically choose to execute this attack against keystroke biometrics, where spoofing is relatively easy as it relies on general-purpose input where the aforementioned protection mechanisms are absent.

The idea of a masterkey attack is the basis of Chapter 5, in which we present the research on this attack. The contributions of this research are the following:

- The design and implementation of novel attack methodologies to generate masterkeys for typing sequences. The most effective masterkey attack leverages a black-box optimization strategy that maximizes the number of individuals impersonated by the 'masterkey typing behavior.'
- To systematically evaluate the effectiveness of masterkeys, we propose a novel evaluation framework. Our framework allows evaluating security against a (biometric) brute-force adversary. It is the biometric brute-force adversary that leverages masterkeys to gain an advantage. A brute-force adversary has also been extensively studied for knowledge-based authentication factors such as PINs and passwords. As such, our evaluation framework allows us to draw a parallel between the security of knowledge factors and biometrics.
- Our attack and framework provided the necessary tools to generate new insights on the impact of demographics on the security of free-text keystroke biometrics. Among others, we can measure the impact of typing speed, age, mother tongue, and the number of fingers used while typing.

1.2.3 Security evaluation

A valid and reliable biometric authentication system requires a minimal recognition performance of the ML model. Because authentication is a security control, the recognition performance implies a security risk. This security risk is expressed as the False Acceptance Rate (FAR), also known as the False Match Rate (FMR). The FAR is the proportion of randomly selected false identity claims the recognition system erroneously accepts. So, this metric indicates how often a model accepts an impostor, a critical security breach.

The problem with the FAR as a security metric is that it cannot easily be compared with the well-known and more intuitive effective key length (bits of security). The effective key length is a logarithmic scale upper-bound security metric used in cryptography that captures resilience against brute-force adversaries that try to break the system by guessing the correct secret. The incompatibility of these two metrics becomes especially prevalent when biometric template protection is used, i.e., fuzzy commitment [72]. The effective key length describes the security of the protected biometric system. However, not only is the effective key length relevant, but an adversary leveraging the accidental mismatches (FAR) remains applicable. The problem of incomparable metrics is exacerbated when the FAR becomes hard to measure due to very low error rates, which leads to systems claiming 0% FAR with a very high effective key length. These claims are very improbable because the high effective key lengths are obtained by adding increasingly correlated bits. Intuitively one feels that the marginal gain of adding another significantly correlated bit is small and limited. However, the current metrics do not allow to quantify this effect.

In Chapter 4, we solve the problem of incomparable metrics by translating the FAR to an effective key length even in the presence of low FAR. Our approach gives us the tools to pinpoint the optimal fusion of correlated data sources. On top of the two practical adversaries - the traditional adversary and the biometric hacker - we describe a third adversary which gains an even further advantage by leveraging correlation in the biometric data and the error-correcting procedures used by the biometric template protection scheme. We do not provide an attack methodology for this adversary, which serves as a theoretical security lower bound.

In Chapter 6, we revisit the metrics of Chapter 4 based on the new insights acquired during our work on masterkey attacks (Chapter 5). Masterkey attacks demonstrate that the *biometric (brute-force) adversary* can gain an advantage due to non-uniformity in the biometric distribution. We did not account for this effect while translating the FAR to an effective key length in Chapter 4. Moreover, masterkey attacks show that the false acceptance rate only captures accidental mistakes and does not inform on the robustness against more deliberate adversaries. Therefore, we develop a new evaluation framework based on guessing metrics for passwords [33]. The new framework provides an absolute and comparable metric that enables comparisons between biometrics, PINs, and passwords.

We summarize our primary contributions as follows:

- We retire FAR, the fraction of accidental false matches, as a security metric in favor of more intuitive and comparable metrics.
- We propose a security evaluation framework for fuzzy commitment schemes considering traditional and biometric hackers. We also derive a theoretical lower

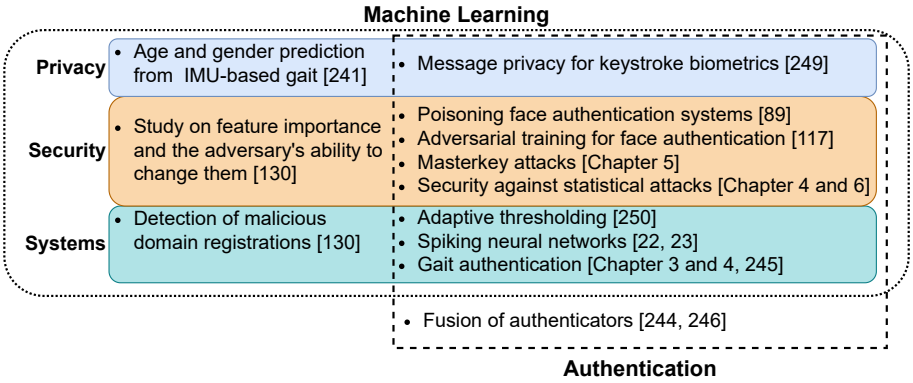


Figure 1.2: Situating research results in the research landscape.

bound for the security level that accounts for the information loss due to the error correction.

- We apply our framework and perform an empirical security analysis for a protected gait authentication system.
- We introduce guessing metrics as a more suitable way to measure the security of biometric systems. Guessing metrics capture the distribution of false accepts rather than solely relying on its mean. Furthermore, by adopting guessing metrics for biometrics, we can compare the security offered by biometrics, passwords, and PINs.
- We revisit our first evaluation framework in light of guessing metrics.
- We revisit our framework and adapt it to guessing metrics. We demonstrate its applicability to keystroke dynamics and facial authentication systems.

1.3 Other contributions

The three pillars (core themes) in this PhD thesis constitute the foundation for reliable biometric authentication systems. However, the pillars contain far more material than the reported works comprising this dissertation's chapters. Throughout this PhD trajectory, there have been contributions to a vast body of work on trustworthy ML and authentication and their intersection. Some of these works stem from efforts guiding new Ph.D. candidates, while others fit the broader research agenda of the lab in advancing the state-of-the-art of ML for security and security for ML. Figure 1.2 visualizes the published research results and situates the area of the contributions.

Privacy

Privacy-preserving Behavioral Authentication on Smartphones. This paper [249] stems from a master thesis and introduces a privacy-enhanced approach to behavioral authentication based on keystroke dynamics. The goal of the approach is to protect the typed message from an honest-but-curious third party that provides the behavioral recognition service. Three privacy-preserving techniques are implemented and evaluated: permutation, substitution, and suppression of the bigram timings.

> G. Vassallo, T. Van hamme, D. Preuveneers, and W. Joosen. "Privacy-preserving Behavioral Authentication on Smartphones". In: *Proceedings of the First International Workshop on Human-Centered Sensing, Networking, and Systems*. 2017, pp. 1–6

A Systematic Comparison of Age and Gender Prediction on IMU Sensor-Based Gait Traces. While inertial measurement unit (IMU) sensors, particularly accelerometers and gyroscopes found in smartphones and wearables, have been widely adopted for activity recognition applications like fall detection and step counting, we caution that these sensors may inadvertently expose sensitive information beyond their intended use cases. Specifically, we explore the extraction of age and gender information from gait traces captured by IMU sensors, employing various feature engineering and machine learning algorithms, including traditional and deep learning methods [241]. Our team achieved the top solutions in the OU-ISIR Wearable Sensor-based Gait Challenge: Age and Gender (GAG 2019) at the 12th IAPR International Conference on Biometrics, excelling in age and gender predictions. The research findings demonstrate the feasibility of reasonably accurate age and gender prediction based on just a few seconds of gait traces.

> T. Van hamme, G. Garofalo, E. Argones Rúa, D. Preuveneers, and W. Joosen. "A Systematic Comparison of Age and Gender Prediction on IMU Sensor-Based Gait Traces". In: *Sensors* 19.13, July 2019, p. 2945. ISSN: 1424-8220. DOI: 10.3390/s19132945

Security

A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints. Malware utilizes Domain Generation Algorithms (DGAs) to generate random domain names for controlling their botnets. To counteract this, law enforcement agencies engage in reverse engineering of these DGAs to proactively take down the corresponding malicious domain names. However, a challenge arises when legitimate domains unintentionally coincide with the generated malicious ones. Therefore, investigators must first distinguish between domains registered for malicious purposes and those registered as contact points for the botnet. Failure to do so can have detrimental effects on website owners and victims of the botnet. To address this issue efficiently, we have developed a machine learning (ML) classifier [130]

that automates the labor-intensive task of classifying thousands of registered DGA domains. Additionally, we have introduced two innovative concepts in leveraging ML for security solutions. The first concept involves analyzing the importance of features and the potential cost for an adversary to modify them. This analysis is crucial to understand the security of the ML-based solution because adversaries may attempt to alter their behavior to evade the classifier. However, making such changes should ideally be impossible or, at the very least, highly costly. Our investigations have revealed that modifying registration behavior imposes an additional cost on the adversary, given that the most significant features are timing-based. Consequently, the adversary would need to register domain names significantly earlier, introducing a higher difficulty level. The second concept revolves around utilizing the classifier in a semi-automated manner. The confidence level of the classifier is utilized to determine when the classifier's decision can be relied upon and when further manual inspection is necessary. By incorporating this semi-automated approach, we aim to enhance the accuracy and reliability of the classification process, reducing the risk of misidentifying domains.

> V. Le Pochat, T. Van hamme, S. Maroofi, T. Van Goethem, D. Preuveneers, A. Duda, W. Joosen, and M. Korczynski. "A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints." In: *NDSS*. 2020

Fishy Faces: Crafting Adversarial Images to Poison Face Authentication. This work [89] researches poisoning attacks against a face authentication system. The study considers attackers with partial control over a subset of the user's enrollment data. The findings demonstrate that the adversary can significantly undermine the effectiveness of the face authentication system by introducing a single meticulously crafted adversarial image into its enrollment samples. This manipulation of the enrollment data enables the attackers to sabotage the system's performance, leading to failed authentications.

> G. Garofalo, V. Rimmer, T. Van hamme, D. Preuveneers, W. Joosen, et al. "Fishy Faces: Crafting Adversarial Images to Poison Face Authentication". In: *12th USENIX Workshop on Offensive Technologies (WOOT 18)*. 2018

Improving Resilience of Behaviometric Based Continuous Authentication with Multiple Accelerometers. This research output [245] is our first study on gait authentication. It challenges the resilience of accelerometer-based gait analysis as a behaviometric under sensor displacement conditions. To improve misauthentication resistance, we present and evaluate a solution using multiple accelerometers on seven positions on the body during different activities and compare the effectiveness. From a security point of view, we investigate the feasibility of zero and non-zero effort attacks on gait analysis as a behaviometric. One of the main areas for improvement in this work is the representation of the authentication problem as a binary classification problem.

> T. Van hamme, D. Preuveneers, and W. Joosen. "Improving Resilience of Behaviometric Based Continuous Authentication with Multiple Accelerometers". In: *Data and Applications Security and Privacy XXXI*. ed. by G. Livraga and S. Zhu. Cham: Springer International Publishing, 2017, pp. 473–485. ISBN: 978-3-319-61176-1

Adversarial Robustness is Not Enough: Practical Limitations for Securing Facial Authentication. The paper [117] uses facial authentication as a case study to examine the relationship between theoretical robustness against adversarial examples and actual system security. It highlights the limitations of prominent robustness methods like Adversarial Training and Guided Complement Entropy in addressing physical adversarial examples. This paper emphasizes the importance of evaluating defenses in applications where adversarial perturbations pose genuine threats to ensure more effective protection in practical scenarios.

> S. Joos, T. Van hamme, D. Preuveneers, and W. Joosen. "Adversarial Robustness is Not Enough: Practical Limitations for Securing Facial Authentication". In: *Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*. 2022, pp. 2–12

Systems

Gait Authentication based on Spiking Neural Networks. This paper [23] introduces a novel approach to gait authentication using spiking neural networks (SNNs), which offer notable advantages in energy consumption and are well-suited to integrate with proposed neuromorphic hardware chips, potentially facilitating broader adoption of artificial intelligence technologies in user devices. However, a key challenge with SNNs lies in training the network due to the limitations of applying conventional error backpropagation, commonly used in traditional artificial neural networks (ANNs). To overcome this limitation, the paper presents a new derivation of error backpropagation for spiking neural networks, incorporating lateral inhibition, and demonstrates competitive performance compared to state-of-the-art ANNs in IMU-based gait authentication. My primary contribution here lies in providing ANN-based networks, which serve as a benchmark.

> E. Argones Rúa, T. Van hamme, D. Preuveneers, and W. Joosen. "Gait Authentication based on Spiking Neural Networks". In: *20th International Conference of the Biometrics Special Interest Group*. Vol. LNI P-315. Gesellschaft für Informatik e.V., Oct. 2021, pp. 51–60. ISBN: 978-3-88579-709-8

Discriminative Training of Spiking Neural Networks Organised in Columns for Stream-based Biometric Authentication. This paper [22] extends our prior work on SNNs [23]. We provide a more elaborate evaluation with a larger number of ANN benchmark systems and more SNN approaches.

> E. Argones Rúa, T. Van hamme, D. Preuveneers, and W. Joosen. "Discriminative Training of Spiking Neural Networks Organised in Columns for Stream-based Biometric Authentication". In: *IET Biometrics* 11.5, 2022, pp. 485–497

Beware the Doppelgänger: Attacks against Adaptive Thresholds in Facial Recognition Systems. In our research, we observed that some users are easier to recognize than others, which is not accounted for by the traditional fixed thresholding schemes in biometric matching. Adaptive thresholding overcomes this limitation but introduces new attack opportunities. This research result [250] introduces a novel attack against dynamic thresholds, which we call doppelgänger attacks. In response, we propose the concept of identity-level thresholding, effectively improving system performance and security compared to fixed and prior adaptive thresholds, even with a large user base. These findings carry important implications for designing and implementing face recognition systems, enhancing their reliability and security.

> W. Verheyen, T. Van hamme, S. Joos, D. Preuveneers, and W. Joosen. "Beware the Doppelgänger: Attacks against Adaptive Thresholds in Facial Recognition Systems". In: *The 18th International Conference on Availability, Reliability and Security Proceedings*. Benevento, Italy, Sept. 2023

Authentication

Frictionless Authentication Systems: Emerging Trends, Research Challenges and Opportunities. This paper [247] stems from our first efforts to build knowledge in the authentication field. It examines the evolving landscape of frictionless authentication systems, considering the increased prevalence of wearable and mobile devices and users' expectations for a seamless experience across diverse environments. It identifies key challenges related to user enrollment and the security, usability, and privacy trade-offs associated with continuous authentication on mobile and wearable devices. The study highlights the need for further research to address the emerging trends and meet consumer demands for secure, privacy-aware access.

> T. Van hamme, V. Rimmer, D. Preuveneers, W. Joosen, M. Mustafa, A. Abidin, and E. Argones Rua. "Frictionless Authentication Systems: Emerging Trends, Research Challenges and Opportunities". In: *SECURWARE 2017*. Vol. abs/1802.07233. IARIA, 2017, pp. 1–5

A Dynamic Decision Fusion Middleware for Trustworthy Context-Aware IoT Applications. Risk-aware authentication relies on the user context to determine a risk level. IoT devices can help to sense the user context. However, due to such devices' decentralized and heterogeneous nature, there is an inherent risk of uncertain, unreliable, and inconsistent observations. We contribute [244] an extension of the optimal Chair-Varshney decision fusion rule that better accommodates the dynamic

nature of IoT devices and adds the ability to detect and mitigate unreliable sensors.

> T. Van hamme, D. Preuveneers, and W. Joosen. "A Dynamic Decision Fusion Middleware for Trustworthy Context-Aware IoT Applications". In: *Proceedings of the 4th Workshop on Middleware and Applications for the Internet of Things*. M4IoT '17. Las Vegas, Nevada: Association for Computing Machinery, 2017, pp. 1–6. ISBN: 9781450351706

Managing Distributed Trust Relationships for Multi-modal Authentication. This paper [246] addresses the security risks in multi-modal authentication schemes that combine multiple behavioral biometrics for identity verification. These schemes often make invalid assumptions, such as fixed probabilities of correct recognition and temporal congruence of behavior metrics. To address this risk, the paper introduces a formal trust model that guides the selection and composition of behavior metrics.

> T. Van hamme, D. Preuveneers, and W. Joosen. "Managing Distributed Trust Relationships for Multi-modal Authentication". In: *Journal of Information Security and Applications* 40, 2018, pp. 258–270. ISSN: 2214-2126

Academic competition

OU-ISIR Wearable Sensor-based Gait Challenge: Age and Gender. Throughout this PhD, we won a research competition. The organizers made a high-level write-up of the different approaches [169].

> T. T. Ngo, M. Atiqur Rahman Ahad, A. D. Antar, M. Ahmed, D. Muramatsu, Y. Makihara, Y. Yagi, S. Inoue, T. Hossain, and Y. Hattori. "OU-ISIR Wearable Sensor-based Gait Challenge: Age and Gender". In: *2019 International Conference on Biometrics (ICB)*. 2019, pp. 1–6. DOI: 10.1109/ICB45273.2019.8987235

1.4 Dissertation outline

The remainder of this thesis is organized in the chapters listed below. We shortly highlight the scope of each chapter along with the scientific publications they are based on.

Chapter 2 provides an overview of related work on AI for biometrics. It introduces the main building blocks of a biometric recognition system and discusses related work for each of them in depth. This chapter also clearly sketches the threat surface of the system as a whole. We consolidate this chapter by indicating research gaps.

Publication Data: T. Van hamme, G. Garofalo, S. Joos, D. Preuveneers, and W. Joosen. "AI for Biometric Authentication Systems". In: *Security and Artificial Intelligence: A Crossdisciplinary Approach*. Ed. by L. Batina, T. Bäck, I. Buhan, and S. Picek. Vol. 13049. 8. Springer Cham, Apr. 2022, pp. 156–180. ISBN: 978-3-030-98795-4

Chapter 3 delves into gait authentication and reports on our first steps into **behavioral authentication**. We bridge the gap to $\approx 1\%$ EER using hidden Markov and universal background models. Furthermore, we protect the system and fuse different mono-modal systems to ensure that recognition performance does not degrade substantially.

Publication Data: T. Van hamme, E. Argones Rúa, D. Preuveneers, and W. Joosen. "Gait Template Protection Using HMM-UBM". in: *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2018, pp. 1–8

In Chapter 4 we extend our work on protected gait authentication (**behavioral authentication**). We explore several deep-learning approaches and report on the impact of different architectures and loss functions. Furthermore, our diverse set of learning approaches allows for demonstrating the importance of biometric template protection. Lastly, in this work, we define new security **metrics** that capture the security against traditional brute-force and biometric brute-force adversaries.

Publication Data: T. Van hamme, E. A. Rúa, D. Preuveneers, and W. Joosen. "On the Security of Biometrics and Fuzzy Commitment Cryptosystems: A Study on Gait Authentication". In: *IEEE Transactions on Information Forensics and Security* 16, 2021, pp. 5211–5224. DOI: 10.1109/TIFS.2021.3124735

Chapter 5 presents a **novel attack**, the masterkey attack. The masterkey attack provides an advantage to brute-force biometric hackers as it is a forged biometric signal that matches a surprisingly large part of the population. We devise a strategy to forge masterkeys for free-text keystroke authentication systems and systematically evaluate their effectiveness in light of guessing metrics. We must account for the fuzzy matching procedure to utilize guessing metrics for biometrics effectively.

Publication Data: T. Van hamme, G. Garofalo, D. Preuveneers, and W. Joosen. "Masterkey Attacks Against Free-Text Keystroke Dynamics and Security Implications of Demographic Factors". In: *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. Los Alamitos, CA, USA: IEEE Computer Society, July 2023, pp. 278–291. DOI: 10.1109/EuroSP57164.2023.00025

In Chapter 6, we retire the FAR as a security metric in favor of guessing metrics. We further study guessing metrics as a security metric, provide upper bounds, and study convergence. Furthermore, we revisit the metrics defined in Chapter 4 and use the new **evaluation framework** to evaluate free-text keystroke dynamics and facial authentication empirically.

Ongoing Work: T. Van hamme, G. Garofalo, E. Argones Rúa, D. Preuveneers and W. Joosen, A novel evaluation framework for biometric security: Assessing guessing difficulty as a metric.

Finally, Chapter 7 encompasses the conclusion of this dissertation, where we provide a summary of the key findings, contemplate limitations, and suggest future directions.

Chapter 2

Background

Preamble

This section provides the required background on biometric authentication. It introduces the primary concepts in biometrics, the main building blocks of a biometric system and reflects on the current state-of-the-art. This chapter is a revised version of a published book chapter on AI for biometrics [242].

Publication Data: T. Van hamme, G. Garofalo, S. Joos, D. Preuveneers, and W. Joosen. "AI for Biometric Authentication Systems". In: *Security and Artificial Intelligence: A Crossdisciplinary Approach*. Ed. by L. Batina, T. Bäck, I. Buhan, and S. Picek. Vol. 13049. 8. Springer Cham, Apr. 2022, pp. 156–180. ISBN: 978-3-030-98795-4

2.1 Introduction

Authentication involves verifying the authenticity of proof of identity. One of the earliest and most well-known examples of (biometric) authentication is handwritten signatures [12]. Signatures are widely used to sign legally binding documents, even though these signatures are rarely verified. For example, how many people check the signature on the ID card of the signing party or even on their credit card in the chip-and-sign payment paradigm? Moreover, signatures are easy to forge, reducing their level of security even further. Nevertheless, signatures work well in the real world due to the physical context and strong legislation against forgery. However, despite the prowess of handwritten signatures in the physical world, adopting them for a digital

future is tricky. The absence of a physical encounter makes it easier to challenge the signature's authenticity. Still, institutions like banks are very interested in automatic signature verification systems due to their pervasiveness.

In online authentication systems, passwords were a much better alternative as they are easier to input and compare for a machine. However, passwords have become a nuisance for users and a security burden [34]. Users develop password headaches due to increasing password lengths and a growing number of password-protected accounts. Additionally, long passwords are troublesome for mobile devices. Furthermore, system providers do not store passwords safely, as password database breaches have unfortunately become common [107]. Consequently, passwords have now become a security risk, as they are susceptible to being leaked through database breaches and phishing attempts. The widespread reuse of passwords also compounds the issue, as the fallout from their leakage can be significant.

Security researchers have searched for alternative authentication schemes for the past twenty years [34]. Biometric technologies have great potential, as they can recognize the user in a near frictionless manner. However, the adoption of biometrics outside phone unlock is still low because of concerns w.r.t. the validity and reliability of biometric recognition, its security, and privacy and data protection.

Recent advances in machine learning drastically increased recognition performance, allowing the design of more reliable and secure biometric systems. This chapter aims to elucidate the various applications of machine learning technology in the biometric authentication pipeline, and acknowledges that this progress comes with a dual effect, as machine learning also expands the potential attack surface of biometric recognition systems.

This chapter is organized as follows. Section 2.2 provides introduces the main concepts of biometric systems. The following sections discuss the primary building blocks more in depth: sensing of biometric data (Section 2.3), storing biometric information (Section 2.4), and comparing biometric data (Section 2.5). Section 2.6 concludes this chapter by summarizing the key challenges in biometric authentication and outlining avenues for future work.

2.2 Biometric system

In this section, we familiarize the reader with the main concepts in biometrics: we first describe the main building blocks [43] and introduce the two primary phases: the *enrollment* and *query* phases. Second, we map the biometric system on a typical ML pipeline. Third, we identify the attack surface. Last, we introduce the evaluation metrics.

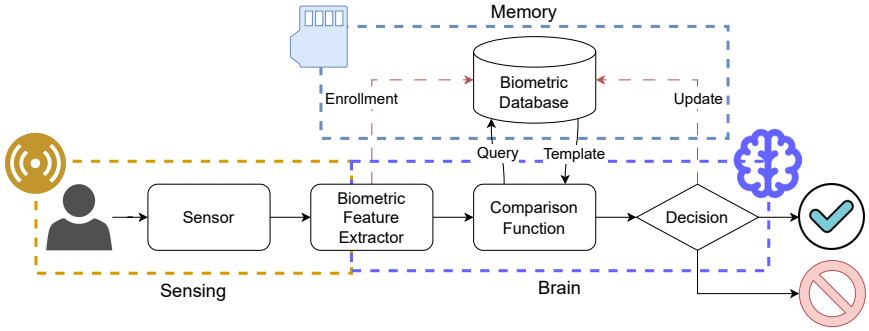


Figure 2.1: Architecture of a biometric authentication system [43]

2.2.1 System design

Biometric authentication requires three important capabilities: firstly, the ability to observe the environment; secondly, the proficiency to discern and compare identifying features; and finally, the memory to recall individuals. In other words, from a high-level, a biometric authentication system consists of senses, brains and memory. Figure 2.1 reflects this high-level abstraction and shows in more detail the typical building blocks in a biometric system, but in the real world, boundaries are more blurred, and components might overlap.

The first step in the chain acquires a biometric measurement. Sensors – such as fingerprint scanners – measure the unique biometric signal and act as the interface between the user and the system. Capturing biometrics can require cooperation by the user, e.g., laying a finger on a fingerprint, or happen in the background without prompting an action from the user, e.g., passive analysis of gait patterns via cameras or motion sensors. The raw sensor data is then cleaned in a pre-processing step, where also its validity might be checked.

The signal is then forwarded to the biometric feature extraction module. Here, the relevant features for user recognition are extracted. These biometric features are stored in a database and retrieved during comparing. Biometric features are also known as biometric templates or biometric embeddings. Typically, multiple biometric enrollment templates are stored for a single individual. They are also known as the template gallery.

The database is populated during the *enrollment phase*. The user presents their biometric trait several times to the sensor. The templates are extracted, encrypted, and stored. The enrollment phase is typically short and does not allow capturing samples in many different contexts. The length of the enrollment phase is limited by usability and security: an enrollment phase that lasts for minutes is not user-friendly.

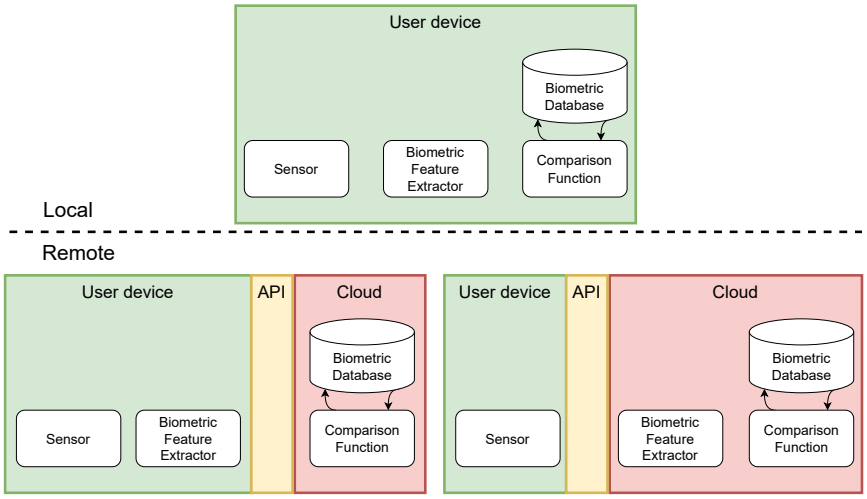


Figure 2.2: Deployment of the building blocks in a biometric pipeline. In the local authentication scenario all building blocks are implemented on the user device. In a remote authentication scenario the sensor is always local, but feature extraction can be executed remotely or locally. The former requires transmitting the raw biometric data to the remote server, while the latter sends only the biometric feature vectors.

Furthermore, the enrollment phase happens in a special security context as it is a trusted phase. Thus, the input obtained during enrollment has a significant impact on the integrity of the system.

After a user is enrolled in the system, their identity can be verified. During the *query phase*, the newly obtained query sample is compared with the biometric template. In a verification scenario, the user claims an identity used to query or index the database. In identification, the database is scanned to search for a match between the query sample and the biometric templates that populate the database.

The comparison is performed within the *comparison function* block, and, based on its output, a decision is made on whether to grant or deny the user access to the protected environment, i.e. accept or reject the user. Optionally, to deal with changes in biometric traits, a second decision block can add the query sample to the enrollment samples in the biometric database [29].

The pipeline can be implemented *local* on the user device or *remote* on a remote server, maintaining its structure. Figure 2.2 shows the most common deployment schemes of biometric pipelines. Most often the whole biometric pipeline is implemented on the user device, preferably using secure hardware. However, this only enables local unlock scenarios such as phone unlock. A more flexible deployment allows for remote

authentication. Even in remote authentication scenarios the sensing always happens locally. The feature extraction can be done either locally or remotely, while the final decision is computed on a remote server.

A biometric system can operate as a *verification* or an *identification* system. The user presents a biometric query sample during the verification and makes an identity claim. The verifier then performs a 1-to-1 comparison between the presented query sample and the enrollment samples stored under the claimed identity's profile. Differently, there is no identity claim during identification, and the identifier performs a 1-to-many comparison, which compares the query biometric to a database of enrollment samples from different individuals. Thus, the tasks answer two different questions: In the first case, it is verified whether the user is who they claim to be. In the latter case, the aim is to identify an unknown person from a set of known identities. Within the scope of this dissertation we are primarily concerned with verification.

2.2.2 The role of ML in biometric authentication

Machine learning (ML) has boosted biometric recognition accuracy [105, 4]. Depending on the chosen modality, the amount of ML used in the steps described in Section 2.2.1 varies. For example, features can be extracted in an automated manner by deep learning or be manually-engineered.

An ML model is a parametric function $f_\theta(x)$, where θ are the parameters of the model and x is the input vector. After pre-processing and possibly manual feature engineering, x is fed to the model. The training phase then searches the parameters θ that minimize a pre-defined loss function.

We can draw a parallel between the ML and biometric pipelines. The sensor measuring the biometric signal provides the input for the ML module x. The subsequent module extracts relevant features for the given task, i.e., the *feature extractor*. As outlined earlier, pre-processing and signal refinement can occur between these two steps. The *model* can incorporate the decision function and the template database. We can learn a model that encodes the identity of a pre-defined set of users as an alternative to traditional distance functions, like computing the *Hamming distance*. Finally, a decision is made based on the output of the model.

2.2.3 Attack surface

The complexity of a biometric system, comprising diverse building blocks, results in a broad attack surface [30]. An adversary can target every building block in Figure 2.1 or the interfaces between them [196]. Depending on the adversary's goal, knowledge, and

capabilities, they have several attack strategies at their disposal. From a macroscopic perspective the attack strategies can be clustered according to which of the capabilities they exploit: fooling the sensing ability with artifacts, compromising the memory, or tricking the brain. From this perspective we discuss the attack strategies:

Sensing. The sensor captures the biometric trait, which is transformed into a machine-comparable representation during pre-processing and feature extraction. Adversaries that disrupt the input do not use or manipulate their own biometrics to evade the system. Instead, they change the input data such that it resembles their victim. The following attack techniques target the input level:

1. *Spoofing or presentation attack.* This type of attack presents fabricated biometric traits to the sensor. Defenses – presentation attack detection or liveness checks – aim to verify whether the submitted data is real or fake.

2. *Replay.* Replay attacks are launched at the interfaces between the building blocks. Not only raw sensor data, for example, an image found on social media, can be replayed. An attacker can operate one step later and replay extracted feature vectors. Defenses rely on secure encrypted communication channels that implement some challenge-response protocol. However, those countermeasures might not fully stop insider threats.

The memory. Attacks on this level focus on corrupting the database or stealing the templates. Biometric templates are valuable as they contain (sensitive) information that can be exploited for other purposes than authentication or can later be used to perform replay attacks or craft presentation attack instruments.

1. *Adversarial examples.* Adversarial examples are not only a threat to the comparison function but to the biometric database as well. As mentioned before, the biometric template is a machine learning model in some biometric authentication schemes. In other words, the biometric database contains a personalized comparison function. Adversarial examples that target the comparison function level are injected during the query phase. In contrast, those that target the biometric database are injected during the enrollment or query phases. Adversarial examples injected during the query phase change the database through an automated update strategy [138]. A secure enrollment phase is paramount in defending against these attacks. Defenses for the update strategy focus on detection [138].

2. *Hill climbing.* The goal of hill climbing attacks is to extract the biometric templates from the template database. The attacker injects specific samples at

the interface, which change depending on the (fine-grained) output from the authentication system. As these attacks also target the interfaces between the building blocks, the same defenses as replay attacks apply.

3. *Insertion, deletion, and theft.* Inserting or deleting elements in the biometric database allows the attacker to manipulate the outcome of the authentication system. Alternatively, an attacker can steal the templates to perform replay attacks, craft spoofing material, or extract private information from the template. Defenses consist of biometric template protection [194] such as secure sketches, fuzzy commitment schemes [21] or cancelable biometrics [233, 199].

The brain. During the query phase, the comparison function provides a score that encapsulates the similarity between the query sample and the enrollment samples stored in the database. Attacks on this level manipulate the output of the comparison function:

1. *Physical adversarial examples.* The attacker slightly modifies the environment captured by the sensor in such a way that the machine learning component produces drastically different output. The adversaries' modifications are less drastic than spoofing attacks in the sense that they still present real biometric data to the sensor [220]. Defenses against this type of attack focus either on detection [54] or aim to add robustness to the ML component [7].

2. *Backdoor attacks.* In machine learning, a backdoor refers to a malicious manipulation or unintended flaw in a trained model. It allows an attacker to control the model's behavior in specific situations, compromising its predictions' integrity. The backdoor is typically introduced during training by injecting manipulated data with hidden triggers. When the model encounters these triggers during inference, it produces biased or compromised results, enabling the attacker to influence the model surreptitiously [132].

3. *Malware infection.* An attacker or insider manipulates the node that runs the comparison function to thwart the outcome of the comparison function. Good security practices, such as access control, secure coding practices, end-point security, and others, defend against this threat. We do not further explore this threat as it is not a specific biometric system vulnerability.

In general, the attacker can influence either the query phase or the enrollment phase. If the system allows template updating, the attacker can influence this mechanism instead of the enrollment phase. An attacker that acts during the query phase aims to evade the system either through spoofing [215] or generating physical realizable adversarial examples [220, 266]. Alternatively, that attacker can launch a hill climbing attack [178] to violate biometric template privacy. Contrary, influencing updates or enrollment leads

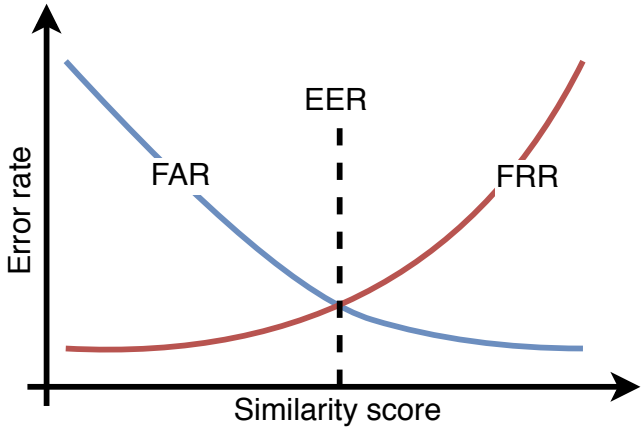


Figure 2.3: The equal error rate is defined by the similarity score at which the FAR and FRR are equal.

to *poisoning* attacks. In a poisoning attack [29, 138], the attacker compromises the biometric template to either violate system integrity or its availability, i.e., rendering it useless.

2.2.4 Evaluation metrics

The Equal Error Rate (EER) measures biometric authentication performance. As illustrated in Fig. 2.3, the EER occurs at the threshold where False Acceptance Rate (FAR), also called False Match Rate (FMR), equals the False Rejection Rate (FRR), also called false non-match rate (FNMR).

$$FAR = \frac{FP}{FP + TN} \tag{2.1}$$

$$FRR = \frac{FN}{FN + TP} \tag{2.2}$$

Determining the threshold for EER post factum is incorrect. In a real system, the threshold for the authentication decision is determined upfront, typically on the validation set. The EER threshold from the validation set is then used in the test set, which determines the FAR and FRR. To express recognition performance as one number, the half total error rate is defined as:

$$HTER = \frac{FAR + FRR}{2} \tag{2.3}$$

When choosing the threshold, there is a trade-off between security and usability. Requiring a more similarity, and thus being less tolerant, offers more security. However, being less tolerant leads to more false rejects, which penalizes usability.

When comparing the recognition performance of different systems it is beneficial to not rely on one FAR/FRR working point. A Detection Error Tradeoff (DET) curve provides a more complete picture and shows the trade-off between the FAR and FRR for different working points on a logarithmic scale.

2.3 Sensing: System input

Like password authentication deals with passwords, a biometric authentication system deals with biometric data. Reference biometric data is obtained during enrollment. This section discusses the steps to provide the steps that provide input to the biometric authentication system: firstly, biometric traits are captured by the biometric sensor; secondly, the raw biometric data is transformed through pre-processing.

2.3.1 Sensors

The sensor interfaces the physical world and the digital biometric authentication system. It captures the biometric trait and presents the digital signal – typically a multivariate time series, a 2D image, or a video – to the feature extractor. The nature of the sensor – the type of data captured, sampling rate, resolution, etc. – impacts the performance of the system drastically, but also its security, as some traits are inherently more costly to forge, for example, face authentication with 2D or 3D images [126]. Fusing sensors further enhances security without hindering authentication accuracy [51].

Since the sensor provides the input to the system, the integrity of the system is highly dependent on the trust placed in the sensor to provide signals originating from an actual individual. If an adversary can bypass the sensor, they can provide arbitrary input to the system, leading to various attacks. Therefore, biometric systems are often implemented as trusted end-to-end pipelines where the identity verifier controls all building blocks [32], e.g., physical access and local authentication on smart devices. However, in remote authentication scenarios, the pipeline is not entirely under the control of the verifier as sensing is done locally, and (some of) the other building blocks are implemented on a remote server. The degree of trust w.r.t. the input depends on the system design, but for all systems, trusted input is paramount for integrity. Therefore, prior work proposed several mechanisms to enhance trust in sensor input, i.e., secure hardware modules [90], sensor authentication by leveraging physically unclonable

functions (PUFs) [47] or sensor authentication by profiling unique sensor noise with ML [239, 149, 59, 58].

A sensor that measures a biometric signal will always capture personal information that can be used for purposes other than authentication [60]. Therefore, a user should trust that the sensor does not leak the raw biometric data and that the obtained biometric signal or a derivative will not be misused further down the biometric authentication chain [206].

2.3.2 Pre-processing

Pre-processing is paramount for recognition performance. In this step, the signal quality is checked and improved by removing noise and performing alignment. To this end, pre-processing leverages signal processing and machine learning. For example, IMU-based gait authentication sometimes requires step segmentation [67, 88, 205, 147]. The segmentation removes phase shifts and provides a single coherent unit to extract features. Furthermore, IMU-based gait authentication systems need to account for variations in sensor orientation, e.g., the multivariate time series signal is transformed into gait dynamic image (GDIs) [272] by estimating the angle between the acceleration vectors on distinct time steps.

Face recognition is another example of a biometric modality with a vast body of literature. Older methods subdivide the pre-processing pipeline of face authentication systems into four steps: face detection [262, 270, 265], pose estimation [164], landmark detection, and finally, face alignment based on these landmarks. Face alignment procedures are typically categorized either as *template fitting* [273] or *regression based* [46]. The downside of these older methods is that they rely on manual feature extraction methods that try to extract shapes, which is challenging in unconstrained environments with varying lighting conditions, occlusions, radical face poses, and different facial expressions. Convolutional neural networks (CNNs) perform much better in these real-world environments. Therefore the most recent face alignment systems rely on deep learning technology. One of the most well-known examples is MTCCN [267] which is used among others in the recent implementations of FaceNet [213]. Still, the alignment of faces remains challenging under different poses. Recently GANs have been proposed to deal with pose variations by performing pose estimation, correction, and finding a pose-invariant representation [236]. Thus, more recent methods no longer subdivide face alignment into different steps and perform the operation in one go.

We provide a last example of fingerprint authentication. A correct alignment is essential as fingerprint comparison is mainly based on extracted minutiae coordinates. Schuch

et al. [214] corrects for rotations by leveraging a siamese networks architecture built on top of CNNs [214].

2.3.3 Attacks and defenses

As mentioned in Section 2.2.3, an attacker can try to attack the system by presenting something different from the real biometric trait to the sensor, i.e., presentation attack or by injecting forged or observed samples at the interfaces between the different components, i.e., replay attack. For example, in face recognition, presentation – also known as spoofing [215] – attacks present a mugshot, a video, or a mask of the victim to the camera. Alternatively, fingerprint authentication can be tricked by fakes made from wood glue or printouts with conductive ink [44]. To perform a replay attack, an adversary must obtain knowledge of their victim's biometrics and access to the interfaces between the different components, for example, by physically opening devices or breaking 'secured' communication channels – insider attacks.

Some authentication modalities are more robust to spoofing as it is more costly to produce a fake. 2D face authentication is the cheapest to spoof, as personal images are often available online, and high-quality printouts are relatively cheap. 3D face authentication is more expensive to spoof, requiring the expertise and effort to build an accurate mask. On top of that, obtaining an accurate 3D representation of a face without the victim's cooperation is more complex.

Presentation attack detection – also known as liveness detection – is an unsolved problem in the biometrics community. The complexity of this problem is adequately demonstrated by the rise of generative AI. Deepfakes have also prompted us humans to look for means to verify the validity of input to our senses. In the field of biometrics the pursuit for a robust method to distinguish between fake and real has dedicated tracks on the major conferences: IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), IAPR International Conference on Biometrics (ICB), and International Joint Conference on Biometrics (IJCB); and a bi-yearly competition LivDet [136, 163]. Typically liveness detection is done either by adding additional hardware or in software by looking for specific characteristics of spoofs. Software-based liveness checks rely on ML; the 2015 LivDet competition winner [175] adopts CNNs to detect spoofs. They show their method is superior to earlier proposed solutions that rely on manually engineered features such as local binary pattern (LBP) extraction and more traditional ML methods like SVMs. Raja et al. [191] further explored the difference in expressiveness between manually engineered features – texture-based, LBP, ...– and automated features learned by a deep learning model and confirmed the earlier observation: PAD benefits from deep learning-based features because they are more robust and lead to better recognition performance. Sadly attackers can also leverage the power of machine learning to produce new attack vectors; for example,

due to the recent successes in speech synthesis, arbitrary speech, i.e., authentication systems that do not rely on a fixed keyphrase, have become easy to spoof. Luckily, Yan et al. [261] showed that it is possible to characterize the difference in sound generated by humans and machines.

Defenses against replay attacks are typically implemented in one of two ways: a *challenge response system* or by correlating the biometric signal with an *additional out-of-band signal* received from another sensor. The challenge-response system relies on applying a signal beyond the attacker's control and retrieving this signal from within the biometric data. For example, one can use the screen of a smartphone to produce a color-coded signal during face authentication. If the reflections match the applied pattern, one can be reasonably sure that the biometric trait is not replayed [225, 230]. Alternatively, some biometrics rely on applying a changeable signal; the biometric system can thus observe the unique reaction to the signal, i.e., the biometric used for authentication, and the signal itself for PAD, an example of such system is hand geometry [131].

Defenses based on out-of-band signal correlation obtain two signals, one with the biometric and another via a different communication channel and different sensor. Both signals are correlated since they capture a similar pattern. For example, the heart rate can be obtained by recording a face and through a PPG sensor or a finger held to the camera [227]. Another example requires moving the camera during face authentication. The movements from the video are then matched to the ones observed by the IMU sensor [133].

2.4 Memory: Biometric database

The biometric database stores the reference template. In this section, we discuss the enrollment phase, point out that some systems store ML models instead of templates, and discuss attacks and defenses.

2.4.1 Template enrollment

The template enrollment phase populates the template DB (cf. Fig. 2.1). These representations vary greatly depending on the modality and present different implementation challenges. The stored template can be either a (set of) feature vector(s) or an ML model. Furthermore, a template update mechanism can be in place.

Static templates are usually updated on-demand, as per the request of the user or the service provider. Differently, self-adapting templates deal with changes in appearance, applying continuous updates. The system can harness the samples the user provides

at verification time, thus becoming invisible to the users themselves. Self-updating strategies help to cope with natural and artificial modifications of one's biometric trait – such as a haircut or old age wrinkles for face images. In practice, the decision made at verification time is fed back to the biometric DB to update the stored template based on a specific update policy (cf. Fig. 2.1). However, unsupervised self-adaption is not without risks as an adversary can misuse this mechanism to poison the enrollment templates [138].

2.4.2 Template comparison

Template comparison is performed in the *Comparison Function* module. However, this module does not always apply its own function. If the template consists of an ML model, for example, OneClass-SVM, the comparison function is embedded in the template itself. Thus, the comparison function is passed to the *Comparison Function* module. Contrary, if the template is a (set of) feature vector(s), the comparison happens based on the function defined in the *Comparison Function* module, which is often some distance function.

2.4.3 Attacks and defenses

In real-world applications, a template should not be stored in plain text because this risks users' privacy and security. Encrypting the reference template protects against data breaches but also requires decryption for comparison, which again risks template leakage at each new authentication attempt. Therefore, biometric template protection (BTP) schemes apply a transformation that does not require the raw biometric for comparison. BTP schemes store derived artifacts that comply with the following three requirements (ISO/IEC FCD 24745):

1. *Irreversibility:* Reconstructing the reference template should be computationally hard.
2. *Unlinkability:* It should not be possible to link different templates derived from the same biometric sample.
3. *Revocability:* The reference template should be revocable so that the original biometric sample is reusable across services.

BTP prevents cross-matching and minimizes the risks associated with breaches by allowing the reuse of a single sample across different services. BTP transforms the biometric data utilizing a secret key chosen during enrolment. The biometric matching

procedure then leverages the transformed biometric enrollment data and the key to compute whether the query sample is sufficiently similar to the enrollment data. Thus, the matching happens in the transformed domain.

Irreversibility and revocability are achieved by using cryptographic-like one-way functions. The most delicate step of the process is the management of the keys. In BTP schemes, this is done in two ways [201]: (1) biometric cryptosystems bind a key to or generate a key from a biometric feature vector. The retrieval and the generation are driven by the so-called *helper data*, which the user provides during both operational phases. (2) Cancelable biometrics apply a transformation that can be invertible or non-invertible. The key used to perform the given transformation is protected via a second factor that can be, e.g., a password or a second biometric trait.

The primary disadvantage of BTP schemes is the reduction in recognition performance, which hinders their adoption [166]. Furthermore, template protection schemes often rely on binary biometric templates [194, 200]. However, face templates extracted by deep networks provide continuous sources. Therefore, recent research proposes new methods to extract keys from continuous sources such that the performance loss due to binarization is minimal [20].

Privacy leakages BTP schemes are effective against *data breaches*. A stolen template can be revoked, reducing the risk of theft and loss. However, other system blocks still leak sensitive data about the user [188]. The attacker might harness the output to learn about the system and the user. The amount of information varies based on the output received from the system: it can be a binary response (accepted/rejected) or a score (e.g. model output). Hill-Climbing attacks [179, 178] uncover biometric data even from protected data, e.g., the reconstruction of encrypted face images from query responses [5].

Adversarial ML: Poisoning Poisoning attacks aim to inject malicious samples within a training set or enrollment data. The adversary targets the self-update strategy of the ML model, such that every re-training step moves the function learned by the model in the direction specified by the adversary. This contamination leads to two threats: a denial of service, which disrupt the service, and a careful modification of the model function that selectively grants or denies access to a target user.

In the context of biometric authentication, this attack targets self-adapting templates during the enrollment phase. The adversary aims to find the best modification to be authenticated as a rightful user (untargeted attack) or deny authentication to a pre-defined victim (targeted attack). In the target poisoning case, Biggio et al. [29] have analyzed the resilience of a biometric template under white box access to the model and no further protection layer. Garofalo et al. [89] have investigated the poisoning setting

for a OneClass-SVM model authenticating a single user. This last attack relies on strong assumptions about the adversary's capabilities. The adversarial sample needs to comply with the following three features: (1) it can bypass the PAD layer, (2) it is accepted as a sample for the victim user, (3) it can trigger a template update (i.e. accepted with high certainty) and (4) acts within a realistic budget in terms of number of queries. The aforementioned assumptions are relaxed by Lovisotto et al. [138], who perform an attack on biometric systems during the query phase, leveraging template updating.

2.5 The brain: extracting and comparing the identifying features

Authentication/identification tasks compare previously enrolled and biometric templates originating from the new query sample. This section gives more details on the most prevalent comparison functions and the attacks against them.

2.5.1 Feature extraction

The feature extraction phase aims to find a low-dimensional representation of the input data to minimize inter-person similarity and maximize intra-person similarity. Finding a discriminate feature set is paramount for recognition performance. Before the introduction of deep learning, feature extraction was performed manually by statistically describing the signal or extracting discriminating features in different domains, e.g., time-domain, frequency-domain, ... Deep learning improved upon this situation by automatically extracting discriminate representations of highly complex signals. Therefore, in most biometric modalities, deep learning is used to find a discriminate representation of the biometric data; a noteworthy counter-example is fingerprints, where most systems still rely on minutiae for matching [183]. Even though minutiae are handcrafted features, due to poor image quality, skin conditions, and other noise sources, minutiae benefit from deep learning [212, 170, 231, 61].

Handcrafted features Robust feature extraction depends heavily on a good preprocessing phase. For example, IMU-based gait authentication highly depends on robust step segmentation [67, 88, 205, 147]. A user's step characteristics are represented by a length-normalized time sequence that captures one step. Matching is then performed by dynamic time warping of the average of the enrollment samples and the query sample.

However, step segmentation is error-prone, therefore later approaches no longer segment steps and extract statistical features derived from sliding or tumbling windows

over the accelerometer and gyroscope traces [258, 139]. Statistical features are powerful in describing the general characteristics of a signal but often fail to capture spatial-temporal relations adequately. Hidden Markov Models can capture these temporal relations by relying on the cyclic nature of gait patterns [173, 240]. HMMs work well in lab conditions. However, they lack the expressiveness to function well in less constrained environments.

Automated feature extraction. CNNs and RNNs successfully capture complex data's visual, spatial, and temporal relations. IMU-based gait authentication also benefits from these techniques. CNNs have been applied successfully on segmented steps [86], gait dynamic images [271] and variable length sequences [241]. Alternatively, RNNs have been adopted on segmented [77] steps and windowed signals [263]. Even when deep learning techniques are applied, signals are pre-processed, and different representations of the signals are fed to the system, e.g., gait dynamic images [272] and the extraction of roll and pitch [241] to accommodate for changes in sensor orientation, or steps segmentation is performed to align the phase of the signal.

