
Mitigating Soft-Biometric Driven Bias and Privacy Concerns in Face Recognition Systems

**Entschärfung soft-biometrischer Bedenken hinsichtlich der Privatsphäre und
Voreingenommenheit von Gesichtserkennungssystemen**

Zur Erlangung des akademischen Grades Doktor-Ingenieur (Dr.-Ing.)
genehmigte Dissertation von Philipp Terhörst aus Bocholt, Germany
Tag der Einreichung: 24.02.2021, Tag der Prüfung: 20.04.2021

1. Gutachten: Prof. Dr. Arjan Kuijper
2. Gutachten: Prof. Dr. Dieter W. Fellner
3. Gutachten: Prof. Vitomir Štruc
Darmstadt



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Computer Science
Department
Interactive Graphics
Systems Group

Mitigating Soft-Biometric Driven Bias and Privacy Concerns in Face Recognition Systems
Entschärfung soft-biometrischer Bedenken hinsichtlich der Privatsphäre und
Voreingenommenheit von Gesichtserkennungssystemen

Doctoral thesis by Philipp Terhörst

1. Review: Prof. Dr. Arjan Kuijper
2. Review: Prof. Dr. Dieter W. Fellner
3. Review: Prof. Vitomir Štruc

Date of submission: 24.02.2021

Date of thesis defense: 20.04.2021

Darmstadt

Bitte zitieren Sie dieses Dokument als:

URN: urn:nbn:de:tuda-tuprints-185152

URL: <http://tuprints.ulb.tu-darmstadt.de/18515>

Dieses Dokument wird bereitgestellt von tuprints,

E-Publishing-Service der TU Darmstadt

<http://tuprints.ulb.tu-darmstadt.de>

tuprints@ulb.tu-darmstadt.de

Attribution 4.0 International (CC BY 4.0)

Be curious, be creative, and have someone at your side.

For Tanja

Erklärungen laut Promotionsordnung

§8 Abs. 1 lit. c PromO

Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§8 Abs. 1 lit. d PromO

Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§9 Abs. 1 PromO

Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§9 Abs. 2 PromO

Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 24.02.2021

P. Terhörst

Abstract

Biometric verification refers to the automatic verification of a person's identity based on their behavioural and biological characteristics. Among various biometric modalities, the face is one of the most widely used since it is easily acquirable in unconstrained environments and provides a strong uniqueness. In recent years, face recognition systems spread world-wide and are increasingly involved in critical decision-making processes such as finance, public security, and forensics. The growing effect of these systems on everybody's daily life is driven by the strong enhancements in their recognition performance.

The advances in extracting deeply-learned feature representations from face images enabled the high-performance of current face recognition systems. However, the success of these representations came at the cost of two major discriminatory concerns. These concerns are driven by soft-biometric attributes such as demographics, accessories, health conditions, or hairstyles.

The first concern is about bias in face recognition. Current face recognition solutions are built on representation-learning strategies that optimize total recognition performance. These learning strategies often depend on the underlying distribution of soft-biometric attributes in the training data. Consequently, the behaviour of the learned face recognition solutions strongly varies depending on the individual's soft-biometrics (e.g. based on the individual's ethnicity).

The second concern tackles the user's privacy in such systems. Although face recognition systems are trained to recognize individuals based on face images, the deeply-learned representation of an individual contains more information than just the person's identity. Privacy-sensitive information such as demographics, sexual orientation, or health status, is encoded in such representations. However, for many applications, the biometric data is expected to be used for recognition only and thus, raises major privacy issues. The unauthorized access of such individual's privacy-sensitive information can lead to unfair or unequal treatment of this individual.

Both issues are caused by the presence of soft-biometric attribute information in the face images. Previous research focused on investigating the influence of demographic attributes on both concerns. Consequently, the solutions from previous works focused on the mitigation of demographic-concerns only as well. Moreover, these approaches require

computationally-heavy retraining of the deployed face recognition model and thus, are hardly-integrable into existing systems.

Unlike previous works, this thesis proposes solutions to mitigating soft-biometric driven bias and privacy concerns in face recognition systems that are easily-integrable in existing systems and aim for more comprehensive mitigation, not limited to pre-defined demographic attributes. This aims at enhancing the reliability, trust, and dissemination of these systems.

The first part of this work provides in-depth investigations on soft-biometric driven bias and privacy concerns in face recognition over a wide range of soft-biometric attributes. The findings of these investigations guided the development of the proposed solutions. The investigations showed that a high number of soft-biometric and privacy-sensitive attributes are encoded in face representations. Moreover, the presence of these soft-biometric attributes strongly influences the behaviour of face recognition systems. This demonstrates the strong need for more comprehensive privacy-enhancing and bias-mitigating technologies that are not limited to pre-defined (demographic) attributes.

Guided by these findings, this work proposes solutions for mitigating bias in face recognition systems and solutions for the enhancement of soft-biometric privacy in these systems. The proposed bias-mitigating solutions operate on the comparison- and score-level of recognition system and thus, can be easily integrated. Incorporating the notation of individual fairness, that aims at treating similar individuals similarly, strongly mitigates bias of unknown origins and further improves the overall-recognition performance of the system.

The proposed solutions for enhancing the soft-biometric privacy in face recognition systems either manipulate existing face representations directly or changes the representation type including the inference process for verification. The manipulation of existing face representations aims at directly suppressing the encoded privacy-risk information in an easily-integrable manner. Contrarily, the inference-level solutions indirectly suppress this privacy-risk information by changing the way of how this information is encoded.

To summarise, this work investigates soft-biometric driven bias and privacy concerns in face recognition systems and proposed solutions to mitigate these. Unlike previous works, the proposed approaches are (a) highly effective in mitigating these concerns, (b) not limited to the mitigation of concerns origin from specific attributes, and (c) easily-integrable into existing systems. Moreover, the presented solutions are not limited to face biometrics and thus, aim at enhancing the reliability, trust, and dissemination of biometric systems in general.

Zusammenfassung

Biometrische Verifizierung verweist auf die automatische Überprüfung der Identität einer Person auf der Grundlage ihrer Verhaltens- und biologischen Merkmale. Unter den verschiedenen biometrischen Modalitäten ist das Gesicht eine der am weitesten verbreiteten, da es in einer unbeschränkten Umgebung leicht zu erfassen ist und gleichzeitig eine starke Einzigartigkeit bietet. In den letzten Jahren haben sich Gesichtserkennungssysteme weltweit verbreitet und werden auch zunehmend in kritische Entscheidungsprozesse wie im Finanzwesen, der öffentlichen Sicherheit oder der Forensik einbezogen. Der wachsende Einfluss dieser Systeme auf das tägliche Leben eines jeden Menschen wird durch die starke Verbesserung ihrer Erkennungsleistung angetrieben.

Die Fortschritte bei der Extraktion von Feature-Repräsentationen mit tiefen neuronalen Netzen aus Gesichtsbildern ermöglichten die hohe Leistungsfähigkeit der aktuellen Gesichtserkennungssysteme. Der Erfolg dieser Darstellungen kam jedoch auf Kosten zweier wesentlicher diskriminierender Bedenken. Diese Bedenken werden durch soft-biometrische Merkmale wie demographische Daten, Accessoires, Gesundheitszustände oder Frisuren hervorgerufen.

Das erste Bedenken bezieht sich auf die Voreingenommenheit von Gesichtserkennungssystemen. Aktuelle Gesichtserkennungslösungen bauen auf Repräsentations-Lernstrategien auf, die auf eine optimale Gesamterkennungsleistung ausgelegt ist. Solche Lernstrategien hängen stark von der zugrundeliegenden Verteilung der soft-biometrischen Attribute in den Trainingsdaten ab und beeinflussen daher stark die Erkennungsleistung verschiedener Individuen abhängig von diesen Attributen.

Das zweite Bedenken betrifft die Privatsphäre der Benutzer in solchen Systemen. Obwohl Gesichtserkennungssysteme darauf trainiert sind, Personen anhand von Gesichtsbildern zu erkennen, enthalten die gelernten Repräsentationen einer Person mehr Informationen als nur ihre Identität. Datenschutzrelevante Informationen wie demographische Daten, sexuelle Orientierung oder Gesundheitszustand der Person sind in solchen Darstellungen enthalten. Bei vielen Anwendungen wird jedoch davon ausgegangen, dass die biometrischen Daten nur zur Erkennung verwendet werden, was große Probleme bezüglich der Privatsphäre aufwirft. Bei vielen Anwendungen wird jedoch erwartet, dass die biometrischen Daten nur zur Erkennung verwendet werden. Das trotzdem solche datenschutzrelevanten

Informationen enthalten sind wirft große Bedenken hinsichtlich der Nutzerprivatsphäre auf. Der unbefugte Zugriff auf die datenschutzsensitiven Informationen einer Person kann zu einer ungerechten oder ungleichen Behandlung dieser Person führen. Dieser unbefugte Zugriff auf die sensiblen Daten einer Person kann zu einer unfairen oder diskriminierenden Behandlung dieser Person führen.

Beide Bedenken werden durch das Vorhandensein von Informationen über soft-biometrische Attribute in den Gesichtsbildern verursacht. Frühere Forschungsarbeiten konzentrierten sich auf die Untersuchung des Einflusses demographischer Merkmale auf beide Bedenken. Folglich fokussierten sich auch die Lösungen aus früheren Arbeiten nur auf die Entschärfung der demographischen Bedenken. Darüber hinaus erfordern diese Ansätze rechenintensive Trainings der eingesetzten Gesichtserkennungsmodelle und sind daher nur schwer in bestehende Systeme integrierbar.

Im Gegensatz zu früheren Arbeiten werden in dieser Dissertation Lösungen zur Entschärfung soft-biometrisch bedingter Voreingenommenheit und Datenschutzbedenken in Gesichtserkennungssystemen vorgeschlagen, die leicht in bestehende Systeme integrierbar sind und auf eine umfassendere Entschärfung abzielen, die sich nicht auf vordefinierte demografische Merkmale beschränkt. Dadurch sollen die Zuverlässigkeit, das Vertrauen und die Verbreitung dieser Systeme verbessert werden.

Der erste Teil dieser Arbeit bietet eingehende Untersuchungen zu der soft-biometrisch bedingten Voreingenommenheit und Datenschutzbedenken bei der Gesichtserkennung über ein breites Spektrum soft-biometrischer Merkmale. Die Erkenntnisse aus diesen Untersuchungen dienen als Grundlage für die Entwicklung der Lösungsvorschläge.

Die Untersuchungen zeigten, dass eine hohe Anzahl von soft-biometrischen und datenschutzrelevanten Attributen in Gesichtsrepräsentationen enthalten ist. Darüber hinaus beeinflusst das Vorhandensein dieser weich-biometrischen Attribute stark das Verhalten der Gesichtserkennungssysteme. Dies zeigt den dringenden Bedarf an weiterführenden Technologien zur Verbesserung der Privatsphäre und zur Verringerung von der systembedingten Voreingenommenheit, die nicht auf vordefinierte (demografische) Attribute beschränkt ist.

Geleitet von diesen Erkenntnissen werden in dieser Arbeit Lösungen zur Entschärfung von Voreingenommenheit in Gesichtserkennungssystemen und Lösungen zur Verbesserung der soft-biometrischen Privatsphäre in diesen Systemen vorgeschlagen. Die vorgeschlagenen Lösungen zur Minderung der Voreingenommenheit arbeiten auf der Vergleichs- und Score-Ebene des Erkennungssystems und können daher leicht in bestehende Systeme integriert werden. Durch die Integration der Notation der individuellen Fairness, die darauf abzielt ähnliche Personen ähnlich zu behandeln, werden Voreingenommenheit unbekannter Herkunft stark abgeschwächt und die Gesamterkennungsleistung des Systems zusätzlich verbessert.

Die vorgeschlagenen Lösungen zur Verbesserung der soft-biometrischen Privatsphäre in Gesichtserkennungssystemen basieren entweder auf der direkten Manipulation bestehender Gesichtsrepräsentationen oder auf der Änderung des Darstellungstyps dieser Repräsentationen einschließlich des Inferenzprozesses zur Verifizierung. Während die Manipulation vorhandener Gesichtsrepräsentationen auf die direkte Unterdrückung der Privatsphäre relevanten Informationen abzielt, unterdrücken die vorgeschlagenen Lösungen, die auf der Inferenz-Ebene des Erkennungssystems arbeiten, indirekt diese Informationen, indem sie die Art und Weise ändern, wie diese Informationen kodiert werden.

Zusammengefasst untersucht diese Arbeit soft-biometrisch bedingte Voreingenommenheit und Datenschutzbedenken in Gesichtserkennungssystemen und präsentiert Lösungen, um diese zu entschärfen. Im Gegensatz zu früheren Arbeiten sind die vorgeschlagenen Ansätze (a) hocheffektiv bei der Verminderung dieser Bedenken, (b) nicht auf die Verminderung von Bedenken beschränkt, die nur von spezifischen Attributen ausgehen, und (c) leicht in bestehende Systeme integrierbar. Außerdem sind die vorgestellten Lösungen nicht auf die Gesichtsbioometrie begrenzt und zielen daher darauf ab, die Zuverlässigkeit, das Vertrauen und die Verbreitung biometrischer Systeme im Allgemeinen zu verbessern.

Acknowledgement

The achievements of the last three years would not have been possible without the support of many people. First of all, I want to thank my Ph.D. supervisor Prof. Dr. Arjan Kuijper for his scientific guidance, positive leadership, and for asking the right questions at the right time. I would also like to express my gratitude to Prof. Dr. techn. Dieter W. Fellner for his strong efforts at creating scientific excellence at Fraunhofer IGD.

A special thanks go to my colleague, supervisor, and friend Dr. Naser Damer for endless hours of scientific discussions and valuable life lessons. He taught me a scientific mindset and inevitable concepts such as the aura of logical distortion.

I would like to express my sincere appreciation to the former and the new head of our department. Dr. Andreas Braun always provided me with open and direct feedback and gave me this position. Florian Kirchbuchner pushed me to participate at the Software Campus program and showed me his endless trust by giving me so much scientific freedom. I would also like to thank all my friends and companions of Fraunhofer IGD. A special thanks go to the members of SLBT, Fadi, Daniel, Meiling, Biying, Uschi, Conny, Andrea, Saied, Julian, Naser, Olaf, Florian, Silvia, Andreas, Javier, Dirk, and Hildegerd. Even if the pressure of an upcoming deadline was imminent, it was always fun with you. A big thanks go also to Viola, Yaza, Aidmar, Alexa, Marius, Cong, Tanja, Timos, Vinh Phuc, Doreen, Nils, and Steffen for creating a comfortable atmosphere at work.

During my PhD, I spend a great time at supervising students. But I guess I learned more from them than they from me. With their endless enthusiasm and the will to extend the research boundaries, we had a fun and productive time. Thank you, André, Paul, Mai Ly, Moritz, Vicky, Adrian, Malte, Michael, Sebastian, Florian, Alexander, Christopher, Ines, Lydia, Serif, and Kevin. A special thanks go to my longest and most dedicated students for their years of trust. Thank you, Jonas, Daniel, Spyderman (Jan) and Marco.

My appreciation is extended to all whole biometrics and computer vision community. Our research fields are developing fast and in many directions thanks to you. My warmest thanks go to Kiran, Frøy, Vito, and Sudipta. Our talks were always fun but also more valuable than you might expect.

Finally, I want to thank my family, friends, and especially Tanja for their continuous support and encouragement. You are the mental pillars that keep my life in balance.

Contents

1. Introduction	1
1.1. Research Questions	2
1.1.1. Reliable Estimation of Soft-Biometrics	2
1.1.2. Mitigation of Soft-Biometric Bias	4
1.1.3. Enhancement of Soft-Biometric Privacy	4
1.2. Thesis Overview	5
2. Background	9
2.1. Biometrics	9
2.1.1. Biometric Systems	10
2.1.2. Soft-Biometrics	14
2.2. Face Recognition	15
2.2.1. Milestones of Face Representations for Recognition	15
2.2.2. Components of a Face Recognition System	16
2.2.3. Deep Face Recognition Models	18
2.3. Performance Metrics	23
2.3.1. Evaluating Verification Performance	24
2.3.2. Evaluating Soft-Biometric Privacy-Preservation	28
2.4. Summary	30
3. Investigation of Soft-Biometric Driven Bias and Privacy Concerns	31
3.1. Introduction	31
3.2. Preliminary Investigations	32
3.2.1. Reliable Estimation of Soft-Biometrics	32
3.2.2. MAAD-Face: A Massively-Annotated Face Dataset	46
3.3. Investigating Bias in Face Recognition	62
3.3.1. Introduction	62
3.3.2. Related Work	63
3.3.3. Experiments on Measuring Differential Performance	64
3.3.4. Results	67

3.3.5. Interim Conclusion	82
3.4. Investigating Bias in Face Quality Assessment	84
3.4.1. Introduction	84
3.4.2. Related Work	85
3.4.3. Evaluated Face Quality Assessment Solutions	86
3.4.4. Experimental Setup	88
3.4.5. Results	90
3.4.6. Interim Conclusion	95
3.5. Analysing Soft-Biometric Characteristics in Face Templates	98
3.5.1. Introduction	98
3.5.2. Related Work	99
3.5.3. Methodology	100
3.5.4. Experimental Setup	103
3.5.5. Results	106
3.5.6. Interim Conclusion	111
3.6. Summary	113
4. Integrable Bias-Mitigation	115
4.1. Introduction	115
4.2. Related Work	116
4.3. Mitigating Bias on Comparison-Level	118
4.3.1. Methodology	118
4.3.2. Experimental Setup	120
4.3.3. Results	122
4.3.4. Interim Conclusion	124
4.4. Mitigating Bias on Score-Level	127
4.4.1. Methodology	128
4.4.2. Experimental Setup	130
4.4.3. Results	132
4.4.4. Interim Conclusion	135
4.5. Summary	142
5. Enhancing Soft-Biometric Privacy	145
5.1. Introduction	145
5.2. Related Work	146
5.2.1. Image-Level Solutions	147
5.2.2. Template-Level Solutions	147
5.2.3. Soft-Biometric Privacy and Cancelable Biometrics	149

5.3. Incremental Variable Elimination	149
5.3.1. Methodology	150
5.3.2. Experimental Setup	153
5.3.3. Results	154
5.3.4. Interim Conclusion	158
5.4. Similarity-Sensitive Noise Transformations	164
5.4.1. Methodology	165
5.4.2. Experimental Setup	167
5.4.3. Results	170
5.4.4. Interim Conclusion	180
5.5. Negative Face Recognition	192
5.5.1. Methodology	192
5.5.2. Experimental Setup	197
5.5.3. Results	200
5.5.4. Interim Conclusion	204
5.6. PE-MIU: Privacy-Enhancement via Minimum Information Units	210
5.6.1. Methodology	210
5.6.2. Experimental Setup	216
5.6.3. Results	219
5.6.4. Interim Conclusion	223
5.7. Summary	229
6. Conclusion and Future Work	231
6.1. Conclusion	231
6.2. Future work	236
A. Appendix	239
B. Publications and Talks	255
B.1. Publications	255
B.2. Invited Talks	260
C. Supervising Activities	261
D. Curriculum Vitae	263
Bibliography	267

1. Introduction

Biometric verification is defined as the automated recognition of individuals based on their behavioural or biological characteristics [12]. In the last decades, biometric identification and verification systems have increasingly gained importance for a variety of enterprise, civilian, and law enforcement applications [SSW09]. Modern electronic passports [15] and IDs [NH08] already contain biometric information of their legitimate holders, such as face images, fingerprints, and iris scans. Among various biometric modalities, the face is one of the most widely used. It is ubiquitous and also acquirable in unconstrained environments. Face recognition systems provide a strong discriminative recognition performance [Mas+18] that led to a world-wide spreading of these systems and a growing effect on everybody's daily life. Moreover, they are increasingly involved in critical decision-making processes, such as in finance, public security, and forensics [WD18].

The high-performance of current face recognition systems is based on the advances in extracting deeply-learned feature representations of face images [JNR16]. These deeply-learned representations of faces, known as face templates, are characterized by high compactness and strong identity discriminability. However, the success of these templates came at the cost of two major discriminatory concerns:

Bias concerns - Many biometric solutions are built on representation-learning strategies that optimize total recognition performance. Since these learning strategies might be strongly dependent on the underlying distribution of the training data, the performance of the learned solution is often depended on the training data properties as well [GNH19b]. Consequently, this can lead to strong discriminatory effects, e.g. in forensic investigations or law enforcement [Dam+18d].

Privacy concerns - The deeply-learned template of an individual contains more information than just the individual's identity. Privacy-sensitive information, such as gender, age, ethnicity, sexual orientation, and health status, is deducible from such a template [DER16]. Since for many applications, these templates are expected to be used for recognition purposes only, this raises major privacy issues. The unauthorized access of an individual's privacy-sensitive information can lead to

unfair or unequal treatment of this individual. Soft-biometric privacy aims to reduce this kind of discriminatory concern.

Several political regulations point out the importance of the right to non-discrimination. These include Article 14 of the European Convention of Human Rights, Article 7 of the Universal Declaration of Human Rights, and Article 71 of the General Data Protection Regulation (GDPR) [VB17]. These political efforts show the importance of mitigating privacy and bias concerns in face recognition systems. Mitigating these concerns could lead to more reliable and trusted face recognition systems [SSW09]. Moreover, it might enhance the public acceptability of face recognition solutions and thus, enable an even broader application of this technology [SP11].

1.1. Research Questions

Based on the current state-of-the-art, this thesis aims at mitigating soft-biometric driven bias and privacy concerns in face recognition systems through a set of unsolved research questions. These questions aim at reducing the discriminatory effects on the users of these systems to make face recognition more reliable, trusted, and secure. To put these questions into a broader perspective and to provide topic-specific answers, these are divided into three groups based on their research area. The first group focuses on reliable estimations of soft-biometric attributes, as these are the origin of the mentioned bias and privacy concerns. The second group of questions focuses on investigating and mitigating soft-biometric bias and the third group focuses on investigating and mitigating soft-biometric driven privacy concerns in face recognition systems.

Figure 1.1 provides an overview of the research questions linked to the main contributions of this work. The research questions and contributions in the blue area deal with soft-biometric bias while the contributions and questions in the green area focus on soft-biometric privacy. On the bottom left of the figure, a legend is shown providing additional details on the contributions.

1.1.1. Reliable Estimation of Soft-Biometrics

To mitigate soft-biometric driven bias and privacy concerns, a tool is needed that is able to reliably estimate soft-biometric attributes. Although the estimation performance reported in previous works have highly increased over time and closely match human-level [HOJ13; Han+15], these models tend to mispredict. This especially holds for predictions under difficult circumstances (e.g. non-frontal pose, one-sided illumination), or when the estimation model faces a sample belonging to a group that was under-represented in the

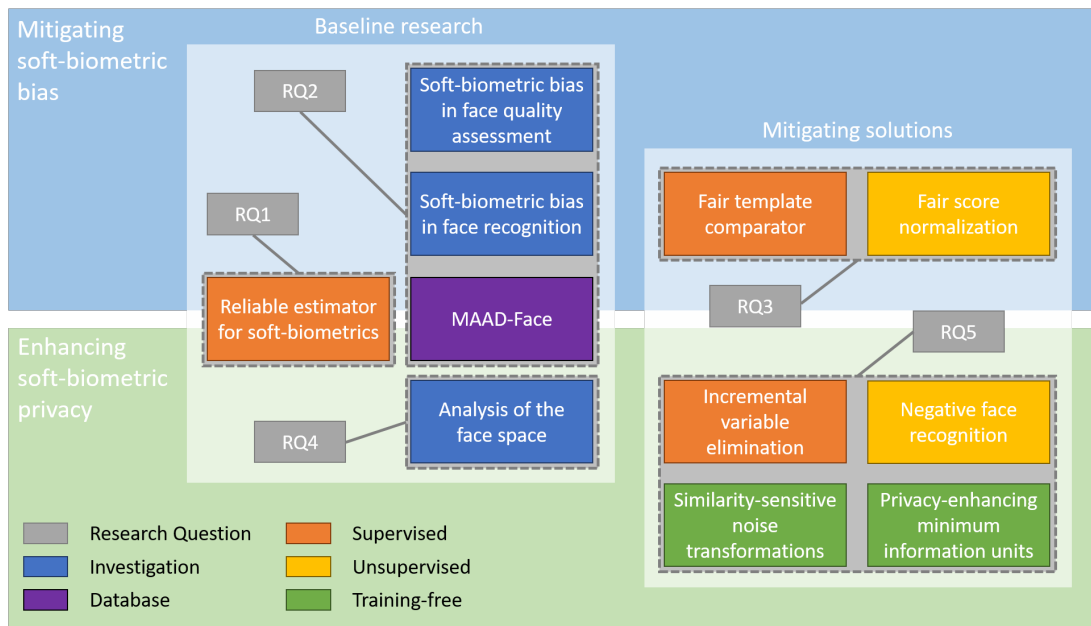


Figure 1.1.: Overview of the key contributions of this work in relation to the corresponding research questions. The green area indicates contributions in the field of soft-biometric privacy while the blue area indicates contributions on the enhancement of soft-biometric privacy.

training data. Current solutions on estimating soft-biometric attributes [DER16], including prediction reliabilities, are based on softmax outputs of the deep learning networks. These outputs are often interpreted as the model’s confidence scores. However, a higher value of such a confidence score does not necessarily imply a higher probability that the classifier is correct as shown in recent works [Guo+17; KL15; NYC14]. In this thesis, soft-biometric driven bias and privacy concerns are analysed based on the prediction reliabilities of soft-biometric attributes to develop efficient mitigation strategies. Consequently, the first research question aims at accurate prediction reliabilities of soft-biometric attribute estimates.

RQ1: How can the prediction confidence (reliability) of a neural network estimator be determined beyond the probabilistic interpretations of the model’s softmax output?

1.1.2. Mitigation of Soft-Biometric Bias

Investigating bias Recent works have shown that commercial, as well as open-source face recognition solutions, show strong differential performances (bias) based on the user’s demographics [GNH19b]. Consequently, several solutions were proposed to mitigate demographic-bias. However, to deploy general non-discriminatory face recognition systems, it is necessary to know the influence of a wide range of soft-biometric attributes on face recognition. This refers to both, face recognition performance as well as the utility estimate of a face image for recognition. Consequently, the second research question aims at investigating the influence of specific soft-biometrics attributes on face recognition to mitigate its discriminatory effects.

RQ2: How do specific soft-biometric attributes affect the behaviour of face recognition systems?

Mitigating bias Driven by the findings that the performances of face recognition systems strongly vary depending on the user’s demographics, previous works proposed solutions to mitigate demographic-bias in face recognition systems. However, these works focused on (a) mitigating demographic-bias based on (b) representation-learning level approaches [Dro+20]. (a) The focus of the mitigation of demographic-bias neglects the discriminatory effects of other soft-biometric attributes on face recognition. (b) The focus of applying representation-learning requires modifying the face recognition model and thus, results in a high workload in real-life applications due to the necessity of a complete replacement of all stored templates. Consequently, the third research question aims at the development of more generalized and integrable bias-mitigating solutions in face recognition.

RQ3: How can soft-biometric bias of various origins in a face recognition system be mitigated without the need for modifying the deployed face recognition model?

1.1.3. Enhancement of Soft-Biometric Privacy

Investigating privacy concerns Despite face representations being trained to enable the recognition of individuals, previous works showed that more information than just the identity is embedded within. They demonstrated that face templates contain information about head pose [Par+17], image characteristics (such as quality [BJ18; Her+19], viewpoint [Hil+18], and illumination [OTo+18]), demographics [DDB18; Ter+19d; ÖAE16], and social traits [Par+19]. However, for many applications, the users do not permit to have access to this information. Thus, the stored data should be exclusively used for recognition purposes [MR17], and extracting such information without a person’s

consent is considered a violation of their privacy [Kin13]. To develop efficient solutions to mitigate soft-biometric privacy concerns, the knowledge about the information encoded in face representations is required. Consequently, the fourth research question aims at investigating what soft-biometric attributes are encoded in face representations.

RQ4: What (soft-biometric) information is stored in biometric face templates?

Mitigating privacy concerns Soft-biometric privacy aims at suppressing or hiding privacy-risk information in face representations to prevent a function creep of encoded information. This is further challenged by simultaneously maintaining a high recognition performance. Previous works mainly tackled this problem by proposing image-level solutions that focus on the suppression of pre-defined (demographic) attributes. However, most biometric representations are stored in templates rather than images [SRB16] and templates offer a less restricted way of encoding information. Moreover, many solutions are limited to the suppression of pre-defined attributes and thus, are vulnerable to unconsidered function creep attacks. Consequently, the fifth research question aims at the development of easily-integrable privacy-enhancement solutions that provide more comprehensive privacy-protection that are not limited to pre-considered attributes.

RQ5: How can soft-biometric privacy be enhanced without the need for modifying the face recognition model?

1.2. Thesis Overview

After motivating and introducing the research focus of this work, an overview of the rest of this thesis is given.

Chapter 2 provides the essential background information to understand the problems and solutions of this work. This includes an introduction to biometric systems and soft-biometrics. A more detailed look is given to face recognition. The historical development of face representations for recognition is discussed, as well as the main components of a face recognition system with a focus on the current deep-learning based face recognition models. Finally, the performance metrics for the evaluation of biometric verification and soft-biometric privacy-preservation are discussed.

Chapter 3 investigates the soft-biometric driven bias and privacy concerns in face recognition. This chapter demonstrates the need for more generalized solutions and provides key findings that guided the development of the proposed solutions. As a response to RQ1, a novel reliability measure [Ter+19d] is proposed to quantify the confidence of the model's prediction. The proposed solution is based on stochastic forward passes through

dropout-reduced neural networks and uses the centrality and dispersion of the network's predictions to derive accurate confidence statements about the model's predictions. Answering RQ2, the influence of soft-biometric attributes on the (biased) behaviour of face recognition systems is analysed. Therefore, the reliability measure from the answer on RQ1 is utilized to create the MAAD-Face database [Ter+20b]. MAAD-Face is a new face annotation database that is characterized by a large number of high-quality attribute annotations. This database is used to demonstrate that the behaviour of a face recognition system is strongly affected by many soft-biometric attributes beyond demographics [Ter+21b; Ter+20e]. The behaviour of a face recognition system refers to both, face recognition performance as well as the utility estimate of a face image for recognition. Lastly, RQ4 is answered in the chapter by investigating what information is stored in face templates [Ter+20a]. The question is answered by investigating the predictability of 113 attributes from face templates at different difficulty-levels with the help of the reliability measure of RQ1. Understandable statements about the stored attribute information are derived by categorizing each attribute into one of three predictability classes demonstrating the need for privacy-enhancing technologies and providing valuable findings for the efficient mitigation of soft-biometric privacy concerns in face recognition.

Answering RQ3, Chapter 4 provides solutions for the efficient mitigation of bias in face recognition. The proposed solutions operate on the comparison- and score-level of the system and thus, can be easily integrated in existing systems. First, a supervised fair template comparator [Ter+20i] is proposed that integrates two notations of fairness at the comparison-level of the system by replacing the deployed similarity function with a fairness-driven similarity model. Second, an unsupervised fair score normalization approach [Ter+20f] is proposed that integrates the notation of individual fairness at the score-level of the system by normalizing the comparison scores of the system to mitigate bias of unknown origins and additionally improving the overall recognition performance.

Chapter 5 aims at answering RQ5 by providing four easily-integrable solutions to enhance soft-biometric privacy of face recognition systems. The proposed solutions either manipulate existing face templates directly or change the type of the identity-representation including its inference for verification. The first type of proposed solutions, the template manipulation approaches, either identifies and eliminates privacy-risk variables from the face templates [Ter+19a] or build on geometric-inspired noise-injections [Ter+19b] to enhance the soft-biometric privacy. The second type of proposed solutions works on the inference-level of a recognition system. In negative face recognition [Ter+20c], the stored (negative) templates contain only information that the individual does not have. For verifying a person's identity, the stored negative template is compared to an ordinary (positive) template and the comparison score is based on a dissimilarity measure. PE-MIU [Ter+20h] is a privacy-enhancing face recognition approach based on minimum informa-

tion units. The approach exploits the structural differences between face recognition and facial attribute estimation by creating templates in a mixed representation of minimal information units. These representations contain the pattern of privacy-sensitive attributes in a highly randomized form. Therefore, the estimation of these attributes becomes hard for function creep attacks. During verification, these units of a probe template are assigned to the units of a reference template by solving an optimal best-matching problem. This allows our approach to maintain a high recognition ability. Unlike previous works, this approach offers a strong and comprehensive privacy-enhancement without the need for training.

Finally, Chapter 6 concludes this work by highlighting its contributions, practical benefits, and key-findings. Moreover, an outlook for future research directions is given.

The contributions of this dissertation are described from the we-perspective, as they are based on published papers.

2. Background

The previous chapter presented a general motivation and an overview of the research problems of this thesis. This chapter provides background information to facilitate a better understanding of the problem and the later proposed solutions.

Section 2.1 gives a general introduction to biometric systems and soft biometrics. This is elaborated with a focus on face biometrics in Section 2.2. The milestones of face representations used for recognition are recaptured to show why all modern state-of-the-art solutions on face recognition involve deep learning. Moreover, the different components of a face recognition system are discussed as well as the core components for their training. This provides the needed information to understand the proposed solutions. Section 2.3 presents evaluation metrics that are commonly used in the literature as well as in this thesis. Finally, we will summarize the core statements of this chapter in Section 2.4.

2.1. Biometrics

Biometrics describes the science of establishing a person's identity based on their behavioural or physical characteristics [JFR10; 12]. It derived from forensic investigations [Rho56] and evolved into several applications scenarios regarding security and convenience. The strong link between identities and individuals is used in security-based applications, such as forensics or border control, or in convenience-based applications, such as automatic log-in and smart home personalization [Dam18].

The goal of biometrics is identity authentication. Traditionally, this is achieved by knowledge proofs of identity (such as passwords or PINs) and/or physical proofs of identity (such as smartcards or keys). However, both proof of identity types may easily be lost, forgotten, or forwarded to someone else. These things become difficult when facing biometric characteristics [JFR10].

Knowledge- or physical-based proof of identity allows a perfect matching to validate a user's identity. For example, in a password-based system, a perfect match between two alphanumeric strings is necessary to validate the claimed identity. Perfect matching usually does not work for a biometric-based proof of identity due to

-
- imperfect sensing conditions (such as various capturing devices and technologies),
 - alterations in the individuals biometric characteristic (such as face ageing),
 - changes in the ambient conditions (such as inconsistent illumination levels),
 - and variations in the user-sensor interaction (such as different head poses).

The observed variability in the set of biometric features of an individual is known as intra-class variations and the variability between the feature sets originating from different individuals is referred to as inter-class variations [JFR10]. A set of biometric features is known as a biometric template of an individual. Usually, these templates are generated by minimizing the intra-class variations and maximizing the inter-class variations.

2.1.1. Biometric Systems

Operation Modes of Biometric Systems

A biometric system is a pattern recognition system operating on acquired biometric data of individuals. Typically, biometric systems operate in three main modes: enrolment, verification, and identification. In all modes, it extracts a feature set (template) from the acquired data of an individual. Depending on the operation mode, it either stores the template in a database as a reference or compares the template against one or more templates that are already stored in the database [JRP04]. Figure 2.1 illustrates the three main operation modes of a biometric system.

During enrolment, a subject is included in the database of the biometric system. The enrolment step includes providing a trusted identity, capturing the biometric characteristics, ensuring high quality of the capture, extracting a distinct template, and storing the templates with the associated identity information in a database [Dam18].

In verification mode, the system validates the claimed identity of an individual by an one-to-one comparison (e.g. "Is the biometric data from Peter?"). The identity can be claimed by utilizing a smart card, a user name, or an identification number. The biometric (probe) data of the individual is captured and used to create a template of the individual. Then, this probe template is compared with the reference template of the claimed identity stored in the database. The comparison results in a continuous comparison score that measures the similarity between the probe and the reference templates. Depending on the comparison score and its decision threshold, the decision of the biometric system might be true or false. The claim is true and the user is genuine, or the claim is false and the user is an imposter. The verification mode is typically used for positive recognition

that aims to prevent multiple individuals using the same identity [JRP04; Way01]. In this work, we will mainly focus on biometric verification.

In identification mode, the system aims to assign an identity to an unknown subject based on its the captured biometrics (e.g. "Whose biometric data is this?") [Dam18]. It aims to recognize an individual by comparing its template against all enrolled templates. Therefore, the system performs a one-to-many comparison [JRP04]. The result of this comparison might be a matched identity or an unidentified user. Identification is critical in negative recognition applications that aims to prevent a single person from using multiple identities.

Properties of Biometric Systems

In ISO/IEC 2382-37, biometric recognition is defined as the automated recognition of individuals based on their behavioural or biological characteristics [12]. These biological properties refer to anatomical and physiological characteristics. Anatomical characteristics refer to the structure of a human body and physiological characteristics refer to their function. Examples for physical or biological traits are face, fingerprint, hand and iris. Behavioural modalities can be represented by keystroke, signature, gait, and voice. However, any human physiological or behavioural characteristic can be used as a biometric characteristic as long as it satisfies the following seven properties [BPJ98; JRP04]:

- *Universality*: a biometric system aims to cover the whole population, which means that every individual should have the biometric characteristic.
- *Uniqueness*: a biometric system aims to represent different individuals distinctively, indicating that no pair of persons should be the same in terms of the characteristic.
- *Permanence*: the characteristic should be time-invariant and thus, the performance of the system.
- *Collectability*: the biometric characteristic can be measured quantitatively.
- *Performance*: a biometric system aims at maximizing its recognition performance and minimizing the computational workload.
- *Acceptability*: a biometric system is convenient for its users and provides high usability.
- *Circumvention*: it is hard to fool the system, e.g. by presenting face biometric samples.

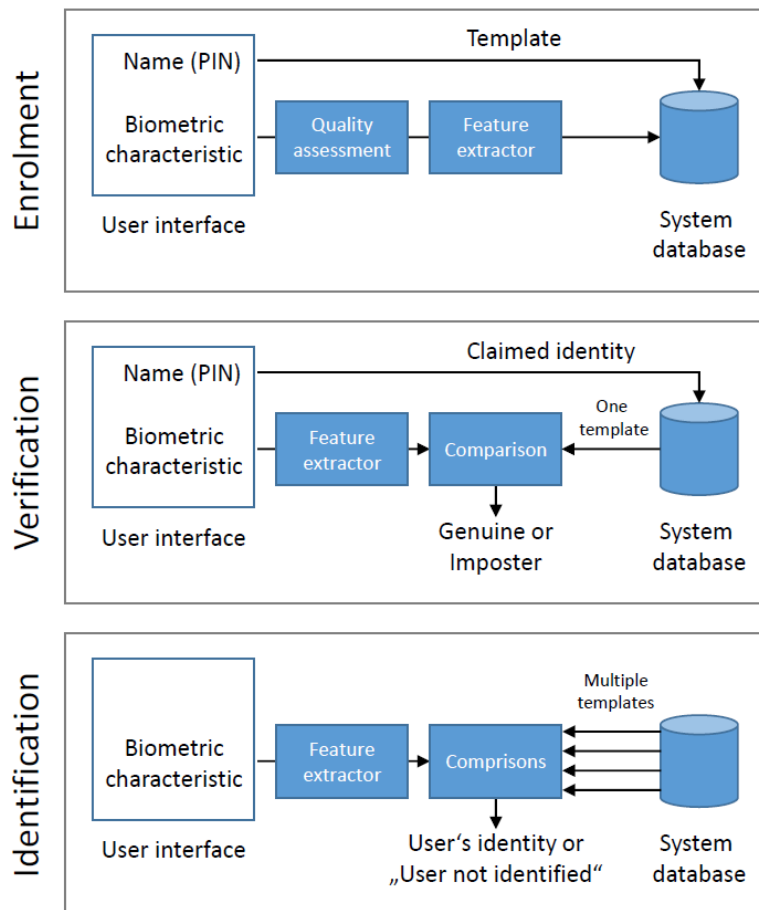


Figure 2.1.: Illustration of the three main operation modes of a biometric system. During the enrolment mode an identity, and its biometric characteristic, is registered in the database. The verification modes verify the claimed identity and in the identification mode, an user's characteristics are compared against multiple stored identities [JRP04; Dam18].

While the first four properties (universality, uniqueness, permanence, and collectability) theoretically define the requirements for a biometric characteristic, the last three (performance, acceptability, and circumvention) describe requirements that should be additionally considered for biometric systems in practice.

These properties are associated differently for different biometric characteristics. This was already discussed in [BPJ98] and is partially presented in Table 2.1. The table reflects the perception of three biometric experts in 1998. For example, at this time, face biometrics is characterized by a high universality (most people have a face), collectability (it can be measured without strong user cooperation), and acceptability (we present our faces on a daily basis). On the other hand, in 1998, it was assigned with medium permanence and low uniqueness, performance, and circumvention. However, this has changed drastically within the last years as we will show in Section 2.2. Current deep-learning based face recognition solutions leverage hierarchical architecture to stitch together pixels into invariant face representations [WD18]. These methods are able to find highly distinctive pixel-patterns in faces and utilize these to produce high-performing face templates. Moreover, deep-learning approaches are also used to enhance the circumvention of face recognition systems, such as for presentation attack detection [RB17]. These recent advances have made face recognition one of the most studied biometric modalities.

Table 2.1.: Comparison of biometric characteristics (H: high, M: medium, L: low) [BPJ98] for different modalities. The data reflects the perception of three biometric experts in 1998.

	Face	Fingerprint	Hand	Keystrokes	Hand Vein	Iris	Retinal Scan	Signature	Voice	DNA	Gait	Ear
Universality	H	M	M	L	M	H	H	L	M	H	M	M
Uniqueness	L	H	M	L	M	H	H	L	L	H	L	M
Permanence	M	H	M	L	M	H	M	L	L	H	L	H
Collectability	H	M	H	M	M	M	L	H	M	L	H	M
Performance	L	H	M	L	M	H	H	L	L	H	L	M
Acceptability	H	M	M	M	M	L	L	H	H	L	H	H
Circumvention	L	H	M	M	H	H	H	L	L	L	M	M

2.1.2. Soft-Biometrics

Biometric data is usually used to recognize individuals. However, it is also possible to deduce the attributes of an individual from the same data. For instance, gender, age, ethnicity, hair color, eye color, height, and weight [WG13] can be deduced from data that was collected for the purpose of biometric recognition.

While these attributes are not necessarily unique to an individual, they can be utilized in a variety of applications, such as surveillance, forensics, and biometric data indexing. Moreover, they can be used in combination with a primary biometric modality to improve recognition performance. This probably led to these attributes being called soft-biometrics [JDN04a; JDN04b; Nix+15].

Formally, soft-biometrics can be defined as follows. *Soft-biometric traits are physical, behavioural, or material accessories, which are associated with an individual, and which can be useful for recognizing an individual. These attributes are typically gleaned from primary biometric data, are classifiable in pre-defined human-understandable categories, and can be extracted in an automated manner* [Dan+11].

Soft-biometrics offer several benefits depending on the use-case. First, they allow generating qualitative descriptions of an individual (e.g. young Asian male with blue eyes and blond hair). This formulation can be easily understood by humans and therefore, this bridges the semantic gap between human and machine descriptions of biometric data. Second, some soft-biometrics, such as gender and ethnicity, can also be deduced from low-quality data. This allows them to be used in a wider range of applications. Lastly, they often can be collected easily since they require less or no cooperation of the observed individual [DER16].

However, this also has serious consequences on the user's privacy. For most biometric systems, the stored data of an individual should be exclusively used for recognition purposes. However, biometric data includes more information than needed for recognition. Moreover, much of this information can also be deduced from biometric templates as we will show in Section 3.5. Therefore, it is necessary to ensure that the stored biometric templates are not used for function creep. This led to the development of soft-biometric privacy-enhancing solutions that aim to suppress privacy-sensitive information (such as gender, ethnicity, health conditions) from biometric templates. In Section 5, we will discuss this topic in more details and provide several solutions to enhance soft-biometric privacy in face recognition.

2.2. Face Recognition

Face recognition is one of the most important topics in computer vision and pattern recognition [WD18]. Among various biometric modalities, the face is one of the most widely used, because it is ubiquitous and acquirable in unconstrained environments. Moreover, it provides a strong and discriminative recognition performance [Mas+18] and has been widely used in many areas, such as finance, public security, forensics, and daily life [WD18].

Face recognition is inherently challenged by large intra-class variations due to the huge facial variability in age [Ort+09], pose [Has+15], illumination [GB03], and expression [LMZ06] (APIE). A big step towards solving the APIE problem in face recognition was done by training deep convolutional neural networks on massive datasets. In 2014, DeepFace [Tai+14] achieved state-of-the-art performance on the LFW benchmark [Hua+07] and demonstrated an unconstrained face recognition performance that for the first time closely matches human-level (DeepFace 97.35% accuracy vs. Humans 97.53% accuracy). In 2017, automated deep face recognition systems already scored above the median of super-recognizers and forensic facial examiners [Phi+18]. Till today, the performance and generalizability of face recognition systems are still improving [Wan+18b; Den+19].

2.2.1. Milestones of Face Representations for Recognition

In this section, we will provide an overview of the key works on face representation for facial recognition. An overview of the milestone is presented in Figure 2.2.

In 1991, Turk and Pentland [TP91] proposed the Eigenface approach that started an era of research on automated face recognition. The early solutions involve holistic approaches that derive low-dimensional representations through specific assumptions of the underlying data distribution [WD18]. These resulted in solution based on linear subspaces [BHK97; MWP98], manifolds [He+05; Yan+07a; Yan+07b], and sparse representations [DHG12; DHG18; Wri+09; ZYF11]. However, these holistic approaches are based on their prior assumptions and thus, fail to address uncontrolled facial variations.

Therefore, face recognition solutions based on local-features are proposed in the early 2000s [WD18]. These approaches include Gabor filters [LW02], local binary patterns (LBP) [AHP06], as well as their multi-level and high-dimensional extensions [Che+13; DHG19; Zha+05]. This results in robust performances due to the invariance properties of local filtering. However, templates based on handcrafted features does not provide the required distinctiveness and compactness [WD18] needed for reliable face recognition.

In the 2010s, learning-based local descriptors were introduced for face recognition [Cao+10; LPL14; Cha+15]. Although these shallow representations follow a learning-

based strategy to enhance the distinctiveness and compactness, they can still not capture complex non-linear facial appearance variations [WD18]. Moreover, no integrated solutions were proposed that jointly addresses the problems of unconstrained face recognition, such as lighting, pose, or expression. As a result, these approaches are not able to extract identity-stable features in real-world scenarios.

However, in recent years, this changed drastically as deeply-learned features for face recognition were introduced. In 2014, DeepFace [Tai+14], a 9-layer convolutional neural network model, was proposed. It achieved state-of-the-art performance on the LFW benchmark [Hua+07] and, for the first time, demonstrated a human-level performance for unconstrained face recognition. From this point, face recognition research focused on deep learning approaches and dramatically improved the performance. Exploiting the strength of deep convolutional neural networks and large face image datasets, these models were trained in an end-to-end fashion to produce face representations that contain strong identity signals and provide significantly stronger robustness to APPE variations. In 2017, automated deep face recognition systems already outperformed forensic facial examiners [Phi+18] and in the following years, the general recognition power strongly improved [Den+19].

2.2.2. Components of a Face Recognition System

A typical face recognition system consists of three modules: a preprocessing module P , a template extraction module T , and a matching module M . In Figure 2.3, the locations of these modules are shown in a face verification pipeline.

Preprocessing The preprocessing module P gets an image I as input and aims to process the image such that facial features can be reliably estimated. Therefore, it consists of a face detector that is used to localize faces in the input image. If no face or multiple ones are detected, it may ask the user for another input.

Depending on the application of the face recognition system it may also consist of sub-modules for face quality assessment (FQA) and presentation attack detection (PAD). The FQA sub-module measures the utility of the input face for recognition. This aims to ensure that only faces of high utility are enrolled or used for verification and thus, it aims to reduce future recognition errors. The PAD sub-module recognizes if the captured face is live or spoofed to avoid wrong decisions by a different type of presentation attacks.

Although face recognition shares similarities with generic object recognition, faces have a well-structured shape and thus, can be better modelled than generic objects [Mas+18]. Consequently, strong domain knowledge can be utilized to ease the face representation

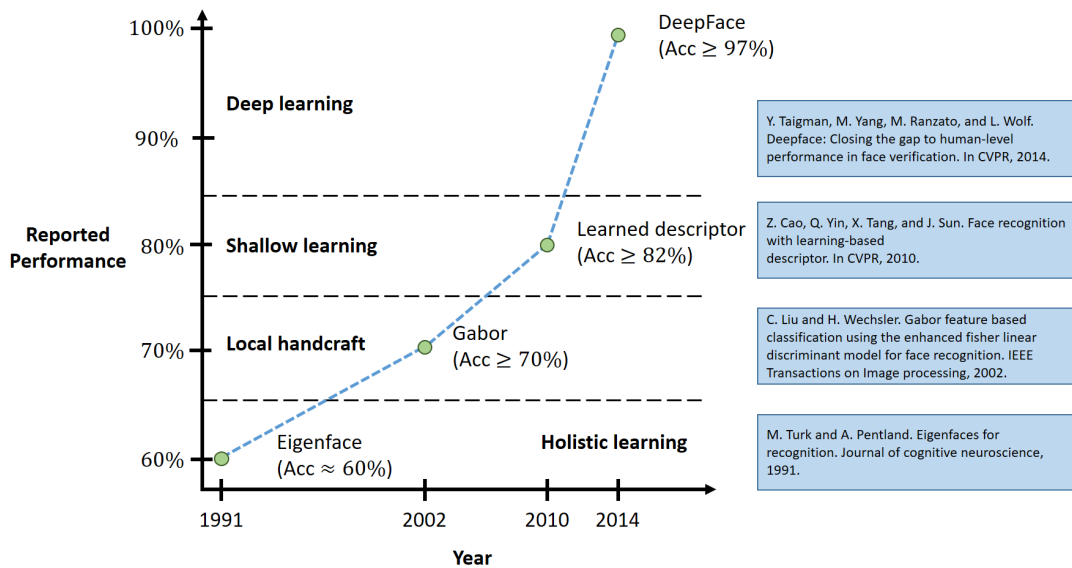


Figure 2.2.: Milestones of face representation for recognition [WD18]. In the 1990s, face recognition was based on holistic approaches. In the 2000s, handcrafted local descriptors dominated the face recognition research, followed by local feature learning. In 2014, DeepFace [Tai+14] achieved state-of-the-art and human-like performance, shifting the research focus on deep learning.

learning. Therefore, the face image is scaled, rotated, translated, and cropped to ensure a consistent alignment between all faces. This significantly simplifies the process of learning and extracting distinctive facial features. The preprocessed image $P(I)$ is then passed to the template extractor.

Template extraction The template extraction module T gets as an input the preprocessed face image $P(I)$, extracts facial features from $P(I)$, and outputs a corresponding face template $x = T(P(I))$. With the era of deep learning, these templates (or embeddings) are created with deep convolutional neural networks. Detailed information on the training, model architectures, and working principles of these face recognition models are described in Section 2.2.3. These models aim to extract identity related information of an individual that is used for recognition. However, as we will show in Section 3.5, these templates also encode privacy-sensitive information that is not necessary for recognition such as gender, age, ethnicity, or accessories.

Face matching In the matching module M , two faces are compared to determine if they belong to the same identity or not. Therefore, the template of the preprocessed face image of the probe $x_{probe} = T(P(I_{probe}))$ is compared against an associated reference template $x_{ref} = T(P(I_{ref}))$ stored in the database. The comparison of both templates is done with a similarity function

$$sim(T(P(I_{probe})), x_{ref}) = s, \quad (2.1)$$

and results in a comparison score s . Usually, the similarity function uses cosine similarity or an (inverse) euclidean distance. Applying a threshold on the comparison score s results in a genuine or imposter decision

$$D = M(T(P(I_{probe})), x_{ref}), \quad (2.2)$$

whether the images belong to the same identity or not.

Please note that (a) the similarity function can also be learned and (b) the comparison score s can also be normalized. (a) refers to metric learning approaches that aim to learn such a similarity function with specific properties. In Section 4.3, we will demonstrate that by proposing a metric learning approach to mitigate ethnic-bias. (b) refers to score normalization approaches that are usually used when combining multiple biometric traits [Agg+08; Dam18]. However, in Section 4.4, we propose fair score normalization and demonstrate that this can be adapted to be an effective bias-mitigation tool even for single-trait biometrics.

2.2.3. Deep Face Recognition Models

Face recognition can be considered as a zero-shot learning task since, for most applications, it is not possible to include candidate faces during training. Therefore, most works perform transfer learning meaning that the network training is based on a closed pool of subjects and is then used as a feature extractor on unseen faces. Despite that difficulty, a high generalization is possible since human faces share a similar shape and texture [WD18]. Generally, deep face recognition solutions mainly differ on three aspects:

- the utilized network architecture that is trained for the task of recognizing faces,
- the loss function that guides the network training,
- and the utilized training data that reflects the inter- and intra-subject variations and thus, builds the fundamentals of the training stage.

In the following, we will discuss each aspect.

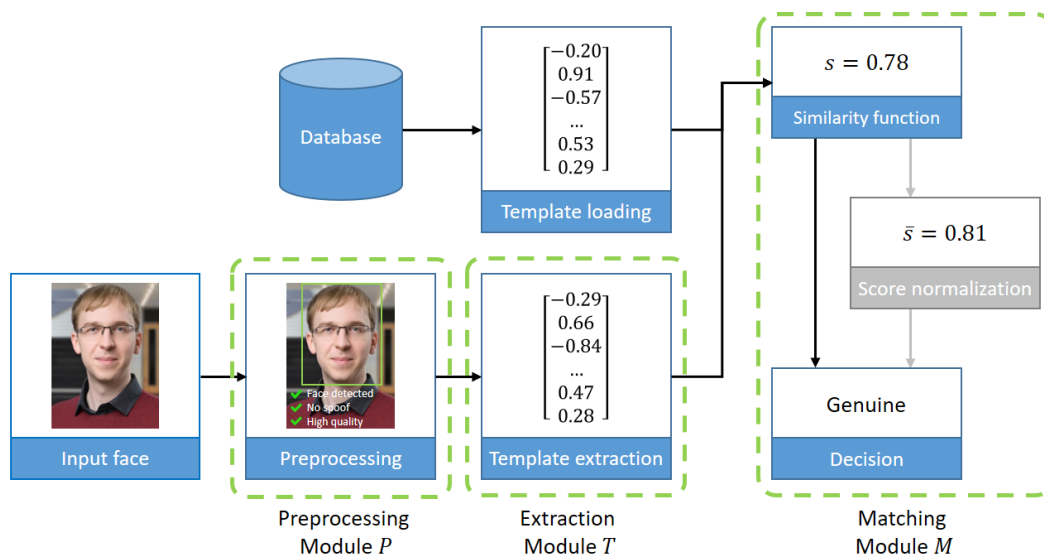


Figure 2.3.: Illustration of a face verification pipeline including the preprocessing, extraction, and matching modules. The score normalization in the matching module M is optional and, for instance, can be used to mitigate bias as proposed in Section 4.4.

Architectures and Databases

The network architectures for deep face recognition usually followed the architecture used in object detection [WD18]. Consequently, often used architectures are AlexNet [San+16; SKP15], VGGNet [Mas+16; PVZ15], and ResNet [Zha+17; Liu+17]. One big trend is the use of deeper networks. However, to enable face recognition on embedded devices, the other trend is to minimize the model size while maintaining as much of its recognition performance [Wu+18; Ge+19].

With the use of deep-learning technologies, a key aspect in developing face recognition systems is the available training data. Although some companies have private face datasets that contain millions of face images (Facebook [Tai+14]) or millions of subjects (Google [SKP15]), the size of publicly available databases is on a significantly lower scale [Mas+18]. Typical datasets for training face recognition model are CASIA-WebFace [Yi+14], VGGFace [PVZ15], VGGFace2 [Cao+18], and MS-Celeb-1M [Guo+16]. CASIA-WebFace [Yi+14] contains around 500K images from 10K subjects. It was automatically collected by looking

at celebrities. The MS-Celeb-1M [Guo+16] dataset contains 10M images from 100k celebrities. It was collected by searching celebrity names in the Bing search engine and retrieving the first 100 images. Since the collection of MS-Celeb-1M was without any filtering, the dataset is strongly biased by label noise, duplicated images, and non-face images [Mas+18]. Consequently, it is hard to use directly. VGGFace [PVZ15] comprises around 2.6M faces of 2.6K individuals. The face images are mostly frontal and of high quality. Later, the improved version VGGFace2 [Cao+18] was proposed. This dataset contains 3.3M images of 9k subjects and additionally covers variations of pose, age, and ethnicity.

Training data form the basis for face recognition performance. However, most databases only cover a partial distribution of face data. Most large-scale datasets are often collected online and consist of celebrities on formal occasions. Therefore, these images are highly different than face images from daily life, surveillance, or security applications. Moreover, demographic cohorts, such as gender, age, and ethnicity, are usually unevenly distributed in these datasets [WD18]. This can lead to significant performance differences based on the individual’s demographics. In Section 3.3, we demonstrate this problem and discuss it in more details. Section 4 describes our proposed solutions for this problem.

Loss functions

The utilized loss function plays a major role in the performance of a face recognition model. It guides the neural network training to extract discriminative facial features. There are basically two ways of training deep face recognition neural networks. In the first case, a multi-class classifier is trained to differentiate between training identities [Tai+14; Wen+16; Liu+17; Wan+18b; Den+19], such as utilizing a softmax loss. In the other case, face templates are directly learned, such as with triplet loss [SKP15; PVZ15].

Triplet Loss Solutions Solutions trained with triplet loss [SKP15; PVZ15] make use of face triplets that consists of an anchor face image x^a , an (positive) image of the same identity x^p , and an (negative) image of a different identity x^n . Triplet loss aims at learning face representations such that the euclidean distance between the anchor template and the template of the positive sample is always smaller than the distance between the anchor and the negative template (including a small margin $\alpha > 0$).

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2 \quad (2.3)$$

Since the distance calibration takes place at the template-level, $f(\cdot)$ defines the network function that maps the input image to the corresponding face template. This leads to the

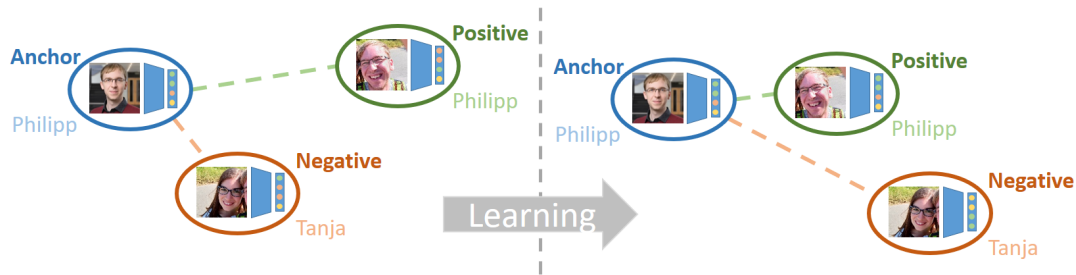


Figure 2.4.: Illustration of the triplet loss learning principle. The distance between anchor and positive sample is reduced, while the the distance between anchor and the negative samples is enlarged.

following loss function

$$\mathcal{L}_{Triplet} = \frac{1}{N} \sum_i \max \{0, \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\}, \quad (2.4)$$

because $f(\cdot)$ only has to be modified for triplets that do not satisfied Equation 2.3. Triplet loss guides a neural network to minimized intra-subject variations as well as maximize the separation between different identities. This can be seen with networks such as FaceNet [SKP15] and VGGFace [PVZ15]. However, this training procedure is not suitable on large datasets since the number of possible triplet pairs grows exponentially and thus, the selection of suitable (semi-hard) triplets becomes difficult.

Softmax Loss Approaches Softmax-based approaches aim at classifying on a closed-set of identities during training and utilizes a previous layer as a feature extractor for unseen faces. The traditional softmax loss

$$\mathcal{L}_{Softmax} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^N e^{W_j^T x_i + b_j}} \right) \quad (2.5)$$

combines a softmax activation on the classification layer with a standard cross-entropy loss. Here, $x_i \in \mathbb{R}^d$ refers to the template of the i^{th} of N training samples that belong to subject y_i . $W_j \in \mathbb{R}^d$ denotes the j^{th} column of the weight matrix $W \in \mathbb{R}^{d \times n}$ with n equals the number of training identities. Moreover, $b_j \in \mathbb{R}^n$ defines the bias term. Early approaches that use this loss, such as DeepFace [Tai+14], doing well in separating training subjects, but do not explicitly minimize the intra-subject variations.

Center loss [Wen+16] tackled this issue by minimizing intra-subjects distances between samples x_i and their corresponding class-centroids c_{y_i} that determines the class center of the deep features. This results in the center loss

$$\mathcal{L}_{Centerloss} = \mathcal{L}_{Softmax} + \frac{\lambda}{2} \mathcal{L}_{Center}, \quad (2.6)$$

with

$$\mathcal{L}_{Center} = \frac{1}{2} \sum_{i=1}^N \|x_i - c_{y_i}\|_2^2, \quad (2.7)$$

and λ to balance between the two losses.

Other approaches are directly based on the softmax loss from Equation 2.5. For simplicity, the bias terms can be fixed to $b_j = 0$ and the individual weights can be normalized $\|W_j\| = 1$ [Wan+18b; Liu+17]. Also the embedding $\|x_i\|$ can be rescaled to $\|x_i\| = r$. This allows to transform the statement $W_j^T x_i + b_i$ to

$$W_j^T x_i + b_i \stackrel{b_i=0}{=} \|W_j\| \|x_i\| \cos(\theta_j) \stackrel{\|W_j\|=1}{=} r \cos(\theta_j), \quad (2.8)$$

where θ_j is the angle between weight W_j and the feature vector x_i . This makes the prediction only dependent on this angle and thus, the embeddings are distributed on a hypersphere with radius r . These modifications lead to the SphereFace loss [Liu+17]

$$\mathcal{L}_{Sphereface} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{e^{r \cos(\theta_{y_i})}}{e^{r \cos(\theta_{y_i})} + \sum_{j=1, j \neq y_i}^N e^{r \cos(\theta_{y_i})}} \right). \quad (2.9)$$

The SphereFace loss introduces the idea of an angular margin and thus, aims to learn angularly discriminative features. Adding a cosine margin penalty to Equation 2.9 leads to the loss function of CosFace [Wan+18b]

$$\mathcal{L}_{CosFace} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{e^{r \cos(\theta_{y_i}) - m}}{e^{r \cos(\theta_{y_i}) - m} + \sum_{j=1, j \neq y_i}^N e^{r \cos(\theta_{y_i})}} \right), \quad (2.10)$$

which achieves a higher generalization due to the added margin principle and thus, a higher performance. By shifting the margin penalty to the angular-level, the loss function of ArcFace [Den+19]

$$\mathcal{L}_{ArcFace} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{e^{r \cos(\theta_{y_i} + m)}}{e^{r \cos(\theta_{y_i} + m)} + \sum_{j=1, j \neq y_i}^N e^{r \cos(\theta_{y_i})}} \right), \quad (2.11)$$

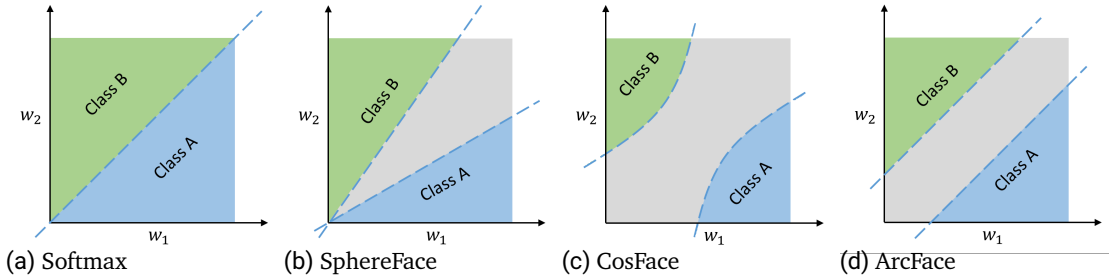


Figure 2.5.: Decision boundaries of different loss functions under a binary classification case [Den+19]. The dashed lines represent the decision boundaries, while the gray areas denote the decision margins.

is constructed. As the representations are distributed around each representation center on the hypersphere of radius r , adding this additive angular margin penalty simultaneously improves the inter-subject separability and the intra-subject compactness. This enhances the distinctiveness of the obtained features as well as stabilises the training process [Den+19].

These small differences between the loss functions still have a strong influence on the achieved decision boundary as visualized in Figure 2.5. Softmax loss (Equation 2.5) creates a linear decision boundary without a margin. SphereFace loss (Equation 2.9) and CosFace loss (Equation 2.10) create a non-linear margin between the decision boundaries. ArcFace loss (Equation 2.11) has a constant linear angular margin. This strongly affects the recognition performance as shown in Table 2.2 on the LFW benchmark [Hua+07].

2.3. Performance Metrics

This section provides performance metrics that are usually used in the literature as well as in this thesis. Section 2.3.1 derives biometrics verification performance measurements that are also recommended in the international standard ISO/IEC 19795-1 [06]. Moreover, two metrics are presented to evaluate subgroup-specific (biased) performance differences. Section 2.3.2 provides the tools needed to investigate privacy-enhancing technologies. This includes metrics to measure the success of function creep attacks as well as a proposed metric to measure how beneficial it is to apply a certain privacy-enhancing technology.

Table 2.2.: Overview of some deep face recognition approaches. The reported performance refers to the accuracy on the LFW [Hua+07] benchmark. The training face data contain information about (the number of images/the number of identities/availability).

Year	Name	Loss Type	Architecture	Training Data	Accuracy
2014	DeepFace [Tai+14]	Softmax	Alexnet	Facebook (4.4M/4K/private)	97.35%
2015	FaceNet [SKP15]	Triplet	GoogleNet-24	Google (500M/10M/private)	99.63%
2015	VGGFace [PVZ15]	Triplet	VGGNet-16	VGGFace (2.6M/2.6K/public)	98.95%
2016	CenterFace [Wen+16]	Softmax	LeNet-7	Multiple DB (0.7M/17K/public)	99.28%
2017	SphereFace [Liu+17]	Softmax	ResNet-64	CASIA-WebFace (0.49M/10K/public)	99.42%
2018	CosFace [Wan+18b]	Angular Margin	ResNet-64	CASIA-WebFace (0.49M/10K/public)	99.33%
2019	ArcFace [Den+19]	Angular Margin	ResNet-100	MS-Celeb-1M (3.8M/85K/public)	99.83%

2.3.1. Evaluating Verification Performance

As mentioned Section 2.1, biometric verification belongs to non-perfect matching. Two samples of the same biometric characteristic of the same identity are not exactly the same due to (a) imperfect sensing conditions, (b) changes in the individual's physiological or behavioural characteristics, (c) alternations of the ambient conditions, or (d) variations in the user-sensor interaction. Consequently, the respond of a biometric recognition system is a comparison score s that quantifies the template similarity of both samples [JRP04]. Typically, a high score refers to a higher certainty that the samples belong to the same individual.

Depending on the system's decision threshold t , a comparison score below t refers to an imposter pair (samples belong to different persons) and a comparison score equal or above t refers to genuine pair (samples belong to the same person). The score distributions of sample pair from the same and from different persons are called genuine and imposter distributions. Figure 2.6a illustrates such a score distribution.

A biometric verification system can make two types of recognition errors:

1. It can mistake the biometric templates from two different individuals to be from the

same one, a false match,

2. or, it can mistake the biometric templates from the same person to be from two different persons, a false non-match.

In this work, we report the verification performance of a biometric system in terms of false non-match rate (FNMR) at fixed false match rates (FMR). Both verification performance measures are defined in the ISO standard [06]. The choice of the system threshold t defines the trade-off between these two errors. Both errors, false match and false non-match, are also often termed as false accept and false reject. Thus, the performance measure of FNMR and FMR is equivalent to false rejection rate (FRR) and false acceptance rate (FAR) [JRP04]. The FNMR is usually reported at a fixed FMR. The European Border Guard Agency Frontex [Fro17] recommends the use of a decision threshold t such that the FMR $\leq 10^{-3}$. Another widely-used verification metric is the equal error rate (EER). The EER is well known as a single-value indicator of the verification performance and equals the FMR at the decision threshold t where FMR and FNMR are the same.

Mathematically, this can be defined as a hypothesis testing formulation. If the biometric template of a captured individual is denoted as x_{probe} and the template of the claimed identity stored in the database is denoted as x_{ref} , the null and alternative hypotheses are [JRP04]

- H_{gen} : the templates x_{probe} and x_{ref} belong to the same individual
- H_{imp} : the templates x_{probe} and x_{ref} belong to different individuals.

Consequently, the decisions are

- D_{gen} : the claimed identity is correct (gen)
- D_{imp} : the claimed identity is not correct (imposter).

To come to one of these decisions, a decision rule

$$D = \begin{cases} D_{gen} & \text{if } s(x_{probe}, x_{ref}) \geq t \\ D_{imp} & \text{if } s(x_{probe}, x_{ref}) < t \end{cases} \quad (2.12)$$

is applied, which is dependent on the comparison score $s(x_{probe}, x_{ref})$ of both templates and the system's decision threshold t . The hypothesis testing formulation inherently contains both mentioned errors. The type 1 error describes a false match (H_{imp} is true, but the decision is D_{gen}). The type 2 error describes a false non-match (H_{gen} is true, but

the decision is D_{imp}). The probability of both errors can be described as their conditional probabilities

$$FMR = p(D_{gen}|H_{imp}) \quad (2.13)$$

$$FNMR = p(D_{imp}|H_{gen}). \quad (2.14)$$

To evaluate the performance of a deployed biometric system, one must collect a set of comparison scores of genuine and imposter comparisons coming from the system. The verification performance of a biometric system

$$FMR = \int_t^\infty p(s(x_{probe}, x_{ref})|H_{imp})ds \quad (2.15)$$

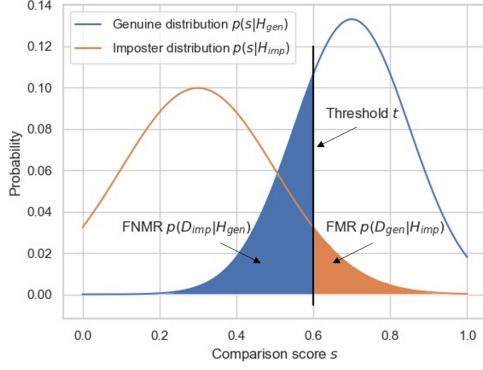
$$FNMR = \int_{-\text{inf}}^t p(s(x_{probe}, x_{ref})|H_{gen})ds, \quad (2.16)$$

can then be defined over the integrals of the genuine score distribution $p(s(x_{probe}, x_{ref})|H_{gen})$ and the score distribution $p(s(x_{probe}, x_{ref})|H_{imp})$ of the imposter. This is further visualized in Figure 2.6a.

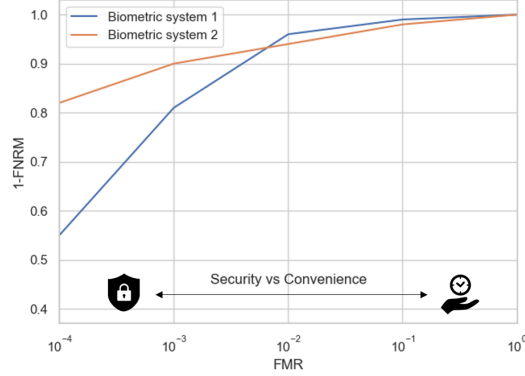
To investigate the system's recognition performance at all decision thresholds, a receiver operating characteristic (ROC) curve is an useful tool. An ROC curve plots the FMR (x-axis) over 1-FNMR (y-axis) as shown in Figure 2.6b. Therefore, the performance of several biometric systems can be compared for different applications. For instance, high security application, such as border control, require a low FMR. In Figure 2.6b, the biometric system 1 would be better suitable for such an application than the biometric system 2. On the other hand, for applications that focus on convenience, such as automatic log-in, the FNMR should be low and thus, usually higher FMR are taken into account. This kind of applications would prefer the biometric system 2 in Figure 2.6b.

Measuring verification bias Biometric systems often possess an unintended biased in form of different performances depending on e.g. the user's demographics. To facilitate a clear discussion on bias in biometric recognition systems, Howard et al. [JRP04] defined the terms of differential performance and differential outcome. Differential performance refers to the difference in the genuine or imposter distribution between specific demographic groups independent of any decision threshold. Differential outcome refers to the differences in the FMR or FNMR between different demographic groups relative to a predefined decision threshold. Despite these definitions, the community often refers to this as recognition bias. In this work, if we will focus on differential outcome.

To measure the performance differences of different ethnic subgroups in terms of the differential outcome, the variation of the group-specific recognition performances have



(a) Score distributions



(b) ROC curves

Figure 2.6.: Evaluating biometric verification performances. On the left, score distributions of genuine and imposter comparisons are shown. ROC curves of two biometric systems are shown on the right.

to be evaluated. This can be done with any measure of statistical dispersion, such as the mean absolute deviation (MAD) or the standard deviation (STD). Given the FNMRs of demographic groups \mathcal{G} at a fixed FMR, each group specific performance (recognition error) is denoted as RE_g for $g \in \mathcal{G}$. The performance differences for the different demographic groups can be evaluated e.g. using STD

$$STD(RE_{g \in \mathcal{G}}) = \sqrt{E[RE^2] - (E[RE])^2}, \quad (2.17)$$

or using MAD

$$MAD(RE_{g \in \mathcal{G}}) = E[|RE - E[RE]|], \quad (2.18)$$

where $E[\cdot]$ refers to the mean operation. Both measures describe a statistical dispersion of the group-specific performance differences. MAD focuses more on the majority of groups, while STD is more outlier-sensitive. Low MAD/STD values indicate that the recognition performances between all groups are similar and thus, less biased. High MAD/STD values indicate strong performance differences between the different groups. In this case, the recognition system is strongly biased in terms of differential outcome.

2.3.2. Evaluating Soft-Biometric Privacy-Preservation

Enhancing soft-biometric privacy describes a trade-off between the desired degradation of the attribute estimation performance by function creep attackers and the desired preservation of the recognition ability. Evaluation metrics for measuring the recognition performance was already introduced in Section 2.3.1. In this section, we will first introduce attribute estimation metrics that are needed for the evaluation of the function creep attacks. Next, we will introduce a metric to jointly investigate the recognition-preservation as well as attribute suppression.

Evaluating function creep attacks In the scenario of soft-biometric privacy, function creep attackers aim to predict privacy-sensitive attributes from biometric templates. These predictions come from the estimation models that are trained with the target labels on the same kind of templates. A simple tool to measure the effectiveness of such an attack is the standard accuracy. However, since training and test data are often highly unbalanced regarding its attribute labels, the balanced accuracy is more suitable. This balanced accuracy is equivalent to the standard accuracy definition with class-balanced sample weights.

In order to evaluate the attribute suppression performance of a privacy-enhancing approach, the suppression rate can be used. The suppression rate

$$sr = \frac{acc_{org} - acc_{mod}}{acc_{org}} \quad (2.19)$$

describes the reduction of the attribute-prediction accuracy of the unmodified (original) templates acc_{org} in comparison to the accuracy of the templates acc_{mod} with privacy-enhancement. A higher suppression rate indicates an advanced privacy-improvement.

Evaluating the soft-biometric privacy trade-off Solutions on soft-biometric privacy aims at degrading the attribute prediction performance of function creep classifiers while preserving its utility for recognition. To evaluate if a privacy-enhancing methodology is beneficial, we propose the privacy-gain identity-loss coefficient (PIC) [Ter+19b]. The PIC

$$PIC = \frac{AE' - AE}{AE} - \frac{RE' - RE}{RE}, \quad (2.20)$$

is defined by attribute prediction errors AE' and AE and the verification errors RE' and RE with and without the privacy-preserving methodology. Positive values indicate that the privacy gain is higher than the loss in the identity preservation performance. Since it

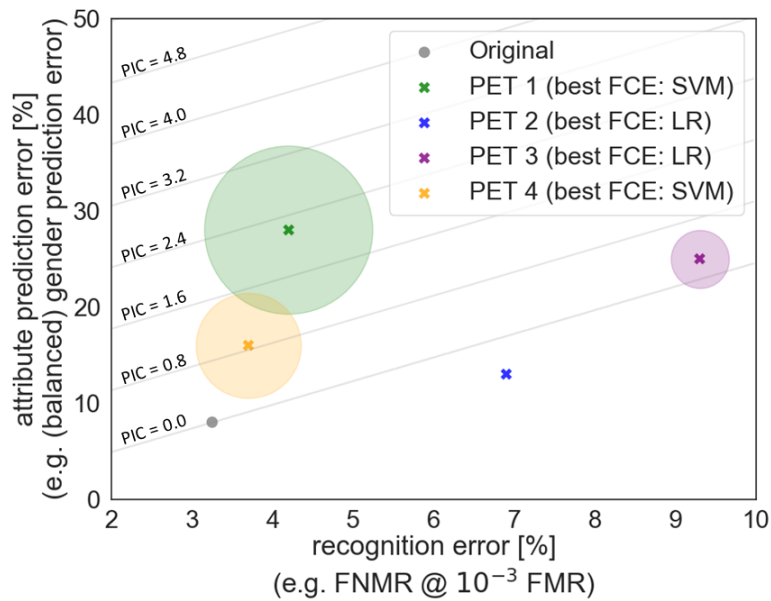


Figure 2.7.: An illustration of an attribute-recognition plot. The attribute prediction error is shown over the recognition error for the unmodified baseline and the different PETs. The attribute error refers to the most successful function creep classifier. The size of the shaded area around a PET refers to its PIC. Additionally, equipotential lines for different PIC-values are shown in grey.

measures how beneficial it is to apply the privacy transformation, a higher PIC indicates a better trade-off and thus, a more beneficial privacy-enhancing technique.

To fully investigate the soft-biometric privacy trade-off, attribute-recognition plots are an useful tool [Bor+20]. An example plot is shown in Figure 2.7. It plots a recognition error (x-axis) over the attribute prediction error (y-axis). In the plot, the unmodified baseline (Original) and different privacy-enhancing technologies (PETs) are represented. The recognition error can be arbitrarily chosen. The attribute prediction error refers to the prediction error of the most successful function creep estimator (FCE) that aims to predict this attribute given the modified template. Each point represents a PET and the area around represents its PIC value. This way it allows to jointly investigate the most important factors for soft-biometric privacy enhancing technologies.

2.4. Summary

This section discussed the background information that is needed to understand the research topic of this work and its proposed solutions.

First, a general introduction to biometrics was given. This included the special properties of a biometric system as well as soft-biometrics. Second, face biometrics was introduced in details. It showed (a) the main components of a biometric system, (b) the key points to train a deep face recognition model, and (c) why current state-of-the-art face recognition approaches are solely based on deep learning. Third, performance metrics are described to measure bias and privacy concerns in face recognition system. These concerns arise from the use of deep learning techniques, as we will show in Section 3.

3. Investigation of Soft-Biometric Driven Bias and Privacy Concerns

3.1. Introduction

This chapter investigates soft-biometric driven bias and privacy concerns in face recognition. The goal of the investigations is to analyse the vulnerabilities of these concerns. The findings of this chapter are used afterwards in Chapter 4 and 5 to develop effective mitigation solutions.

To enable such in-depth investigations, in Section 3.2 two preliminary works are proposed. The first preliminary work is a novel reliability measure [Ter+19d] that is able to accurately quantify the confidence of a model's prediction. This answers RQ1 and is used for the definition of predictability classes in Section 3.5 to answer RQ4. Moreover, it is used in the second preliminary work. The second proposed preliminary work is MAAD-Face [Ter+20b], a novel face database that is characterized by its large number of soft-biometric attribute annotations. Using the novel reliability measure from RQ1 ensures a correctness of these annotations. The MAAD-Face annotations database proposes the large amount of high-quality attribute annotations needed for the investigation of soft-biometric driven performance differences (bias) in face recognition systems (Section 3.3).

The investigations of soft-biometric driven bias and privacy concerns in face recognition aim to answer RQ2 and RQ4. In Section 3.3, the influence of soft-biometric attributes on the performance of face recognition systems is analysed. In Section 3.4, the influence of soft-biometric attributes on face quality assessment is investigated [Ter+20e]. Together, both investigations analyse of the effect of soft-biometric bias on the behaviour of face recognition systems answering RQ2. Finally, Section 3.5 investigates which soft-biometric attributes are stored in biometric face templates [Ter+20a] to answer RQ4.