2.5.2 Distance functions

There are a few different options for the comparison function. When the biometric templates are binary [18], the hamming distance is used, while for continuous templates, the most popular choice is the cosine distance [252]. The choice of distance function is depends on the loss function. For example, in facial authentication systems, the current state-of-the-art feature extractors are variants of the well-known cross-entropy loss, which inherently optimizes angular similarity [66]. On the other end, some feature extractors for facial authentication systems are trained on losses that emit features separated in L_2 distance [213], which leads to the use of Euclidean distance functions to compare biometric templates.

These distance functions are often used in open-ended systems, where not all classes (users) are known during training. Furthermore, they allow for both identification and verification. For verification, the enrollment template of the claimed identity is recovered from the database and compared with the query template. The system accepts the user if their distance is below a pre-defined threshold. The process is similar for identification, but now the distance with all stored identities is calculated. The identity with the lowest distance is returned if it falls below a specified centroid. When none of the enrolled users meet this criterion, the system returns that the user is unknown.

Regardless of distance function, different schemes exist to fuse multiple enrollment samples. Since using a single enrollment sample does not represent an individual

well, many are collected. Thus, instead of comparing the query template to a single enrollment sample, it is compared against a weighted combination called the centroid. The weighting scheme can be simple, where each feature vector is weighted equally. Alternatively, the most recent feature vectors can be given a higher weight. These are the most prevalent options, but many more exist [27].

2.5.3 Learned functions

Learned comparison functions provide an alternative for distance functions. For example, SVMs are a common decision function in biometric authentication. Since authentication cannot be modeled as a binary classification task (it is hard to model the negative class since all other users belong to it, also those that still need to be born), one-class SVMs are used [86, 89]. One-class SVMs are trained on biometrics from the legitimate user only and do not require examples of the negative class.

Another way to perform open set recognition is by calculating the loglikelihood ratio of the probability of the biometric sample under a Universal Background Model (UBM) and a personally adapted model [21, 240]. The main idea of this approach is to learn the common patterns in biometric data by using many people's data. The obtained model is known as the UBM. The UBM is adapted to a personal model utilizing only a few data samples of a single individual. During authentication, the probability of the query sample given the personal model and the general model (UBM) is computed and compared. This approach was later refined in speaker verification by extracting the most discriminating features into a lower dimensional representation called i-vectors (identity vectors). I-vectors are compared using linear discriminant analysis (LDA) [64].

2.5.4 Attacks and defenses

Adversarial perturbations

Often, a distance function compares the enrollment and query sample. In this case, to impact the matching result, the adversary has to modify the biometric templates being compared, which it can achieve by fooling the feature extractor. A feature extractor can be fooled by the use of adversarial examples [229]. Therefore, an adversary can leverage adversarial examples to change their perceived identity, i.e., execute an evasion attacks.

Evasion attacks aim to change the model's perception of the adversary's identity. A few variations exist under the umbrella of evasion attacks, such as flipping the decision in authentication and changing the perceived identity in identification [30, 220]. The former requires a targeted attack [128, 266] because the adversary wants to be perceived

as their victim. In contrast, the latter requires an untargeted attack because the adversary does not want to be recognized as themself.

Defenses against adversarial input focus on improving the robustness of the machine learning classifier during training [99, 254, 208] or on detecting adversarial input [54]. The two main strategies to improve model robustness are adversarial training or adapting the loss function to obtain better separated and more condensed classes in feature space. Detection systems look for adversarial noise, which is typically constrained to be physically realizable.

We point out that adversarial examples require knowledge of the victim's biometric and the ability to influence the system's input. An adversary capable of digitally modifying the system's input will not use an adversarial example, as this adversary can inject the victim's biometric data directly, which is much simpler. Therefore, the classic adversarial examples of small perturbation on top of a digital image are improbable attack vectors in the context of biometric authentication. However, a weaker adversary can typically modify their environment. Therefore, physically realizable adversarial examples [220] added to the scene are realistic attack vectors.

Backdoor attacks

Another way for the adversary to thwart the output of the decision is to install a backdoor in the biometric feature extractor or decision model [132, 55]. A backdoor is a malicious manipulation of the machine learning model that is activated by adding a specific trigger to the input. Upon receiving input with the trigger the machine learning model will provide an output chosen by the adversary who installed the backdoor rather than the output corresponding to the rest of the input. An adversary can install this trigger by influencing the training phase as a malicious infrastructure provider, by releasing an open source dataset with poisoned samples, or by releasing an open source model. The biometric backdoor remains active even when the open source model is retrained [264]. Again one of the primary challenges is to generate triggers that are physically realizable and can be added to the scene, for example glasses, hats or jewelry worn by the adversary [53, 255]. There is a multitude of defenses defined that aim to sanitize the data, and remove installed backdoors. However, it seems that, just like with adversarial examples, defenses and attacks are impractical requiring relevant real world use cases to go beyond purely academic research [55].

2.6 Conclusion

Biometric authentication could very well emerge as the future for authentication systems. They are more user-friendly and offer unique security properties. Machine learning significantly improves recognition performance by constructing complex comparison functions or providing more discriminate presentations of highly complex biometric data. Furthermore, machine learning aids in solving some of the security problems, such as spoofing and replay attacks. However, machine learning also enlarges the attack surface introducing new attack vectors that can be exploited.

To study the impact of machine learning on biometric authentication, we introduced the different building blocks of such a system. We then showed how these biometric building blocks map onto a machine learning pipeline. Within this comprehensive model, we discussed the security and privacy challenges of the biometric authentication system. In doing so, we provide the reader with a comprehensive model to further study the application of machine learning for biometric system security. We end this chapter by summarizing the learning problem introduced by biometric authentication, why deep learning provides benefits and introduces new threats, and what (research) challenges remain.

2.6.1 Biometric authentication as an open set problem

Biometric authentication, i.e., the verification of a claimed identity, is within the security community sometimes modeled as a closed-world binary classification problem [245, 226, 268]. However, this has several disadvantages: (1) from a practical perspective, it is hard to train your classifier as the data set is highly unbalanced due to short enrollment phases. Thus, there is a limited amount of positive data – you –, while there is lots of data necessary to represent the negative class – not you. (2) accurately representing the negative class is nearly impossible. The classification of a previously unseen person will depend heavily on the population used to represent the negative class. If that specific user has traits more similar to the positive class than any of the negative classes, that person will be authenticated. (3) Zhao et al. [269] proved that closed-world binary classifiers do not sufficiently constrain their decision boundaries, which leads to poor performance against random inputs.

Biometric authentication is an open-world problem where a limited amount of enrollment data is available to learn the representation of a new individual. Furthermore, an open-world assumption is a more realistic problem representation, as only part of the population is known at training time. The most used open-world techniques are dynamic time warping, which measures the similarity of two temporal sequences; one-class support vector machines; universal background models and few-shot learning with siamese networks or other deep learning-based feature extractors.

2.6.2 Threats linked to new factors and deep learning

Deep learning has paved the way for the use of novel behavioral and biometrics for authentication, like finger-vein [129], ECG [121], hand geometry [131], typing [4, 42], or eye movement [74, 140]. However, these techniques enlarge the attack surface. The use of deep learning in critical infrastructure sparked a whole new domain of research, i.e., adversarial examples. Additionally, human bias is represented in machine learning algorithms, which leads to unfair algorithms. Due to this human bias, security levels of authentication systems are different depending on skin color, ethnicity, and gender [40, 174]. Lastly, due to deep learning, function creep has become a more significant threat because more complex functions can be learned, thus also functions that map biometric feature vectors to soft biometrics or even the raw biometric data [154, 93]. In general, due to deep learning, personal data has become more sensitive, as the amount of leakage is unclear. The other way round, as deep learning enables us to estimate biometric features from other biometric traits, sharing any personal data poses a security risk for biometric authentication [73].

2.6.3 Future directions

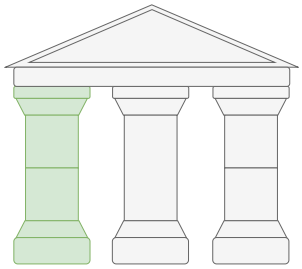
Biometric authentication systems still have a long way to go before becoming satisfactory authentication methodologies that can be applied in a broad set of use cases. More research is needed to reduce user interaction further, limit the attack surface and ensure that biometric authentication does not violate user privacy. Ross et al. [206] described the future directions for biometrics in general. We focus here on biometric authentication specifically.

To make biometric authentication systems entirely frictionless, they should be able to run unobtrusively in the background, constantly monitoring the user context and natural human-computer interactions [41]. Biometrics must help shift the current single-shot authentication paradigm to a paradigm with continuous authentication. As a biometric modality that fits all use cases does not exist, the future biometric authentication system will probably leverage many different biometric modalities and sensors [84, 97]. Such a system should be adaptive, deal with missing modalities and temporal discrepancies, and decide when to perform step-up authentication at the penalty of user interaction [246].

Throughout this chapter, we have shown the elaborate attack surface of biometric authentication systems. The vast literature on presentation attack detection still needs to describe a flexible, bullet-proof solution. We are in need of anti-spoofing methods that generalize over different attack types and datasets [180]. Defenses that are successful against replay attacks and spoofing rely on challenge-response methodologies [131, 230] or correlate different biometric signals [227]. However, these techniques add

more friction or are not universally applicable. Lastly, using deep learning introduces new threats such as deep fooling, back door attacks – where the model is learned to react differently for pre-defined input patterns. New research can focus on showing the extent of the threats or developing novel defense strategies, but should take into account that it is not just the model that has to be compromised or defended, but the entire system, of which the model is only a part [16].

Biometric authentication keeps struggling with privacy threats [179]. More research is necessary to understand better the extent to which private attributes are leaked and how to limit this leakage without sacrificing recognition performance. Alternatively, future research could focus on the security threat to biometric authentication due to the sharing of (soft) biometric traits in other contexts, for example, medical data, fitness trackers, social media, online DNA tests, etc.



Chapter 3

Gait template protection using HMM-UBM

> "Well sir, I have a silly walk and I'd like to obtain a government grant to help me develop it."
>
> *— Mr. Pudey - Ministry of Silly Walks*

Preamble

This chapter presents a Hidden Markov Model-Universal Background Model gait authentication system. The privacy and security of the gait templates is protected by a fuzzy commitment scheme. We show that with limited enrollment data the HMM-UBM system achieves a very competitive equal error rate of $\approx 1\%$ using one sensor. The proposed template protection scheme benefits from eigenfeatures coming from multiple Universal Background Model systems fused with a novel technique that minimizes the bit error rate for genuine attempts. This allows the protected system to achieve a false rejection rate below 5% with an effective key length of 64 bits.

The content of this chapter is an extended version of our published paper on IMU-based gait authentication with template protection [240]. It represents our first mature research result on behavioral authentication that respects user privacy. This research result has been a pivotal milestone in building and expanding the research lab's expertise in biometric authentication. Among others, the research reports on the feature set that

was crucial in unlocking good recognition performance, which gave us a competitive advantage in the GAG 2019 competition [169, 241] and was leveraged in several research projects. This work contributes to the core theme of behavioral authentication.

Publication Data: T. Van hamme, E. Argones Rúa, D. Preuveneers, and W. Joosen. "Gait Template Protection Using HMM-UBM". in: *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2018, pp. 1–8

3.1 Introduction

Gait, the steady state movement pattern during locomotion, has been considered as behaviometric in the last decade [11]. Gait can be observed with Inertial Measurement Unit (IMU) sensors incorporated in personal and wearable devices. Furthermore, gait has proven to be challenging to imitate [87]. Although gait authentication achieves remarkable performance, there are only a few examples of attempts to reliably and securely protect gait templates in the literature [103]. Template protection [39] is one of the most sensible approaches for integrating biometrics into secure and privacy-friendly authentication solutions. In the case of biometric cryptosystems such as fuzzy extractors [72], this also enables the integration of biometrics into cryptographic protocols.

In this work, we aim to design a high-performance gait template protection scheme based on a biometric cryptosystem. We adopt a technique called Hidden Markov Model Universal Background Models (HMM-UBM), a well-known and effective method for biometric authentication across different biometric data types [19, 81, 204]. This technique builds two models: a universal one (UBM) which represents how the general population walks, and a personal one which captures the individual's unique way of walking. Authentication is then performed by assessing the probability of the newly acquired query walking sequence under the two models. We are applying this technique specifically to IMU-based gait authentication. This technique adapts UBM parameters to the enrollment data, providing accurate results even when the enrollment data is scarce. This system will be used as a baseline and starting point for the protected system, based on fusing UBM eigenfeatures [124] from multiple UBMs in a fuzzy commitment scheme [119]. In addition, we use a novel continuous feature fusion technique to obtain binary templates from their continuous counterparts.

We evaluate the unprotected and protected HMM-UBM systems on the largest publicly available gait dataset OU-ISIR [168]. It contains gait data from IMU sensors located on three body positions around the waist from 495 subjects performing two level walk sequences, a sequence for an up-slope walk and a sequence for a down-slope walk. OU-ISIR enables authentication in an open-world scenario with scarce training and authentication data.

Related work	Original database sensor information	No of subject	EER (%)	OU-ISIR EER (%)
Mantyjarvi et al., 2005 [147]	Accelerometer at back waist	36	7	NA
Rong et al., 2007 [205, 168]	Accelerometer at back waist	21	5.6	14.3
Derawi et al., 2010 [67, 168]	Accelerometer at left hip	60	5.7	14.3
Gafurov et al., 2010 [88, 168]	Accelerometer at ankle	30	1.6-6.1	15.8
Trung et al, 2011 [237, 168]	Accelerometer and gyroscope in backpack	32	6.0	20.2
Nickel et al., 2011 [173]	Accelerometer at right hip	48	10.3	NA
Lu et al., 2014 [139]	Accelerometer and gyroscope in pocket	47 + 12	14.0	NA
Zhong and Deng, 2014 [272]	NA	NA	NA	5.6
San-Segundo et al., 2016 [209]	Accelerometer and gyroscope at left hip	30	19.1	NA
San-Segundo et al., 2017 [210]	Accelerometer and gyroscope at left hip	30	6.1	NA
Zhao and Zhou, 2017 [271]	Accelerometer and gyroscope in pocket	20	Not reported	NA
Yao et al., 2017 [263]	Accelerometer and gyroscope in all kinds of places	9	Not reported	NA
Gadaleta and Rossi, 2018 [86]	Accelerometer and gyroscope in right pocket	35 + 15	0.15	NA

Table 3.1: Comparing the EER of related work on accelerometer based gait authentication schemes.

In the open-world scenario, the proposed HMM-UBM approach using simple features derived from the provided IMU data obtains authentication performances close to 1% Equal Error Rate. Most important, the protected system using UBM eigenfeatures from several HMM-UBM systems also obtains a remarkable performance of $\approx\ 0\%$ False Acceptance Rate and False Rejection Rate below 5% for an effective key length of 64 bits, which is the best-achieved compromise regarding security and usability for this modality.

3.2 Related work

Researchers had been studying accelerometer-based gait recognition for over a decade when Ngo et al. [168] published their OU-ISIR benchmark dataset. Prior to this dataset new methods for accelerometer based gait authentication were evaluated on relatively few individuals, making the results hard to compare. The OU-ISIR dataset contains around 500 unique individuals which is an order of magnitude more than datasets used by related work. Ngo et al. [168] used their benchmark to re-evaluate four gait authentication methodologies. Table 3.1 gives an overview of the achieved EERs. For completeness, we added earlier and more recent work. The remainder of this section elaborates on the more recent works elicited in Table 3.1.

A common theme among the four works re-evaluated by Ngo et al. [168] is the extraction of gait cycles to build a template gallery. However, this is the limiting factor, as robustly extracting gait cycles is hard. The best work achieved an equal error rate (EER) of 14.3%.

Lu et al. [139] steered away from extracting gait cycles using a GMM-UBM verification method. They extract standard statistical features in both time and frequency domains

from a 5-second window. They achieve an EER of 14.0%. San-Segundo et al. [209] use the same verification framework but drop the MAP adaptation approach in favor of i-vector adaptation. Features are based on perceptual linear prediction (PLP) coefficients. An EER of 6.1% is reached.

Zhong and Deng [272] exploit i-vector-adapted GMM-UBM models as well. They propose another feature extraction method called gait dynamic images (GDI). It is a visual rotation-invariant representation of gait. When evaluated on the benchmark OU-ISIR dataset, an EER of 5.6% is attained. Zhao and Zhou [271] use GDIs to train a CNN. They train a binary classification model per user for authentication purposes. Unfortunately, Zhao and Zhou do not report on achieved EERs. Furthermore, this approach requires 15 minutes of enrollment data. On top of that, this work focuses on a small closed-world scenario, where it is impossible to ensure the feasibility of the proposed approach for a more general large population open-world scenario. Especially since the advantage of UBMs w.r.t. classical ML and DL techniques is that they, by definition, make the 'open world assumption'. Gadaleta and Rossi [86] address this by first training a CNN to obtain a gait representation. The representation serves as input for a one-class SVM, but the work returns to extracting individual gait cycles. A 0.15% EER is achieved by querying the OSVM multiple times with different gait cycles. The work has only been evaluated with 15 users. Long gait sequences are necessary for training the representation network, which is unavailable in the OU-ISIR database. Besides, their results suggest that five gait cycles are enough to provide accurate authentication in 80% of the cases, which is approximately the length of each gait sequence in the OU-ISIR database.

Regarding template protected gait authentication, Hoang et al. [103] adopt a fuzzy commitment scheme. They built a template gallery by extracting gait cycles. The work was evaluated with only 38 subjects. The system requires eight templates consisting of four gait cycles per subject. Which is around four times more training data than our system needs. They acquire a 139-bit effective key length at a less favorable 16.18% FRR and a 0% FAR. Their authentication system without error correcting codes performs worse than ours, with a 3.5% EER.

3.3 Feature extraction

In a gait system based on accelerometer and gyroscope sensory data, both linear acceleration including gravity $\mathbf{a}^g(n) = \left(a_x^g(n), a_y^g(n), a_z^g(n)\right)$ and gyroscopic velocity $\boldsymbol{\omega}(n) = (\omega_x(n), \omega_y(n), \omega_z(n))$ the sensors provide signals at a given sample rate. In our case, sensors provide samples at 100 Hz rate. From these signals, we derived the following additional ones:

Linear acceleration This is just the linear acceleration after the subtraction of the gravity, i.e. $\mathbf{a}(n) = (a_x(n), a_y(n), a_z(n)) = \left(a_x^g(n), a_y^g(n) - g, a_z^g(n)\right)$. This subtraction can be done easily due to the controlled conditions of the database acquisition, where sensors were well aligned.

Linear velocity $\mathbf{v}(n) = (v_x(n), v_y(n), v_z(n))$, where each component $v_d(n)$ is computed from its corresponding linear acceleration component $a_d(n)$ as a high-pass version of the accumulated acceleration, i.e. $v_d(n) = h(n) * \sum_{k=0}^{n} a_d(n)$. The linear filter $h(n)$ is characterized by its z-transform $H(z) = \frac{0.81 - 1.62z^{-1} + 0.81z^{-2}}{1 - 1.8z^{-1} + 0.81z^{-2}}$, and its purpose is to remove slow drifts and the integration offset. This filter is specific to the sampling rate of our system, and its cut-off frequency is 1.95 Hz.

Angular acceleration $\boldsymbol{\alpha}(n) = (\alpha_x(n), \alpha_y(n), \alpha_z(n))$, where each component $\alpha_d(n)$ is computed from its corresponding angular velocity component $\omega_d(n)$ as its first order differences, i.e. $\alpha_d(n) = \omega_d(n) - \omega_d(n-1)$.

Magnitudes The magnitudes of the angular and linear dynamics were also derived:

- Linear acceleration magnitude $a(n) = |\mathbf{a}(n)|$.
- Linear velocity magnitude $v(n) = |\mathbf{v}(n)|$.
- Angular velocity magnitude $\omega(n) = |\boldsymbol{\omega}(n)|$.
- Angular acceleration magnitude $\alpha(n) = |\boldsymbol{\alpha}(n)|$.

Roll and Pitch from Accelerometer The roll and pitch are estimated from the accelerometer signals without removing the gravity, as $r_a(n) = \operatorname{atan2}\left[a_z^g(n), a_y^g(n)\right]$ and $p_a(n) = \operatorname{atan2}\left[a_x^g(n), a_y^g(n)\right]$ respectively. Its associated roll and pitch velocities and accelerations $\dot{r_a}(n)$, $\dot{p_a}(n)$, $\ddot{r_a}(n)$, and $\ddot{p_a}(n)$ are obtained as the first and second order differences of $r_a(n)$ and $p_a(n)$ respectively.

Roll and Pitch from Accelerometer and Gyroscope If we define the pitch and roll estimated from the gyroscope signals as $p_g(n) = \sum_{k=0}^{n} \omega_x(n)$ and $r_g(n) = \sum_{k=0}^{n} \omega_z(n)$, we derive a joint estimate of the pitch and the roll signals as follows:

$$p(n) = 0.98\left[p(n-1) + p_g(n)\right] + 0.02p_a(n)$$

$$r(n) = 0.98\left[r(n-1) + r_g(n)\right] + 0.02r_a(n)$$

Vertical and Horizontal components The vertical and horizontal acceleration components are derived from the accelerometer data as $a_v(n) = a_y(n)$ and $a_h(n) = \left|a_x^2(n) + a_z^2(n)\right|$ respectively. Their corresponding velocities $v_v(n)$, $v_h(n)$ are computed by the same procedure as the linear velocities from the linear accelerations, and their corresponding jerks $j_v(n)$, $j_h(n)$ are computed by the same procedure as the angular accelerations from the angular velocities.

We define the set of features directly derived from the linear dynamics provided by the sensors as:

$$\mathrm{LD} = \{v_x, v_y, v_z, v, a_x, a_y, a_z, a\} \ , \tag{3.1}$$

and the gyroscopic dynamics provided by the sensors as:

$$\mathrm{GD} = \{\omega_x, \omega_y, \omega_z, \omega, \alpha_x, \alpha_y, \alpha_z, \alpha\} \ . \tag{3.2}$$

We also define the set of features involving roll, pitch, vertical and horizontal components as:

$$\mathrm{RPVH} = \{r_a, \dot{r_a}, \ddot{r_a}, p_a, \dot{p_a}, \ddot{p_a}, r, \dot{r}, \ddot{r}, p, \dot{p}, \ddot{p}, v_v, a_v, j_v, v_h, a_h, j_h\} \ . \tag{3.3}$$

3.4 HMM-UBM gait recognition

In this section, we first explain the unprotected system for gait authentication based on UBM-HMMs. We continue to use this system as a starting point for designing a protected system based on the fuzzy commitment scheme.

3.4.1 Unprotected system

In the proposed gait recognition system, we use HMMs as generative models of the gait sequences. Our approach is based on the work by Reynolds et al. [204] for speaker verification. Gait acquisitions are defined as variable-length sequences of feature vectors, which can be denoted in matrix notation as $G = [\mathbf{g}_1 \mid \ldots \mid \mathbf{g}_{T_G}]$, where T_G denotes the length of the sequence G. An HMM is fully specified by the set of parameters $\mathcal{H} = \{\boldsymbol{\pi}, A, B\}$, where $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_S]$ is the initial state probability vector, the $A = [a_{ij}]$ is the state transition probabilities matrix, $B = \{b_1(\mathbf{g}), \ldots, b_S(\mathbf{g})\}$ is the set of output probability density functions associated to each state, and S is the number of states. Each of the output distributions in B is a Gaussian mixture model (GMM), defined by the set of parameters $\mathcal{B}_s = \{\Lambda_s, \mathcal{M}_s, \mathcal{S}_s\}$, where $\Lambda_s = \left\{\lambda_s^1, \ldots, \lambda_s^M\right\}$ is the mixture weights set, $\mathcal{M}_s = \left\{\boldsymbol{\mu}_s^1 \ldots, \boldsymbol{\mu}_s^M\right\}$ is the mean vectors set, $\mathcal{S} = \left\{\Sigma_s^1, \ldots, \Sigma_s^M\right\}$ is the covariance matrices set, and M is the number of mixtures. The output probability density functions are defined as $b_s(\mathbf{g}) = \sum_{m=1}^{M} \lambda_s^m \mathcal{N}(\mathbf{g}; \boldsymbol{\mu}_s^m, \Sigma_s^m)$, where $\mathcal{N}(\mathbf{g}; \boldsymbol{\mu}, \Sigma)$ denotes the multivariate Gaussian probability density function with mean $\boldsymbol{\mu}$ and covariance matrix Σ. In the proposed system, we use diagonal covariance matrices.

In an HMM-UBM system, instead of training a user-specific HMM for each user, a user-independent UBM $\mathcal{H}^{\mathrm{UBM}}$ is trained using gait sequences from a large population

of subjects. Samples could be initially allocated to different states using any sensible approach. For simplicity and generality, we used a standard clustering algorithm for this purpose [135]. Then, the UBM parameters are estimated using the Baum-Welch algorithm, as described in [31]. Given the nature of gait, the resulting UBM is ergodic. User models are then generated by Maximum a Posteriori (MAP) adaptation of the means of each state output probability density function of the UBM to a set of enrollment gait sequences from the user $\mathcal{E}^{\mathrm{U}} = \left\{G^{\mathrm{U},1}, \ldots, G^{\mathrm{U},E}\right\}$. If we define the probability of the UBM state s at time t in the enrollment sequence $G^{\mathrm{U},e}$ as $\gamma_t^{\mathrm{U}}(s,e)$, which can be computed as indicated in [135], then the soft count of samples associated to the Gaussian mixture m in state s of the UBM can be defined as:

$$c_{s,m}^{\mathrm{U}} = \sum_{e=1}^{E} \sum_{t=1}^{T_{G_e^{\mathrm{U}}}} \frac{\gamma_t^{\mathrm{U}}(s,e)\, \lambda_s^m \mathcal{N}\left(\mathbf{g}_t^{\mathrm{U},e}; \boldsymbol{\mu}_s^{\mathrm{UBM},m}, \Sigma_s^{\mathrm{UBM},m}\right)}{\sum_{k=1}^{M} \lambda_s^k \mathcal{N}\left(\mathbf{g}_t^{\mathrm{U},e}; \boldsymbol{\mu}_s^{\mathrm{UBM},k}, \Sigma_s^{\mathrm{UBM},k}\right)}.$$

Then, the MAP estimate of the mean of mixture m of the output distribution of state s of the user HMM $\mathcal{H}^{\mathrm{U}}$ can be computed as:

$$\boldsymbol{\mu}_s^{\mathrm{U},m} = \boldsymbol{\mu}_s^{\mathrm{ML,U},m} \frac{c_{s,m}^{\mathrm{U}}}{c_{s,m}^{\mathrm{U}} + \tau} + \boldsymbol{\mu}_s^{\mathrm{UBM},m} \frac{\tau}{c_{s,m}^{\mathrm{U}} + \tau},$$

where $\boldsymbol{\mu}_s^{\mathrm{ML,U},m}$ is the Maximum Likelihood estimate of the mean of mixture m in state s, and τ is a relevance factor. In this work, we set $\tau = 14$, and only one sequence will be available for enrollment purposes, i.e. $E = 1$.

Gait verification of a test gait sequence G^V for user U is then formulated as the following hypothesis test:

H_0: G^V has been produced by U
H_1: G^V has been produced by any other user.

The null hypothesis H_0 likelihood is represented by $\mathrm{p}\left(\mathcal{H}^{\mathrm{U}} \mid G^V\right)$, whereas the alternative hypothesis H_1 likelihood is represented by $\mathrm{p}\left(\mathcal{H}^{\mathrm{UBM}} \mid G^V\right)$. The verification is done by comparing the hypothesis loglikelihood ratio $s_{G^V}^{\mathrm{U}}$ to a threshold, i.e.:

$$\mathrm{H}_0 \text{ is accepted} \iff s_{G^V}^{\mathrm{U}} \geq \Theta, \tag{3.4}$$

where $s_{G^V}^{\mathrm{U}} = \log\left[\mathrm{p}\left(G^V \mid \mathcal{H}^{\mathrm{U}}\right)\right] - \log\left[\mathrm{p}\left(G^V \mid \mathcal{H}^{\mathrm{UBM}}\right)\right]$. The involved loglikelihoods are approximated by the loglikelihood of the gait sequence given the most likely state sequence, i.e. $\log\left[\mathrm{p}\left(G^V \mid \mathcal{H}\right)\right] \approx \max_{\mathbf{s}}\left\{\log\left[\mathrm{p}\left(G^V \mid \mathcal{H}, \mathbf{s}\right)\right]\right\}$. This is calculated using the Viterbi algorithm.

3.4.2 Protected system

Using a UBM-HMM unprotected system as the underlying system of a fuzzy commitment scheme is possible. Below we explain how to obtain low-dimensional representations of the users in a UBM-HMM system and how to use these representations under the restrictions imposed by a fuzzy commitment scheme for template protection.

UBM eigenfeatures The protected system we propose for gait authentication uses HMM-UBM eigenfeatures. These eigenfeatures are a low-dimensional representation of the adapted models. The differences between the means of the adapted models and the UBM for a set of users representing the whole population are reduced in dimensionality using Probabilistic Principal Components Analysis. The obtained representation has several advantages for its use in a biometric fuzzy commitment cryptosystem: preserving most of the discriminability; and having a fixed length. The procedure for obtaining these features was first introduced in [124] for GMMs, and it is explained in detail in [21] for the general case of HMMs.

Feature protection scheme The feature protection scheme is illustrated in Figure 3.1, where $\mathsf{ECC}_{\mathsf{encode}}$ and $\mathsf{ECC}_{\mathsf{decode}}$ stand for the encoding and decoding procedures of an error correcting code (ECC), and $\mathrm{h}()$ is a cryptographically secure hash function. The binary representation $\mathbf{b}$ is not derived from the biometrics but randomly generated. The authentication succeeds only when the Hamming distance between the binary representation obtained during authentication $\hat{\mathbf{b}}$ and the one generated during the enrollment is equal or less than the error correcting capability t of the employed ECC. This scheme is derived from a fuzzy commitment scheme [119], but adding a quantization metadata $\mathcal{Q}$ computed during enrollment. It relates the enrollment biometrics and the binary representation, and it is later used during verification to assist in quantizing the probe template. Although the quantization metadata minimizes the Bit Error Rate (BER) for genuine attempts, it will not provide any information about the binary representation $\mathbf{b}$. Therefore, it does not pose a risk for security, i.e., knowing the metadata does not provide any advantage to a security adversary trying to guess the secret string $\mathbf{m}$. The following section thoroughly explains the metadata's composition and computation.

Quantization metadata Our protected system uses eigenfeatures from different UBMs. Many cannot provide a BER usable within a fuzzy commitment scheme after quantization. The BER must be limited to the one the employed error correcting code (ECC) can handle. Therefore, the proposed approach fuses different eigenfeatures to obtain each bit.

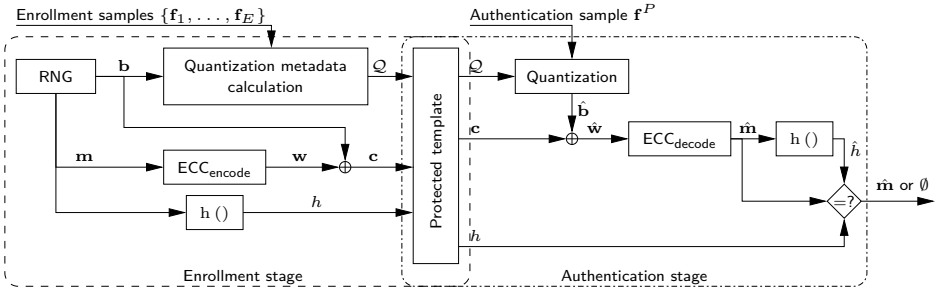


Figure 3.1: Template protection approach incorporating quantization metadata to the Fuzzy Commitment

If we define n as the number of bits we need to extract–the length of the codeword $\mathbf{w}$ in Figure 3.1–from our set of available features $\mathcal{F} = \{f_1, \ldots, f_F\}$, the algorithm that we propose computes a quantization metadata $\mathcal{Q}$ constituted by n disjoint sets of features $\mathcal{A}_i$, and a vector of weights $\mathbf{w} = [w_1, \ldots, w_F]^t$, i.e., $\mathcal{Q} = (\{\mathcal{A}_1, \ldots, \mathcal{A}_n\}, \mathbf{w})$. The randomly chosen binary representation is denoted as $\mathbf{b} = [b_1, \ldots, b_n]$. During authentication, each bit b_i is estimated from the probe vector $\mathbf{f}^P = \left[f_1^P, \ldots, f_F^P\right]^t$ used for authentication as follows:

$$b_i = \frac{\operatorname{sign}\left(\sum_{f_j \in \mathcal{A}_i} w_j f_j^P\right) + 1}{2} \tag{3.5}$$

The weights involved in this computation and the disjoint sets defining which features contribute to each bit are calculated during the enrollment from a set of E enrollment vectors of features and the randomly chosen binary representation $\mathbf{b}$.

Let us assume that the features in $\mathcal{F}$ are uncorrelated Gaussians $f_i \sim \mathcal{N}(\mu_i, \sigma_i)$. Let us also assume that σ_i is known, and we have an estimate $\hat{\mu}_i$ from E enrollment samples, and the feature mean is distributed population-wise as $\mu_i \sim \mathcal{N}(0, 1)$. If the set of features contributing to bit i is defined as $\mathcal{A}_i = \left\{f_{i_1}, \ldots, f_{i_{|\mathcal{A}_i|}}\right\}$ then the optimal BER for the genuine attempts is obtained when the weights are computed as:

$$w_{i_1} = \quad (2b_i - 1) \operatorname{sign} \hat{\rho}_{i_1} \tag{3.6}$$

$$w_{i_k} = \quad \frac{\hat{\rho}_{i_k} \operatorname{sign}\left[(2b_i - 1)\, \hat{\rho}_{i_k}\right] \sqrt{\sum_{j=1}^{k-1} w_{i_j}^2 \sigma_{i_j}^2}}{\sigma_{i_k} \sqrt{\sum_{j=1}^{k-1} \hat{\rho}_{i_j}^2}}, \; \forall k > 1 \tag{3.7}$$

In this case, the BER associated with the bit b_i for genuine attempts is as follows:

$$\mathrm{BER}_i\left(\hat{\rho}_i, E\right) = \sqrt{\frac{E}{2\pi}} \int_{\rho_i=-\infty}^{+\infty} \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\hat{\rho}_i}{\sqrt{2}}\right) e^{-\frac{E(\rho_i-\hat{\rho}_i)^2}{2}}\right] \mathrm{d}\rho_i\,, \tag{3.8}$$

where the reliability of the binary feature $\hat{\rho}_i$ is defined as:

$$\hat{\rho}_i = \sum_{j=1}^{|\mathcal{A}_i|} w_{i_j}\hat{\mu}_{i_j} \Big/ \sqrt{\sum_{j=1}^{|\mathcal{A}_i|} w_{i_j}^2 \sigma_{i_j}^2} \tag{3.9}$$

Obtaining the sets $\mathcal{A}_1, \ldots, \mathcal{A}_n$ in such a way that the average BER is minimized is a number partitioning problem [127]. We adopted a greedy approach to solving it, which provides a good approximation to the optimal solution. This approach iterates on the available features–those who have not yet been chosen–, choosing the most reliable one among them, i.e., the one with the maximum $|\hat{\rho}|$. The chosen feature is then incorporated into the least reliable set $\mathcal{A}_i$–the set with minimum $\hat{\rho}_i$. Weights are computed abiding Eqs. 3.6 and 3.7.

It is easy to check that the quantization metadata does not provide any information about $\mathbf{b}$ since the features in $\mathcal{F}$ are uncorrelated and with a median of 0.

3.5 Dataset and experimental protocol

The work is evaluated on the multisensor part of the OU-ISIR dataset [168]. This dataset consists of 495 subjects with an almost equal gender distribution. The age of test subjects range from 8 to 78 years old. Subjects wore a belt with 3 IMU sensors: left hip, right hip and back. 3D angular velocity and 3D acceleration data was collected at a sampling rate of 100 Hz. Each subject performs two walking sequences for different types of walking: on a flat surface, up and down the stairs, and up and down a slope. The flat surface walking parts have an average length of 5.24 s. Relative location and orientation of the sensors between subjects and between sequences are invariant.

In a preprocessing step very short sequences were removed, as well as subjects with one or no walk sequences for a specific sensor. After preprocessing 483 subjects remained.

The biggest disadvantage of this dataset are the constrained lab conditions. On top of that, the database owners did not provide demographic information per subject, which prevents further analysis. Yet, the OU-ISIR database is the largest inertial gait database w.r.t. the amount of subjects. It has proven challenging for gait authentication systems to deal with this diversity of subjects. Furthermore, the amount of data per subject is very limited. In this sense, OU-ISIR poses a realistic challenge. Since ground truth data is hard to obtain, and is most of the time obtained during a controlled and assumed

to be safe enrollment phase, it is reasonable to assume that a real world authentication system will be favored if enrollment is short, in other words, when only a small amount of user specific gait information is available.

In this work we only used the gait sequences from the OU-ISIR database labeled as *level walk*. Users usually have two level walk sequences. The first of them is used for enrollment, and the second one for test. Only the 483 users with two valid level walk sequences captured with the Left, Center, and Right sensors were considered. These users were divided into two experimental groups by lexicographic order of their pseudonyms, since no demographic data was available for a more sensible split. The first group is constituted by the first 242 users, labeled as $\mathrm{U}_0^1, \ldots, \mathrm{U}_{241}^1$, and the second one by the remaining 241 users, labeled as $\mathrm{U}_0^2, \ldots, \mathrm{U}_{240}^2$. Regarding impostor attempts, each user is impersonated by the 5 next users from his group, i.e., the set of users impersonating user U_i^g is $I_{\mathrm{U}_i^g} = \left\{\mathrm{U}_{(i+1)\bmod N_g}^g, \ldots, \mathrm{U}_{(i+5)\bmod N_g}^g\right\}$, with $N_1 = 242$ and $N_2 = 241$. Only the second walks are used as impostor attempts, therefore group 1 has a total 242 genuine and 1210 impostor attempts, while group 2 has a total 241 genuine and 1205 impostor attempts.

Each group's UBM is trained using all the captures from the other group. Verification performance is evaluated as the average a posteriori Equal Error Rate (EER), i.e.:

$$\mathrm{EER} = \left(\mathrm{EER}_1 + \mathrm{EER}_2\right)/2.$$

If we define the set of impostor and genuine scores from group g as $\left\{s_g^{\text{Impostor}}\right\}$ and $\left\{s_g^{\text{Impostor}}\right\}$ respectively, and the cardinality of a set as the operator $\#$, then EER_g can be computed as follows:

$$\mathrm{EER}_g = \frac{\mathrm{FAR}_g\left(\Theta_g^{\mathrm{EER}}\right) + \mathrm{FRR}_g\left(\Theta_g^{\mathrm{EER}}\right)}{2}, \tag{3.10}$$

where:

$$\begin{aligned}
\mathrm{FAR}_g\left(\Theta_g\right) &= \frac{\#\left\{s_g^{\text{Impostor}} \geq \Theta_g\right\}}{\#\left\{s_g^{\text{Impostor}}\right\}} \\
\mathrm{FRR}_g\left(\Theta_g\right) &= \frac{\#\left\{s_g^{\text{Genuine}} < \Theta_g\right\}}{\#\left\{s_g^{\text{Genuine}}\right\}} \\
\Theta_g^{\mathrm{EER}} &= \underset{\Theta_g}{\operatorname{argmin}} \left|\mathrm{FAR}_g\left(\Theta_g\right) - \mathrm{FRR}_g\left(\Theta_g\right)\right|
\end{aligned}$$

For the protected system, where the employed ECC fixes the working point, FAR and FRR are reported for both groups.

3.6 Experiments

This study reports on different experiments to analyze the performance of HMM-UBM systems in the OU-ISIR dataset. The first experiment studies the effect of the feature set on authentication performance for the unprotected system. The second experiment researches the impact of model complexity. The first two experiments consider the performance of the unprotected system, while the third and last experiment investigates protected systems. The proposed gait authentication system leverages feature fusion of eigenfeatures from different gait recognition models (mono-modal multimodel fusion).

Feature sets selection experiments First, we analyzed the performance of different feature sets using all the sensors. This experiment is aimed at selecting the most suited feature set for HMM-UBMs using the simple features introduced in Section 3.3. Although more experiments were performed, we show in this chapter a significant selection of them using the feature sets defined in Equations 3.1, 3.2, and 3.3, which are reflected in Table 3.2. It can be seen that the combination of linear dynamics and gyroscopic dynamics ($\mathrm{LD} \cup \mathrm{GD}$) is outperformed by the combination of roll, pitch, vertical and horizontal acceleration (RPVH). Regarding more complete combinations, the combination of all of them outperforms either $\mathrm{LD} \cup \mathrm{GD}$ and RPVH. However, the best results are obtained when the linear dynamics are removed, using the roll, pitch, horizontal, vertical and gyroscopic dynamics features ($\mathrm{RPVH} \cup \mathrm{GD}$). This is not surprising, since the LD feature set is providing information very similar to the one provided by the vertical and horizontal components already included in the features RPVH.

Feature Set	Left	Center	Right
$\mathrm{LD} \cup \mathrm{GD}$	3.56	3.67	4.06
RPVH	2.92	2.49	3.11
$\mathrm{RPVH} \cup \mathrm{GD}$	**1.87**	**1.46**	**2.40**
All features	1.89	2.14	2.69

Table 3.2: Performance in terms of EER$\%$ of 4 states, 32 Gaussian mixtures per state HMM-UBM system using different sets of features for the different sensors (Left, Center, and Right).

Number of states and complexity experiments A second experiment explores the influence of the number of states in the overall performance of the HMM-UBM approach. For this, we tested HMM topologies ranging from $S = 1$ up to $S = 16$ states, while keeping the complexity of the model constant, represented by the product

of the number of states by the number of Gaussian mixtures in each state $C = S \cdot M$. In our experiments we set $C = 128$. The results of this experiment using the Center sensor and the feature set $\mathrm{RPVH} \cup \mathrm{GD}$ are shown in Table 3.3. From the results shown in this table we can infer that the authentication performance in terms of EER remains mostly unaffected by the number of states when the complexity C of the model remains unaltered.

S	M	EER%
1	128	1.45
2	64	1.60
4	32	1.46
8	16	1.42
16	8	1.49

Table 3.3: Performance in terms of EER% using the Center sensor of HMM-UBM configurations with S states and M Gaussian mixtures per state for constant $C = S \cdot M = 128$ using the $\mathrm{RPVH} \cup \mathrm{GD}$ feature set.

We also explored the influence of the complexity C of the HMM-UBM for the 4-states case. In this experiment, the number of Gaussian mixtures per states ranges from $M = 8$ up to $M = 64$, therefore the tested complexities ranges from $C = 32$ up to $C = 256$. The results of this experiment are shown in Table 3.4. It can be seen that the lowest complexity $C = 32$ provides comparatively the worse authentication performance, although it remains remarkably competitive. However, the best performances correspond to higher model complexities, although this tendency seems to saturate for $C \geq 128$. This complexity at saturation depends on the amount of available data for UBM training, so we can infer that for this particular dataset and the experimental protocol employed in this study, the best trade-off between complexity and performance is provided by a complexity $C = 128$.

M	EER%
8	2.18
16	1.63
32	1.46
64	1.35
128	1.54

Table 3.4: Performance in terms of EER% using the Center sensor of HMM-UBM configurations with $S = 4$ states and different M Gaussian mixtures per state, using the $\mathrm{RPVH} \cup \mathrm{GD}$ feature set.

Protected system The third experiment analyzes the authentication performance and effective key length (EKL) obtained by the protected system introduced in Section 3.4.2. For this experiment, the UBM eigenfeatures are obtained from different unprotected systems trained on the center sensor using the feature set $\mathrm{RPVH} \cup \mathrm{GD}$, as shown in Table 3.5. We tested two different configurations of the protected system: *System 1*, incorporating all the UBM eigenfeatures from 16 unprotected systems, and *System 2*, incorporating all the UBM eigenfeatures from 8 unprotected systems. The total number of UBM eigenfeatures is $F = 3840$ for System 1 and $F = 1920$ for System 2.

System	S	M	D	N	System	S	M	D	N
1	1	128	240	4	2	1	128	240	2
	2	64	120	4		2	64	120	2
	4	32	60	4		4	32	60	2
	8	16	30	4		8	16	30	2

Table 3.5: Unprotected systems incorporated into the protected systems. D stands for the number of eigenfeatures computed for each state, and N is the number of unprotected systems with these parameters incorporated into the protected system.

Table 3.6 shows the performance of the protected system for different codes. The parameter k coincides with the EKL of the protected scheme. The EKL is the security parameter of the fuzzy commitment scheme. By using 8 UBM eigenfeatures sets, it is possible to reach an EKL of 36, while when using 16 UBM eigenfeatures sets, it is possible to reach up to 64 bits with a low FRR.

	BCH Code							BCH Code					
System	n	k	t	Group	FAR	FRR	System	n	k	t	Group	FAR	FRR
1	127	64	10	1	0.08	4.96	2	127	29	21	1	0.00	8.68
				2	0.08	3.32					2	0.00	10.79
		57	11	1	0.08	3.72			22	23	1	0.08	5.37
				2	0.08	2.49					2	0.08	7.47
		50	13	1	0.25	2.07			15	27	1	0.25	2.89
				2	0.25	1.66					2	0.25	3.73
	63	57	1	1	0.41	2.07		63	36	5	1	0.08	4.96
				2	0.41	3.32					2	0.17	3.32
		51	2	1	0.74	1.24			30	6	1	0.25	3.31
				2	0.66	1.66					2	0.25	2.49
		45	3	1	1.32	0.83			24	7	1	0.33	2.07
				2	1.08	1.24					2	0.58	2.07

Table 3.6: Protected systems percentual performance figures for different code parameters.

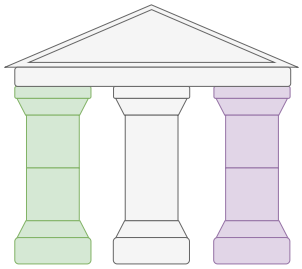
3.7 Discussion and Conclusions

In this work, we proposed HMM-UBM systems using simple and computationally inexpensive features for gait recognition. We analyzed their performance in an open world and a scarce enrollment and authentication data scenario using the OU-ISIR dataset.

The authentication performance of the proposed HMM-UBM system reaches EERs around 1% using a single sensor. This performance level is remarkable compared to the ones reported in the literature for this dataset. Although other works report low error rates, they are tested in closed-set scenarios with fewer users, which hinders a fair comparison with the one proposed here.

The quantization employed in the proposed protected method enables the fusion of multiple sets of features coming from multiple HMM-UBM systems. Furthermore, this fusion provides binary features with a BER that practical ECCs can handle for EKLs equivalent to the requirements for passwords [95] and acceptable FRR.

These results demonstrate the feasibility of the proposed approaches in open-world scenarios, even only with around 5 seconds of enrollment and authentication data.



Chapter 4

On the security of biometrics and fuzzy commitment cryptosystems: a study on gait authentication

Preamble

As biometric templates consist of highly correlated features, the real security level offered by biometric authentication systems remains an open research question. In this work we provide new approximations and a lower bound of the security offered by fuzzy commitment schemes. Fuzzy commitment cryptosystems and in general biometric template protection schemes play an important role in allowing for remote storage and processing of biometric data, as they mitigate the threat of biometric template leakage. The use of such schemes would alleviate some of the usability constraints imposed by the state-of-practice local use of biometrics. As such we conduct an in-depth security analysis for IMU based gait authentication systems, where we evaluate the effectiveness of attacks within the scope of two well-defined threat models that target both unprotected (no template protection) and protected (fuzzy commitment) systems. A pivotal enabler of our analysis is the development of nine different approaches to gait authentication, which allows us to perform intramodal fusion on these distinct, yet highly correlated biometric templates, and to protect them with a strengthened fuzzy commitment scheme. Our analysis clearly demonstrates the high correlation between

the different biometric templates, which, among others, further showcases the threat of biometric template leakage. Furthermore, as our analysis incorporates a threat model that assumes biometric template leakage, it provides metrics for the security provided by the biometric modality itself.

This chapter is entirely based on our published work on evaluating the security offered by biometric authentication systems protected by fuzzy commitment template protection schemes [248]. It extends our work on privacy-enhanced IMU-based gait authentication from Chapter 3 with more learning approaches and delineates an evaluation framework that drives a more in-depth security evaluation. The evaluation framework provides new insights into the validity and reliability of the protected biometric system and enables us to explore the magnitude of advantages gained from intramodal feature fusion. This research result contributes to the behavioral authentication and security metrics pillars.

Publication Data: T. Van hamme, E. A. Rúa, D. Preuveneers, and W. Joosen. "On the Security of Biometrics and Fuzzy Commitment Cryptosystems: A Study on Gait Authentication". In: *IEEE Transactions on Information Forensics and Security* 16, 2021, pp. 5211–5224. DOI: 10.1109/TIFS.2021.3124735

4.1 Introduction

Biometrics has become a popular technology for person authentication due to its usability benefits. However, one of the major concerns when using biometrics is the vulnerability of biometric templates, which can pose a severe threat to both the security and privacy of users [193]. For instance, a leaked biometric template might thwart security for any future use of that trait, and it may also be used to track users among services even if it is irreversible [223]. Furthermore, templates leak private attributes [241, 154].

These critical leakages are often mitigated by using local security elements on the user's device, e.g., Trusted Platform Modules, to store and process the biometric acquisitions and templates. Remote authentication then uses secret keys unlocked by the local biometric authentication protocol [80]. However, the local nature of the templates and secret keys impedes adoption because the resilience of the credentials depends on a single physical device [48]. Moreover, lost or stolen devices burden the user with repeating the whole enrollment process, i.e., biometric enrollment and possibly key exchange. Additionally, for some use cases of biometric authentication, it is impractical to bind a single user with a single device, such as for biometric authentication at multiple entrances of secure facilities.

Remote storage of biometric templates would accommodate more resilient and flexible authentication systems no longer bound to a specific (user) device. Instead of leveraging secure hardware, these systems can rely on biometric template protection schemes [201] to limit the impact of leaking a biometric template. The most secure methods integrate biometric authentication in multifactor authentication systems, where fixed authentication factors, such as passwords or cryptographic tokens, are combined with biometrics. With this approach, it is possible to provide the guarantees recommended in standard best practices [112], which requires: (1) *confidentiality* and *integrity* for the transmission of the biometric templates, and (2) *irreversibility*, *unlinkability*, and *renewability* of the stored biometric templates. These latter constitute the main requirements for template protection schemes.

However, using biometric template protection can result in degraded authentication performance. Intramodal fusion alleviates this concern [20]. This fusion combines biometric representations from different systems, even from the same biometric captures, providing less intraclass biometric template variability. However, there is a limited understanding of how this intramodal fusion of highly correlated data can affect the security properties of the underlying fuzzy extractor template protection scheme. Therefore, a thorough security analysis must be carried out in advance to allow the deployment of systems relying on remote authentication, which offers advantages in usability.