3.2. Preliminary Investigations

This section proposes two preliminary works that are needed for the investigations on the soft-biometric driven bias and privacy concerns. In Section 3.2.1, a novel reliability measure [Ter+19d] is presented and analysed on the example of age and gender estimation. This reliability measure is needed in Section 3.5 to answer RQ4 and to provide reliable attribute annotations needed in Section 3.2.2 for answering RQ2. Section 3.2.2 proposes MAAD-Face [Ter+20b], a new face database that contains a large number of high-quality annotations. These are required to analyse the influence of soft-biometric attributes on face recognition systems [Ter+20e] in Section 3.3.

3.2.1. Reliable Estimation of Soft-Biometrics

Introduction

To investigate the bias and privacy concerns driven by soft-biometric attributes, a tool is needed to reliably estimate these attributes. Although the estimation performance reported in previous work has highly increased overtime and closely match human-level [HOJ13; Han+15], these models tend to mispredict. This especially holds for predictions under difficult circumstances (e.g. pose, illumination), or when the trained model faces a sample belonging to a miss-represented group that was under-represented in the training data.

Intuitive solutions for this problem include rejecting face images based on the quality of the images [Han+15] or based on the model’s confidence scores. However, rejecting low-quality face images does not take into account a potential model bias. While it looks reasonable to use the model’s confidence scores, a higher confidence score does not necessarily imply a higher probability that the classifier is correct as shown in recent work [Guo+17; KL15; NYC14].

In practice, knowing the reliability of a prediction has several advantages. The reliability scores allow us to discover and prevent model biases. If a model is not sure about the decision (low reliability), it can reject the sample without a decision or ask another model or a human operator to make the decision instead [Jia+18]. Especially in forensic scenarios, having a reliability measure about the model decision is of great significance, since the assessment of the strength of evidence is a central activity in forensic case work [Zei+18]. Also for system monitoring tasks, having a reliability measure has great benefits. During deployment, these measures allow monitoring the classifier to detect distributions shifts and thus, it is possible to detect when the classifier is no longer as useful as it was when first deployed.

In this section, we propose a novel reliability measure [Ter+19d] and proof its effectiveness on the example of age and gender estimation. This solution is able to make highly accurate predictions and further stating the reliability of these predictions. By applying multiple stochastic forward passes through a dropout-reduced network, a score set of stochastic age and gender predictions are successfully created. Based on the centrality and the dispersion of these scores, the reliability and thus, the model’s confidence about the prediction is accurately specified.

We evaluate our solution on the Adience benchmark [EEH14] and show that our proposed neural network architecture reaches and goes beyond state-of-the-art results. Furthermore, we demonstrate that the proposed reliability measure correlates with the age and gender classification performance and thus, demonstrating the effectiveness of the proposed reliability measure.

Related Work

In recent years, many works were published solutions for age and gender estimation from face. For age estimation, one of the first works was published by Kwon and Lobo [KL99] in 1999. They evaluated an age group classification tasks between babies, young adults, and senior adults, based on the ratios between hand-crafted features. Today, most approaches replaced the hand-crafted features by convolutional neural network (CNN) features, because these were able to capture the complex patterns needed for age estimation tasks. These features enable a prediction performance which is even surpassing human-level performance [HOJ13]. CNN-based solutions for age estimation tasks are either focusing on age group classification [LH15; WGK15] or on age regression problems [Hue+15; YLL14]. The presented solutions cover a wide range of mechanisms such as domain-adaption [RTV18], cascade CNN’s [Che+16], autoencoders [ZBB18], and deep regression trees [She+17].

Similar to age estimation, the first works on gender estimation started in the early nineties. In 1991, Golomb *et al.* [GLS90] proposed a neural network to identify gender and reported a gender decision performance comparable to humans. However, the investigation contained only 90 subjects in a controlled environment and the whole preprocessing was performed manually. Similar to age estimation, a pipeline consisting of feature extraction and a stacked classifier was used in multiple works. The feature extraction part utilized weber’s local descriptors (WLD) [Ull+12], local binary pattern (LBP) [Sha12], and biologically inspired features (BIF) [HJ14]. More recent approaches are built on CNN features. Wolfshaar *et al.* [WKW15] fine tuned a pre-trained CNN with a stacked support vector machine (SVM), while Mansanet *et al.* [MAP16] proposed a local deep neural network with a voting scheme for the local contributions.

Most of the recent works exploit the fact that age and gender estimation tasks share similar features. In [GM13], Guo *et al.* investigated the estimation of age, gender, and ethnicity using canonical correlation analysis and partial least squares. Han *et al.* [Han+15] demonstrated that age, gender and race estimation is a challenging problem for machines and humans. They showed, on four different datasets, that the automated attribute estimation closely matches human level performance. In [EEH14], a drop-out SVM approach is used in combination with a robust face alignment technique. Considering the over-sensitiveness on facial variations of CNN's, Rodríguez *et al.* [Rod+17] proposed a CNN-based feed-forward attention mechanism. This mechanism allows a deeper focus on particular regions of the face image leading to more accurate predictions. In [Han+18], Han *et al.* proposed a deep multi-task learning approach for face attribute estimation by considering the attribute correlation and heterogeneity during the feature representation learning.

So far, previous work solely focused on enhancing the age and gender estimation performance by proposing new solutions for these tasks. However, these models still tend to mispredict under difficult circumstances or due to model or data bias. Therefore, we propose a solution that predicts the age and gender of a given face image and additionally offers an accurately reliability measure for the prediction. Consequently, weak predictions can be detected and rejected before the attribute prediction system can show an erroneous behaviour.

Methodology

Our approach is built on dropout predictions. By applying multiple stochastic forward passes through dropout-reduced networks, a stochastic set of age and gender class predictions can be obtained. In [GG16], Gal *et al.* proofed theoretically that the use of dropout predictions in NN can be interpreted as an approximation of the well known probabilistic Gaussian process model. The proposed reliability measure quantifies the confidence of the model by determining the centrality and the dispersion of the stochastic prediction set. A high centrality and a low dispersion of the stochastic predictions indicate high confidence, while a low centrality and a high dispersion characterize low confidence of the model's prediction.

The proposed network In Figure 3.1 the proposed NN architecture is shown. Given an aligned and cropped face image, a pretrained FaceNet¹¹ [SKP15] trained on MS-Celeb-1M [Guo+16] was utilized. The FaceNet embedding of this image is extracted and passed to

¹¹<https://github.com/davidsandberg/facenet>

the proposed NN. The network consists of five layers. The 128-dimensional input layer is followed by three fully connected hidden layer with 128, 16, and 128 dimensions. After each layer batchnormalization [IS15] is applied with $m = 0.99$ and $\epsilon = 10^{-3}$. Further, dropout [Sri+14] is applied on each layer with a dropout probability $p = 33.5\%$. For the activation functions, we choose leaky rectified linear units (leaky ReLU) with $\alpha = 0.1$ in order to enable the network to recognize non-linear behaviour. The last layer is split up into two softmax layers, an 8-dimensional layer for the 8 age classes and a 2-dimensional layer for both gender classes. The output of the softmax function with an input vector w is given by

$$\sigma_j(w) = \frac{\exp(w_j)}{\sum_{c=1}^C \exp(w_c)}, \quad (3.1)$$

and normalizes vector w into a probability distribution for C classes.

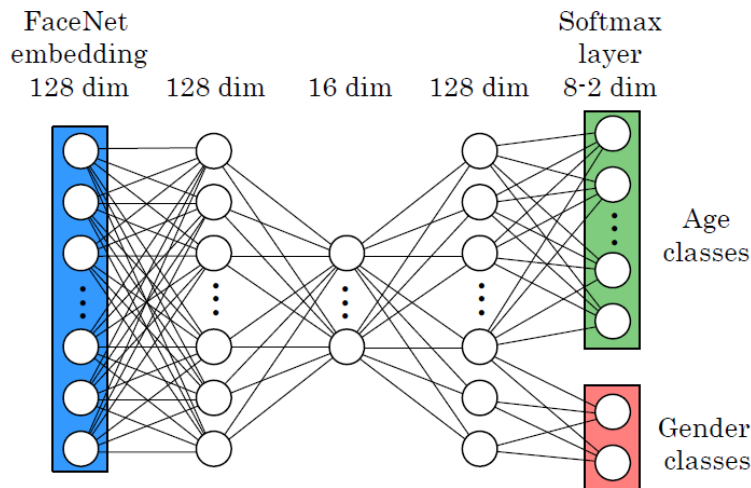


Figure 3.1.: Proposed neural network architecture: a 128-dimensional input layer is followed by three hidden layer with 128, 16, and 128 dimensions. The last layer is split up into two softmax layers, an 8-dimensional softmax age layer and a 2-dimensional softmax gender layer. After every dense layer batchnormalization is applied as well as dropout with a dropout probability of 33.5%. For activation functions, leaky rectified linear units (leaky ReLU) are utilized with $\alpha = 0.1$.

Reliability measure During deployment, the face image is aligned using MTCNN [Zha+16] and cropped. The resulting face image is passed into FaceNet [SKP15] to extract an embedding. This embedding is used as an input for the proposed network. By applying dropout during the prediction, $m = 100$ stochastic forward passes are performed to produce m different softmax outputs for the age and gender classes, denoted as $v_{i,j} \forall i \in \{1, \dots, m\}, j \in \{1, \dots, C\}$. Parameter C describes the number of classes for age ($C = 8$) and gender ($C = 2$) classification. Computing the mean of all m softmax outputs v_i lead to an averaged softmax output \bar{v}_j . The predicted class \hat{c} of the NN is given by the arguments of the maximum averaged softmax value

$$\hat{c} = \underset{j}{\operatorname{argmax}} \bar{v}_j = \underset{j}{\operatorname{argmax}} \left(\frac{1}{m} \sum_{i=1}^m v_{i,j} \right). \quad (3.2)$$

In order to develop a reliability measure that values the confidence of the model's predictions, we propose a novel reliability measure rel . This confidence measure takes into account the probability interpretation of the softmax layer as well as the agreement of the stochastic predictions. Given the outputs of the m stochastic forward passes of the predicted class \hat{c} denoted as $x = v_{i,\hat{c}}$, the proposed reliability measure

$$rel(x) = \underbrace{\frac{1 - \alpha}{m} \sum_{i=1}^m x_i}_{\text{Measure of centrality}} - \underbrace{\frac{\alpha}{m^2} \sum_{i=1}^m \sum_{j=1}^m |x_i - x_j|}_{\text{Measure of dispersion}}, \quad (3.3)$$

consists of two parts. The first part is a measure of centrality and computes the mean of the m stochastic softmax outputs for the class. This aims at utilizing the probability interpretation of the softmax output. A higher value can be interpreted as a high probability that the prediction is correct. However, this assumes that the stochastic outputs follow a Boltzmann distribution. Therefore, the second part, the measure of dispersion, quantifies the agreement of the stochastic outputs x . It calculates the mean distances between all score combinations of x . If all values in x are close to each other the measure of dispersion is low. This illustrates a case of high agreement between the stochastic predictions and can be interpreted as high confidence of the model. If the model has low confidence about the prediction, the stochastic scores in x will vary in a wider range and the measure of dispersion will be higher. The parameter $\alpha \in [0, 1]$ allows counterweighting the measure of centrality and dispersion. Low values of α will focus more on the probability interpretation of the softmax function, while high values will focus more on the variation of the stochastic outputs.

For completeness, it should be mentioned that the measures of centrality and dispersion can easily be replaced by other functions. For instance, centrality function can be replaced

by a median, while the dispersion function can be replaced by a simple Gaussian variance. However, this assumes x to be normally distributed.

Experimental Setup

Database - In order to evaluate the performance of our approach on face images under real-world conditions, the Adience dataset [EEH14] was used. The dataset consists of over 26.5k images from over 2.2k different subjects. For the experiments, every sample was considered that is labelled with a gender and an age group. The age groups consists of eight classes (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+). For both, age and gender estimation, subject-exclusive 5-fold cross-validation is used to ensure comparability with previous works. Figure 3.2 shows the data distribution over the different age and gender cases.

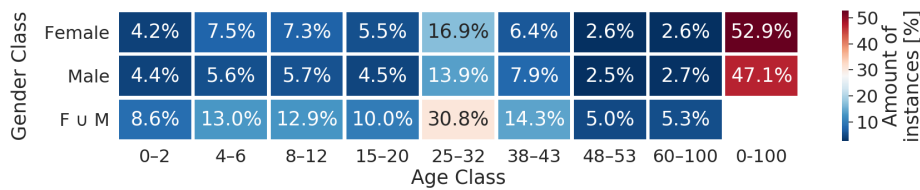


Figure 3.2.: Instance distribution of the Adience dataset. The samples were grouped by age and gender.

Investigations - The investigations in this work are separated into two parts. The first part analyses the age and gender estimation performance to compare the proposed method with state-of-the-art approaches. This includes confusion matrices of the proposed model for the age and gender estimation task. The second part of the investigation aims at analysing the effect of the prediction reliabilities. Therefore, the prediction performance is shown over several reliability thresholds to measure the correlation between the reliability and the prediction performance. To get a deeper understanding of the reliability effect, a more detailed analysis is carried out over the different gender and age classes. Finally, multiple sample images are shown for different age and gender reliabilities to visualize the understanding of the model’s reliability.

Details about the methodology - Since age and gender share many facial features, the network training was done by simultaneously optimizing the age and gender classification

task (multi-task learning) using a categorical cross-entropy loss function

$$\mathcal{L}(y^a, \hat{y}^a, y^g, \hat{y}^g) = - \sum_{i=1}^2 y_i^a \log(\hat{y}_i^a) - \sum_{i=1}^8 y_i^g \log(\hat{y}_i^g).$$

Here, y^a and y^g describe binary indicators if the age and gender class prediction was correct, while \hat{y}^a and \hat{y}^g describe the predicted probabilities for the age and gender classes. Using the adam optimization algorithm [KB14], the network was trained with a batch size of 128 over 120 epochs. These training settings resulted in the most stable results.

To calculate the reliability of a prediction, the balance between the measure of centrality and the measure of dispersion have to be determined. For the sake of simplicity, we evaluate the proposed solution for equally balanced weights ($\alpha = 0.5$).

Results

Before we come to the evaluation of the main contribution, a novel reliability measure, we show that the proposed NN solution is comparable with state-of-the-art approaches. Therefore, Figure 3.3 shows the performance of our solution together with the best performances reported by previous work. All solutions are reported on the Adience benchmark. With a gender decision accuracy of about 90%, the proposed solution shows comparable performance to previous work. For age classification, the proposed approach is slightly superior to the best performing state-of-the-art model and reaches an age class classification performance of $(64.3 \pm 2.2)\%$.

To get a deeper insight into the model behaviour, in Figure 3.4 and 3.5 the confusion matrices for age and gender are shown. The confusion matrices are presented as in the work from Rodríguez *et al.* [Rod+17]. For age, the model predicts the right age class in most cases. Errors made by the model occur mainly by predicting an adjacent age class resulting in a one-off age accuracy of 95.3%. It is noticeable that the age class (48-53) show the weakest results. This can be explained by the bias induced by the number of training samples since this age class contains the lowest number of training instances (see Figure 3.2). Moreover, the age class (25-32) can be estimated with a performance of 83.5%, which is significantly higher than other age classes. This is due to the fact that this age class is overrepresented with 30.8% of all instances. Consequently, the model optimized itself more on this age class further resulting in lower performance in adjacent age classes.

In Figure 3.5, the confusion matrix for the gender classification tasks can be seen. The results show that the model predicts female face images with slightly higher accuracy

Figure 3.3.: Best reported performances from previous work and our approach. All results are reported based on the Adience benchmark.

Model	Accuracy (%)		
	Age	One-off age	Gender
Eidinger [EEH14]	45.1 ± 2.6	80.7 ± 1.1	77.8 ± 1.3
Levi [LH15]	50.7 ± 5.1	84.7 ± 2.2	86.8 ± 1.4
Wolfshaar [WKW15]	-	-	87.2 ± 0.7
Chen [Che+16]	52.9 ± 6.0	88.5 ± 2.2	-
Liu [Liu+18]	60.2 ± 5.3	93.7 ± 2.3	-
Rodríguez [Rod+17]	61.8 ± 2.1	95.1 ± 0.0	93.0 ± 1.8
Rothe [RTV18]	64.0 ± 4.2	96.6 ± 0.9	-
Ours	64.3 ± 2.2	95.3 ± 1.5	89.8 ± 2.5

than male images, probably because the dataset contains 12% more female face images. This yields to an overall gender decision performance of $(89.8 \pm 2.5)\%$.

Figure 3.6 shows the age and gender decision performance over all possible class combinations. For age class classification (Figure 3.6a), the performance for male face images is significantly higher than for female faces. Furthermore, the best performing age class (25-32) is also the age class with the most provided training samples, while the worst performing age class (48-53) is also the one with the least amount of training samples. Considering the one-off age accuracies (Figure 3.6b), a performance in the high nineties can be observed in most cases. Only the two oldest age classes show a weaker performance, probably to a model bias to predict younger ages caused by the data distribution of the training set. For gender classification (Figure 3.6c), a very high gender recognition performance can be seen for instances of 15 years. In contrast, for lower age classes, the gender decision performance is significantly lower. This is probably due to the fact that gender-specific characteristics are more developed at older ages.

The average reliability values are shown for each age and gender class combination in Figure 3.7. It is noticeable that the proposed reliability measure correlates with the performance and the number of training samples. For instance, in Figure 3.7a, the lowest reliability values for age estimation occur in the age classes (48-53) and (60-100). These age classes show the least accuracies in age and one-off age estimation (Figure 3.6a and 3.6b). Moreover, these age classes have the lowest amount of training samples (Figure 3.2). The same can be observed for the highest reliability values, which appear in the

Figure 3.4.: *Age confusion matrix* presented as in [Rod+17]. The table represents the mean values over all cross-validation folds. Bold values indicate the class accuracy.

		Predicted							
		0-2	4-6	8-12	15-20	25-32	38-43	48-53	60+
Real	0-2	58.9	40.8	0.1	0.0	0.2	0.0	0.0	0.0
	4-6	13.1	77.0	9.1	0.5	0.2	0.1	0.0	0.0
	8-12	1.2	19.6	66.6	5.2	7.2	0.2	0.0	0.0
	15-20	0.0	1.0	6.5	40.7	49.4	1.6	0.3	0.5
	25-32	0.0	0.1	0.9	5.4	83.5	9.2	0.2	0.7
	38-43	0.0	0.0	0.3	0.8	47.2	44.9	3.0	3.8
	48-53	0.0	0.0	0.1	0.4	12.5	38.3	17.3	31.4
	60+	0.0	0.0	0.4	0.0	7.0	14.1	12.9	65.6

Figure 3.5.: *Gender confusion matrix*. The table represents the mean accuracies over all cross-validation folds.

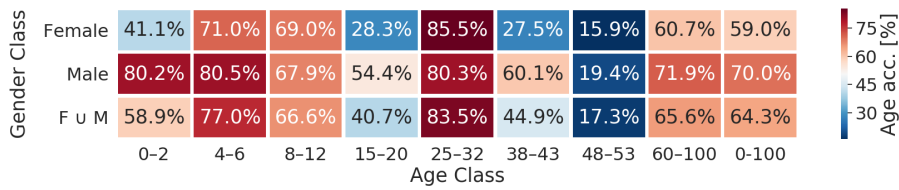
		Predicted	
		Female	Male
Real	Female	91.7	8.3
	Male	11.8	88.2

age classes (0-2) and (4-6). These age classes also show the highest performance in the one-off accuracy (Figure 3.6b).

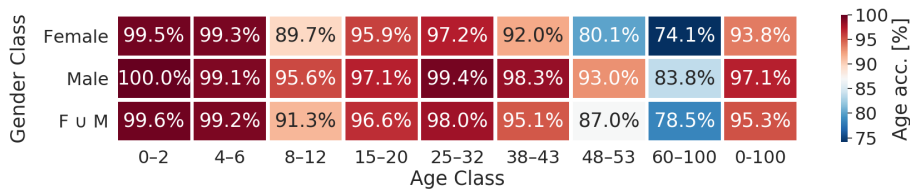
The reliability values for the task of gender estimation are shown in Figure 3.7b. Very high reliabilities can be found for age classes with 15 years and older, while younger age classes show significantly lower reliabilities. The same pattern can be found for the gender decision accuracy in Figure 3.6c. Very high accuracies are achieved for age classes over 15 years, while the performance for lower ages is significantly lower.

These results demonstrate that the proposed reliability measure correlates with the performance and thus, captures the confidence of the predictions.

One of the goals of this work is to show that utilizing the proposed reliability measure for threshold leads to better predictions. In Figure 3.8, the age and gender estimation



(a) Age accuracy per class



(b) One-off age accuracy per class

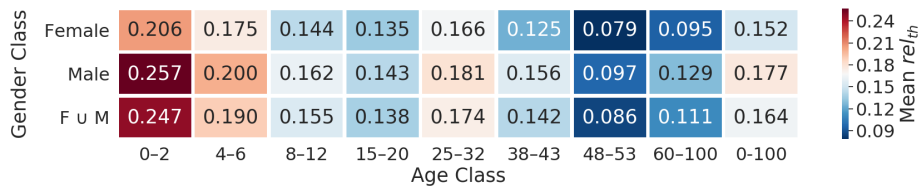


(c) Gender accuracy per class

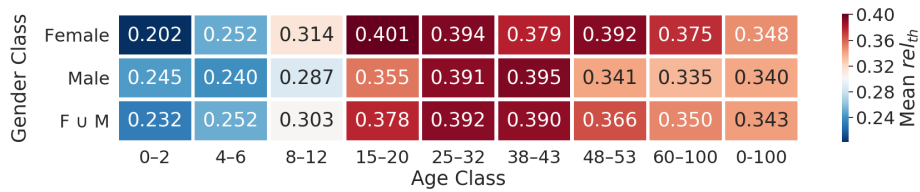
Figure 3.6.: Age and gender decision performance for all possible class combinations. The reported results are averaged accuracies over all cross-validation folds.

performance is shown for different reliability thresholds. The blue lines indicate the attribute estimation performances, while the red line indicates the percentage of instances which are over the threshold. The shaded areas represent the standard deviations over the cross-validation folds. In all cases, the performance grows with a higher reliability threshold. This indicates that the proposed reliability measure is able to value the confidence of the model's prediction.

The age classification performance is shown for different reliability thresholds in Figure 3.8a. Choosing a threshold that rejects 20% of the instances, results in a performance increase from 64% to around 70%. Rejecting 50% of the instances with the lowest reliabilities results in an age classification performance of over 75%. Figure 3.8b shows the one-off age classification accuracy over different reliability thresholds. Selecting a



(a) Age reliability per class



(b) Gender reliability per class

Figure 3.7.: Mean age and gender reliability values for all possible class combinations. The reported reliabilities show the mean values over all cross-validation folds.

threshold that rejects 20% of the instance changes the performance from 95.5% to 98%. The influence of the reliability on the gender classification accuracy is shown in Figure 3.8c. Here, rejecting 20% of the lowest reliable instance lead to an performance increase from 89.8% to over 95% and choosing a gender reliability threshold that rejects 50% of the instances results of over 98.5%. This demonstrates that rejecting low reliable samples from a biometric system is an effective approach to prevent the system from malicious predictions.

In order to get a visual understanding of the model’s reliability, Figure 3.9 and 3.10 show some random sample images for three different reliabilities. In Figure 3.9 face images are shown for three age reliabilities. The top row shows 10 random samples with a maximum age reliability around $rel \approx 0.42$. It can be seen that the model is very confident to predict the age of a baby. This is probably due to the fact that these differ most visually from the other age groups. Taking a look at the images with the highest age reliabilities, over 90% of the images show babies. The middle row presents 10 random samples with an average age reliability $rel \approx 0.14$. These images are all of high quality and show mainly young adults as it can be expected from Figure 3.7a. The bottom row presents 10 random samples with a low age reliability of around $rel \approx 0$. These images are often of bad quality, show challenging illuminations and occlusions.

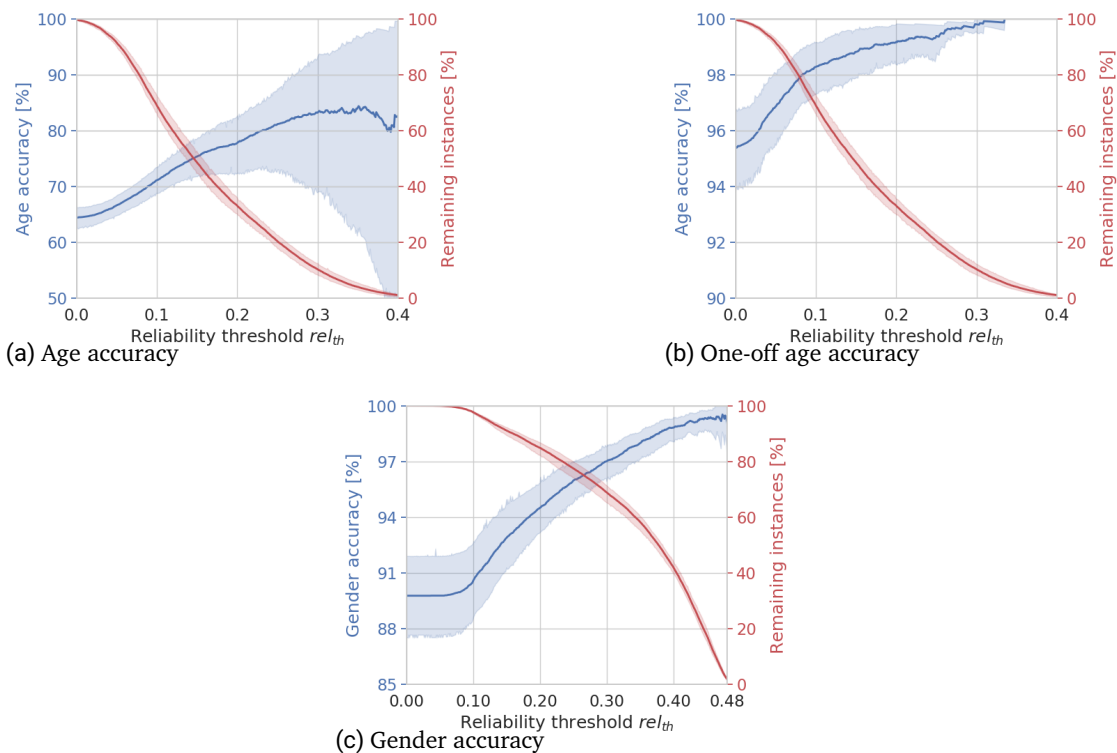


Figure 3.8.: Age and gender accuracy for different reliability thresholds. Shaded areas describe the standard deviations over the cross-validation folds. For age and gender, their own reliability values and thresholds are used.

Figure 3.10 shows face images categorized for three gender reliabilities. The top row shows 10 random samples with a high gender reliability of around $rel \approx 0.49$, the middle row presents 10 random samples with an average gender reliability of $rel \approx 0.38$, while the bottom row shows 10 random samples with a low gender reliability of around $rel \approx 0.07$. For the two upper rows, the gender can be determined easily. Only for the average reliabilities (bottom row), some images show slightly challenging illuminations. The bottom row presents samples where the classifier is least confident about the gender and also for humans classifying the gender of these images is a challenging task.

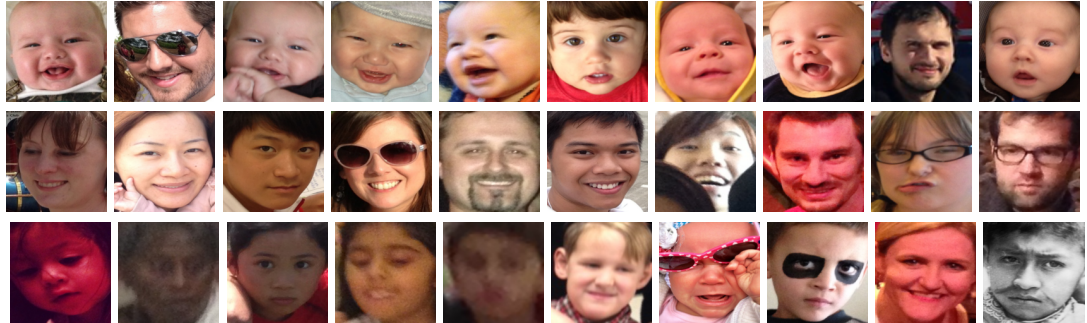


Figure 3.9.: Image samples of three different reliabilities for age estimation. The top row shows 10 random samples with a high reliability $rel \approx 0.42$. The middle row contains 10 random samples of the average reliability $rel \approx 0.14$. The bottom row contains samples with the lowest age reliability $rel \approx 0$.

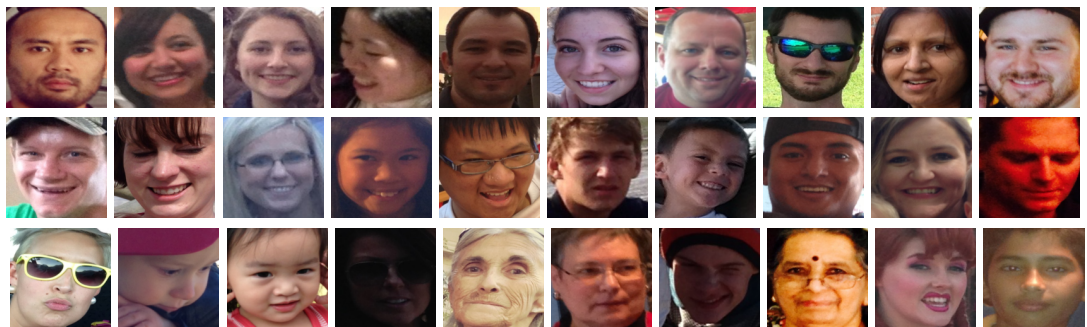


Figure 3.10.: Image samples of three different reliabilities for gender estimation. The top row shows 10 random samples with a high reliability $rel \approx 0.49$. The middle row contains 10 random samples of the average reliability $rel \approx 0.38$. The bottom row contains samples with the lowest gender reliability $rel \approx 0.07$.

Interim Conclusion

In this section, we proposed a novel and accurate measure to quantify the prediction reliability of a neural network model. Based on multiple stochastic predictions from dropout-reduced neural networks, the reliability of the model's prediction is derived by determining the centrality and the dispersion of these predictions. The experiments were conducted on the publicly available Adience benchmark. We showed that the proposed solution reaches and exceeds state-of-the-art performance. We further demonstrated that the proposed reliability measure can provide high-quality confidence statements about the predictions. Especially in forensic, these statements are of great significance, since the assessment of the strength of evidence is a central activity in forensic case work. During deployment, these statements can be further used to prevent malicious model behaviour or to monitor the system by detecting distributions shifts. Our proposed solution can be easily integrated into the many approaches that have been trained with dropout. We will use the proposed reliability measure in Section 3.2.2 to guarantee the required high-quality annotations for MAAD-Face dataset and in Section 3.5 to determine the attribute predictability and thus, to determine which attributes are stored within biometric face templates.

3.2.2. MAAD-Face: A Massively-Annotated Face Dataset

Introduction

In order to analyse the influence of soft-biometric attributes on the behaviour of face recognition systems, a database is required that contains a large number of soft-biometric attribute labels of high quality. Recent face databases are specifically constructed for the development of face recognition systems. Consequently, these contain large numbers of faces under diverse settings but lack annotations.

In this section, we propose the MAAD-Face annotations dataset [Ter+20b]. MAAD-Face is a novel face annotations database that is characterized by its large number of high-quality face annotations. Utilizing our novel annotation-transfer pipeline, we transfer the attribute labels from the source-databases (LFW [Hua+07] and CelebA [Liu+15]) to the target-database (VGGFace2 dataset [Cao+18]). The pipeline trains a massive attribute classifier to accurately predict the attributes of the source-database. Since the MAC makes use of prediction reliabilities [Ter+19d] from Section 3.2.1, the pipeline neglects annotations of low-confident predictions. MAAD-Face consists of 3.3M faces of over 9k individuals. With 123.9M attribute annotations of 47 different binary attributes, it provides 15 and 137 times more attribute labels than CelebA and LFW. To analyse the quality of the attribute annotations, three human evaluators investigated the correctness of the labels of CelebA, LFW, and MAAD-Face. The results demonstrate the superiority of the MAAD-Face annotations over the other databases. The MAAD-Face dataset is also publicly available².

To summarize, this section presents *three main contributions*:

1. A novel annotation transfer-pipeline is proposed that is able to transfer highly-accurate attribute labels from source-databases to a target-database. We use this pipeline to create MAAD-Face.
2. We propose the MAAD-Face annotations dataset based on VGGFace2 [Cao+18]. MAAD-Face is a new face annotations database consisting of 123.9M attribute annotations of 47 different binary attributes. It provides 15 and 137 times more labels than CelebA and LFW, while the attribute annotations are of higher quality.
3. A human evaluation of the annotation correctness of three large-scale annotation face databases, LFW, CelebA, and MAAD-Face, is conducted. These demonstrate the superiority of the MAAD-Face annotations over the other investigated databases.

²<https://github.com/pterhoer/MAAD-Face>

The rest of this section is structured as follows. First, an overview of annotated face datasets is provided. Second, a human evaluation of the annotation-correctness of three relevant datasets is provided and discussed. Third, the label-transfer pipeline is explained and how it is used to create MAAD-Face. Finally, the statistical properties of the MAAD-Face annotation database are discussed.

Review of Annotated Face Datasets

In recent years, a number of face databases have been released. These mainly aimed at providing a large dataset for developing face recognition solutions. With the use of deep-learning techniques in face recognition, the required data for training these solutions has grown strongly and thus, the sizes of face databases. However, less attention was given to the estimation of facial attributes. These soft-biometric characteristics can be of high importance in applications such as access control [DER16], human-computer interaction [Ter+19d], and law enforcement [GZS07]. Current face databases only provide insufficient numbers of training labels for training accurate solutions. Moreover, these labels often lack in their correctness and thus, prevent the development of soft-biometric solutions. In the following, we discuss popular face databases that also contain attribute information.

ColorFeret [Phi+00] consists of 14.1k images of 1.2k different individuals with different poses under controlled conditions. The dataset includes a variety of face poses, facial expressions, and lighting conditions. Each image contains labels of the individual's gender, ethnicity, head pose, age, glasses, and beard. In total, ColorFeret provides around 183k soft-biometric labels.

The Adience dataset [EEH14] consists of over 26.5k images of over 2.2k different individuals in unconstrained environments. In total, the dataset provides around 263k annotations for gender and age. These labels that were manually labelled.

The Morph dataset [RT06b] contains 55.1k frontal face images of more than 13.6k individuals. For each image, it provides information about the person's gender, ethnicity, age, beard, and glasses. 80.4% of the faces belong to the ethnicity black, 19.2% to white, and 0.4% to others. The individual's age varies from 16-77 years. 79.4% of the faces are within an age-range of [20, 50]. In total, the Morph database provides over 0.5M labels for soft-biometric attributes.

VGGFace [PVZ15] and VGGFace2 [Cao+18] are two databases from the University of Oxford. VGGFace [PVZ15] contains 2.6M images from 2.6k individuals and provides information about the head pose (frontal, profile). VGGFace2 [Cao+18] contains faces from over 9k subjects with over 3M images. The dataset contains a large variety of pose, age, and ethnicity. Over 40% of the face are frontal and over 50% are half-frontal. Most

images belong to individuals over 18 years old and around 40% belong to the age group of [25, 34]. For each image, gender annotations are available. A subset of 30k images of celebrities was additionally labelled with 10 further attributes about the individual’s hair, beard, glasses, and hat. In total, VGGFace2 provides 3.6M labels about the person’s face.

Labeled Faces in the Wild (LFW) [Hua+07] contains 13.2k images of 5.7k different identities from unconstrained environments. It contains variability in pose, lighting, expression, and demographics. With 74 binary attributes, it provides a large diversity on binary attribute annotations. However, as we show in Section 3.2.2, the correctness of these labels are often weak (72% accuracy compared to human annotations). In total, LFW provides over 0.9M attribute labels.

The CelebFaces Attributes Dataset (CelebA) [Liu+15] contains over 202k images of 10.0k different subjects. It covers large pose variations and background clutter and provides rich annotations for 40 binary attributes. In total, CelebA provides over 8M labels for soft-biometric attributes.

In this section, we propose the MAAD-Face database [Ter+20b]. Using our novel label-transfer technique we are able to create highly accurate face annotations building upon VGGFace2. Consequently, it contains over 3.3M face images from over 9.1k different subjects with a large variety of poses, ages, and ethnicities. MAAD-Face provides labels for 47 binary attributes. In total, it consists of over 123.9M attribute annotations, which is over 15 times higher than the second-largest face annotation dataset. Moreover, its label quality is significantly higher than related databases, such as LFW and CelebA, as we will show in the following Section 3.2.2.

Evaluating Label-Correctness of Related Face Datasets

We evaluate the quality of attribute labels from three face datasets, LFW, CelebA, and MAAD-Face. The quality refers to the correctness of the labels compared to the annotations of human evaluators. The label-correctness of each attribute in LFW, CelebA, and MAAD-Face was manually evaluated by three human evaluators. For each attribute, the evaluators got 50 positively-labelled and 50 negatively-labelled images. These were randomly chosen. Then, each evaluator was asked to carefully label these images for the given attribute. This led to over 16k manually annotated labels³. The manually annotated labels are used to compute the accuracy, precision, and recall for each attribute of database. The accuracy refers to the percentage of correct labels, where the ground truth is determined by the human evaluators. Precision is defined as the number of true positives over the number of

³Please note that this only represents a small fraction of all labels and additionally reflects the subjective opinion of the three evaluators. Therefore, the results should not be considered as absolute values but should rather be used as indicators.

Table 3.1.: Statistics of related face annotation databases. Distinctive attributes refer to the number of attributes that are labelled while the number of labels refers to the total number of (attribute) annotations in the database. The correctness of the attribute annotations (see Section 3.2.2) are shown for the most relevant databases, LFW, CelebA, and MAAD-Face, since these contain labels for a high number of distinct attributes. In total, MAAD-Face provides the highest number of attribute annotations. Moreover, MAAD-Face additionally provides labels of much higher quality than related databases.

Database	Num. of subjects	Num. of images	Attribute labels				
			Distinctive attributes	Number of labels	Accuracy	Precision	Recall
ColorFeret	1.2k	14.1k	13*	0.2M			
Adience	2.3k	26.6k	10*	0.3M			
Morph	13.6k	55.1k	10*	0.6M			
VGGFace	2.6k	2.6M	1	2.6M			
VGGFace2	9.1k	3.3M	11	3.6M			
LFW	5.7k	13.2k	74	0.9M	0.72	0.61	0.84
CelebA	10.0k	0.2M	40	8.0M	0.85	0.83	0.89
MAAD-Face	9.1k	3.3M	47	123.9M	0.91	0.87	0.94

true and false positives. In our context, precision refers to "What proportion of positive labelled-samples in the database is also positively-labelled by the human evaluators?". Recall is defined as the number of true positives over the number of true positives and false negatives. In our context, recall refers to "What proportion of positive labels annotated by the human evaluators are identified correctly?". Tables 3.3, 3.4, and 3.2 present the results for this analysis on LFW, CelebA, and MAAD-Face.

LFW For LFW (Table 3.3), many attributes show a very weak performance and thus, a low correlation with the annotations of the human evaluators. Young age group labels (baby, child, youth) are close to a random accuracy and additionally often have a small precision. This is also observed e.g. for *frowning*, *chubby*, *curly hair*, *wavy hair*, *bangs*, *goatee*, and *square face*. Moreover, labels for *attractive man* are mostly placed on female faces. In general, there is a big mismatch between the labels of LFW and the annotations of the human evaluators. The accuracy for most attributes is below 80% and only 5 out of 76 attributes have an accuracy of over 90%. Over all attributes this leads to an accuracy

of 72%, a precision of 61%, and a recall of 84%. The high gap between the low precision and the relatively high recall indicates that there are a lot of falsely positive annotated labels in LFW.

CelebA The attribute performance for CelebA is shown in Table 3.4. It has labels for 40 binary attribute, which is a lower number than LFW. However, these annotations are of much higher quality. Only 2 attributes have an accuracy of less than 70% and 14 attributes even reach over 90%. Over all attributes, the accuracy is 85%, the precision is 83%, and the recall is 89%. Similar to LFW, there is a tendency that most of the wrong labels are within in the positives.