In this paper, we study the impact of leveraging these highly correlated data sources for mono-modal systems by performing an extensive security analysis for both unprotected and protected authentication system architectures. With this in mind, we designed nine different approaches for biometric template extraction for inertial-based gait, which serves as an example of a biometric modality implementable in contemporary mobile devices: five *generative* approaches built on Hidden Markov Models (HMM) and four *discriminative* approaches that rely on Artificial Neural Networks (ANN). These approaches will be used separately or as part of an ensemble. The security analysis follows the following methodology:

- We describe two multifactor authentication schemes: (i) a remote authentication system relying on protected multifactor authentication, based on a Strengthened Fuzzy Commitment, and (ii) a local authentication scenario using unprotected biometrics, which relies on local hardware security.
- We propose two threat models, i.e., the *Traditional Hacker* and the stronger *Biometric Hacker*, where the biometric hacker has the same goal and capabilities as the traditional hacker, but obtained additional knowledge on population-wise biometric template statistics and leaked templates.
- We analyze and evaluate different attacks that these adversaries can perform assuming a disclosure of the fixed factor.

From this analysis, we provide a theoretical and empirical quantitative description of the actual security of both schemes and the negative impact of internal template correlations. From a theoretical point of view, we contribute new approximations and bounds for the complexity of the attacks for protected systems based on fuzzy commitment schemes beyond the commonly used *effective key length*. These measures consider the impact of internal correlations in the biometric representation. They constitute a powerful tool to assess risks in remote biometric authentication systems based on similar approaches. Although the concrete design of the possible attacks depends on the specific protection scheme, the proposed methodology can be applied to any other protection scheme. Applying our methodology to other protection schemes would lead to the specific attacks these adversaries can perform and a well-founded quantitative description of the security these systems can provide.

4.2 Related work

This section presents and discusses relevant previous works on gait authentication, template protection, and fuzzy extraction.

4.2.1 On gait authentication

Disclaimer: This section overlaps with Section 3.2.

Gait authentication with inertial sensors is a well-studied subject with over a decade of work. However, only in 2014 the first large-scale dataset that serves as a reliable benchmark was released, i.e. the OU-ISIR dataset by Ngo et al. [168]. Leveraging this benchmark dataset, Ngo et al. re-evaluate four established gait authentication systems [67, 237, 205, 88]. These systems approach gait authentication similarly: (1) extract gait cycles, (2) pre-process and length normalize these cycles, (3) average the acquired enrollment samples, and (4) at query time, compare the query sample with the stored template via some distance function. Ngo et al. conclude that the most performant gait authentication system reaches an Equal Error Rate (EER) of $\approx 14.3\%$, which is quite far from the required ISO/IEC FAR/FRR working point of 1 in 10k and 1 in 100, respectively[112].

The aforementioned methods rely on robust gait cycle extraction and alignment, a non-trivial task. Therefore, approaches based on Gaussian Mixture Models (GMM) [139, 272, 209, 210] and Hidden Markov Models (HMM) [173, 240] have been proposed. These methods model a window of the gait sequence as the output of an HMM, which consists of a set of states with different GMMs as output probability density functions. Verification is performed by comparing the loglikelihood of the query sample under a

personal model and a general - universal background - model (UBM). A key difference between the early HMM work by Nickel et al. [173] and the other systems are that the former trains the personalized models from scratch, while the latter adapts them from the UBM. Adapting personalized models from the UBM requires only a limited amount of enrollment data compared to training from scratch. Other differences between the approaches reside in feature extraction and how the personal models are adapted from the UBM. Lu et al. [139] demonstrated the strength of the GMM approach by obtaining an EER of $\approx 5.6\%$ on the OU-ISIR dataset. Van hamme et al. [240] improved upon this result with HMMs, achieving an EER of $\approx 1.4\%$.

Since 2017, approaches based on deep learning have been proposed. One of the first works by Nguyen et al. [172] extracts high-level features with a set of different 1-layer Convolutional Neural Networks (CNN), which are fed to a Support Vector Machine (SVM). This system achieves an EER of $\approx 11.6\%$ on the OU-ISIR dataset in a closed-world scenario. IDNet [86] achieved a more promising result for the open-world scenario, which relies on CNNs (3 layers) to extract features fed to a one-class SVM for authentication. However, the evaluation is performed on a small dataset with only 50 users. Later work by Zou et al. [275] outperforms IDNet by combining CNN and Long Short-Term Memory (LSTM) features. Moreover, they show that their approach is superior to IDNet for identification. Zou et al. also evaluate their approach on OU-ISIR, where they achieve an accuracy of $\approx 99.6\%$, which, due to their experimental setup, is equivalent to a Half Total Error Rate (HTER) of $\approx 0.4\%$. Sadly, this number is optimistic, as the input to the network consists of positive pairs – i.e. the enrollment sample and query sample belong to the same user – which can originate from the same sequence and even contain the same step.

Fernandez et al. [77] perform a more rigorous evaluation. They evaluate their Recurrent Neural Network (RNN) based approach in an open-world scenario on the OU-ISIR dataset and obtain an EER of $\approx 6.7\%$. The methods above all rely on robust step segmentation, which is a process that relies on heuristics tuned explicitly for the dataset. Delgado-Escãno et al. [65] do not introduce such bias and achieve an EER of $\approx 1.1\%$ in a closed-world scenario on the OU-ISIR dataset.

Prior work demonstrates the widespread use of the OU-ISIR as a benchmark dataset for inertial sensor-based gait recognition research. At this point, the best-performing system in the open-world scenario is the HMM-UBM with an EER of $\approx 1.4\%$. On the other hand, deep learning-based approaches show much promise with results of around $\approx 1.1\%$ when no new users are introduced during testing, i.e., under a closed world assumption.

4.2.2 On template protection

Template protection has been approached from many different angles. The different methods can be classified into the following main categories [146]:

Cryptosystems. Cryptographic keys are derived from or bound to the biometric information. Some suitable methods are BioHashing [115] and Fuzzy Extractors [72], which will be further explained in the next section.

Non-Invertible transformations. These template protection schemes rely on non-invertible parametric transformations. Some relevant examples are the Cartesian and Polar transformations [6], Wavelet-based methods [253, 100], or Random Projections [186].

Filter-based methods Filter-based methods could also be understood as non-invertible transformations since they rely on the non-invertibility of the chosen filter operations. Non-invertibility is a property of Bloom filters [198] or Log Gabor filters [122].

Hybrid approaches. Hybrid approaches usually rely on applying a non-invertible transformation and a subsequent cryptography-based method [106, 21].

Multibiometrics-based methods. Features from different biometric modalities are fused at the feature level before applying template protection [181, 165].

4.2.3 On fuzzy extraction

In this work, we focus on Fuzzy Extractors [72] since it is a well-analyzed method that smoothly integrates into cryptographic security protocols. It is a method to convert a fuzzy authentication factor, such as biometrics, into a fixed key that can be used in cryptographic protocols. Fuzzy extractors exist for different kinds of fuzzy factors, like the fuzzy commitment [118] for fixed-length binary strings or the fuzzy vault [118] for variable length sets. The previous examples use Error Correcting Codes (ECCs) to deal with the variability of the fuzzy factors. However, other recently proposed approaches exist, such as using Learning with Errors in Computational Fuzzy Extractors [85]. There are also examples for different biometric modalities, as fingerprint [17], online

signature [21], iris [10], or gait [240] authentication, and even for multibiometrics, such as [49].

Argones Rúa et al. [20] have recently shown that it is possible to easily integrate fuzzy extractors with additional fixed factors in strengthened, fuzzy extractors (SFE). The fixed factor encrypts the public helper data, which can be designed to allow the best authentication performance. The evaluation is done for Gaussian sources protected by a fuzzy commitment. The effective key length describes fuzzy commitment security. The effective key length is the size of the message space of the ECC used to commit the biometric binary string, i.e., the size of the stored hash preimage space. This and the previous works rely on the fuzzy extractor assumptions for providing security, i.e., the fuzzy source should have a bound entropy (or min-entropy [72]). On the other hand, biometric security is informed in terms of the authentication rates regarding biometrics, thus making it difficult to assess the overall security of the proposed solution. Although these measurements are sufficient when the biometric source abides by the fuzzy extractor entropy requirements, a serious in-depth security analysis regarding the statistical properties of the biometric templates and their impact on the overall security is missing. Ideally, this analysis should lead to not only an estimation of the actual biometric entropy finally incorporated into the system but also of the actual security of the template protection scheme and a quantitative methodology to maximize it.

4.3 Gait authentication systems

This section describes the gait authentication systems employed in this work. First, we describe the composition of the protected and unprotected systems. Then, we describe the joint parts, i.e., the employed gait representation and the pattern recognition experts, one generative and the other discriminative. Finally, we briefly describe the intramodal fusion employed within the SFE for the system relying on protected templates.

4.3.1 Remote biometric authentication systems architectures

Figure 4.1 shows the composition of the remote biometric authentication systems, one based on unprotected templates and the other on protected templates using SFE.

Unprotected system

As shown in Figure 4.1, a biometric sensor acts as the interface between the physical and digital world by capturing the biometric trait. From the raw biometric data, a biometric template is computed, which is compared to the biometric templates obtained

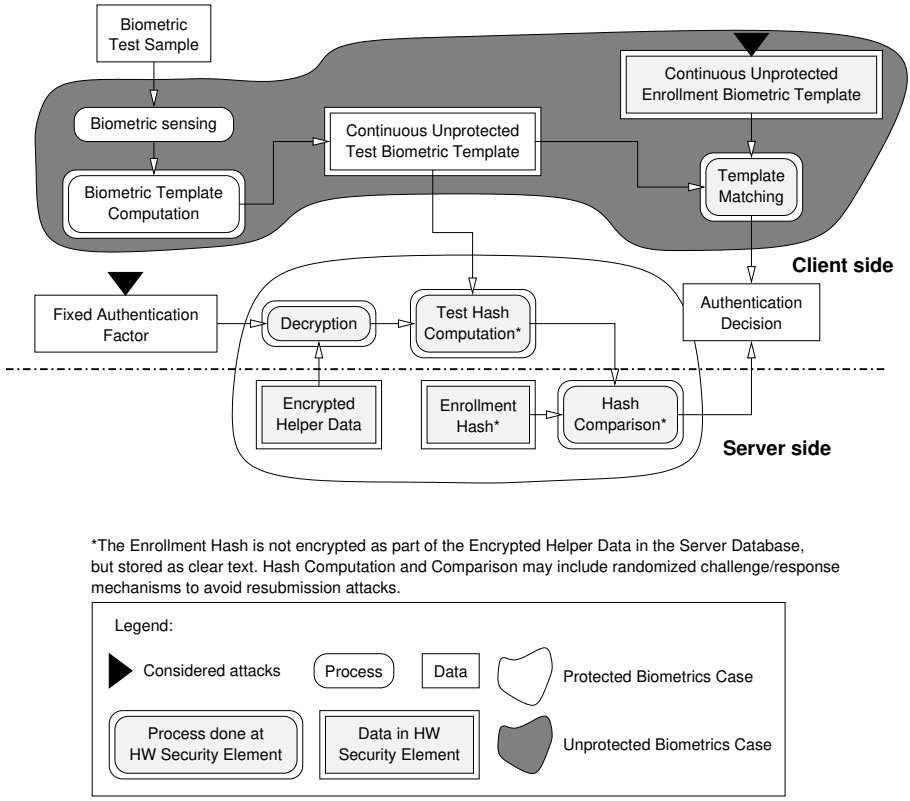


Figure 4.1: Biometric Authentication Architectures, including hardware security features usage, SFE-protected/unprotected templates-based authentication, and the attack points considered in the security analysis.

during the enrollment phase. This comparison yields a single value, which is the basis of the authentication decision.

In our system, the biometric sensor is an IMU sensor that captures the angular velocity and linear accelerations over time. We provide more information on how the streams are represented in Section 4.3.2. We implemented two entirely different approaches for biometric template computation, i.e., a discriminative and generative approach, described in Sections 4.3.3 and 4.3.4, respectively. The matching is based either on cosine similarity or a loglikelihood ratio test. The final decision depends on a pre-determined threshold.

The main threats to unprotected biometric systems are lost or stolen biometric templates, which can be leveraged in spoofing and replay attacks. They also pose a privacy risk. To

avoid such attacks and prevent biometric templates from leaking, unprotected systems are mainly intended for use cases where the whole biometric pipeline can be trusted, e.g., when the system is implemented end-to-end in secure hardware in a single device [91]. However, such trust assumptions are hard to justify for remote biometric authentication systems, where templates are stored remotely, thereby effectively limiting the use cases of unprotected systems and motivating the use of biometric template protection schemes.

Protected system

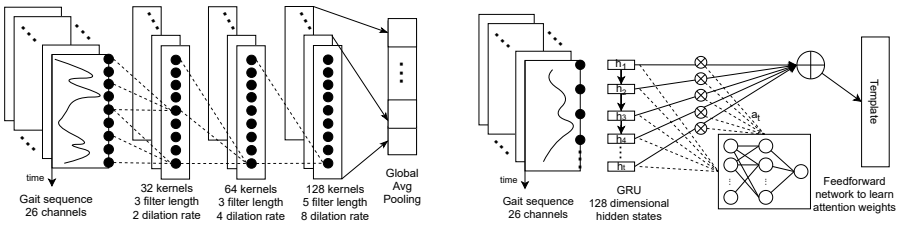
Using biometric template protection seems a good match for systems relying on remote biometric authentication, where biometric credentials are stored server-side. Server-side template storage allows for more flexible usage of biometrics, unbinding the template to any specific capture device. An essential requirement in these systems is the secure storage of biometric credentials. Secure storage implies irreversible, unlinkable, and renewable templates.

In this study, we focus on SFE, a multifactor-protected biometric system that satisfies the requirements [20] mentioned above, where a fixed authentication factor is used to encrypt an *enriched* fuzzy extractor helper data, consisting of:

- *Quantization Helper Data* (QHD) allows for a more efficient and effective quantization of the continuous biometric template into a less variable binary string and unbinds the resulting binary representation from the user's biometrics.
- *Fuzzy Commitment* (FC), computed during enrollment as the xor of the biometric binary representation and a random codeword from an ECC.

As shown in Figure 4.1, this encrypted helper data is stored in the server database. As the encrypting key is only known to the user, the helper data remains opaque to the server. The server also stores the hash of the message used to encode the codeword bound in the fuzzy commitment. Authentication can directly compare the hash stored server and decoded user sides. However, more advanced implementations avoid replay attacks by using the hash in cryptographic protocols to enable cryptographic challenge-response or zero-knowledge proof protocols.

This system meets the required irreversibility, unlinkability, and renewability properties. However, unlinkability is not preserved when the fixed authentication factor is disclosed. Even the provided security can be degraded below the effective key length of the underlying fuzzy commitment in the presence of internal correlations in the biometric binary representation. The impact of this will be quantified during our analysis.



(a) TCN (b) GRU with attention

Figure 4.2: The ANN architectures for the TCN and GRU.

4.3.2 Stream representation

Disclaimer: This section is a summary of Section 3.3.

Inertial Measurement Units (IMUs) measure movement as a six-dimensional time series: 3D linear acceleration and 3D angular velocity. To help our system extract relevant features, we represent the streams as a 26-dimensional time series as described in our earlier work [240]:

- *Gyroscopic dynamics (8D)*: angular velocity, the first-order difference of angular velocity, and their magnitude.
- *Roll and pitch from accelerometer (6D)*: calculated as the arctan between the appropriate accelerations axes. Roll and pitch velocity and acceleration are approximated by first- and second-order differences.
- *Roll and pitch from accelerometer and gyroscope (6D)*: fusing an estimation of roll and pitch from the accelerometer and the gyroscope with a complementary filter. Again, velocity and acceleration are approximated.
- *Vertical and horizontal acceleration (6D)*: the estimated vertical and horizontal acceleration in the world plane, obtained by aligning with the gravitational component.

4.3.3 Discriminative approach: Neural networks

Open set recognition with only a few enrollment samples has been successfully tackled in face recognition by applying metric learning [151]. In particular, siamese networks trained with triplet loss [213] have achieved excellent results. However, recent successes in face recognition are achieved by training a classifier with a variation

of the well-known cross-entropy loss, with the most performant loss function being arcface loss [66]. In our work, we adopt the aforementioned loss functions to create gait authentication systems. Thus, we use neural networks to obtain a low-dimensional representation of the gait sequences. Authentication decisions are then made by comparing a pre-determined threshold with the output of a distance function computed on the representation of the enrollment and query sequences.

Besides defining a function to optimize, we require a method to capture the characteristics of the gait sequences. The two most logical choices are Temporal Convolutional Networks (TCNs) and Recurrent Neural Networks (RNNs). TCNs have been shown to often outperform RNNs for sequence modeling tasks [25]. Moreover, to ensure a sufficiently large receptive field, we build upon dilated convolutions as defined by van den Oord et al. [177]. Recently, due to attention mechanisms, RNNs have also increased their receptive field, making them more suited for long sequences [24], such as gait.

After extensive tuning, we settled on the following two architectures, also shown in Figure 4.2: 1) a TCN with three layers, where the dilation rate and the number of kernels are multiplied by two in every layer. The final representation is achieved by applying global average pooling. 2) an RNN with one layer of Gated Recurrent Units (GRUs), with an additive attention layer [24] which is aligned with the input sequence. We train both architectures with the two aforementioned loss functions, i.e., triplet loss and arcface loss, which gives four different systems.

4.3.4 Generative approach: HMM-UBM

Gait acquisitions are variable-length sequences of feature vectors, which we model as the output of a Continuous Density Hidden Markov Model (CD-HMM), as done in [240]. Each state output probability density function is a Gaussian Mixture Model (GMM). As authentication deals with scarce enrollment data, a Universal Background Model (UBM) approach is a good fit because it requires limited data to obtain personalized models from the UBM. The UBM is initialized [135] and trained [189] from multiple-person gait sequences, representing the whole population. Since we will protect the system with a fuzzy extractor which requires a fixed number of features to characterize each user, we extract the well-studied eigenfeatures from the HMM-UBM. These eigenfeatures are a low-dimensional representation of the adapted models, obtained by applying Probabilistic Principal Components Analysis to the differences between the means of the adapted models and the UBM for a set of users representing the whole characterized population, as first introduced in [124] for GMMs, and explained in detail in [21] for the more general case of HMMs.

4.3.5 Intramodal fusion

In this work, we follow the strategy indicated in [20] for features combination for fuzzy extraction, where real-valued features are combined into bits according to their reliability. In our case, there are no real-valued features reliable enough to obtain more than 1 bit per feature, so the relevant reliability measure is the quotient $\frac{\mu}{\sigma}$, where μ is the average, and σ is the standard deviation of the user-specific real-valued features, once they are normalized to have a population-wise 0 mean. The feature combination is thoroughly explained in work mentioned above. It consists of a greedy algorithm combining the real-valued features into the required number of bits, seeking a balanced error probability among them.

We assume that the fused features are uncorrelated for this study. However, we know beforehand that this assumption only sometimes holds when dealing with intramodal fusion, especially when the acquisition used for verification is the same for all the systems. There are several reasons to make this assumption: (i) it may be challenging to fully decorrelate these features since they are modeling the same phenomena, and trying to do it can harm the complementary information the fused systems convey; (ii) it has been observed that proceeding this way has a positive impact in both authentication rates and fuzzy extractor parameters eligibility, as observed in [21]; and (iii) although we know that this will have an impact on the fuzzy extractor security, we are interested in quantifying this impact.

4.4 Security analysis

In this section, we provide a security analysis of biometric systems using unprotected and protected templates relying on SFE in remote authentication scenarios. We first define two threat models, which we use to design attacks against unprotected and protected systems.

4.4.1 Threat models

The performance of an approach for biometric authentication is traditionally described with metrics such as the False Acceptance Rate (FAR), False Rejection Rate (FRR), Equal Error Rate (EER), and Half Total Error Rate (HTER). Alternatively, the FAR and FRR are sometimes called the False Match Rate (FMR) and False Non-Match Rate (FNMR). On top of the FAR and FRR, a protected system that extracts keys is additionally characterized by the *effective key length*. Even though these metrics are widespread, they are insufficient for a security evaluation because such evaluations must be executed within a well-defined threat model for the metrics to be meaningful.

Therefore, we first define two different threat models where the attacker spoofs biometric data: the *traditional hacker* and the *biometric hacker*. The main difference is that the biometric hacker is a stronger adversary that leverages biometric knowledge, such as leaked biometric templates or the distribution of biometric representations.

Traditional hacker

The traditional hacker's *goal* is to obtain access to the system as a specific user. Furthermore, the traditional hacker is *capable* of spoofing input data to every process in Figure 4.1. This adversary has no *knowledge* on the distribution of biometric templates.

Unprotected system The traditional hacker is a baseline for the unprotected system that compares our approach to IMU-based gait authentication with related work. In this threat model, the FAR, as it is traditionally reported, captures the probability of the traditional hacker triggering a positive authentication response when using actual raw biometric data, such as their own.

Protected system Also, for the protected system, an evaluation within the threat model of the traditional hacker serves as a baseline where the effective key length captures the system's security.

Biometric hacker

The biometric hacker has the same *goal* and *capabilities* as the traditional hacker but has superior *knowledge*. I.e., the biometric hacker knows the distribution of biometric templates and has access to leaked biometric templates. Since the biometric hacker is a stronger attacker than the traditional hacker, they can execute all the attacks the traditional hacker can.

Unprotected system In the case of the unprotected system, if the biometric template of a specific user leaks, the adversary achieves their goal. Therefore we assume the attacker obtained the target user's template leaked from another system. Moreover, the template was computed by a different template extraction function and thus cannot be used directly in the target system.

Protected system This adversary gains control of the user's fixed authentication factor, obtaining the decrypted SFE helper data by normal interaction with the

authentication system. In the system described above, this helper data consists of the fuzzy commitment (the n-bit xor of a specific ECC codeword and the biometric binary array representation) and the QHD.

4.4.2 Attacks against unprotected system

Zero-effort attack

In a zero-effort attack, the traditional hacker injects their own biometric data into the template computation module. Then, as mentioned earlier, the FAR indicates the success rate of the adversary. Furthermore, analyzing this attack allows us to study the security-usability, i.e., FAR-FRR, trade-off and compare our gait authentication systems with related work. Thus, during our evaluation, we report the traditional metrics, i.e., FAR, FRR, EER, and HTER.

Biometric template correlation attack

We construct an attack that leverages the correlation between biometric templates extracted by different biometric 'template computation' approaches, called the correlation attack. This attack is executed by the biometric hacker, who will try to impersonate a target user $\mathcal{U}_T$ in a target biometric system $\mathcal{S}_T$ by injecting a forged template. The adversary obtains a leaked template from the target user, namely $\mathcal{T}_L^{\mathcal{U}}$, and a set of pairs of templates $\mathcal{P} = \left\{ \left(\mathcal{T}_L^{\mathcal{U}_i}, \mathcal{T}_T^{\mathcal{U}_i} \right) \right\}$ from both leaked and target systems, extracted from the same biometric samples, being representative of the biometric population. The adversary constructs n functions $\phi_n(.) : \mathbb{R}^m \rightarrow \mathbb{R}$ with m and n the number of features of the leaked $\mathcal{T}_L^{\mathcal{U}_i}$ and target templates $\mathcal{T}_L^{\mathcal{U}_i}$, respectively. The functions $\phi_n(.)$ are chosen such that $[\phi_0(\mathcal{T}_L^{\mathcal{U}_i}), \phi_1(\mathcal{T}_L^{\mathcal{U}_i}), \ldots, \phi_n(\mathcal{T}_L^{\mathcal{U}_i})]$ approximates $\mathcal{T}_T^{\mathcal{U}_i}$. The adversary then uses the functions $\phi_n(.)$ to forge a template $\mathcal{T}^{\mathcal{A}}$ from the leaked template $\mathcal{T}_L^{\mathcal{U}_i}$, which they inject into the biometric template matching module. The matching module then compares the forged biometric template $\mathcal{T}^{\mathcal{A}}$ with the enrollment template of the target user $\mathcal{E}_T^{\mathcal{U}}$. We will report the success rate of the correlation attack with the imposter acceptance rate IAR = $\frac{\text{\# successful attempts}}{\text{total \# of attempts}}$.

4.4.3 Attacks against protected system

Hash preimage brute force attack

Even an adversary without information on the distribution of the biometric binary templates can try a brute force attack on the hash preimage space, with a complexity $\mathcal{C}_{\text{hacker}}$ equal to the effective key length (k, the size of the message space of the (n, k, t) ECC with key length n and error correcting capability t) of the fuzzy extractor:

$$\mathcal{C}_{\text{hacker}} = k \tag{4.1}$$

The risk associated with this attack in the remote authentication scenario is very low for even moderate values of k, since the server can very likely detect the attack, which consists of repeatedly trying different messages' hash values, thus triggering a convenient defensive action.

Biometric binary representation statistical attack

Since the binary biometric representation committed in the fuzzy commitment is random and uncoupled from the continuous biometrics in the SFE scheme, the adversary needs access to the QHD and biometric template statistics to get an advantage beyond the brute force hash preimage attack. Our assumption includes that the biometric hacker can decrypt the QHD. To estimate the biometric template, the adversary can randomly sample from the biometric distribution, ideally leading to a complexity equal to the biometric template entropy. This attack is similar to the *guessing attack* proposed in [123, 202], where the authors use a representative population of templates to perform the attack. We must note that since the adversary samples from the biometrics distribution, its success probability $\mathrm{P}^{\text{template}}_{\text{bio-hacker}}(\text{success})$ is the False Acceptance Rate (FAR), resulting in the following exact value of this attack's complexity:

$$\mathcal{C}^{\text{template}}_{\text{bio-hacker}} = -\log_2(\text{FAR}) \tag{4.2}$$

This exact value may not be accurately empirically estimated for low values of the FAR. Thus, we propose both a theoretical bound and an estimate of this.

A common way to estimate the biometrics distribution entropy is computing the number of degrees of freedom (*DoF*) of the best binomial fit to the inter-class (impostor attempts) Hamming distance distribution, as shown in [63, 197], and computed as $DoF = p(1-p)/\sigma^2$, where p is the average bit error rate and σ is the standard deviation of the inter-class binary representations average Hamming distance. Without intra-class

variability, this only figure would sufficiently describe the scheme's security. However, some intraclass variability must be allowed, and this is done by using ECCs. The allowed intraclass variability decreases the complexity of this attack since the biometric hacker does not need to get an exact binary representation either. For example, suppose the system employs a code with error correcting capability t errors. In that case, we can extrapolate the proportion of allowed errors t/n to the corresponding value in the biometric information space to obtain an upper bound of the probability of success of the attack. This attack's success probability will always be smaller than the cumulative distribution of the binomial in the $\lfloor DoF \rfloor$-dimensional biometric information space evaluated at $\left\lceil DoF\frac{t}{n} \right\rceil$:

$$\mathrm{P}^{\text{template}}_{\text{bio-hacker}}\left(\text{success}\right) \leq \frac{1}{2^{\lfloor DoF \rfloor}} \sum_{k=0}^{\left\lceil DoF\frac{t}{n} \right\rceil} \binom{\lfloor DoF \rfloor}{k}, \tag{4.3}$$

resulting in a lower bound of the complexity of this attack:

$$\mathcal{C}^{\text{template}}_{\text{bio-hacker}} \geq \lfloor DoF \rfloor - \log_2 \left(\sum_{k=0}^{\left\lceil DoF\frac{t}{n} \right\rceil} \binom{\lfloor DoF \rfloor}{k} \right) \tag{4.4}$$

This bound's first and second terms quantify the impact of inter-class variability and the allowed intra-class variability on the attack's complexity. However, this bound is pessimistic due to the necessary rounding operations. Thus, we propose using a parametric approximation, provided by a β-binomial fit of the impostor Hamming distance distribution instead of these. This distribution is a generalization of the binomial distribution, adequate for over-dispersed distributions, and already employed to estimate biometrics statistics in [256]. In this case, we have an approximation of the complexity given by:

$$\mathcal{C}^{\text{template}}_{\text{bio-hacker}} \approx -\log_2 \left[CDF_{\beta\text{-Bin}\left(n,\hat{\alpha},\hat{\beta}\right)}\left(t\right) \right], \tag{4.5}$$

with $CDF_{\beta\text{-Bin}\left(n,\hat{\alpha},\hat{\beta}\right)}$ the cumulative distribution function of the β-binomial distribution with n degrees of freedom, and parameters $\hat{\alpha}$ and $\hat{\beta}$ estimated from data using the method of moments.

Hash preimage statistical attack

The other strategy available to the biometric hacker is to attack the ECC message space, as in the brute force hash preimage attack but reduce the complexity by using knowledge of the biometric templates distribution. To determine the theoretical lower

bound of the complexity of such an attack, we hypothesize an equivalent protected biometric system in the biometric information space. Hence, we assume that there exists an ideal transformation from the n-dimensional binary template space to the DoF-dimensional biometric information space that ideally compresses the biometric information $\mathcal{T} : \{0,1\}^n \longrightarrow \{0,1\}^{DoF}$. The adversary can obtain no further advantage after the impostor distribution is binomial, as shown in [134]. We also presume the existence of an error-correcting code with the same (n, k, t) ratios as the one used by the original system: $ECC(DoF, k\frac{DoF}{n}, t\frac{DoF}{n})$. In this equivalent theoretical system, the complexity of the hash preimage attack is the equivalent effective key length $k\frac{DoF}{n}$. However, in practice, finding the optimal transformation is infeasible, and the best practical solution can only ideally achieve this complexity:

$$\mathcal{C}_{\text{bio-hacker}}^{\text{hash-preimage}} \geq k\frac{DoF}{n} \tag{4.6}$$

In this equation, inter-class variability determines the factor *DoF*. The allowed intra-class variability determines the factor k/n through the relation between k and t for the specific chosen ECC. Generally, when the allowed intra-class variability t increases, the factor k/n decreases.

The results shown in Eqs. 4.1, 4.2, and 4.6 describe the complexity of the attacks against the protected systems in order to get the binary biometric template from the SFE public helper data. More specifically, $\mathcal{C}_{\text{hacker}}$ is the strength against a brute force preimage attack against the hash, not relevant in our scenarios, where the hash is not disclosed. However, an adversary with knowledge of biometrics, or having access to an adequate population of biometric templates, can perform more elaborated attacks by obtaining a similar biometric template, with complexity $\mathcal{C}_{\text{bio-hacker}}^{\text{template}}$, or by trying to exploit biometric statistical properties to perform a hash preimage attack, with complexity $\mathcal{C}_{\text{bio-hacker}}^{\text{hash-preimage}}$. We define for convenience $\mathcal{C}_{\text{bio-hacker}} = \min\{\mathcal{C}_{\text{bio-hacker}}^{\text{template}}, \mathcal{C}_{\text{bio-hacker}}^{\text{hash-preimage}}\}$ as the complexity the biometric hacker faces to impersonate a user.

4.5 Experiments

In this section, we describe and report on the different experiments that allow us to empirically quantify the security offered by the protected and unprotected systems within their respective threat models.

4.5.1 Database

Disclaimer: This section has some overlap with Section 3.5.

To train and evaluate our systems, we use the IMU sequences in the OU-ISIR dataset labeled as *level walk* and captured by the center sensor. Usually, for each user, there are two sequences, of which we use one for enrollment and one for testing. Therefore, we only consider the 483 users with two valid walking sequences for all three sensors, i.e., left, right, and center-located sensors.

The sampling rate is 100Hz, and the sequences are, on average, 5.2 seconds long. We use the entire length of the gait sequence for training and evaluation, except for training the ANNs. In the case of ANNs, we select a part of the original sequence of fixed length, i.e., 1.8 seconds, and for each epoch, we select a random starting point within the original sequence.

4.5.2 Experimental setup

We adopt a 10-fold cross-validation protocol with a test and validation set, where we randomly split the 483 users into 10 disjoint sets which contain either 48 or 49 users: $S_s = \{U_0^s, U_1^s, \ldots, U_x^s\}$ with $x \in [47, 48]$ and $s \in [0, 9]$. Thus, we do 10 rounds, where during each round k, we select S_k as the test set, $S_{k+1 (\mathrm{mod}\ 10)}$ as the validation set, and the remaining 8 sets as the training set. The validation set serves the purpose of tuning hyperparameters and selecting the authentication threshold of the system. We select the threshold in two ways: 1) as the threshold at which the equal error rate (EER) is obtained – the threshold at which false acceptance rate (FAR) is equal to false rejection rate (FRR) – and 2) as the threshold for which the FRR is at most 5%.

For unprotected systems, we report for these thresholds respectively the half total error rate (HTER $= \frac{FAR+FRR}{2}$) and FAR in the test set. Additionally, we report the EER in the test set to compare with related work. As each set contains x users for whom there are two walking sequences, x benign authentication attempts and $x \cdot (x-1)$ malicious authentication attempts can be simulated.

For protected systems, each ensemble is tested using the available BCH codes with code lengths $n = \{63, 127, 255, 511\}$. At each code length, we are interested in finding the code that provides the highest security, defined as:

$$\mathcal{S} = \min \left\{ \mathcal{C}^{\text{hash-preimage}}_{\text{bio-hacker}}, \mathcal{C}^{\text{template}}_{\text{bio-hacker}} \right\}, \tag{4.7}$$

while ensuring an average FRR in the validation sets below 5%, which can be considered a sensible working point regarding usability. Besides the attack complexities, we also report the average FAR and FRR in the test sets.

4.5.3 Experimental results

Unprotected systems security evaluation

We first evaluate the security of the unprotected systems for zero-effort attacks utilizing the traditional metrics, i.e., FAR, EER, and HTER. Afterward, we evaluate system security in the presence of a biometric hacker. In zero-effort attacks, adversaries have no biometric knowledge and use their own biometric traits. Finally, we build and evaluate authentication systems with 9 approaches: 5 based on the HMMs and 4 on ANNs. We also create an ensemble authentication system that fuses the scores of different approaches.

Algorithm		EER%	HTER%	FAR@5%FRR%
HMM LLKR	1 S 128 M	1.24 ± 0.60	2.21 ± 0.50	0.23 ± 0.09
	2 S 64 M	1.33 ± 0.44	2.29 ± 0.47	0.16 ± 0.08
	4 S 32 M	1.47 ± 0.36	2.10 ± 0.72	0.21 ± 0.08
	8 S 16 M	1.42 ± 0.31	2.35 ± 0.59	0.24 ± 0.12
	16 S 8 M	1.37 ± 0.36	2.16 ± 0.63	0.32 ± 0.27
HMM COS	1 S 128 M	1.18 ± 0.48	2.06 ± 0.84	0.52 ± 0.57
	2 S 64 M	1.27 ± 0.80	1.77 ± 0.83	0.89 ± 1.27
	4 S 32 M	1.45 ± 0.54	1.88 ± 0.73	0.99 ± 0.68
	8 S 16 M	1.55 ± 0.61	1.83 ± 0.58	0.83 ± 0.52
	16 S 8 M	1.69 ± 0.53	2.44 ± 0.55	1.14 ± 0.40
ANN	TCN triplet	1.77 ± 0.72	2.23 ± 0.79	0.96 ± 0.83
	TCN arcface	1.20 ± 0.64	2.16 ± 0.91	0.75 ± 0.65
	GRU triplet	2.13 ± 0.75	2.58 ± 1.13	0.87 ± 1.05
	GRU arcface	0.97 ± 0.60	1.83 ± 0.73	0.42 ± 0.72
Ensemble	HMM COS 6 systems	0.95 ± 1.03	1.58 ± 1.09	0.63 ± 0.95
	ANN 12 systems	0.50 ± 0.80	1.41 ± 1.28	0.21 ± 0.63
	HMM & ANN 23 systems	0.40 ± 0.64	1.22 ± 1.49	0.0 ± 0.0

Table 4.1: Mean authentication performance in the test fold of the different systems in terms of EER(%), HTER(%) for the validation set EER threshold, and FAR(%) when the FRR in the validation set is at most 5%.

Zero-effort attack We construct five different HMM-UBM systems to keep the model complexity consistent, i.e., $S \cdot M = cte$, with S the number of states in the HMM and M the number of mixtures per state. Specifically we choose $M = 2^{7-i}$ with $i \in [0, 4]$ and choose S such that $S \cdot M = 128$. For each of the 5 mixture-state combinations, we repeated the experimental protocol described in Section 4.3.3 four times. The matching scores are obtained in two ways: 1) by calculating the loglikelihood ratio (LLKR) between the adapted user model and the UBM. 2) by calculating the cosine similarity (COS) between the templates, i.e., the eigenmodel coefficients. The results in Table 4.1 are obtained by averaging system performance within every fold and then averaging over the folds, where we assume that the standard deviation is equal to the square root of the average of the variances. We obtain well-performing systems with EERs close to 1%, which is in line with related work, cfr. Section 4.2.1. Table 4.1 also shows that both scoring methodologies, LLKR and COS, lead to authentication systems with similar performance in terms of EER and HTER. However, the LLKR provides better FRRs for low FARs.

We also evaluate the 4 different ANN systems described in Section 4.3.3: TCN with triplet loss, TCN with arcface loss, GRU with triplet loss, and GRU with arcface loss. For each of these systems, we repeat the experimental protocol five times. Table 4.1 summarizes the system performance. In general, the ANNs perform on par with the HMMs. Furthermore, the arcface loss seems to be the superior loss function.

We fuse the scores (of the cosine similarity function) of the different approaches and create ensembles of HMMs, ANNs, and the combination of HMMs and ANNs. As we create 4 different systems for each of the 5 HMM topologies, and 5 different systems for each ANN architecture, we have 20 generative and 20 discriminative systems. The ensemble authentication system is obtained by summing the scores, which are normalized with *tanh-estimators* [207], as implemented by the bob suite [13]. The tanh-estimators' parameters are obtained through the validation set's positive scores - benign authentication attempts. We gradually added systems until all the available systems were fused. Then we select the ensemble with the best EER and the minimal amount of systems; for this ensemble, we show the performance in Table 4.1 and report the number of fused systems.

Biometric template correlation attack As previously stated, the most significant security risk in biometric systems relying on unprotected templates is their leakage, especially when they are obtained by a hacker knowledgeable on the underlying biometric algorithms.

As described in Section 4.4.2, a motivated biometric attacker can construct an attack where they feed the biometric system an estimate of the biometric template of their victim. This estimate is obtained by learning a mapping between pairs of templates

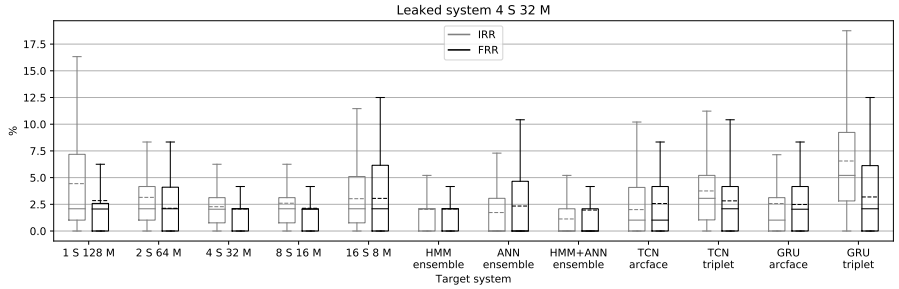


Figure 4.3: The FRR of the system and the imposter rejection rate (IRR) of the template estimation attack based on linear regressors, where the leaked system is the HMM with 4 states and 32 mixtures, with the target systems on the x-axis. The dotted line indicates the mean, while the full line represents the median.

from a leaked and target system: $\mathcal{P} = \left\{ \left(\mathcal{T}_L^{\mathcal{U}_i}, \mathcal{T}_T^{\mathcal{U}_i} \right) \right\}$. Each template $\mathcal{T}^{\mathcal{U}_i}$ consists of 128 continuous features: $\mathcal{T}^{\mathcal{U}_i} = [f_0, f_1, \ldots, f_{127}]$. Thus we can learn a set of linear regressors such that $\mathcal{T}_T^{\mathcal{U}_i} = [\phi_0(\mathcal{T}_L^{\mathcal{U}_i}), \phi_1(\mathcal{T}_L^{\mathcal{U}_i}), \ldots, \phi_{127}(\mathcal{T}_L^{\mathcal{U}_i})]$. Specifically, we apply principal component regression, where we select the 64 most principal components of the leaked templates $\mathcal{T}_L^{\mathcal{U}_i}$. As a significant amount of data is required to train the regressors, we train, for each fold, the regressors with the training data, while the test fold is used to evaluate the effectiveness of the attack. An attack is successful if the system accepts the estimated template, where the authentication threshold is set as the EER threshold in the validation set. We performed the attack for each fold and each sensible combination of leaked and target systems. We illustrate the attack success rate for a typical example in Figure 4.3, i.e., the biometric templates for the 4 states 32 mixture HMM is leaked. We also report results when ensemble systems are targeted, which fuse the scores of 4 different HMMs - we excluded the leaked system -; the 4 ANNs, or both the HMMs and ANNs. It is clear that with mean Imposter Rejection Rates (IRR) that are close to the FRR, this is a very successful attack. The same observation holds when biometric templates from other systems are leaked.

Figure 4.3 also shows that ensemble systems are not harder to attack than individual systems because they have lower mean IRR than each fused system. Furthermore, Figure 4.3 also shows that the more similar two systems are, the easier it is to target.

Protected systems security evaluation

The security evaluation results of SFE-protected ensembles are summarized in Table 4.2. In this table, E stands for the ensemble size. $\mathcal{C}_{\text{hacker}}$ is the effective key length,

$\mathcal{C}^{\text{template, }\beta\text{-Bi}}_{\text{bio-hacker}}(\text{val})$ is the theoretical estimation of the biometric hacker template attack complexity computed as shown in Equation 4.5 in the validation sets, $\mathcal{C}^{\text{template, emp.}}_{\text{bio-hacker}}(\text{val})$ and $\mathcal{C}^{\text{template, emp.}}_{\text{bio-hacker}}(\text{test})$ are the empirical values of this complexity obtained in the validation and test sets respectively, and $\mathcal{C}^{\text{hash-preimage}}_{\text{bio-hacker}}$ is the theoretical bound of the biometric hacker hash preimage attack computed as shown in Equation 4.6 in the validation sets.

The provided security figures remain stable for the different binary representation lengths, slightly below 7 bits for both HMM and ANNs, while reaching slightly more than 7 bits for the mixed ensembles. The maximum provided security is 6.97 bits for HMM ensembles, 6.55 bits for ANN ensembles, and 7.36 bits for mixed ensembles. Results also demonstrate that the hash preimage attack complexity lower bound is significantly below the template attack complexity theoretical and empirical estimates in all cases.

We show the evolution of the attack complexity on the ensemble size for mixed ANN and HMM ensembles in Figure 4.4 as a representative example.

In both biometric hacker attacks, complexity generally decreases when increasing the ensemble size above a specific size which depends on the binary representation length. Similar behavior occurs for ensembles with only HMM or ANN systems.

Figure 4.5 shows the evolution of the degrees of freedom of the binomial fit of the interclass Hamming distances, together with the obtained effective key length and the required effective key length that would provide a biometric hacker hash preimage attack complexity greater than 6 bits, according to Equation 4.6, for ensembles of HMM systems with binary templates of length 127. As we can see, the obtained effective key length grows more or less in line with the required one, but it is clearly below this requirement for too small or too high ensemble sizes. Therefore, we can infer that there exists an ensemble size that provides a maximum security working point. We should note that for $n = 511$, this ensemble size is bigger than the available number of systems in our ensembles, which is 20 for HMM and ANN ensembles and 40 for mixed ones. However, considering that security remains approximately constant for the different code lengths, the computational cost associated with the increase in the ensemble size does not seem to be justified.

Regarding the validity of the entropy estimation by computing the degrees of freedom of the binomial fit of the interclass Hamming distances, we show a representative example in Figure 4.6 for a code length $n = 255$ and ensemble size $E = 8$, where the binomial and β-binomial fits for these distances distribution on the test sets are shown. In this example, $DoF = 16$. The remarkably good approximation of the empirical distributions shown by the binomial fit supports the assumptions made on the theoretical security bound for $\mathcal{C}^{\text{hash-preimage}}_{\text{bio-hacker}}$ in Eq. 4.6. The β-binomial fit demonstrates a remarkable precision across the entire domain. However, as shown in Table 4.2,

the tail of the actual distribution of Hamming distances seems to be slightly heavier than the β-binomial tail, which makes the theoretical approximation $\mathcal{C}_{\text{bio-hacker}}^{\text{template, }\beta\text{-Bi}}(\text{val})$ slightly optimistic. As shown in Table 4.2, the corresponding empirical security figures are consistently around 1 bit less than the values computed from the β-binomial fit. This can indicate the presence of different types of outliers [260], but it also indicates the need for being conservative regarding this theoretical approximation.

On a final note, we would like to point out that the behavior of the systems with binary length 511 is limited by the available number of systems in the ensemble, which tend to make the observed $\mathcal{C}_{\text{bio-hacker}}^{\text{hash-preimage}}$ smaller than for smaller binary lengths. For these systems, the optimal security values are expected only to be reached for considerably larger ensembles. Also, some fluctuations are due to the discrete nature of the ECC's error-correcting capability t, which determines the FAR/FRR working point. For example, in Figure 4.5, the effective key length drops around ensemble size 11 due to fluctuations in the evaluation set performance since the ECC parameters were chosen such that the FRR in the validation set is under 5%.

System	n	E	$\mathcal{C}_{\text{hacker}}$	$\mathcal{C}_{\text{bio-hacker}}^{\text{template, }\beta\text{-Bi}}(\text{val})$	$\mathcal{C}_{\text{bio-hacker}}^{\text{template, emp.}}(\text{val})$	$\mathcal{C}_{\text{bio-hacker}}^{\text{template, emp.}}(\text{test})$	$\mathcal{C}_{\text{bio-hacker}}^{\text{hash-preimage}}$	FAR%	FRR%
HMM	63	4	36	12.04	11.89	10.89	6.97	0.05	6.21
	127	10	85	13.41	$+\infty$	11.48	6.62	0.03	6.63
	255	19	155	14.71	13.48	12.48	6.55	0.02	8.07
	511	15	76	13.84	12.48	11.31	3.66	0.17	2.28
ANN	63	3	30	10.39	9.89	8.81	5.62	0.22	4.35
	127	8	92	12.32	10.78	9.78	6.18	0.11	4.55
	255	16	187	14.00	11.48	10.23	6.55	0.08	4.14
	511	20	211	14.69	11.31	10.57	5.75	0.07	4.55
Fusion	63	6	30	12.39	11.89	10.67	6.79	0.06	5.38
	127	16	92	14.62	13.48	12.89	7.36	0.01	8.70
	255	32	179	15.15	13.48	12.48	6.98	0.02	6.83
	511	40	220	16.09	14.48	13.48	6.28	0.01	7.25

Table 4.2: Security evaluation of protected ensembles for different binary template sizes n.

4.6 Discussion

This section further discusses the results of the security analysis we performed for unprotected and protected systems.

4.6.1 On the security of the unprotected systems

We further discuss the impact of our security evaluation on the unprotected system. The security analysis focused on two threat models, i.e., the traditional and biometric

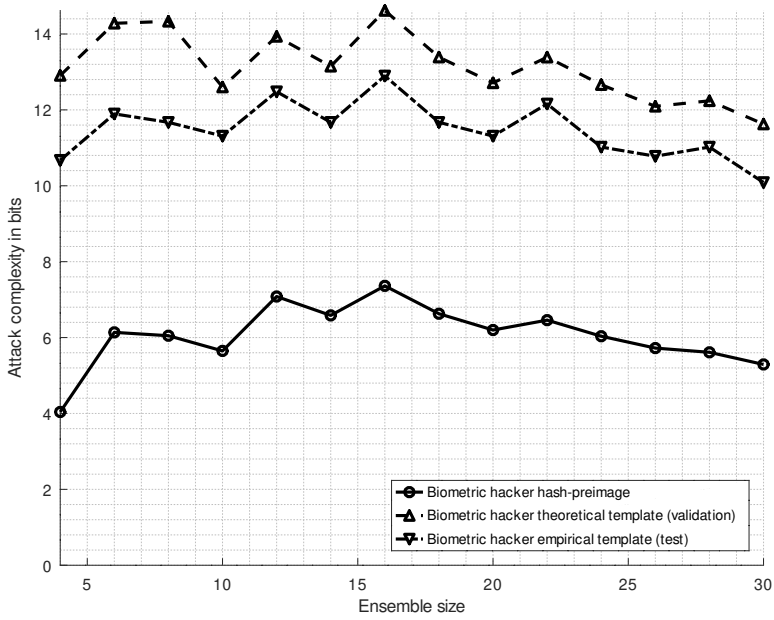


Figure 4.4: Complexity of biometric hacker attacks on mixed ANN and HMM ensembles with binary template size $n = 127$

hacker. Analogously we further discuss both.

First, to perform the zero-effort attack, we construct nine (5 HMMs and 4 ANNs) distinct biometric template extraction methods with competitive performance compared to the state-of-the-art. The OU-ISIR database is recorded in lab conditions, which introduces a bias in the evaluation concerning the real-worldness of the conditions. However, the open-world experimental protocol ensures that there is no overlap between individuals in training, evaluation, and test sets, which makes the experimental results a good representation of the performance of this biometric trait under these acquisition conditions. We can see that the a priori HTER is slightly worse in general than the a posteriori EER, which is in the best case less than 1% (GRU network trained with arcface loss). This highlights the importance of reporting a priori performance figures since the systems' setup is based on limited data. Nevertheless, the improvement in biometric recognition performance obtained by fusing systems is remarkable, with test FAR of 0.63%, 0.21% and even 0% at FRR 5%.