MAAD-Face Table 3.2 shows the attribute performance of MAAD-Face. MAAD-Face has labels for 47 binary attributes. In the evaluation against the human annotations, 3 attributes reach a performance of below 70%. However, also 34 attributes reach over 90% accuracy with the majority of close to 100%. Over all attributes, this leads to an accuracy of 91%, a precision of 87%, and a recall of 94%.

Summary In Table 3.1 the properties of the investigated databases are shown including the overall performance of our annotation-correctness study. Although LFW provides the highest number of binary attributes, it provides the lowest number of attribute labels with the lowest annotation qualities. Only 72% of the investigated labels match the annotations of the human evaluators. CelebA consists of 40 binary labels with a total of 8.0M attribute annotations. Moreover, with an accuracy of 85%, the quality of these annotations is significantly higher. In terms of the number of labels and label-quality, MAAD-Face exceeds the other databases. It provides 47 binary attributes with a total of 123.9M labels. This is 15 times higher than CelebA and 137 times higher than LFW. Moreover, the labels quality (in terms of accuracy, precision, and recall) is significantly higher than the other databases. 91% of the MAAD-Face labels match the annotations of human evaluators. Consequently, MAAD-Face provides significantly more and higher-quality attribute annotations.

Labels-Transfer Pipeline

We will present one of the main contributions of this work, a novel label-transfer pipeline that is able to create highly reliable and accurate attribute annotations. We will explain this pipeline based on the example of the MAAD-Face annotations database. The MAAD-Face

database that was created by transferring the labels of CelebA and LFW on the images of VGGFace2.

An overview of the proposed label-transfer pipeline is shown in Figure 3.11. The pipeline consists of five steps that aim to transfer the labels of source-databases to the target database.

1. A massive attribute classifier (MAC) is trained on the training-part of the source-datasets. Besides making predictions about the estimated labels of a given image, the MAC is able to additionally providing a reliability statement that states the model's prediction confidence for each label.
2. The MAC predicts the labels on the test-part of the source-datasets including the prediction reliabilities.
3. Based on this performance, the reliability threshold for each attribute is determined. Moreover, a performance-reliability mapping is calculated that allows assigning an attribute reliability with its expected correctness (performance).
4. The MAC predicts the attribute labels as well as the corresponding reliabilities for each image in the target-dataset. Predicted labels below the attribute threshold will be rejected to guarantee a high-quality of the transferred source annotations.
5. Finally, the source annotations (with their reliabilities) are aggregated using the corresponding performance-reliability mapping. If the source annotations for an image produces different labels, the label is used as the target label that has the higher expected correctness.

In the following, we describe how (a) the MAC training procedure is conducted on the source-datasets, (b) the prediction reliability statements of the MAC are calculated, and (c) how this results in the final labels for target-database.

Massive Attribute Classifier (MAC) To transfer the labels for each attribute from source databases to a target database, we (a) train a MAC jointly on all attributes of a source-database to make use of a shared embedding space and (b) construct the MAC such that it is able to produce accurate reliability measures for each attribute-label prediction.

The MAC is a neural network that is trained to predict the attributes of the source-dataset. The network architecture is chosen to maximize the prediction accuracy. As it will be demonstrated in Section 3.2.2, the only requirement for the MAC is trained with at least one dropout-layer [Sri+14]. We will need this layer to determine the reliability of a prediction. Each source-database is subject-exclusively divided into an 80% training set

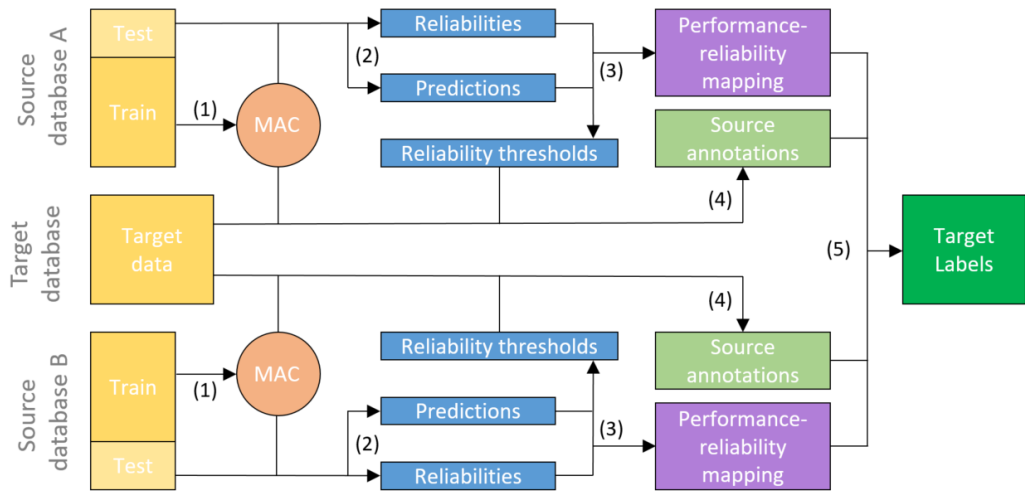


Figure 3.11.: Overview of the proposed label-transfer pipeline. For each source dataset, (1) a MAC is trained on the training part of a source dataset. (2) The MAC produces predictions and prediction reliabilities on the test set. (3) These are used to determine the reliability thresholds per attribute and to calculate the performance-reliability mapping. (4) The MAC and the reliability thresholds are used to create (source) attribute annotations for the target dataset. Finally, (5) the source annotations from each source dataset are aggregated using the corresponding performance-reliability mappings to construct the final target labels for the target dataset.

and a 20% test set. A separate MAC is trained for each source-training set. To construct MAAD-Face, we use VGGFace2 as the target-database and CelebA and as source-databases for training two MACs.

In the following, we describe the structure and the training details of the MAC, as well as the data cleaning process used. As we demonstrated in Section 3.2.2, many labels of LFW are wrongly assigned. To prevent a confusion of the MAC trained on these labels, we filter out labels that are wrongly assigned with a high probability.

MAC training Generally, the training of the MAC can vary and should be task and data-dependent. In order to prepare the MAC for our label-transfer pipeline, it needs to be trained with at least one dropout-layer [Sri+14] and consists of a soft-max layer as the output.

For the construction of MAAD-Face, we build the MAC on the templates of face images. As shown in [Ter+20a], one can easily and accurately predict many attributes from such templates. Based on these results, we trained a neural network model on FaceNet [SKP15] embeddings to jointly predict multiple attributes of the source-database. However, a MAC can also be trained end-to-end or by fine-tuning an existing network. The utilized network structure follows the one used in [Ter+20a]. It consists of two initial layers, the input layer of size n_{in} and the second dense layer of size 512. The size of the utilized face embedding is denoted by n_{in} and for our FaceNet model⁴ refers to 128 dimensions. Starting from the second layer, each attribute a has an own branch consisting of two additional layers of size 512 and $n_{out}^{(a)}$, where $n_{out}^{(a)}$ refers to the number of classes per attribute. Each layer has a ReLU activation, except for the output-layers. These have softmax activations. Moreover, Batch-Normalization [IS15] and dropout [Sri+14] ($p_{drop} = 0.5$) is applied to every layer. The dropout allows to generalize the performance but also enables us to derive reliability statements about the predictions as we will describe in Section 3.2.2. The training of the MAC was done in a multi-task learning fashion by applying a categorical cross-entropy loss for each attribute branch and use an equal weighting between each of these attribute-related losses. For the training, an Adam optimizer [KB14] was used with $e = 200$ epochs, an initial learning rate $\alpha = 10^{-3}$, and a learning-rate decay of $\beta = \alpha/e$. The parameter choices followed [Ter+20a]. The batch size b was chosen according to the amount of available training data, $b = 1024$ for CelebA and $b = 16$ for LFW.

Cleaning training attribute labels For the label-transfer pipeline, this step is only necessary if a source-database consists of attribute annotations of low quality. As we demonstrated in Section 3.2.2, this is the case for LFW. However, the quality of the input data of a model is important for the quality of its output data [Gei+20]. Therefore, in this section, we will describe a label-cleaning process that was used on the LFW dataset.

While in CelebA the attributes are of binary nature, the labels in LFW originate from the prediction probabilities of a binary classifier [Hua+07]. Therefore, these labels are continuous and measure the degree of the attribute [Kum+09; Kum+11]. Positive values represent "true" labels and negative values represent "false" labels. However, using the prediction probabilities of a binary classifier does not necessarily reflect the correctness of the prediction as shown in recent works [Guo+17; KL15; NYC14]. Consequently, a wide range of the LFW labels centred around a value of zero is unreliable.

To ensure that our MAC learns on meaningful LFW-labels, we manually removed these centred labels as described in [Ter+20a]. Therefore, we assigned an upper and lower score threshold for each attribute. Images with a score over the upper threshold are

⁴<https://github.com/davidsandberg/facenet>

assigned as true, images with a score under the lower threshold are assigned as false, images with scores within the range are rejected. The upper and lower thresholds for one attribute are manually determined by moving potential thresholds away from zero. At each potential threshold, ten images with the closest attribute scores are investigated. Here, the original LFW labels of the images are manually investigated for correctness. If only eight or fewer attributes are investigated as correct, the potential threshold is further moved away from the starting point and the procedure is repeated. If a potential threshold returns images with 9 or more correct labels, it is chosen as the limit. Repeating this over all attributes will result in a lower and an upper threshold for each of these attributes. By binaryzing the scores with these upper and lower thresholds, reduces the amount of labels by 51,7%. However, it also ensures an error-minimizing data basis of the MAC. Thus, it allows us to train the MAC on meaningful and mostly correctly labelled data.

Deriving reliability statements To ensure that the target-database will only get annotations of high quality, the prediction reliability is additionally estimated for each prediction (target-label). Therefore, we follow the methodology described in Section 3.2.1 [Ter+19d] to enable our MAC to accurately state its own prediction confidence (reliability). To derive the reliability statement additionally to an attribute prediction, $m = 100$ stochastic forward passes are performed. In each forward pass, a different dropout-pattern is applied, resulting in m different softmax outputs $v_i^{(a)}$ for each attribute a . Given the outputs of the m stochastic forward passes of the predicted class \hat{c} denoted as $x^{(a)} = v_{i,\hat{c}}^{(a)}$, the reliability measure is given as

$$rel(x^{(a)}) = \frac{1 - \alpha}{m} \sum_{i=1}^m x_i^{(a)} - \frac{\alpha}{m^2} \sum_{i=1}^m \sum_{j=1}^m |x_i^{(a)} - x_j^{(a)}|,$$

with $\alpha = 0.5$, following the recommendation in [Ter+19d]. The first part of the equation is a measure of centrality and utilizes the probability interpretation of the softmax output. A higher value can be interpreted as a high probability that the prediction is correct. The second part of the equation is the measure of dispersion and quantifies the agreement of the stochastic outputs x . In [Ter+19d], this was shown to be an accurate reliability measure.

Attribute label generation We combine the MAC models of the source-datasets and the reliability measure to create high-quality target annotations. First, we will describe how to set the reliability thresholds for each attribute and MAC. Then, we will describe how this can be used to create the annotations on the target-dataset.

Defining reliability thresholds For each source-database, a MAC model \mathcal{M} was already trained on the training-part as described in Section 3.2.2. Now, the MAC predicts the source-annotations on the test-part including the prediction reliabilities. Moreover, the MAC repeats this step on the target-database. For each attribute a of the source-database, the reliability threshold $thr_{Source}^{(a)}$ is chosen such that the (balanced) prediction accuracy of a is over $acc_{min}\%$ and at least $d_{min}\%$ of the target-samples are over this threshold. Consequently, acc_{min} defines the quality of the target-labels while d_{min} define the amount of the labels in the target-database. If an attribute does not accomplish this requirement, the attribute is discarded.

For the creation of MAAD-Face, we set $acc_{min} = 90\%$ and $d_{min} = 50\%$ to receive a large number of high-quality annotations. This results in manually chosen reliability thresholds $thr_{CelebA}^{(a)}$ and $thr_{LFW}^{(a)}$ for each attribute $a \in \mathcal{A}$.

Creating target-labels After defining the reliability thresholds for each MAC and attribute $a \in \mathcal{A}$, we can create the target-annotations. Therefore, each MAC computes its predictions p_{Source} and prediction reliabilities r_{Source} on the target-dataset. The prediction *True* is defined as 1, the prediction *False* is defined as -1. If an attribute-prediction $p_{Source}^{(a,i)}$ for an image i has a prediction-reliability below the threshold $r_{Source}^{(a,i)} < thr_{Source}^{(a)}$, the annotation is set to 0 (*undefined*). In this case, the MAC is not confident enough about its prediction and rejecting these predictions guarantee high-quality remaining labels. For each source-dataset, this procedure results in a set of labels l_{Source} for the target dataset images. Finally, this set of labels have to be combined to create the target-annotations. If an attribute just appears in one of the source-datasets, the source-labels l_{Source} are directly used for the target-dataset. If an attribute appears in multiple source-datasets, we have to decide which label to use as the target-annotation. In this case, the reliability r_{Source} is mapped back to the performance of the test set $acc(r_{Source})$ and the label assigned with the highest map-back performance is used for the target-annotation. Please note that such a decision can not be made based on the reliability-level only since the range of the reliability values vary between each MAC. Mapping back the reliability values to the test-set performances allow an aligned comparison of the label-quality.

Algorithm 1 summarizes the label generation procedure. The inputs are the predictions $\{p_{Source}\}$, the corresponding reliabilities $\{r_{Source}\}$, the reliability thresholds $\{thr_{Source}\}$, as well as a set of all attribute \mathcal{A} . The output of the algorithm is the annotations l_{Target} of the target-dataset. The *transfer* function transforms the predictions p_{Source} into the source-labels l_{Source} based on the prediction reliabilities r_{Source} and the corresponding attribute reliability thresholds thr_{Source} . If an attribute appears in multiple source databases, the *highest* function maps back the reliability to the test-set performance $acc(r_{Source}^{(a,i)})$ and

returns the label $l_{Source}^{(a,i)}$ with the highest map-back performance.

The last step (*obtainPlausability*) performs a plausibility check including required corrections, given the target labels l_{Target} , the attributes \mathcal{A} , and the corresponding attribute classes. For each attribute, at maximum one class can be true. For instance, for the attribute gender, either the class male or female can be true. A list of the attributes with the corresponding classes is shown in Table 3.2. Due to this restriction, we set all attribute class labels for an image i to undefined (0) if more than one attribute class showed true before. This aims at maintaining high-quality labels.

Algorithm 1 - Label Generation

Input: $\{p_{Source}\}, \{r_{Source}\}, \{thr_{Source}\}, \mathcal{A}$

Output: Target-dataset labels l_{Target}

```

1: for  $a \in \mathcal{A}$  do
2:   for each source dataset do
3:      $l_{Source}^{(a)} \leftarrow transfer(p_{Source}^{(a)}, r_{Source}^{(a)}, thr_{Source}^{(a)})$ 
4:   end for
5: end for
6:  $l_{MAAD} = zeros(|\mathcal{A}|, |\mathcal{I}|)$ 
7: for  $a \in \mathcal{A}$  do
8:   for  $i \in \mathcal{I}$  do
9:      $l_{Target}^{(a,i)} \leftarrow highest(\{l_{Source}^{(a,i)}\}, \{acc(r_{Source}^{(a,i)})\})$ 
10:  end for
11: end for
12:  $l_{Target} \leftarrow obtainPlausability(l_{Target}, \mathcal{A})$ 
13: return  $l_{Target}$ 

```

MAAD-Face Statistics

The biggest advantage of MAAD-Face is its large number of high-quality attribute labels. Since it builds on the VGGFace2 database, it consists of over 9.1k identities with over 3.3M face images of various poses, ages, and illuminations. MAAD-Face has labels for 47 distinctive attributes with a total of 38.3M labels. On average 37.5 ± 3.7 labels are defined per image. Figure 3.12 shows the label distribution of MAAD-Face for all 47 attributes. For each attribute, green indicates the percentage of positive labels, red indicates the percentage of negatively labelled images, and grey represents the percentage of images with undefined labels. Some attributes have a low number of positive labels, such as

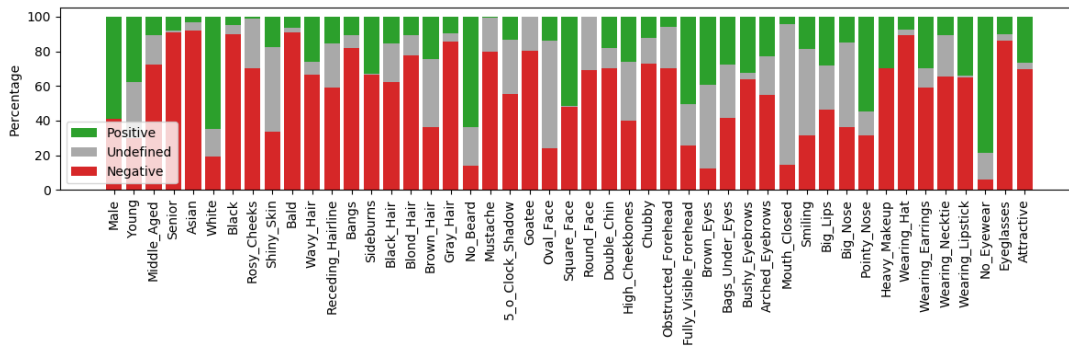


Figure 3.12.: Label distribution of the proposed database MAAD-Face. For each of the 47 attributes, green indicates the percentage of positive labels, red indicates the percentage of negatively labelled images, and grey represents the percentage of images that have an undefined label for the attribute.

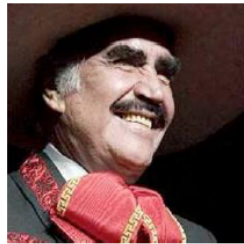
Mustache (16.6k) or *Goatee* (9.2k) and instead, a higher number of undefined labels. This way, we can ensure high correctness of the labels as explained in Section 3.2.2 (accuracy *Mustache* 98%, accuracy *Goatee* 95%). In total, this leads to MAAD-Face having 23.1% positive, 56.6% negative, and 20.3% undefined labels. A list of all attributes with the correctness analysis was already discussed with Table 3.2 in Section 3.2.2. The high quality of the attribute labels is also observable in Figure 3.13. There, five random sample images are shown with their corresponding attribute labels.

Interim Conclusion

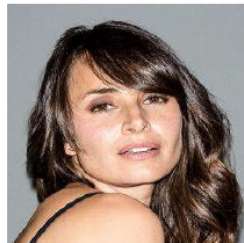
This section presented three contributions: (1) A novel annotation transfer pipeline is proposed that allows to transfer attribute labels of high accuracy from multiple source-datasets to a target-dataset. This pipeline is used to create MAAD-Face. (2) MAAD-Face is a novel face annotations database that provides over 3.3M faces with 123.9M annotations of 47 different attributes. To the best of our knowledge, MAAD-Face is the publicly available database that provides the largest number of attribute annotations. (3) The correctness of the attribute labels of three relevant annotated face databases, CelebA, LFW, and MAAD-Face are evaluated. This evaluation was performed manually by three human evaluators and demonstrated that the attribute annotations of MAAD-Face are of significantly higher quality than related databases. In the next section, MAAD-Face will be used to analyse the influence of soft-biometric attributes on face recognition systems.



Male	1	Bangs	-1	Round Face	-1	Big Lips	0
Young	0	Sideburns	-1	Double Chin	-1	Big Nose	0
Middle Aged	1	Black Hair	1	High Cheekbones	-1	Pointy Nose	1
Senior	-1	Blond Hair	-1	Chubby	-1	Heavy Makeup	-1
Asian	-1	Brown Hair	0	Obstructed Forehead	0	Wearing Hat	-1
White	1	Gray Hair	-1	Fully Visible Forehead	0	Wearing Earrings	-1
Black	-1	No Beard	0	Brown Eyes	0	Wearing Necktie	1
Rosy Cheeks	-1	Mustache	0	Bags Under Eyes	1	Wearing Lipstick	-1
Shiny Skin	0	5 o Clock Shadow	0	Bushy Eyebrows	1	No Eyewear	1
Bald	-1	Goatee	0	Arched Eyebrows	-1	Eyeglasses	-1
Wavy Hair	-1	Oval Face	0	Mouth Closed	-1	Attractive	-1
Receding Hairline	0	Square Face	1	Smiling	1		



Male	1	Bangs	-1	Round Face	0	Big Lips	0
Young	-1	Sideburns	1	Double Chin	1	Big Nose	1
Middle Aged	-1	Black Hair	0	High Cheekbones	0	Pointy Nose	-1
Senior	1	Blond Hair	-1	Chubby	1	Heavy Makeup	-1
Asian	-1	Brown Hair	-1	Obstructed Forehead	1	Wearing Hat	1
White	0	Gray Hair	1	Fully Visible Forehead	-1	Wearing Earrings	-1
Black	-1	No Beard	-1	Brown Eyes	0	Wearing Necktie	-1
Rosy Cheeks	0	Mustache	1	Bags Under Eyes	0	Wearing Lipstick	-1
Shiny Skin	1	5 o Clock Shadow	-1	Bushy Eyebrows	1	No Eyewear	1
Bald	-1	Goatee	-1	Arched Eyebrows	-1	Eyeglasses	-1
Wavy Hair	-1	Oval Face	-1	Mouth Closed	0	Attractive	-1
Receding Hairline	0	Square Face	1	Smiling	0		



Male	-1	Bangs	1	Round Face	0	Big Lips	0
Young	1	Sideburns	-1	Double Chin	-1	Big Nose	-1
Middle Aged	-1	Black Hair	-1	High Cheekbones	1	Pointy Nose	1
Senior	-1	Blond Hair	-1	Chubby	-1	Heavy Makeup	1
Asian	-1	Brown Hair	1	Obstructed Forehead	0	Wearing Hat	-1
White	1	Gray Hair	-1	Fully Visible Forehead	-1	Wearing Earrings	1
Black	-1	No Beard	1	Brown Eyes	0	Wearing Necktie	-1
Rosy Cheeks	0	Mustache	-1	Bags Under Eyes	-1	Wearing Lipstick	1
Shiny Skin	0	5 o Clock Shadow	-1	Bushy Eyebrows	-1	No Eyewear	1
Bald	-1	Goatee	-1	Arched Eyebrows	-1	Eyeglasses	-1
Wavy Hair	1	Oval Face	0	Mouth Closed	0	Attractive	1
Receding Hairline	-1	Square Face	-1	Smiling	0		



Male	-1	Bangs	-1	Round Face	-1	Big Lips	-1
Young	1	Sideburns	-1	Double Chin	-1	Big Nose	-1
Middle Aged	-1	Black Hair	0	High Cheekbones	1	Pointy Nose	1
Senior	-1	Blond Hair	0	Chubby	-1	Heavy Makeup	1
Asian	-1	Brown Hair	0	Obstructed Forehead	-1	Wearing Hat	-1
White	1	Gray Hair	0	Fully Visible Forehead	1	Wearing Earrings	1
Black	-1	No Beard	1	Brown Eyes	-1	Wearing Necktie	-1
Rosy Cheeks	0	Mustache	-1	Bags Under Eyes	-1	Wearing Lipstick	1
Shiny Skin	0	5 o Clock Shadow	-1	Bushy Eyebrows	-1	No Eyewear	1
Bald	-1	Goatee	-1	Arched Eyebrows	0	Eyeglasses	-1
Wavy Hair	1	Oval Face	1	Mouth Closed	0	Attractive	1
Receding Hairline	-1	Square Face	-1	Smiling	1		



Male	-1	Bangs	-1	Round Face	0	Big Lips	1
Young	0	Sideburns	-1	Double Chin	-1	Big Nose	-1
Middle Aged	1	Black Hair	1	High Cheekbones	1	Pointy Nose	0
Senior	-1	Blond Hair	-1	Chubby	-1	Heavy Makeup	1
Asian	1	Brown Hair	0	Obstructed Forehead	-1	Wearing Hat	-1
White	-1	Gray Hair	-1	Fully Visible Forehead	1	Wearing Earrings	1
Black	-1	No Beard	1	Brown Eyes	1	Wearing Necktie	-1
Rosy Cheeks	-1	Mustache	-1	Bags Under Eyes	-1	Wearing Lipstick	1
Shiny Skin	1	5 o Clock Shadow	-1	Bushy Eyebrows	-1	No Eyewear	0
Bald	-1	Goatee	-1	Arched Eyebrows	1	Eyeglasses	-1
Wavy Hair	1	Oval Face	0	Mouth Closed	-1	Attractive	1
Receding Hairline	-1	Square Face	-1	Smiling	-1		

Figure 3.13.: Samples images from MAAD-Face with the corresponding 47 attribute-labels.

Table 3.2.: Attribute label quality analysis of MAAD-Face. Main source describes from which dataset most of the labels are transferred from.

Main source	Category	Attribute	Class	Accuracy	Precision	Recall	
CelebA	Demographics	Gender	Male	0.99	0.98	1.00	
CelebA		Age	Young	0.99	1.00	0.98	
LFW			Middle Aged	0.93	0.98	0.89	
LFW			Senior	0.97	0.96	0.98	
LFW	Race		Asian	0.90	0.88	0.92	
LFW			White	0.89	1.00	0.82	
LFW			Black	0.94	0.90	0.98	
CelebA	Skin	Rosy Cheeks	Rosy Cheeks	0.99	0.98	1.00	
LFW		Shiny Skin	Shiny Skin	0.77	0.84	0.74	
CelebA	Hair	Hairstyle	Bald	0.96	0.92	1.00	
CelebA			Wavy Hair	0.99	1.00	0.98	
CelebA		Receding Hairline	Receding Hairline	0.77	0.54	1.00	
CelebA		Bangs	Bangs	0.98	0.96	1.00	
CelebA		Sideburns	Sideburns	0.93	0.88	0.98	
CelebA		Haircolor		Black Hair	0.98	0.96	1.00
CelebA				Blond Hair	1.00	1.00	1.00
CelebA				Brown Hair	0.97	0.94	1.00
CelebA				Gray Hair	0.95	0.90	1.00
CelebA		Beard	Beard	No Beard	0.98	1.00	0.96
CelebA	Mustache			0.98	0.98	0.98	
CelebA			5 o Clock Shadow	0.97	0.94	1.00	
CelebA			Goatee	0.95	0.90	1.00	
LFW	Face Geometry	Face Shape	Oval Face	0.81	0.90	0.76	
LFW				Square Face	0.80	0.78	0.81
LFW				Round Face	0.69	0.56	0.76
CelebA			Double Chin	Double Chin	0.94	0.88	1.00
CelebA			High Cheekbones	High Cheekbones	0.92	0.92	0.92
CelebA			Chubby	Chubby	0.94	0.88	1.00
LFW			Forehead visibility	Obstructed Forehead	0.91	0.94	0.89
LFW				Fully Visible Forehead	0.80	0.75	1.00
LFW		Periocular	Brown Eyes	Brown Eyes	0.68	0.44	0.85
LFW			Bags Under Eyes	Bags Under Eyes	0.68	0.40	0.91
CelebA	Bushy Eyebrows		Bushy Eyebrows	0.95	0.94	0.96	
CelebA		Arched Eyebrows	Arched Eyebrows	1.00	1.00	1.00	
LFW	Mouth	Mouth Closed	Mouth Closed	0.84	0.80	0.87	
CelebA		Smiling	Smiling	0.95	1.00	0.91	
LFW		Big Lips	Big Lips	0.70	0.50	0.83	
CelebA	Nose	Nose type	Big Nose	0.97	0.98	0.96	
LFW			Pointy Nose	0.88	0.88	0.88	
CelebA	Accessories	Heavy Makeup	Heavy Makeup	0.98	0.98	0.98	
CelebA		Wearing Hat	Wearing Hat	0.92	0.84	1.00	
CelebA		Wearing Earrings	Wearing Earrings	0.83	0.70	0.95	
LFW		Wearing Necktie	Wearing Necktie	0.91	0.84	0.98	
CelebA		Wearing Lipstick	Wearing Lipstick	0.95	0.90	1.00	
LFW		Eyeglasses	No Eyewear	0.98	0.98	0.98	
CelebA			Eyeglasses	0.90	0.80	1.00	
CelebA	Other	Attractive	Attractive	1.00	1.00	1.00	
		Total			0.91	0.87	0.94

Table 3.3.: Attribute label analysis of LFW based on the ground truth of three human evaluators. The annotation quality is reported in terms of accuracy, precision, and recall.

Class	Acc	Precision	Recall	Class	Acc	Precision	Recall
Male	0.89	0.96	0.84	Eyes Open	0.73	0.96	0.66
Asian	0.86	0.74	0.97	Big Nose	0.75	0.54	0.93
White	0.74	0.98	0.66	Pointy Nose	0.80	0.82	0.79
Black	0.91	0.84	0.98	Big Lips	0.73	0.56	0.85
Baby	0.54	0.08	1.00	Mouth Closed	0.86	0.82	0.89
Child	0.55	0.10	1.00	Mouth Slightly Open	0.79	0.88	0.75
Youth	0.56	0.14	0.88	Mouth Wide Open	0.93	0.88	0.98
Middle Aged	0.67	0.90	0.62	Teeth Not Visible	0.86	0.78	0.93
Senior	0.87	0.94	0.82	No Beard	0.69	1.00	0.62
Black Hair	0.78	0.88	0.73	Goatee	0.62	0.24	1.00
Blond Hair	0.91	0.84	0.98	Round Jaw	0.77	0.76	0.78
Brown Hair	0.70	0.60	0.75	Double Chin	0.66	0.34	0.94
Bald	0.74	0.50	0.96	Wearing Hat	0.69	0.40	0.95
No Eyewear	0.91	0.98	0.86	Oval Face	0.59	0.78	0.57
Eyeglasses	0.91	0.88	0.94	Square Face	0.55	0.12	0.86
Sunglasses	0.86	0.72	1.00	Round Face	0.81	0.72	0.88
Moustache	0.84	0.72	0.95	Color Photo	0.57	1.00	0.54
Smiling	0.87	0.80	0.93	Posed Photo	0.64	0.32	0.89
Frowning	0.61	0.22	1.00	Attractive Man	0.62	0.26	0.93
Chubby	0.53	0.16	0.62	Attractive Woman	0.75	0.50	1.00
Blurry	0.69	0.90	0.63	Indian	0.65	0.32	0.94
Harsh Lighting	0.64	0.92	0.59	Gray Hair	0.89	0.94	0.85
Flash	0.73	0.66	0.77	Bags Under Eyes	0.75	0.76	0.75
Soft Lighting	0.75	0.66	0.80	Heavy Makeup	0.88	0.76	1.00
Outdoor	0.83	0.82	0.84	Rosy Cheeks	0.63	0.30	0.88
Curly Hair	0.51	0.02	1.00	Shiny Skin	0.66	0.44	0.79
Wavy Hair	0.50	0.08	0.50	Pale Skin	0.82	0.90	0.78
Straight Hair	0.60	0.78	0.54	5 o Clock Shadow	0.59	0.18	1.00
Receding Hairline	0.75	0.62	0.84	Strong Nose-Mouth Lines	0.86	0.88	0.85
Bangs	0.54	0.08	1.00	Wearing Lipstick	0.81	0.64	0.97
Sideburns	0.61	0.40	0.69	Flushed Face	0.61	0.28	0.82
Fully Visible Forehead	0.79	1.00	0.70	High Cheekbones	0.81	0.70	0.90
Partially Visible Forehead	0.82	0.80	0.83	Brown Eyes	0.44	0.46	0.44
Obstructed Forehead	0.62	0.24	1.00	Wearing Earrings	0.79	0.58	1.00
Bushy Eyebrows	0.64	0.42	0.75	Wearing Necktie	0.76	0.66	0.83
Arched Eyebrows	0.79	0.80	0.78	Wearing Necklace	0.61	0.22	1.00
Narrow Eyes	0.69	0.46	0.85	Total	0.72	0.61	0.84

Table 3.4.: Attribute label analysis of CelebA based on the ground truth of three human evaluators. The annotation quality is reported in terms of accuracy, precision, and recall.

Class	Acc	Precision	Recall
5 o Clock Shadow	0.85	0.74	0.95
Arched Eyebrows	0.89	0.92	0.87
Attractive	0.81	0.74	0.86
Bags Under Eyes	0.80	0.80	0.80
Bald	0.84	0.68	1.00
Bangs	0.75	0.50	1.00
Big Lips	0.73	0.84	0.69
Big Nose	0.79	0.86	0.75
Black Hair	0.87	0.96	0.81
Blond Hair	0.94	0.94	0.94
Blurry	0.88	0.78	0.98
Brown Hair	0.90	0.88	0.92
Bushy Eyebrows	0.81	0.78	0.83
Chubby	0.83	0.66	1.00
Double Chin	0.76	0.58	0.91
Eyeglasses	0.96	0.92	1.00
Goatee	0.93	0.94	0.92
Gray Hair	0.98	0.98	0.98
Heavy Makeup	0.90	0.92	0.88
High Cheekbones	0.88	0.86	0.90
Male	1.00	1.00	1.00
Mouth Slightly Open	0.90	0.88	0.92
Mustache	0.95	0.94	0.96
Narrow Eyes	0.86	0.82	0.89
No Beard	0.91	1.00	0.85
Oval Face	0.62	0.92	0.58
Pale Skin	0.85	0.92	0.81
Pointy Nose	0.83	0.94	0.77
Receding Hairline	0.66	0.38	0.86
Rosy Cheeks	0.78	0.70	0.83
Sideburns	0.84	0.88	0.81
Smiling	0.94	0.92	0.96
Straight Hair	0.83	1.00	0.75
Wavy Hair	0.82	0.66	0.97
Wearing Earrings	0.93	0.88	0.98
Wearing Hat	1.00	1.00	1.00
Wearing Lipstick	0.91	0.90	0.92
Wearing Necklace	0.86	0.80	0.91
Wearing Necktie	0.85	0.72	0.97
Young	0.75	0.52	0.96
Total	0.85	0.83	0.89

3.3. Investigating Bias in Face Recognition

3.3.1. Introduction

Recent works [Orc16; AZN18; FPO02; Phi+11; BG18; Gar+16] showed that current face recognition solutions possess demographic-biases leading to discriminatory performance differences based on the user’s demographics. Driven by these findings, several approaches were proposed to mitigate demographics-bias in face recognition technologies. This was achieved through adversarial learning [GLJ19; Lia+19], margin-based approaches [WD19; Hua+18], data augmentation [Wan+19; Kor+19; Yin+19], metric-learning [Ter+20i], or score normalization [Ter+20f]. However, the strong research focus on demographic-bias does only tackle a minor proportion of all possible discriminatory effects. Without knowing the influence of all facial attributes on the face recognition performance, an accurate and non-discriminatory face recognition system might not be possible.

This section is based on Terhörst et al. [Ter+21b] and aims at the necessity of investigating the face recognition bias based on a wide range of attributes beyond demographics to answer RQ2. To be precise, we analyse the performance differences of two popular face recognition models (FaceNet [SKP15] and ArcFace [Den+19]) with regard to 47 attributes. The experiments are conducted on the MAAD-Face⁵ database [Ter+20b; Cao+18] from Section 3.2.2. It consists of over 3.3M face images with over 120M high-quality attribute annotations. For the experiments, several decision thresholds are taken into account to cover a wide range of applications. To prevent misinterpretations of the results origin from testing data with (a) unbalanced label distributions or (b) attribute correlations, we (a) introduce control groups to derive a validity value for the recognition performance in the presence of a specific attribute and (b) analyse the pairwise correlations of the attribute annotations. While (a) allows us to quantify results that arise from unbalanced testing data and prevent falsified statements about the attribute-related bias, (b) emphasize if an attribute bias might originate from a different (correlated) attribute. Besides a detailed analysis, we present a visual summary that states the performance difference between samples with and without a specific attribute over the validity of the results. This aims to present the results in a compact and simply understandable manner.

The results support the findings of previous works stating that face recognition systems have to deal with demographic-biases. However, the results demonstrate that also many of the non-demographic attributes strongly affect the recognition performance, such as accessories, hairstyles and -colors, face shapes, or facial anomalies. Investigating two face recognition models that differ only in the loss function used during training, we showed the effect of the underlying training principles on recognition. While the

⁵<https://github.com/pterhoer/MAAD-Face>

triplet-loss based FaceNet model showed attribute-related performance differences that are relatively constant on several decision thresholds, the angular margin based ArcFace model showed performance differences that are often dependent on the used decision threshold. The many performance differences affected by attributes could be explained through the attribute's relation to the visibility of a face, the temporal variability, and the degree of abnormality. However, our experiment also reveals many surprising results that future work have to address. Our findings strongly demonstrate the need for further advances in making face recognition systems more robust, explainable, and fair. This section demonstrates the strong need for the development of generalized bias-mitigating face recognition solutions motivating the proposed solutions in Chapter 4.

3.3.2. Related Work

The phenomena of bias in face biometrics were found in several disciplines such as presentation attack detection [Fan+20], the estimation of facial characteristics [Ter+19d; DDB18], and the assessment of face image quality [Ter+20g]. In general, one of the main reason for bias might be the induction of non-equally distributed classes in training data [Kor+19; Hua+18] that leads to differences in the recognition performance and thus, might have an unfair impact, e.g. on specific subgroups of the population. Previous works on bias in face recognition [Dro+20] mainly focused on the influence of demographics. However, in Section 3.5 we demonstrated that more (non-demographic) characteristics are stored in face templates that might have an impact on the face recognition performance. In the following, we will shortly discuss related works on estimating bias in face biometrics. A overview of related work on bias-mitigation is shown in Section 4.2.

Estimating Bias in Face Recognition

In recent years, several works have been published that demonstrated the influence of demographics on commercial and open-sources face recognition algorithms. Studies [JBS15; MYM18; DNJ18; Sri+19b] analysing the impact of age demonstrated a lower biometric performance on faces of children. Studies [Ver+19; AZB20; AB20] analysing the effect of gender on face recognition showed that the recognition performance of females is weaker than the performance on male faces. Experiments without unbalanced data distributions and with an unbalanced towards female faces resulted in similar results [AZB20]. In [AB20], experiments with a PCA-decomposition showed that females faces are intrinsically more similar than male ones. Research analysing the impact of the user's ethnicity showed faces of ethnicities which were under-represented in the training process

perform significantly weaker. The same was found for darker-skinned cohorts in general [Kri+20].

More recent studies [Kla+12; Sri+19a; Coo+19; HT19; Cav+19; HSV19a; Rob+20; GNH19b; Bal+20] focused on jointly investigating the effects of user demographics on face recognition. These studies showed that the effects lead to an exponential face recognition error increase when facing the same biased race, gender, and age factors [HSV19a]. Particular attention deserves the Face Recognition Vendor Test (FRVT) [GNH19b], a large-scale benchmark of commercial algorithms analysing the face recognition performance with regards to demographics. They consistently elevated false positives for female subjects and subjects at the outer ends of the age spectrum.

How This Work Contributes to State-of-the-Art

So far, the majority of research in estimating bias in face recognition focused on demographic factors, such as age, gender, and race. However, to achieve a generally accurate and fair face recognition model, it is necessary to know all potential origins of performance differences. Therefore, this section aims at closing this knowledge gap by analysing the performance differences on a much wider attribute range than previous works. More precisely, this work investigates the influence of 47 attributes on the face recognition performance of two popular face embeddings.

3.3.3. Experiments on Measuring Differential Performance

Database and Considered Attributes

To get reliable statements on the effect of different attributes on face recognition, we need a database that (a) provides a high number of face images with (b) many attribute annotations of (c) high quality. For the experiments, we choose the publicly available MAAD-Face⁶ database [Ter+20b; Cao+18] since this database fulfils our experimental requirements. MAAD-Face provides over 120M high-quality attribute annotations of 3.3M face images of over 9k individuals. It provides annotations for 47 distinct attributes of various kinds such as demographics, skin types, hair-styles and -colors, face geometry, annotations for the periocular, mouth, and nose area, as well as annotations for accessories. An exact list of the annotation attributes can be obtained from Table 3.5 and 3.6. The attribute annotations proofed to have a higher quality than comparable face annotation databases [Ter+20b].

⁶<https://github.com/pterhoer/MAAD-Face>

Face Recognition Models

For the experiments, we use two popular face recognition models, FaceNet [SKP15] and ArcFace [Den+19]. To create a face embedding for a given face image, the image has to be aligned, scaled, and cropped. Then, the preprocessed image is passed to a face recognition model to extract the embeddings. For FaceNet, the preprocessing is done as described in [KS14]. To extract the embeddings, a pretrained model⁷ was used. For ArcFace, the image preprocessing was done as described in [Guo+18] and a pretrained model⁸ is used, which is provided by the authors of ArcFace. Both models use a ResNet-100 architecture and were trained on the MS1M database [Guo+16]. The identity verification is done by comparing two embeddings using cosine-similarity.