Second, the biometric template correlation attack, executed by the biometric

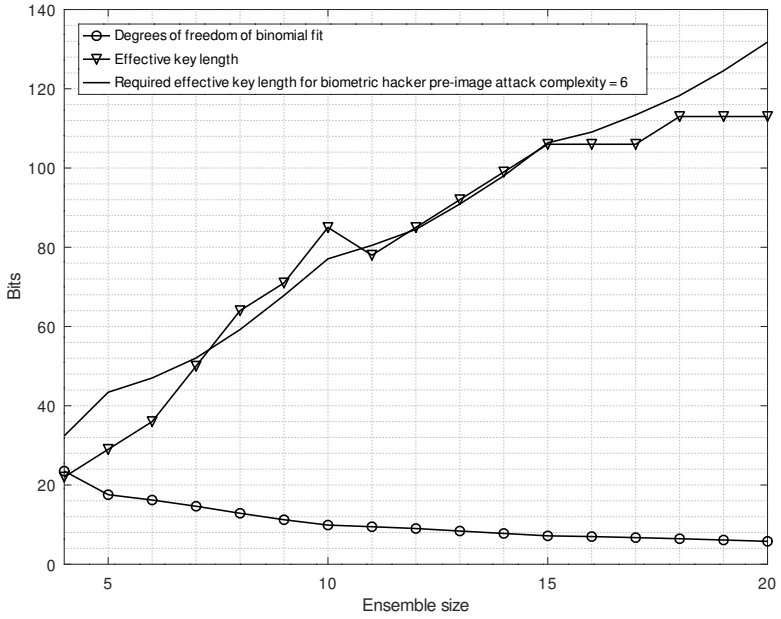


Figure 4.5: Degrees of freedom of binomial fit of impostor attempts Hamming distance, maximum effective key length for FRR $\leq 5\%$, and required effective key length for a biometric hacker hash preimage attack complexity bounded above 6 bits for HMM ensembles and binary template length 127.

hacker, clearly demonstrates that leaking biometric templates does not only affect authentication systems that use the same feature extractor but affects all future uses of the leaked biometric trait. Moreover, history has proven that credential databases are leaked frequently. As such, it would be naive to assume that templates do not leak. Furthermore, if template leakage is the norm, the security of the biometric authentication system relies solely on the assumption that system input can be trusted. Therefore, storing biometric templates in a protected form that does not require reconstructing the original biometric data for biometric template matching is essential. Lastly, due to the high success rate of the biometric template correlation attack, even if biometric templates are protected with, for example, an SFE, biometric template leakage from an unprotected system will reduce the security of the protected SFE system to the security of the fixed authentication factor.

The ensemble systems relying on score fuse seem not more resilient to biometric template correlation attacks. Moreover, since the mean IRR is lower than the FRR,

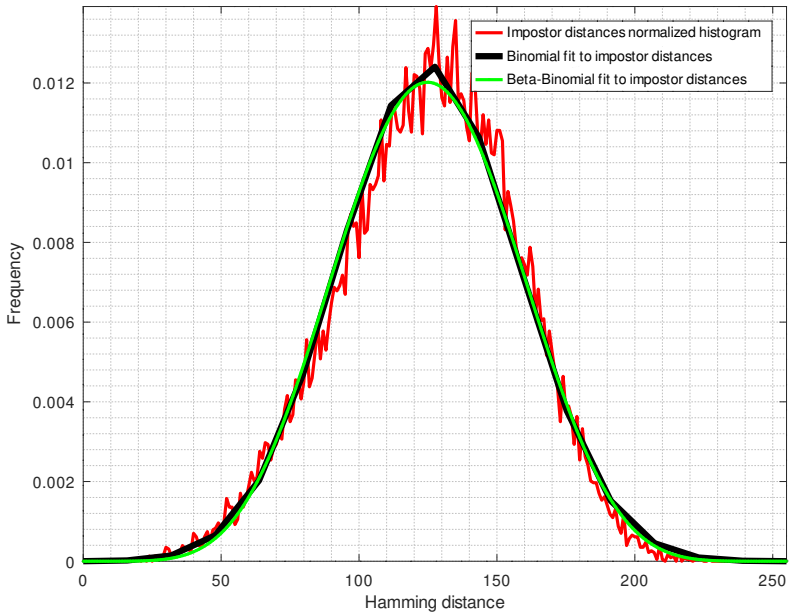


Figure 4.6: Histogram of Hamming distances of impostor attempts in validation sets and its corresponding binomial and β-binomial fits, for $n = 255$, and the ensemble of $E = 8$ neural networks.

while for the individual systems, IRR is higher than the FRR, we found evidence that the ensemble systems are more accessible to attack than the individual systems. As the biometric template correlation attack is already very successful, we did not execute any attacks where the ensemble system is leaked, which is at least as effective as leaking a single system and probably even more effective because the attacker has more information to estimate the target template. A small not on generalizability of this observation: when composing the ensembles differently, e.g., bagging, the results might be drastically different.

4.6.2 On the security of the protected system

The performed security analysis of the protected systems and its quantitative experimental evaluation shows some relevant results

1. The biometric hacker attack on the biometric template becomes eventually easier for a given binary template length for bigger ensembles, i.e., its complexity may initially grow due to the compelling evidence provided by adding systems to the ensemble. However, as the ensemble size grows, these contributions get compensated by the strong correlations between the bits in the binary representation.

2. There exists ensemble sizes that provide maximum security working points, taking into account both biometric hacker template and hash preimage attacks.

3. The provided security slightly increases when ANN and HMM systems are mixed in the ensemble for a given complexity, but this increment is limited.

4. The ensemble size providing maximum security grows rapidly with the binary representation length. However, the provided security does not significantly increase with it, so the growth in system complexity is not justified from a security point of view.

The analyzed attacks provide a more complete view of the security of the protected system. The effective key length only provides an accurate view (it is equivalent to $\mathcal{C}_{\text{hacker}}$) of the fuzzy commitment security when bit independence holds. Although the analyzed systems incorporate evidence from several gait authentication systems, it must be noted that these analyses are also relevant for mono-expert systems, with no intramodal fusion, since correlations among the biometric template's bits may still be present, as shown by Daugman [63]. $\mathcal{C}_{\text{hacker}}$ describes the strength against an adversary performing a brute force attack on the hash. The hash should be stored and checked remotely, decreasing the risk of this attack. In this case, we need to evaluate $\mathcal{C}_{\text{bio-hacker}}$, which describes the strength against an adversary with knowledge of the biometric distribution and having access to the decrypted helper data of an SFE, containing the xor of a codeword and the biometric binary representation. Since the impostor can only perform a limited number of consecutive failed attempts without raising an alarm, this setup is comparable to the one accepted for authentication factors such as PINs. A 4-digit PIN provides maximum security of $\log_2(10^4) \approx 13.29$ bits if uniformly random distributed. However, due to non-uniformity this is notably not the case in the real world [251]. For instance, it is shown that the success probability of 4-PIN and 6-PIN digits on one guess is around 3% [148], which is roughly equivalent to $\log_2(0.03) \approx 5$ bits. Also, the proposed bound on the hash preimage attack is a theoretical bound, and it is still to be determined how close to this bound a practical attacker can get. When we consider only the biometric hacker template attack, the security provided by the tested biometric system is approximately the same as that provided by an ideal randomly distributed 4-digit PIN.

Although the main goal in our threat model for protected systems is impersonation, by doing this, the adversary gets valuable information on the biometric template.

Moreover, as shown in the correlation attack for unprotected systems, this also poses a severe threat to the security of any template extracted from the same trait, making the revokability of templates only rely on the fixed factor in SFE systems.

As a final remark, the SFE-protected biometric system is a two-factor authentication scheme. To study the security the biometric modality offers, we assumed in our threat model that the second factor is broken. This assumption is reasonable because standard attack vectors exist for most of the widely deployed authentication methodologies [34], e.g., passwords are very prone to phishing or re-use, security tokens get lost or stolen, etc. Moreover, if we assume that a password protects the second factor, it has been shown that with an offline dictionary attack, 50% of passwords offer less than 20 bits of security [33]. This further proves that a broken second factor is more than reasonable. On top of that, it further motivates the appreciation for the security the biometric modality offers.

4.6.3 On generalizability of the security analysis

This research aims to perform a security analysis of a biometric authentication system protected with fuzzy commitment schemes. Our analysis considers the correlation between the bits of the key produced by the system. There are two primary sources of correlation: 1) the binary features extracted from the fuzzy data, and 2) the ECC code used to correct errors. Our security analysis should be performed for any biometric cryptosystem dealing with these sources of correlation. However, how they exactly influence the metrics requires a new analysis.

In our case, the fuzzy data source is an intramodal fusion of IMU-based gait. Our primary motivation for performing intramodal fusion for a single modality (and single capture) is to precisely analyze the influence of correlation in mono-modal systems. A similar analysis is relevant for other biometric modalities and fusion strategies. Prior work demonstrated that other biometric modalities such as iris and face [63, 202] exhibit statistical properties similar to IMU-based gait. On top of that, other fusion strategies will lead to different amounts of correlation.

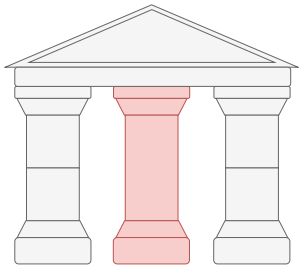
Other biometric cryptosystems such as fuzzy vault [118] also rely on ECC codes; thus, the same security analysis can be executed. However, because fuzzy vault uses a specific binarization that deals with variable length feature sets, the biometric hacker must perform statistical analysis for the fuzzy vault scheme.

4.7 Conclusion

In this work, we have analyzed the security of biometric authentication systems using monomodal information captured by discriminative and generative authentication algorithms, both for unprotected templates relying on hardware security features and for protected templates in a remote authentication architecture.

For the unprotected system, we have shown that a biometric hacker can successfully use correlations between templates in different systems to forge templates as effective as genuine ones. This concerns the leakage of templates, which affects future usages of these biometrics. The use of ensembles does not help to mitigate this attack. A prophylactic measure that limits the risk associated with template leakage is the use of template protection schemes integrated with fixed factors since the fixed factor would offer another security layer against sophisticated template regression attacks when templates are leaked and reduce the risk of this leakage to happen.

We have also thoroughly analyzed the security for protected biometrics, specifically SFEs derived from fuzzy commitment. We have shown new attacks that can be performed by an adversary knowledgeable of biometrics, taking into account internal template correlations and establishing bounds and approximations for the complexity of these attacks. From these attacks, we derived new security measurements. Furthermore, we analyzed the behavior of these security figures with respect to the size of the intramodal ensemble, showing that security does not, in general, grow with the ensemble size, thus showing that there is complexity for a given biometric trait where the increase in system complexity is not justified in security terms. The attack complexity measures proposed in this paper provide valuable information that must be considered for a secure design of authentication systems based on protected biometric templates, either in the case of ensembles or mono-expert systems.



Chapter 5

Masterkey attacks against free-text keystroke dynamics and security implications of demographic factors

Preamble

This chapter presents and systematically evaluates the first masterkey attack against free-text keystroke dynamics. A masterkey is a typing sequence that matches, hence successfully impersonates, a large part of the population. Therefore, masterkeys are effective tools for an adversary who aims to impersonate someone without knowledge of their typing behavior. On top of the attack itself, we present a new unifying evaluation framework for masterkey attacks that allow for the comparison with knowledge-based authentication factors. In other words, we unify the evaluation of password security with that of masterkey attacks and demonstrate that, for a skilled adversary, typing biometrics is approximately 20 times less secure than passwords and two times less secure than a 4-digit pin. Lastly, we study the effect of demographics on typing biometrics, which, among others, provides novel insights into the effect of being a well-versed typist on security.

The material covered in this chapter originates from our research paper on biometric masterkeys on free-text keystroke dynamics [243]. This research is an excellent example of how offensive work improves our understanding of the defensive prowess of

security solutions. The concept of biometric masterkeys allows us to pivot the security narrative for biometrics from *"How easily can fake biometric data be generated?"*, to *"How hard is it to guess biometric data?"*. This new viewpoint allows us to compare the security offered by biometrics with that offered by knowledge factors. Because this chapter primarily describes and evaluates masterkey attacks against a behavioral authentication system, it contributes to the 'novel attacks' pillar. Nevertheless, this work served as the primary inspiration to evaluate the security offered by biometrics from a different perspective. An idea that is further pursued in Chapter 6.

Publication Data: T. Van hamme, G. Garofalo, D. Preuveneers, and W. Joosen. "Masterkey Attacks Against Free-Text Keystroke Dynamics and Security Implications of Demographic Factors". In: *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. Los Alamitos, CA, USA: IEEE Computer Society, July 2023, pp. 278–291. DOI: 10.1109/EuroSP57164.2023.00025

5.1 Introduction

A masterkey is a key that opens multiple locks. Recently, this concept received lots of attention in biometrics due to its prowess of impersonating users without knowledge about the victim's appearance. For example, a masterface [171, 221] is a face image that matches a surprisingly large part of the population, making it an exciting candidate to break a face recognition system. Moreover, this concept is not constrained to the face domain; examples exist for fingerprints [36] and keystrokes [218, 167]. The fact that biometric recognition performance is not uniform among individuals was already demonstrated in 1998 by Doddington [71]. Doddington identified different categories of users; among them are wolves, i.e., those users that tend to impersonate others more easily. A decade later, in 2007, the concept of wolf attacks was first introduced [238]. A wolf attack comprises selecting wolf samples from a given dataset to achieve a high probability of impersonating others. Masterface and masterprint attacks go one step further and create synthetic samples to improve the attack.

Masterkey attacks pose an important threat to behavioral authentication systems because preventing spoofing is significantly more complicated compared to conventional biometric authentication solutions: First and foremost, behavioral biometrics typically rely on general-purpose hardware. As such behavioral authentication systems cannot rely upon dedicated secure hardware widely deployed in traditional biometric authentication systems such as phone unlock. Second, since imitating behavior is complex [162], liveness detection – presentation attack detection – systems are less prevalent. Third, use cases for behavioral biometrics go beyond the traditional local unlock scenarios for which mainstream biometrics are popular. In other words, they

trigger a paradigm shift from local to global, manifesting in behavioral authentication products being implemented as APIs[1].

A clear example of a relevant behavioral identity verification system is keystroke dynamics. Companies like TypingDNA[2] and Behaviosec[3] provide keystroke authentication services for various applications, such as fraud prevention, proctoring, and continuous and multi-factor authentication. Given the increasing significance of keystroke biometrics, prior research has developed masterkey attacks against keystroke dynamics when used as a frictionless second factor for password authentication [218, 167]. These attacks assume fixed typing sequences and target machine learning systems that can only handle those fixed-length sequences. However, modern solutions can also manage text-independent typing sequences [4]. Moreover, the metrics used in previous work on masterkey attacks don't allow for a comparison with traditional knowledge-based authentication methods. Nevertheless, such a comparison is relevant because masterkey attacks are essentially intelligent dictionary attacks, which are familiar in the context of password cracking. This raises the question of how biometrics measure up against knowledge-based factors. This study addresses these two gaps. First, it broadens the focus from fixed-text single-shot keystroke authentication to free-text single-shot keystroke authentication. Second, it assesses keystroke biometrics using a novel evaluation framework that enables a direct comparison with knowledge-based authentication factors.

The contributions of this work can be summarized as follows:

1. We design, implement, evaluate, and successfully execute masterkey attacks against free-text single-shot keystroke authentication.
2. We propose an evaluation framework that unifies security metrics for knowledge-based factors and single-shot biometric authentication systems to compare their strengths systematically.
3. We perform a large-scale study and offer new insights into the effect of demographics on the security level offered by single-shot keystroke biometrics.

5.2 Background and related work

In this section, we describe the authentication system pipeline and describe the related work on keystroke dynamics. Afterward, we describe the threat surface of biometric

[1] https://info.behaviosec.com/behaviosec-developer-portal-page
https://api.typingdna.com/

[2] https://www.typingdna.com/

[3] https://www.behaviosec.com/

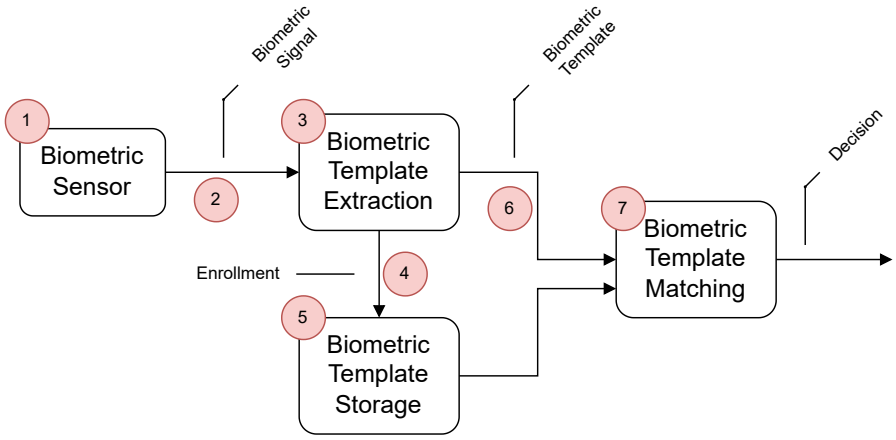


Figure 5.1: Typical biometric pipeline. The four main attacks target are the sensor (1), the interfaces (2,4,6), the processing blocks (3,7), and the storage (5).

authentication systems, which allows us to position the related work on masterkey attacks. Lastly, we describe the related work on wolf/masterkey attacks.

5.2.1 Biometric authentication pipeline

Biometric authentication aims to assert a claimed identity based on evidence inherent to the individual, e.g., a face image, fingerprint, gait acquisition, sequence of keystrokes, et cetera. In contrast to knowledge and possession factors, verifying this proof requires fuzzy matching rather than exact matching. As such, it is possible to make accidental matches (type I errors) and wrong non-matches (type II errors).

The biometric authentication pipeline consists of two phases and four main building blocks. The enrollment phase populates the biometric database with reference samples. After enrollment, the system verifies identity claims by matching new biometric acquisitions to the reference samples. The process involves a sensor to capture biometric samples and a feature extractor that transforms raw biometric data into compact identity representations. Figure 5.1 depicts the typical biometric pipeline.

Sensor The sensor is the interface between the physical and digital worlds. It captures the physical properties of persons and represents them digitally. We denote the digital biometric as $b \in B$, where b refers to a specific biometric signal and B to the space of all biometric signals.

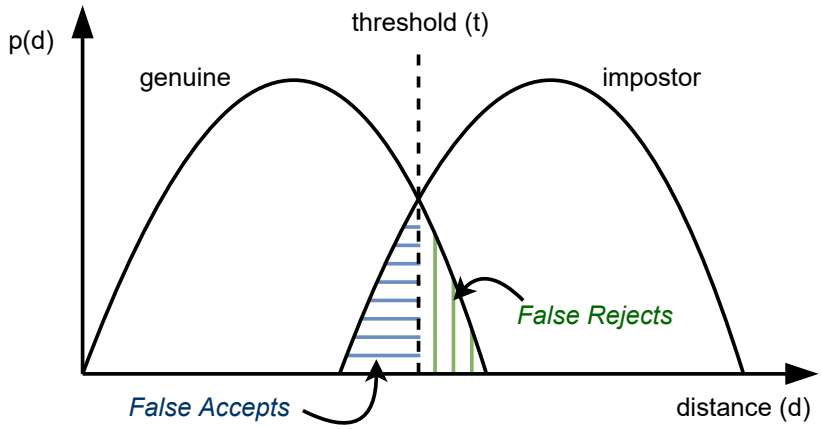


Figure 5.2: Trade-off between false accepts and false rejects when selecting threshold t.

Feature extractor $f(\cdot)$ The biometric feature extractor extracts the discriminative information from the biometric signal. The feature extractor is often a learned function $f(\cdot)$. It maps biometric signals from input space B to the n-dimensional feature space $f : B \rightarrow \mathbb{R}^n$: $x = f(b)$. In this work, we normalize the feature vectors $\|x\| = 1$. The vector x goes by many names, e.g., feature vector, biometric embedding, biometric template, and identity vector.

The state-of-the-art feature extractors are primarily advanced in face recognition [252]. Nowadays, deep learning architectures are trained by some margin-based loss functions such as the triplet loss [213] or arcface loss [66].

(Biometric) Template database The template database stores the biometric templates obtained during enrollment $x_j^e = f(b_j^e)$. There are multiple ways to combine the enrollment samples to perform the matching later. A popular one computes the mean of the biometric enrollment vectors to obtain a biometric centroid $c_j = \frac{1}{E_j} \sum_{i=1}^{E_j} x_j^i$. The subscript j refers to the j^{th} individual, and E_j refers to the number of enrollment samples for person j.

Matcher $m(\cdot, \cdot, \cdot)$ The matcher performs biometric matching and answers the question $d(f(b), c_j) < t$. $d(\cdot, \cdot)$ is a distance function, i.e., cosine distance, c_j is the centroid of individual j and t is a system parameter called the threshold.

$$m(f(b), c_j, t) = \begin{cases} 1 & \text{if } d(f(b), c_j) < t \\ 0 & \text{if } d(f(b), c_j) > t \end{cases} \tag{5.1}$$

The threshold trades off the number of false accepts and false rejects. False accepts pose a security risk, while false rejects hamper the user experience. Figure 5.2 shows the trade-off between false accepts and false rejects.

We keep an overview of the mathematical notation in Table A.1.

5.2.2 Keystroke biometrics

Forsen et al. [82] introduced the idea that typing is a unique property. However, the first critical evaluation of typing behavior for authentication was performed in 2009 by Killourhy and Maxion [125]. Killourhy and Maxion contributed the first benchmark dataset (DSN-2009) for fixed-length keystroke dynamics, which they used to benchmark different anomaly detection algorithms. Their dataset contained 400 typing sequences of 51 users that typed the password '.tie5Roan1'. They demonstrate a best-case equal error rate (EER) of around 9.6%.

In 2016 the keystroke biometrics ongoing competition (KBOC) [160] promoted new work to improve the state-of-the-art for fixed-text keystroke biometrics further. The competition winner Vinnie Monaco further improved the state-of-the-art by an additional pre-processing step which normalizes the input samples based on the genuine distribution.

Free-text biometrics have also been studied since 2009 by Tappert et al. [232] who collected the first large-scale dataset: 144 users provided at least five samples each in different recording sessions. Samples contain an average of 746 keystrokes, mixing free-text and fixed-text input. Monaco et al. later used this dataset to develop a continuous authentication system [157]. Finally, Acien et al. [4], under the supervision of Monaco, scaled the evaluation to a more realistic number of 100k users, and applied more novel deep learning techniques.

Keystroke biometrics have also been studied for mobile platforms. The original DSN dataset [125] also has a mobile equivalent, i.e., the mobile-DSN [14]. Mobile keystroke dynamics ML models are very similar to physical keyboard dynamics, with the difference that there are no key release timing data [143]. To accommodate this data loss, IMU sensor data is typically included. The advantage of IMU sensor data is that it works for other ways to operate keyboards, such as swiping. However, IMU sensor data also complicates pre-processing and is even more dependent on the context of the device, i.e., is the mobile device held in hand, or does it lie flat on the table?

5.2.3 Single-shot and continuous authentication

Behavioral biometrics, such as free-text keystroke dynamics, is compelling because behavior can be captured in the background. As such, behavioral authentication introduces no user friction, and the authentication process becomes transparent. Moreover, because the authentication process has become imperceptible, it can be repeated continuously. Consequently, behavioral authentication allows for continuous authentication. Unlike traditional single-shot authentication, continuous authentication constantly repeats the authentication process during a (privileged) session. When a session is considered risky, the session is invalidated, and the user is prompted for step-up authentication.

A continuous authentication system relies upon the biometric recognition systems used for single-shot authentication. They can be implemented as a tumbling window [161], where the authentication process is repeated for non-overlapping time windows in which the behavioral data is collected. Different authentication decisions can be fused before deciding upon session termination. Alternatively, a sliding window can be used. As the biometric recognition model outputs are highly correlated, they are used to increase or decrease a risk level [68]. When the risk level is too high, the session is invalidated.

The performance of continuous authentication hinges on the performance of the biometric recognition model. Therefore, related work on continuous authentication often mentions the traditional single-shot authentication performance metrics such as false match (false acceptance) rate, false non-match (false rejection) rate, and equal error rate [228]. However, as argued by Mondal and Bours, more useful metrics are related to how quickly sessions are invalidated correctly and how often this phenomenon occurs incorrectly [159, 37, 158]. Within the scope of this work, we primarily discuss free-text keystroke dynamics as a single-shot authentication methodology, which allows for direct comparison with password and pin-code security. This is a necessary step to describe the security of continuous authentication systems. Section 5.7.2 discusses extending our work to accommodate continuous authentication.

5.2.4 Threat surface of biometric systems

Biometric authentication systems rely on fuzzy matching, which leads to accidental false matches that pose a security risk. On top of that, biometric authentication systems have an elaborate attack surface. The attack surface is easily grasped by considering the biometric system building blocks as shown in Figure 5.1. Biggio et al. [30] determine four major attack locations shown in Figure 5.1, i.e., attacks to the sensor, the signal processing algorithms, the biometric template database, and the interfaces between them. Standard security practices such as encrypted channels,

specialized hardware, and database encryption secure the integrity of communication channels, the template database, and processing algorithms. The remaining attack location is the sensor which is subject to presentation attacks. There is no silver bullet solution for presentation attack detection, which makes it an active research area [192, 108]. Hence, assuming that the input to the biometric system comes exclusively from trusted sensors, its security is mainly described by the effectiveness of the deployed anti-spoofing solution and the probability of errors inherent to the fuzzy matching. This assumption is also clearly stated in many different biometric security evaluation standards, e.g. ISO/IEC [113, 114], Common Criteria [57], FIDO [78], and even Android [155]. Therefore, current security evaluations mainly focus on a zero-effort attacker who accidentally obtains access to the system by using their biometric signal and a strong adversary who knows the biometric signal of their victim and uses this information to generate presentation attacks. They omit a threat model in which a masterkey/wolf sample can be exploited. Such a threat model breaks the assumption that an adversary can spoof arbitrary input but assumes no adversary knowledge of the biometric signal of the victim. As we argued in Section 5.1, this threat model is very relevant for keystroke biometrics, as they rely on typing input from the browser. Therefore, keystrokes can be spoofed on the browser or OS levels. This fact has been exploited by Rahman et al. [190] in a replay attack. In this attack, the adversary observes the typing behavior of their victim and (optionally) generates new typing sequences from them. This threat model differs from masterkey attacks, described in the following section, where the adversary has no such knowledge.

5.2.5 Masterkey attacks

Doddington [71] first introduced the fact that biometric recognition performance is not uniform among individuals. Serwadda et al. [218] exploit this fact in a masterkey attack against fixed-text keystrokes. They guess the typing behavior of their victim by studying the typing behavior of other users typing the same password. Their guess successfully impersonates 5 to 30% of the population. One of the main drawbacks of this attack is that it requires access to data from a considerable population typing the same password. Negi et al. [167] improve upon this by estimating bigram timings of the password from a large corpus of free typed text. Their algorithm uses k-Means++, the initialization algorithm of k-Means clustering, to determine optimal guesses for the keystroke timings.

Masterkey attacks have also been introduced in the domain of fingerprint [36] and face recognition [171, 221]. The complexity of attacks in these domains is higher as the machine learning models are more complicated than the fixed-text biometric models. These attack methodologies rely on black-box optimization methods to find masterkeys and on GANs to make them look realistic. However, to run the black-box optimization, it is assumed that the adversary has access to the model or can make many queries.

Furthermore, the adversary also knows the population they want to generate masterkeys for.

5.3 Masterkey attack for free-text keystroke dynamics

One of our key contributions is a masterkey attack for free-text keystroke dynamics. In this section, we first formalize the threat model of our attack. Second, we describe how masterkey samples are constructed.

5.3.1 Threat model

The adversary executing a masterkey attack has the **goal** of impersonating an arbitrary victim. The adversary has the **capability** of spoofing arbitrary key timings to the system. The adversary has **knowledge** of the typing behavior of the general population. Specifically, the adversary has access to a large dataset with key timings of people typing free text. Furthermore, the users in their dataset differ from those enrolled in the system they want to compromise. Another way to look at this attack is as a dictionary attack. The adversary aims to create a dictionary of the most probable typing behaviors, which they use to perform a brute-force attack.

The threat model above describes the most realistic attack. However, there is value in relaxing this threat model to gain insights into the maximal effectiveness of these attacks and the impact of acquiring more knowledge. Therefore, in this research, we start from a stronger threat model and relax it until the ideal threat model is reached. We distinguish between four adversary models:

1. The **insider adversary** knows the target keystroke biometrics model and the enrolled users and can execute their attack in feature space. In other words, they do not have to find keystroke timings. Therefore, this adversary model is only realistic for an insider.

2. The **feature space adversary** knows the exact keystroke biometrics model and executes the attack in feature space. This adversary does not know the enrolled users. This adversary model is relevant if the biometric system is set up to extract features client-side to ensure that the typed text remains unknown to the biometric system provider.

3. The **input space adversary** performs their attack in input space with knowledge of the target keystroke model and the enrolled users. We also describe this

adversary to quantify the effect of optimizing in input space rather than feature space.

4. The **pragmatic adversary** performs their attack in input space without knowledge of the target model nor the enrolled users. Therefore, this threat model is the weakest and most probable.

5.3.2 Attack methodology

The attack described above requires the adversary to understand how an average user types. Once they have this knowledge, they need a procedure to find or create sequences matching the typing behavior of as many individuals as possible. We will denote this more formally using the expressions introduced in Section 5.2.1. Table A.1 provides an overview of these notations.

Each adversary has access to typing sequences b_j^i with $j \in [1, U]$ the user number and $i \in [1, N_j]$ the sample number. U denotes the number of users the adversary has typing sequences of, and N_j denotes the number of samples of user j. Furthermore, the adversary also has access to a feature extractor $f(\cdot)$ that embeds typing behavior into an identity vector. Based on this information, the adversary proceeds as follows: First, they compute a representation for each user $c_j = \frac{1}{N_j} \sum_{i=1}^{N_j} b_j^i$. As such, the adversary has an approximation of the distribution of the typing behavior in the feature space. Note that the *insider adversary* knows the exact centroids of the enrolled users and thus does not have to approximate this distribution. Second, the adversary looks for masterkey samples that compromise the maximum amount of users, i.e., a sample mk that maximizes $\max_{mk} \sum_{j=1}^{U} m(c_j, f(mk), t)$ with $m(\cdot, \cdot, \cdot)$ a biometric matching function. The biometric matching function returns 1 if there is a match and 0 in case of no match. The matcher function can be tuned based on the threshold parameter t. Of course, when the masterkey is in the feature space, the function to optimize becomes $\max_{mk} \sum_{i=1}^{U} m(c_j, mk, t)$.

The easiest way to approximate this function is to select the sample $mk = b_j^i$ that maximizes it. However, this limits the search space to the collected typing sequences. We call this the wolf attack, as we select wolf samples rather than create them. Another simple approach that only works in feature space is performing k-Means clustering [171]. The main parameter of this attack is the number of clusters k. This number also determines the number of masterkeys found.

Prior work on masterkey attacks in the face [171, 221] and fingerprint [36] domain use a black-box (gradient-free) optimization approach. A successful and well-tested one is an evolutionary algorithm called covariance matrix adaptation (CMA-ES) [102, 101]. This algorithm only works on continuous values; thus, it can be used in feature space.

We can also use it in input space, but because keycodes are discrete, we must select the sentence to be typed ourselves.

The challenge is not to find one masterkey but a sequence of masterkeys such that the number of accounts compromised is maximized for a specific number of tries. As discussed in Section 5.4.3, this is an NP-hard problem. Therefore, we opt for a brute force approach where the optimization process is repeated with the centroids of accounts compromised by earlier masterkeys dropped.

5.4 Guessing metric evaluation framework

Our evaluation framework starts from the observation that an adversary who tries to crack passwords and the masterkey adversary – described in Section 5.3.1 – are very similar. These similarities inspire us to adapt the evaluation metrics for password security to biometrics. The evaluation metrics for passwords quantify the uncertainty a brute force adversary faces when trying to guess the correct authentication secret, i.e., the password [33]. As such, they are also called guessing metrics. In the remainder of this section, we draw the parallel between a brute force adversary who tries to guess passwords and the masterkey adversary. Afterward, we introduce the guessing metrics and describe our adaptation for biometrics. Finally, we discuss applying them to practical security evaluations of authentication systems.

5.4.1 Brute force adversaries

The brute force adversary tries to unrightfully obtain access to a system by guessing. In the case of knowledge factors, they try to guess the secret that would be accepted by the authentication system, e.g., a pin or password. The expected efficiency of brute force attacks is measured by guessing metrics defined in Section 5.4.2. To ensure that these metrics are unbiased, comparative, and repeatable, the adversary is assumed to know the exact distribution of the secrets. As such, they can make optimal guesses.

More formally, in the case of knowledge factors, guessing metrics capture the efficiency with which an attacker who knows the distribution of secrets χ can determine the secret x from a set of unknown secrets $\{X_1, X_2, \ldots, X_k\}$ formulating queries of the form "$X_i = x$?" to an oracle [33] – where $x \in \chi$ denotes a specific event (e.g. password). The probability of each event x is denoted as p_x, with $0 < p_x \leq 1$ and $\sum p_x = 1$. We rank events from most likely to least likely and encode this information in the subscript: the most likely event is x_1 with probability p_1 and least likely event is x_K with probability p_K, thus $p_1 \geq p_2 \geq \cdots \geq p_K$. This ordering depicts the optimal guessing order. K also denotes the number of unique events.

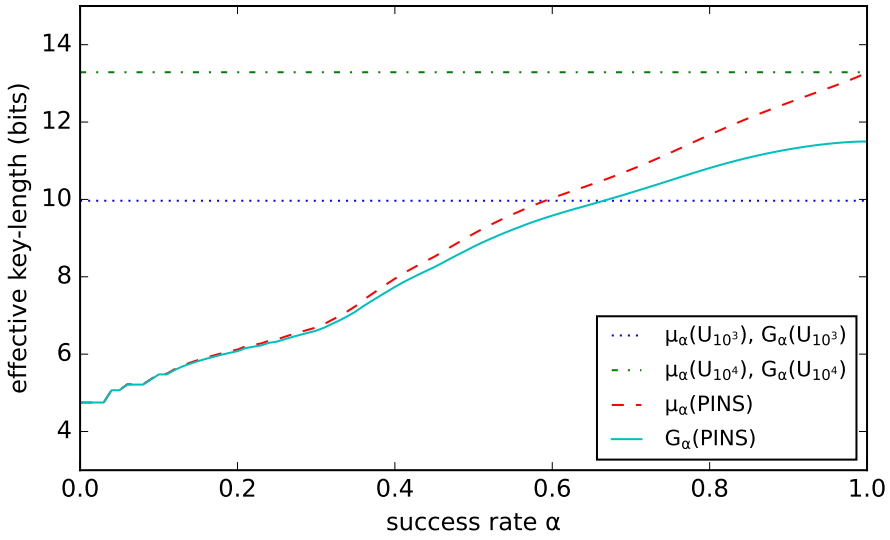


Figure 5.3: Comparison of guessing difficulty between two uniform probability distributions of 10^3 and 10^4 users, and the probability distribution of 4-digit PINs [35].

Determining an optimal guessing order for biometric authentication based on the frequency of occurrence is impossible for the following reasons: firstly, $x \in \chi$ events are no longer discrete, they are continuous-valued feature vectors, hence uncountable. Secondly, biometric matching is fuzzy. Instead of trying to determine how efficiently an adversary can guess a secret X_i while having access to an oracle that answers "$X_i = x$?", the oracle now replies to the question "$d(f(b), c) < t$?". These two aspects complicate finding an optimal guessing sequence and estimating the attack success rate p_x of guess x. In Section 5.4.3 we overcome these challenges by interpreting p_i as the attack success rate of the i^{th} guess and devising a sub-optimal but pragmatic guessing strategy.

5.4.2 Guessing metrics for knowledge factors

The most important guessing metrics are described below, and their relationship is depicted in Figure 5.3. These metrics do not depend on the number of unique events K. Thus, they converge quickly with an increasing sample size of the evaluation set U drawn randomly from χ: $X \xleftarrow{R} \chi$.

Min-entropy The min-entropy [203] defined as

$$H_\infty = -\log p_1 \tag{5.2}$$

depends solely on p_1 and therefore reflects security against an attacker who only guesses the most likely secret, thereby providing a good worst-case security metric.

β-success-rate The β-success-rate [38] measures the expected success for an attacker limited to β guesses per account.

$$\lambda_\beta(\chi) = \sum_{i=1}^{\beta} p_i \tag{5.3}$$

α-work-factor α-work-factor [187] is the number of guesses needed by an attacker with an optimal dictionary to break at least $\alpha * 100\%$ of the accounts. For example, if α is 0.25 the adversary can break 25% of accounts with α-work-factor guesses. Alternatively, the adversary has a probability of 25% to break an arbitrary account.

$$\mu_\alpha(\chi) = \min\left\{ j | \sum_{i=1}^{j} p_i \geq \alpha \right\} \tag{5.4}$$

α-guesswork The metrics above do not take into account that an attacker can stop early after a successful attempt. Therefore, α-guesswork [33] gauges the expected number of guesses per account to effectively break a fraction α accounts.

$$G_\alpha(\chi) = (1 - \lambda_{\mu_\alpha})\mu_\alpha + \sum_{i=1}^{\mu_\alpha} p_i \dot{i} \tag{5.5}$$

Note on Shannon entropy and other metrics Although Shannon entropy [219] is a widely used measure of information (or uncertainty), it has no value in assessing guessing difficulty [33] since it estimates the average number of *subset queries* needed to identify *X*. Furthermore, Bonneau [33] demonstrated that Hartley entropy [203] and guesswork [153] are unworkable as they depend on the number of unique events N.

Effective key length Computer scientists prefer to assess security on a logarithmic scale, i.e., in bits, because this enables a better understanding by directly linking to attacker workloads for guessing cryptographic keys. As a result, guessing metrics are converted to bits by computing the logarithm of the size of a discrete uniform distribution that yields the same guessing metric value [33]:

$$\tilde{\lambda}_\beta(\chi) = \log\left(\frac{\beta}{\lambda_\beta(\chi)}\right) \tag{5.6}$$

$$\tilde{\mu}_\alpha(\chi) = \log\left(\frac{\mu_\alpha(\chi)}{\lambda_{\mu_\alpha}}\right) \tag{5.7}$$

$$\tilde{G}_\alpha(\chi) = \log\left(\frac{2G_\alpha(\chi)}{\lambda_{\mu_\alpha}} - 1\right) + \log\left(\frac{1}{2-\lambda_{\mu_\alpha}}\right) \tag{5.8}$$

For reference, a 4-digit pin code offers ≈ 13.3 bits of security when the pin codes are uniformly distributed. We added the newly introduced symbols to Table A.1.

5.4.3 Guessing metrics for biometrics

Passwords and biometrics differ primarily in two ways. The first distinction is that $x \in \chi$ events are no longer discrete, hence uncountable. The events are instead biometric signals from which continuous valued feature vectors are derived using a learned function $f(\cdot)$. We refer the reader to Table A.1 for an overview of mathematical notations. The second distinction is that biometric matching is fuzzy. Instead of trying to determine how efficiently an adversary can guess a secret X_i while having access to an oracle that answers "$X_i = x$?", the oracle now replies to the question "$d(f(b_i), c) < t$?". Here the threshold t and the distance function $d(\cdot,\cdot)$ are system parameters, and c corresponds to the biometric representation of a specific person. This representation is the mean of the enrollment samples to which we refer as the centroid $c_j = \frac{1}{E_j}\sum_{i=1}^{E_j} f(b_j^i)$, with E_j the number of enrollment samples for person j.

Sub-optimal guessing strategy We want to devise a guessing strategy based on the probability that a guess satisfies the query, i.e., that biometric sample b satisfies "$d(f(b), c) < t$". Additionally, samples can meet this constraint for many individuals. In other words, every biometric sample matches a set of individuals. Thus, the aim is to maximize the number of covered individuals by selecting at most a given amount of sets. This problem is known as the maximum coverage problem and is NP-hard [104]. More formally, the maximum coverage problem aims to find a subset $S' \subseteq S$ of a given collection of sets $S = \{S_1, S_2, \ldots, S_m\}$ that maximizes the elements covered $|\bigcup_{S_i \in S'} S_i|$, under the constraint that the cardinality of the subset is less or equal to a given number $|S'| \leq k$. The best-known polynomial time solution is the greedy algorithm. The greedy algorithm selects, at each stage, the set with the largest number of uncovered elements. Then, the probability p_i for the i^{th} guess is the attack success rate of that specific guess, i.e., the number of uncovered accounts impersonated by this guess divided by the total number of accounts. This algorithm is formulated in Algorithm 1.

5.4.4 Parameterization of guessing metrics

The adversary capabilities determine the most relevant values for the parameters of the guessing metrics, i.e. the number of guesses β, and the fraction of accounts to break α. To determine these parameters for password cracking, they differ between online (*throttled*) and offline (*unthrottled*) attacks. Online adversaries are restricted in the number of guesses. The exact number depends on the rate-limiting strategy. Prior work on passwords concluded 10 guesses is reasonable [33], hence λ_{10} is considered. In the case of phone unlock with pin there is an exponential time penalty for guesses [148], i.e., the first 5 guesses are free, and subsequent guesses are costly. Therefore they consider λ_5 up to λ_{100}. Biometric phone unlock typically limits the number of tries to 5 (iOS [2, 3]) or 3 (Android unlock class 1 and 2 [155]) before falling back to pin codes. The number of tries for class 3 Android biometrics is unspecified and is probably higher. Alternatively, for remote authentication systems where keystroke biometrics are used as a second factor on top of passwords, we expect the number of guesses allowed to be similar to stand-alone password authentication. Therefore in this work, we consider λ_3 up to λ_{10}. We also consider $\lambda_1 = H_\infty$ as a worst-case metric which acts as a lower bound. It will be used to describe the effectiveness of wolf and masterkey samples.

Offline attacks typically occur when credentials are leaked due to database breaches. However, this does not apply to biometric template database leaks as biometric templates are either encrypted with traditional encryption, which does not allow for matching templates without decryption or protected with biometric template protection schemes. In the first case, a brute-force adversary is obsolete. In contrast, in the latter case, other security bounds are defined, e.g., Van hamme et al. [248] described a security lower bound for fuzzy commitment schemes. However, offline attacks match an adversary who gained access to the interfaces and can therefore guess in an unthrottled manner. Furthermore, the offline adversary model and its metric α-guesswork, provide a complete picture of the biometric distribution. Therefore, to compare feature extractors, we consider all values of α and plot the *guessing curve*. Additionally, to compare biometrics with knowledge factors, we report results for $\alpha = 0.5$ and $\alpha = 0.25$.

While free-text keystroke recognition models are used for continuous and single-shot authentication, the parametrization above primarily focuses on the single-shot authentication use case. The single-shot and continuous authentication paradigms mainly differ in how often the keystroke recognition model is queried and how decisions are used. Continuous authentication is, thus, a more complicated form of single-shot authentication, which we will not evaluate within the scope of this work. However, we discuss adapting the framework to accommodate continuous authentication in Section 5.7.2.

Algorithm 1: Algorithm to determine the guessing order of biometric samples and corresponding attacker success rates p_i.

// C is the matrix with the centroids with dimensions $[c, f]$. X is the matrix with the biometric vectors with dimensions $[m, f]$. t is the threshold.
Data: C, X, t
// Compute the angular distance between unit length biometric samples and centroids
$A = 1 - XC^T$;
// 1 for all entries bigger than threshold, 0 for the others
$A = Ones[A > t]$;
$guesses$ = [];
// Find greedy guesses
while A *!= 0* **do**
  // Sum along the rows and find the maximum and its index
  $guesses.append(\max_i(\sum_j^c A_{i,j})/c)$;
  $ind = argmax_i(\sum_j^c A_{i,j})$;
  // Change matrix to accommodate for broken centroids
  $A = A * (1 - A[ind, :])^T$ // element wise multiplication over all rows
end
return $guesses$

5.5 Experimental setup

The evaluation and execution of our attack require a keystrokes model and an extensive dataset of free-typed text. In this section, we discuss these building blocks.

5.5.1 Keystroke data

Keystrokes from physical keyboards are collected as the keycode, the timing of the downstroke (press), and the upstroke (release). A very typical preprocessing step consists of extracting the relative timings of the bigrams [9, 4]:

- **Hold latency** is the time between the key down and key upstroke of the same key.

- **Inter key latency** is the delay between the key upstroke of the current key and the key down stroke of the following key.
- **Press latency** is the time between two consecutive key down strokes.
- **Release latency** is the time between two consecutive key up strokes.

Together with the keycode, this provides five features for each key pressed. In other words, the input to our model are sequences of variable length N, which leads to an input dimensionality of $N \times 5$. The keystroke timings are converted to seconds, and the javascript keycodes are divided by 255.

We perform our evaluation and train our models with the dataset collected by Dhakal et al. [70]. This dataset contains 15 sentences typed by over $168,000$ users, thereby being the largest scale-free text-typing dataset. The typing behavior was collected through a web form. Participants memorized a common sentence and afterward typed it out. The sentences were selected randomly from a corpus of 1525 common English sentences and contained at least three words.

After preprocessing the data, we had approximately 161k individuals, each typing 15 sequences. We select 160k individuals and split them into the following subsets:

- **Training set** of $40,000$ users used for training the keystroke dynamics system.
- **Validation set** of $10,000$ used for early stopping and to determine the authentication system threshold.
- **Test set** of $110,000$ users used to evaluate the attack and perform the demographics analysis.

Input When training a second keystroke algorithm, we further divide the test set into a second training set of $40,000$ users and a test set of $70,000$ users.

5.5.2 Keystroke model

The keystroke model is a function $f(.)$ that extracts identity vectors from keystroke timings. These identity vectors are matched by computing the cosine distance between them. The function $f(.)$ is learned by training a deep learning architecture with the triplet loss function [213]. Our architecture consists of two cascaded gated recurrent units (GRUs) with 64 units with a batch normalization layer between them. The output of the last layer is normalized. Thus, the final identity vectors have norm 1 and are of length 64. The models are implemented using TensorFlow.

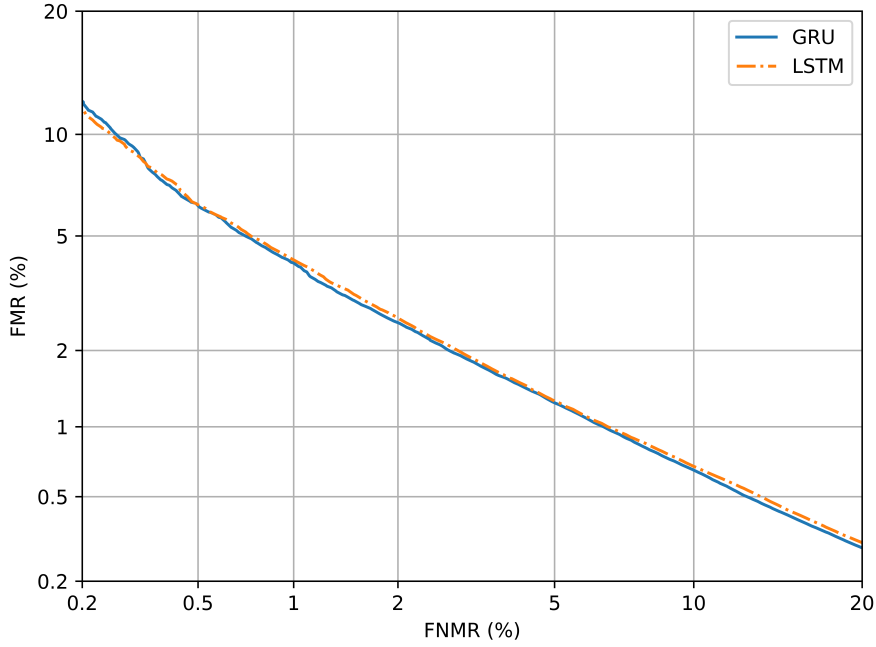
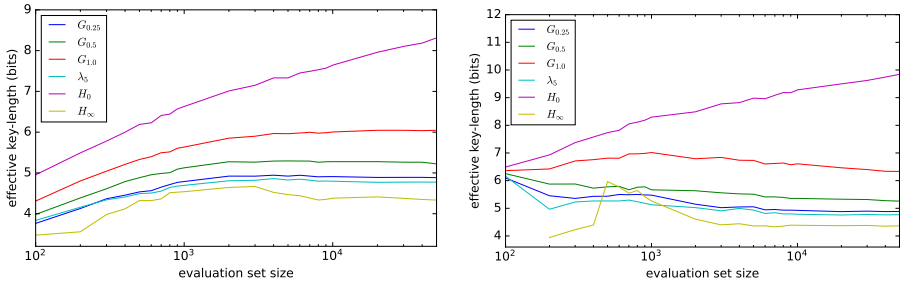


Figure 5.4: DET curve (FMR in function of FNMR) of the keystroke biometrics models.

Our second model is an exact clone of the TypeNet model [4]. This model is also trained with triplet loss and uses two LSTMs with 128 units. Thus, these biometric feature vectors also have a norm equal to 1 and are of length 128.

The matcher function $m(\cdot, \cdot, t)$ of our biometric system relies on the cosine distance function $1 - cos(\theta)$ where theta is the angle between $f(b_j^i), f(mk)$. The threshold t is chosen such that the false non-match rate (false rejection rate) in the validation set is 5%. At this threshold, the GRU and LSTM systems have a FAR in the validation set of 1.25% and 1.28%, respectively. The DET curves of both systems are shown in Figure 5.4. To compute the DET curve and throughout the experiments, we use 5 enrollment samples, i.e. typed sentences, for each enrolled user.



(a) Insider adversary

(b) Feature space adversary

Figure 5.5: Guessing metrics expressed as effective key lengths on the y-axis, the number of users in the evaluation set (on a logarithmic scale) on the x-axis. In 5.5a the adversary knows the enrolled users, while in 5.5b users are unknown.

5.6 Evaluation

Our evaluation is split into three main parts: first, we research the adversaries that generate masterkeys in feature space. Second, we investigate the adversaries that are not capable of injecting features into the system and are therefore limited to spoofing key timings. Third, we perform a demographics analysis.

5.6.1 Masterkeys in feature space

With our first set of experiments, we research the maximal potential of masterkey attacks and the effect of the size of the typing corpus. The number of unique individuals in the typing corpus impacts our adversary's ability to approximate the distribution of typing behavior for the whole population. Furthermore, the typing corpus's size impacts the metrics' reproducibility.

We conduct our first experiments with two of the adversaries described in Section 5.3.1: the *insider adversary* and *feature space adversary*. As the *insider adversary* knows which users are enrolled and executes their attack in feature space, they have optimal knowledge to create master keys. As such, researching the *insider adversary* provides insights into the minimal security offered by the keystroke dynamics system and, thus, the maximal effect of the attack. The *feature space adversary* cannot optimize their attack for the users enrolled in the biometric system. However, the adversary has a corpus of free-typed text that gives them an idea of how the population types and has access to the keystroke dynamics model used by the target system. Studying these

		H_∞	λ_3	λ_5	λ_{10}	$G_{0.25}$	$G_{0.5}$
Insider adversary	Wolf	4.38	4.69	4.8	5.02	4.91	5.28
	k-Means	3.93	4.21	4.28	4.46	4.27	4.46
	CMA-ES	3.66	3.82	3.92	4.18	3.86	4.06
Feature space adversary	Wolf	4.39	4.63	4.78	5.06	4.93	5.36
	k-Means	3.89	4.15	4.25	4.47	4.24	4.48
	CMA-ES	3.73	3.81	4.0	4.31	3.93	4.22
Input space adversary	CMA-ES	4.25	4.29	4.51	4.92	4.67	5.94
Pragmatic adversary	Wolf	4.91	5.42	5.47	5.6	5.65	5.82
	CMA-ES	6.64	7.1	7.74	8.67	8.91	8.91
Knowledge factors	Password [33]	6.5	/	/	9.1	17.6	21.6
	4-digit Pin [35]	4.75	5.22	5.5	5.91	6.32	8.78

Table 5.1: Guessing metrics computed on $10,000$ users for several combinations of adversary and attack. The results are compared with the entropy of a zero-effort attacker as well as traditional knowledge factors such as passwords [33] and PINs [35]

adversaries gives us insights into the impact of the size of the typing corpus used to execute and evaluate the attack.