Evaluation Metrics

The face verification performance is reported in terms of (a) false non-match rates (FNMR) at a fixed false match rate (FMR) and (b) equal error rates (EER). The EER equals the FMR at the threshold where $FMR = 1 - FNMR$ and is well known as a single-value indicator of the verification performance. The used error rates are specified for biometric verification evaluation in the international standard [16]. In the experiments, the face verification performance is reported on three operating points to cover a wide range of potential applications. This includes EER, as well as, the FNMR at 10^{-3} and 10^{-4} FMR as recommended by the best practice guidelines for automated border control of the European Boarder Guard Agency Frontex [Fro17]. For each operating point and attribute, the verification performance is computed on all samples with positive and all samples with negative annotations. This will allow to compare the performance differences of face embeddings regarding binary attributes, such as bald vs non-bald faces.

Control Groups

During the experiments, the number of testing samples with positive and negative labels might be significantly different. To prevent misleading conclusions from such unbalanced annotation distributions, we introduce positive and negative control groups for each attribute. For each attribute, six positive and negative control groups are created by randomly selecting samples from the database. This control group creation is done such that the synthetic control groups have the same number of samples as their positive and negative counterparts.

⁷<https://github.com/davidsandberg/facenet>

⁸<https://github.com/deepinsight/insightface>

Comparing the verification performance of the positive and negative control groups allow us to state the validity of our (real) attribute performance. If the performances of the negative and positive control groups is very similar, the (real) attribute recognition performance is treated as valid. In this case, the unbalanced testing data distribution shows no effect on the performance. If the relative performance of the control groups differ strongly, the recognition performance might be significantly affected from unbalanced distribution of the positively and negatively annotated samples. Consequently, the (real) attribute recognition performance might be affected as well and statements about the influence of this attribute on the recognition are of low validity. In the experiments, the validity val of an attribute a

$$val(a) = 1 - \frac{err_{control}^{(+)}(a)}{err_{control}^{(-)}(a)}, \quad (3.4)$$

is defined over the relative performance differences between the control groups. The terms $err_{control}^{(+)}(a)$ and $err_{control}^{(-)}(a)$ represent the recognition errors of the positive (+) and the negative (−) control groups of attribute a . For the experiments, we consider attributes with a validity of < 0.9 as *not valid*. However, we will also present the performance differences with the corresponding validity values such that the readers are able to choose more suitable validity threshold for their applications.

Investigations

To analyse the influence of different attributes on the recognition performance of two popular face recognition models, the investigations are divided into five parts.

1. A correlation analysis between the attribute annotations is performed to emphasizes if an attribute bias might originate from correlated attribute annotations.
2. For each attribute, the recognition performance of its positively- and negatively-labelled attribute groups are compared to investigate the influence of this attribute on the recognition performance. The results are discussed in the context of the corresponding validity values to avoid misinterpretations occurring from unbalanced testing data.
3. A visual summary is provided that relates the impact of the attributes on the face recognition systems to the validity of the results. This aims at providing an compact and easily-understandable overview of the findings of this work.

-
4. We discuss possible explanations causing the performance differences and highlight behaviours of the face recognition systems that remain unclear.
 5. Lastly, we use the observations to derive future research challenges for face recognition systems.

3.3.4. Results

Investigating the Correlation of Attribute Annotations

To understand the quality of the used labels and potential biases in the attribute space, Figure 3.14 shows a selection of specific attribute-label correlations. The attributes are chosen to show the 15 most positive and negative pairwise correlations. It can be seen that *Wearing Lipstick*, *Wearing Earrings*, *Heavy Makeup*, *Young*, and *Attractive* correlates highly positively with *Arched Eyebrows*, *Wavy Hair*, and *Rosy Cheeks*. In contrast, these attributes correlates negatively with *Square Face*, *Male*, and *Bags Under Eyes*. These correlations have to be considered when comparing the performance differences for the different attributes. However, the correlation matrix also approves the quality of some labels that semantically excludes each other. For instance, *5 o Clock Shadow* negatively correlates with *No Beard* and *Eyeglasses* negatively correlates with *No Eyewear*.

The Impact of Facial Attributes on Recognition

The main contribution of this work is an analysis of the effect of 47 distinct attributes on two popular face recognition models. This aims at investigating model biases. For each attribute, the face verification performance is calculated on positively-labelled samples, as well as on negatively-labelled samples. This is done on three operating points as explained in Section 3.3.3. The relative performance between the positive and negative groups allows to investigate potential biases of the face recognition model towards the analysed attribute. To determine if performance differences result from unbalanced data distributions we introduced control groups as explained in Section 3.3.3.

In Tables 3.5, 3.6, 3.7, 3.8, 3.9 and 3.10 the performance of the positive and negative class is shown for each attribute. The performance of the annotated data is referred as Real while the performance of the control groups is referred as Control. The relative performance (Rel. Perf.) shows the relative performance difference between the positive and negative attribute classes. If the relative performance between the control classes are below 10% ($val \geq 0.9$), the result is considered as *valid* (green highlighting). Otherwise, the result is considered as *not valid* indicated by a grey highlighting. Positive values for the relative performance of an attribute represents a positive effect of the attribute on

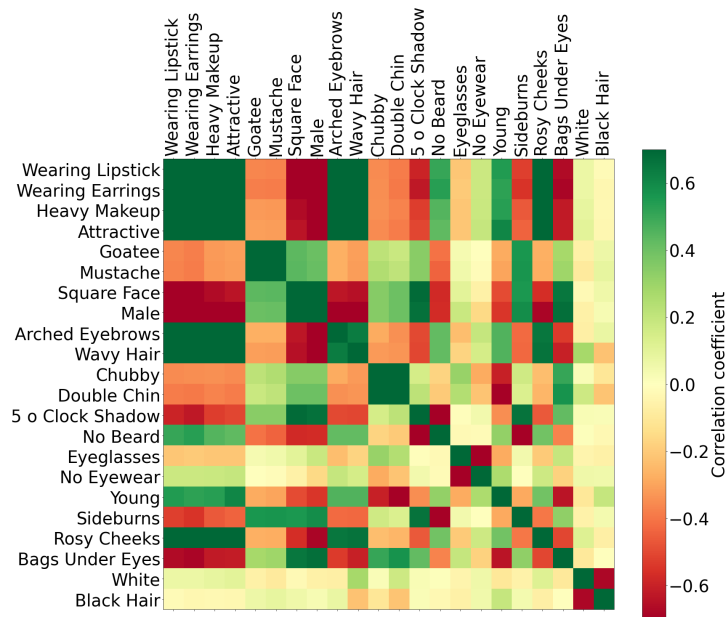


Figure 3.14.: Compressed annotation correlations of the used MAAD-Face database. The attributes are chosen such that the 15 most positive and negative pairwise are visible. Green indicate positive correlations, while red indicate a negative correlation. The correlation is based on the Pearson coefficient. When interpreting the results from Section 3.3.4 highly-correlated attributes should be considered to prevent misinterpretations.

the face recognition performance. Negative values indicate a negative influence of the attribute on the recognition performance. In the following, we present the results of our study on bias on FaceNet and ArcFace embeddings.

Biases in FaceNet Embeddings The results of our attribute-related study on performance differences of the FaceNet model are shown in Table 3.5, 3.6 and 3.7.

Previous works focused on performance differences affected by the user’s demographics. The results on FaceNet confirms the observations of these works. Demographics strongly affect the recognition performance. One of the strongest impacts on FaceNet is observed for ethnicities. For the investigated FaceNet model, *Asian* and *Black* faces lead to significantly lower recognition rates than *White* faces. Also *Young* ones perform significantly weaker than e.g. *Middle-aged* faces. Concerning gender, we observe that *Male* face perform better

then *Female* ones. These findings are intensively discussed in previous works [AZB20; AB20]. However, the experimental results show that there are many more aspects that strongly affect the recognition performance.

One factor leading to performance differences is the user's hair. While *Bald* faces and *Receding Hairlines* lead to an improved recognition performance, *Wavy Hair* styles or *Bangs* are observed to degrade the performance. This can be explained by the visibility of the face. In general, *Wavy Hair* and *Bangs* are more likely to cover parts of the face while *Bald* faces or faces with *Receding Hairlines* do not occlude part of the faces.

A contrarily observation can be made for facial hair. Faces with *No-Beard* perform worse than faces with a beard, such as a *5 o Clock Shadow*. A reason for this can be that people might consider their beard as a part of their identity and preserve it such that it can be used for recognition. However, it also has to be considered that this performance difference might come from the correlation of beards with male faces.

Also the color of the hair has an impact on the FaceNet embeddings. While *Blond Hair* shows a strongly degraded face recognition performance, *Gray Hair* leads to the strongest performances.

The results indicate that the shape of a face only has a minor impact on the face recognition performance. For *Oval Faces*, no significant differences to non-oval faces could be observed. Although, a positive effect on the recognition performance is shown for *Square Faces*, in Section 3.3.4 a strong correlation between *Square Face* and *Male* was shown. This might explain the behaviour.

Faces with *High Cheekbones*, *Double Chins*, and *Chubby* faces also perform better for FaceNet features than the inverted counterparts. Probably because these properties provide additional information that can be used for recognition. In contrast to this, an *Obstructed Forehead* strongly degrades the recognition performance while a *Fully Visible Forehead* provides additional (uncovered) information that supports the recognition process.

Anomalous properties in the periocular area, such as *Bags Under Eyes*, *Bushy Eyebrows*, or *Arched Eyebrows*, lead to better recognition rates compared to face images without these attributes. The same goes for *Big Nose* and *Pointy Nose*.

The reason that *Smiling* and a *Mouth Closed* lead to an improvement in the recognition performance might be explainable through the issue of facial expressions. In [Dam+18d], the opposite effect was already shown by demonstrating that crazy faces result in comparably low comparison scores.

Interestingly, accessories have a strong influence on the recognition performance of FaceNet. *Wearing Hat*, *Wearing Earrings*, or *Eyeglasses* degrade the face recognition performance significantly and might be explained by the fact that these accessories cover discriminative parts of the face. However, manually checking some images of MAAD-Face does not support this hypothesis. Consequently, we conclude that FaceNet produces

less-reliable embeddings when facing such attributes.

Biases in ArcFace Embeddings The results of our attribute-related study on performance differences of the ArcFace model are shown in Table 3.8, 3.9 and 3.10.

Previous works focused on performance differences affected by the user’s demographics. The results on ArcFace partly confirms the observations of these works. Demographics sometimes have a strongly effect the recognition performance. For the investigated ArcFace model, *Young* faces perform weaker than *Middle-aged* or *Senior* faces. Interestingly, the intensively discussed gender bias is strongly dependent on the used decision threshold. Especially for lower FMRs the performance differences between *Male* and *Female* increases. Concerning the ethnic-bias on the ArcFace model, we are not able to confirm the observations from previous works. For *White* faces, the performance is significantly higher than for non-white faces. For *Asian* and *Black* faces, a strong degradation in the recognition performance can be observed. However, we have to consider this results as *not valid*, since we can observe strong performance differences on the control groups. This indicates that these results are strongly influenced by the unbalanced training data.

Similar to FaceNet, the user’s hair shows to have a significant impact on the face recognition performance. While, *Receding Hairlines*, *Wavy Hair* and *Sideburns* supports the recognition process, faces with *Bangs* show a strong degradation. Again, the performance differences on ArcFace show to be threshold-dependent. For *Wavy Hair*, the positive effect on face recognition vanishes for lower FMRs, and for *Bangs*, the negative effect increases drastically for higher security settings.

Also the color of the user’s hair have an impact on the recognition performance. *Gray Hair* performs significantly above average, while *Black Hair* performs significantly below average. *Blond Hair* and *Brown Hair* lead to performance differences depending on the decision threshold. For high FMRs, *Blond Hair* improves the recognition performance, while for lower FMRs, the recognition performance changes to below-average. For faces with *Brown Hair*, the positive effect on recognition vanishes for lower FMRs.

The effect of wearing a beard on the performance of ArcFace is similar to FaceNet. Having *No Beard* decreases the recognition performance and having a beard, such as a *5 o Clock Shadow*, enhances the recognition. These effects magnify for lower FMRs.

On contrast to FaceNet, the face shape determines the recognition performance of ArcFace. Both, *Oval Faces* and *Square Faces* have positive effect on the recognition performance, which is dependent on the utilized decision threshold. *Round Faces* show a strongly degraded recognition. However, a large fraction of this performance differences can be explained by the unbalanced data distribution and thus, we have to neglect the results for *Round Faces*.

Similar to FaceNet, *High Cheekbones*, *Double Chin*, *Chubby*, and a *Fully Visible Forehead* lead to improved face recognition performances. While a *Fully Visible Forehead* refers to no partial occlusions of the face that might negatively infer, the other attributes provide anomalous points that might help for recognition.

Surprisingly, faces with *Brown Eyes* perform drastically weaker than faces with non-brown eyes. For *Bags Under Eyes*, *Bushy Eyebrows*, and *Arched Eyebrows*, an improved face recognition performance can be observed. These attributes can be treated as anomalies and thus, can support the recognition process. The same goes for *Big Nose* and *Pointy Nose*.

Interestingly, a *Smiling Face* strongly enhanced the face recognition performance. This might be explained by (a) the fact that many training databases contain *Smiling* faces and (b) *Smiling* faces often appear on posed photos.

Similar to FaceNet, accessories have a strong impact on the performance differences of ArcFace. While having *Heavy Makeup*, such as *Wearing Lipstick*, improves the recognition, faces with *Eyeglasses* or *Wearing Hat* lead to strong degradations in the face recognition performance. A reason for this might be that people wearing *Heavy Makeup* perceive this is part of their identity and will wear this makeup more permanently. Consequently, a person in the training data might either have no or only *Heavy Makeup* images. On the other side, people tend to change their *Eyeglasses* or (*Wearing*) *Hats* more frequently. Moreover, these attributes might lead to partial occlusions of the face leading to less identity-information available and thus, to a degraded face recognition performance.

Lastly, the results show that the recognition performance is higher on faces that are perceived as *Attractive* compared to faces that are labelled as not attractive. However, this might be explained by the positive correlation with attribute that showed the same positive effect, such as *Lipstick*, *Heavy Makeup*, and *Arched Eyebrows*.

Summary

To provide an overview of the findings, Figure 3.15 shows the relative performance differences on FaceNet and ArcFace features based on the investigated attributes. The shown relative performance is based on the FMR at 10^{-3} FNMR as recommended by the European Boarder Guard Agency Frontex [Fro17]. The validity describes the performance difference between the positive and negative attribute-related control groups as shown in Equation 3.4. An attribute performance with a validity of less than 90% is considered as *not valid* (grey area) since the unbalanced data annotations might affect the reported performance. The red area indicates that the recognition performance of the positive attribute class is significantly weaker than the performance of the negative class. In contrast, the green area indicates a significant improvement of recognition performance of

Table 3.5.: FaceNet - Part 1/3. Face recognition performance based on several attributes.

Category	Attribute	Class	EER		FNMR@FMR=10 ⁻³		FNMR@FMR=10 ⁻⁴	
			Real	Control	Real	Control	Real	Control
Demographics	Male	Positive	6.64%	6.49%	33.28%	32.51%	53.64%	52.44%
		Negative	7.87%	6.46%	42.47%	32.40%	62.55%	52.32%
		Rel. Perf.	15.56%	-0.42%	21.63%	-0.35%	14.24%	-0.21%
	Young	Positive	6.91%	6.46%	39.39%	32.39%	60.37%	52.30%
		Negative	5.73%	6.47%	28.93%	32.37%	48.97%	52.18%
		Rel. Perf.	-20.58%	0.12%	-36.19%	-0.08%	-23.27%	-0.22%
	Middle_Aged	Positive	5.41%	6.33%	28.77%	31.70%	48.75%	51.25%
		Negative	6.96%	6.48%	36.77%	32.52%	57.70%	52.45%
		Rel. Perf.	22.29%	2.33%	21.74%	2.52%	15.52%	2.28%
	Senior	Positive	6.01%	6.23%	30.26%	31.19%	50.52%	50.52%
		Negative	6.69%	6.49%	34.19%	32.54%	54.58%	52.53%
		Rel. Perf.	10.16%	3.93%	11.50%	4.16%	7.44%	3.82%
	Asian	Positive	11.16%	5.91%	69.46%	29.52%	88.48%	48.20%
		Negative	6.33%	6.49%	31.91%	32.55%	51.27%	52.54%
		Rel. Perf.	-76.30%	8.88%	-117.66%	9.33%	-72.58%	8.28%
	White	Positive	5.97%	6.48%	31.28%	32.51%	50.15%	52.50%
		Negative	7.51%	6.44%	46.82%	32.16%	69.44%	51.94%
		Rel. Perf.	20.54%	-0.56%	33.18%	-1.11%	27.79%	-1.07%
Black	Positive	8.85%	6.02%	52.50%	30.20%	73.61%	49.32%	
	Negative	6.61%	6.49%	33.47%	32.54%	53.34%	52.52%	
	Rel. Perf.	-33.98%	7.14%	-56.89%	7.19%	-37.99%	6.09%	
Skin	Rosy_Cheeks	Positive	1.29%	5.46%	3.76%	26.05%	9.46%	42.72%
		Negative	7.36%	6.48%	37.03%	32.51%	57.65%	52.47%
		Rel. Perf.	82.42%	15.80%	89.86%	19.87%	83.59%	18.59%
Shiny_Skin	Positive	6.08%	6.41%	36.43%	32.05%	57.46%	51.83%	
	Negative	7.90%	6.47%	41.33%	32.37%	62.43%	52.29%	
	Rel. Perf.	23.06%	0.85%	11.86%	0.99%	7.97%	0.88%	
Hair	Bald	Positive	5.10%	6.13%	30.52%	30.69%	52.37%	49.93%
		Negative	6.70%	6.49%	34.13%	32.54%	54.47%	52.52%
		Rel. Perf.	23.89%	5.55%	10.59%	5.70%	3.85%	4.94%
	Wavy_Hair	Positive	7.55%	6.46%	40.97%	32.29%	60.69%	52.10%
		Negative	6.82%	6.48%	34.23%	32.50%	54.84%	52.48%
		Rel. Perf.	-10.68%	0.34%	-19.69%	0.65%	-10.65%	0.73%
	Receding_Hairline	Positive	4.93%	6.43%	26.02%	32.06%	44.95%	51.75%
		Negative	7.35%	6.47%	39.92%	32.46%	61.12%	52.43%
		Rel. Perf.	32.90%	0.74%	34.82%	1.23%	26.46%	1.29%
	Bangs	Positive	6.82%	6.34%	45.53%	31.63%	69.28%	51.14%
		Negative	6.43%	6.49%	32.02%	32.54%	51.86%	52.52%
		Rel. Perf.	-5.99%	2.25%	-42.17%	2.79%	-33.59%	2.62%
	Sideburns	Positive	6.68%	6.46%	34.33%	32.34%	54.08%	52.23%
		Negative	6.75%	6.49%	35.47%	32.52%	55.96%	52.46%
		Rel. Perf.	1.04%	0.43%	3.24%	0.53%	3.35%	0.44%
	Black_Hair	Positive	7.13%	6.42%	42.35%	32.06%	65.73%	51.73%
		Negative	6.20%	6.48%	32.46%	32.49%	52.06%	52.47%
		Rel. Perf.	-15.04%	1.02%	-30.47%	1.34%	-26.26%	1.40%
Blond_Hair	Positive	9.63%	6.34%	52.00%	31.66%	71.71%	51.18%	
	Negative	6.45%	6.48%	32.63%	32.52%	52.96%	52.48%	
	Rel. Perf.	-49.35%	2.17%	-59.37%	2.66%	-35.41%	2.48%	
Brown_Hair	Positive	7.40%	6.45%	39.73%	32.26%	59.12%	52.06%	
	Negative	6.19%	6.47%	35.13%	32.41%	57.09%	52.30%	
	Rel. Perf.	-19.52%	0.26%	-13.08%	0.49%	-3.55%	0.48%	
Gray_Hair	Positive	5.32%	6.29%	26.00%	31.50%	44.11%	50.99%	
	Negative	6.72%	6.49%	34.60%	32.54%	55.25%	52.52%	
	Rel. Perf.	20.83%	3.05%	24.83%	3.20%	20.17%	2.90%	

Table 3.6.: FaceNet - Part 2/3. Face recognition performance based on several attributes.

Category	Attribute	Class	EER		FNMR@FMR=10 ⁻³		FNMR@FMR=10 ⁻⁴	
			Real	Control	Real	Control	Real	Control
Beard	No_Beard	Positive	7.20%	6.48%	37.97%	32.49%	58.83%	52.44%
		Negative	6.13%	6.40%	31.07%	31.94%	51.01%	51.60%
		Rel. Perf.	-17.53%	-1.38%	-22.20%	-1.74%	-15.33%	-1.62%
	Mustache	Positive	6.45%	4.93%	50.77%	22.55%	73.71%	36.74%
		Negative	6.90%	6.48%	35.54%	32.52%	56.12%	52.49%
		Rel. Perf.	6.41%	23.94%	-42.88%	30.68%	-31.34%	30.01%
	5_o_Clock_Shadow	Positive	6.16%	6.38%	30.98%	31.87%	50.01%	51.53%
		Negative	7.49%	6.48%	39.91%	32.46%	60.86%	52.39%
		Rel. Perf.	17.78%	1.54%	22.37%	1.83%	17.83%	1.64%
	Goatee	Positive	2.59%	4.69%	18.78%	20.11%	38.17%	32.98%
		Negative	6.92%	6.49%	35.49%	32.54%	56.11%	52.55%
		Rel. Perf.	62.59%	27.63%	47.09%	38.19%	31.97%	37.24%
Face Geometry	Oval_Face	Positive	8.14%	6.40%	45.16%	31.97%	64.96%	51.64%
		Negative	8.26%	6.46%	45.11%	32.30%	67.44%	52.08%
		Rel. Perf.	1.45%	1.01%	-0.11%	1.00%	3.68%	0.84%
	Square_Face	Positive	6.32%	6.48%	31.37%	32.49%	51.25%	52.44%
		Negative	7.81%	6.47%	41.51%	32.43%	61.90%	52.38%
		Rel. Perf.	19.13%	-0.12%	24.42%	-0.16%	17.20%	-0.12%
	Round_Face	Positive	16.53%	4.52%	88.11%	19.03%	93.33%	31.40%
		Negative	5.31%	6.49%	27.06%	32.52%	45.05%	52.46%
		Rel. Perf.	-211.14%	30.27%	-225.65%	41.49%	-107.17%	40.14%
	Double_Chin	Positive	5.45%	6.44%	26.28%	32.15%	44.43%	51.85%
		Negative	7.09%	6.48%	38.20%	32.51%	59.43%	52.46%
		Rel. Perf.	23.05%	0.71%	31.20%	1.10%	25.24%	1.18%
	High Cheekbones	Positive	5.99%	6.46%	33.69%	32.27%	53.73%	52.10%
		Negative	8.10%	6.47%	41.66%	32.41%	62.29%	52.32%
		Rel. Perf.	26.11%	0.20%	19.13%	0.43%	13.73%	0.43%
	Chubby	Positive	5.11%	6.38%	26.98%	31.81%	47.76%	51.48%
		Negative	6.85%	6.49%	36.65%	32.54%	57.76%	52.49%
		Rel. Perf.	25.35%	1.62%	26.38%	2.23%	17.31%	1.92%
Obstructed Forehead	Positive	8.85%	6.11%	60.01%	30.67%	80.51%	49.92%	
	Negative	6.02%	6.49%	31.14%	32.52%	50.70%	52.50%	
	Rel. Perf.	-46.87%	5.75%	-92.69%	5.69%	-58.79%	4.91%	
Fully_Visible Forehead	Positive	5.47%	6.48%	28.25%	32.46%	47.36%	52.35%	
	Negative	7.82%	6.45%	44.34%	32.28%	66.70%	52.09%	
	Rel. Perf.	30.01%	-0.43%	36.29%	-0.55%	28.99%	-0.49%	
Periocular	Brown_Eyes	Positive	7.54%	6.48%	42.04%	32.44%	63.89%	52.36%
		Negative	6.12%	6.36%	33.59%	31.83%	52.03%	51.50%
		Rel. Perf.	-23.28%	-1.81%	-25.15%	-1.94%	-22.78%	-1.67%
	Bags_Under_Eyes	Positive	5.90%	6.45%	31.51%	32.31%	52.50%	52.16%
		Negative	8.03%	6.47%	42.47%	32.42%	62.85%	52.31%
		Rel. Perf.	26.47%	0.36%	25.79%	0.37%	16.48%	0.29%
	Bushy_Eyebrows	Positive	5.66%	6.47%	29.86%	32.36%	49.67%	52.29%
		Negative	7.26%	6.48%	37.79%	32.51%	58.28%	52.45%
		Rel. Perf.	22.03%	0.23%	21.00%	0.44%	14.77%	0.31%
	Arched_Eyebrows	Positive	5.99%	6.46%	33.71%	32.28%	52.99%	52.06%
		Negative	7.59%	6.48%	38.64%	32.47%	59.96%	52.40%
		Rel. Perf.	21.10%	0.37%	12.75%	0.58%	11.62%	0.64%

Table 3.7.: FaceNet - Part 3/3. Face recognition performance based on several attributes.

Category	Attribute	Class	EER		FNMR@FMR=10 ⁻³		FNMR@FMR=10 ⁻⁴		
			Real	Control	Real	Control	Real	Control	
Mouth	Mouth Closed	Positive	5.25%	5.99%	27.84%	29.97%	46.77%	48.87%	
		Negative	7.05%	6.41%	46.08%	32.00%	68.38%	51.71%	
		Rel. Perf.	25.49%	6.53%	39.60%	6.34%	31.60%	5.50%	
	Smiling	Positive	6.08%	6.44%	34.06%	32.17%	53.51%	51.91%	
		Negative	8.67%	6.46%	47.88%	32.36%	70.12%	52.23%	
		Rel. Perf.	29.86%	0.28%	28.87%	0.58%	23.68%	0.61%	
Big_Lips	Positive	6.79%	6.45%	39.95%	32.33%	61.39%	52.19%		
	Negative	6.97%	6.47%	34.09%	32.44%	53.99%	52.36%		
	Rel. Perf.	2.58%	0.32%	-17.20%	0.31%	-13.72%	0.31%		
Nose	Big_Nose	Positive	6.28%	6.42%	36.68%	32.04%	59.22%	51.82%	
		Negative	8.40%	6.48%	46.15%	32.43%	67.05%	52.32%	
		Rel. Perf.	25.23%	0.90%	20.52%	1.20%	11.67%	0.94%	
	Pointy_Nose	Positive	6.04%	6.48%	32.67%	32.48%	51.66%	52.44%	
		Negative	7.80%	6.46%	43.90%	32.32%	65.97%	52.22%	
		Rel. Perf.	22.56%	-0.33%	25.57%	-0.49%	21.69%	-0.42%	
Accessories	Heavy Makeup	Positive	6.25%	6.46%	35.96%	32.31%	55.91%	52.17%	
		Negative	7.08%	6.49%	34.76%	32.52%	54.97%	52.48%	
		Rel. Perf.	11.70%	0.46%	-3.44%	0.62%	-1.71%	0.59%	
	Wearing Hat	Positive	9.01%	6.24%	55.58%	31.23%	77.17%	50.65%	
		Negative	6.05%	6.49%	30.40%	32.54%	49.86%	52.55%	
		Rel. Perf.	-48.74%	3.78%	-82.84%	4.03%	-54.77%	3.60%	
	Wearing Earrings	Positive	7.54%	6.46%	41.92%	32.35%	61.83%	52.25%	
		Negative	6.78%	6.48%	33.84%	32.49%	54.34%	52.45%	
		Rel. Perf.	-11.15%	0.25%	-23.89%	0.43%	-13.79%	0.37%	
	Wearing Necktie	Positive	3.99%	6.36%	19.72%	31.65%	37.81%	51.23%	
		Negative	7.53%	6.48%	41.03%	32.52%	62.50%	52.47%	
		Rel. Perf.	47.05%	1.88%	51.93%	2.65%	39.51%	2.37%	
	Wearing Lipstick	Positive	6.74%	6.46%	38.36%	32.37%	58.49%	52.29%	
		Negative	7.01%	6.49%	34.54%	32.51%	54.78%	52.49%	
		Rel. Perf.	3.91%	0.39%	-11.05%	0.44%	-6.78%	0.39%	
	No_Eyewear	Positive	5.77%	6.48%	29.39%	32.53%	48.75%	52.51%	
		Negative	6.64%	6.11%	37.21%	30.64%	63.01%	49.90%	
		Rel. Perf.	13.10%	-6.09%	21.03%	-6.16%	22.63%	-5.24%	
	Eyeglasses	Positive	7.79%	6.33%	43.15%	31.57%	65.99%	51.15%	
		Negative	5.70%	6.49%	29.16%	32.54%	48.78%	52.52%	
		Rel. Perf.	-36.65%	2.51%	-47.99%	3.00%	-35.27%	2.61%	
	Other	Attractive	Positive	6.27%	6.45%	36.28%	32.31%	56.11%	52.10%
			Negative	7.05%	6.49%	34.77%	32.51%	54.96%	52.50%
			Rel. Perf.	11.16%	0.51%	-4.35%	0.61%	-2.09%	0.77%

Table 3.8.: ArcFace - Part 1/3. Face recognition performance based on several attributes.

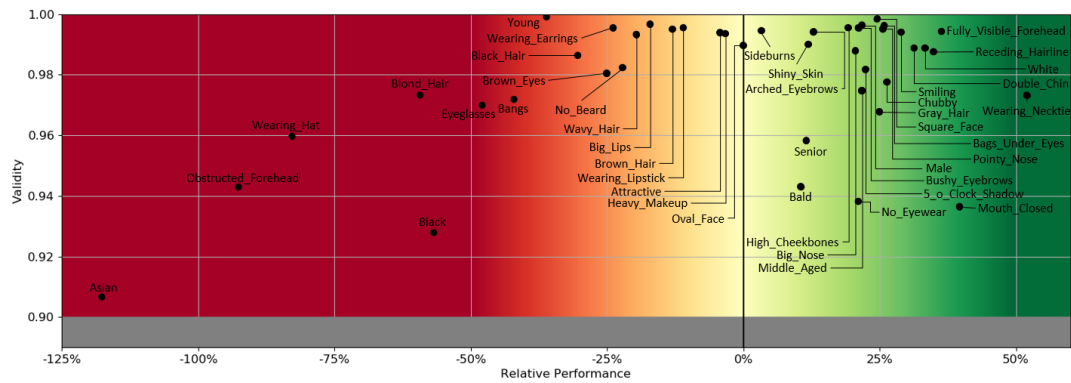
Category	Attribute	Class	EER		FNMR@FMR=10 ⁻³		FNMR@FMR=10 ⁻⁴	
			Real	Control	Real	Control	Real	Control
Demographics	Male	Positive	3.98%	3.98%	7.07%	7.22%	9.71%	10.17%
		Negative	3.82%	3.96%	7.99%	7.20%	12.33%	10.13%
		Rel. Perf.	-4.35%	-0.38%	11.54%	-0.38%	21.24%	-0.37%
	Young	Positive	3.74%	3.97%	7.30%	7.20%	11.08%	10.14%
		Negative	3.70%	3.95%	6.32%	7.17%	8.52%	10.11%
		Rel. Perf.	-0.86%	-0.46%	-15.42%	-0.46%	-30.08%	-0.28%
	Middle_Aged	Positive	3.01%	3.81%	5.05%	6.93%	6.93%	9.80%
		Negative	4.07%	3.98%	7.79%	7.22%	11.36%	10.17%
		Rel. Perf.	26.14%	4.05%	35.20%	4.04%	39.04%	3.56%
	Senior	Positive	2.95%	3.62%	4.52%	6.58%	6.15%	9.38%
		Negative	4.02%	3.98%	7.47%	7.24%	10.62%	10.18%
		Rel. Perf.	26.60%	9.02%	39.44%	9.09%	42.13%	7.87%
	Asian	Positive	7.99%	3.29%	16.68%	6.01%	22.59%	8.69%
		Negative	3.73%	3.98%	6.75%	7.23%	9.61%	10.18%
		Rel. Perf.	-114.49%	17.22%	-147.13%	16.84%	-134.94%	14.60%
	White	Positive	3.27%	3.98%	5.84%	7.23%	8.55%	10.18%
		Negative	5.80%	3.91%	11.69%	7.10%	16.03%	10.01%
		Rel. Perf.	43.50%	-1.66%	50.08%	-1.87%	46.66%	-1.74%
Black	Positive	5.72%	3.40%	10.90%	6.21%	15.02%	8.95%	
	Negative	3.85%	3.98%	7.06%	7.23%	10.11%	10.18%	
	Rel. Perf.	-48.63%	14.53%	-54.43%	14.16%	-48.64%	12.08%	
Skin	Rosy_Cheeks	Positive	0.98%	2.91%	1.17%	5.12%	1.31%	7.47%
		Negative	4.39%	3.98%	8.33%	7.23%	11.77%	10.16%
	Rel. Perf.	77.61%	26.88%	85.99%	29.13%	88.86%	26.51%	
Shiny_Skin	Positive	3.50%	3.93%	6.33%	7.13%	9.27%	10.04%	
	Negative	4.17%	3.96%	8.13%	7.18%	11.89%	10.11%	
	Rel. Perf.	16.14%	0.61%	22.13%	0.72%	22.04%	0.73%	
Hair	Bald	Positive	2.79%	3.50%	4.48%	6.38%	6.07%	9.14%
		Negative	4.01%	3.98%	7.43%	7.23%	10.62%	10.18%
		Rel. Perf.	30.40%	12.13%	39.77%	11.78%	42.83%	10.21%
	Wavy_Hair	Positive	3.03%	3.95%	6.34%	7.17%	10.28%	10.09%
		Negative	4.35%	3.97%	7.92%	7.23%	10.82%	10.17%
		Rel. Perf.	30.46%	0.73%	19.95%	0.84%	4.96%	0.80%
	Receding_Hairline	Positive	3.03%	3.92%	4.68%	7.12%	6.13%	10.04%
		Negative	4.10%	3.97%	8.20%	7.22%	12.25%	10.15%
		Rel. Perf.	26.21%	1.18%	42.90%	1.28%	49.98%	1.17%
	Bangs	Positive	4.03%	3.80%	8.79%	6.91%	13.94%	9.78%
		Negative	3.83%	3.98%	6.77%	7.23%	9.42%	10.17%
		Rel. Perf.	-5.11%	4.43%	-29.80%	4.44%	-47.96%	3.89%
	Sideburns	Positive	3.72%	3.97%	6.51%	7.21%	9.10%	10.13%
		Negative	3.98%	3.97%	7.62%	7.22%	11.10%	10.16%
		Rel. Perf.	6.58%	0.08%	14.56%	0.12%	18.07%	0.30%
	Black_Hair	Positive	5.12%	3.92%	9.85%	7.11%	13.47%	10.01%
		Negative	3.48%	3.97%	6.36%	7.21%	9.28%	10.15%
		Rel. Perf.	-47.25%	1.28%	-54.86%	1.46%	-45.17%	1.38%
Blond_Hair	Positive	3.09%	3.81%	7.38%	6.92%	12.43%	9.76%	
	Negative	4.09%	3.98%	7.34%	7.23%	10.16%	10.18%	
	Rel. Perf.	24.53%	4.22%	-0.57%	4.25%	-22.38%	4.07%	
Brown_Hair	Positive	3.24%	3.96%	6.46%	7.18%	10.26%	10.10%	
	Negative	4.12%	3.97%	7.59%	7.20%	10.59%	10.14%	
	Rel. Perf.	21.36%	0.35%	14.93%	0.26%	3.11%	0.36%	
Gray_Hair	Positive	2.68%	3.76%	4.01%	6.82%	5.40%	9.67%	
	Negative	4.07%	3.98%	7.57%	7.23%	10.77%	10.17%	
	Rel. Perf.	34.09%	5.58%	47.01%	5.70%	49.87%	4.97%	

Table 3.9.: ArcFace - Part 2/3. Face recognition performance based on several attributes.

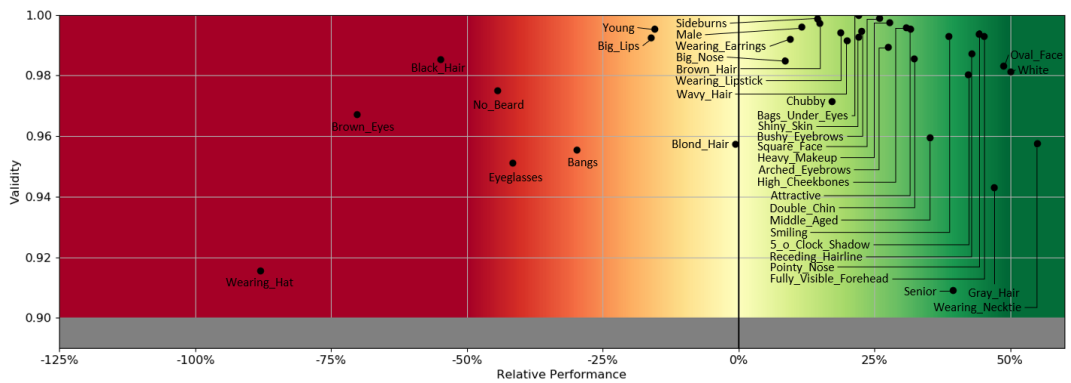
Category	Attribute	Class	EER		FNMR@FMR=10 ⁻³		FNMR@FMR=10 ⁻⁴	
			Real	Control	Real	Control	Real	Control
Beard	No_Beard	Positive	4.13%	3.98%	8.10%	7.23%	11.93%	10.18%
		Negative	3.31%	3.89%	5.61%	7.06%	7.90%	9.95%
		Rel. Perf.	-25.05%	-2.23%	-44.32%	-2.49%	-50.91%	-2.28%
	Mustache	Positive	4.89%	2.63%	9.62%	4.61%	13.54%	6.68%
		Negative	4.06%	3.98%	7.62%	7.23%	10.97%	10.17%
		Rel. Perf.	-20.46%	33.85%	-26.25%	36.24%	-23.41%	34.31%
	5 o Clock Shadow	Positive	2.96%	3.90%	4.94%	7.06%	7.08%	9.96%
		Negative	4.24%	3.96%	8.55%	7.20%	12.68%	10.14%
		Rel. Perf.	30.18%	1.74%	42.23%	1.97%	44.16%	1.75%
	Goatee	Positive	1.18%	2.46%	1.68%	4.00%	2.67%	5.89%
		Negative	4.08%	3.98%	7.67%	7.23%	11.03%	10.18%
		Rel. Perf.	71.16%	38.16%	78.13%	44.74%	75.83%	42.12%
Face Geometry	Oval_Face	Positive	2.73%	3.90%	5.69%	7.07%	9.65%	9.97%
		Negative	5.40%	3.96%	11.10%	7.19%	15.61%	10.12%
		Rel. Perf.	49.55%	1.59%	48.72%	1.67%	38.22%	1.39%
	Square_Face	Positive	3.73%	3.97%	6.37%	7.22%	8.68%	10.16%
		Negative	4.13%	3.97%	8.61%	7.21%	13.02%	10.14%
		Rel. Perf.	9.65%	0.03%	25.96%	-0.10%	33.37%	-0.15%
	Round_Face	Positive	7.04%	2.30%	22.68%	3.89%	35.87%	5.53%
		Negative	3.17%	3.98%	5.30%	7.22%	7.43%	10.16%
		Rel. Perf.	-122.46%	42.18%	-328.09%	46.22%	-383.05%	45.63%
	Double_Chin	Positive	3.34%	3.93%	5.32%	7.12%	7.00%	10.04%
		Negative	4.08%	3.98%	7.84%	7.23%	11.50%	10.17%
		Rel. Perf.	18.22%	1.23%	32.24%	1.43%	39.15%	1.29%
High Cheekbones	Positive	3.34%	3.95%	5.96%	7.17%	8.63%	10.10%	
	Negative	4.28%	3.97%	8.60%	7.20%	12.70%	10.13%	
	Rel. Perf.	21.87%	0.48%	30.76%	0.42%	32.08%	0.34%	
Chubby	Positive	3.70%	3.87%	6.11%	7.01%	7.86%	9.90%	
	Negative	3.90%	3.97%	7.37%	7.22%	10.79%	10.17%	
	Rel. Perf.	5.18%	2.62%	17.14%	2.85%	27.15%	2.58%	
Obstructed Forehead	Positive	5.48%	3.51%	13.03%	6.39%	20.40%	9.17%	
	Negative	3.52%	3.97%	6.10%	7.21%	8.56%	10.15%	
	Rel. Perf.	-55.61%	11.62%	-113.74%	11.37%	-138.28%	9.66%	
Fully Visible Forehead	Positive	3.30%	3.97%	5.47%	7.21%	7.49%	10.15%	
	Negative	4.64%	3.95%	9.98%	7.16%	15.06%	10.06%	
	Rel. Perf.	28.85%	-0.46%	45.15%	-0.70%	50.30%	-0.86%	
Periocular	Brown_Eyes	Positive	4.69%	3.97%	9.13%	7.21%	12.88%	10.14%
		Negative	2.63%	3.85%	5.36%	6.98%	8.73%	9.85%
		Rel. Perf.	-78.48%	-2.96%	-70.17%	-3.28%	-47.51%	-2.92%
	Bags_Under_Eyes	Positive	3.78%	3.96%	6.37%	7.19%	8.48%	10.11%
		Negative	3.87%	3.96%	8.17%	7.19%	12.63%	10.12%
		Rel. Perf.	2.20%	0.13%	22.05%	0.01%	32.84%	0.10%
	Bushy_Eyebrows	Positive	3.51%	3.96%	6.05%	7.19%	8.28%	10.11%
		Negative	4.00%	3.97%	7.81%	7.22%	11.58%	10.17%
		Rel. Perf.	12.35%	0.20%	22.54%	0.54%	28.46%	0.60%
	Arched_Eyebrows	Positive	3.21%	3.94%	6.08%	7.15%	9.29%	10.05%
		Negative	4.42%	3.97%	8.38%	7.23%	11.79%	10.17%
		Rel. Perf.	27.52%	0.81%	27.46%	1.05%	21.20%	1.14%

Table 3.10.: ArcFace - Part 3/3. Face recognition performance based on several attributes.