In the first experiment, we executed a wolf attack as the *insider adversary*. The attack was carried out with varying numbers of enrolled users, and the results are depicted in Figure 5.5a. This figure demonstrates that the most important guessing metrics converge around $10,000$ unique individuals. Only the Hartley entropy, which is the $\log_2$ of the number of guesses required to impersonate all users, keeps growing. The guesswork (α-guesswork with $\alpha = 1$), which in the case of passwords does not converge, seems to converge only later.

In a second experiment, we repeat the wolf attack as the *feature space adversary*. Again, the number of users enrolled in the system is varied. The typing corpus used by the adversary contains different users and also varies in size. We keep the number of individuals of both sets the same and call it the evaluation set size. Figure 5.5b demonstrates the effect of the evaluation set size on the guessing metrics. A similar effect is observed in the first experiment, where the guessing metrics converge around $10,000$ unique individuals. Compared with the first experiment, more security is measured when the number of persons in the evaluation set is small ($< 10,000$). The small evaluation set does not provide a good approximation for the distribution of the samples in feature space. Therefore, the wolf samples transfer poorly. This also has

the effect that the adversary selects wolf samples that are less successful over more successful counterparts. In other words, the attack success rates are no longer ordered when computing the guessing metrics. In the case of large evaluation sets, wolf samples transfer well, and the order of the success rates is maintained.

In a third experiment, we compare three attack methodologies: wolf attack, k-Means, and CMA-ES. We execute this attack as the *insider adversary* and *feature space adversary* with evaluation sets of $10,000$ individuals. We plot the α-guesswork for different values of α (the fraction of users impersonated by the adversary) in Figure 5.6 and summarize the most important metrics in Table 5.1. Our results indicate that CMA-ES outperforms the other attack approaches. Furthermore, an adversary with access to the enrolled users is not more successful than an adversary who collects their typing corpus. Only in the case of CMA-ES do we observe an increase of 0.16 bits for $G_{0.5}$. Figure 5.6 stops at $\alpha = 0.7$ as we ran k-Means with $k = 20$, and instructed CMA-ES to stop when it had generated enough masterkeys to impersonate 70% of the population.

5.6.2 Masterkeys in input space

In Section 5.3.1, we defined two adversaries who operate in input space: the *input space adversary* and the *pragmatic adversary* These adversaries must inject masterkeys by spoofing key presses and releases. Compared to the adversaries who operate in the feature space discussed above, their adversary model is more attainable. Adversaries operating in input space can no longer freely navigate the feature space and are limited by the typed text. We research to what extent this dependency on the typed text influences the ability of the adversary to generate effective masterkeys.

We selected a random typing sequence from the typing database. This sequence is: 'shift', 'W', 'E', ' ', 'S', 'O', 'L', 'D', ' ', 'T', 'H', 'E', ' ', 'shift', 'U', 'K', ' ', 'B', 'U', 'S', 'I', 'N', 'E', 'S', 'S', ' ', 'T', 'O', ' ', 'shift', 'A', 'E', 'shift', 'F', 'forward slash', 'backspace'. Our attack within the *input space adversary* threat model relies on CMA-ES to generate timings of key presses and releases. The results of the attack evaluation are shown in Table 5.1. The *input space adversary* is less successful than the insider adversary, especially when more than five guesses can be made. We could only use the CMA-ES technique in the input space, as the k-Means only works in feature space, and the wolf attack yields the same results. Selecting another typing sequence might influence the effectiveness of the adversary. We discuss the effect of semantics more in-depth in Section 5.7.

A second experiment tests to what extent the masterkeys transfer to other models. The *pragmatic adversary* has no access to the target model and constructs masterkey samples with the help of a surrogate model. As discussed in Section 5.5.2, we created

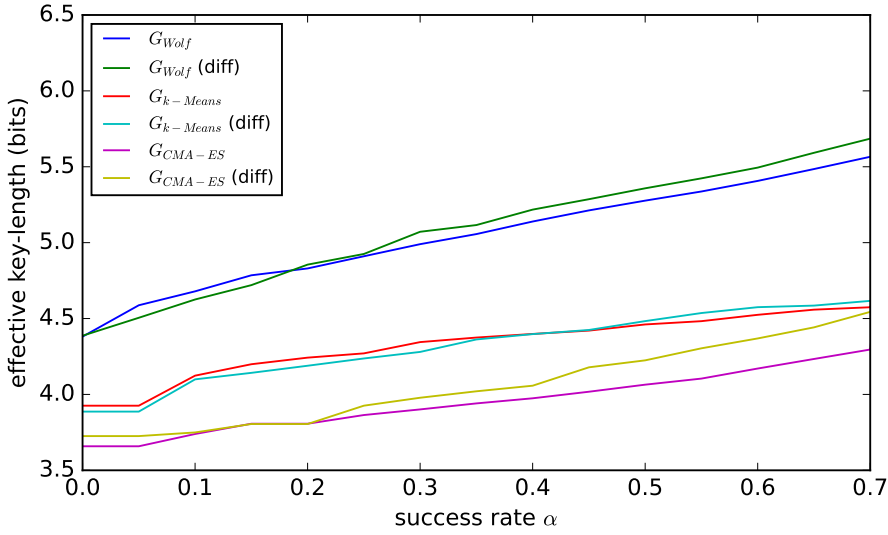


Figure 5.6: Change in α-guesswork (G) with increasing success rate α for the three attack methodologies. *diff* indicates knowledge of the enrolled users.

a second model which is trained with a different set of users, uses a different type of recurrent neural network, and is significantly bigger. This second network will serve as the target model for the masterkeys generated in the first experiment. We also measure the effectiveness of wolf samples. The wolf samples correspond to real typing sequences, which can be used for this evaluation. Of course, using such samples in the real world would severely limit the attacker's capabilities as they are restricted to the texts that occur in their corpus. When the *pragmatic adversary* uses CMA-ES to determine key press and release timings that lead to master keys, we observe that these transfer quite poorly to other models. For example, comparing the *pragmatic adversary* with the *input space adversary*, the keystroke dynamics system offers between 5 and 6 bits more security for all the relevant guessing metrics. In other words, the adversary must increase their guesses to break the same number of accounts with a factor of 32 to 64. It seems that the black-box optimization approach leverages model-specific patterns.

5.6.3 Demographics analysis

It is nearly impossible to change your appearance and behavior. When these aspects are used for security, you get what you see. This contrasts sharply with knowledge-

based factors, where you can drastically improve the security level provided by those solutions by attaining a healthy security posture. Therefore, it is useful to study the effect of demographics on the security level of biometric solutions. In our study, we took into account three attributes related to demographics and two attributes that are related to typing skills:

- **Age**, split into 5 different categories.
- **Sex**.
- **Native language**, between 6 common languages.
- **Number of fingers** used while typing (self-reported), a category per even number (up to 10).
- **Typing speed** (measured), which we subdivided into three nearly equal splits, i.e. slow, median, and fast typers.

We relied on our strongest attacker for these experiments: the *insider adversary* who uses CMA-ES to create masterkeys. We used the strongest adversary because we believe the minimum security level is the most informative. As shown in previous experiments, the offered security within other adversary models is (slightly) better. Another reason to use the *insider adversary* threat model is that we want to retain the maximal size of our evaluation set and avoid splitting it up into a set of enrolled users and the adversary's typing corpus. We only retained categories that had at least 1000 individuals. We randomly sampled 1000 individuals for each category and computed the insider adversary's guessing metrics. We repeated this process 10 times and report the average for each metric. Moreover, we performed a one-way ANOVA test to examine whether the security level differs significantly. In the native language category, most categories have a low number of individuals, which leads to a high overlap between the evaluation sets. The high overlap influences the standard deviation and violates the ANOVA tests' requirement of homoscedasticity. Therefore, for the native language category, we perform an Alexander-Govern test. The results can be found in Table 5.2. The p-values of the statistical hypothesis testing are included in Appendix A.2 (Table A.2). Table 5.2 summarizes this information and indicates for what groups the null hypothesis could not be rejected with high confidence ($p < 0.01$)

Our results conclude that age and sex do not affect the security level of free-typed keystroke dynamics. The age categories of > 50 and $35 - 50$ get slightly less security, but this could be related to typing skills. That age and sex do not affect security is a good sign. This implies no bias for these categories and that typing biometrics can be used for a large part of the population without discriminating.

Native language, however, does have a significant effect, with security dropping almost a bit for $G_{0.5}$ when the individual's mother tongue is Hindi instead of English. On the

other hand, this is an expected result, as being less fluent in the language you are typing in will influence typing behavior.

Our results also indicate that using more fingers seems to improve security. Using two fingers even lowers security significantly when only a few – 1 or 3 – guesses are allowed. Furthermore, being an exceptionally fast or slow typist reduces security. For security, it is better to type at an average speed.

5.7 Discussion

5.7.1 Comparison with knowledge-based factors

Comparison of guessing metrics. We also included the guessing metrics for 4-digit pins and passwords in Table 5.1. As such, we can compare their security level with the security offered by free-text biometrics. Because of the poor transferability of the masterkeys found with CMA-ES to other models, free-text keystroke biometrics are resilient against the pragmatic adversary. Moreover, free-text biometrics offer more security than knowledge-based factors when the adversary is limited to 10 guesses or less. When the adversary knows the typing biometrics model or can perform a significant amount of queries (over 200k queries are required), their masterkeys become immensely more effective. This reduces the security of the keystroke biometrics system to only 4.8 bits, which is 1.1 bits less compared to a 4-digit pin (5.91 bits). In other words, a 4-digit pin offers two times as much security. Similarly, passwords (9.1 bits) are around 18 times more resilient to guessing than free-text keystroke dynamics (4.92 bits). We believe that the security against the *input space adversary* is the most crucial metric to describe the system's security because it does not rely on the secrecy of the biometrics model but still requires the adversary to generate key press and release timings.

Other threat models. The attacks described in this work assume that the adversary does not know the typing behavior of the individual they want to impersonate. If this assumption is violated, the adversary becomes drastically more successful. Rahman et al. [190] demonstrated that by snooping between 50 and 200 keystroke timings, they had an attack success rate for impersonation as high as 88%. On top of that, most browsers do not, by default, disable the precise timings of key events. Therefore users are, by default, at risk of having their keystroke biometrics spoofed when filling in web forms. Nevertheless, passwords and pin codes have similar threat models when considering phishing. The main difference is that the adversary still needs to trick the user during phishing, while capturing key events can be done by an honest-but-curious adversary.

Guessing metrics approximations. Due to many uncommon passwords, Bonneau et al. [33] needed a parametric approximation of α-guesswork. However, no such approximations are necessary in the case of keystroke biometrics, as no user seems entirely unique. More information on this in Section 5.7.4

5.7.2 Continuous authentication

Within the scope of this work, we primarily focused on keystroke recognition for single-shot authentication. To extend our evaluation framework to continuous authentication, we must consider how a continuous authentication system differs from a single-shot authentication system.

As described in Section 5.2.3, one way to build a continuous authentication system is by using a tumbling window over long typing sequences and fusing the authentication decisions by requiring three consecutive rejects to invalidate the session. In such case, an adversary who swaps between the three most popular typing behaviors expects to keep typing indefinitely for a fraction of λ_3 of the population. Important here is "typing indefinitely." As argued by Bours and Mondal [37], the number of actions before locking a session is a more meaningful metric for continuous authentication. Therefore, a more significant change to the evaluation framework considers the number of actions before session invalidation.

An adversary leveraging a masterkey attack against a continuous authentication system will want to reach a specific goal. For example, sending an e-mail with the victim's laptop. This action requires a specific number of keystrokes. The adversary should then develop typing sequences that maximize the fraction of individuals impersonated by it. The most important parametrization of the guessing metrics is when only one guess is allowed, i.e., H_∞. However, comparing $G_0.25$ and $G_0.5$ of different systems still provides a good idea of the performance of a system, as with single-shot authentication.

5.7.3 Limitations

Practicality of the attack. As argued in Section 5.4, masterkey attacks are intelligent dictionary attacks. To successfully execute a masterkey attack on free-text keystroke biometrics, the adversary has to generate the key press and release timings for a specific sequence of keycodes. The desired typing sequence depends on the application. For example, when keystroke dynamics are used for checkout fraud detection in online transactions, the keystroke dynamics system will monitor some form that needs to be filled in. The adversary, thus, has to generate masterkeys for that specific form. If these masterkeys are not reusable, plenty of effort is lost. In other words, the semantics of

the text can serve the same purpose as salting in passwords. Our attack requires too many queries to the system to generate key press and release timings in real-time.

Semantics. We did not examine how the typed text (actual keycodes/words) impacts the adversary's ability to carry out a masterkey attack. This involves semantics specific to the application and necessitates the collection of datasets under adversarial conditions.

5.7.4 Future work

Besides addressing the limitations of this work and extending the framework to support continuous authentication, we believe our evaluation framework can be generalized to **describe biometric security better**. Une et al. [238] attempted to introduce the wolf attack probability as a default metric to describe biometric system performance. The existence of wolves is currently not accounted for by mainstream metrics such as FMR, as they are aggregated metrics. We did not reach an evaluation set size with the number of individuals required to make the hartley entropy converge during our evaluation. However, because biometric models project the biometric samples on the n-ball (norm of the feature vectors is 1), a point must be reached where all the guessed points are at most separated by the threshold. The dimensionality of this n-ball is probably less than the dimensionality of the feature space as this is a parameter of the architecture (in this research 64 and 128).

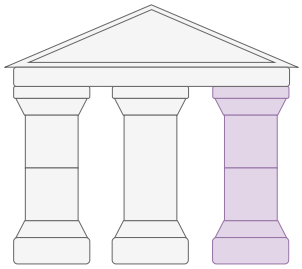
5.8 Conclusion

There is an increased interest in behavioral authentication due to its potential to assess identity continuously without added friction. Keystroke biometrics is a mainstream behavioral biometric part of several companies' portfolios. However, as such systems are used over the web, they have very different threat models compared to mainstream phone unlock biometrics. This research performs a systematic security evaluation of free-text keystroke biometrics. To this end, we first design a masterkey attack against free-text keystroke biometrics. Masterkeys have great potential to be leveraged in attack vectors against remote behavioral authentication systems because they match a large part of the population, making them ideal candidates to execute impersonation attacks when the victim's behavior is unknown to the adversary. Our attack leverages black-box optimization algorithms to find such masterkeys. Second, we propose a novel evaluation framework based on guessing metrics used for the security evaluation of passwords. Therefore, we can directly compare biometric and knowledge-based authentication factors and conclude that keystroke biometrics offer several magnitudes less security than passwords (≈ 4 bits less) and 4-digit pin codes (≈ 1 bits less). Lastly, we perform a demographics analysis that sheds new light on the security of free-text

typing biometrics w.r.t. typing skill, native language, age, and sex. We conclude that using more fingers to type while maintaining an average typing speed (around 58 wpm) is beneficial for security.

		size	$\hat{\tilde{H}}_\infty$	$\hat{\tilde{\lambda}}_3$	$\hat{\tilde{\lambda}}_5$	$\hat{\tilde{\lambda}}_{10}$	$\hat{\tilde{G}}_{0.25}$	$\hat{\tilde{G}}_{0.5}$
Age	<18	33457	3.56	3.72*	3.84*	4.16*	3.77*	4.01*
	18-25	37790	3.39	3.69*	3.87*	4.23*	3.75*	4.1*
	25-35	24453	3.22	3.65*	3.86*	4.22*	3.71*	4.07*
	35-50	10410	3.11	3.5	3.7	4.07	3.49	3.82
	>50	3647	2.75	3.13	3.4	3.95	2.99	3.41
Sex	male	44864	3.46*	3.67*	3.87*	4.22*	3.75*	4.08*
	female	50788	3.46*	3.68*	3.86*	4.21*	3.74*	4.07*
Native language	en	92433	3.45	3.77	3.93	4.22	3.84	4.11
	es	2157	2.69	3.36	3.69	4.23	3.18	4.05
	hi	1402	2.59	3.0	3.31	4.02	2.8	3.24
	tl	2649	2.95	3.43	3.66	4.07	3.4	3.78
	zh	1721	3.35	3.67	3.84	4.18	3.69	4.0
Number of fingers	1-2	14050	2.35	3.02	3.43	4.01	2.71	3.44
	3-4	14149	2.82	3.11	3.41	3.96	2.98	3.42
	5-6	11770	3.21	3.53	3.73	4.13	3.55	3.91
	7-8	17487	3.73	3.89	4.04	4.33	3.97	4.27
	9-10	52473	2.92	3.4	3.64	4.11	3.36	3.8
Words per minute (speed)	<39.5	36767	2.3	2.74	3.06	3.82	2.52	2.81
	39.5-58.5	35931	2.98	3.21	3.47	3.98	3.16	3.54
	>58.5	37232	2.16	2.65	2.95	3.76	2.38	2.67

Table 5.2: Estimates of the guessing metrics for relevant demographic splits. We only provide groups with more than 1000 unique users, which excludes several *native languages*. Values with an asterisk (*) are not significantly different, i.e., p-value ≥ 0.01.



Chapter 6

A novel evaluation framework for biometric security: Assessing guessing difficulty as a metric

> "The only man I know who behaves sensibly is my tailor; he takes my measurements anew each time he sees me. The rest go on with their old measurements and expect me to fit them."
>
> — *George Bernard Shaw*

Preamble

This chapter reports on our ongoing study on replacing the false match rate as a security metric. The primary issue is that the existing metric used to evaluate biometric security, the False Match Rate (FMR), does not account for the distribution of mismatches and fails to capture the phenomenon of masterkeys. This metric is unable to provide insights into the number of unique individuals the biometric system can differentiate, making it an ineffective measure for evaluating the biometric brute-force adversary's effectiveness. As a result, the chapter proposes a new methodology to evaluate biometric security that includes (1) a theoretical analysis of the effective key space of biometrics, (2) a

security evaluation framework to replace the FMR. The utility of the framework is demonstrated through a large-scale security evaluation of face and free-text keystroke authentication. The goal is to provide a comprehensive methodology for evaluating biometric security, accounting for interpersonal differences and allowing comparisons with other authentication factors, while considering inputs beyond those observed in the wild.

Ongoing Work: T. Van hamme, G. Garofalo, E. Argones Rúa, D. Preuveneers and W. Joosen, A novel evaluation framework for biometric security: Assessing guessing difficulty as a metric.

6.1 Introduction

Authentication is one of the most important security controls to prevent unauthorized access to confidential information and sensitive resources. Nevertheless, authentication poses a complex challenge that lacks a one-size-fits-all solution that is user-friendly, secure, deployable, and free from privacy risks. Passwords have become a nightmare [79, 34] to handle, and despite enormous efforts of the community [8, 92], adoption of possession factors is low [184, 56]. Biometric authentication has proven to be a compelling alternative that relieves users from remembering complex passwords and managing physical tokens. Therefore, biometrics have been widely adopted to unlock personal devices.

Despite the growing importance of biometrics, their security is often challenged. For example, recently, the security community was caught off-guard by the concept of biometric masterkeys [171, 221, 36]. Masterkeys are biometric inputs that match an unexpectedly large portion of the population. Consequently, masterkeys have tremendous potential for a brute-force adversary that aims to impersonate a victim without knowledge of their biometrics. The main reason masterkey attacks took aback the biometric community is the lack of consensus on evaluating the effectiveness of a *biometric brute-force adversary* who tries to guess the correct biometric data. Currently, the primary metric relied upon is the False Match Rate (FMR), also known as the False Acceptance Rate (FAR). However, the FMR only captures the proportion of mismatches when random individuals make false identity claims.

The FMR does not correspond to the effectiveness of the biometric brute-force adversary for three main reasons: First, the FMR is an aggregate metric that does not account for the distribution of mismatches. Therefore, it fails to account for the existence of masterkeys. Second, the FMR alone does not provide insights into the number of unique individuals the biometric recognition system can tell apart. For this reason, it cannot be interpreted as an 'effective key length', the most commonly used metric to capture the effectiveness of a brute-force adversary. Third, the FMR is evaluated using biometric input signals observed in the wild. Because collecting biometric data

is time-consuming, researchers are tempted to rely on a limited number of users to evaluate the FMR, thereby missing among others the impact of corner-case inputs.

The phenomenon of masterkey attacks can be explained by the non-uniform distribution of biometrics [71] and the fact that biometric matching is not exact. Thus, there are 'hotspots' in the biometric space that cover a larger portion of the population. This information is not captured by the FMR, which only provides insights into the average occurrence of mismatches. Because the mismatches are systematic [75], they are more prevalent for certain individuals. As a consequence, a brute-force adversary leveraging masterkey attacks will impersonate these users at a rate that is significantly higher than the FMR. Accounting for these personal differences during security evaluations is important as, unlike passwords, users cannot improve their biometrics security level by following proper security guidelines.

The FMR does not account for the distribution of biometrics. However, it is expected to provide insights into the robustness against biometric brute-force adversaries because the FMR is used to predict the size of the effective key space [176, 248]. The cardinality of the effective key space is the number of unique values the brute-force adversary can choose from and is a popular metric to describe security against brute-force adversaries. For example, in the case of a 4-digit PIN code, there are 10.000 unique inputs. Thus, the expected success rate of a random guess is 10^{-4}. Following similar reasoning, the effective key space of biometrics is FMR^{-1}. Nonetheless, this reasoning is false. It was shown that despite low FMR, an adversary providing random input to the biometric system obtains significantly higher odds of impersonating their victims [268]. As such, another method is required to compare different approaches to authentication.

Biometric systems might contain design flaws, such as corner-case inputs that match a large part of the population or being too tolerant during the matching procedure. For example, it was shown that an all-grey image is always accepted by a finger vein authentication system [238]. Or, the region of acceptance in the biometric template space was not restrictive when building binary classifiers for authentication [268]. These problems are hard to discover while solely relying on 'naturally' occurring biometric data that is costly to collect. Therefore, a security evaluation should consider uncommon input.

In this work, we research the *biometric brute-force adversary* and propose a methodology to measure their effectiveness. Our research comprises three fundamental elements that form the core of our contribution:

- A **theoretical analysis** of the size of the effective key space of biometrics. The theoretical analysis provides an upper limit to the number of unique guesses a *biometric brute-force adversary* has to perform. It lays the foundation required to understand the impact of sample size on the security evaluation of biometrics.
- A **security evaluation framework** for the *biometric brute-force adversary*

that replaces the FMR and related metrics. Our framework accounts for interpersonal differences, facilitates comparisons with other authentication factors, and considers inputs beyond those observed in the wild. To achieve these properties, our methodology leverages guessing metrics. Guessing metrics have been proposed to evaluate password security [33]. They quantify the effort required by a brute-force adversary who estimates the distribution of 'secrets' (passwords) and formulates guesses by advancing in the optimal order. Guessing metrics acknowledge that such distribution can be very unbalanced and therefore account for the fact that an adversary can aim to break only a fraction of all accounts. The primary challenges in adapting guessing metrics to biometrics are accounting for fuzzy matching and the often non-discrete nature of biometric templates.

- A **large-scale evaluation** of the security of face and free-text keystroke authentication against a *biometric brute-force adversary*. Our large-scale analysis finally answers questions such as: "How does the security of face recognition compare with pin codes?" and "How much does adversary knowledge of soft-biometric characteristics (e.g., gender, age, mother tongue, etc.) influence security?".

6.2 Related work

Evaluating the security of biometric systems is an important problem for which standardization organizations describe best practices, e.g., Common Criteria [57], Android [155], FIDO [113, 114], etc. They typically define two major threat models for *impersonation*:

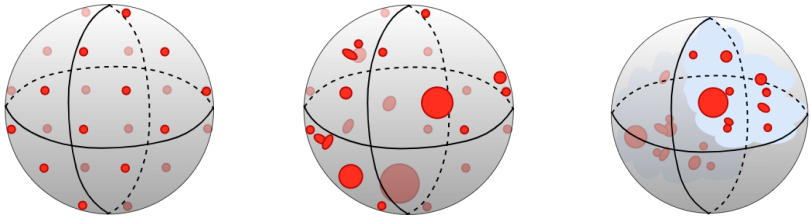
- **Zero-effort adversary.** This adversary counts on their own biometric data to impersonate their victim. They have no *knowledge* of the biometric distribution or the biometric of their victim. They have no *capabilities* beyond presenting their own biometric signal to the sensor. The metric used to capture the effectiveness of the zero-effort adversary is the False Match Rate (FMR), also known as the False Acceptance Rate (FAR).
- **Presentation adversary.** This adversary *knows* the biometric signal of their victim. Furthermore, they are *capable* of crafting fake biometric signals presented to the sensor. The Impostor Attack Presentation Accept Rate (IAPAR) captures the effectiveness of the presentation adversary.

These adversaries are at opposite ends of the spectrum: the zero-effort adversary is very weak. Weak to the extent that one can wonder whether the "adversary" in "zero-effort

adversary" is justified. In contrast, the presentation adversary is very knowledgeable and capable. This gap leaves room for a middle-ground adversary who does not know the biometric signal of their victim but is not limited to only using their own biometric data. We call this middle-ground adversary the **biometric brute-force adversary**. The biometric brute-force adversary is similar to the zero-effort adversary as they leverage the fuzziness in pattern matching to obtain access to the system more easily. Unlike the zero-effort adversary, the biometric brute-force adversary is an active adversary who presents biometric signals based on their knowledge of the biometric distribution. Brute-force adversaries have two attack strategies at their disposal [144]: (i) the *naive attack*, where inputs are guessed with no particular order, and (ii) the *dictionary attack*, where more probable inputs are guessed first. Within the scope of this work, we are mostly interested in the effectiveness of the dictionary attack.

6.2.1 Masterkey attacks

The threat model of the biometric brute-force adversary is relevant only if the attacker can gain an advantage compared to the zero-effort adversary. Serwadda and Phoha [217] find such an advantage against a biometric recognition system that leverages swiping behavior. They demonstrate that the FMR becomes meaningless when instructing a "Lego" robot to perform swipes based on population-wide statistics. Moreover, Serwadda and Phoha show that the FAR is increased between 339% and 1004% depending on the type of touch stroke. Serwadda and Phoha extended this idea to other biometrics biometrics [216] and introduced the term masterkey. Similar to a physical masterkey that can open multiple locks, the biometric masterkey is biometric input that works for multiple individuals. Serwadda and Phoha construct a masterkey against keystroke biometrics [216] by choosing keystroke timings based on population-wide statistics of bigrams, i.e., the delays between two consecutive key up and down events. Their masterkeys are accepted by 5% to 30% of the population, while the equal error rates were considerably lower. However, their approach did not allow them to find masterkeys for users further from the mean. Negi et al. [167] address this problem using the popular initialization algorithm of k-means, k-means++. They describe two types of attacks: targeted and indiscriminate. Their targeted attack requires the adversary to collect a large dataset of users typing the specific password for which they want to generate the typing behavior. The targeted attack allows impersonating 40%-70% of the users with ten tries. The indiscriminate version does not require such a specific dataset and only requires the adversary to know the distribution of the bigrams contained in the password. This freedom comes at a small performance penalty as the adversary can impersonate 30%-50% of the users in ten tries. Van hamme et al. [243] rely on a black-box optimization approach to extend prior work on masterkey attacks from fixed-text keystroke dynamics to free-text. Habib et al. [98] further improved on the previous approaches by using a mathematically grounded



(a) Uniform (b) Non-uniform (c) Manifold

Figure 6.1: Example distributions of biometrics on the surface of a sphere. 6.1a shows an ideal uniform distribution. 6.1b demonstrates the non-uniformity. 6.1c illustrates that the biometrics do not populate the entire surface of the sphere but only a subspace known as the manifold.

approach that finds overlap in 'acceptance' regions of n-dimensional hypervolumes. However, their approach comes at a big computational overhead.

The statistical approaches of the works mentioned above are computationally heavy and infeasible in high-dimensional spaces, such as 2d images. Therefore, for facial images [171, 221] and fingerprints [36], a generative model is leveraged. The latent space of the generative model is explored with a black-box optimization approach, e.g., evolutionary algorithms. The optimized function is chosen such that the generated images match a large part of the population. Bontrager et al. [36] use this approach to generate DeepMasterPrints that break 1.11% of the population with a single attempt while the FMR is 0.01%. In the face authentication domain Shmelkin et al. [221] and Nguyen et al. [171] concurrently published work on "master faces" that are highly successful. For example, with 10 master faces, over 40% of the population is covered.

Throughout this chapter, we will use the term **masterkey attack** to refer to forged biometric samples optimized to match a large part of the population. Hence, we do not change the term depending on the targeted biometric modality.

Masterkey attacks are possible due to the non-uniformity of biometric traits, which is illustrated in Figure 6.1b. Doddington already demonstrated this effect in 1998 [238]. Specifically, he introduced animal categories for individuals who are easier to impersonate, match poorly to themselves, and more easily impersonate others. This last animal is called a wolf, an individual that easily impersonates others. Brute-force adversaries can also gain an advantage by copying biometric traits from wolves. We refer to **wolf attacks** when the adversary selects the most effective attack sample from a biometric database rather than creating one themselves.

6.2.2 Evaluating the effect of masterkeys

The FMR, being an aggregate metric, cannot capture the non-uniformity of the biometric distribution. Eberz et al. [75] recognize this problem and again show that errors in biometric recognition tend to be systematic. Moreover, they show that optimizing biometric recognition models based on the mean error rate can enlarge the prevalence of systematic errors, increasing the opportunity for wolf and masterkey attacks. As a solution, they propose to report the Gini Coefficient along with the FMR and FRR. However, the Gini Coefficient lacks any interpretation within the context of security.

Une et al. [238] proposed a metric with a security interpretation called the wolf attack probability (WAP). The WAP looks for the most effective input sample and measures the probability that it matches the set of enrolled users. The metric is called *wolf* attack probability even though Une et al. [238] do consider masterkey attacks. For example, they show that inputting an all-grey image matches 100% of the individuals enrolled in a finger vein authentication system. Of course, safeguards can be added to prevent this type of out-of-distribution input. This demonstrates that it is important to distinguish between wolf attacks, which are in distribution, and masterkey attacks, which can be out of distribution. In 2007, when Une et al. [238] proposed WAP this distinction was not made. WAP is a worst-case security metric because it only considers the most effective masterkey sample. Therefore, it does not weigh the effectiveness of subsequent guesses and disregards the distribution.

The prior work on masterkey attacks uses a plethora of evaluation metrics. Some only consider the most effective masterkey and acknowledge that some users will never be impersonated [217, 216]. Others use an attack gallery of 10 masterkey attacks and report the percentage of enrolled individuals impersonated [221]. Yet another way to express the effectiveness of the masterkey is to report the increase w.r.t. the FMR [217, 36], which is great for denoting the relevance of the attack, but not so good for comparing the security the biometric offers. All are reasonable choices, but the attack curves plotting the percentage of impersonated users w.r.t. the number of masterkeys used, proposed by Habib et al. [98] paints the most complete picture. The main limitations of these figures are that their interpretation is not straightforward. and furthermore, they cannot be used to compare with other authentication factors, such as knowledge-based factors. This study builds upon the concept of attack curves introduced by Habib et al. [98] and determines the 'optimal' strategy for a biometric brute-force adversary generating multiple masterkeys. It links this 'optimal' strategy and the well-known guessing metrics used in knowledge-based factors. Consequently, our metrics facilitate a comparison between knowledge-based and biometric-based authentication methods. Moreover, the biometric guessing metrics allow for the measurement of the impact of the adversary's knowledge about the victim's soft-biometric characteristics on the guessing difficulty.

6.2.3 Effective key length of biometrics

Effective key length k is the most commonly used metric to describe the effectiveness of a brute-force adversary. The effective key length originates from cryptography, describing the number of unique secret keys 2^k. The effective key space size is the maximum number of attempts required to guess a secret. Moreover, assuming that every key is equally likely, the success rate of a single guess is 2^{-k} Using a similar reasoning, prior work [176, 248] estimates the key space of biometrics as FMR^{-1}. However, determining the unique number of biometric values is more complicated because biometrics requires fuzzy matching instead of exact matching. We will show in Section 6.3.2 that due to fuzzy matching, the effectiveness of subsequent guesses will decrease due to overlap.

Another approach by Zhao et al. [268] estimates the acceptance region of the biometric system. The acceptance region is determined by randomly sampling biometric feature vectors in the feature space and computing the fraction the authentication system accepts. Zhao et al. introduced this approach as they observed that some authentication systems with small false match rates allowed adversaries supplying random input to gain access with better odds than a coin flip. We need to point out that their approach to authentication exacerbated this problem because they built binary classifiers that model the negative class with only a few examples of other users. Therefore, they do not maximally restrict the acceptance region. Furthermore, their evaluations of the error rates were done with only a few users. Nonetheless, the concept of the acceptance region captures the effectiveness of a brute-force adversary providing random input to the system. Yet, the metric does not capture the minimal number of acceptance region volumes required to cover the whole feature space. To achieve coverage of the whole template space, an overlap of these regions might be unavoidable.

Terhörst et al. [234] research the generalizability of masterkey attacks and describe an approach to compute the maximum effectiveness of masterkeys in an 'ideal' setting. To this end, they approximate the expected success rate of a brute-force adversary in an 'ideal' setting where the biometrics are uniformly distributed in template space. The authors define a constrained optimization problem to find the maximum number of points that can be distributed in template space such that the distance between them is greater than a threshold. In other words, they look for the number of individuals that can be distributed over the template space such that there is no overlap, i.e., the capacity of the biometric template space. They then solve another constrained optimization problem to determine the 'maximum coverage' of a masterkey in this 'ideally' separated space. The maximum masterface effectiveness is then the ratio between the masterkey coverage and the capacity of the template space. However, masterkeys are mainly effective due to the non-uniform distribution of the biometric templates. Thus, the more relevant question is: "What is the minimum number of guesses required to cover the whole biometric template space?". Additionally, we

would like to point out that the dimensionality of the biometric template space is a system parameter. To not lose performance, this system parameter is larger than the dimensionality of the manifold the biometric templates populate. Figure 6.1c illustrates the distribution of the biometrics on only part of the surface of the sphere, i.e., the manifold.

6.2.4 The research gap

Prior work clearly demonstrates that the FMR is not a reliable security metric:

- Mismatches are systematic [75] and the probability of a mismatch is not distributed uniformly [71]. The FMR is an aggregate metric that does not capture the distribution of mismatches. As such, it cannot adequately capture the effectiveness of masterkey attacks.
- The FMR has been used to estimate the 'effective key length' to compare biometrics and other authentication factors [176, 248]. As the FMR is the probability that a random person successfully impersonates another random individual, this seems reasonable, but it is incorrect (cf. Section 6.3.2). Biometrics rely on fuzzy matching rather than exact matching. Therefore, overlap between successive guesses is unavoidable, which leads to a decrease in the guessing success rate. Therefore, another approach must be taken to estimate the number of unique guesses required to cover the whole biometric template space.
- The FMR is evaluated using only observed biometric samples. Because collecting biometric data is costly and time-consuming, it can lead to evaluation with too few users. This led to systems that were believed to be secure but were vulnerable to random input [268]. Another disadvantage of only relying on biometric samples observed in the wild is that it does not allow spotting valid but 'dangerous' inputs with the potential to match all users [238].

There is a real need for an evaluation framework that captures the effectiveness of a brute-force adversary. The evaluation framework should:

1. **Account for the distribution of biometric mismatches.** An adversary will leverage non-uniformity to gain an advantage.
2. **Furnishing practical revelations.** The assessment framework offers valuable information, serving as a reliable and comprehensible yardstick to facilitate decision-making endeavors. Notably, it empowers the comparison of security

levels provided by different authentication solutions and permits the quantification of the influence of supplementary adversary knowledge, such as knowledge pertaining to the victim's demographics. This capability contributes to enhancing the efficacy and precision of security-related determinations and strategies.

3. **Incorporates diverse input scope.** The proposed approach encompasses input data beyond the biometric information typically encountered in real-world scenarios. This inclusive perspective proves beneficial in the identification of potential design flaws that might not surface when restricted to observed data in natural settings. By deliberately examining unconventional inputs, the system gains the ability to detect corner cases, thereby contributing to a more robust and comprehensive evaluation of the system's performance and capabilities.

6.3 The brute-force adversary

In this research, we study the brute-force adversary targeting biometric authentication systems. The brute-force adversary model is well-studied for the other authentication factors, i.e., cryptographic keys and knowledge-based factors, where they identified two attack strategies (cf. Section 6.2):

- **Naive attack.** Inputs are chosen randomly.
- **Dictionary attack.** More probable inputs are picked first.

When all inputs to the authentication mechanism are equally probable, the dictionary attack does not provide an advantage over the naive attack. In the case of uniform input, the expected number of guesses is $\frac{N+1}{2}$ (cf. Equation 6.10) with N the size of the input space. The input space size is often expressed on a logarithmic scale as the 'effective key length' k, and thus the expected number of guesses is $\frac{2^k+1}{2}$. However, the distribution is never fully uniform. For example, humans choose dictionary words in their passwords, pick dates as PIN codes, and even random number generators are only pseudo-random. A brute-force adversary who estimates the distribution, i.e., makes a dictionary, gains an advantage.

Non-uniformity reduces the guessing difficulty and must be accounted for. In the case of biometrics, two additional factors have to be considered: (i) the tolerance allowed during matching, also called fuzzy matching, (ii) the number of distinct biometrics N is unknown. Moreover, the set of all possible inputs is uncountable. The remainder of this section dives deeper into those three challenges. An overview is presented in Table 6.1, and Figure 6.2 visually represents the challenges.

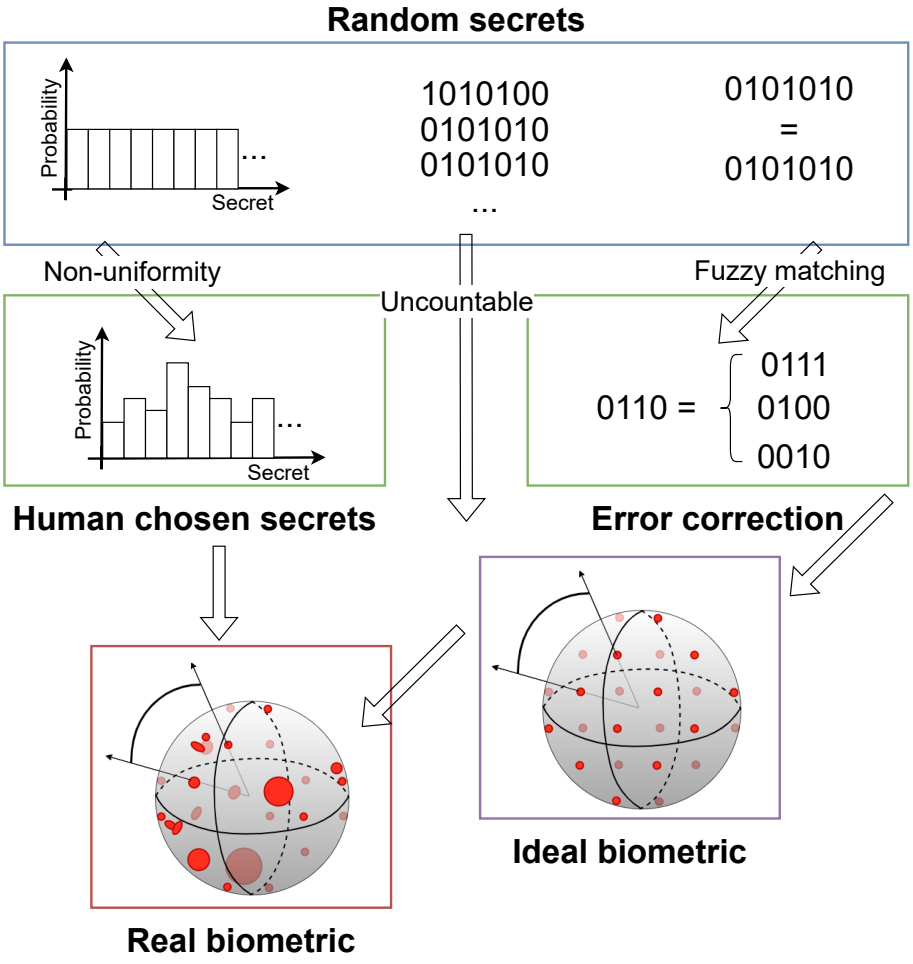


Figure 6.2: Illustration of the primary challenges when computing guessing difficulty of secret data: non-uniformity, fuzzy matching, and uncountability. For randomly chosen secrets, every codeword is equally probable, and each codeword matches only with itself. Human-chosen secrets introduce non-uniformity. Error correction allows one codeword to match multiple codewords. For an ideal biometric, the secrets are no longer discrete but continuous feature vectors, i.e., the codepoints are distributed uniformly on an n-ball. Matching happens by computing the angular distance. When dealing with real biometrics, all challenges come together as the data is non-uniform, the events are uncountable, and the matching is fuzzy.

	Uniform	Exact matching	Countable	Measure of guessing difficulty
Perfectly random secrets	✓	✓	✓	Guesswork (cf. Equation 6.10)
Human chosen secrets	✗	✓	✓	Guessing metrics [33] (cf. Section 6.5.1)
BCH codes (ECC (n, k, t))	✓	✗	✓	Effective key length k
Ideal biometric	✓	✗	✗	Section 6.4
Real biometric	✗	✗	✗	Section 6.5.2

Table 6.1: Overview of the factors impacting the evaluation of guessing difficulty and the approach used to evaluate them.

6.3.1 Non-uniformity

Non-uniformity complicates computing the guessing difficulty. An intuitive choice to account for non-uniformity relies on a measure of uncertainty, i.e., the Shannon entropy. However, it has been shown that this belief is wrong [144]. On the other hand, the generalization of Shannon entropy, i.e., Réyni entropy, does provide a lower and upper bound on guessing difficulty. Specifically:

- **min-entropy** [203] $H_\infty = -\log p_1$ is a lower-bound that is solely based on the probability of occurrence of the most frequent secret p_1. Note that the min-entropy has also been proposed as a security metric in biometrics and is equivalent to the logarithm of the Wolf Attack Probability (WAP) [238].
- **Hartley entropy** [203] $H_0 = \log N$ is an upper-bound and is based on the size of the input space (number of unique inputs) N.

Both metrics disregard the biometric distribution and can provide very different perspectives. Furthermore, the entropy metrics introduced above do not consider that the brute-force adversary can decide to stop early. Thereby, the adversary accepts that some accounts will never be broken, but they do not keep spending effort on very improbable inputs. In Section 6.5, we introduce guessing metrics that capture the effectiveness of such a pragmatic brute-force adversary. Specifically, we present the metrics used for knowledge-based factors in Section 6.5.1 and adapt them for biometrics in Section 6.5.2. The main challenge to adapting the guessing metrics for biometrics is accounting for fuzzy matching and the continuous nature of biometric templates, which makes them uncountable.

6.3.2 Fuzzy matching

The size of 'effective key space' N upper bounds the guessing difficulty. Fuzzy matching complicates determining N because a single guess eliminates multiple options. However, if overlap between guesses can be avoided, N can be computed.

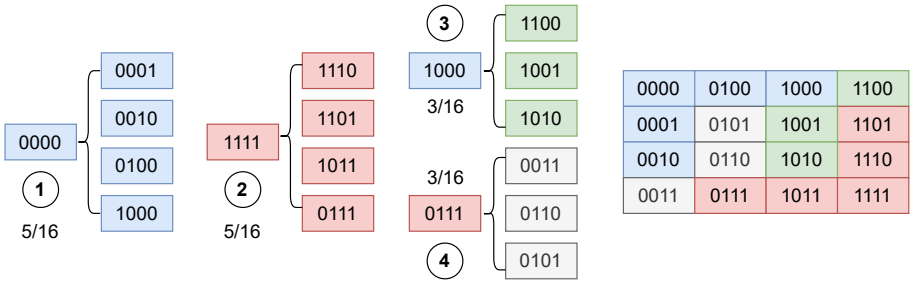


Figure 6.3: An optimal guessing sequence of 4 guesses that exhaust all 4-bit codewords, where codewords match if their Hamming distance is 1 or less. The initial success rate of a guess is $\frac{5}{16}$ which drops to $\frac{3}{16}$ for the third and fourth guesses.

For example, consider BCH codes, a class of error-correcting codes. An (n,k,t)-code contains 2^k unique messages and can correct t bit flips in an n bit codeword. Such BCH codes relate to biometrics because they are used in biometric template protection schemes. Biometric enrollment templates are protected by xor-ing an encoded codeword (randomly chosen among the 2^k) with the discretized biometric representation (of n bits). During authentication, a new biometric representation is captured from the same individual. This new biometric sample will be similar to the enrollment sample but might contain some errors. Then the new representation is again xor-ed with the protected biometric template, possibly introducing errors on the n-bit codeword. The n-bit codeword with possible errors is decoded. If the Hamming distance between the original biometric representation and the one collected during verification is less than t, the error-correcting capability of the ECC, the correct codeword, will be found, and the verification system will accept the user. Otherwise, the codeword will not be decodable, or the decoding process will yield the wrong k-bit codeword. This authentication procedure is the primary idea behind fuzzy key commitment [119].

In other words, if the distribution of the n-bit biometric templates is uniform and the n bits are uncorrelated, the protected biometric can represent 2^k unique individuals and has effective key length k. However, typically the n bits extracted from biometric data are correlated. Daugman [62] first noticed that despite extracting 2048 bits from an iris scan, the score distribution of the Hamming distance between imposter pairs resembled a binomial with 244 degrees of freedom (DoF). Van hamme et al. [248] relied upon this approach to estimate the effective key length for protected biometrics as $\frac{k}{n}\mathrm{DoF}$.

BCH codes avoid overlap between the $N = 2^k$ different messages. However, as shown by the formula above, avoiding overlap comes at the cost of performance, as the protection schemes decrease the number of unique messages N with a factor of $\frac{k}{n}$ (the code rate). When not relying on BCH codes, the FMR has been used to

determine the size of the effective key space as FMR^{-1}. This reasoning follows a similar train of thought as when determining the expected success rate of guess for uniformly distributed secrets with exact matching: If there are N distinct possibilities, the expected success rate of a guess is $\frac{1}{N}$. However, this is incorrect in the case of fuzzy matching because the overlap between already tried guesses and new guesses is unavoidable, which reduces the average success rate of a single guess. As an illustrative example, consider the set of 4-bit codewords, where all codewords are equally likely, and two codewords match if their Hamming distance is one or less. Then the expected success rate of a guess, i.e., FMR [148] is:

$$\mathrm{FMR} = \frac{\binom{4}{0} + \binom{4}{1}}{2^4} = \frac{5}{16} \tag{6.1}$$

However, as shown in Figure 6.3, the success rate of consecutive guesses decreases due to overlap. Figure 6.3 depicts an optimal guessing sequence that limits the overlap and thus maximizes the average success rate of guesses. It can be seen that only the first two guesses achieve a success rate of $\frac{5}{16}$, In other words, FMR^{-1} is a lower bound for N.

In this chapter, we propose a different approach to computing the 'effective key space' of a biometric recognition system. Moreover, in Section 6.4, we deduce an upper bound on the size of the 'effective key space' N. Our approach starts from the observation that most biometric feature extractors project the captured biometric signal on the n-ball. Then we compute the minimal number of points on the n-ball such that every point is at most a pre-determined angle apart. The two most important parameters in this function are the angle, directly related to the tolerance of the matching, and the dimensionality n of the n-ball. The acceptable False Rejection Rate (FRR) determines the tolerance of matching. n is a system parameter, thus n can be chosen arbitrarily large. This would falsely increase the estimate of N. Therefore, we require a way to estimate the true dimensionality of the biometric data.

6.3.3 Uncountable biometrics

The number of unique biometric signals cannot be counted. Yet counting is essential to determine the security against a brute-force adversary. Particularly, computing guessing metrics requires estimating the probability of a specific password by counting the number of times, it occurs (cf. Section 6.5). Additionally, the number of unique PIN codes is counted to determine an upper bound on guessing difficulty. Thus, the uncountable nature of biometrics complicates computing the guessing metrics (cf. Section 6.5.2) and requires a different approach to determining the size of the effective key space (cf. Section 6.4).

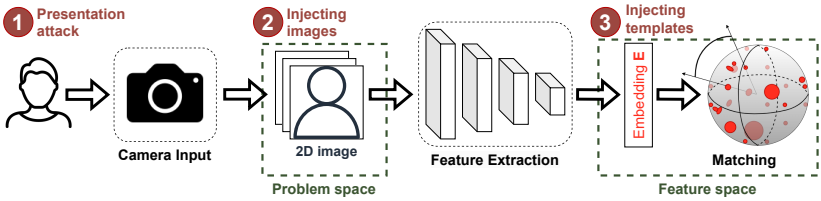


Figure 6.4: The biometric pipeline. The problem space and feature space are clearly indicated, as well as the attack points where an adversary can inject masterkeys.

Another important aspect to consider while 'counting' biometrics is the different 'spaces' biometrics are presented in (shown in Figure 6.4). For example, a 2D facial recognition system will capture a human face in an image. This image is fed to a biometric feature extractor which summarizes the high dimensional input data such that they can be more easily compared by a distance metric, e.g., Hamming distance, Euclidean distance, or cosine distance. Following the terminology introduced by Pierazzi et al. [185], the space of all images considered human faces is the problem space, which is projected into the feature space by the feature extractor. In biometrics, the feature space is also known as the template space. Ideally, we want to consider the unique number of human faces in the input space. However, this is a Sisyphean labor. We do know that a picture consists of 3*width*height pixels and that there are thus $2^{8*3*\text{width}*\text{height}}$ possible inputs. But how many of these images resemble human faces is unclear. Moreover, the number of distinct faces is even harder to determine. Consequently, approaching this problem from the feature space makes the problem more manageable because the distance between two feature vectors resembles similarity. Furthermore, a pre-determined threshold decides upon similarity and dissimilarity.

The feature space provides an attractive alternative for performing the security analysis. However, not every feature vector represents a realistic face. Determining the subspace of the feature space that resembles human faces, also called the manifold (cf. Figure 6.1c), poses yet another challenge. In the past, this challenge has been tackled in discrete feature space by considering the distribution of the Hamming distance between distinct faces [62]. Furthermore, an adversary who finds a masterkey in feature space is not guaranteed to find a 2D facial image that is projected on the target feature vector. Additionally, making a presentation attack instrument (PAI) that resembles the 2D facial image adds more noise. Thus the capabilities of the guessing adversary are also important. The following attack positions exist (also shown in Figure 6.4):

1. Presenting (fake) biometric input to the sensor.
2. Circumventing the sensor and directly inputting captures of biometric data.
3. Injecting biometric templates (feature vectors).