Category	Attribute	Class	EER		FNMR@FMR=10 ⁻³		FNMR@FMR=10 ⁻⁴		
			Real	Control	Real	Control	Real	Control	
Mouth	Mouth Closed	Positive	3.06%	3.37%	5.40%	6.13%	7.70%	8.85%	
		Negative	3.88%	3.90%	7.79%	7.08%	11.93%	9.99%	
	Smiling	Positive	21.21%	13.69%	30.62%	13.38%	35.48%	11.39%	
		Negative	3.35%	3.94%	5.93%	7.14%	8.48%	10.05%	
	Big_Lips	Positive	4.62%	3.96%	9.65%	7.19%	14.57%	10.12%	
		Negative	27.32%	0.56%	38.59%	0.71%	41.81%	0.65%	
Rel. Perf.		4.15%	3.94%	8.12%	7.17%	11.91%	10.10%		
Nose	Big_Nose	Positive	3.88%	3.97%	7.00%	7.22%	9.84%	10.16%	
		Negative	3.88%	3.97%	7.00%	7.22%	9.84%	10.16%	
	Pointy_Nose	Positive	-6.93%	0.78%	-16.08%	0.75%	-21.01%	0.63%	
		Negative	4.39%	3.90%	7.89%	7.07%	10.48%	9.95%	
	Accessories	Heavy Makeup	Positive	3.90%	3.95%	8.62%	7.18%	13.58%	10.10%
			Negative	3.90%	3.95%	8.62%	7.18%	13.58%	10.10%
Wearing Hat		Positive	-12.67%	1.49%	8.52%	1.51%	22.80%	1.46%	
		Negative	3.15%	3.97%	5.84%	7.22%	8.86%	10.16%	
Wearing Earrings		Positive	5.28%	3.96%	10.46%	7.18%	14.62%	10.11%	
		Negative	5.28%	3.96%	10.46%	7.18%	14.62%	10.11%	
Other	Attractive	Positive	40.44%	-0.43%	44.19%	-0.63%	39.44%	-0.49%	
		Negative	3.08%	3.96%	5.79%	7.20%	9.00%	10.13%	
	Wearing Necktie	Positive	4.32%	3.97%	8.02%	7.22%	11.14%	10.16%	
		Negative	4.32%	3.97%	8.02%	7.22%	11.14%	10.16%	
	Wearing Lipstick	Positive	28.75%	0.18%	27.75%	0.24%	19.27%	0.32%	
		Negative	5.51%	3.66%	12.28%	6.62%	18.45%	9.44%	
No_Eyewear	Positive	3.71%	3.98%	6.53%	7.23%	9.14%	10.18%		
	Negative	3.71%	3.98%	6.53%	7.23%	9.14%	10.18%		
Eyeglasses	Positive	-48.79%	8.09%	-88.01%	8.45%	-101.94%	7.29%		
	Negative	3.25%	3.95%	6.64%	7.17%	10.59%	10.10%		
Other	Attractive	Positive	4.08%	3.98%	7.33%	7.23%	10.10%	10.17%	
		Negative	4.08%	3.98%	7.33%	7.23%	10.10%	10.17%	
Other	Attractive	Positive	20.23%	0.83%	9.44%	0.80%	-4.92%	0.78%	
		Negative	2.72%	3.82%	3.84%	6.92%	4.72%	9.79%	
Other	Attractive	Positive	4.25%	3.97%	8.52%	7.22%	12.68%	10.16%	
		Negative	4.25%	3.97%	8.52%	7.22%	12.68%	10.16%	
Other	Attractive	Positive	35.95%	4.00%	54.94%	4.24%	62.77%	3.73%	
		Negative	3.28%	3.96%	6.38%	7.19%	9.93%	10.11%	
Other	Attractive	Positive	4.27%	3.98%	7.85%	7.23%	10.83%	10.18%	
		Negative	4.27%	3.98%	7.85%	7.23%	10.83%	10.18%	
Other	Attractive	Positive	23.21%	0.40%	18.74%	0.57%	8.25%	0.65%	
		Negative	3.64%	3.98%	6.39%	7.23%	8.92%	10.18%	
Other	Attractive	Positive	3.86%	3.50%	6.62%	6.35%	8.99%	9.12%	
		Negative	3.86%	3.50%	6.62%	6.35%	8.99%	9.12%	
Other	Attractive	Positive	5.75%	-13.57%	3.42%	-13.84%	0.82%	-11.60%	
		Negative	4.60%	3.79%	9.13%	6.88%	13.03%	9.75%	
Other	Attractive	Positive	3.68%	3.98%	6.45%	7.23%	8.99%	10.18%	
		Negative	3.68%	3.98%	6.45%	7.23%	8.99%	10.18%	
Other	Attractive	Positive	-25.08%	4.88%	-41.53%	4.89%	-44.86%	4.24%	
		Negative	2.95%	3.96%	5.49%	7.19%	8.60%	10.10%	
Other	Attractive	Positive	4.30%	3.97%	8.02%	7.22%	11.14%	10.17%	
		Negative	4.30%	3.97%	8.02%	7.22%	11.14%	10.17%	
Other	Attractive	Positive	31.44%	0.20%	31.57%	0.47%	22.82%	0.65%	
		Negative	31.44%	0.20%	31.57%	0.47%	22.82%	0.65%	



(a) FaceNet



(b) ArcFace

Figure 3.15.: Visual summary on the performance differences affected by each attribute. Figure 3.15a visualizes the results for FaceNet, while Figure 3.15b visualizes the results for ArcFace. The relative performance is based on the recognition performance on the positively-labelled data versus the performance of negatively-labelled data. The validity is based on the performance differences of the control groups. Validity values below 0.9 (more than 10% performance differences between the control groups) are considered as *not valid* (grey area) and are not shown in this figure. The red areas indicate an attribute-related bias that leads to a degraded face recognition performance for faces with the specific attribute. Green areas indicate that faces possessing a specific attribute enhances the recognition performance. It can be observed that the majority of the investigated attributes strongly affects the recognition performance.

the positive attribute class over the negative class. If an attribute has only a minor effect on the recognition performance, the relative performance is close to 0% (yellow area).

FaceNet vs ArcFace The main difference between FaceNet and ArcFace are the underlying training-principles. FaceNet uses triplet-loss learning [SKP15] that aims solely at minimizing the intra-class variations while maximizing the inter-class variations. In contrast, ArcFace introduces an angular large-margin principle [Den+19] that additionally aims at enhancing the robustness of recognition model. The utilized training principle together with the used network structure and the training data determines the recognition behaviour. This includes the effect of performance differences appearing when certain attributes of the face are present. Since the used FaceNet and ArcFace models share the same network structure and training data, the observed performance differences might arise from the training principles.

The Effect of Attributes on Recognition It turns out that the majority of the investigated attributes strongly affect the recognition performance of both, FaceNet and ArcFace. For FaceNet, many faces that are perceived as *Attractive* or make use of *Heavy Makeup* do not show to alter the recognition performance. The same goes for *Oval Faces* and faces with *Sideburns*. For ArcFace, *Blond Hair*, *Big Nose*, *Big Lips*, *Wearing Earrings*, and *Young* faces show only a minor effect on the recognition performance. However, especially for ArcFace the used decision threshold (here for a FMR of 10^{-3}) determines the performance difference. For both recognition models, the majority of the investigated attributes strongly affect the recognition performance. Some of the observations might be explainable.

- **Demographics:** Recent works [HSV19a; Rob+20; GNH19b; Bal+20] extensively discussed the impact of demographic attributes on face recognition. Our results support the findings from previous works. We observe an improved recognition performance for the attributes *Middle Aged*, *Senior*, *White*, and *Male*. Contrarily, a degraded recognition performance is observed for *Young*, *Asian*, *Black*, and *Female* faces. For FaceNet, the observed performance differences are stronger than for ArcFace. Moreover, we could not show that *Asian* or *Black* faces performance weaker than *White* faces on ArcFace, since the data unbalance lead to a low validity for our results.
- **Visibility-related attributes:** We observe that attributes that indicate a fully visible face lead to an improved face recognition performance. This includes the attributes *Fully Visible Forehead*, *Receding Hairline*, *No Eyewear*, and *Bald*. In contrast, attributes that might lead to small partial occlusions of the face lead to significantly degraded

recognition performances. For FaceNet, this includes faces with an *Obstructed Forehead*, *Bangs*, and *Wavy Hair*. For ArcFace, this includes samples with *Eyeglasses* or *Bangs*.

- **Temporary attributes:** For faces with temporary attributes, such as for accessories, a degraded face recognition performance can be observed. This includes *Wearing Hat*, *Wearing Earrings*, *Wearing Lipstick*, and *Eyeglasses*. Beside a partial-occlusion of small parts of the face, these attributes are non-permanent and can quickly change the appearance of the face.
- **Anomalous characteristics:** It turns out that conspicuous characteristics that is only possessed by a small proportion of the population lead to strongly enhanced recognition performances. This includes *Arched Eyebrows*, *Big Nose*, *Pointy Nose*, *Bushy Eyebrows*, *Double Chin*, and *High Cheekbones*.
- **Facial expressions:** Faces that are *Smiling* or that have their *Mouth Closed* perform above average for face recognition. Contrary, faces with non-neutral expressions lead to degraded face recognition performances. This bias might come from the data utilized for training that usually contains neutral or smiling faces and was discussed in more details by previous works [CBF06; CBF05].

While these attribute-dependent performance differences might be explainable, the reason for the impact of other attributes on recognition is currently unclear.

- **Colors:** The results demonstrate strong performance differences based on the user's hair- and eyecolor. For FaceNet, faces with *Blond Hair*, *Black Hair*, and *Brown Hair* show strongly degraded recognition performances. In contrast, faces with *Gray Hair* lead to an improved recognition. For ArcFace, *Gray Hair* also strongly improves the recognition performance while *Black Hair* decreases it. The performance differences for *Blond Hair* and *Brown Hair* strongly varies dependent on the used decision threshold. For instance, for high FMRs, *Blond Hair* has a positive effect on recognition, for a lower FMR (e.g. 10^{-4}) the same attribute changes to a negative effect. The same can be observed for eyecolors. Faces with *Brown Eyes* perform weaker than faces from the opposite group. The performance differences of these attributes does not reflect the distribution of the training data and thus, might arise from a different origin.
- **Beard:** As we discussed before, attributes that might induce a partially occluded face lead to a degraded face recognition performance. Although, beards cover parts of the face that should lead to a degraded performance, the results demonstrate the

opposite. Faces with *No Beard* perform below-average, while faces with e.g. a *5 o Clock Shadow* achieve much higher recognition rates.

- **Wearing Necktie:** Unlike other accessories, *Wearing Necktie* improved the face recognition performance drastically. We assume that this results from a data collection bias induced by the correlation with hidden factors, such as environment. Persons wearing a necktie often have to present themselves in public (e.g. celebrities) and thus, photos are often taken with frontal poses and full lightning. However, the high validity and the strong performance differences makes it hard to argue in this direction.
- **Antagonistic Behaviour:** Some attributes result in performance differences of the opposite direction depending on the used training-principle (triplet-loss vs angular margin loss). For instance, faces with *Wavy Hair* lead to a negative performance on FaceNet and to positive performance on ArcFace. Also the attributes *Attractiveness*, *Heavy Makeup*, and *Oval Faces* negatively affect the face recognition performance on FaceNet, but show some strong positive impacts on the recognition performance of ArcFace.


As mentioned earlier, the resulting performance of a face recognition model is mainly determined by its loss-function, its network architecture, and the utilized training data. Since both investigated models have the last two points in common, the observed differences in the performance might arise from the underlying training principles. Generally, we observe that the large angular margin loss from ArcFace leads to a significantly stronger overall recognition performance compared to FaceNet. The loss aiming to enhance the model robustness also shows a clearly visible effect on the attribute-related performance differences. On ArcFace, slightly less attributes negatively affect the recognition performance than on FaceNet. However, the performance differences that origins from the affected (biased) attributes are still of high impact. A remarkable observation is the fact that the performance differences remain relatively constant over several decision thresholds for FaceNet, while for ArcFace the performance differences often significantly vary for different decision thresholds. This can be observed for instance for faces with *Bangs*, *Blond Hair*, or a *Double Chin*.

Future challenges for face recognition The observations of the experiment point out some critical issues of current face recognition solutions in terms of robustness, fairness, and explainability.

-
- **Need for robustness:** Face recognition systems need to become more robust against partial occlusions (from accessories or hair), facial expressions (beyond neutral and smiling faces), and temporary attributes that might change the daily appearance of a face. This can greatly enhance the applicability in more real-life scenarios.
 - **Need for fairness:** Face recognition systems need to enhance the user-fairness. We observed performance differences based on the user-demographics (demographic-bias), anomalous characteristics (such as pointy noses, bushy eyebrows, and high cheekbones), beard types, and accessories. This can lead to discriminative decisions of face recognition systems that several political regulation, such as the GDPR [VB17], try to prevent.
 - **Need for explainability:** Face recognition models need to explain themselves. Why do colors/face shapes/beards/accessories lead to performance differences? Why can we observe an antagonistic behaviour between the two different learning principles for some attributes? In order to enhance the model transparency and to enable efficient model-debugging, future work have to elaborate on the explainability of face recognition models.

3.3.5. Interim Conclusion

The growing effect of face recognition systems on everybody's daily life, including critical decision-making processes, shows the need of non-discriminative face recognition solutions. Previous works focused on estimating and mitigating demographic-bias. However, to deploy non-discriminatory face recognition systems it is necessary to know which performance differences appear in the presences of certain facial attributes beyond demographics. Driven by this need, this section analysed the performance differences on two popular face recognition models concerning 47 different attributes to answer RQ2. To prevent misleading statements of attribute biases, we consider attribute correlations and minimize the effect of unbalanced testing data via control group based validity values. We investigated the effect of two different learning-principles on the performance differences originating from facial attributes. The results show that, besides demographics, many attributes strongly affect the recognition performance of both investigated face recognition models, FaceNet and ArcFace. While for FaceNet the observed performance differences originated by several attributes remain relatively constant, these differences strongly depend on the used decision threshold for ArcFace. We provided explanations for many observed performance differences. However, the reason for some observations remain unclear and have to be addressed by future work. The findings of this work strongly demonstrate the need for further advances in making face recognition systems more robust,



explainable, and fair. The results of this section show the strong demand for generalized bias-mitigating solutions. In Chapter 4, easily-integrable solutions are presented that fulfil this need.

3.4. Investigating Bias in Face Quality Assessment

3.4.1. Introduction

RQ2 aims at investigating the influence of soft-biometric attributes on the behaviour of face recognition systems. In the previous section, we demonstrated that the performance of face recognition systems is highly different depending on the user’s soft-biometrics. In this section, we analyse the influence of soft-biometric attributes on face quality assessment since this also determines the behaviour of a face recognition system. The study is based on the work from Terhörst et al. [Ter+20e]. In general, biometric sample quality is defined as the utility of a sample for the purpose of recognition [Her+19; Phi+13; Gao+07; BJ18] and is crucial for many applications. Recent work [Ter+20g] has shown that the accuracy and the robustness of face quality estimation can be enhanced drastically by adapting the face quality assessment algorithm to the deployed face recognition model. However, this can lead to biased face quality assessment algorithms as well.

There are several political regulations to prevent discriminatory decisions. Article 14 of the European Convention of Human Rights and Article 7 of the Universal Declaration of Human Rights ensure people the right to non-discrimination. Also, the General Data Protection Regulation (GDPR) [VB17] aims at preventing discriminatory effects (article 71). Despite these political efforts, several works [Phi+11; BG18; Orc16; AZN18; FPO02; Gar+16] showed that open-source [Orc16; Ser+20] as well as commercial [BG18] face recognition solutions, are strongly biased towards different demographic groups. The more accurate terms of “differential performance” and “differential outcome” were presented in [HSV19b] to avoid the unintended interpretation of bias by policy makers and statisticians. Based on these terms, a number of recent works are supporting the notion of differential performance in face recognition systems [GNH19b; Co0+19].

Face quality assessment solutions can possess intended and unintended kinds of biases, e.g. non-demographic and demographic bias. While non-demographic bias enhances the quality estimation process without discriminative consequences, transferring demographic bias unintentionally to face quality assessment algorithms can have a serious impact on society. During the enrolment of an individual or for quality-based fusion approaches (e.g. in surveillance scenarios), face quality assessment is needed. Consequently, a transferred bias to the quality estimation will directly increase discriminative decisions of such quality-based subsystems.

Moreover, in the operation, face quality estimation can be used as a separate processing step [20] and can be trained while having in mind a face recognition system different than the one used in the field. Therefore, having a biased quality estimation can add to the bias of the face recognition system, as it might have different biases.

In this section, we present a detailed analysis of the correlation between bias in face recognition systems and the corresponding face quality assessment. To the best of our knowledge, this is the first work analysing this relation. The experiments were conducted on two publicly available datasets under diverse image capturing conditions. The correlation analysis was done using two different face recognition solutions with four state-of-the-art face quality assessment algorithms from academia and industry. Investigating different head poses, ethnicities, and age classes, we found degraded performances, and thus biases, towards certain subclasses for both face recognition systems. The experiments demonstrated a strong correlation between face recognition bias and face quality assessment. Face images from the classes affected by the bias were estimated with lower quality values than unbiased images. Consequently, the bias is transferred to the quality assignment process.

The goal of this work is to point out that current face image quality assessment approaches have to deal with similar bias-related problems than in face recognition. We specify that the quality of a face image points out a biased ground of a faulty decision. Especially in a controlled environment, such as ABC gates where the image is of good quality, a low face quality must alarm the operator to a high probability of a faulty decision, whether a false match or a false non-match, which might require manual inspection. This faulty decision can be a bias issue given the controlled capture conditions.

3.4.2. Related Work

Several standards have been proposed to insure face image quality by constraining the capture requirements, such as ISO/IEC 19794-5 [11] and ICAO 9303 [15]. In these standards, quality is divided into *image-based* qualities (such as illumination, occlusion) and *subject-based* quality measures (such as pose, expression, accessories). These mentioned standards influenced many face quality assessment approaches that have been proposed recently.

The first generation of face quality assessment algorithms defines quality metrics based on image quality factors [Gao+07; Fer+12; Was+17; ZG17; Phi+13; Aba+14; HSM06; AHB12; DVS14]. However, these approaches have to consider every possible factor manually, and since humans may not know the best characteristics for face recognition systems, recent research focuses on learning-based approaches.

End-to-end learning approaches for face quality assessment were first presented in 2011. Aggarwal et al. [Agg+11] proposed an approach for predicting the face recognition performance using a multi-dimensional scaling approach to map space characterization features to genuine scores. In [Won+11], a patch-based probabilistic image quality approach was designed to work on 2D discrete cosine transform features and trains a

Gaussian model on each patch. In 2015, a rank-based learning approach was proposed by Chen et al. [Che+15]. They define a linear quality assessment function with polynomial kernels and train weights based on a ranking loss. In [KLR15], face quality assessment was performed based on objective and relative face image qualities. While the objective quality metric refers to objective visual quality in terms of pose, alignment, blurriness, and brightness, the relative quality metric represents the degree of mismatch between training face images and a test face image. Best-Rowden and Jain [BJ18] proposed an automatic face quality prediction approach in 2018. They proposed two methods for quality assessment of face images based on (a) human assessments of face image quality and (b) quality values from similarity scores. Their approach (b) is based on support vector machines applied to deeply learned representations. In 2019, Hernandez-Ortega et al. proposed FaceQnet [Her+19], which adapts the quality label generation from Best-Rowden [BJ18] and applies it to fine-tune a face recognition neural network to predict face qualities in a regression task. Stochastic embedding robustness (SER-FIQ) is a novel face image quality measurement concept proposed in [Ter+20g]. Their method determines the embedding variations generated from random subnetworks of the deployed face recognition model. The magnitude of these variations define the robustness and thus, the quality. Their method avoids the need for training and further allows to take into account the decision patterns of the deployed face recognition model.

So far, the best quality estimates were achieved when the systems adapt to the utilized face recognition model. However, there is a risk of transferring the face recognition bias towards the quality assessment. Therefore, this work analyses the correlation between face quality assessment and face recognition bias. To the best of our knowledge, this is the first work that analyses this relationship and its implications on the real use of the technology.

3.4.3. Evaluated Face Quality Assessment Solutions

Face quality assessment aims at estimating the usability of an image for the purpose of recognition [Her+19; Phi+13; Gao+07; BJ18]. For our correlation study between face quality and face recognition bias, we choose the four of the latest face quality assessment approaches from academia and industry. These approaches will be shortly discussed in the following.

COTS COTS [Neu19] refers to an commercial off the shelf industry product from Neurotechnology, the used version is published in 2019. Unfortunately, it only provides the application and does not provide any information about its working principles. However, in

[Ter+20g], the authors show that COTS predicted quality synchronise well with FaceNet [SKP15] performance, and to a much lower degree with ArcFace [Den+19] performance.

Best-Rowden In 2018, Best-Rowden and Jain [BJ18] presented two approaches to face quality estimation, with and without human assessments. We evaluate their approach based on quality labels coming from comparison scores, because the features and comparison scores are matcher dependent and thus, it adapts to the deployed face recognition model. They define a quality label for query j of subject i as

$$z_{ij} = (s_{ij}^G - \mu_{ij}^I) / \sigma_{ij}^I, \quad (3.5)$$

where s_{ij}^G is the genuine score and μ_{ij}^I and σ_{ij}^I are the mean and the standard deviation of the imposter scores. They use the face embeddings of the deployed face recognition model and, based on these features, they train a support vector regressor to estimate the quality score of an input image. Following their methodology, we train this approach on the MORPH [RT06b] dataset. The hyperparameters are determined beforehand by a 5-fold cross-validation on this dataset.

FaceQnet FaceQnet [Her+19] by Hernandez-Ortega et al. was published in 2019. They adapted the idea of using the comparison score labels (see Equation 3.5) from Best-Rowden et al. and combined them with a ResNet-based deep neural network structure. Their approach is based on FaceNet embeddings and is trained on VGGFace2 [Cao+18]. In [Ter+20g], it was shown that even if the approach was trained on FaceNet embeddings, FaceQnet shows better synchronisation with ArcFace [Den+19] performance, indicating some overfitting on FaceNet [SKP15] embeddings. For our experiments, we used the pretrained FaceQnet model⁹ provided by the authors.

SER-FIQ Stochastic embedding robustness (SER) is a face image quality (FIQ) estimation concept presented in [Ter+20g], which avoids the use of inaccurate quality labels. They defined face image quality based on the robustness of deeply learned features. Calculating the variations of embeddings coming from random subnetworks of the deployed face recognition model, their solution defines the magnitude of these variations as a robustness measure, and thus, image quality. Given an input image I and the deployed face recognition model \mathcal{M} , their method applies $m = 100$ different dropout patterns [Sri+14] to the neural network. This results in m random subnetworks of \mathcal{M} . Each of these networks

⁹<https://github.com/uam-biometrics/FaceQnet>

produces a stochastic embedding x_i . The quality

$$q(I) = 2\sigma\left(-\frac{2}{m^2}\sum_{i<j}d(x_i, x_j)\right), \quad (3.6)$$

of an input I is then defined as the sigmoid of the negative mean Euclidean distance $d(x_i, x_j)$ between all stochastic embedding pairs. A greater variation between the stochastic embeddings indicates a lower robustness of the representation and thus, a lower sample quality q . Lower variations between the stochastic embeddings indicate a high robustness in the embedding space and are considered as a high sample quality q . Since it can be directly applied on the deployed face recognition model, it completely avoids any training and further adapts to the decision patterns of the model. The authors showed that this concept leads to significantly better quality estimations than previous work. We follow their procedure that applies the dropout pattern repetitively on the last layer of the face recognition network. A more detailed explanation of SER-FIQ is presented in the appendix Section A.

3.4.4. Experimental Setup

Database To evaluate the correlation between face quality assessment and bias in face recognition systems under controlled and unconstrained conditions, we conducted experiments on the two publicly available datasets, ColorFeret [Phi+00] and Adience [EEH14]. ColorFeret [Phi+00] consists of 14k images of 1.2k different individuals with different poses under controlled conditions. The dataset further includes a variety of face poses, facial expressions, and lighting conditions. The Adience dataset [EEH14] consists of over 26.5k images of over 2.2k different individuals in an unconstrained environment. Both databases contain information about identity, gender and age. ColorFeret also provides labels regarding the subject’s ethnicities and head poses. In the experiments, this information is used to investigate how face quality assessment algorithms affect the recognition performance under diverse circumstances.

The investigated face quality assessment solutions are based on three databases, MORPH [RT06b], VGGFace2 [Cao+18], and MS1M [Guo+16]. MORPH [RT06b] contains 55k frontal face images of more than 13k individuals. 80.4% of the faces belong to the ethnicity black, 19.2% to white, and 0.4% to others. The individual’s age vary from 16-77 years. 79.4% of the faces are within an age-range of [20, 50]. The VGGFace2 [Cao+18] database contains faces from over 9k subjects with over 3 million images. The dataset contains a large variety of pose, age, and ethnicity. Over 40% of the face are frontal and over 50% are half-frontal. Most images belong to individuals over 18 years old and around 40%

belong to the age group of [25, 34]. The MS1M [Guo+16] contains over 100k subjects with 10 million images. The faces cover a large variance of age. Over 50% of the faces belong to white subjects. The faces are mostly frontal. This information will be used to discuss the influence of the training data on quality predictions.

Evaluation metrics In order to evaluate the face quality assessment performance, we follow the methodology by Grother et al. [GT07] using error versus reject curves. These curves show the verification error-rate (y-axis) achieved when unconsidering a certain percentage of face images (x-axis). Based on the predicted quality values, these unconsidered images are these with the lowest predicted quality and the error rate is calculated on the remaining images. Error versus reject curves indicate good quality estimation when the verification error decreases consistently when increasing the ratio of unconsidered images.

In order to prove that a face recognition system is biased towards some classes, the verification error is reported for all classes. The verification error is reported in terms of false non-match rate (FNMR) at fixed false match rates (FMR). The FMR is reported at 0.1% FMR threshold as recommended by the best practice guidelines for automated border control of European Border Guard Agency Frontex [Fro17]. To show the correlation between face quality assessment and biased face verification performance, the proportion of subgroups is continuously analysed over quality thresholds. The proportion of biased subgroups will decrease fast if the face quality assessment algorithm assigns them lower quality values than unbiased subgroups. To get a deeper understanding of the correlation between biased and quality, quality distributions for the different subgroups are illustrated. These allow validating shifts and separations between biased and unbiased subgroups.

Face recognition networks To get the face embedding for a given face image, the image has to be aligned, scaled, and cropped. Then, the preprocessed image is passed to a face recognition model to extract the embeddings. In this work, we use two face recognition models, FaceNet [SKP15] and ArcFace [Den+19]. For FaceNet, the preprocessing is done as described in [KS14]. To extract the embeddings, a pretrained model¹⁰ was used. For ArcFace, the image preprocessing was done as described in [Guo+18] and a pretrained model¹¹ is used, which is provided by the authors of ArcFace. Both models were trained on the MS1M database [Guo+16]. The output size is 128 for FaceNet and 512 for ArcFace. The identity verification is done by comparing two embeddings using cosine-similarity.

¹⁰<https://github.com/davidsandberg/facenet>

¹¹<https://github.com/deepinsight/insightface>

Investigations This work aims at investigating the correlation between face recognition bias and face quality estimation. This is done using two popular face embeddings, FaceNet [SKP15] and ArcFace [Den+19]. Since the face quality assessment performance strongly influences the interpretation of the correlation analysis, the quality estimation performance is analysed in the first step. The second step aims at demonstrating that there is bias in the utilized face recognition systems. Therefore, the face verification performance of these systems is analysed based on poses, ethnicities, and age classes. After the bias between these classes is identified, the correlation between the face quality assessment and the face recognition bias is investigated in the third step. Moreover, the separability in the quality space of the biased and unbiased classes is analysed.

3.4.5. Results

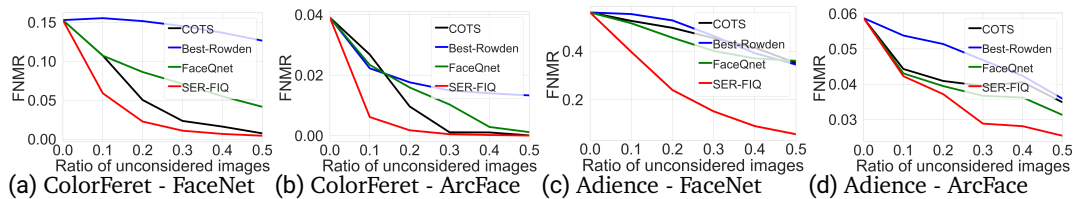


Figure 3.16.: Face quality assessment performance on ColorFeret and Adience using two face embeddings, FaceNet and Arcface. The FNMR is reported at a FMR of 0.1%. The plots are at different scales, but show similar quality assessment behaviours.

Face quality assessment performance Figure 3.16 shows the face quality assessment performance for the four discussed solutions. The performance is reported in terms of FNMR at FMR of 0.1% as recommended by the European Border Guard Agency Frontex [Fro17]. It can be seen that COTS shows a better quality estimation performance under constrained scenarios (Figure 3.16a and 3.16b). The approach of Best-Rowden shows a better quality prediction performance on ArcFace embeddings than on FaceNet. This might be because Best-Rowden was trained on a frontal face database and ArcFace is more robust to these variations. FaceQnet uses the same kind of training labels than Best-Rowden, but trained a deep learning model to make more advanced predictions. This approach shows a solid performance in all cases. Similar to the results from [Ter+20g], SER-FIQ shows the best performance in all scenarios. This is probably because this method exploits the decision patterns of the deployed model is therefore able to estimate how

robust the model is about the input.

Table 3.11.: Face verification performance within certain subgroups. The FNMR is evaluated at two FMR thresholds for two face recognition models. In each category, at least one subgroup shows a significantly higher error rate indicating a strong bias in the face embeddings.

		FaceNet		ArcFace		
Classes		0.1%FMR	1%FMR	0.1%FMR	1%FMR	
Colorferet	Pose	Frontal	0.40%	0.00%	0.00%	0.00%
		Half	1.78%	0.15%	0.07%	0.04%
		Profile	30.95%	10.14%	12.29%	7.55%
		Rotated	0.07%	0.03%	1.39%	0.00%
	Ethnicity	White	10.79%	3.41%	2.55%	1.80%
		Asian	33.90%	12.06%	6.63%	4.19%
		Black	31.34%	16.54%	6.41%	3.66%
		Others	12.15%	6.29%	3.53%	2.08%
All		16.22%	3.92%	4.15%	2.98%	
Adience	Age	[0,2]	80.02%	59.88%	18.81%	9.73%
		[4,6]	63.95%	36.80%	13.19%	6.46%
		[8,12]	37.16%	17.27%	9.92%	4.79%
		[15,20]	89.78%	52.51%	10.30%	6.15%
		[25,32]	28.37%	4.58%	5.31%	4.81%
		[38,43]	16.48%	4.07%	2.68%	2.07%
		[48,53]	20.94%	5.85%	1.92%	1.39%
		[60,100]	11.32%	2.97%	1.67%	0.66%
All		55.99%	16.28%	5.99%	3.24%	

Identifying biases in pose, ethnicity, and age In order to identify biased classes in the two utilized face embeddings, Table 3.11 shows the face verification performance at two decision thresholds for FaceNet and ArcFace embeddings. The performance is evaluated over four different head poses, four ethnicities, and eight age classes. In the case of poses, all poses show very low error rates, with the exception of the profile view. Here, the error rates are more than 10 times higher than the next highest class. This shows that there is a strong bias towards profile face images. In the case of ethnicities, face images of white individuals show the smallest error rate, followed by the class others. For the ethnicities asian and black, the error rates are strongly increased and thus, indicate a strong ethnic

bias. This might come from a training process that mainly involved face images of white individuals. In the cases of the age classes, there are higher error rates among young people (below 7 years) compared to older individuals. This bias might come from the lack of appropriate training material as well as the fact that faces at this age are not yet fully developed.

Over all three attributes pose, ethnicity, and age, ArcFace shows significantly lower error rates than FaceNet. However, for both face embeddings, it is demonstrated that there exists high biases towards certain classes.

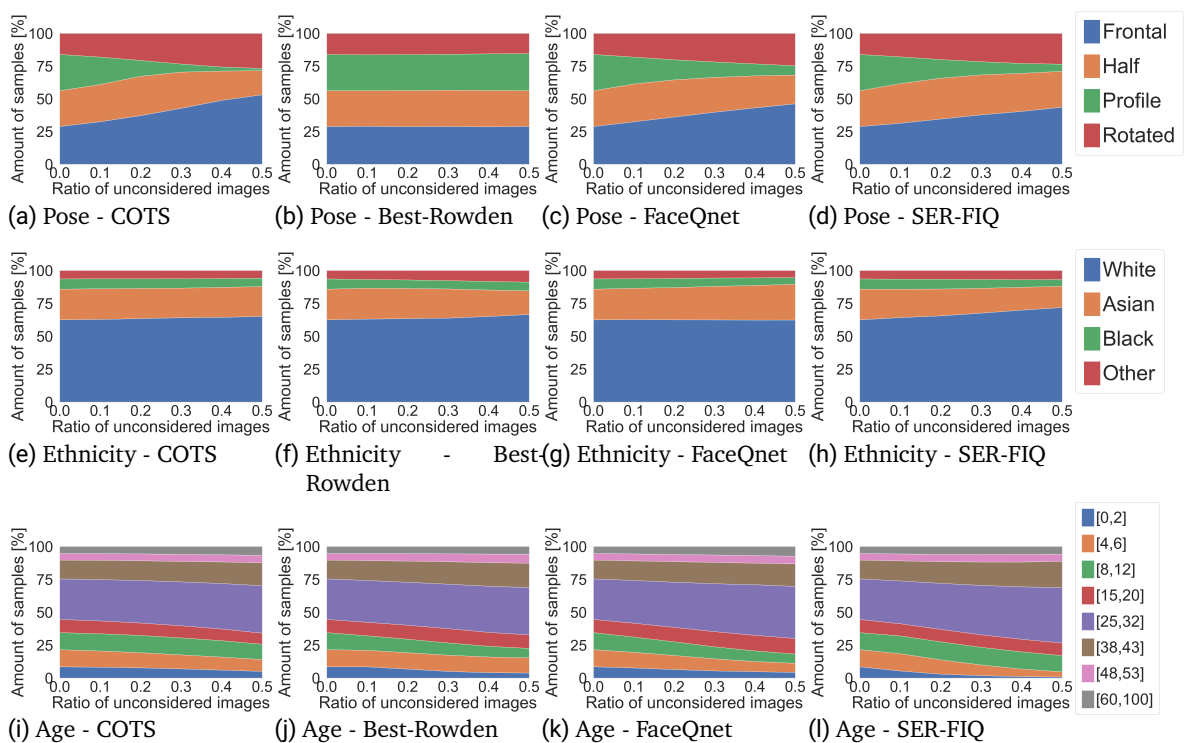


Figure 3.17.: Analysis of the proportion of subgroups for FaceNet embeddings. The pose (a-d), ethnicities (e-h), and age (i-l) proportions are shown when applying several quality thresholds.

The correlation study - bias versus quality In order to analyse which kind of images will be assigned low face image qualities, Figure 3.17 and 3.18 show an analysis of

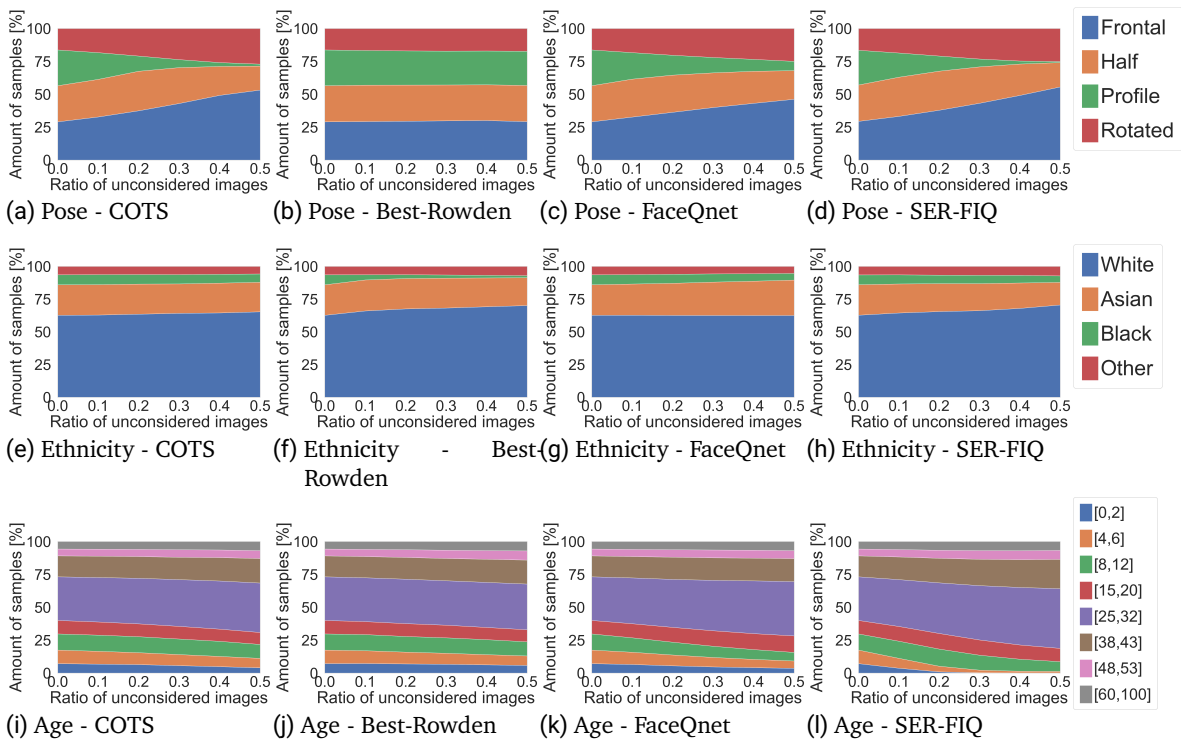


Figure 3.18.: Analysis of the proportion of subgroups for ArcFace embeddings. The pose (a-d), ethnicities (e-h), and age (i-l) proportions are shown when applying several quality thresholds.

the proportion of subclasses remaining when applying several quality thresholds. An unbiased face quality estimator will result in a stable proportion of subclasses over different quality thresholds. A biased estimator will cause classes effected by the bias to shrink, since these classes are mainly assigned with low quality values. To get a more detailed understanding of the correlation between quality scores and affected classes, Figure 3.19 show quality score distributions for the different subclasses. Based on the experiment before, we observed bias to frontal poses, at asian and black ethnicities and to face images of individuals below 7 years.