Assuming the strongest adversary provides a lower bound on security, i.e., the adversary cannot be more effective. Therefore, while evaluating the guessing metrics for biometrics, we consider masterkeys in template space. To represent a weaker adversary, we also perform an evaluation relying solely on biometric captures observed in the wild.

We also provide a lower bound on their effectiveness by considering an adversary presenting biometric data of real individuals (attack point 1 in Figure 6.4). A new approach to masterkey attacks that generate masterkeys in problem space should then express their effectiveness in relation to these bounds.

6.4 Effective size of the biometrics space

The effective key space size N is closely related to guessing difficulty. As discussed in Section 6.3, in the case of uniformity, the expected number of guesses is $\frac{N+1}{2}$. Otherwise, N upper bounds the number of guesses [203]. Determining the number of unique biometric signals is challenging because the biometric template space is uncountable, and the fuzzy matching introduces unavoidable overlaps between neighboring templates.

In this section, we tackle those challenges and describe an approach to estimate the size of the biometric template space. Particularly, we look for the minimal number of points that can be distributed in biometric template space such that the overlap is minimized. Our approach builds upon the upper bound of the number of spherical code points [120]. Since our method assumes continuous templates, we demonstrate that it also holds for discrete templates.

6.4.1 Upperbound on the guessing difficulty for biometrics

Determining the number of 'unique' biometrics in input space is infeasible. Although, we can aim to estimate the number of points required to cover the biometric template space. To this end, we draw a parallel between the biometric feature space and spherical codes. Spherical code theory defines an upper bound on the number of code points. Or, in other words, the maximum number of code points that can be distributed on the surface of an n-ball such that every code point is at least an angle θ apart. We built upon this upper bound to determine an upper bound on the minimal number of points such that each point is at most an angle θ apart. Or, in other words, the number of guesses required to cover the whole template space, i.e., the upper bound for guessing difficulty.

Biometric feature extractors and spherical codes

Biometric feature extractors often project biometric signals on the surface of an n-dimensional ball $f : B \to S^{n-1}$: $X_i = f(B_i)$ with radius 1, instead of on the whole $\mathbb{R}^n$. Or, at the very least biometric templates are compared by computing the cosine distance. The cosine distance can be computed by first normalizing the templates (projecting on the unit length n-ball) and then computing the dot product.

The feature extractor is trained to minimize intra-cluster variability and maximize inter-cluster variability. Thus, the feature extractor aims to distribute the training samples on the surface of an n-ball such that biometric samples of the same person are constrained to a dedicated region, while biometric samples of other individuals fall outside this region. This problem is similar to finding a spherical code. Spherical codes are a set of points $X = \{X_1, X_2, \ldots, X_k\}$ on the n-ball $X \subset S^{n-1}$ such that the dot product between every two points is smaller than a given threshold. When each code point is viewed as the centroid of an individual in the training set, the biometric feature extractor can be interpreted as a function that projects the biometric training samples onto a spherical code. This insight allows comparing biometric matching with correcting a specific point on the sphere. Both check whether a point on the sphere falls within a specific spherical cap. If so, the user is accepted, or the point is corrected to the code point.

Despite the many similarities between biometrics and spherical codes, biometric centroids are not spherical code points. Biometric centroids do not satisfy the constraint that the distance between two biometric centroids is less than a threshold. This is to be expected as the feature extractor only knows a small portion of the whole population during training and cannot guarantee perfect separation. Moreover, doing a perfect job is most likely impossible, and trying to do so impedes generalization. Hence, spherical caps belonging to individuals overlap, which introduces false matches.

Upper bound for the number of spherical code points

Finding spherical codes is a challenging problem with optimal lattices only known for $n \leq 5$ and 'special' dimensions 8, 12, and 24 [224]. Even though optimal packings are not known for other dimensions, the vast body of literature on spherical codes does provide an upper bound for the maximal cardinality $M(n, \phi)$ of $X \subset S^{n-1}$ that satisfies any of the following equivalent properties [145]:

1. Spherical caps of angular radius $\phi/2$ around two different code points do not intersect.

2. The Euclidean distances between different code points are $\geq 2\sin(\phi/2)$.

3. The scalar products between two different code points are $\leq \cos(\phi)$.

Those upper bounds are determined for two angle regions: the *small angle* region $0 < \phi \leq \pi/2$ and the *large angle* region $\pi/2 < \phi \leq \pi$. For the the *small angle* region Kabatiansky and Levenshtein [120] formulated the upper bound for $n \to \infty$ as $2^{n*H(\phi)}$ with [145]:

$$H(\phi) = \frac{1+s}{2s}\log_2(\frac{1+s}{2s}) - \frac{1-s}{2s}\log_2(\frac{1-s}{2s}) \tag{6.2}$$

where $s = sin(\phi)$.

For the *large angle* region there are two bounds, one for $-1/n \leq \cos(\phi) < 0$ and one for $-1 \leq \cos(\phi) \leq -1/n$ [195]. This gives the upper bound $G(\phi)$ with [145]:

$$G(\phi) = min\{n+1, \frac{\cos(\phi)-1}{\cos(\phi)}\} \tag{6.3}$$

The problem we are trying to solve to estimate guessing difficulty is related. Instead of finding an upper bound for the number of points on an n-ball such that the angle between every point is at least an angle ϕ, we are looking for the minimal number of points that can be distributed on the surface of the n-ball such that they are at most an angle ϕ apart. To solve our related problem, we look for the spherical code that guarantees that the angle between every two points is at most ϕ. It turns out this holds for a spherical code with angle $\phi/2$.

Lemma 6.4.1. *For a spherical code X with code parameters n and ϕ that maximized its number of code points, the dot product between every two code points is maximum* $\cos 2\phi$.

Proof. Assume a code point x_1 of the spherical code X and its closest, but a different code point is x_2, and the angle between them is $\geq 2\phi$. This implies there exists a third point x_3 where the angle between x_3 and both x_1 and x_2 is $\geq \phi$. Therefore x_3 has to be part of the spherical code, which is a contradiction. In other words, no two code points exist where the angle between them is more than 2ϕ. □

$$\begin{aligned} &\forall x_1, x_2 \in X \mid x_1 \cdot x_2 \leq \cos(2\phi)\ \exists x_3 \in X\ s.t. \\ &\quad x_1 \cdot x_3 \geq \cos(2\phi)\ \ \&\ \ x_2 \cdot x_3 \geq \cos(2\phi) \\ &\qquad\qquad \&\ \ x_1 \neq x_2 \neq x_3 \end{aligned} \tag{6.4}$$

Thus, we can compute the upper bound of the maximal cardinality $M(n, \phi/2)$, which ensures that the angle between every two points on the n-ball is at most ϕ. Most likely,

the number of points found in this manner is an overestimate for our problem. However, this is fine as we are concerned with finding an upper boundary.

Determining the dimensionality n and the angle ϕ

The function to compute the upper bound on the guessing difficulty has two variables, the number of dimensions n and the angle ϕ, the minimal distance between two code points. We determine these two variables by devising ϕ based on the False Non-Match Rate (FNMR), also known as the False Rejection Rate (FRR). Then, we ascertain the dimension n by sampling points on different n-balls, determining the score distribution of the pairwise distances, and computing the FMR with angle ϕ.

In analogy with spherical codes, we choose ϕ as the desired error-correcting capability. Since we know the genuine distribution, we choose ϕ to achieve a specific FNMR. Note that ϕ and the threshold of the biometric system are directly related as $t = 1 - \cos(\phi)$.

At this point, we make an essential assumption. We assume that the error-correcting capability remains constant when the number of dimensions n changes as long as the dimensionality suffices to encode all relevant information. We base our assumption on the fact that the amount of noise on biometric acquisitions of the same individual is bounded and remains constant. Therefore, spreading the same information over more dimensions will not change the intra-cluster variability.

The number of dimensions of the n-ball is computed by randomly sampling points on its surface and calculating the dot product between them. This distribution represents the imposter distribution of a uniformly distributed biometric. Therefore, we can use ϕ to compute the FMR. We then select the lowest n such that the FMR of the original biometric is lower than the computed FMR. Alternatively, when the FMR is very low and hard to determine accurately, the distribution of distance between imposters can be fitted to the distribution of distances of randomly sampled points on the n-ball with, for example, Kolmogorov-Smirnov.

Lastly, the upper bound on the guessing difficulty is computed as $M(n, \phi)$. Note that ϕ is used rather than 2ϕ. 2ϕ provides an upper bound for the number of code points, so there is no overlap between the spherical caps with angle ϕ. However, this also leaves a lot of points on the sphere that will not be corrected to any of the code points. As explained above, to ensure that the guesses are not more than angle 2ϕ apart, we utilize ϕ.

An algorithm to estimate the guessing difficulty

Above, we described the methodology's rationale for estimating the number of guesses required to impersonate the whole population. The algorithm is summarized in Algorithm 2. The estimate obtained by this algorithm is an upper bound on the guessing difficulty. However, a practical adversary that follows the greedy guessing approach described in Section 6.5.2 exceeds this upper bound due to being sub-optimal (cf. Section 6.5.3). In other words, the upper bound found with Algorithm 2 is an upper bound on the number of guesses an optimal adversary would have to perform.

6.4.2 Relationship between discrete and continuous biometric templates

Every binary code of length n can be projected on the unit n-ball by constructing vectors where zero is replaced with -1 and normalizing them (dividing by $\sqrt{n}$). For binary codes, the tolerance t is expressed as the Hamming distance, while the length normalized vectors are compared through the angular distance $1 - \cos(\phi(n))$. If the Hamming distance between two codes a and b is t, the angular distance between the corresponding length normalized vectors is:

$$1 - \cos(\phi(n)) = 1 - a \cdot b = 1 - \frac{n - 2t}{n} = \frac{2t}{n} \tag{6.5}$$

Thus:

$$\phi(n) = \arccos(\frac{n - 2t}{n}) \tag{6.6}$$

We can then apply Algorithm 2 to compute the upper bound of the effective key space. Additionally, the probability mass function (PMF) of the Hamming distance between two random n-bit codewords can be modeled as a binomial with $p = 0.5$ and n equal to the number of (uncorrelated) bits, which implies that the distribution of angular distances between two points on the surface of the n-ball follows a similar shape to the binomial distribution.

As the binomial models the PMF of the Hamming distance between two random n-bit codewords, it should also resemble the imposter distribution. However, Daugman [62] observed that the imposter distribution of n-bit biometric iris templates does not follow the binomial distribution of size n, due to correlations. Instead, Daugman determined the degrees of freedom (DoF) of the binomial that best fits the imposter distribution $\mathrm{DoF} = \frac{p(1-p)}{\sigma^2}$.

In other words, there is a theoretical equivalent system that has DoF uncorrelated bits. Thus, the key length n is reduced to DoF, or it is multiplied with a factor $k = \frac{DoF}{n}$ with $k < 1$. The threshold t is reduced with the same factor, thus the new threshold

Algorithm 2: Algorithm to determine the effective key space of a biometric.

// t is the threshold at the desired FNMR, which determines the angle ϕ
$\phi = \arccos(1 - t)$;
// Randomly sample points on n-balls, compute the angular distance between them and compute the FMR_n at the threshold. Stop if the computed FMR is lower than the FMR of the biometric.
$n = 3$;
// minimal dimension is 3
$\mathrm{FMR}_c = 1$ // initial value for the FMR
while *FMR* $< \mathrm{FMR}_c$ **do**
  // sample k vectors of length n from a uniform distribution
  $X = random.uniform((k, n))$;
  $X = normalize(X)$;
  // normalize the vectors
  // compute the angular distance between the sampled vectors and flatten the upper triangle without the main diagonal in a 1D array
  $D = upper.triangle(1 - XX^T)$;
  $\mathrm{FMR}_c = \frac{length(D[D<t])}{length(D)}$;
  $n = n + 1$;
end
// If it is hard to determine the FMR accurately, n can be computed by finding the best fitting distribution of distances D for I, the angular distances between biometric samples from different individuals, i.e., represents the impostor distribution.
// Based on n and ϕ compute the effective key space size of the biometric.
$s = sin(\phi)$;
$H = n\frac{1+s}{2s}\log_2(\frac{1+s}{2s}) - \frac{1-s}{2s}\log_2(\frac{1-s}{2s})$;
return H

$T = kt$. This implies that the code points projected on the DoF-ball are separated by $\cos(\phi(DoF)) = \frac{DoF-2T}{DoF}$. In Algorithm 2, we assumed that ϕ is constant when the dimensionality is reduced, which holds:

$$\cos(\phi(DoF)) = \cos(\phi(n)) \tag{6.7}$$

$$\iff \frac{DoF - 2T}{DoF} = \frac{n - 2t}{n} \tag{6.8}$$

$$\iff \frac{kn - 2kt}{kn} = \frac{n - 2t}{n} \tag{6.9}$$

6.5 Measuring guessing difficulty

Guessing metrics capture the effectiveness of a brute-force adversary. The brute-force adversary aims to impersonate their victims by guessing the authentication 'secret' (e.g., a password). The brute-force adversary optimizes their guessing strategy by estimating the distribution of these authentication secrets. The distribution can be gauged based on publicly available data, e.g., password database leaks.

Notation: The assumption is that the brute-force adversary knows the distribution of secrets χ. However, more realistically, the adversary estimates the real distribution χ by collecting a set of secrets $\hat{\chi}$ of size M. The probability $p(x)$ of individual secrets x in the attacker dataset $\hat{\chi}$ estimates the expected attack success rate for guess x. We denote the estimated attack success probability of x as p_x. The adversary then determines the optimal guessing strategy by ranking the events x in descending order: $p_1 \geq p_2 \geq \cdots \geq p_N$. N corresponds to the least successful guess, and $\sum_i p_i = 1$. Therefore, we refer to specific events x by their index i in the optimal guessing sequence.

6.5.1 Guessing metrics for knowledge-based factors

Disclaimer: This section overlaps with Section 5.4.2.

The first large-scale study on guessing metrics for knowledge-based factors was performed by Bonneau et al., both on passwords [33] and on PIN codes [35]. Their approach was later used to study the effects of design decisions on the security of knowledge-based authentication systems. For example, Markert et al. [148] study the influence of blacklists on pin code security.

Bonneau et al. model password selection as a random draw $X \xleftarrow{R} \chi$. The authentication system then serves as an oracle that answers the question: "is $x_i = X$?". Guessing

metrics then capture the efficiency with which an adversary can learn individual passwords X by formulating queries x_i to the oracle. As such, the guessing metrics can be defined as a function of the success rate of individual guesses p_i. Ordering them as monotonically decreasing sequences provides the optimal guessing sequence. As knowledge-based factors are countable events, the expected success rate of a guess p_i is equal to the rate of occurrence of a specific password $x_i \in \chi$.

One of the main challenges when defining guessing metrics for passwords is not letting them depend on the number of unique passwords N in χ. Due to the long tails representing strong passwords that occur infrequently, N converges slowly or w.r.t. the attacker set size (sample size) M. Therefore, **Hartley entropy** [203] $H_0 = \log N$ cannot be used. Furthermore, on top of being hard to determine accurately, the significant fraction of strong infrequent passwords leads to optimistic values for metrics such as **guesswork** [152]:

$$G(\chi) = E[\#_{guesses}(X \xleftarrow{R} \chi)] = \sum_{i=1}^{N} p_i \cdot i \tag{6.10}$$

Alternatively, **min-entropy** [203] $H_\infty = -\log p_1$ can be estimated effectively but only serves as a worst-case metric as it is solely based on the most frequent password.

In addition to being hard to measure, the aforementioned metrics do not consider the fact that an adversary can decide to cease guessing hard passwords. An adversary can give up on hard-to-guess passwords and instead focus on easier cases. **β-success-rate** [38] is the fraction of accounts that can be broken by an adversary limited to β guesses:

$$\lambda_\beta(\chi) = \sum_{i=1}^{\beta} p_i \tag{6.11}$$

Alternatively, the **α-work-factor** [187] defines the number of guesses the adversary needs to do to achieve a desired success rate α, or put differently, to break a fraction α accounts:

$$\mu_\alpha(\chi) = \min\left\{ j | \sum_{i=1}^{j} p_i \geq \alpha \right\} \tag{6.12}$$

Lastly, the α-work-factor and β-success-rate do not account for the fact that the adversary stops guessing as soon as they guessed correctly. This is accounted for by the **α-guesswork** [33]:

$$G_\alpha(\chi) = (1 - \lambda_{\mu_\alpha})\mu_\alpha + \sum_{i=1}^{\mu_\alpha} p_i \cdot i \tag{6.13}$$

In cryptography, the complexity of guessing a secret correctly is often gauged as the *effective key length*. The effective key length provides a logarithmic scale expressed

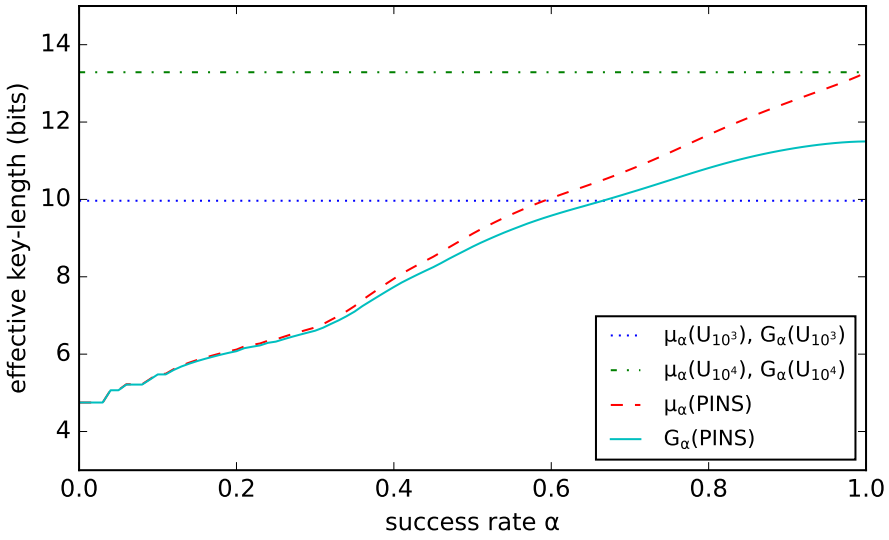


Figure 6.5: Comparison of guessing difficulty between two uniform probability distributions of 10^3 and 10^4, and the probability distribution of real-world 4-digit PINs [35].

in bits. Thus, adding a bit of security requires double the work. The guessing metrics are transformed into bits by calculating the size of the uniform distribution that yields the same values for the guessing metrics. The $\log_2$ of the population size provides a measure directly comparable to the effective key length. The guessing metrics are converted to bits as follows:

$$\tilde{\lambda}_\beta(\chi) = \log\left(\frac{\beta}{\lambda_\beta(\chi)}\right) \tag{6.14}$$

$$\tilde{\mu}_\alpha(\chi) = \log\left(\frac{\mu_\alpha(\chi)}{\lambda_{\mu_\alpha}}\right) \tag{6.15}$$

$$\tilde{G}_\alpha(\chi) = \log\left(\frac{2G_\alpha(\chi)}{\lambda_{\mu_\alpha}} - 1\right) + \log\left(\frac{1}{2 - \lambda_{\mu_\alpha}}\right) \tag{6.16}$$

α-guesswork and α-work-factor and their relationship for 4-digit pincodes are depicted in Figure 6.5. For references, we added the α-guesswork and α-work-factor for the uniform distributions of size 10^3 and 10^4.

6.5.2 Guessing metrics for biometrics

Disclaimer: This section overlaps with Section 5.4.3.

As described above, guessing metrics enable repeatable and mathematically sound security evaluations that account for non-uniformity. Adapting them for biometrics requires dealing with two other challenges, i.e., uncountability of the events x and fuzzy matching (cf. Section 6.3). These challenges impact how to determine and interpret p_i and how to deduce an ordered guessing sequence. p_i can no longer be assessed through how frequently a secret x occurs. Furthermore, fuzzy matching complicates establishing an optimal guessing sequence $p_1, p_2, \ldots, p_k$.

Van hamme et al. [243] first described how to deal with uncountability and fuzzy matching. They established that the greedy approach is the most optimal known approach and define p_i:

> p_i is the probability of a (false) match at the i-th guess given that the previous guesses were non-matches. The optimal sequence can be approximated using a greedy approach by choosing the guess x_i at iteration i as $x_i = \text{argmax}_{x_i}\{p_i\}$.

Algorithm 3 computes the guessing sequence $p_1, p_2, \ldots, p_k$ using the greedy approach. Once the guessing sequence is established, the guessing metrics can be calculated.

The proof that the greedy approach is the best-known algorithm to solve this problem is straightforward. In essence, determining the optimal guessing list for k guesses is solving a maximum coverage problem. The maximum coverage problem is a well-known optimization problem that is NP-hard, and for which the greedy algorithm is the best possible polynomial time approximation [76]. The objective of the maximum coverage problem is to find a collection of exactly k subsets $S' \subseteq S$, such that the number of elements covered by these subsets is maximized, provided that:

- The universe U is a finite set that consists of n elements: $U = \{e_1, e_2, \ldots, e_n\}$.
- There is a collection S with m subsets of U: $S = \{S_1, S_2, \ldots, S_m\}$ where each $S_i \subseteq U$.
- k, the number of subsets that can be chosen: $1 \leq k \leq m$.

For biometrics, the universe consists of the users to be impersonated. The subsets are defined by the biometric input samples, where each subset contains the users for which that biometric input sample would be accepted. And, as mentioned above, k is the adversary's guessing budget.

Algorithm 3: Algorithm to determine the guessing order of biometric samples and corresponding attacker success rates p_i.

```
// C is the matrix with the centroids with dimensions
   [c, f]. B is the matrix with the biometric vectors
   with dimensions [m, f]. t is the threshold.
```
Data: C, X, t
```
// Compute the angular distance between unit length
   biometric samples and centroids
```
$A = 1 - XC^T$;
```
// 1 for all entries bigger than threshold, 0 for the
   others
```
$A = Ones[A > t]$;
$guesses$ = [];
```
// Find greedy guesses
```
while A *!= 0* **do**
```
   // Sum along the rows and find the maximum and its
      index
```
   $guesses.append(\max_i(\sum_j^c A_{i,j})/c)$;
   $ind = argmax_i(\sum_j^c A_{i,j})$;
```
   // Change matrix to accommodate for broken
      centroids
```
   $A = A * (1 - A[ind, :])^T$ `// element wise multiplication over all rows`

end
return $guesses$

As mentioned in Section 6.3.3 we consider two types of biometric input samples that determine the subsets:

1. The biometric samples observed in the dataset. We refer to this evaluation as the **wolf attack**. This could be considered a lower-bound on the effectiveness of masterkey attacks. A biometric sample that is specifically forged to increase its impersonation rate should be more effective than naturally occurring ones.

2. Forged biometric samples that only exist in feature space. We call this the **feature space masterkey attack**. This evaluation provides an upper-bound on the effectiveness of masterkey attacks. Because the found attack point is most optimal in feature space, a masterkey attack constructing a biometric sample in feature space will never be more effective as the optimal attack point in feature space is not guaranteed to be reachable from problem space.

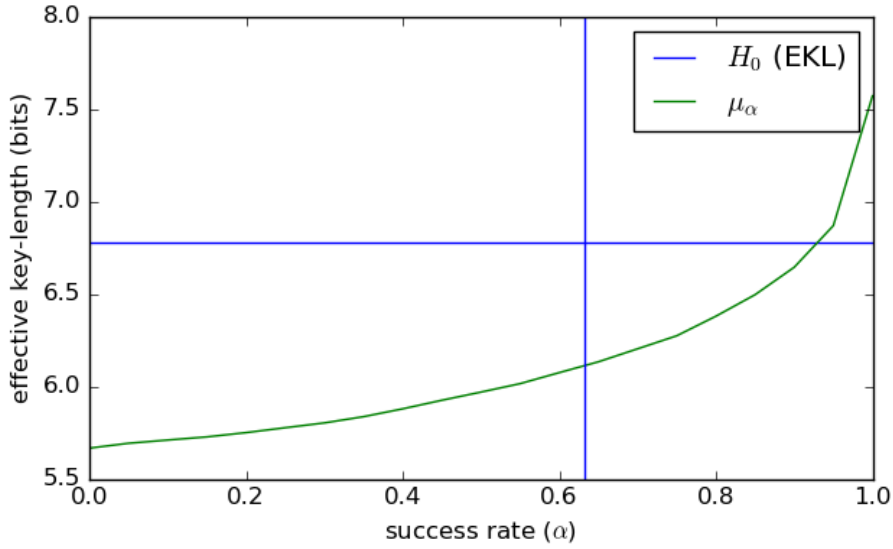


Figure 6.6: The relationship between the effective key length (EKL) found with Algorithm 2 and the α-work-factor μ_α. These metrics are computed for a hypothetical biometric where the users are sampled uniformly on the 15-ball and the threshold is set to $\frac{7}{15}$

6.5.3 Relationship with effective key length

We suggested two approaches to estimate the security offered by biometrics: (1) the upper bound on the size of the effective key space N and (2) guessing metrics. These two approaches are related. It has been shown that the greedy approach to the set coverage problem achieves an approximation ratio of $f = 1 - \frac{1}{e} \approx 0.63$ [104]. This implies that the number of impersonated accounts x for a specific number of guesses k is between $f * OPT < x < OPT$, with OPT the number of impersonated accounts for an optimal solution. Thus, in the worst-case scenario, all accounts could have been guessed correctly at $\alpha = f$ (but it is more likely for higher α). This implies that the upper bound should be bigger than μ_α for $\alpha < 63\%$. Due to the suboptimality introduced by the greedy algorithm, μ_α can pass the upper bound on N when $\alpha \geq 63\%$. Figure 6.6 depicts this relationship for a hypothetical biometric sampled uniformly on the 15-ball with a threshold of $\frac{7}{15}$.

6.5.4 Usage in security evaluations

Disclaimer: This section is a summary of Section 5.4.4.

The selection of appropriate parameters for the guessing metrics, including the number of guesses β and the fraction of accounts targeted for attack α, depends on the capabilities of the adversary. Previous research in password cracking has found that a reasonable choice is around 10 guesses (referred to as λ_{10}) [33]. However, in the case of phone unlock with a PIN, there exists an exponential time penalty for successive guesses, where the initial 5 guesses are unrestricted and subsequent ones become increasingly costly [148]. Therefore, a range of values from λ_5 to λ_{100} is considered in such scenarios.

Regarding biometric phone unlock, the usual practice is to limit the number of consecutive attempts to 5 (for iOS) or 3 (for Android unlock classes 1 and 2) before fallback to PIN codes [155]. The specific number of attempts allowed for class 3 Android biometrics is unspecified, though it is presumed to be higher. In cases where keystroke biometrics are employed as a second factor in remote authentication systems, the number of permitted guesses is likely to be comparable to that of stand-alone password authentication, typically set to 10. Consequently, this work examines the parameter range from λ_3 to λ_{10}. Additionally, $\lambda_1 = H_\infty$ is used as a worst-case metric, providing a lower bound and evaluating the effectiveness of wolf and masterkey samples.

Adversaries that obtained access to a database leak or have access to interfaces can conduct unthrottled guessing. For an unthrottled adversary, α-guesswork offers a comprehensive assessment of the guessing difficulty. Compare different approaches for authentication can be done through a "guessing curve", where all values of α are considered. Instead of providing the whole curve, prior work often reports on two specific values for α, i.e., $\alpha = 0.25$ and $\alpha = 0.5$.

6.6 Experimental setup

6.6.1 Keystroke dynamics

Disclaimer: This section overlaps with Section 5.5.

Pre-processing We analyze keystroke data collected from physical keyboards, which include the keycode, downstroke (press) timing, and upstroke (release) timing. To prepare the data for further analysis, we perform a typical preprocessing step,

focusing on extracting the relative timings of bigrams [9, 4]. The following four relative timings are considered:

- **Hold latency**: Time between the key down and key upstroke of the same key.
- **Inter-key latency**: Delay between the key upstroke of the current key and the key downstroke of the following key.
- **Press latency**: Time between two consecutive key downstrokes.
- **Release latency**: Time between two consecutive key upstrokes.

These relative timings, along with the corresponding keycode, result in five features for each key pressed. Consequently, our input to the model consists of sequences of variable length N, leading to an input dimensionality of $N \times 5$. The keystroke timings are converted to seconds, and the javascript keycodes are divided by 255 for standardization.

Dataset and model architecture For our evaluation and model training, we utilized the dataset collected by Dhakal et al. [70], which comprises 15 sentences typed by over 168k users. This dataset is the largest scale-free text-typing dataset available. The typing behavior was recorded through a web form where participants were required to memorize a common sentence and then reproduce it by typing it out. The sentences were randomly selected from a corpus of 1525 common English sentences, each containing at least three words.

After preprocessing the data, we had approximately 161k individuals, each providing 15 sequences. We divided this data into different subsets as follows:

1. **Training set**: Comprising 40k users used for training the keystroke dynamics system.
2. **Validation set**: Consisting of 10k users used for early stopping and determining the authentication system threshold.
3. **Test set**: Comprising 110k users used for evaluating the attack and performing demographics analysis.

Additionally, for training a second keystroke algorithm, we further divided the test set into a second training set of 40k users and a test set of 70k users.

Our keystroke model is designed to extract identity vectors from keystroke timings. The identity vectors are compared by computing the cosine distance between them. To train

the model, we employ a deep learning architecture with the triplet loss function [213]. This architecture comprises two cascaded gated recurrent units (GRUs) with 64 units, with a batch normalization layer placed between them. The output of the last layer is normalized, resulting in final identity vectors with a norm of 1 and a length of 64.

Furthermore, we adopt a second model identical to the TypeNet model [4]. This model is also trained with the triplet loss and employs two Long Short-Term Memory (LSTM) units, each containing 128 units. As a result, the biometric feature vectors from this model also have a norm of 1 and a length of 128.

Biometric system and threshold selection The comparison function of our biometric system utilizes the cosine distance function $1 - \cos(\theta)$, where θ is the angle between the enrollment samples and the query sample. To determine the appropriate threshold, we aim to achieve a false non-match rate (false rejection rate) of 5% in the validation set. At this threshold, the GRU and LSTM systems exhibit a false acceptance rate (FAR) of approximately 1.25% and 1.28%, respectively, on the validation set. The Detection Error Tradeoff (DET) curves of both systems are presented in Figure 6.7. Throughout the experiments, we use five enrollment samples (i.e., typed sentences) for each enrolled user.

6.6.2 Face recognition

Dataset and model architecture For face recognition, we used a facenet model available online [211] trained on the VGGFace2 [45] dataset. The model was used to extract biometric templates from a cleaned version of the MS-Celeb-1M dataset [116], which contains around 4.5 million images of $78,000$ users. As we require 5 enrollment samples per user, around $20,000$ users remained, which gives a total of 1 million face images.

Biometric system and threshold selection Again we use the cosine distance function to compare users and set the threshold such that an FNMR of 5% is achieved. We also discretize the continuous biometric templates, which is the first step to applying biometric template protection. The technique used for discretized is described by Argones et al. [20]. The discretized features are compared with the Hamming distance. We set the threshold such that the FNMR is as close to 5% as possible. This leads to an FMR of 0.80% for the continuous face biometric system. The discrete face recognition system outperforms the continuous system and has an FMR of 0.44%. The DET curves of both biometric systems are shown in Figure 6.7. Note that all systems have very similar EERs, and would often be considered equally good.

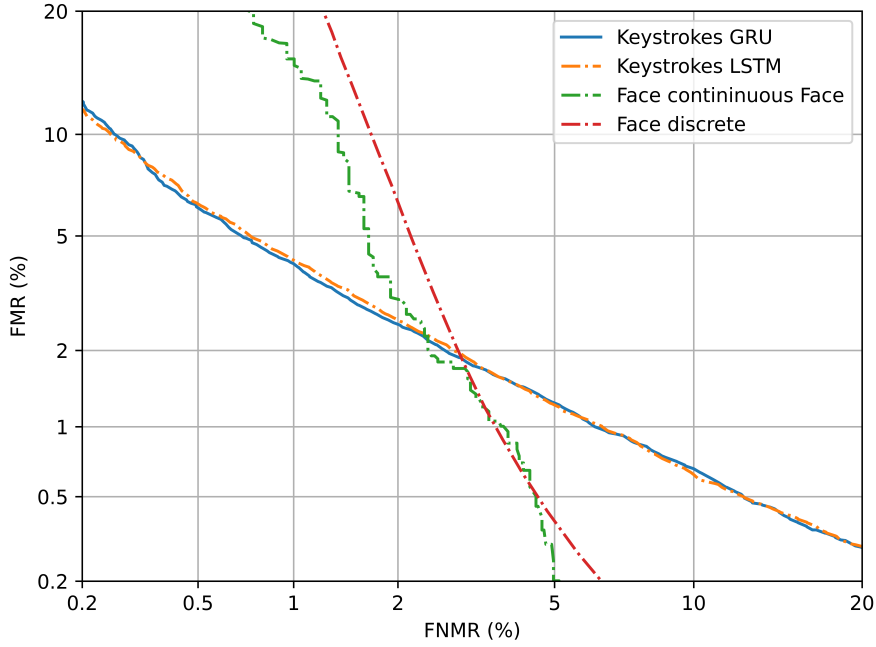


Figure 6.7: DET curve of the keystroke and face biometric systems.

6.7 Effect of sample size

The probability of a match for the i-th guess p_i is not exact and is approximated based on a random draw from the biometric distribution[1]. In the context of passwords, it has been observed that computing the guessing metrics directly from the random draw tends to underestimate their values. For biometrics, we have found a similar effect. Figure 6.8 illustrates this phenomenon for free-text keystroke dynamics using the wolf attack. The graph clearly demonstrates that the guessing metrics increase as the number of unique users (M) in the sample increases.

Just like in the case of passwords, there exists a stability region concerning the sample size for certain metrics, as shown in Figure 6.8. The partial guessing metrics appear to stabilize when there are about 2,000 unique users in the sample. However, one exception is $H_0 = \log(N)$, which suggests that biometric samples from more remote areas of the n-ball are present in the sample. As the sample size increases, the probability

[1]Theoretically, it is possible that an adversary knows the exact distribution of enrolled users. However, an adversary with such knowledge probably does not have to guess in the first place.

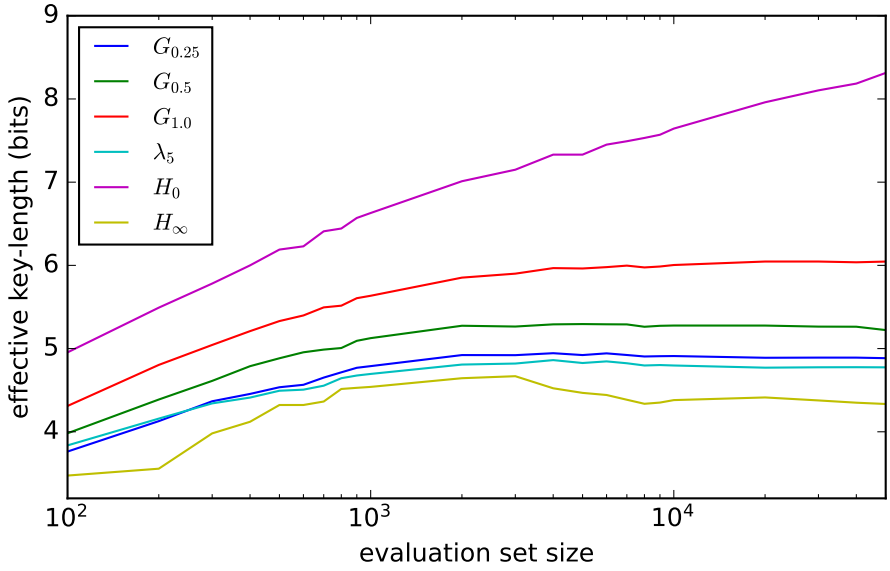


Figure 6.8: Wolf Attack Guessing Effectiveness: The y-axis displays guessing metrics represented as effective key lengths, while the x-axis shows the number of users in the evaluation set (logarithmic scale). This illustrates the adversary's performance executing a wolf attack.

of observing more remote individuals increases. These individuals are remote in the sense that they fall just outside of more dense regions on the n-ball. As a result, more inefficient guesses with significant overlap with prior guesses are required to cover them. Consequently, the guessing metrics are not suitable for estimating N, mainly due to the suboptimal greedy approach used to select guesses. Fortunately, Section 6.4 provides a description of how to estimate N accurately.

The main challenge lies in determining the number of unique users necessary to compute stable performance metrics. In the case of passwords, the primary challenge arises from the large number of infrequent passwords, which requires a heuristic that demands a minimum number of occurrences of a specific password to compute the guessing metrics accurately. In biometrics, this methodology is not feasible since the events are uncountable. The rest of this section investigates why p_i is overestimated and proposes a heuristic to determine the sample size required for stable guessing metrics.

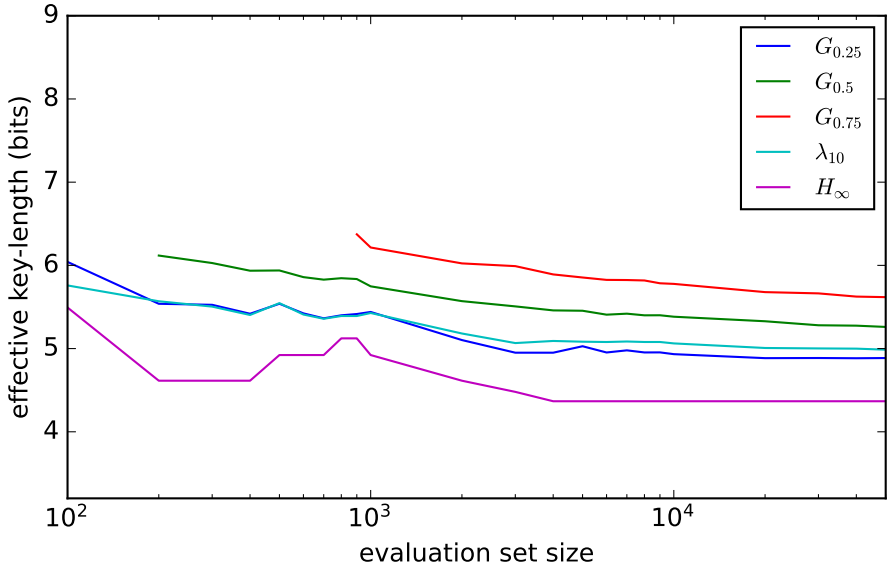


Figure 6.9: Wolf Sample Transferability: Adversarial guessing effectiveness when using wolf samples from an attack dataset to impersonate users in a target dataset. The x-axis shows the unique number of users in the attack dataset on a logarithmic scale, while the y-axis represents guessing metrics expressed as effective key lengths. The target dataset contains 50,000 unique users, and its size is fixed.

6.7.1 Impact of statistical hazards

Figure 6.8 provides clear evidence that the guessing metrics are underestimated. This implies that p_i tends to be overestimated. We explore two reasons for this overestimation:

Accidental wolves

Wolves can occur by accident in their sample. To demonstrate this effect, we assume a perfect biometric system with a uniform distribution on the n-ball. In such a scenario, no wolves should exist, and every user should have the same chance to impersonate others. However, when computing the guessing metrics, initial guesses tend to be more effective, suggesting that some individuals are mistakenly considered more effective than their peers. We illustrate the effect of accidental wolves by sampling uniformly from the 15-ball 1000 times with different sample sizes (M) and calculating

Sample size	1^{st} percentile p_1	99^{th} percentile p_1
100	5%	9%
500	3.2%	4.6%
1.000	2.9%	3.7%
5.000	2.26%	2.58%
10.000	2.11%	2.33%
50.000	1.88%	1.97%
100.000	1.83%	1.89%

Table 6.2: The 1^{st} and 99^{th} percentiles of computing p_1 for 1.000 different samples taken uniformly from the 15-ball. The sample size is varied from 100 to 50.000. The angular distance is $\frac{7}{15}$, which gives an FMR of $\approx 1.67\%$.

p_1. Table 6.2 presents the observed values in the 1st and 99th percentiles, clearly showing the existence of accidental wolves, whose impact diminishes with increasing sample size

Transferability of wolves

Determining the midpoint of clusters on the n-ball accurately requires a sufficiently large number of samples. The adversary does not want to compute the guess that maximizes their attack success rate, only to find out it does not generalize to the biometric database they are attacking. We demonstrate the impact of sample size on the transferability of wolves by using keystroke biometrics and sampling two non-overlapping datasets. The first dataset serves as an attacker dataset to determine the wolf samples in the guessing sequence. Then, this guessing sequence is used to compute the guessing metrics within the second dataset (the target dataset). Figure 6.9 shows the results for varying sizes of the attacker dataset, while the target dataset remains fixed at $50,000$ unique users. The experiment reveals that the guessing metrics decrease with the sample size of the attacker dataset, indicating that wolf samples found with a small number of users in the attacker dataset are ineffective in the target dataset. As the sample size reaches $5,000$ unique users, the metrics stabilize, suggesting that the guessing difficulty is accurately gauged.

In Figure 6.9, $G_{0.5}$ can only be computed with at least 200 individuals in the attacker dataset, while calculating $G_{0.75}$ requires 1,000 individuals. This inability to compute these metrics is due to the wolf samples found in the attack dataset impersonating an insufficient number of individuals. When the adversary has access to a bigger dataset containing more biometric acquisitions, they will be able to establish longer guessing sequences. Table 6.3 shows the fraction of individuals that could not be impersonated

Attacker dataset size	100	200	500	1.000	2.000	5.000	10.000	20.000	50.000
Percentage individuals not impersonated in the target dataset	63.68	46.14	31.22	19.42	11.74	6.68	4.24	2.35	1.21
Length of the guessing sequence (= max number of guesses)	31	45	73	99	129	161	200	245	314

Table 6.3: The target dataset's fraction of non-impersonated individuals is measured during the transfer of wolves from an attack dataset. This transfer involves exploring different attack dataset sizes and the adversary's utilization of various numbers of wolves in the guessing sequence.

in the target dataset and the length of the guessing sequence established in the attack dataset.

6.7.2 The region of stability

In the case of passwords, the law of large numbers allows us to estimate p_i for frequently observed events reliably. As the sample size increases, the relative number of events surpassing the minimal number of observations also increases. Beyond a certain point, further increasing the sample size only marginally affects the fraction of events exceeding the minimal number of observations, enabling the determination of guessing metrics for larger α values.

In biometric systems, p_i tends to be overestimated, but this error diminishes with an increase in the number of unique individuals in the sample. Instead of using the number of biometric acquisitions, we express the sample size as the count of unique individuals (M). To determine the appropriate sample size M leading to stable guessing metrics, we aim to ensure that p_i is not significantly overestimated and the midpoints of the clusters on the surface of the n-ball are approximated.

From experiments on keystroke dynamics shown in Figures 6.8 and 6.9, it seems that the metrics stabilize around a sample size of $10,000$. To verify this, we randomly sampled 100 subsets from our corpus of 110,000 users, repeating this process four times with varying subset sizes. The results indeed show that as α increases, the metrics stabilize more rapidly. The most effective attack point, which requires around $10,000$ users to achieve a fault of approximately 0.1 bits, presents the greatest challenge. On average, p_i with $i < 10$ is accurately determined from a sample size of $2,000$ users. The effective key length is approximately 8.07 bits, thus, about 270 guesses are required to impersonate all users through guessing. Hence, $2,000$ users are only around 7 times that number. For stable measures of H_∞, around $10,000$ users are required, which is approximately 37 times the effective key length. The heuristic we propose is to require anywhere between 10 and 40 times the effective key length in the number of users depending on what guessing metrics must be gauged within a 0.1 bit error. In the

	110.000	10.000	5.000	2.000	1.000
H_∞	4.32	4.25 - 4.56	4.17 - 4.61	4.16 - 4.66	4.03 - 4.68
$G_{0.25}$	4.85	4.85 - 4.97	4.83 - 4.97	4.73 - 4.97	4.66 - 4.95
$G_{0.5}$	5.20	5.22 - 5.34	5.20 - 5.34	5.12 - 5.31	5.00 - 5.24
$G_{0.75}$	5.55	5.60 - 5.68	5.57 - 5.68	5.46 - 5.62	5.32 - 5.52

Table 6.4: The guessing metrics were calculated on a complete dataset consisting of 110,000 users, as well as on random subsamples of a specific size drawn 100 times. The results include the 1^{st} and 99^{th} percentiles for the subsamples.

remainder of this chapter we will use 10.000 users to compute the guessing metrics for keystroke biometrics (when possible) and around 20.000 users for face authentication, which has an effective key length of 9.09 bits.

6.8 Analysis of face and keystroke biometrics

6.8.1 Comparing face and keystrokes with knowledge factors

We assess the security of the face and keystroke authentication systems described in Section 6.6. To establish a lower bound on security, as mentioned in Section 6.5.2, we evaluate the systems using wolf attacks, which represent biometric data in real-world scenarios. Conversely, the masterkey attack provides an upper bound.

For the masterkey attack, we employ the CMA-ES [101] optimization framework to search the template space and identify the most effective attack points. However, it is essential to note that CMA-ES is a black-box optimization method and may not guarantee finding the most optimal point. Previous research [243] explored alternative approaches like k-means clustering, but CMA-ES seemed to be the best option. Due to difficulties with convergence in the original 512-dimensional space of face templates, we applied PCA to reduce the dimensionality to 64. Once a candidate masterkey was found, its effectiveness was tested by transforming it back to the original 512 dimensions and calculating the percentage of individuals it could impersonate. The values of H_0 for biometrics were computed using Algorithm 2.

In line with the heuristic described in Section 6.7.2, we used a sample size of 10,000 users for evaluating keystroke dynamics and doubled that number to 20,000 for face authentication, considering the effective key space is 1 bit higher for faces.

The analysis results are presented in Table 6.5. The findings indicate that face recognition offers similar security to a 4-digit PIN code against adversaries with

a lower number of guesses, which is often the case in biometrics and 4-digit PIN codes due to rate limiting. Although the effective key space of a 4-digit PIN is considerably higher than that of a face recognition system, the security levels are similar, suggesting that non-uniformity affects biometrics significantly less than PIN codes.

Furthermore, we observed that the masterkey attack is much less effective for face recognition compared to keystroke dynamics. This discrepancy could be attributed to two reasons: (1) CMA-ES's performance is hindered due to dimensionality reduction, and (2) CMA-ES may not be able to find attack points that are even more effective. We hypothesize that semantics in keystroke dynamics play a crucial role. Previous work on swiping behavior [217] has shown that certain swipe patterns can match many users, and a similar effect may be present in keystroke dynamics, where specific keycode combinations serve as better candidates for masterkeys.

The False Match Rate (FMR) of the keystroke system is 1.25%, while that of the face recognition system is 0.8%. Taking the $-\log(\mathrm{FMR})$ yields effective key lengths of 6.32 and 6.9 bits, respectively. As subsequent guesses become less efficient due to overlap, higher effective key lengths are computed with Algorithm 2.

6.8.2 Choosing the superior model

The most crucial factor when determining superior recognition capabilities between models is the False Match Rate (FMR). However, as discussed in Section 6.2, relying solely on FMR as an aggregate metric may not provide a comprehensive understanding of the system's performance. To support this claim, we computed the guessing metrics for four different biometric systems, adopting a wolf attack perspective.

For keystroke biometric systems, we used a sample size of 10,000 users to calculate the guessing metrics, while for face systems, we used 20,000 users. All systems exhibited a False Non-Match Rate (FNMR) of 5

Based on the FMR, one might favor the GRU keystroke system over the LSTM and the discretized face system over the continuous counterpart. However, when considering the guessing metrics illustrated in Figure 6.10, a different choice emerges.

For instance, the LSTM provides slightly more security than the GRU for lower values of α, which is a crucial region as adversaries are often throttled, and more security is desired in such cases. As for face biometrics, the evidence supporting the choice of the continuous system over the discretized system is even more compelling. The continuous face authentication system offers significantly more security for all values of α.

		H_∞	λ_3	λ_5	λ_{10}	$G_{0.25}$	$G_{0.5}$	H_0
Keystroke dynamics	Wolf	4.38	4.69	4.8	5.02	4.91	5.28	8.07
	Masterkey	3.66	3.82	3.92	4.18	3.86	4.06	
Face recognition	Wolf	4.98	5.33	5.57	5.9	6.31	7.37	9.09
	Masterkey	4.82	5.03	4.93	5.37	5.52	6.71	
Knowledge factors	Password [33]	6.5	/	/	9.1	17.6	21.6	NA
	4-digit Pin [35]	4.75	5.22	5.5	5.91	6.32	8.78	13.29

Table 6.5: Guessing difficulty for keystroke dynamics and face recognition. Two different attack techniques are reported, the wolf attack, which provides an upper bound, and the masterkey attack, which provides a lower bound. The results are compared with the entropy of traditional knowledge factors such as passwords [33] and PINs [35]

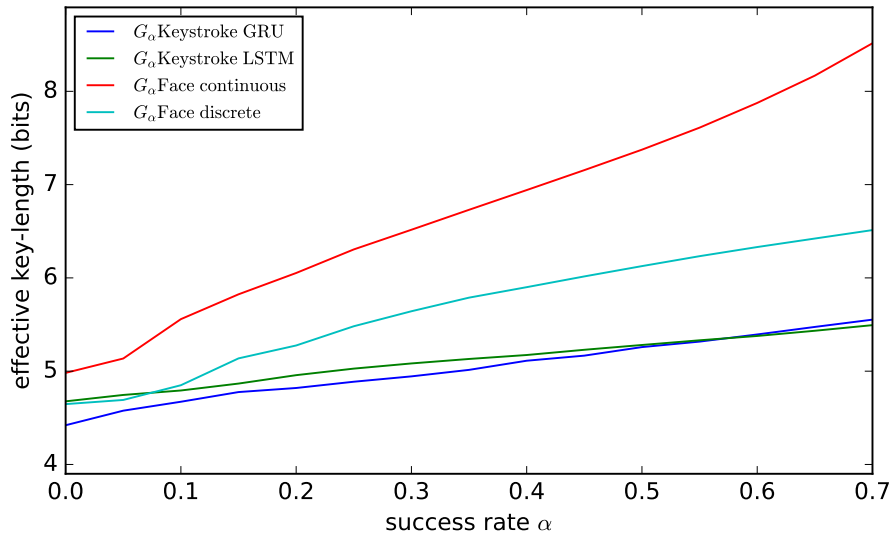


Figure 6.10: Change in α-guesswork (G) with increasing success rate α for the four biometric authentication systems (2 keystrokes and 2 face).