COTS The industry product COTS from Neurotechnology shows a very strong performance in filtering profile face images. This can be observed with FaceNet and ArcFace

embeddings. The score distributions further show high peaks around the lowest quality values for profile images indicating a weak recognition performance for this pose. For age, the number of samples of biased age classes affected by the bias are reduced with higher quality thresholds. The score distributions of the class [0,2] and [4,6] are shifted towards lower qualities. Consequently, the quality assessment is biased towards age. For ethnicity, the proportions for the different classes mainly stagnates. Furthermore, in the corresponding score distributions the distributions show a large overlap in both cases. Consequently, the face quality is mainly biased towards pose and age.

Best-Rowden The approach from Best-Rowden shows biased decision towards ethnicity and age. For different head poses, the quality predictions do not differ and the quality distributions are very similar. This can be explained by training on the frontal face database MORPH. For FaceNet embeddings, a slightly biased behaviour is seen for asian and black faces. For ArcFace embeddings, a strong bias towards black faces is observable. Despite training the approach on a database with 80.4% black ethnics, the major influence comes from the utilized embeddings that were used for the training. Both embeddings were trained on MS1M, a database with mainly white ethnicities. For age, it can be observed that age classes under 12 years are getting lower quality estimates.

FaceQnet FaceQnet [Her+19] shows a bias in all three investigated cases. For pose and age, the method reduces the number of samples of the classes affected by the bias showing that also the quality assessment possesses the same bias. This is supported by the quality distributions in Figure 3.19. The profile distribution is clearly separated and the distributions for young individuals (till 12 years) are shifted towards smaller quality values. For ethnicity, the number of samples from the classes affected by bias increases on FaceNet as well as ArcFace embeddings. Moreover, the quality score distributions strongly overlap and assign the asian distributions to the highest qualities. The age-bias can be explained by the used training database VGGFace2, which consists of mainly young adults. However, this does not explain the quality prediction differences for pose and ethnicity, since VGGFace2 contains more non-frontal than frontal images and contains a large variance of ethnics. The resulting bias can be better explained by the utilized embeddings. The FaceQnet model was trained on comparison scores from FaceNet embeddings based on the MS1M dataset. MS1M contains mostly frontal faces of white adults.

SER-FIQ SER-FIQ shows the best face quality assessment performance in all investigated cases, since it directly measures the quality based on the deployed face recognition model. Therefore, it is able to consider the model decision patterns including biased decisions.

This effect can be observed in all evaluated cases for face quality assessment. In all of these cases, the classes affected by the bias are strongly reduced with a growing quality threshold, while the ratio of the classes with a good face verification performance increases. This can be observed for frontal head poses, asian and black faces, and faces from individuals below 7 years. The quality score distributions in Figure 3.19 further strengthen the suspicion of bias. In all cases, the distributions are clearly separated from each other. Consequently, SER-FIQ adapts to the bias from the deployed face recognition model, which arises from the unbalanced MS1M training data. For non-demographic attributes, a potential bias transfer fulfils the task of quality estimation in a non-discriminative manner. However, for demographic attributes, SER-FIQ exactly fulfils the utility definition of face quality estimation including a discriminating bias transfer. Future works have to come up with a solution to this problem.

Summary For all evaluated face quality assessment algorithms, biased quality estimates are observed. We point out that if the face quality assessment approach is trained on face embeddings, the major influence of the quality estimation bias was observed to originate from the face embeddings, not the data for training the quality assessment algorithms. It was shown that the classes that are affected by face recognition bias are also getting lower quality assignments. The utility definition of face quality assessment causes this bias transfer and future work have to come up with a solution to this problem. This (a) might be a development of face quality assessment solution that does not adapt demographic bias or (b) strengthen the focus on bias mitigating face recognition models, since an unintended bias transfer will not happen with an unbiased face recognition model.

3.4.6. Interim Conclusion

Current definitions of face quality assessment are based on the suitability of a face image for the task of face recognition. Optimizing this suitability estimation can be achieved when the face quality assessment is built on the deployed face recognition. This leads to more robust and accurate quality predictions as recent work has shown. However, this can lead to an unintended bias transfer towards the face quality assessment including its discriminatory effects on the society. In this section, we presented a profound investigation between face recognition bias and face quality estimation. The experiments were conducted on two publicly available databases and involved four state-of-the-art face quality assessment algorithms from academia and industry and two widely-used face recognition systems. The results showed that face image quality highly correlates with demographic, as well as non-demographic, bias by demonstrating that current face quality

assessment methods already adapted the bias. Consequently, every enrolment process, as well as quality-based fusion approach, possess the bias as well. The current definition of face quality allows this bias transfer. The ethical questions concerning fairness and discrimination that arises with this definition, however, have to be discussed by future work. Possible solutions for this problem include (a) a development of face quality assessment approach that, by design, prevents a demographic bias transfer or (b) a strong focus on bias-mitigating face recognition models, since an unintended bias transfer will not happen with an unbiased face recognition model.

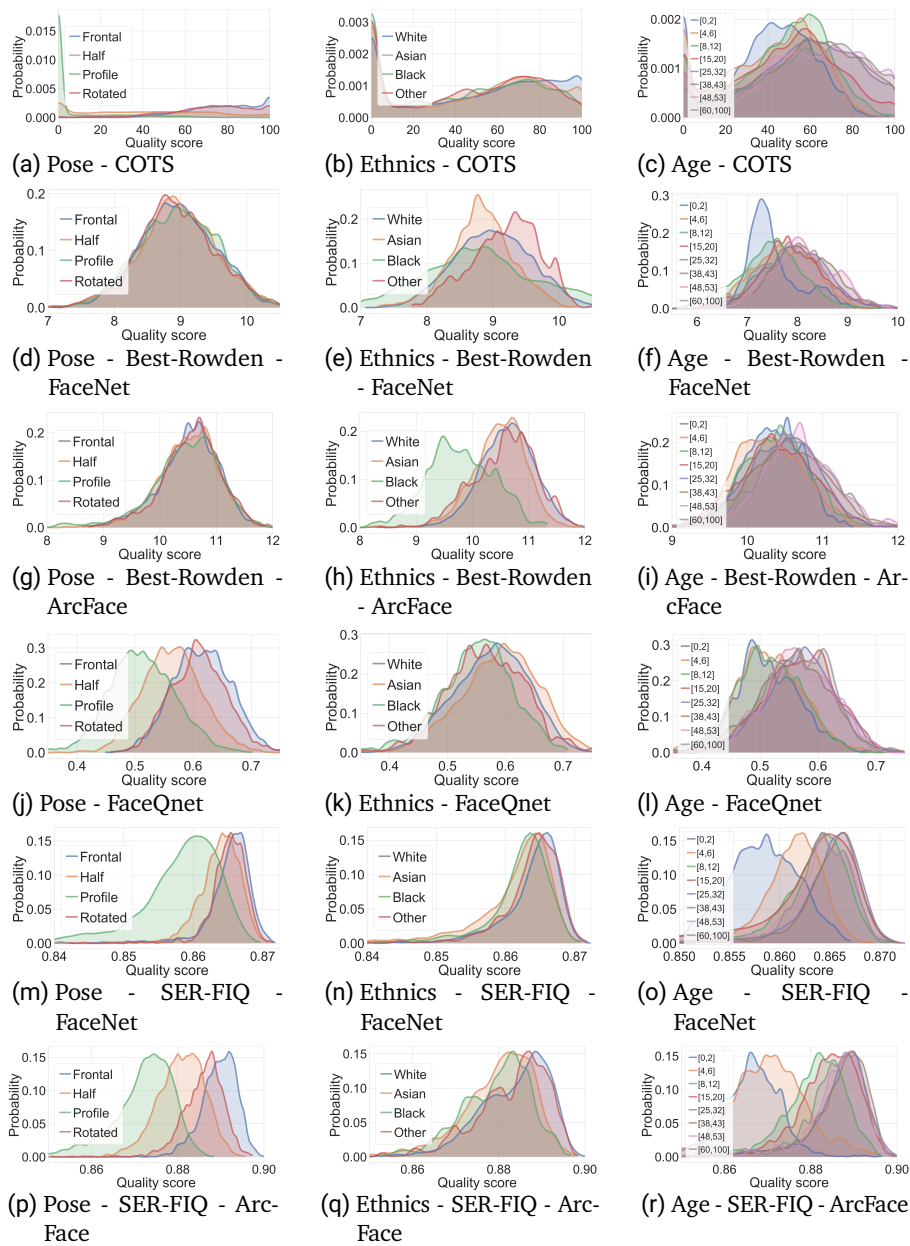


Figure 3.19.: Quality score distributions for several poses (left), ethnicities (middle), and age classes (right). The quality scores are shown for the different face quality assessment approaches. While COTS and FaceQnet work on image level, Best-Rowden and SER-FIQ are applied on FaceNet and ArcFace features.

3.5. Analysing Soft-Biometric Characteristics in Face Templates

3.5.1. Introduction

In this chapter RQ4 "What information is stored in biometric face templates?" is answered and is based on the work of Terhörst et al. [Ter+20a]. This aims at demonstrating the need for enhancing soft-biometric privacy in face recognition systems and further provides key-insights for the development of more effective solutions in Chapter 5.

The advances of deep neural representations lead to high-performing face recognition solutions [GNH18]. Due to the achieved performance, face recognition systems spread world-wide and increasingly affect our daily life [Dam+18d]. Despite that these face representations are trained to enable recognition of individuals, previous works showed that more information than just the identity is embedded. They demonstrated face templates contain information about head pose [Par+17], image characteristics (such as quality [BJ18; Her+19], viewpoint [Hil+18], and illumination [OTo+18]), demographics [DDB18; Ter+19d; ÖAE16], and social traits [Par+19]. However, for many applications, the users do not permit to have access to this information. Thus, the stored data should be exclusively used for recognition purposes [MR17], and extracting such information without a person's consent is considered a violation of their privacy [Kin13]. This problem is known as soft-biometric privacy [MR17] and solutions are either build on image- [OR14; MRR19; MRR20] or template-level [Ter+19a; Ter+20c; Ter+20h; Bor+20].

Since the knowledge about encoded attributes in face template is required to develop more advanced bias-mitigating solutions [GLJ19; Lia+19; Ter+20i; Ter+20f; Yin+19] and more comprehensive privacy-enhancing technologies, in this section, we investigate the predictability of 113 attributes from face templates at different difficulty-levels. We jointly trained a massive attribute classifier (MAC) with a high number of attributes to take advantage of a shared feature space. The MAC is modified such that it is able to accurately state its prediction reliability [Ter+19d]. This allows us to make predictions at two reliability levels and thus, to derive more fine-grained statements about the predictability of attributes in face templates. The experiments were conducted on two publicly available databases, CelebA [Liu+15] and LFW [Hua+07], and on two popular face embeddings, FaceNet [SKP15] and ArcFace [Den+19]. To derive understandable statements about the stored attribute information, we categorized each attribute into one of three predictability classes: easily-predictable, predictable, and hardly-predictable. The results show that 39 attributes are assigned to the easily-predictable class and 74 of the 113 investigated attributes are at least predictable. Despite that face templates are learned to be robust to non-permanent factors, the results demonstrate that especially these attributes are easily-predictable. This includes information about age, hairstyles, haircolors, beards, and

accessories, such as makeup, lipstick, and glasses.

3.5.2. Related Work

The development of deep neural network representations for faces led to strong performance boosts for face recognition [GNH18]. However, since these representations are derived from black-box models, it is not clear which kind of information is stored in these representations.

In 2017, Parde et al. [Par+17] investigated face representations in terms of head position and source of the image. The results demonstrated that the investigated representations contain accurate information of the yaw and pitch of a face and about whether the input-face origins from a still image or a video frame. They suggest that image-quality information might be available in these features as well. This hypothesis was proofed to be correct [Ter+20g; BJ18; Her+19]. In [Ter+20g; BJ18; Her+19], face image quality was successfully predicted based on face embeddings.

In [Par+19], Parde et al. analysed if face representations retain information in faces that supports social-trait inferences. In their experiments, they investigated 11 social traits such as talkative, assertive, shy, quiet, warm, artistic, efficient, careless, impulsive, anxious, and lazy. They trained linear classifiers to predict these human-assigned social trait profiles and demonstrated that these traits can be determined from face embeddings to a high degree. The best-predicted traits were impulsive, warm, and anxious.

Hill et al. [Hil+18] analysed the representations of caricature faces. They examined the organization of viewpoint (0° , 20° , 30° , 45° , 60°), illumination (ambient vs spotlight), gender (male vs female), and identity in the embedding space. Their results showed that the utilized face recognition model creates a highly organized, hierarchical, similarity structure in which information about face identity and imaging characteristics coexist. These results were summarized by O’Toole et al. [OTo+18]. They reviewed what properties are known about the face space and ground them in the context of previous-generation face recognition algorithms.

In [ZSL16a; ZSL16b] Zhong et al. demonstrated that the use of various mid-level representations from face recognition networks leads to highly accurate facial attribute estimation performances. This indicates that also high-level representations, such as face recognition templates, might contain a significant amount of facial attribute information. In [DDB18; Ter+19d; Bou+19; ÖAE16], it is shown that demographic attributes such as gender, age, and race can be derived from face templates.

So far, previous works showed that head pose, image characteristics (such as quality, source of the image, viewpoint, illumination), demographic attributes (gender, age, race), and social traits (e.g. impulsive, warm, and anxious) can be found in face templates.

In contrast to previous work that investigated only specific characteristics, in this section, we analyse a wide range of attributes (up to 113) in face representations. Moreover, we analyse the predictability of these attributes under different levels of prediction reliabilities. This allows us to state more generally which attributes are encoded in face templates.

3.5.3. Methodology

This section aims at analysing the set of soft-biometric information that is stored in face templates. To do so, we train a classifier to jointly predict these attributes. If the classifier can successfully predict these, we conclude that these attributes are stored in the face templates. However, this only allows us to answer the question of what information is embedded. A statement about what information is not included is not possible, because the reverse conclusion is not necessarily logical. If an estimator is not able to learn the pattern of an attribute, it does not imply that the pattern does not exist. The classifier might just not be able to deal with the complexity of the attribute pattern or the data variability and representation might be low.

To answer RQ4, the following three subsections explain the different steps of the investigation methodology. In Section 3.5.3, we will first explain the classifier training procedure that allows a joint prediction of a large number of attributes. Learning these attributes in a multi-task learning approach will enhance the performance, since many attributes share similar features. In Section 3.5.3, we explain how this classifier can accurately state its predictions confidence. This prediction confidence determines the quality of a prediction and enables us to derive predictability classes in Section 3.5.3. These predictability classes allow us to generalize our findings into easily understandable statements.

Massive attribute classifier (MAC)

To investigate what attribute-information is stored in face templates, we train a classifier model to predict multiple attributes. If the classifier can correctly predict these attributes given face templates, we can draw conclusions about what attributes are encoded in the investigated representation.

Therefore, we trained a neural network model to jointly predict multiple attributes given face templates of the training set. Due to the large number of predicted attributes, we refer to this model as the massive attribute classifier (MAC). To find an optimal network structure for our MAC, we evaluated multiple models with various number of dense layers and layer sizes. To be precise, we evaluated random network structures with 1-3 initial layers and 1-3 branch layers that connects the last initial layer with the the softmax layers

of each attribute. For each layer a size of 128, 256, and 512 was evaluated. We choose the structure with the most stable results as the layout of our MAC. However, despite the large variations in the investigated network structures, we observed that, in most cases, the predicted performance per attribute only varies within a range of 1-2%.

The chosen MAC-network consists of two initial layers, the input layer of size n_{in} and the second dense layer of size 512. Here, n_{in} refers to the size of the utilized face embedding. Starting from the second layer, each attribute a has an own branch consisting of two additional layers of size 512 and $n_{out}^{(a)}$, where $n_{out}^{(a)}$ refers to the number of classes per attribute. Each layer has a ReLU activation, except for the output-layers, which have softmax activations. Moreover, Batch-Normalization [IS15] and dropout [Sri+14] with a dropout-probability of $p_{drop} = 0.5$ is applied to every layer. The dropout allows to generalize the performance, but also enables us to derive reliability statements about the predictions (described in Section 3.5.3). The training of the MAC was done in a multi-task learning fashion by applying a categorical cross-entropy loss for each attribute branch and use an equal weighting between each of these attribute-related losses. For the training, an Adam optimizer [KB14] was used with $e = 200$ epochs, an initial learning rate $\alpha = 10^{-3}$, and a learning-rate decay of $\beta = \alpha/e$. These parameter choices are guided by [Ter+19d]. The batch size b was chosen according to the amount of data available, $b = 1024$ for CelebA and $b = 16$ for LFW.

Reliability statements

To derive statements about the predictability of an attribute in a face template, we use prediction reliabilities to simulate close-to-optimal classifier circumstances. Therefore, we follow the methodology in [Ter+19d; Ter+19c] to enable our MAC to state its prediction confidence (reliability). Following this approach, we trained the MAC with dropout. To derive a reliability statement additionally to an attribute prediction, $m = 100$ stochastic forward passes are performed. In each forward pass, a different dropout-pattern is applied, resulting in m different softmax outputs $v_i^{(a)}$ for each attribute a . Given the outputs of the m stochastic forward passes of the predicted class \hat{c} denoted as $x^{(a)} = v_{i,\hat{c}}^{(a)}$, the reliability measure is given as

$$rel(x^{(a)}) = \frac{1 - \alpha}{m} \sum_{i=1}^m x_i^{(a)} - \frac{\alpha}{m^2} \sum_{i=1}^m \sum_{j=1}^m |x_i^{(a)} - x_j^{(a)}|,$$

with $\alpha = 0.5$, following the recommendation in [Ter+19d]. The first part of the equation is a measure of centrality and utilizes the probability interpretation of the softmax output. A higher value can be interpreted as a high probability that the prediction is correct. The

second part of the equation is the measure of dispersion and quantifies the agreement of the stochastic outputs x . In [Ter+19d], this was shown to be an accurate reliability measure.

We use this reliability measure to simulate more idealistic circumstances. For each attribute, we calculate the prediction and corresponding reliability of each instance. Then we take the predictions of 100% and 50% of the highest reliabilities to evaluate the performance. This performance refers to the ratio of considered predictions (RCP) of 100% and 50%. The performance at 100% RCP refers to the general performance of the whole dataset. The performance at 50% RCP refers to the performance on the predictions with 50% of the highest reliabilities. Consequently, this refers to the performance based on the prediction on which the MAC is most confident about. The unconsidered 50% of the predictions might contain factors of variances (such as blur, non-frontal head poses) that lead to unstable, and thus inaccurate, attribute estimates.

Predictability classes

To derive more understandable statements about which attribute information is stored in a face template, we categorize each attribute into one of three predictability classes:

- **Easily-predictable (+ +):** an attribute is categorized as easily-predictable if, and only if, the balanced accuracy at 100% RCP is above 90%. This means that *highly accurate predictions are possible even under non-ideal circumstances* such as bad illuminations and non-frontal head poses.
- **Predictable (+):** an attribute is categorized as predictable if, and only if, the balanced accuracy at 100% RCP is under 90%, but the balanced accuracy at 50% RCP is above 90%. This indicates that *highly accurate predictions are possible under close-to-optimal conditions*, since it only takes into account 50% of the most confident MAC predictions.
- **Hardly-predictable (0):** an attribute is categorized as hardly-predictable if the balanced accuracy is below 90% at both, 100% and 50% RCP. Even *under close-to-optimal circumstances, the MAC is not able to reach high accuracies*. Consequently, the attribute patterns might be too complex for the MAC to handle or it does not exist a meaningful pattern for this attribute.

While the first two categorizes (Easily-predictable and Predictable) allow making confident statements about the amount of attribute information in face templates, the same does not apply for the third category (Hardly-predictable). The last category only states that



Figure 3.20.: Sample images from CelebA (top row) and LFW (bottom row)

the classifier is not able to accurately learn the pattern, but this might be due to several reasons: (1) the pattern does not exist, (2) the pattern does exist, but it is too complex for the model to learn, or (3) the pattern does exist but the amount of data and its representation is not appropriate for the classifier to learn. Consequently, for the third case, we can not determine if the attribute pattern exists.

3.5.4. Experimental Setup

Databases

For the analysis of the face space, we chose the Labeled Faces in the Wild (LFW) [Hua+07] and the CelebFaces Attributes (CelebA) [Liu+15] datasets because of their large and rich attribute annotations. The large number of different soft-biometric labels allows to deeply investigate which of these attributes are encoded in face templates. Figure 3.20 shows sample images from both datasets. The CelebA dataset [Liu+15] is a large-scale dataset with more than 200k images of over 10k celebrities. It covers large variations in pose and background. Moreover, each image is labelled with 40 binary attributes. LFW [Hua+07] contains over 13k images from over 5k individuals and exhibits variability in pose, lighting, focus, resolution, facial expression, age, gender, race, accessories, make-up, occlusions, background, and photographic quality. The face images are 250x250 pixels and mostly in color. Each image is annotated with up to 73 attributes.

The attribute labels of both databases [Hua+07; Liu+15] cover a wide range of characteristics such as the person’s demographics, skin, hair, beard, face geometry, periocular area, mouth, nose, accessories, and environment.

Cleaning attribute labels

In contrast to CelebA, where the attribute labels are of binary nature, in LFW, the labels come from the prediction probabilities of a binary classifier [Hua+07]. Each label value

measures the degree of the attribute and thus, are continuous [Kum+09; Kum+11]. E.g. for the attribute male, a higher label score indicates that the person appears more masculine than a person with a lower label score. Consequently, the top rank images for an attribute represent the label true, while the lowest rank images indicate the label false. A value around zero means that the corresponding attribute has little meaning on this image.

To make sure that our MAC performs well when training on LFW, we manually converted the continuous attribute labels to binary labels. Therefore, we assigned an upper and lower score threshold for each attribute. Images with a score over the upper threshold are assigned as true, images with a score under the lower threshold are assigned as false, images with scores within the range are assigned as undefined. The upper and lower thresholds for one attribute are manually determined by moving potential thresholds away from zero. At each potential threshold, ten images with the closest attribute scores are investigated. Here, the original LFW labels of the images are manually investigated for correctness. If only eight or fewer attributes are investigated as correct, the potential threshold is further moved away from the starting point and the procedure is repeated. If a potential threshold returns images with 9 or more correct labels, it is chosen as the limit. Repeating this over all attributes will result in a lower and an upper threshold for each of these attributes. By binaryzing the scores with these upper and lower thresholds, we ensure an error-minimizing data basis of the MAC. This allows us to train and test on meaningful and correctly labelled data.

Please note that the label-cleaning process reduces the amount of used labels by 51,7% that might induce a bias in our evaluation. To avoid biased conclusions that might result from this process, we evaluate on another binary labelled database. After the label-cleaning, we found 15 attribute labels of either a low number of positively and negatively labelled samples (<100). These are listed in Table 3.12 with the number of positively and negatively labelled samples in the test and training set. We will mark these attributes (in grey) in the following investigations to consider their low expressiveness during the face analysis.

Evaluation metrics

In this work we derive what information is contained in the face templates based on prediction accuracies. In machine learning, accuracy is defined by the ratio of the number of correct predictions to the total number of predictions [Mur13]. To be robust to attribute-imbbalances, we report the prediction performance in terms of balanced accuracy. This refers to the standard accuracy with class-balanced sample weights [KND15].

The train/test data is defined by dividing the databases in a 70%/30% subject-exclusive

Table 3.12.: Train/test sample distribution on LFW for *selected attributes* that are found insufficient for a meaningful attribute analysis *after label-cleaning*. Pos and Neg refers to the number of positively and negatively labelled samples for the train and test set. The listed 15 attributes are found to be insignificant for the analysis due to a low number of samples in either the positive or negative class.

Attribute	Train		Test	
	Pos	Neg	Pos	Neg
Color Photo	8806	29	3772	24
Mouth Slightly Open	674	109	315	57
Round Face	9	588	3	250
Goatee	20	3346	10	1557
Baby	23	9137	15	3913
Bangs	89	5238	44	2080
Bald	114	4413	47	1953
Big Lips	101	751	48	318
Sunglasses	74	8583	50	3631
Partially Visible F.	124	1501	55	601
Mouth Wide Open	107	6593	56	2925
Double Chin	154	172	57	136
Harsh Lighting	113	914	62	487
Outdoor	173	510	63	243
Teeth Not Visible	125	2209	66	1089

split. To analyse the prediction performance of an estimator under more ideal circumstances, we chose a classifier for the attribute prediction task that is additionally able to accurately state its prediction confidence. For each face template, this classifier predicts the associated attributes and their prediction reliabilities. To get the prediction performance under more ideal circumstances, for each attribute, only the predictions with 50% of the highest reliabilities are considered for the balanced accuracy. This balanced accuracy refers to a ratio of considered predictions (RCP) of 50%. Since this relates to the MAC prediction confidence, the balanced accuracy should be higher at lower RCP-levels.

Face template extraction

In this work, we utilize two widely-used face recognition models, FaceNet [SKP15] and ArcFace [Den+19]. We use pre-trained models trained on the MS1M database [Guo+16] for both networks, FaceNet¹² and ArcFace¹³. To get the face template for a given face image, the image has to be aligned, scaled, and cropped. For FaceNet, the preprocessing is done as described in [KS14]. For ArcFace, we follow the preprocessing as described in [Guo+18]. The preprocessed image is passed to a face recognition model to extract the embeddings. The output size is 128 for FaceNet and 512 for ArcFace.

Investigations

This work aims at understanding what kind of soft-biometric information is stored in face templates. Therefore, our investigations are divided into three parts:

1. We validate the attributes labels of both datasets by studying the correlations between the attributes.
2. We analyse what attributes are contained in face representations by investigating the attribute prediction performances on both datasets and face embeddings. To get a more complete perspective on the problem, the prediction performances on different confidence-levels of the classifier are investigated.
3. We obtain an overview of which kind of information is encoded in face templates by categorizing each attribute into one of three predictability classes based on their two-level prediction performances.

3.5.5. Results

This section is divided into three subsections, each focusing on one investigation point: (1) analysis of the attribute correlation, (2) investigation of the attribute predictability, and (3) summarize findings.

Attribute-correlation analysis

To understand the quality of the labels and potential biases in the attribute labels, Figure 3.21 shows a selection of attribute-label correlations. The attributes are chosen to show

¹²<https://github.com/davidsandberg/facenet>

¹³<https://github.com/deepinsight/insightface>

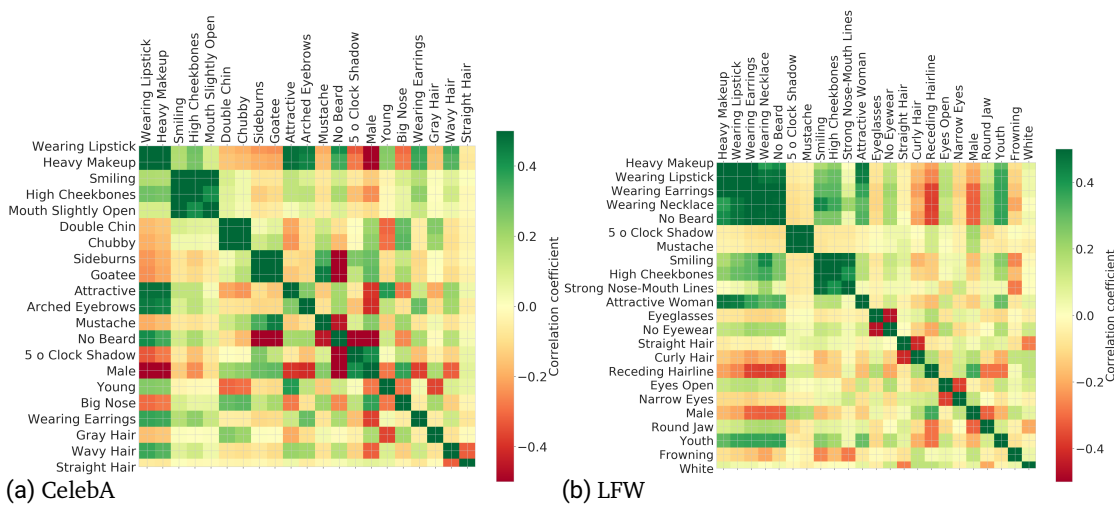


Figure 3.21.: Label-correlation for CelebA and LFW. The attributes are chosen to show the 15 most positive and negative pairwise correlations. The attribute-correlation for LFW is shown after the label-cleaning process. Green indicate positive correlations, while red indicate a negative correlation. The correlation is based on the Pearson coefficient.

the 15 most positive and negative pairwise correlations. For CelebA, the correlation in Figure 3.21a shows that the large majority of male faces in the database do not wear lipstick, earrings, and makeup. These attributes mostly belong to female faces. Moreover, it shows some biases in the database labels. The majority of male faces have a beard. If a face is labelled as *attractive*, it belongs to a young female face most likely wearing accessories and makeup. However, this figure also approves the quality of some labels, e.g. *No Beard* negatively correlates with all kinds of beards such as *Sideburns*, *Goatee*, and *Mustache*.

Figure 3.21b shows the attribute correlation for LFW. It shows that the attributes *Heavy Makeup*, *Wearing Lipsticks*, *Wearing Earrings*, and *Wearing Necklace* belongs together with *Youth* and *Attractive Woman*, *Smiling*, and *High Cheekbones*. Moreover, this set of attributes does not correlate with a *Receding Hairline* and *Male*. Nevertheless, it also approves the quality of other labels such as *No Eyewear* (negatively correlates with *Eyeglasses*) and *Curly Hairs* (negatively correlates with *Straight Hair*).

Attribute-analysis of the face space

100% RCP (hard) refers to the use of all samples under the given circumstances. 50% RCP (easy) refers to the 50% the predictions of which the classifier is most sure about its correctness. In Table 3.13 the prediction performance is shown for CelebA including the assigned predictability classes. Two general observations are made. First, the performance at the 50% RCP-level is always higher than for 100% RCP showing that MAC learned reliable predictions on the dataset. Second, even if the prediction performance on FaceNet (FN) and ArcFace (AF) is very similar, the performance on FN is always slightly higher. This can be explained by ArcFace’s margin-principle during training that distorts the feature space more incoherently and thus, makes it harder for pattern learning. In total, many of CelebA attributes can be predicted with high accuracy from face templates. This includes demographic characteristics such as gender, characteristics of the person’s hairstyle, haircolor, and beard. Moreover, the deeply encoded features also contain highly-detailed information about the person’s accessories, such as necklace and earrings.

Table 3.14 shows the same evaluation setting on the LFW database. The grey highlights refer to results with limited significance since the label-cleaning process eliminated many samples with low-quality labels. The low number of train- and testing-samples explains some of the weak performance such as for *Baby*, *Sunglasses*, and the *Mouth* category. However, comparing the results of LFW with the results of CelebA (Table 3.13) shows similar performances on attributes which occur in both datasets, such as demographic attributes, haircolors, face geometry etc. Consequently, our label-cleaning process removed low-quality attribute-labels but did not result in a large bias of the data. Due to the entangled patterns encoded in the templates some attributes, such as *Bold*, *Bangs*, and *Goatee*, are easy to learn and thus, achieve high performances. Generally, the prediction performance using ArcFace embeddings is significantly weaker than using FaceNet. ArcFace embeddings contain more complex attribute patterns and for the experiments on LFW less data was available for training, since we manually filtered low-quality labels. Consequently, it can be expected that with more training data the performance on ArcFace is better. Nevertheless, similar to CelebA, many attributes can be predicted with high accuracies from the templates only. This goes for demographic attributes such as gender, age, and race, as well as for hairstyle, haircolor, beard, and accessories. Moreover, characteristics about the face geometry such as face shape, double chin, and forehead visibility can be determined. Factors that do not belong to the person, such as lighting conditions and blurriness, can not be predicted reliably with the MAC. It is interesting to note that the high predictability of *Attractive woman* can be explained by the high correlation to accessories.

Table 3.13.: Prediction performance on CelebA: the performance is based on FaceNet (FN) and ArcFace (AF) embeddings and is reported in terms of balanced accuracies at two difficulty scenarios: 100% RCP (hard) and 50% RCP (easy). ++, +, and ⁰ state the assigned predictability class.

Attribute	100% RCP		50% RCP		
	FN	AF	FN	AF	
Demo	Male ⁺⁺	98.9%	98.4%	99.9%	99.9%
	Young ⁺	85.5%	83.6%	96.4%	94.5%
Skin	Pale Skin ⁰	76.0%	71.9%	87.1%	83.0%
	Rosy Cheeks ⁺	83.4%	78.2%	96.3%	81.7%
Hairstyle	Bald ⁺⁺	95.7%	94.0%	100.0%	100.0%
	Bangs ⁺⁺	91.7%	89.3%	99.4%	98.3%
	Receding Hairline ⁺	85.4%	82.5%	96.4%	94.2%
	Sideburns ⁺⁺	92.8%	92.1%	90.0%	99.7%
	Straight Hair ⁰	68.6%	70.7%	79.9%	82.0%
	Wavy Hair ⁰	74.4%	76.6%	86.4%	89.4%
Haircolor	Black Hair ⁺	83.7%	81.5%	96.6%	94.3%
	Blond Hair ⁺⁺	91.9%	90.1%	99.3%	98.3%
	Brown Hair ⁺	76.5%	75.9%	90.1%	88.3%
	Gray Hair ⁺⁺	93.0%	91.1%	99.6%	98.8%
Beard	5 o Clock Shadow ⁺	86.9%	85.8%	99.6%	99.0%
	Goatee ⁺⁺	93.4%	91.8%	97.2%	98.9%
	Mustache ⁺⁺	92.2%	89.7%	100.0%	98.8%
	No Beard ⁺⁺	92.1%	90.8%	99.4%	99.0%
Face Geo.	Chubby ⁺	86.5%	83.1%	96.5%	95.4%
	Double Chin ⁺	86.6%	82.9%	96.9%	95.4%
	High Cheekb. ⁺	78.5%	72.2%	91.6%	82.6%
	Oval Face ⁰	63.4%	61.9%	70.8%	68.1%
Periocular	Arched Eyebrows ⁺	79.8%	77.0%	93.3%	89.5%
	Bags Under Eyes ⁰	72.1%	70.7%	80.6%	80.7%
	Bushy Eyebrows ⁺	83.4%	78.5%	95.9%	91.9%
	Narrow Eyes ⁰	66.5%	60.7%	75.4%	66.7%
Mouth	Big Lips ⁰	74.6%	68.8%	86.4%	78.7%
	Mouth Slightly Open ⁰	74.5%	67.5%	86.5%	76.5%
	Smiling ⁺	80.1%	71.7%	92.9%	82.1%
Nose	Pointy Nose ⁰	71.7%	69.3%	83.1%	78.9%
	Big Nose ⁰	77.4%	75.8%	88.1%	87.1%
Accessories	Eyeglasses ⁺⁺	97.3%	90.6%	99.8%	98.7%
	Heavy Makeup ⁺⁺	90.1%	88.7%	99.2%	98.5%
	Wearing Earrings ⁺	79.2%	77.0%	94.8%	91.6%
	Wearing Hat ⁺⁺	95.4%	92.8%	99.4%	99.0%
	Wearing Lipstick ⁺⁺	92.8%	91.4%	99.4%	98.7%
	Wearing Necklace ⁰	71.8%	71.4%	86.9%	84.2%
Other	Wearing Necktie ⁺	83.7%	82.1%	98.5%	98.0%
	Blurry ⁰	74.3%	68.2%	85.2%	78.4%
	Attractive ⁺	79.6%	77.9%	92.4%	89.6%

Table 3.14.: Prediction performance on LFW: the performance is based on FaceNet (FN) and ArcFace (AF) embeddings and is reported in terms of balanced accuracies at two difficulty scenarios: 100% RCP (hard) and 50% RCP (easy). ++, +, and ⁰ state the assigned predictability class. Grey highlighting refers to reduced expressiveness due to limited data after the label-cleaning process.

		100% RCP		50% RCP				100% RCP		50% RCP		
Attribute		FN	AF	FN	AF	Attribute		FN	AF	FN	AF	
Demographics	Male ⁺⁺	98.3%	83.9%	99.5%	94.2%	Periocular	Eyes Open ⁰	60.4%	54.4%	63.6%	54.8%	
	Baby ⁰	55.1%	49.9%	50.0%	50.0%		Brown Eyes ⁺	82.1%	64.0%	92.8%	66.8%	
	Child ⁰	68.8%	57.5%	75.8%	52.4%		Bags Under Eyes ⁺	87.2%	73.7%	95.4%	83.5%	
	Youth ⁺	79.9%	70.5%	93.1%	79.8%		Narrow Eyes ⁰	77.1%	66.2%	86.3%	74.1%	
	Middle Aged ⁺	88.4%	74.0%	95.2%	82.9%		Bushy Eyebrows ⁺⁺	96.3%	83.8%	99.1%	91.7%	
	Senior ⁺⁺	99.6%	83.9%	100.0%	88.4%		Arched Eyebrows ⁺	85.3%	71.6%	94.5%	76.8%	
	Asian ⁺⁺	95.5%	66.2%	100.0%	69.6%		Mouth	Mouth Closed ⁰	73.2%	64.0%	83.9%	72.4%
	White ⁺⁺	97.4%	73.6%	99.4%	81.4%			Mouth Slightly Open ⁰	73.8%	61.8%	83.0%	65.1%
	Black ⁺⁺	95.3%	63.2%	98.3%	53.6%			Mouth Wide Open ⁰	66.6%	50.8%	59.9%	50.0%
Indian ⁺	85.2%	50.2%	92.5%	54.7%	Teeth Not Visible ⁰	70.0%		65.2%	75.3%	58.3%		
Skin	Rosy Cheeks ⁰	67.2%	58.8%	73.0%	64.3%	Smiling ⁰		72.0%	67.9%	81.3%	75.9%	
	Shiny Skin ⁰	82.1%	67.9%	89.7%	75.6%	Big Lips ⁺		87.6%	57.3%	98.0%	57.8%	
	Pale Skin ⁰	68.0%	62.9%	79.9%	67.2%	Nose	Big Nose ⁺	84.5%	71.6%	93.6%	81.5%	
	Flushed Face ⁰	66.5%	55.5%	77.5%	52.3%		Pointy Nose ⁺⁺	96.5%	71.5%	100.0%	71.3%	
Hairstyle	Curly Hair ⁰	69.0%	61.7%	77.8%	68.7%		Nose-Mouth Lines ⁰	70.0%	61.7%	80.7%	71.6%	
	Wavy Hair ⁺⁺	95.0%	80.5%	99.7%	83.3%	Accessories	Heavy Makeup ⁺⁺	96.7%	69.9%	99.0%	57.1%	
	Straight Hair ⁰	67.5%	59.8%	76.8%	65.5%		Wearing Hat ⁺	87.2%	67.9%	96.9%	53.8%	
	Receding Hairline ⁺	83.3%	73.0%	93.5%	84.9%		Wearing Earrings ⁺⁺	91.7%	73.3%	97.9%	72.9%	
	Bald ⁺⁺	93.6%	75.8%	97.9%	75.0%		Wearing Necktie ⁺	84.6%	72.8%	93.5%	75.2%	
	Bangs ⁺⁺	97.0%	64.1%	100.0%	50.0%		Wearing Necklace ⁺	83.7%	74.1%	92.1%	82.5%	
Sideburns ⁺⁺	98.9%	84.1%	99.7%	89.2%	Wearing Lipstick ⁺⁺		98.5%	75.9%	99.5%	74.0%		
Haircolor	Black Hair ⁺⁺	90.4%	65.6%	96.5%	61.5%	No Eyewear ⁺⁺	95.5%	86.1%	98.2%	90.3%		
	Blond Hair ⁺⁺	95.2%	71.7%	98.8%	55.6%	Eyeglasses ⁺⁺	96.1%	90.0%	98.4%	95.6%		
	Brown Hair ⁺	81.5%	71.9%	91.9%	82.7%	Sunglasses ⁰	71.6%	50.8%	62.4%	50.0%		
	Gray Hair ⁺⁺	98.8%	88.4%	100.0%	93.9%	Environment	Blurry ⁰	61.4%	57.2%	66.3%	58.6%	
Beard	No Beard ⁺⁺	98.1%	83.9%	100.0%	92.1%		Harsh Lighting ⁰	76.0%	61.3%	89.1%	57.9%	
	Mustache ⁺⁺	98.5%	79.7%	99.3%	78.1%		Flash ⁰	78.3%	58.3%	88.3%	51.5%	
	5 o Clock Shadow ⁺⁺	96.5%	83.8%	99.6%	92.4%		Soft Lighting ⁰	65.7%	60.2%	72.3%	66.1%	
	Goatee ⁺⁺	94.5%	70.0%	100.0%	100.0%		Outdoor ⁰	77.2%	60.8%	81.9%	65.9%	
	Face Geometry	Oval Face ⁺	82.7%	71.6%	95.4%	75.8%	Others	Frowning ⁰	78.3%	72.4%	88.8%	79.5%
Square Face ⁺⁺		99.1%	89.1%	100.0%	96.3%	Color Photo ⁰		72.8%	54.0%	75.0%	60.0%	
Round Face ⁺		84.2%	49.6%	100.0%	50.0%	Posed Photo ⁰		76.0%	60.7%	80.9%	63.0%	
Round Jaw ⁰		70.6%	60.8%	81.1%	58.4%	Attractive Man ⁰		74.4%	65.0%	85.1%	74.2%	
Double Chin ⁺⁺		91.5%	81.1%	100.0%	88.7%	Attractive Woman ⁺⁺		95.3%	75.1%	100.0%	71.4%	
High Cheekbones ⁺		79.9%	73.3%	90.4%	81.8%							
Chubby ⁺		85.5%	74.3%	98.0%	79.4%							
Obstructed Forehead ⁺		85.9%	65.0%	99.9%	61.3%							
Partially Visible F. ⁺		85.2%	65.9%	94.0%	50.0%							
Fully Visible F. ⁺		85.9%	71.8%	95.4%	82.2%							

Summary

From 113 investigated attributes, we found that 39 attributes belong to easily-predictable, 35 belong to predictable and 39 to hardly-predictable. To obtain a more general overview of the encoded information in face templates, Table 3.15 summarizes the categories of the attributes in the three predictability classes. The assignment of the categories to the individual attributes is shown in Table 3.14. Providing a more complete view of the problem, this table also includes findings from related works. Since the face templates are trained with the purpose of recognition, it seems logical that categories such as *Face Geometry*, *Periocular Area*, *Nose*, and *Mouth* are easily-predictable. Surprisingly, this is not the case. Instead, non-permanent factors such as *Hairstyle*, *Haircolor*, *Beard*, *Accessories*, *Head Pose*, and *Social Traits* are easily-predictable. Modern face recognition systems aim to be robust against these factors and still, these factors are strongly present in face templates.