6.8.3 Dictionary specificity

When the adversary possesses knowledge of certain soft-biometric traits of their victim, they can use a dictionary (subset) of individuals sharing the same trait. To investigate the impact of this knowledge on the demographics of the victim, we conducted demographic splits in our experiment, leveraging the masterkey perspective and evaluating it on keystroke dynamics.

We set a requirement of at least 2,000 users in a demographic split for it to be considered. However, for subsamples with more than 10,000 users, we limited the evaluation to 10,000 randomly selected users. As shown in Table 6.6, a more specific dictionary indeed enhances the attacker's guessing efficacy. Nevertheless, not all information is equally useful from the adversary's perspective.

Our findings indicate that knowledge of the victim's typing speed, the number of fingers used, and even their native language proves to be most useful for the attacker. Additionally, having information about whether the victim is above 50 years old provides the adversary with a significant advantage. Knowing the victim's sex is still useful but to a considerably lesser extent.

6.9 Discussion

In this discussion section, we critically examine the extent to which the proposed evaluation framework meets the identified requirements outlined in Section 6.2.4. The focus is on evaluating the framework's effectiveness in addressing the *distribution of biometric mismatches*, *furnishing practical revelations*, and *incorporating a diverse input scope*.

Now turning to the evaluation framework itself, we introduce a methodology designed to assess the security of biometric systems against brute-force adversaries. Leveraging guessing metrics, our framework facilitates a comprehensive evaluation, considering the distribution of biometric mismatches. The resulting metrics are tailored for direct comparison with knowledge-based factors, contributing to a holistic understanding. Additionally, the framework extends its reach beyond the immediate dataset by incorporating inputs not explicitly present in the evaluation set.

As we evaluate the proposed framework against the identified requirements, it becomes evident that our approach makes notable strides in addressing the distribution of biometric mismatches, providing practical insights, and incorporating a diverse input scope. However, a nuanced exploration reveals certain limitations that warrant consideration and pave the way for future investigations

The cost of guessing. The proposed framework lacks consideration for the cost of guessing. For instance, the costs associated with inputting a password differ significantly from creating presentation attack instruments (PAI), such as printing out a photo. This introduces a nuanced dimension where 1 bit of security for passwords may not be equivalent to 1 bit of security for biometrics. Additionally, the effectiveness of presentation attacks is mitigated by liveness detection, increasing the overall cost of these attempts.

Further exploration from an economic perspective should account for various aspects of threat models, including the efforts required to select an optimal candidate and the cost associated with inputting that candidate. Selecting an optimal candidate involves the adversary obtaining knowledge about the distribution of 'secrets,' requiring the construction of a biometric feature extractor. Although the impact of the choice of feature extractor might be minimal due to the generalization of systems [243], variations in the approach to the recognition problem could emphasize different traits.

Constructing PAIs for biometrics involves additional considerations, particularly in bypassing liveness checks. Depending on the biometric modality, this process varies in complexity and cost. For example, presentation attacks against 3D face recognition may require masks, significantly increasing the cost compared to attacks against keystroke biometrics, which might not necessitate any PAI."

Identifying uncommon input. Our evaluation framework relies on a black-box optimization algorithm (CMA-ES) to identify masterkeys in feature space. This optimization approach plays a crucial role in determining the candidates considered for guessing. It's important to note that the efficiency of this optimization algorithm directly impacts the identification of effective masterkeys, thereby influencing the overall security of the system.

Ubiquity of biometric wolves. An intriguing question arises when considering more distinctive biometrics, such as the iris: do these biometrics also have 'wolves' and masterkeys? The concept of biometric wolves and masterkeys hinges on the fact that different individuals may exhibit similarities in their biometrics, leading to collisions in the feature space. The likelihood of such collisions diminishes when dealing with highly distinctive biometrics. Consequently, our evaluation framework may require a larger user dataset to provide reliable estimates. While a perfectly uniform biometric distribution simplifies the estimation of brute-force adversary effectiveness, we contend that this scenario is improbable, given the inherent non-randomness often present in biometric data, such as genetic influences."

6.10 Conclusion

This chapter introduces novel metrics for assessing the security of biometric authentication systems, replacing the traditional false match rate with a more robust security analysis. By employing these metrics, different models would be selected compared to relying solely on the false match rate (FMR).

The proposed evaluation framework addresses biometric security from the perspective of guessing difficulty. To achieve this, we tackle challenges arising from the non-uniformity and uncountability of biometric data, along with the fuzzy matching process. Our framework overcomes these challenges by offering two approaches for estimating guessing difficulty. Firstly, we present an algorithm to determine the effective key length of biometrics. Secondly, we adapt pragmatic guessing metrics commonly used in password security to the context of biometrics and establish a relationship between the outcomes of both approaches.

We apply our evaluation framework to assess face and free-text keystroke authentication systems. We demonstrate that face authentication is as secure as 4-digit PIN codes and also explore the extent to which partial attacker knowledge can reduce the security of free-text keystroke authentication.

		size	$\hat{\tilde{H}}_\infty$	$\hat{\lambda}_3$	$\hat{\lambda}_5$	$\hat{\lambda}_{10}$	$\hat{\tilde{G}}_{0.25}$	$\hat{\tilde{G}}_{0.5}$
Baseline	NA	10000	3.66	3.82	3.92	4.18	3.86	4.06
Age	<18	33457	3.56	3.72	3.84	4.16	3.77	4.01
	18-25	37790	3.39	3.69	3.87	4.23	3.75	4.1
	25-35	24453	3.22	3.65	3.86	4.22	3.71	4.07
	35-50	10410	3.11	3.5	3.7	4.07	3.49	3.82
	>50	3647	2.75	3.13	3.4	3.95	2.99	3.41
Sex	male	44864	3.46	3.67	3.87	4.22	3.75	4.08
	female	50788	3.46	3.68	3.86*	4.21*	3.74*	4.07*
Native language	en	92433	3.45	3.77	3.93	4.22	3.84	4.11
	es	2157	2.69	3.36	3.69	4.23	3.18	4.05
	tl	2649	2.95	3.43	3.66	4.07	3.4	3.78
Number of fingers	1-2	14050	2.35	3.02	3.43	4.01	2.71	3.44
	3-4	14149	2.82	3.11	3.41	3.96	2.98	3.42
	5-6	11770	3.21	3.53	3.73	4.13	3.55	3.91
	7-8	17487	3.73	3.89	4.04	4.33	3.97	4.27
	9-10	52473	2.92	3.4	3.64	4.11	3.36	3.8
Words per minute (speed)	<39.5	36767	2.3	2.74	3.06	3.82	2.52	2.81
	39.5-58.5	35931	2.98	3.21	3.47	3.98	3.16	3.54
	>58.5	37232	2.16	2.65	2.95	3.76	2.38	2.67

Table 6.6: Estimates of guessing difficulty for keystroke dynamics when the adversary has knowledge of soft-biometric characteristics. We only provide groups with more than 2.000 unique users, which excludes several *native languages*.

Chapter 7

Conclusion

In the present digital era, identity verification plays an indispensable role. However, despite its significance, authentication remains challenging as contemporary solutions strive to strike a balance between security, usability, and privacy. Biometric authentication promises a new optimum in this tension field by reducing user friction and introducing novel security benefits such as theft resilience, non-transferability, and continuous authentication. However, these advantages are not without their own set of challenges, including the sensitive nature of biometric data, the inherent fuzziness of matching, and the non-uniformity of the biometric distribution.

This dissertation aims to address those challenges through research comprising three pillars (depicted in Figure 7.1): *behavioral authentication*, *novel attacks*, and *security evaluation*. The concluding chapter of this thesis begins by summarizing the contributions of the research, followed by a discussion that delves into the broader implications of these contributions within the realm of biometric authentication. Within this discourse, we elucidate the primary limitations and chart a course for future research endeavors. During this discussion, we outline the primary limitations and pave the road for future work. Ultimately, the dissertation concludes with a retrospective analysis of six years of research, encompassing the entirety of this PhD trajectory.

7.1 Summary of the contributions

The central aim of this dissertation is to forge a path toward reliable biometric authentication. Biometric authentication entails methods that accurately recognize individuals based on their inherent characteristics while upholding their privacy.

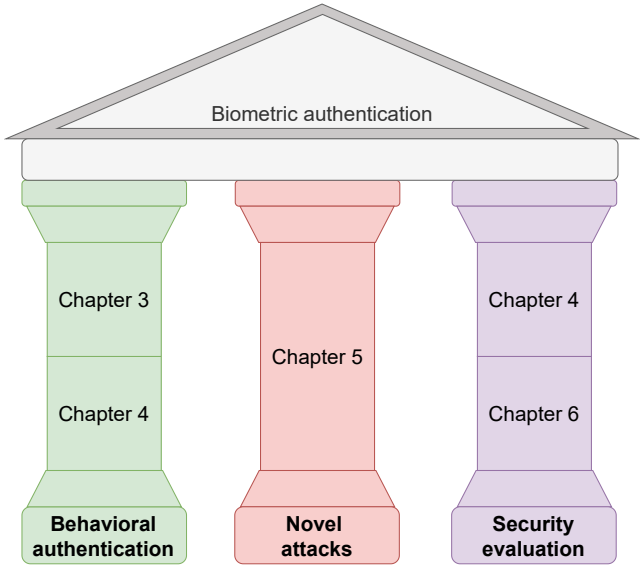


Figure 7.1: [Repeated picture Figure 1.1] The three pillars that reflect the research activities of this PhD. The pillars support the end goal of valid and reliable, secure, and private biometric authentication.

Furthermore, because authentication is a security-critical application, adversarial manipulation must be accounted for. Three foundational pillars were erected to achieve this overarching goal: advancing precise behavioral authentication, exploring novel attack methods, and establishing comprehensive security metrics.

This thesis dedicates considerable effort to researching behavioral authentication, which relies on behavioral traits rather than appearance. Behavioral authentication offers several advantages over traditional biometrics, as behavior can be collected unobtrusively while the user interacts with the system. Therefore, behavioral authentication can become transparent to the user, while appearance-based biometrics typically still require a small interaction, such as presenting a finger to the sensor. Furthermore, the continuous monitoring of behavior in the background allows for a paradigm shift from single-shot authentication to continuous authentication, an effective countermeasure to session hijacking.

Accurate behavioral authentication The drawback of behavioral authentication is distinguishing stable and relevant identifying characteristics from unwanted noise. Behavior is very context-dependent and thus tends to produce a more volatile signal. Furthermore, behavioral data is collected over time.

In Chapter 3, we explored using traditional generative machine learning to develop an IMU-based gait authentication system. We relied on Hidden Markov Models (HMM) and Universal Background Models (UBM) to learn the unique individual characteristics of human locomotion, even with limited enrollment data. Our approach reaches $\approx 1.5\%$ Equal Error Rate (EER) in the challenging open-world setting. A key element to successfully learning the distinguishing characteristics is our high-level representation of the IMU data. We dedicated substantial effort to engineering a feature set that maximizes the signal's information and enhances learning. The experience gained from our rigorous efforts in developing an optimized feature set proved to be instrumental in achieving victory in the "2019 OU-ISIR Wearable Sensor-based Gait Challenge: Age and Gender (GAG 2019) [1] — a testament to our breakthrough beyond the state-of-the-art. For more details on our approach and overarching conclusions of the competition, we refer the reader to our write-up [241] and the summary of the competition organizers [169].

In Chapter 4, we extend the generative approach with discriminative approaches based on deep learning. Specifically, we develop four different methods which combine one of two different network architectures and one of two different loss functions. The architectures are based on a temporal convolutional neural network (TCN) and a Recurrent Neural Network with attention, while triplet loss [213] and arcface [66] are used as loss functions. Our empirical results show that the discriminative and generative approaches achieve similar recognition performance, with EERs between $\approx 1-2\%$. However, we should be critical and wonder whether the found results generalize. On the one hand deep learning seems to especially outperform traditional ML at scale, thus more training data could boost performance. On the other hand, it makes sense that provided a given data representation and learning task there is maximum utility that can be achieved.

As mentioned earlier, biometric systems should respect the user's privacy. Therefore, sensitive biometric data must be protected with biometric template protection schemes that guarantee the biometric template's irreversibility, unlinkability, and renewability. These features ensure that the user's privacy is guaranteed in case of biometric template leakage and is protected against an honest-but-curious authentication service provider. Furthermore, leaked biometric templates can be revoked, reducing their security threat. Chapters 3 & 4 apply a fuzzy commitment scheme to the gait templates.

A downside of template protection is the loss of recognition performance. We counteract this performance loss by fusing the extracted features from different models. For this fusion, we consider models with the same learnable parameters but converged to a different optimum and models with different learnable parameters. The results in Chapter 3 demonstrate that relying solely on HMMs, an effective key length of 64 bits is achieved at an FRR of 5%. In Chapter 4, we research how far intra-modal fusion

[1] http://www.am.sanken.osaka-u.ac.jp/GAG2019/

can push the offered security. Our analysis demonstrates the existence of an optimal ensemble size that provides a maximum security working point. Adding more systems beyond this working point decreases security as the effect of the ensemble contribution is negated by the strong correlations between the additional bits. Furthermore, the optimal working point requires fusing many systems, while the increase in security compared to smaller-sized ensembles is modest. Therefore, the additional complexity of achieving the optimal working point is not justified from a security point of view.

Security evaluation The security evaluation that drives our analysis of the benefits of intramodal fusion for fuzzy commitment cryptosystems in Chapter 4 improves drastically upon our analysis presented in Chapter 3. While the evaluation methodology in Chapter 3 was based on prior work, it seems to encourage unreasonably long keys. Moreover, because the evaluation considers primarily a traditional hacker who tries to guess the secret key hidden in the protected biometric template, it ignores other threat actors that leverage the knowledge of the biometric distribution.

Chapter 4 presents an evaluation framework considering a biometric hacker who leverages the population-wise biometric template statistics. The evaluation framework accounts for a biometric hacker who performs a *guessing attack* by sampling biometric templates from the distribution. The effectiveness of this adversary is captured by estimating the FAR and converting it to an effective key length. A second attack strategy for this adversary is to attack the ECC message space directly. To capture the effectiveness of this adversary in a theoretical lower bound, we hypothesized an equivalent protected system in the biometric information space. The size of the biometric information space was gauged by relying on prior work by Daughman, which estimates it by determining the degrees of freedom from the binomial that best fits the score distribution of impostors.

Our evolution framework sheds new light on biometric authentication systems protected with fuzzy commitment schemes. The methodology is not specific to gait authentication systems or intramodal fusion. It generalizes to any system that requires accounting for correlations between the bits of the keys produced by the biometric authentication system.

Novel attacks Our work on gait authentication (cf. Chapters 3 & 4) considers a biometric adversary guessing the biometric input by sampling from the biometric distribution. Such an adversary has an expected success rate equal to the False Match Rate (FMR). However, biometric mismatches are systematic and non-uniform. An adversary can leverage these properties to gain an additional advantage by executing "masterkey" attacks. Much like how a physical masterkey fist multiple locks, a biometric masterkey matches several people.

Chapter 5 contributes a masterkey attack against free-text keystroke biometrics. The attack leverages a black-box optimization algorithm that optimizes keystroke timings for a desired attack sentence. Chapter 5 also contributes a methodological approach to evaluating the successfulness of masterkey attacks. The approach adapts the scientific foundation of guessing metrics in password security evaluations to biometrics. Our systematic evaluation allows us to accurately compare the security offered by knowledge-based and inherence-based authentication factors. Furthermore, it enables research on the effect of demographic factors. We perform such analysis for free-text keystroke biometrics and show that against the most potent adversary, free-text keystroke biometrics provide several magnitudes less security than a 4-digit PIN code. However, for less capable and more realistic adversaries, the guessing difficulty of free-text keystroke biometrics is on par with a 4-digit PIN code.

Security evaluation revisited Masterkey attacks provide a different perspective to guessing the difficulty of biometrics. As such, we revisit the use - and our own earlier use in Chapter 4 - of False Match Rate to assess the guessing difficulty of biometrics.

Chapter 6 extends our evaluation of masterkey attacks (cf. Chapter 5) and contributes a security evaluation framework for biometrics that replaces the False Match Rate. We first describe an upper bound on the security offered by biometric authentication systems by determining the minimal number of guesses required to exhaust the biometric template space. Determining the upper bound is a theoretical exercise that relies heavily on mathematical foundations from spherical code literature. This theoretical analysis complements and relates to the empirical analysis that relies on guessing metrics. By combining the empirical and theoretical approaches, our evaluation framework paints a complete picture of the resilience of biometric authentication to guessing. Such analysis finally provides the tools to contrast different authentication factors. Moreover, the framework allows us to measure the security impact of adversaries' partial knowledge of their victim's physiological and behavioral characteristics.

7.2 Charting future paths

Within the scope of this PhD thesis, we researched behavioral authentication as a promising technology that can improve the security and usability of the lock on the front door to digital systems. While in research, we go to great lengths to leave no stone unturned; inevitably, our efforts reach their limitations. Acknowledging those limitations paves the road ahead. To this end, we first acknowledge the limitations of our work and use them as an anchor point to suggest directions for future work.

Personal differences in biometric performance The security evaluation framework proposed in this dissertation accounts for the biometric distribution's non-uniformity while estimating the adversary's guessing difficulty. A different viewpoint analyzes the effect of non-uniformity from the user's perspective, who is concerned with their 'personal' security level.

Thus, while the biometric-brute force adversary considered in this work happily exploits non-uniformity to gain an additional advantage, we did not consider non-uniformity from the defender's perspective. Non-uniformity originates from people being more or less unique in light of the specific biometric. Therefore, studying the biometric template space to estimate a user's uniqueness also measures their resilience to manipulation, e.g., adversarial perturbations. An evaluation of the user's uniqueness can drive defensive actions such as a dynamic thresholding scheme. Such thresholding scheme accounts for interpersonal differences by being more restrictive when the users' biometric data is less unique and less restrictive when the users' biometric data is more unique. This way, a new optimum between security and usability can be achieved.

Measuring security offered by continuous authentication Behavioral authentication enables continuous authentication, where the user's identity is continuously assessed. Continuous authentication provides an alternative to single-shot authentication, where the user's identity is only verified before privileged actions. Continuous authentication secures sessions against session hijacking by requiring step-up authentication when the behavior has diverged. Our security evaluations primarily focussed on the single-shot authentication paradigm. As correctly pointed out by Bours and Mondal [37], the evaluation of continuous authentication should account for the number of actions before a session is locked. This number should then be compared to the number of actions an adversary requires to obtain their end goal.

Future work can pick up on the concept of masterkeys and look for behavior that allows the adversary to evade detection by the continuous authentication system indefinitely. Avoiding detection for a prolonged time can be achieved by cycling through different types of common behaviors or by relying on less informative behavioral actions, e.g., it has been shown that certain types of swipes reduced detection rates of swiping behavior recognition systems [217]. From the defensive perspective, security evaluations should not rely solely on averages, i.e., the average number of impostor actions before being locked out. Just like the evaluation for single-shot authentication scenarios, the security evaluation should account for the distribution, i.e., the distribution of the number of impostor actions before being locked out.

Securing the whole biometric pipeline While this dissertation delves deeply into the adversarial exploitation of the fuzzy matching scheme, it pays less attention to other plausible threats. In Chapter 2, we acknowledged several other threats, such as physical

adversarial examples and the leakage of biometric enrollment templates. Physically realizable adversarial examples are specially crafted perturbations that can be added to the scene, e.g., a user carrying glasses [220]. These glasses are crafted with the goal of changing the identity of the individual wearing them ("Hi there, Clark Kent"), or at least changing their identity according to the biometric recognition system. This attack strategy carries great potential as it is an entirely new attack not caught by current security measures such as liveness detection.

The main limitation of adversarial perturbations is that they are brittle. It is hard to ensure that they work for an unknown target network and to physically realize them such that they are robust to changes in lighting conditions, rotations, and small translations. Future research that aims to leverage physical adversarial examples in an attack vector must overcome these challenges. Furthermore, adversarial examples are primarily researched in the visual domain. Generalizing this concept to other types of data remains challenging [1].

From the defensive perspective, robustness against adversarial manipulation is a topic that enjoys considerable attention. However, most approaches proposed in academia, such as adversarial training and certified adversarial robustness, are expensive and do not provide security [117]. Moreover, robustness typically describes performance a bit outside the normal i.i.d. (independent and identically distributed) assumption rather than providing the worst-case performance in the presence of an adversary [274]. To shift the narrative from what is known as 'adversarial robustness' to actual security, we require offensive work to stop attacking ML models in isolation and start attacking ML-based systems where the whole pipeline is considered [16]. The approaches taken in offensive work will inspire the design of pragmatic defensive controls such as detection and response, e.g., stateful defenses against black-box attacks [52], providing the ML pipeline the ability to abstain [69], and creating conditions which make executing adversarial attacks challenging. We see biometric authentication systems as an intriguing playground for ML security, where the research community can explore this arms race between attacker and defender.

We have already proven that the leakages of biometric templates pose security concerns (cf. Chapter 4). However, industry practitioners do not recognize this threat. For example, Apple claims that *Touch ID does not store any images of your fingerprint and instead relies only on a mathematical representation. It is not possible for someone to reverse engineer your actual fingerprint image from this stored data.* [15]. A similar argument is made by Microsoft on Windows Hello: *Additionally, even if an attacker was actually able to get the biometric data from a device, it cannot be converted back into a raw biometric sample that could be recognized by the biometric sensor.* [156] Novel research falsifies such claims [142, 150] and prompts lawmakers to recognize the security and privacy risks of mass storage of biometric templates. Therefore, the research community must keep investing in demonstrating novel template inversion attacks and defining defensive controls such as biometric template protection.

Multidisciplinary nature of biometrics Throughout this dissertation, we have mainly focussed on the technological side of biometrics. However, recognizing people through biometrics poses multidisciplinary challenges. For example, in this dissertation, we did not conduct user studies to demonstrate the usability benefits of biometrics. More importantly, we have yet to show that society accepts the use of behavior for authentication purposes. From the legal perspective, biometric data is sensitive and, when collected on a large scale, poses societal risks of mass surveillance. A great example is clearview.ai, which offers a search engine for faces by law enforcement. A similar debate is held within many governmental institutions which extend their biometric recognition capabilities by collecting more biometric data on ID cards and equipping cities and highways with smart cameras. Technological advances in generative AI offer a completely different perspective on biometrics. Even today, deepfakes undermine the trust in the human perception of biometrics by creating new attack opportunities for social engineering and magnifying the concerns around misinformation.

In general, the technological advances driven by AI move at an unprecedented pace. The industry has adopted a *fail fast, fail forward* approach while they fight for dominance in the AI war. Their carefree approach is beautifully illustrated by the introduction of AI-driven chatbots which are not well-rounded products but large-scale experiments (and flashes of marketing geniality). The speed of technological advances is in sharp contrast with the relatively slow-moving legal landscape. A strong collaboration between legal and technical research is required to bridge this gap and avoid ineffective legal frameworks. Research partnerships between technological and legal experts ensure actionable legislation that leverages the correct terminology and relies on realistic technological controls.

Going outside the lab Accurate biometric recognition necessitates the development of a function capable of extracting distinctive characteristics from biometric data. This involves the collection of raw training data, the adoption of a learning approach that models the desired function, and the preprocessing of data into a format optimizing learning efficacy.

Within the confines of this thesis, significant effort was dedicated to researching an effective learning pipeline for gait recognition. It's important to note that while our focus was on this aspect, we did not directly collect data. Instead, similar to previous studies, we relied on well-established benchmark datasets known for their satisfactory data quality. The evaluation of quality was based on comprehensive documentation, methodological rigor in data collection, and representation of a sufficiently large and diverse population.

However, reliance on benchmark datasets introduces abstractions that may overlook real-world constraints, such as variations in sensing hardware, computational resource

limitations, and energy consumption restrictions. Additionally, these datasets, collected under controlled lab conditions, lack crucial contextual information, including the user's emotional and physical state, changes in behavior over time, and sensor displacement.

To truly assess the impact of context shifts, data collection should extend beyond lab conditions and align with the operational context of the biometric authentication system. Understanding the effects of challenging operational conditions is crucial for building system resilience. We anticipate enhancing system trustworthiness by diversifying training data, implementing template updating, and sensing environmental changes to uphold data quality. Notably, existing practices in domains like face [111] and fingerprint [110] specify data quality measures, allowing the authentication system to operate only when the data meets minimal quality standards. Furthermore, operational conditions are carefully managed to maximize data quality.

In conclusion, translating theoretical research into a functioning system requires navigating a myriad of additional constraints often abstracted in fundamental research. This navigation demands a pragmatic mindset, combining theoretical approaches from works like the research conducted in this PhD thesis with a practical understanding of real-world challenges."

From mono-modal to multi-modal and multifactor tackling overarching limitations of (biometric) authentication solutions. With any authentication system, the threat of getting locked out of your assets is daunting. For example, access tokens get lost, or you forget your password. Nevertheless, these threats feel more controllable than the randomness of accidental mismatches in biometrics or the misfortune of physical accidents that temporarily or permanently change appearance and behavior.

The distress of losing control over our accounts is mitigated by implementing account recovery mechanisms. The main problem with account recovery is that it should be at least as secure as the primary authentication mechanism, albeit more user friction is acceptable. For example, biometrics for phone unlock are complemented with PIN codes (it is typically required to enter the PIN code once every 48 hours, most likely to ensure they are remembered).

An authentication system that optimizes the tension between usability, security, and privacy achieves a composition of authentication factors that maximally complement each other. The search for original blends of authentication methodologies provides opportunities for novel research and business. For example, an original and appreciated research result studies the unique way in which users pick up their phones as a solution to puppet attacks where a dormant victim's finger is put on the fingerprint sensor by the adversary [257]. Alternatively, voice authentication is being integrated into banks' call centers responsible for account recovery. Voice recognition gives the call centers

an additional anchor point to verify the person's identity on the other line. Within the confines of this dissertation, we primarily researched mono-modal systems. When going multi-modal and multifactor new challenges emerge, such as liveness detection solutions that influence each other and certain factors being unavailable.

7.3 Concluding remarks

Throughout six years of research encompassing this PhD we researched the trustworthiness of machine learning through the lens of biometric authentication. Machine learning systems can sometimes fail to maintain the trust we place in them, making it crucial to comprehend these failure modes in order to implement preventive and reactive measures. One primary area of concern lies in the reality that machine learning models can never achieve absolute accuracy. Consequently, it becomes imperative to establish robust model validation processes that ascertain the probability of erroneous outputs. However, merely expressing the likelihood of mistakes does not provide a comprehensive understanding of their impact on the overall system's utility. To properly assess the validity and reliability of a system, it is essential to evaluate the consequences of unavoidable errors. To address this issue, this PhD research has proposed methods to evaluate the impact of accidental mistakes made by machine learning algorithms on the security of biometric systems.

Another significant failure mode is the susceptibility of machine learning to adversarial manipulation of input data. While machine learning performs well under stable and predefined data distributions, it becomes fragile when faced with data outside these bounds. Adversaries exploit this vulnerability by manipulating input to the system, gaining an advantage in the process. The methodologies presented in this work take into account the potential for adversarial interference, which is particularly vital for authentication systems operating in adversarial contexts.

A critical factor underpinning all our research efforts in this PhD is the availability of high-quality data. Without a reliable dataset, training and evaluating biometric authentication would have been infeasible. This necessitates significant computational resources and a substantial amount of data. However, the reluctance of the industry to share data poses a hindrance to further research progress. Moreover, the current data landscape exhibits a substantial inequality, with major industry players and tech companies amassing the majority of data. Nevertheless, positive strides are being made to address this issue through initiatives such as Hugging Face, which aim to democratize access to high-quality datasets and models. And still, especially in the realm of security, collaborative efforts between academia and industry must be strengthened to ensure the secure and trustworthy adoption of machine learning technologies.

Appendix A

Additional materials for masterkey attacks against keystroke dynamics

A.1 Mathematical symbols

We provide an overview of the mathematical notations used in Table A.1.

A.2 Results of statistical analysis

We report the p-values of the ANOVA and Alexander-Govern tests in Table A.2. The ANOVA test examines the null hypothesis that the sample means from the subgroups are the same. We perform these tests for a specific data split based on age, gender, sex, typing speed, and the number of fingers used. We redid the ANOVA test for the age category excluding two age groups, i.e., $35 - 50$ and < 50. In the latter case, we could not reject the null hypothesis with high confidence for most of the metrics.

Table A.1: An overview of the mathematical symbols used and the section where they were introduced.

Symbol	Definition	Section
B	biometric signal space, e.g., the keystroke timings of sentences typed by users	5.2.1
b	a specific biometric signal, e.g., a specific sequence of keystroke timings of a single user typing	5.2.1
b_j^i	the i^{th} biometric signal of user j	5.2.1
$f(\cdot)$	the feature extractor which projects the biometric signal onto a biometric feature vector (biometric template)	5.2.1
x_j^i	the feature vector (biometric template) with norm 1, e.g., 64 real numbers	5.2.1
c_j	the biometric centroid of user j, this is stored in the biometric database	5.2.1
E_j	the number of enrollment samples for user j	5.2.1
$m(\cdot,\cdot,\cdot)$	the matcher function, it provides an authentication decision based on the biometric query sample, the centroid of the claimed identity and the threshold	5.2.1
$d(\cdot,\cdot)$	the distance function which takes as arguments two feature vectors, i.e., the angular distance $1 - cos(\theta)$ with θ the angle between the two feature vectors	5.2.1
t	the threshold, i.e., a real number between [0, 2]	5.2.1
U	the number of users in the dataset	5.3.2
N_j	the number of biometric signals in the dataset for user j	5.3.2
mk	a masterkey, depending on the adversary type mk is a biometric signal or a biometric template	5.3.2
p_n	the attack success rate of the n^{th} guess	5.4.1
H_∞	min-entropy	5.4.2
λ_β	β-success-rate, the expected success rate for an adversary limited to β guesses per account	5.4.2
μ_α	α-work-factor, the number of guesses required per account to break a fraction α accounts, does not take into account early stopping	5.4.2
G_α	α-guesswork, the expected number of guesses to break a fraction α accounts, takes into account early stopping	5.4.2
$\tilde{\lambda}_\beta$	β-success-rate expressed as an effective key length	5.4.2
$\tilde{\mu}_\alpha$	α-work-factor expressed as an effective key length	5.4.2
$\tilde{G}_\alpha$	α-guesswork expressed as an effective key length	5.4.2

Table A.2: p-values for the ANOVA tests. In the case of the native language, the p-values are obtained with an Alexander-Govern test.

		$\hat{\hat{H}}_\infty$	$\hat{\hat{\lambda}}_3$	$\hat{\hat{\lambda}}_5$	$\hat{\hat{\lambda}}_{10}$	$\hat{\hat{G}}_{0.25}$	$\hat{\hat{G}}_{0.5}$
Age	<18 18-25 25-35 35-50 >50	3.3694e-15	3.0348e-21	4.5953e-21	8.3900e-16	1.9489e-20	1.3815e-21
Age	<18 18-25 25-35	1.2932e-04	1.0744e-01	6.6417e-01	1.9093e-02	3.4794e-01	9.8144e-02
Sex	male female	8.8907e-01	9.3115e-01	7.7717e-01	5.2012e-01	7.9905e-01	7.7068e-01
Native language	en es hi tl zh	7.7542e-09	4.9536e-13	5.2485e-11	4.1183e-09	1.8021e-15	3.4445e-11
Number of fingers	1-2 3-4 5-6 7-8 9-10	5.9072e-23	7.7611e-23	7.3345e-22	5.5800e-18	6.8246e-24	1.6557e-19
Words per minute (speed)	<39.5 39.5-58.5 >58.5	2.0164e-14	1.6491e-17	5.6501e-20	2.8917e-10	2.5035e-14	5.9870e-17

Bibliography

[1] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor. "SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems". In: *2021 IEEE Symposium on Security and Privacy (SP)*. 2021, pp. 730–747. DOI: 10.1109/SP40001.2021.00014.

[2] *About Face ID advanced technology*. https://support.apple.com/en-us/HT208108. Apr. 2022.

[3] *About Touch ID advanced security technology*. https://support.apple.com/en-us/HT204587. Sept. 2017.

[4] A. Acien, A. Morales, R. Vera-Rodriguez, J. Fierrez, and J. V. Monaco. "Typenet: Scaling up keystroke biometrics". In: *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE. 2020, pp. 1–7.

[5] A. Adler. "Images can be regenerated from quantized biometric match score data". In: *Canadian Conference on Electrical and Computer Engineering 2004 (IEEE Cat. No. 04CH37513)*. Vol. 1. IEEE. 2004, pp. 469–472.

[6] T. Ahmad and Fengling Han. "Cartesian and polar transformation-based cancelable fingerprint template". In: *IECON 2011 - 37th Annual Conference of the IEEE Industrial Electronics Society*. 2011, pp. 373–378. DOI: 10.1109/IECON.2011.6119339.

[7] N. Akhtar, A. Mian, N. Kardan, and M. Shah. "Advances in adversarial attacks and defenses in computer vision: A survey". In: *IEEE Access* 9, 2021, pp. 155161–155196.

[8] *Alliance overview*. https://fidoalliance.org/overview/. Accessed: 2023-04-06.

[9] A. Alsultan and K. Warwick. "Keystroke dynamics authentication: a survey of free-text methods". In: *International Journal of Computer Science Issues (IJCSI)* 10.4, 2013, p. 1.

[10] F. Álvarez, L. Hernandez Encinas, and C. Sánchez Ávila. "Biometric Fuzzy Extractor Scheme for Iris Templates". In: *Proceedings of WORLDCOMP'09, The 2009 World Congress in Computer Science, Computer Engineering, and Applied Computing*. July 2009, pp. 563–569.

[11] A. Alzubaidi and J. Kalita. "Authentication of Smartphone Users Using Behavioral Biometrics". In: *IEEE Communications Surveys Tutorials* 18.3, thirdquarter 2016, pp. 1998–2026. ISSN: 1553-877X. DOI: 10.1109/COMST.2016.2537748.

[12] R. Anderson. "Chapter 17: Biometrics". In: *Security engineering: a guide to building dependable distributed systems*. John Wiley & Sons, 2020, pp. 571–598.

[13] A. Anjos, M. Günther, T. de Freitas Pereira, P. Korshunov, A. Mohammadi, and S. Marcel. "Continuously Reproducing Toolchains in Pattern Recognition and Machine Learning Experiments". In: *International Conference on Machine Learning (ICML)*. Aug. 2017.

[14] M. Antal and L. Z. Szabó. "An Evaluation of One-Class and Two-Class Classification Algorithms for Keystroke Dynamics Authentication on Mobile Devices". In: *2015 20th International Conference on Control Systems and Computer Science*. 2015, pp. 343–350. DOI: 10.1109/CSCS.2015.16.

[15] Apple. *About Touch ID advanced security technology*. https://support.apple.com/en-us/HT204587. Last accessed on 24 October 2022. Sept. 2017.

[16] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. A. Roundy. "Position:"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice". In: *First IEEE Conference on Secure and Trustworthy Machine Learning*. 2022.

[17] A. Arakala, J. Jeffers, and K. J. Horadam. "Fuzzy Extractors for Minutiae-Based Fingerprint Authentication". In: *Advances in Biometrics*. Ed. by S.-W. Lee and S. Z. Li. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 760–769. ISBN: 978-3-540-74549-5.

[18] E. Argones Rua and J. L. Alba Castro. "Online Signature Verification Based on Generative Models". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4, 2012, pp. 1231–1242. DOI: 10.1109/TSMCB.2012.2188508.

[19] E. Argones Rúa, D. Pérez-Piñar López, and J. Alba Castro. "Ergodic HMM-UBM System for On-Line Signature Verification". In: *Biometric ID Management and Multimodal Communication*. Springer Berlin Heidelberg, 2009, pp. 340–347. ISBN: 978-3-642-04391-8.

[20] E. Argones Rúa, A. Abidin, R. Peeters, and J. Romme. "Efficient and Privacy-Preserving Cryptographic Key Derivation From Continuous Sources". In: *IEEE Transactions on Information Forensics and Security* 14.11, 2019, pp. 2834–2847. DOI: 10.1109/TIFS.2019.2907459.

[21] E. Argones Rúa, E. Maiorana, J. L. Alba Castro, and P. Campisi. "Biometric Template Protection Using Universal Background Models: An Application to Online Signature". In: *IEEE Transactions on Information Forensics and Security* 7.1, 2012, pp. 269–282. DOI: 10.1109/TIFS.2011.2168213.

[22] E. Argones Rúa, T. Van hamme, D. Preuveneers, and W. Joosen. "Discriminative Training of Spiking Neural Networks Organised in Columns for Stream-based Biometric Authentication". In: *IET Biometrics* 11.5, 2022, pp. 485–497.

[23] E. Argones Rúa, T. Van hamme, D. Preuveneers, and W. Joosen. "Gait Authentication based on Spiking Neural Networks". In: *20th International Conference of the Biometrics Special Interest Group*. Vol. LNI P-315. Gesellschaft für Informatik e.V., Oct. 2021, pp. 51–60. ISBN: 978-3-88579-709-8.

[24] D. Bahdanau, K. Cho, and Y. Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015.

[25] S. Bai, J. Z. Kolter, and V. Koltun. *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling*. 2018. arXiv: 1803.01271 [cs.LG].

[26] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. "Evasion attacks against machine learning at test time". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*. Springer. 2013, pp. 387–402.

[27] B. Biggio, L. Didaci, G. Fumera, and F. Roli. "Poisoning attacks to compromise face templates". In: *2013 International Conference on Biometrics (ICB)*. IEEE. 2013, pp. 1–7.

[28] B. Biggio, G. Fumera, F. Roli, and L. Didaci. "Poisoning Adaptive Biometric Systems". In: *Structural, Syntactic, and Statistical Pattern Recognition*. Ed. by G. Gimel'farb, E. Hancock, A. Imiya, A. Kuijper, M. Kudo, S. Omachi, T. Windeatt, and K. Yamada. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 417–425. ISBN: 978-3-642-34166-3.

[29] B. Biggio, G. Fumera, F. Roli, and L. Didaci. "Poisoning adaptive biometric systems". In: *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer. 2012, pp. 417–425.

[30] B. Biggio, P. Russu, L. Didaci, F. Roli, et al. “Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective”. In: *IEEE Signal Processing Magazine* 32.5, 2015, pp. 31–41.

[31] J. Bilmes. *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*. Tech. rep. International Computer Science Institute, Berkeley, Department of Electrical Engineering, and Computer Science, U.C. Berkeley, 1998.

[32] J. Blasco, T. M. Chen, J. Tapiador, and P. Peris-Lopez. “A Survey of Wearable Biometric Recognition Systems”. In: *ACM Comput. Surv.* 49.3, Sept. 2016. ISSN: 0360-0300. DOI: 10.1145/2968215. URL: https://doi.org/10.1145/2968215.

[33] J. Bonneau. “The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords”. In: *2012 IEEE Symposium on Security and Privacy*. 2012, pp. 538–552. DOI: 10.1109/SP.2012.49.

[34] J. Bonneau, C. Herley, P. C. v. Oorschot, and F. Stajano. “The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes”. In: *2012 IEEE Symposium on Security and Privacy*. 2012, pp. 553–567. DOI: 10.1109/SP.2012.44.

[35] J. Bonneau, S. Preibusch, and R. Anderson. “A birthday present every eleven wallets? the security of customer-chosen banking pins”. In: *International Conference on Financial Cryptography and Data Security*. Springer. 2012, pp. 25–40.

[36] P. Bontrager, A. Roy, J. Togelius, N. Memon, and A. Ross. “DeepMasterPrints: Generating MasterPrints for Dictionary Attacks via Latent Variable Evolution”. In: *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. 2018, pp. 1–9. DOI: 10.1109/BTAS.2018.8698539.

[37] P. Bours and S. Mondal. “Performance evaluation of continuous authentication systems”. In: *Iet Biometrics* 4.4, 2015, pp. 220–226.

[38] S. Boztas. “Entropies, guessing, and cryptography”. In: *Department of Mathematics, Royal Melbourne Institute of Technology, Tech. Rep* 6, 1999, pp. 2–3.

[39] J. Breebaart, B. Yang, I. Buhan-Dulman, and C. Busch. “Biometric template protection”. In: *Datenschutz und Datensicherheit - DuD* 33.5, May 2009, pp. 299–304. ISSN: 1862-2607. DOI: 10.1007/s11623-009-0089-0. URL: https://doi.org/10.1007/s11623-009-0089-0.

[40] J. Buolamwini and T. Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by S. A. Friedler and C. Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 23–24 Feb 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[41] A. Buriro, B. Crispo, and M. Conti. “AnswerAuth: A bimodal behavioral biometric-based user authentication scheme for smartphones”. In: *Journal of Information Security and Applications* 44, 2019, pp. 89–103. ISSN: 2214-2126. DOI: https://doi.org/10.1016/j.jisa.2018.11.008. URL: https://www.sciencedirect.com/science/article/pii/S2214212618304435.

[42] A. Buriro, B. Crispo, F. Del Frari, and K. Wrona. “Touchstroke: Smartphone User Authentication Based on Touch-Typing Biometrics”. In: *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*. Ed. by V. Murino, E. Puppo, D. Sona, M. Cristani, and C. Sansone. Cham: Springer International Publishing, 2015, pp. 27–34. ISBN: 978-3-319-23222-5.

[43] P. Campisi. *Security and privacy in biometrics*. Vol. 24. Springer, 2013.

[44] K. Cao and A. K. Jain. “Hacking mobile phones using 2D printed fingerprints”. In: *Department of Computer Science and Engineering, Michigan State University*, 2016.

[45] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. “Vggface2: A dataset for recognising faces across pose and age”. In: *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE. 2018, pp. 67–74.

[46] X. Cao, Y. Wei, F. Wen, and J. Sun. “Face alignment by explicit shape regression”. In: *International Journal of Computer Vision* 107.2, 2014, pp. 177–190.

[47] Y. Cao, L. Zhang, S. S. Zalivaka, C. Chang, and S. Chen. “CMOS Image Sensor Based Physical Unclonable Function for Coherent Sensor-Level Authentication”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 62.11, 2015, pp. 2629–2640.

[48] M. Casey, M. Manulis, C. J. P. Newton, R. Savage, and H. Treharne. “An Interoperable Architecture for Usable Password-Less Authentication”. In: *Emerging Technologies for Authorization and Authentication*. Ed. by A. Saracino and P. Mori. Cham: Springer International Publishing, 2020, pp. 16–32. ISBN: 978-3-030-64455-0.

[49] D. Chang, S. Garg, M. Hasan, and S. Mishra. "Cancelable Multi-Biometric Approach Using Fuzzy Extractor and Novel Bit-Wise Encryption". In: *IEEE Transactions on Information Forensics and Security* 15, 2020, pp. 3152–3167. DOI: 10.1109/TIFS.2020.2983250.

[50] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu. "Who is real bob? adversarial attacks on speaker recognition systems". In: *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2021, pp. 694–711.

[51] S. Chen, A. Pande, and P. Mohapatra. "Sensor-assisted facial recognition: an enhanced biometric authentication system for smartphones". In: *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. 2014, pp. 109–122.

[52] S. Chen, N. Carlini, and D. Wagner. "Stateful detection of black-box adversarial attacks". In: *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*. 2020, pp. 30–39.

[53] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. "Targeted backdoor attacks on deep learning systems using data poisoning". In: *arXiv preprint arXiv:1712.05526*, 2017.

[54] E. Chou, F. Tramèr, and G. Pellegrino. "SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems". In: *2020 IEEE Security and Privacy Workshops (SPW)*. 2020, pp. 48–54. DOI: 10.1109/SPW50608.2020.00025.

[55] A. E. Cinà, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, B. Biggio, M. Pelillo, and F. Roli. "Wild patterns reloaded: A survey of machine learning security against training data poisoning". In: *ACM Computing Surveys*, 2022.

[56] J. Colnago, S. Devlin, M. Oates, C. Swoopes, L. Bauer, L. Cranor, and N. Christin. ""It's not actually that horrible" Exploring Adoption of Two-Factor Authentication at a University". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–11.

[57] Common Criteria BIO-iTC. *Biometrics Security iTC (BIO-iTC)*. https://biometricitc.github.io/. Nov. 2021.

[58] D. Cozzolino, F. Marra, D. Gragnaniello, G. Poggi, and L. Verdoliva. "Combining PRNU and noiseprint for robust and efficient device source identification". In: *Eurasip Journal on Information Security* 2020.1, 2020, pp. 1–12. ISSN: 2510523X. DOI: 10.1186/s13635-020-0101-7. arXiv: 2001.06440.

[59] D. Cozzolino and L. Verdoliva. "Noiseprint: a CNN-based camera model fingerprint". In: *IEEE Transactions on Information Forensics and Security* 15, 2019, pp. 144–159.

[60] A. Dantcheva, P. Elia, and A. Ross. “What Else Does Your Biometric Data Reveal ? A Survey on Soft Biometrics”. In: 11.3, 2016, pp. 441–467.

[61] L. N. Darlow and B. Rosman. “Fingerprint minutiae extraction using deep learning”. In: *IEEE International Joint Conference on Biometrics, IJCB 2017* 2018-January, 2018, pp. 22–30. DOI: 10.1109/BTAS.2017.8272678.

[62] J. Daugman. “Statistical richness of visual phase information: update on recognizing persons by iris patterns”. In: *International Journal of computer vision* 45, 2001, pp. 25–38.

[63] J. Daugman. “The importance of being random: statistical principles of iris recognition”. In: *Pattern Recognition* 36.2, 2003. Biometrics, pp. 279–291. ISSN: 0031-3203. DOI: https://doi.org/10.1016/S0031-3203(02)00030-4.

[64] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. “Front-End Factor Analysis for Speaker Verification”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 4, 2011, pp. 788–798.

[65] R. Delgado-Escaño, F. M. Castro, J. R. Cózar, M. J. Marín-Jiménez, and N. Guil. “An End-to-End Multi-Task and Fusion CNN for Inertial-Based Gait Recognition”. In: *IEEE Access* 7, 2019, pp. 1897–1908. DOI: 10.1109/ACCESS.2018.2886899.

[66] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. “Arcface: Additive angular margin loss for deep face recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4690–4699.

[67] M. O. Derawi, P. Bours, and K. Holien. “Improved Cycle Detection for Accelerometer Based Gait Authentication”. In: *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. Oct. 2010, pp. 312–317. DOI: 10.1109/IIHMSP.2010.84.

[68] I. Deutschmann, P. Nordström, and L. Nilsson. “Continuous authentication using behavioral biometrics”. In: *IT Professional* 15.4, 2013, pp. 12–15.

[69] L. Devos, W. Meert, and J. Davis. *Adversarial Example Detection in Deployed Tree Ensembles*. 2022. arXiv: 2206.13083 [cs.LG].

[70] V. Dhakal, A. Feit, P. O. Kristensson, and A. Oulasvirta. “Observations on Typing from 136 Million Keystrokes”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18)*. ACM, 2018. DOI: https://doi.org/10.1145/3173574.3174220.

[71] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. *Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation*. Tech. rep. National Inst of Standards and Technology Gaithersburg Md, 1998.

[72] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith. “Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data”. In: *SIAM Journal on Computing* 38.1, 2008, pp. 97–139. DOI: 10.1137/060651380. eprint: https://doi.org/10.1137/060651380.

[73] S. Eberz, G. Lovisotto, A. Patanè, M. Kwiatkowska, V. Lenders, and I. Martinovic. “When Your Fitness Tracker Betrays You: Quantifying the Predictability of Biometric Features Across Contexts”. In: *2018 IEEE Symposium on Security and Privacy (SP)*. 2018, pp. 889–905.

[74] S. Eberz, G. Lovisotto, K. B. Rasmussen, V. Lenders, and I. Martinovic. “28 Blinks Later: Tackling Practical Challenges of Eye Movement Biometrics”. In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. CCS ’19. London, United Kingdom: Association for Computing Machinery, 2019, pp. 1187–1199. ISBN: 9781450367479. DOI: 10.1145/3319535.3354233. URL: https://doi.org/10.1145/3319535.3354233.

[75] S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic. “Evaluating behavioral biometrics for continuous authentication: Challenges and metrics”. In: *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 2017, pp. 386–399.

[76] U. Feige. “A Threshold of Ln n for Approximating Set Cover”. In: *J. ACM* 45.4, July 1998, pp. 634–652. ISSN: 0004-5411. DOI: 10.1145/285055.285059.

[77] P. Fernandez-Lopez, J. Liu-Jimenez, K. Kiyokawa, Y. Wu, and R. Sanchez-Reillo. “Recurrent Neural Network for Inertial Gait User Recognition in Smartphones”. In: *Sensors* 19.18, 2019. ISSN: 1424-8220. DOI: 10.3390/s19184054.

[78] *FIDO Biometrics Requirements*. https://fidoalliance.org/specs/biometric/requirements/Biometrics-Requirements-v2.0-fd-20201006.html#BioTerms. Oct. 2020.

[79] *FIDO Consumer Research Report*. https://media.fidoalliance.org/wp-content/uploads/2020/05/FIDO-Consumer-Research-Report.pdf. Accessed: 2023-04-06.

[80] *FIDO2: Web Authentication (WebAuthn)*. https://fidoalliance.org/fido2/fido2-web-authentication-webauthn/. Accessed: 2021-07-01.

[81] J. Fierrez, J. Ortega-Garcia, D. Ramos, and J. Gonzalez-Rodriguez. “HMM-based on-line signature verification: Feature extraction and signature modeling”. In: *Pattern Recognition Letters* 28.16, 2007, pp. 2325–2334. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2007.07.012. URL:

http://www.sciencedirect.com/science/article/pii/S0167865507002395.

[82] G. E. Forsen, M. R. Nelson, and R. J. Staron Jr. *Personal Attributes Authentication Techniques.* Tech. rep. Pattern Analysis and Recognition Corp Rome Ny, 1977.

[83] M. Fredrikson, S. Jha, and T. Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures". In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 2015, pp. 1322–1333.

[84] L. Fridman, A. Stolerman, S. Acharya, P. Brennan, P. Juola, R. Greenstadt, and M. Kam. "Multi-modal decision fusion for continuous authentication". In: *Computers & Electrical Engineering* 41, 2015, pp. 142–156.

[85] B. Fuller, X. Meng, and L. Reyzin. "Computational Fuzzy Extractors". In: *Advances in Cryptology - ASIACRYPT 2013*. Ed. by K. Sako and P. Sarkar. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 174–193. ISBN: 978-3-642-42033-7. DOI: 10.1007/978-3-642-42033-7_10.

[86] M. Gadaleta and M. Rossi. "IDNet: Smartphone-based gait recognition with convolutional neural networks". In: *Pattern Recognition* 74, 2018, pp. 25–37. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2017.09.005. URL: http://www.sciencedirect.com/science/article/pii/S0031320317303485.

[87] D. Gafurov, E. Snekkenes, and P. Bours. "Spoof Attacks on Gait Authentication System". In: *IEEE TIFS* 2.3, Sept. 2007, pp. 491–502. ISSN: 1556-6013. DOI: 10.1109/TIFS.2007.902030.

[88] D. Gafurov, E. Snekkenes, and P. Bours. "Improved Gait Recognition Performance Using Cycle Matching". In: *24th IEEE International Conference on Advanced Information Networking and Applications Workshops, WAINA 2010*. Jan. 2010, pp. 836–841.

[89] G. Garofalo, V. Rimmer, T. Van hamme, D. Preuveneers, W. Joosen, et al. "Fishy Faces: Crafting Adversarial Images to Poison Face Authentication". In: *12th USENIX Workshop on Offensive Technologies (WOOT 18)*. 2018.

[90] G. Gilchrist and S. D. Viavant. *Trusted biometric client authentication*. US Patent 6,167,517. Dec. 2000.

[91] GlobalPlatform. "TEE Biometric System PP-Module Version 1.0". In: GlobalPlatform, 2020, pp. 28–62. URL: https://globalplatform.org/specs-library/tee-biometric-system-pp-module-v1-0-gpd%5C_spe%5C_091/.

[92] M. Golla, G. Ho, M. Lohmus, M. Pulluri, and E. M. Redmiles. "Driving 2FA adoption at scale: Optimizing two-factor authentication notification design patterns". In: *30th USENIX Security Symposium*. USENIX. 2021, pp. 109–126.

[93] M. Gomez-Barrero and J. Galbally. “Reversing the irreversible: A survey on inverse biometrics”. In: *Computers & Security* 90, 2020, p. 101700.

[94] P. Gother, M. Ngan, and K. Hanaoka. *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*. Tech. rep. National Institute of Standards and Technology Interagency or Internal Report 8280. Washington, D.C.: U.S. Department of Commerce, 2019. DOI: https://doi.org/10.6028/NIST.IR.8280.

[95] P. A. Grassi, J. L. Fenton, E. M. Newton, R. A. Perlner, A. R. Regenscheid, and W. E. B. andJustin P. Richer. *SP 800-63B–Digital Identity Guidelines–Authentication and Lifecycle Management*. Tech. rep. 2017.

[96] T. Gu, B. Dolan-Gavitt, and S. Garg. “Badnets: Identifying vulnerabilities in the machine learning model supply chain”. In: *arXiv preprint arXiv:1708.06733*, 2017.

[97] S. Gupta, A. Buriro, and B. Crispo. “DriverAuth: A risk-based multi-modal biometric-based driver authentication scheme for ride-sharing platforms”. In: *Computers & Security* 83, 2019, pp. 122–139. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2019.01.007. URL: https://www.sciencedirect.com/science/article/pii/S0167404818310113.

[98] S. Habib, H. Khan, A. Hamilton-Wright, and U. Hengartner. “Revisiting the Security of Biometric Authentication Systems Against Statistical Attacks”. In: *ACM Transactions on Privacy and Security* 26.2, 2023, pp. 1–30.

[99] L. G. Hafemann, R. Sabourin, and L. S. Oliveira. “Characterizing and Evaluating Adversarial Examples for Offline Handwritten Signature Verification”. In: *IEEE Transactions on Information Forensics and Security* 14.8, 2019, pp. 2153–2166.

[100] J. Hämmerle-Uhl, E. Pschernig, and A. Uhl. “Cancelable Iris-Templates using Key-dependent Wavelet Transforms”. English. In: *Proceedings of the 6th IAPR International Conference on Biometrics (ICB’13)*. 2013.

[101] N. Hansen, Y. Akimoto, and P. Baudis. *CMA-ES/pycma on Github. Zenodo*. Feb. 2019. DOI: https://doi.org/10.5281/zenodo.2559634.

[102] N. Hansen and A. Auger. “Principled Design of Continuous Stochastic Search: From Theory to Practice”. In: *Theory and Principled Methods for the Design of Metaheuristics*. Ed. by Y. Borenstein and A. Moraglio. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 145–180. ISBN: 978-3-642-33206-7. DOI: 10.1007/978-3-642-33206-7_8. URL: https://doi.org/10.1007/978-3-642-33206-7_8.

[103] T. Hoang, D. Choi, and T. Nguyen. "Gait authentication on mobile phone using biometric cryptosystem and fuzzy commitment scheme". In: *International Journal of Information Security* 14.6, Nov. 2015, pp. 549–560. ISSN: 1615-5270. DOI: 10.1007/s10207-015-0273-1. URL: https://doi.org/10.1007/s10207-015-0273-1.

[104] D. S. Hochbaum. "Approximating covering and packing problems: set cover, vertex cover, independent set, and related problems". In: *Approximation algorithms for NP-hard problems*. 1996, pp. 94–143.

[105] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales. "When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition". In: *Proceedings of the IEEE international conference on computer vision workshops*. 2015, pp. 142–150.

[106] Hua-Hong Zhu, Qian-Hua He, and Yan-Xiong Li. "A two -step hybrid approach for voiceprint-biometric template protection". In: *2012 International Conference on Machine Learning and Cybernetics*. Vol. 2. 2012, pp. 560–565. DOI: 10.1109/ICMLC.2012.6358984.

[107] T. Hunt. *haveibeenpwnd*. https://haveibeenpwned.com/. Accessed: 15-06-2020.

[108] A. Husseis, J. Liu-Jimenez, I. Goicoechea-Telleria, and R. Sanchez-Reillo. "A survey in presentation attack and presentation attack detection". In: *2019 International Carnahan Conference on Security Technology (ICCST)*. IEEE. 2019, pp. 1–13.

[109] *ISO/IEC 24745:2022 Information security, cybersecurity and privacy protection - Biometric information protection*. Geneva, Switzerland: International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), 2022.

[110] *ISO/IEC 29794-4:2017 Information technology - Biometric sample quality - Part 4: Fingerprint image data*. ISO/IEC standard. Geneva, Switzerland: International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), 2017.

[111] *ISO/IEC CD 29794-5.2 Information technology - Biometric sample quality - Part 5: Face image data*. ISO/IEC Committee Draft. Geneva, Switzerland: International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), 2023.

[112] ISO/IEC JTC 1, Information technology, SC 37, Biometrics. *Information technology – Performance testing of biometric template protection schemes*. ISO/IEC 30136:2018. Geneva, CH: International Organization for Standardization, Mar. 2018.

[113] ISO/IEC JTC 1/SC 27. *ISO/IEC 19792:2009 Information technology — Security techniques — Security evaluation of biometrics*. https://www.iso.org/standard/51521.html. Aug. 2009.

[114] ISO/IEC JTC 1/SC 27. *ISO/IEC 19989-1:2020 Information security — Criteria and methodology for security evaluation of biometric systems — Part 1: Framework*. https://www.iso.org/standard/72402.html. Sept. 2020.

[115] A. T. B. Jin, D. N. C. Ling, and A. Goh. "Biohashing: two factor authentication featuring fingerprint data and tokenised random number". In: *Pattern Recognition* 37.11, 2004, pp. 2245–2255. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2004.04.011.

[116] C. Jin, R. Jin, K. Chen, and Y. Dou. "A community detection approach to cleaning extremely large face database". In: *Computational intelligence and neuroscience* 2018, 2018.

[117] S. Joos, T. Van hamme, D. Preuveneers, and W. Joosen. "Adversarial Robustness is Not Enough: Practical Limitations for Securing Facial Authentication". In: *Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*. 2022, pp. 2–12.

[118] A. Juels and M. Sudan. "A Fuzzy Vault Scheme". In: *Designs, Codes and Cryptography* 2, Feb. 2006, pp. 237–257. DOI: 10.1007/s10623-005-6343-z.

[119] A. Juels and M. Wattenberg. "A Fuzzy Commitment Scheme". In: *Proceedings of the 6th ACM Conference on Computer and Communications Security*. CCS '99. Kent Ridge Digital Labs, Singapore: ACM, 1999, pp. 28–36. ISBN: 1-58113-148-8. DOI: 10.1145/319709.319714. URL: http://doi.acm.org/10.1145/319709.319714.

[120] G. A. Kabatiansky and V. I. Levenshtein. "On bounds for packings on a sphere and in space". In: *Problemy peredachi informatsii* 14.1, 1978, pp. 3–25.

[121] N. Karimian, D. Woodard, and D. Forte. "ECG Biometric: Spoofing and Countermeasures". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020, pp. 1–1.

[122] H. Kaur and P. Khanna. "Cancelable Features Using Log-Gabor Filters for Biometric Authentication". In: *Multimedia Tools Appl.* 76.4, Feb. 2017, pp. 4673–4694. ISSN: 1380-7501. DOI: 10.1007/s11042-016-3652-3.

[123] D. Keller, M. Osadchy, and O. Dunkelman. "Fuzzy Commitments Offer Insufficient Protection to Biometric Templates Produced by Deep Learning". In: *CoRR* abs/2012.13293, 2020. arXiv: 2012.13293.

[124] P. Kenny, G. Boulianne, and P. Dumouchel. "Eigenvoice Modeling With Sparse Training Data". In: *IEEE Transactions on Speech and Audio Processing* 13.3, May 2005.

[125] K. S. Killourhy and R. A. Maxion. “Comparing anomaly-detection algorithms for keystroke dynamics”. In: *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*. IEEE. 2009, pp. 125–134.

[126] S. Komkov and A. Petiushko. “AdvHat: Real-world adversarial attack on ArcFace Face ID system”. In: *arXiv preprint arXiv:1908.08705*, 2019.

[127] R. E. Korf. “Multi-way number partitioning”. In: *Proceedings of the 21st International Jont Conference on Artifical Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009, pp. 538–543.

[128] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet. “Fooling End-To-End Speaker Verification With Adversarial Examples”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 1962–1966.

[129] R. S. Kuzu, E. Piciucco, E. Maiorana, and P. Campisi. “On-the-Fly Finger-Vein-Based Biometric Recognition Using Deep Neural Networks”. In: *IEEE Transactions on Information Forensics and Security* 15, 2020, pp. 2641–2654.

[130] V. Le Pochat, T. Van hamme, S. Maroofi, T. Van Goethem, D. Preuveneers, A. Duda, W. Joosen, and M. Korczynski. “A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints.” In: *NDSS*. 2020.

[131] J. Li, K. Fawaz, and Y. Kim. “Velody: Nonlinear Vibration Challenge-Response for Resilient User Authentication”. In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. CCS ’19. London, United Kingdom: Association for Computing Machinery, 2019, pp. 1201–1213. ISBN: 9781450367479. DOI: 10.1145/3319535.3354242. URL: https://doi.org/10.1145/3319535.3354242.

[132] S. Li, S. Ma, M. Xue, and B. Z. H. Zhao. “Deep Learning Backdoors”. In: *Security and Artificial Intelligence: A Crossdisciplinary Approach*. Ed. by L. Batina, T. Bäck, I. Buhan, and S. Picek. Cham: Springer International Publishing, 2022, pp. 313–334. ISBN: 978-3-030-98795-4. DOI: 10.1007/978-3-030-98795-4_13. URL: https://doi.org/10.1007/978-3-030-98795-4_13.

[133] Y. Li, Y. Li, Q. Yan, H. Kong, and R. H. Deng. “Seeing your face is not enough: An inertial sensor-based liveness detection for face authentication”. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 2015, pp. 1558–1569.

[134] M.-H. Lim and P. C. Yuen. “Entropy Measurement for Biometric Verification Systems”. In: *IEEE Transactions on Cybernetics* 46.5, 2016, pp. 1065–1077. DOI: 10.1109/TCYB.2015.2423271.

[135] Y. Linde, A. Buzo, and R. Gray. “An Algorithm for Vector Quantizer Design”. In: *IEEE Transactions on Communications* 28.1, 1980, pp. 84–95. DOI: 10.1109/TCOM.1980.1094577.

[136] *LivDet*. https://livdet.org. Accessed: 11-06-2020.

[137] C. López, J. Solano, E. Rivera, L. Tengana, J. Florez-Lozano, A. Castelblanco, and M. Ochoa. "Adversarial attacks against mouse-and keyboard-based biometric authentication: black-box versus domain-specific techniques". In: *International Journal of Information Security*, 2023, pp. 1–21.

[138] G. Lovisotto, S. Eberz, and I. Martinovic. *Biometric Backdoors: A Poisoning Attack Against Unsupervised Template Updating*. 2019. arXiv: 1905.09162 [cs.CR].

[139] H. Lu, J. Huang, T. Saha, and L. Nachman. "Unobtrusive Gait Verification for Mobile Phones". In: *Proceedings of the 2014 ACM International Symposium on Wearable Computers*. ISWC '14. Seattle, Washington: ACM, 2014, pp. 91–98. ISBN: 978-1-4503-2969-9. DOI: 10.1145/2634317.2642868.

[140] S. Luo, A. Nguyen, C. Song, F. Lin, W. Xu, and Z. Yan. "OcuLock: Exploring Human Visual System for Authentication in Virtual Reality Head-mounted Display". In: *Proceedings 2020 Network and Distributed System Security Symposium*. Internet Society, 2020.

[141] T. Madiega and H. Mildebrath. *Regulating facial recognition in the EU*. https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/698021/EPRS_IDA(2021)698021_EN.pdf. 2021. DOI: 10.2861/140928.

[142] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain. "On the reconstruction of face images from deep face templates". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 5 May 2019, pp. 1188–1202. ISSN: 19393539. DOI: 10.1109/TPAMI.2018.2827389.

[143] E. Maiorana, H. Kalita, and P. Campisi. "Mobile keystroke dynamics for biometric recognition: An overview". In: *IET Biometrics* 10.1, 2021, pp. 1–23. DOI: https://doi.org/10.1049/bme2.12003. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/bme2.12003. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/bme2.12003.

[144] D. Malone and W. Sullivan. "Guesswork is not a substitute for Entropy". In: *Irish Information Technology and Telecommunication conference, IT&T 2005*. 2005, pp. 1–5. URL: https://mural.maynoothuniversity.ie/6302/.

[145] Y. I. Manin and M. Marcolli. "Asymptotic bounds for spherical codes". In: *Izvestiya: Mathematics* 83.3, 2019, p. 540.

[146] Manisha and N. Kumar. "Cancelable Biometrics: a comprehensive survey". In: *Artificial Intelligence Review* 53.5, June 2020, pp. 3403–3446. DOI: 10.1007/s10462-019-09767-8.

[147] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S. M. Makela, and H. A. Ailisto. "Identifying users of portable devices from gait pattern with accelerometers". In: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* Vol. 2. 2005, ii/973–ii/976 Vol. 2. DOI: 10.1109/ICASSP.2005.1415569.

[148] P. Markert, D. V. Bailey, M. Golla, M. Dürmuth, and A. J. Aviv. "This PIN Can Be Easily Guessed: Analyzing the Security of Smartphone Unlock PINs". In: *2020 IEEE Symposium on Security and Privacy (SP)*. 2020, pp. 286–303. DOI: 10.1109/SP40000.2020.00100.

[149] F. Marra, G. Poggi, C. Sansone, and L. Verdoliva. "Blind PRNU-Based Image Clustering for Source Identification". In: *IEEE Transactions on Information Forensics and Security* 12.9, 2017, pp. 2197–2211.

[150] J. G. Marta Gomez-Barrero. "Reversing the irreversible: A survey on inverse biometrics". In: *Computers & Security* 90, Mar. 2020.

[151] I. Masi, Y. Wu, T. Hassner, and P. Natarajan. "Deep face recognition: A survey". In: *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*. IEEE. 2018, pp. 471–478.

[152] J. Massey. "Guessing and entropy". In: *Proceedings of 1994 IEEE International Symposium on Information Theory*. 1994, pp. 204–. DOI: 10.1109/ISIT.1994.394764.

[153] J. L. Massey. "Guessing and entropy". In: *Proceedings of 1994 IEEE International Symposium on Information Theory*. IEEE. 1994, p. 204.

[154] R. Matovu and A. Serwadda. "Your substance abuse disorder is an open secret! Gleaning sensitive personal information from templates in an EEG-based authentication system". In: *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. 2016, pp. 1–7. DOI: 10.1109/BTAS.2016.7791210.

[155] *Measuring Biometric Unlock Security*. https://source.android.com/security/biometric/measure. June 2022.

[156] Microsoft. *Windows Hello biometrics in the enterprise*. https://learn.microsoft.com/en-us/windows/security/identity-protection/hello-for-business/hello-biometrics-in-enterprise. Last accessed on 24 October 2022. Sept. 2022.

[157] J. V. Monaco, N. Bakelman, S.-H. Cha, and C. C. Tappert. "Developing a keystroke biometric system for continual authentication of computer users". In: *2012 European Intelligence and Security Informatics Conference*. IEEE. 2012, pp. 210–216.

[158] S. Mondal and P. Bours. "A study on continuous authentication using a combination of keystroke and mouse biometrics". In: *Neurocomputing* 230, 2017, pp. 1–22.

[159] S. Mondal and P. Bours. “Continuous authentication using mouse dynamics”. In: *2013 International Conference of the BIOSIG Special Interest Group (BIOSIG)*. IEEE. 2013, pp. 1–12.

[160] A. Morales, J. Fierrez, R. Tolosana, J. Ortega-Garcia, J. Galbally, M. Gomez-Barrero, A. Anjos, and S. Marcel. “Keystroke biometrics ongoing competition”. In: *IEEE Access* 4, 2016, pp. 7736–7746.

[161] A. Mosenia, S. Sur-Kolay, A. Raghunathan, and N. K. Jha. “CABA: Continuous authentication based on BioAura”. In: *IEEE Transactions on Computers* 66.5, 2016, pp. 759–772.

[162] M. Muaaz and R. Mayrhofer. “Smartphone-Based Gait Recognition: From Authentication to Imitation”. In: *IEEE Transactions on Mobile Computing* 16.11, 2017, pp. 3209–3221. DOI: 10.1109/TMC.2017.2686855.

[163] V. Mura, G. Orrù, R. Casula, A. Sibiriu, G. Loi, P. Tuveri, L. Ghiani, and G. L. Marcialis. “LivDet 2017 fingerprint liveness detection competition 2017”. In: *2018 International Conference on Biometrics (ICB)*. IEEE. 2018, pp. 297–302.

[164] E. Murphy-Chutorian and M. M. Trivedi. “Head pose estimation in computer vision: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 31.4, 2008, pp. 607–626.

[165] A. Nagar, K. Nandakumar, and A. K. Jain. “Multibiometric Cryptosystems Based on Feature-Level Fusion”. In: *IEEE Transactions on Information Forensics and Security* 7.1, 2012, pp. 255–268. DOI: 10.1109/TIFS.2011.2166545.

[166] K. Nandakumar and A. K. Jain. “Biometric template protection: Bridging the performance gap between theory and practice”. In: *IEEE Signal Processing Magazine* 32.5, 2015, pp. 88–100.

[167] P. Negi, P. Sharma, V. Jain, and B. Bahmani. “K-means++ vs. Behavioral Biometrics: One Loop to Rule Them All.” In: *NDSS*. 2018.

[168] T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi. “The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication”. In: *Pattern Recognition* 47.1, 2014, pp. 228–237. ISSN: 0031-3203. DOI: http://dx.doi.org/10.1016/j.patcog.2013.06.028.

[169] T. T. Ngo, M. Atiqur Rahman Ahad, A. D. Antar, M. Ahmed, D. Muramatsu, Y. Makihara, Y. Yagi, S. Inoue, T. Hossain, and Y. Hattori. “OU-ISIR Wearable Sensor-based Gait Challenge: Age and Gender”. In: *2019 International Conference on Biometrics (ICB)*. 2019, pp. 1–6. DOI: 10.1109/ICB45273.2019.8987235.

[170] D. L. Nguyen, K. Cao, and A. K. Jain. “Robust minutiae extractor: Integrating deep networks and fingerprint domain knowledge”. In: *Proceedings - 2018 International Conference on Biometrics, ICB 2018*, 2018, pp. 9–16. DOI: 10.1109/ICB2018.2018.00013. arXiv: 1712.09401.

[171] H. H. Nguyen, J. Yamagishi, I. Echizen, and S. Marcel. “Generating Master Faces for Use in Performing Wolf Attacks on Face Recognition Systems”. In: *2020 IEEE International Joint Conference on Biometrics (IJCB)*. Houston, TX, USA: IEEE Press, 2020, pp. 1–10. DOI: 10.1109/IJCB48548.2020.9304893. URL: https://doi.org/10.1109/IJCB48548.2020.9304893.

[172] K.-T. Nguyen, T.-L. Vo-Tran, D.-T. Dinh, and M.-T. Tran. “Gait Recognition with Multi-region Size Convolutional Neural Network for Authentication with Wearable Sensors”. In: *Future Data and Security Engineering*. Ed. by T. K. Dang, R. Wagner, J. Küng, N. Thoai, M. Takizawa, and E. J. Neuhold. Cham: Springer International Publishing, 2017, pp. 197–212. ISBN: 978-3-319-70004-5.

[173] C. Nickel, C. Busch, S. Rangarajan, and M. Möbius. “Using Hidden Markov Models for accelerometer-based biometric gait recognition”. In: *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. Mar. 2011, pp. 58–63. DOI: 10.1109/CSPA.2011.5759842.

[174] NIST. *Face Recognition Vendor Test (FRVT)*. https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing. Accessed: 15-06-2020.

[175] R. F. Nogueira, R. de Alencar Lotufo, and R. Campos Machado. “Fingerprint Liveness Detection Using Convolutional Neural Networks”. In: *IEEE Transactions on Information Forensics and Security* 11.6, 2016, pp. 1206–1213.

[176] L. O’Gorman. “Comparing passwords, tokens, and biometrics for user authentication”. In: *Proceedings of the IEEE* 91.12, 2003, pp. 2021–2040. DOI: 10.1109/JPROC.2003.819611.

[177] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. “Wavenet: A generative model for raw audio”. In: *arXiv preprint arXiv:1609.03499*, 2016.

[178] E. Pagnin, C. Dimitrakakis, A. Abidin, and A. Mitrokotsa. “On the Leakage of Information in Biometric Authentication”. In: *Progress in Cryptology – INDOCRYPT 2014*. Ed. by W. Meier and D. Mukhopadhyay. Cham: Springer International Publishing, 2014, pp. 265–280. ISBN: 978-3-319-13039-2.

[179] E. Pagnin and A. Mitrokotsa. “Privacy-preserving biometric authentication: challenges and directions”. In: *Security and Communication Networks* 2017, 2017.

[180] K. Patel, H. Han, and A. K. Jain. “Cross-database face antispoofing with robust feature representation”. In: *Chinese Conference on Biometric Recognition*. Springer. 2016, pp. 611–619.

[181] P. P. Paul and M. Gavrilova. “Multimodal Cancelable Biometrics”. In: *2012 IEEE 11th International Conference on Cognitive Informatics and Cognitive Computing*. 2012, pp. 43–49. DOI: 10.1109/ICCI-CC.2012.6311208.

[182] L. Peets, M. Hansen, S. J. Choi, M. Drake, and J. Ong. *European Parliament Votes in Favor of Banning the Use of Facial Recognition in Law Enforcement*. https://www.insideprivacy.com/artificial-intelligence/european-parliament-votes-in-favor-of-banning-the-use-of-facial-recognition-in-law-enforcement/. 2021.

[183] D. Peralta, M. Galar, I. Triguero, D. Paternain, S. GarcÃa, E. Barrenechea, J. M. BenÃtez, H. Bustince, and F. Herrera. “A survey on fingerprint minutiae-based local matching for verification and identification: Taxonomy and experimental evaluation”. In: *Information Sciences* 315, 2015, pp. 67–87. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2015.04.013. URL: http://www.sciencedirect.com/science/article/pii/S0020025515002819.

[184] T. Petsas, G. Tsirantonakis, E. Athanasopoulos, and S. Ioannidis. “Two-factor authentication: is the world ready? Quantifying 2FA adoption”. In: *Proceedings of the eighth european workshop on system security*. 2015, pp. 1–7.

[185] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro. “Intriguing Properties of Adversarial ML Attacks in the Problem Space”. In: *2020 IEEE Symposium on Security and Privacy (SP)*. 2020, pp. 1332–1349. DOI: 10.1109/SP40000.2020.00073.

[186] J. K. Pillai, V. M. Patel, R. Chellappa, and N. K. Ratha. “Secure and Robust Iris Recognition Using Random Projections and Sparse Representations”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.9, 2011, pp. 1877–1893. DOI: 10.1109/TPAMI.2011.34.

[187] J. O. Pliam. “On the incomparability of entropy and marginal guesswork in brute-force attacks”. In: *International conference on cryptology in India*. Springer. 2000, pp. 67–79.

[188] P. Punithavathi and G. Subbiah. “Can cancellable biometrics preserve privacy?” In: *Biometric Technology Today* 2017.7, 2017, pp. 8–11.

[189] L. R. Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proceedings of the IEEE* 77.2, 1989, pp. 257–286. DOI: 10.1109/5.18626.

[190] K. A. Rahman, K. S. Balagani, and V. V. Phoha. “Making impostor pass rates meaningless: A case of snoop-forge-replay attack on continuous cyber-behavioral verification with keystrokes”. In: *CVPR 2011 WORKSHOPS*. 2011, pp. 31–38. DOI: 10.1109/CVPRW.2011.5981729.

[191] K. B. Raja, R. Raghavendra, S. Venkatesh, M. Gomez-Barrero, C. Rathgeb, and C. Busch. “A study of hand-crafted and naturally learned features for fingerprint presentation attack detection”. In: *Handbook of Biometric Anti-Spoofing*. Springer, 2019, pp. 33–48.

[192] R. Ramachandra and C. Busch. “Presentation attack detection methods for face recognition systems: A comprehensive survey”. In: *ACM Computing Surveys (CSUR)* 50.1, 2017, pp. 1–37.

[193] S. Rane. “Standardization of Biometric Template Protection”. In: *IEEE MultiMedia* 21.4, 2014, pp. 94–99. DOI: 10.1109/MMUL.2014.65.

[194] S. Rane, Y. Wang, S. C. Draper, and P. Ishwar. “Secure biometrics: Concepts, authentication architectures, and challenges”. In: *IEEE Signal Processing Magazine* 30.5, 2013, pp. 51–64. ISSN: 10535888. DOI: 10.1109/MSP.2013.2261691.

[195] R. A. Rankin. “The closest packing of spherical caps in n dimensions”. In: *Glasgow Mathematical Journal* 2.3, 1955, pp. 139–144.

[196] N. K. Ratha, J. H. Connell, and R. M. Bolle. “Enhancing security and privacy in biometrics-based authentication systems”. English. In: *IBM Systems Journal* 40.3, 2001. Copyright - Copyright International Business Machines Corporation 2001; Last updated - 2012-02-17; CODEN - IBMSA7; SubjectsTermNotLitGenreText - United States; US, pp. 614–634. URL: https://search.proquest.com/docview/222418906?accountid=17215.

[197] C. Rathgeb and A. Uhl. “Statistical attack against iris-biometric fuzzy commitment schemes”. In: *CVPR 2011 WORKSHOPS*. 2011, pp. 23–30. DOI: 10.1109/CVPRW.2011.5981720.

[198] C. Rathgeb, J. Wagner, B. Tams, and C. Busch. “Preventing the cross-matching attack in Bloom filter-based cancelable biometrics”. In: *3rd International Workshop on Biometrics and Forensics (IWBF 2015)*. 2015, pp. 1–6. DOI: 10.1109/IWBF.2015.7110226.

[199] C. Rathgeb, F. Breitinger, C. Busch, and H. Baier. “On application of bloom filters to iris biometrics”. In: *IET Biometrics* 3.4, 2014, pp. 207–218.

[200] C. Rathgeb, M. Gomez-Barrero, C. Busch, J. Galbally, and J. Fierrez. “Towards cancelable multi-biometrics based on bloom filters: a case study on feature level fusion of face and iris”. In: *3rd international workshop on biometrics and forensics (IWBF 2015)*. IEEE. 2015, pp. 1–6.

[201] C. Rathgeb and A. Uhl. “A survey on biometric cryptosystems and cancelable biometrics”. In: *EURASIP Journal on Information Security* 2011.1, 2011, p. 3.

[202] C. Rathgeb and A. Uhl. “Statistical attack against fuzzy commitment scheme”. In: *IET Biom.* 1.2, 2012, pp. 94–104. DOI: 10.1049/iet-bmt.2011.0001.

[203] A. Rényi et al. “On measures of entropy and information”. In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 547-561. Berkeley, California, USA. 1961.

[204] D. Reynolds, T. Quatieri, and R. Dunn. “Speaker Verification Using Adapted Gaussian Mixture Models”. In: *Digital Signal Processing* 10, 2000, pp. 19–41.

[205] L. Rong, Z. Jianzhong, L. Ming, and H. Xiangfeng. “A Wearable Acceleration Sensor System for Gait Recognition”. In: *2007 2nd IEEE Conference on Industrial Electronics and Applications*. May 2007, pp. 2654–2659. DOI: 10.1109/ICIEA.2007.4318894.

[206] A. Ross, S. Banerjee, C. Chen, A. Chowdhury, V. Mirjalili, R. Sharma, T. Swearingen, and S. Yadav. “Some research problems in biometrics: The future beckons”. In: *Proc. of 12th IAPR International Conference on Biometrics (ICB)*, 2019.

[207] A. Ross and K. Nandakumar. “Fusion, Score-Level”. In: *Encyclopedia of Biometrics*. Boston, MA: Springer US, 2009, pp. 611–616. ISBN: 978-0-387-73003-5. DOI: 10.1007/978-0-387-73003-5_158.

[208] A. Rozsa, M. GÃŒnther, E. M. Rudd, and T. E. Boult. “Facial attributes: Accuracy and adversarial robustness”. In: *Pattern Recognition Letters* 124, 2019. Award Winning Papers from the 23rd International Conference on Pattern Recognition (ICPR), pp. 100–108. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2017.10.024. URL: http://www.sciencedirect.com/science/article/pii/S0167865517303926.

[209] R. San-Segundo, R. Cordoba, J. Ferreiros, and L. D’Haro-Enriquez. “Frequency features and GMM-UBM approach for gait-based person identification using smartphone inertial signals”. In: *Pattern Recognition Letters* 73, 2016, pp. 60–67. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2016.01.008. URL: http://www.sciencedirect.com/science/article/pii/S0167865516000209.

[210] R. San-Segundo, J. D. Echeverry-Correa, C. Salamea-Palacios, S. L. Lutfi, and J. M. Pardo. “I-vector analysis for Gait-based Person Identification using smartphone inertial signals”. In: *Pervasive and Mobile Computing* 38, 2017, pp. 140–153. ISSN: 1574-1192. DOI: https://doi.org/10.1016/j.pmcj.2016.09.007. URL: http://www.sciencedirect.com/science/article/pii/S1574119216302140.

[211] D. Sandberg. *facenet*. https://github.com/davidsandberg/facenet. 2022.

[212] A. Sankaran, P. Pandey, M. Vatsa, and R. Singh. "On latent fingerprint minutiae extraction using stacked denoising sparse AutoEncoders". In: *IEEE International Joint Conference on Biometrics*. 2014, pp. 1–7.

[213] F. Schroff, D. Kalenichenko, and J. Philbin. "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[214] P. Schuch, J. M. May, and C. Busch. "Unsupervised Learning of Fingerprint Rotations". In: *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2018, pp. 1–6.

[215] S. Schuckers. "Presentations and attacks, and spoofs, oh my". In: *Image and Vision Computing* 55, 2016. Recognizing future hot topics and hard problems in biometrics research, pp. 26–30. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2016.03.016. URL: http://www.sciencedirect.com/science/article/pii/S026288561630052X.

[216] A. Serwadda and V. V. Phoha. "Examining a large keystroke biometrics dataset for statistical-attack openings". In: *ACM Transactions on Information and System Security (TISSEC)* 16.2, 2013, pp. 1–30.

[217] A. Serwadda and V. V. Phoha. "When kids' toys breach mobile phone security". In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 2013, pp. 599–610.

[218] A. Serwadda, V. V. Phoha, and A. Kiremire. "Using Global Knowledge of Users' Typing Traits to Attack Keystroke Biometrics Templates". In: *Proceedings of the Thirteenth ACM Multimedia Workshop on Multimedia and Security*. MM&Sec '11. Buffalo, New York, USA: Association for Computing Machinery, 2011, pp. 51–60. ISBN: 9781450308069. DOI: 10.1145/2037252.2037263. URL: https://doi.org/10.1145/2037252.2037263.

[219] C. E. Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3, 1948, pp. 379–423.

[220] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition". In: *Proceedings of the 2016 acm sigsac conference on computer and communications security*. 2016, pp. 1528–1540.

[221] R. Shmelkin, T. Friedlander, and L. Wolf. "Generating Master Faces for Dictionary Attacks with a Network-Assisted Latent Space Evolution". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. 2021, pp. 01–08. DOI: 10.1109/FG52635.2021.9666968.

[222] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. "Membership inference attacks against machine learning models". In: *2017 IEEE symposium on security and privacy (SP)*. IEEE. 2017, pp. 3–18.

[223] K. Simoens, P. Tuyls, and B. Preneel. "Privacy Weaknesses in Biometric Sketches". In: *2009 30th IEEE Symposium on Security and Privacy*. 2009, pp. 188–203. DOI: 10.1109/SP.2009.24.

[224] N. J. A. Sloane. *Spherical Codes: Nice arrangements of points on a sphere in various dimensions*. http://neilsloane.com/packings/. Accessed: 2023-04-06.

[225] D. F. Smith, A. Wiliem, and B. C. Lovell. "Face recognition on consumer devices: Reflections on replay attacks". In: *IEEE Transactions on Information Forensics and Security* 10.4, 2015, pp. 736–745.

[226] J. Solano, L. Tengana, A. Castelblanco, E. Rivera, C. E. Lopez, and M. Ochoa. "A Few-Shot Practical Behavioral Biometrics Model for Login Authentication in Web Applications". In: 2020.

[227] J. Spooren, D. Preuveneers, and W. Joosen. "PPG2Live: Using dual PPG for active authentication and liveness detection." In: *Proceedings of the 12th IAPR International Conference on Biometrics (ICB 2019)*. IEEE. 2019.

[228] I. C. Stylios, O. Thanou, I. Androulidakis, and E. Zaitseva. "A review of continuous authentication using behavioral biometrics". In: *Proceedings of the SouthEast European Design Automation, Computer Engineering, Computer Networks and Social Media Conference*. 2016, pp. 72–79.

[229] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. "Intriguing properties of neural networks". In: *2nd International Conference on Learning Representations, ICLR 2014*. 2014.

[230] D. Tang, Z. Zhou, Y. Zhang, and K. Zhang. "Face Flashing: a Secure Liveness Detection Protocol based on Light Reflections". In: *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. URL: http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018%5C_03B-5%5C_Tang%5C_paper.pdf.

[231] Y. Tang, F. Gao, J. Feng, and Y. Liu. "FingerNet: An unified deep network for fingerprint minutiae extraction". In: *IEEE International Joint Conference on Biometrics, IJCB 2017* 2018-January, 2018, pp. 108–116. DOI: 10.1109/BTAS.2017.8272688. arXiv: 1709.02228.

[232] C. Tappert, M. Villani, and S.-H. Cha. "Keystroke Biometric Identification and Authentication on Long-Text Input". In: *Int. Journal Information Security and Privacy (IJISP)*, Jan. 2010.

[233] A. B. Teoh, Y. W. Kuan, and S. Lee. "Cancellable biometrics and annotations on biohash". In: *Pattern recognition* 41.6, 2008, pp. 2034–2044.

[234] P. Terhorst, F. Bierbaum, M. Huber, N. Damer, F. Kirchbuchner, K. B. Raja, and A. Kuijper. "On the (Limited) Generalization of MasterFace Attacks and Its Relation to the Capacity of Face Representations". In: *2022 IEEE International Joint Conference on Biometrics (IJCB)*, 2022, pp. 1–9.

[235] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. "Stealing Machine Learning Models via Prediction APIs." In: *USENIX security symposium*. Vol. 16. 2016, pp. 601–618.

[236] L. Tran, X. Yin, and X. Liu. "Disentangled representation learning gan for pose-invariant face recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1415–1424.

[237] N. T. Trung, Y. Makihara, H. Nagahara, R. Sagawa, Y. Mukaigawa, and Y. Yagi. "Phase registration in a gallery improving gait authentication". In: *2011 International Joint Conference on Biometrics (IJCB)*. Oct. 2011, pp. 1–7. DOI: 10.1109/IJCB.2011.6117527.

[238] M. Une, A. Otsuka, and H. Imai. "Wolf Attack Probability: A New Security Measure in Biometric Authentication Systems". In: *Advances in Biometrics*. Ed. by S.-W. Lee and S. Z. Li. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 396–406. ISBN: 978-3-540-74549-5.

[239] D. Valsesia, G. Coluccia, T. Bianchi, and E. Magli. "User Authentication via PRNU-Based Physical Unclonable Functions". In: *IEEE Transactions on Information Forensics and Security* 12.8, 2017, pp. 1941–1956.

[240] T. Van hamme, E. Argones Rúa, D. Preuveneers, and W. Joosen. "Gait Template Protection Using HMM-UBM". In: *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2018, pp. 1–8.

[241] T. Van hamme, G. Garofalo, E. Argones Rúa, D. Preuveneers, and W. Joosen. "A Systematic Comparison of Age and Gender Prediction on IMU Sensor-Based Gait Traces". In: *Sensors* 19.13, July 2019, p. 2945. ISSN: 1424-8220. DOI: 10.3390/s19132945.

[242] T. Van hamme, G. Garofalo, S. Joos, D. Preuveneers, and W. Joosen. "AI for Biometric Authentication Systems". In: *Security and Artificial Intelligence: A Crossdisciplinary Approach*. Ed. by L. Batina, T. Bäck, I. Buhan, and S. Picek. Vol. 13049. 8. Springer Cham, Apr. 2022, pp. 156–180. ISBN: 978-3-030-98795-4.

[243] T. Van hamme, G. Garofalo, D. Preuveneers, and W. Joosen. "Masterkey Attacks Against Free-Text Keystroke Dynamics and Security Implications of Demographic Factors". In: *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. Los Alamitos, CA, USA: IEEE Computer Society, July 2023, pp. 278–291. DOI: 10.1109/EuroSP57164.2023.00025.

[244] T. Van hamme, D. Preuveneers, and W. Joosen. “A Dynamic Decision Fusion Middleware for Trustworthy Context-Aware IoT Applications”. In: *Proceedings of the 4th Workshop on Middleware and Applications for the Internet of Things*. M4IoT ’17. Las Vegas, Nevada: Association for Computing Machinery, 2017, pp. 1–6. ISBN: 9781450351706.

[245] T. Van hamme, D. Preuveneers, and W. Joosen. “Improving Resilience of Behaviometric Based Continuous Authentication with Multiple Accelerometers”. In: *Data and Applications Security and Privacy XXXI*. Ed. by G. Livraga and S. Zhu. Cham: Springer International Publishing, 2017, pp. 473–485. ISBN: 978-3-319-61176-1.

[246] T. Van hamme, D. Preuveneers, and W. Joosen. “Managing Distributed Trust Relationships for Multi-modal Authentication”. In: *Journal of Information Security and Applications* 40, 2018, pp. 258–270. ISSN: 2214-2126.

[247] T. Van hamme, V. Rimmer, D. Preuveneers, W. Joosen, M. Mustafa, A. Abidin, and E. Argones Rua. “Frictionless Authentication Systems: Emerging Trends, Research Challenges and Opportunities”. In: *SECURWARE 2017*. Vol. abs/1802.07233. IARIA, 2017, pp. 1–5.

[248] T. Van hamme, E. A. Rúa, D. Preuveneers, and W. Joosen. “On the Security of Biometrics and Fuzzy Commitment Cryptosystems: A Study on Gait Authentication”. In: *IEEE Transactions on Information Forensics and Security* 16, 2021, pp. 5211–5224. DOI: 10.1109/TIFS.2021.3124735.

[249] G. Vassallo, T. Van hamme, D. Preuveneers, and W. Joosen. “Privacy-preserving Behavioral Authentication on Smartphones”. In: *Proceedings of the First International Workshop on Human-Centered Sensing, Networking, and Systems*. 2017, pp. 1–6.

[250] W. Verheyen, T. Van hamme, S. Joos, D. Preuveneers, and W. Joosen. “Beware the Doppelgänger: Attacks against Adaptive Thresholds in Facial Recognition Systems”. In: *The 18th International Conference on Availability, Reliability and Security Proceedings*. Benevento, Italy, Sept. 2023.

[251] D. Wang, Q. Gu, X. Huang, and P. Wang. “Understanding Human-Chosen PINs: Characteristics, Distribution and Security”. In: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ASIA CCS ’17. Abu Dhabi, United Arab Emirates: Association for Computing Machinery, 2017, pp. 372–385. ISBN: 9781450349444. DOI: 10.1145/3052973.3053031.

[252] M. Wang and W. Deng. “Deep face recognition: A survey”. In: *Neurocomputing* 429, 2021, pp. 215–244.

[253] S. Wang and J. Hu. "A Hadamard transform-based method for the design of cancellable fingerprint templates". In: *2013 6th International Congress on Image and Signal Processing (CISP)*. Vol. 03. 2013, pp. 1682–1687. DOI: 10.1109/CISP.2013.6743947.

[254] Y. Wang, C. Wu, K. Zheng, and X. Wang. "Improving Reliability: User Authentication on Smartphones Using Keystroke Biometrics". In: *IEEE Access* 7, 2019, pp. 26218–26228.

[255] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao. "Backdoor attacks against deep learning systems in the physical world". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 6206–6215.

[256] G. Á. Werner and L. Hanka. "Using the beta-binomial distribution for the analysis of biometric identification". In: *2015 IEEE 13th International Symposium on Intelligent Systems and Informatics (SISY)*. 2015, pp. 209–215. DOI: 10.1109/SISY.2015.7325381.

[257] C. Wu, K. He, J. Chen, Z. Zhao, and R. Du. "Liveness is Not Enough: Enhancing Fingerprint Authentication with Behavioral Biometrics to Defeat Puppet Attacks". In: *Proceedings of the 29th USENIX Conference on Security Symposium*. SEC'20. USA: USENIX Association, 2020. ISBN: 978-1-939133-17-5.

[258] M. M.-A. Wu, O. S. Schneider, I. Karuei, L. Leong, and K. MacLean. "Introducing GaitLib: a library for real-time gait analysis in smartphones". In: 2014.

[259] Y. Xu, K. Raja, R. Ramachandra, and C. Busch. "Adversarial attacks on face recognition systems". In: *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Springer International Publishing Cham, 2022, pp. 139–161.

[260] N. Yager and T. Dunstone. "The Biometric Menagerie". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.2, 2010, pp. 220–230. DOI: 10.1109/TPAMI.2008.291.

[261] C. Yan, Y. Long, X. Ji, and W. Xu. "The Catcher in the Field: A Fieldprint Based Spoofing Detection for Text-Independent Speaker Verification". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. CCS '19. London, United Kingdom: Association for Computing Machinery, 2019, pp. 1215–1229. ISBN: 9781450367479. DOI: 10.1145/3319535.3354248. URL: https://doi.org/10.1145/3319535.3354248.

[262] M.-H. Yang, D. J. Kriegman, and N. Ahuja. "Detecting faces in images: A survey". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.1, 2002, pp. 34–58.

[263] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher. "DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 351–360. ISBN: 978-1-4503-4913-0. DOI: 10.1145/3038912.3052577. URL: https://doi.org/10.1145/3038912.3052577.

[264] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao. "Latent Backdoor Attacks on Deep Neural Networks". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. CCS '19. London, United Kingdom: Association for Computing Machinery, 2019, pp. 2041–2055. ISBN: 9781450367479. DOI: 10.1145/3319535.3354209.

[265] C. Zhang and Z. Zhang. "A survey of recent advances in face detection". In: 2010.

[266] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu. "Dolphinattack: Inaudible voice commands". In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017, pp. 103–117.

[267] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks". In: *IEEE Signal Processing Letters* 23.10, Oct. 2016, pp. 1499–1503. ISSN: 1070-9908. DOI: 10.1109/LSP.2016.2603342.

[268] B. Zhao, H. Asghar, and D. Kaafar. "On the Resilience of Biometric Authentication Systems against Random Inputs". In: Jan. 2020. DOI: 10.14722/ndss.2020.24210.

[269] B. Z. H. Zhao, H. J. Asghar, and M. A. Kaafar. "On the Resilience of Biometric Authentication Systems against Random Inputs". In: *Proceedings 2020 Network and Distributed System Security Symposium*, 2020. DOI: 10.14722/ndss.2020.24210. URL: http://dx.doi.org/10.14722/ndss.2020.24210.

[270] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. "Face recognition: A literature survey". In: *ACM computing surveys (CSUR)* 35.4, 2003, pp. 399–458.

[271] Y. Zhao and S. Zhou. "Wearable Device-Based Gait Recognition Using Angle Embedded Gait Dynamic Images and a Convolutional Neural Network". In: *Sensors* 17.3, 2017. ISSN: 1424-8220. DOI: 10.3390/s17030478. URL: http://www.mdpi.com/1424-8220/17/3/478.

[272] Y. Zhong and Y. Deng. "Sensor orientation invariant mobile gait biometrics". In: *IEEE International Joint Conference on Biometrics*. Sept. 2014, pp. 1–8. DOI: 10.1109/BTAS.2014.6996246.

[273] X. Zhu and D. Ramanan. “Face detection, pose estimation, and landmark localization in the wild”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 2879–2886.

[274] K. Zico. *Robustness in Machine Learning: A Five-Year Retrospective*. IEE Secure and Trustworthy ML (SaTML) 2023 keynote - YouTube video. Feb. 2023. URL: %5Curl%7Bhttps://www.youtube.com/watch?v=jm4pfAP_hPs%7D.

[275] Q. Zou, Y. Wang, Q. Wang, Y. Zhao, and Q. Li. “Deep learning-based gait recognition using smartphones in the wild”. In: *IEEE Transactions on Information Forensics and Security* 15, 2020, pp. 3197–3212.

List of publications

- W. Verheyen, T. Van hamme, S. Joos, D. Preuveneers, and W. Joosen. “Beware the Doppelgänger: Attacks against Adaptive Thresholds in Facial Recognition Systems”. In: *The 18th International Conference on Availability, Reliability and Security Proceedings*. Benevento, Italy, Sept. 2023.
- T. Van hamme, G. Garofalo, D. Preuveneers, and W. Joosen. “Masterkey Attacks Against Free-Text Keystroke Dynamics and Security Implications of Demographic Factors”. In: *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. Los Alamitos, CA, USA: IEEE Computer Society, July 2023, pp. 278–291. DOI: 10.1109/EuroSP57164.2023.00025.
- E. Argones Rúa, T. Van hamme, D. Preuveneers, and W. Joosen. “Discriminative Training of Spiking Neural Networks Organised in Columns for Stream-based Biometric Authentication”. In: *IET Biometrics* 11.5, 2022, pp. 485–497.
- S. Joos, T. Van hamme, D. Preuveneers, and W. Joosen. “Adversarial Robustness is Not Enough: Practical Limitations for Securing Facial Authentication”. In: *Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*. 2022, pp. 2–12.
- T. Van hamme, G. Garofalo, S. Joos, D. Preuveneers, and W. Joosen. “AI for Biometric Authentication Systems”. In: *Security and Artificial Intelligence: A Crossdisciplinary Approach*. Ed. by L. Batina, T. Bäck, I. Buhan, and S. Picek. Vol. 13049. 8. Springer Cham, Apr. 2022, pp. 156–180. ISBN: 978-3-030-98795-4.
- T. Van hamme, E. A. Rúa, D. Preuveneers, and W. Joosen. “On the Security of Biometrics and Fuzzy Commitment Cryptosystems: A Study on Gait Authentication”. In: *IEEE Transactions on Information Forensics and Security* 16, 2021, pp. 5211–5224. DOI: 10.1109/TIFS.2021.3124735.

- E. Argones Rúa, T. Van hamme, D. Preuveneers, and W. Joosen. "Gait Authentication based on Spiking Neural Networks". In: *20th International Conference of the Biometrics Special Interest Group*. Vol. LNI P-315. Gesellschaft für Informatik e.V., Oct. 2021, pp. 51–60. ISBN: 978-3-88579-709-8.
- V. Le Pochat, T. Van hamme, S. Maroofi, T. Van Goethem, D. Preuveneers, A. Duda, W. Joosen, and M. Korczynski. "A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints." In: *NDSS*. 2020.
- T. Van hamme, G. Garofalo, E. Argones Rúa, D. Preuveneers, and W. Joosen. "A Systematic Comparison of Age and Gender Prediction on IMU Sensor-Based Gait Traces". In: *Sensors* 19.13, July 2019, p. 2945. ISSN: 1424-8220. DOI: 10.3390/s19132945.
- G. Garofalo, V. Rimmer, T. Van hamme, D. Preuveneers, W. Joosen, et al. "Fishy Faces: Crafting Adversarial Images to Poison Face Authentication". In: *12th USENIX Workshop on Offensive Technologies (WOOT 18)*. 2018.
- T. Van hamme, E. Argones Rúa, D. Preuveneers, and W. Joosen. "Gait Template Protection Using HMM-UBM". In: *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2018, pp. 1–8.
- T. Van hamme, D. Preuveneers, and W. Joosen. "Managing Distributed Trust Relationships for Multi-modal Authentication". In: *Journal of Information Security and Applications* 40, 2018, pp. 258–270. ISSN: 2214-2126.
- T. Van hamme, D. Preuveneers, and W. Joosen. "A Dynamic Decision Fusion Middleware for Trustworthy Context-Aware IoT Applications". In: *Proceedings of the 4th Workshop on Middleware and Applications for the Internet of Things*. M4IoT '17. Las Vegas, Nevada: Association for Computing Machinery, 2017, pp. 1–6. ISBN: 9781450351706.
- T. Van hamme, D. Preuveneers, and W. Joosen. "Improving Resilience of Behaviometric Based Continuous Authentication with Multiple Accelerometers". In: *Data and Applications Security and Privacy XXXI*. Ed. by G. Livraga and S. Zhu. Cham: Springer International Publishing, 2017, pp. 473–485. ISBN: 978-3-319-61176-1.
- T. Van hamme, V. Rimmer, D. Preuveneers, W. Joosen, M. Mustafa, A. Abidin, and E. Argones Rua. "Frictionless Authentication Systems: Emerging Trends, Research Challenges and Opportunities". In: *SECURWARE 2017*. Vol. abs/1802.07233. IARIA, 2017, pp. 1–5.
- G. Vassallo, T. Van hamme, D. Preuveneers, and W. Joosen. "Privacy-preserving Behavioral Authentication on Smartphones". In: *Proceedings of the First International Workshop on Human-Centered Sensing, Networking, and Systems*. 2017, pp. 1–6.



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
DISTRINET
Celestijnenlaan 200A box 2402
B-3001 Leuven
tim.vanhamme@kuleuven.be
http://www.distrinet.cs.kuleuven.be