For many applications, the user of a face recognition system solely provides his biometric data for recognition. To prevent a function creep of his data, face templates should contain only identity-related information. However, the experiment showed that many privacy-sensitive attributes are encoded in face templates. This raises a major privacy risk. Consequently, future works might analyse the reason for this rich encodings and find solutions to preserve privacy in face recognition systems.

Table 3.15.: Categorized summary of the predictability classes including findings of related works.

Easily-predictable	Predictable	Hardly-predictable
Demographics	Face Geometry	Skin
Hairstyle	Periocular	Mouth
Haircolor	Nose	Environment
Beard	Image Quality [BJ18]	
Accessories		
Head Pose [Par+17]		
Social Traits [Par+19]		

3.5.6. Interim Conclusion

The success of current face recognition systems is based on the advances of deeply-learned templates. Recent works have shown that demographics, image characteristics, and social traits are encoded in these templates. This can lead to biased decisions in face recognition

systems and raises major privacy issues. In many applications, these templates are expected to be used for recognition purposes only and deducing information that is not required for recognition is considered as a violation of their privacy. The knowledge of the encoded information in face templates is necessary to develop effective bias-mitigating and privacy-preserving technologies. The main contribution of this section is an analysis of what information is stored in face templates. This aims at answering RQ4. More precisely, 113 attributes are analysed towards their predictability from face templates. The experiments were conducted on two popular face templates under two difficulty-levels. To facilitate the understandability of the results, each attribute was further categorized into one of three predictability classes. Results reveal that about one third of the analysed attributes are easily-predictable, another third is predictable, and one third is hardly-predictable. Despite that face recognition templates are trained to be robust against non-permanent factors, the results demonstrate that especially these attributes are accurately predictable from face templates. In the Chapters 4 and 5, the knowledge of this analysis is used to develop comprehensive bias-mitigating and privacy-preserving solutions for face recognition.

3.6. Summary

In this chapter, soft-biometric driven bias and privacy concerns in face recognition systems were investigated. Therefore, two preliminary works were introduced before analysing these issues in more detail. First, a novel reliability measure [Ter+19d] was proposed that allows to accurately quantify a model's prediction reliability. This measure is used on further investigations and aims at answering RQ1. Second, a large-scale face annotations dataset, MAAD-Face [Ter+20b], was proposed with the use of the novel reliability measure. MAAD-Face contains a large number of high-quality attribute annotations associated with the face images. Unlike related databases, this allows analysing face recognition bias over a wide range of soft-biometric attributes.

Answering RQ2, the influence of soft-biometric attributes on the behaviour of face recognition systems was analysed [Ter+21b]. Previous works showed that the performance of face recognition systems is strongly dependent on the user-demographics. Our investigations demonstrated that a wide range of soft-biometrics attributes strongly affects the recognition performance. Moreover, we demonstrated that this biased behaviour also appears in the quality assessment of face images [Ter+20e]. Consequently, bias-mitigating solutions are needed for face recognition systems that are not limited to the pre-defined demographic attributes. In Chapter 4, approaches are proposed to solve this problem.

Answering RQ4, it was investigated what information is stored in biometric face templates [Ter+20a]. This aims at analysing soft-biometric privacy concerns in face recognition. Previous works showed that pose, image characteristics, demographics, and social traits are embedded in biometric face templates. We investigated a wider range of soft-biometric attributes that might be stored in face embeddings. Using a massive attribute classifier and the proposed reliability measure, 113 attributes were analysed and assigned with a predictability class. It was shown that many soft-biometric attributes are embedded in face templates. Although face templates are learned to be robust to non-permanent factors, the results demonstrate that especially these are easily-predictable. However, the stored data should be exclusively used for recognition purposes and extracting such information without consent is considered as a violation of their privacy. Consequently, soft-biometric privacy-enhancing solutions for this problem are needed and proposed in Chapter 5.

4. Integrable Bias-Mitigation

4.1. Introduction

Face recognition systems are spreading worldwide and have a growing effect on everybody's daily life. Moreover, these systems are involved in critical decision-making processes, such as in forensics and law enforcement [Dam+18d]. However, current biometric solutions are based on deeply-learned features that are mainly optimized for maximum accuracy [JNR16]. Consequently, many biometric systems show a strong biased performance, such as for certain demographics [Orc16; AZN18; FPO02; Phi+11; BG18; Gar+16; Dro+20].

To prevent discriminatory decisions, several regulations were introduced, such as Article 14 of the European Convention on Human Rights and Article 7 of the Universal Declaration of Human Rights. These aim to ensure individuals the right to non-discrimination. Moreover, the General Data Protection Regulation (GDPR) [VB17] aims at preventing discriminatory effects (article 71). Despite these regulatory efforts, several works [Phi+11; BG18; Orc16; AZN18; FPO02; Gar+16] showed that open-source [Orc16], as well as commercial [BG18] face recognition systems, are strongly biased towards different demographics.

Previous works tried to solve this problem by learning less-biased face embeddings [GLJ19; Lia+19; WD19; Hua+18; Wan+19; Kor+19; Yin+19]. However, this kind of approaches require a computationally expensive replacement of every template in the database and can not be integrated into existing systems that only store the face templates of enrolled individuals. Consequently, more integrable solutions are needed.

In this chapter, we propose two bias-mitigating solutions that operate beyond template-level. More precisely, we propose bias-mitigating solution on comparison- and score-level. This allows an easy integration of these solutions into existing face recognition systems without the need for a full database replacement. Section 4.2, provides a summary of related works on bias-mitigation in face recognition including the proposed approaches. Afterwards, the proposed solutions are presented chronologically in the following sections.

- Section 4.3: **Fair Template Comparison (FTC)** [Ter+20i] is the first bias-mitigating solution that works on the comparison-level of a biometric system. Replacing the

systems similarity function by a fairness-driven network that is trained with a novel penalization term allows reducing bias in the decision process. The proposed penalization term allows to include the notation of individual and group fairness during training. This forces the score distributions of different ethnicities to be more similar. The results showed that especially the notation of individual fairness leads to bias reduction rates between 15.35% and 52.67%, while it preserves a high recognition ability.

- Section 4.4: **Fair Score Normalization (FSN)** [Ter+20f] is the first bias-mitigation approach that operates on the score-level of a biometric system. In contrast to previous works, this unsupervised solution is specifically designed to jointly reduce the effect of bias of unknown origins and enhances the overall recognition performance. FSN builds on the notation of individual fairness and thus, aims at treating similar individuals similarly. The results demonstrate that the proposed solution mitigates bias by up to 82.7%. Moreover, it mitigates bias more consistently than existing works and, in contrast to these approaches, enhances the overall recognition performance by 53.2% at an FMR of 10^{-3} and by 82.9% at an FMR of 10^{-5} .

4.2. Related Work

The phenomena of bias in biometrics was found in several disciplines such as presentation attack detection [Fan+20] and the estimation of facial characteristics [Ter+19c; Ter+19d; DDB18]. In face biometrics, bias might be induced by non-equally distributed classes in training data [Kor+19; Hua+18]. Klare et al. [Kla+12] showed that the performance of face recognition algorithms is strongly influenced by demographic attributes. In [BG18; Orc16], the authors came to the same conclusions for commercial and open-sources face recognition algorithms. They demonstrated that the person's gender and ethnicity strongly determines their face recognition performance.

These findings motivated research towards mitigating demographic-bias in face recognition approaches. For more unbiased face recognition, Zhang and Zhou [ZZ10] formulate the face verification problem as a multiclass cost-sensitive learning task and demonstrated that this approach can reduce a different kind of faulty decisions of the system. In 2017, range loss [Zha+17] was proposed to learn robust face representations that can deal with long-tailed training data. It is designed to reduce overall intrapersonal variations while enlarging interpersonal differences simultaneously. Recent works published in 2019 aimed at mitigating bias in face recognition through adversarial learning [GLJ19; Lia+19], margin-based approaches [WD19; Hua+18], data augmentation [Wan+19; Kor+19;

Yin+19], metric-learning [Ter+20i], or score normalization [Ter+20f].

In [GLJ19], Gong, Liu, and Jain proposed de-biasing adversarial network. This network consists of one identity classifier and three demographic classifiers (gender, age, race) and aims at learning disentangled feature representations for unbiased face recognition. Liang et al. [Lia+19] proposed a two-stage method for adversarial bias mitigation. First, they learn disentangled representations by a one-vs-rest mechanism and second, they enhance the disentanglement by additive adversarial learning.

Also margin-based approaches were proposed to reduce bias in face recognition systems. In [WD19], Wang et al. applied reinforcement learning to determine a margin that minimizes ethnic bias. Huang et al. [Hua+18] proposed a cluster-based large-margin local embedding approach to reduce the effect of local data imbalance and thus, aims at reducing bias coming from unbalanced training data.

Finally, data augmentation methods were presented for fairer face recognition. In [Wan+19], Wang et al. proposed large margin feature augmentation to balance class distributions. Kortylewski et al. [Kor+19] proposed a data augmentation approach with synthetic data generation and Yin et al. [Yin+19] proposed a center-based feature transfer framework to augment under-represented samples.

So far, previous work mainly focused on learning less-biased face representations. However, mitigating the bias of a real face recognition system with one of these approaches will require a computationally expensive template-replacement of the whole database. Furthermore, it requires that for every enrolled individual a face image is additionally stored to the persons face template. For many face recognition systems, this is not the case [Dey+14; SRB16]. Consequently, more easily-integrable solutions are needed. Therefore, we propose two bias-mitigating solutions that work beyond template-level.

In [Ter+20i], we propose the first bias-mitigating face recognition approach that operates on the comparison-level. Learning a fairness-driven comparison metric, our solution includes the notations of individual and group fairness into the system's decision process. The results demonstrate that especially the notation of individual fairness is able to effectively reduce ethnic bias.

In [Ter+20f], we propose the first bias-mitigating face recognition on the score-level of a biometric system. Our novel fair score normalization approach aims at treating similar individuals more similarly and thus, more fairly. In contrast to previous works, this unsupervised solution jointly (a) reduces the biases of unknown origins and (b) enhance the overall recognition performance.

4.3. Mitigating Bias on Comparison-Level

In this section, we propose the first bias-mitigating face recognition approach that operates on the comparison-level of a biometric recognition system [Ter+20i]. Previous works solely focused on learning less-bias face templates. However, integrating one of these approaches in existing systems is a computationally expensive task that further requires stored face images for each enrolled identity. Consequently, we propose a fair template comparison approach that aims at mitigating ethnic-bias in a face recognition system.

Our contribution is a fairness-driven neural network model that is used to determine if two face templates belong to the same identity or not. This fair comparator is trained with two novel loss functions that introduce two fairness criteria into the decision process: individual and group fairness. While individual fairness aims to treat similar individuals similarly, group fairness extends this statement to groups of individuals. The novel loss functions force the score distributions of individuals belonging to different ethnics to be similar and thus, it reduces the performance differences of different ethnics. The experiments were conducted on two publicly available databases, ColorFeret and Labelled Faces in the Wild. These databases contain ethnicity labels of up to 4 classes. The results demonstrated that our proposed approaches are able to maintain a high recognition rate, while significantly reducing the performances differences between the demographic subgroups. While the loss function based on group fairness is able to achieve bias reduction rates of up to 41.22%, introducing individual fairness to the fairness-driven neural network reaches bias reduction rates of up to 52.67%.

4.3.1. Methodology

The main idea of this work is to learn a similarity function that treats individuals of different ethnicities similarly and thus, reduces its verification performance differences between different ethnics. To achieve this goal, we propose two novel loss functions that allow to learn a less biased similarity function. Each of these loss functions incorporates a different fairness criteria to the model. This results in a bias-mitigating neural network that is able to produce fair comparison scores given two biased face embeddings. The fairness definitions that we incorporate into the models are group and individual fairness. While individual fairness aims at treating similar individuals similarly, the definition of group fairness extends this statement to the context of groups of individuals.

In this work, we transfer these fairness definitions into a similarity learning approach that forces the score distributions of different and same ethnicities to be similar. This idea is adapted from Berk et al. [Ber+17] who build these two criteria into linear classification and regression models. Let S be a training set of biased face embeddings and \mathcal{E} is a set

of ethnicities that appear in S . Then, S_g describes a set of samples that belong to the ethnicity $g \in \mathcal{E}$. Since we want to learn a fair similarity function model, the input of the model is given by the absolute difference of the embeddings $x_{i,j} = |e_i - e_j|$ of sample i and j . The $|\cdot|$ operator ensures that the model is invariant to input permutations. For each sample pair $x_{i,j}$, $y \in \{-1, 1\}$ describes its target output relation, where 1 represents a genuine pair and -1 represents an imposter pair.

The fair template comparison model is trained with the following loss

$$\mathcal{L} = (1 - \lambda) H(y) + \lambda f(y, \hat{y}) + \gamma l_2. \quad (4.1)$$

The binary cross-entropy function $H(\hat{y})$ ensures that the model is learning to differentiate between genuine and imposter pairs. Our novel fair penalization term $f(y, \hat{y})$ either incorporates group or individual fairness into the model training. The fairness parameter $\lambda \in [0, 1]$ controls the trade-off between learning the most accurate (and thus, the most biased) comparison score decisions from $H(y)$ and focusing on a fair treatment of the different demographic subgroups. A high fairness parameter λ will force the model to focus more on fulfilling the fairness criteria. The last term of the loss function γl_2 describes a simple l_2 regularization to prevent overfitting.

The novel fair penalization term $f(y, \hat{y})$ either manipulates the model to learn a decision pattern that meets the requirements of group fairness or individual fairness. The penalization term for group fairness is given by

$$f_G = \left(\sum_{i,j \in \mathcal{E}} \frac{1}{|S_i| \cdot |S_j|} \sum_{\substack{(\hat{y}_k, y_k) \in S_i \\ (\hat{y}_l, y_l) \in S_j}} \delta(y_k, y_l) (\hat{y}_k - \hat{y}_l) \right)^2, \quad (4.2)$$

where the delta function

$$\delta(y_k, y_l) = \begin{cases} 1 & \text{if } y_k = y_l \\ 0 & \text{if } y_k \neq y_l \end{cases} \quad (4.3)$$

ensures that genuine and imposter pairs are treated separately. In total, f_G computes the mean (comparison score) prediction differences per genuine and imposter pair and per ethnic pair combination. These values are further considered quadratically to keep convexity. If the model overestimates a pair $x_{i,j}$, it is still able to compensate by underestimating a pair of the same ethnic combination.

The penalization term for individual fairness is given by

$$f_I = \sum_{i,j \in \mathcal{E}} \frac{1}{|S_i| \cdot |S_j|} \sum_{\substack{(\hat{y}_k, y_k) \in S_i \\ (\hat{y}_l, y_l) \in S_j}} \delta(y_k, y_l) (\hat{y}_k - \hat{y}_l)^2. \quad (4.4)$$

As in Equation 4.2, Equation 4.4 computes the mean score differences between the predictions of different groups. However, this time the prediction differences are taken into account quadratically. This strongly forces the score distributions of the same and different ethnicities pairs to be similar without the option to compensate. Therefore, individual fairness forces a bias reduction in a more strictly manner than group fairness. It should be mentioned that this formulation of fairness penalization is not restricted to ethnic bias.

4.3.2. Experimental Setup

Databases

In order to evaluate the ethnic performance differences (ethnic bias) under constrained and less-constrained capturing conditions, we conducted experiments on two publicly available datasets, ColorFeret [Phi+00] and Labeled Faces in the Wild (LFW) [Hua+07]. ColorFeret [Phi+00] consists of 11,283 images of 944 different individuals with different poses under controlled conditions. To ensure a more balanced distribution with a significant number of images per ethnicity, we have aggregated the ethnicities into Black, White, Asian and Other. The LFW dataset [Hua+07] contains more than 13,142 web images of 5,721 individuals. Both datasets include labels about the person’s ethnicity, as shown in Figure 4.1. For the experiments, a subject-exclusive 3-fold cross-validation setting is utilized.

Table 4.1.: A summary of the used datasets, ColorFeret and LFW. The number of images and identities is shown in total and per ethnicity.

Ethnicity	Databases			
	ColorFeret		LFW	
	Images	Identities	Images	Identities
White	7,050	618	10,955	4,650
Black	882	78	762	416
Asian	2,629	225	1,425	655
Other	722	73	-	-
Total	11,283	944	13,142	5,721

Evaluation metrics

In this work, the face verification performance is measured in terms of true acceptance rate (TAR) at a fixed false acceptance rate (FAR). We follow the FAR thresholds as recommended by the best practice guidelines for automated border control of European Border Guard Agency Frontex [Fro17]. The equal error rate (EER) is well known as a single-value indicator of the verification performance and equals the FAR at the threshold where $FAR = 1 - TAR$. To measure the performance differences of different ethnic subgroups, we use the mean absolute deviation (MAD)

$$MAD(TAR) = E [TAR - E [TAR]] \quad (4.5)$$

at a fixed FAR threshold. The MAD computes the mean of the absolute deviation of the TAR of each ethnic subgroup towards the mean TAR $E[TAR]$ over all subgroups. Consequently, it measures the performance differences (in terms of TAR) between the ethnic subgroups and thus, we will refer to this as a measure of bias. A low MAD indicates that all ethnicities have similar performances (low ethnic bias), while higher MAD show strong differences between the verification performances of the ethnic subgroups (high ethnic bias).

Face verification pipeline

In order to verify the identity of an individual, its face template is computed from the given face image and compared against the template of the claimed identity that is stored in the database. This results in a comparison score which is used to verify if the two faces belong to the same identity.

To create the face template from a given face image, first, the image is aligned, scaled, and cropped as described in [KS14]. Second, the preprocessed image is forwarded to a FaceNet [SKP15] model. This model outputs a 128-dimension face embedding representing the identity properties of the face images. In this work, we use a pretrained model¹ that was developed on the MS1M database [Guo+16].

Traditionally, the decision if two face templates come from the same identity or not, is calculated with the cosine similarity of these templates that acts as the comparison score. In our experiments, we refer to this as the "baseline" approach. Our approach offers a fairness-driven neural network approach that replaces the cosine similarity to mitigate bias in the comparison process.

¹<https://github.com/davidsandberg/facenet>

Fairness-driven neural network training

The fairness-driven neural network that we use in this work consists of an 128-dimensional input layer followed by an 256-dimensional and two 512-dimensional intermediate layers. The output of the network is a binary layer with sigmoid activation. All intermediate layers possess a ReLU activation function [NH10] and dropout [Sri+14] (with $p = 0.3$) is applied to these layers. The network was trained with a batchsize of $b = 200$ over 50 epochs using an Adam optimizer [KB14] with a learning rate of 10^{-3} . The distribution of genuine and imposter pairs, as well as the distribution between ethnicities, it kept at the same level per batch to avoid issues that may appear with unbalanced data.

Investigations

To the best of our knowledge, this is the first work that aims at mitigating bias of face recognition at the comparison-level. Consequently, the contribution of this work is not fairly comparable to previous works. In order to still be able to make a fair analysis of the bias-mitigation performance, we compare different variants of our approach against the popular baseline that utilizes the cosine-similarity. The variants of our approach include two different fairness criteria for the loss function (group and individual fairness) as well as different values of the fairness trade-off parameter λ .

The investigations in this work is divided into two steps. First, the intra-ethnic face verification performances are analysed to show that there are significant performance differences for the utilized FaceNet features. Second, the baseline and the proposed approaches are analysed in compared in terms of verification maintainability and the reduction of the ethnic bias.

4.3.3. Results

In order to demonstrate that the verification performances differ between different demographic subgroups, Figure 4.1 shows the verification performances over all individuals (All) as well as the performances on every intra-class ethnicity. On both databases, a large deviation in the demographic-specific performances can be observed. This motivates the need for a bias-mitigating approach in the face recognition pipeline.

Figures 4.2 and 4.3 show ROC-curves and bias reduction plots for the baseline and the variations of our proposed approach on both databases. While the ROC curves (Figure 4.2a, 4.2c, 4.3a, 4.3c) allow a detailed investigation of the verification performance, the MAD plots (Figure 4.2b, 4.2d, 4.3b, 4.3d) allow simultaneous analysis of the bias reduction. The plots for the ColorFeret database (Figure 4.2) show that for fairness parameters

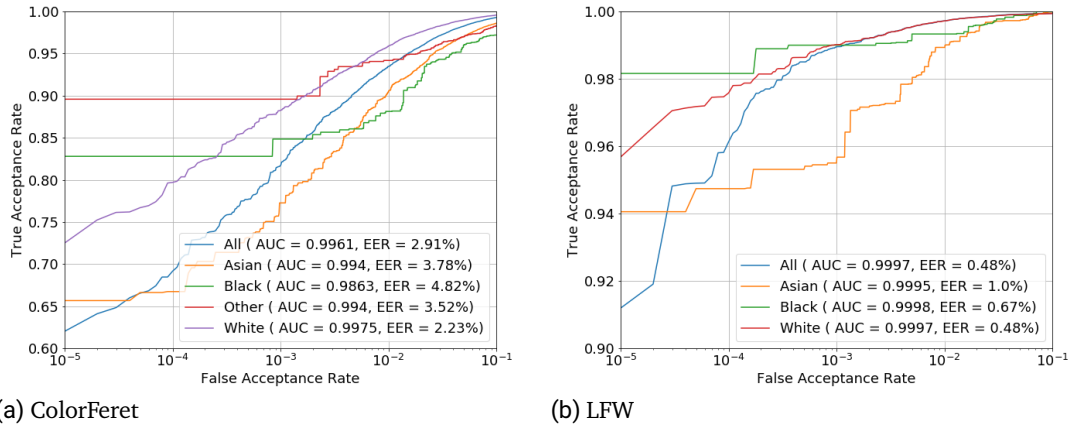


Figure 4.1.: Verification performance (ROC curves) on ColorFeret and Adience: the performance is shown on the whole dataset (all) and for every intra-ethnicity experiment. The recognition performance between different ethnic sub-groups differs significantly.

of $\lambda = \{0.4, 0.5, 0.6\}$, the verification performances of our approach stay close to the verification performance of the biased baseline. This holds for both fairness criteria, group (Figure 4.2a) and individual fairness (Figure 4.2c). On the other hand, in Figures 4.2b and 4.2d, it can be observed that the performances differences (bias) is reduced. The same can be observed for the LFW dataset in Figure 4.3. Here, the results for the fairness parameters of $\lambda = \{0.80, 0.85, 0.90\}$ are shown. Figures 4.3a and 4.3c demonstrate that the verification performance is well preserved for both fairness criteria. In Figures 4.3b and 4.3d, it is noticed that the performance differences between the ethnic subgroups (ethnic bias) are reduced as well.

In order to quantitatively evaluate the bias reduction of our proposed approach, Table 4.2 shows the bias reduction on both databases and fairness criteria for three FAR thresholds. The bias reduction is measured in terms of MAD at these FAR thresholds. The choice of the FAR thresholds follows the best practice guidelines for automated border control of the European Border and Coast Guard Agency Frontex [Fro17]. It can be seen that ethnicity-based performance differences (bias) are significantly reduced in most cases. For all FAR thresholds and on both databases, the loss based on individual fairness shows a higher bias mitigation than the approach based on group fairness. For individual fairness, our proposed solution achieves bias reduction rates of 10.48% to 25.06% on ColorFeret

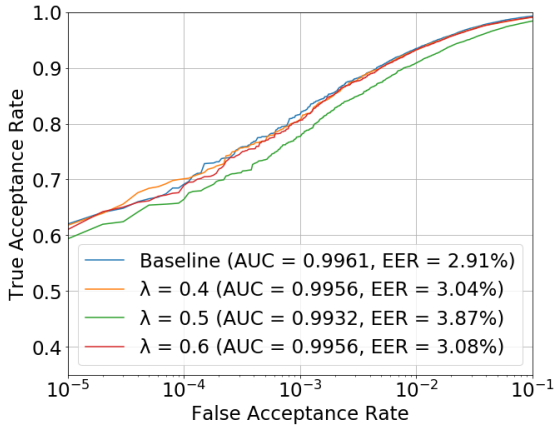
and 15.35% to 52.67% on LFW. This demonstrates that our individual-fairness-based solution is able to significantly mitigate bias, while it preserves very high verification rates as shown before.

Table 4.2.: Bias reduction performances for different FAR on ColorFeret and LFW. The reduction is measured in MAD@FAR. The benefits of our proposed approaches are shown for group fairness and individual fairness. The showed performances reflect fairness parameters of $\lambda = 0.5$ and $\lambda = 0.9$.

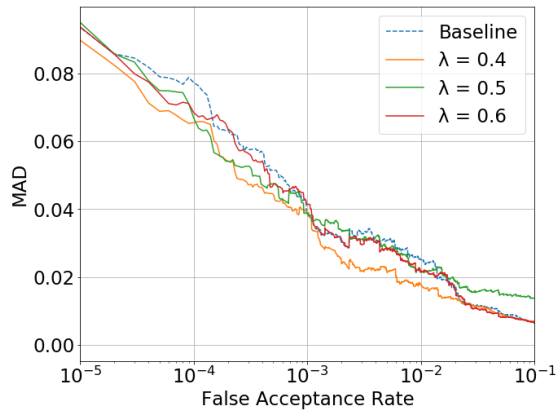
Database	Fairness Criteria	Bias reduction at FAR of		
		10^{-5}	10^{-4}	10^{-3}
ColorFeret	Group	4.06%	15.28%	-1.03%
	Individual	10.48%	27.46%	25.06%
LFW	Group	6.98%	8.20%	41.22%
	Individual	15.35%	16.39%	52.67%

4.3.4. Interim Conclusion

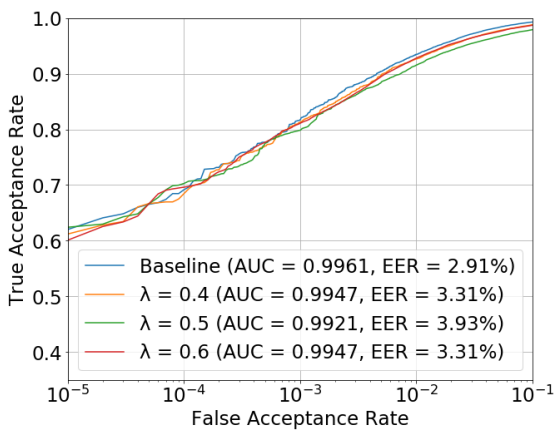
In this section, we successfully introduced fair template comparison for mitigating ethnic-bias in face recognition systems. Previous works proposed solutions that aim at mitigating bias only at the template-level of a biometric system. However, integrating this approach into existing systems requires a complete face image collection of all enrolled identities whose face templates have to be replaced at high computational costs. In this section, we propose a fair face template comparator that was trained with a novel loss function. This loss function is able to incorporate two different fairness criteria in the comparison process: group and individual fairness. The experiments were conducted on two publicly available databases labelled with up to four different ethnicities. For both fairness criteria, the proposed solution is able to significantly reduce the ethnic bias, while maintaining high recognition rates. The results demonstrate that our solution based on individual fairness is able to achieve high bias reduction rates between 15.35% and 52.67%. Unlike previous works that solely aim at learning less biased face representations, our solution is the first work to provide a bias-mitigating solution at the comparison-level of face recognition systems. Moreover, this solution is not restricted to the mitigation of ethnic bias. However, this approach still requires annotated training data to mitigate the bias and affects the overall recognition performance of the system.



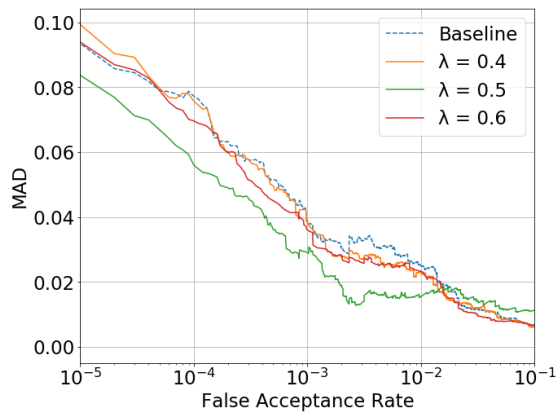
(a) CF - ROC - Group



(b) CF - MAD - Group

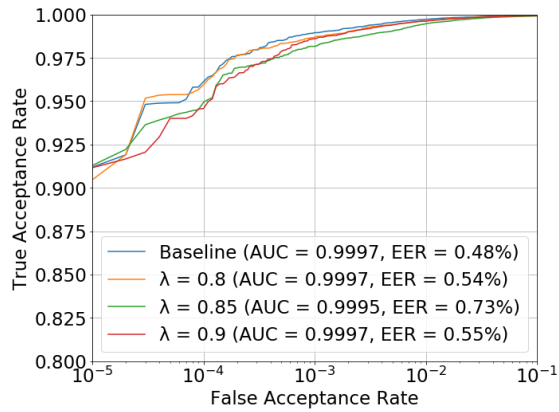


(c) CF - ROC - Individual

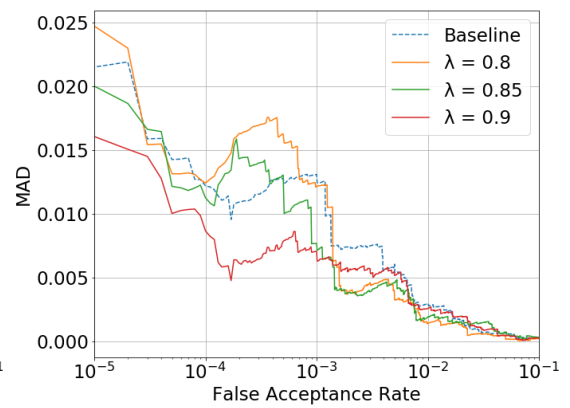


(d) CF - MAD - Individual

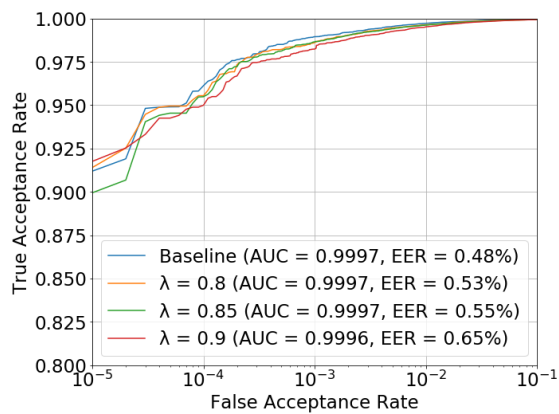
Figure 4.2.: ROC curves (a,c) and bias reduction plots (b,d) on ColorFeret. The performance of the original FaceNet embeddings are shown (Baseline) as well as the proposed comparison approaches based on group fairness (Group) and individual fairness (Individual). The bias reduction plots show the MAD of the performance of different ethnics over different FAR.



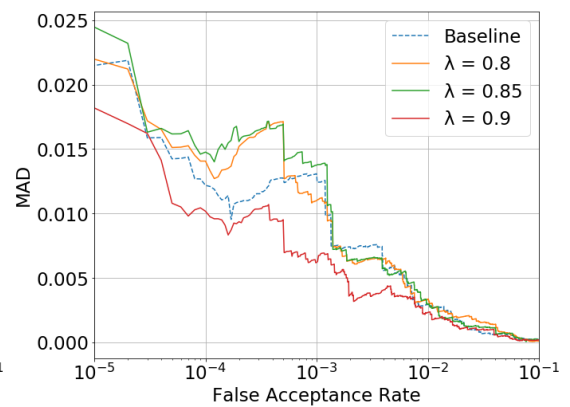
(a) LFW - ROC - Group



(b) LFW - MAD - Group



(c) LFW - ROC - Individual



(d) LFW - MAD - Individual

Figure 4.3.: ROC curves (a,c) and bias reduction plots (b,d) on LFW. The performance of the original FaceNet embeddings are shown (Baseline) as well as the proposed comparison approaches based on group fairness (Group) and individual fairness (Individual). The bias reduction plots show the MAD of the performance of different ethnics over different FAR.

4.4. Mitigating Bias on Score-Level

Previous works on bias-mitigating face recognition proposed solutions to produce less-biased templates. To avoid the need for storing additional face images for each enrolled individual and the high computational workload when integrating one of these approaches into existing systems, we proposed an easily-integrable solution in Section 4.3. However, also this solution (a) requires annotated training data, (b) mitigates only the bias of the annotated attributes, and (c) negatively affects the overall face recognition performance.

In this section, we propose a novel and unsupervised fair score normalization [Ter+20f] approach that mitigates bias in face recognition systems. Unlike previous works, our solution jointly

- a) works with unlabelled training data,
- b) effectively mitigates bias of unknown origins,
- c) and strongly enhances the overall recognition performance.

Its theoretical motivation is based on the notation of individual fairness [Dwo+12], resulting in a solution that treats similar individuals similarly and thus, more fairly. The proposed approach clusters samples in the embedding space such that similar identities are categorized without the need for pre-defined demographic classes. For each cluster, an optimal local threshold is computed and used to develop a score normalization approach that ensures a more individual, unbiased, and fair treatment. The experiments are conducted on three publicly available datasets captured under controlled and in-the-wild conditions and on two face embeddings. To justify the concept of our fair normalization approach, we provide a visual illustration that demonstrates (a) the suitability of the notation of individual fairness for face recognition and (b) the need for more individualized treatment of face recognition systems. Experiments were conducted on three publicly available datasets captured under controlled and in-the-wild circumstances. The results demonstrate that our solution reduces demographic biases, e.g. by up to 82.7% in the case when gender is considered. Moreover, it mitigates the bias more consistently than existing works. In contrast to previous works, our fair normalization approach enhances the overall performance by up to 53.2% at a false match rate of 10^{-3} and up to 82.9% at a false match rate of 10^{-5} . Additionally, it is easily integrable into existing recognition systems and not limited to face biometrics.

4.4.1. Methodology

The goal of this work is to enhance the fairness of existing face recognition systems in an easily-integrable manner. In this work, we follow the notation of individual fairness [Dwo+12]. This notation emphasizes that similar individuals should be treated similarly. We transfer this idea to the embedding and score level to propose a novel fair group-based score normalization method, without the need for pre-defined demographic groups. The proposed approach is able to treat all identities more individually and therefore, increase the group-related, as well as the total, recognition performance.

Fair group score normalization

Our proposed solution is presented assuming a set of face embeddings $X = (X_{train} \cup X_{test})$ with the corresponding identity information $y = (y_{train} \cup y_{test})$, both partitioned into test and training set.

Training phase During training phase, a k-means cluster algorithm [HW79] is trained on X_{train} to split the embedding space into k clusters ($k = 100$ in our experiment). For each cluster $c \in \{1, \dots, k\}$, an optimal threshold for a false match rate of 10^{-3} is computed using the genuine and imposter scores of cluster c

$$gen_c = \{s_{ij} \mid ID(i) = ID(j), i \neq j, \forall i \in C_c\} \quad (4.6)$$

$$imp_c = \{s_{ij} \mid ID(i) \neq ID(j), \forall i \in C_c\}. \quad (4.7)$$

The genuine score set gen_c of cluster c includes the all comparison scores of samples i and j that come from the same identity ($ID(i) = ID(j)$), where at least one sample lies within cluster c ($i \in C_c$). Conversely, the imposter score set imp_c of cluster c is defined as all comparison scores s_{ij} from different identity pairs ($ID(i) \neq ID(j)$) where at least one sample lies within cluster c ($i \in C_c$). The local threshold for each cluster c is denoted as $thr(c)$. Furthermore, the threshold for the whole training set X_{train} is calculated and denoted as the global threshold thr_G .

Operation phase During the operation phase, the normalized comparison score \hat{s}_{ij} should be computed to determine if sample i and j belong to the same identity. Firstly, the corresponding clusters for both samples are computed. The cluster thresholds for sample i and j are denoted as thr_i and thr_j . Secondly, these cluster thresholds, as well as the global threshold thr_G , are used to calculate the normalized score

$$\hat{s}_{ij} = s_{ij} - \frac{1}{2} (\Delta thr_i + \Delta thr_j), \quad (4.8)$$

where

$$\Delta thr_i = thr_i - thr_G, \quad (4.9)$$

describes the local-global threshold difference for sample i .

Discussion

The goal of this score normalization approach is to introduce individual fairness in a biometric system and thus, reduce the discriminatory behavior of face recognition systems. The notation of individual fairness emphasizes that similar individuals should be treated similarly. We incorporate this statement in our normalization method using clustering and local thresholds. Clustering in the embedding space identifies similar individuals and local cluster thresholds enable approximately individual treatment.

The choice of the individuality parameter k defines the number of clusters for our fair score normalization and is crucial for the recognition performance. A small k (e.g. $k = 2$) refers to a less individual normalization of the score, while a very large k reduces the number of samples per clusters and thus, the quality of the local thresholds.

How does fair normalization affect different sample pairs?

In the following, we discuss how the proposed fair normalization approach affects biased and unbiased genuine and imposter pair comparisons.

Biased genuine pair - Assuming that an identity \mathcal{I} , with samples i and j , belongs to a *biased* group, their comparison score $s_{ij} = 0.4$ will be low. Since this is lower than the global threshold $thr_G = 0.6$, the decision for this genuine pair will be falsely made towards imposter. Since these samples belong to a biased cluster, the recognition performance within is low and so are the local thresholds $thr_i = thr_j = 0.3$. The low local thresholds lead to a negative local-global threshold difference $\Delta thr_i = \Delta thr_j = 0.3 - 0.6 = -0.3$ and thus, the normalized comparison score $\hat{s}_{ij} = 0.4 - \frac{1}{2}(-0.3 - 0.3) = 0.7$ increases. Since $\hat{s} = 0.7 > thr_G = 0.6$, the system now comes to the correct genuine decision with the proposed normalization method.

Unbiased genuine pair - Assuming that an identity \mathcal{I} , with samples i and j , belongs to an *unbiased* group, their comparison score $s_{ij} = 0.9$ will be high. Since these samples belong to an unbiased cluster, the performance within is high and so are the local thresholds $thr_i = thr_j = 0.7$. The low local thresholds leads to a positive $\Delta thr_i = \Delta thr_j = 0.9 - 0.6 = +0.2$ and thus, the normalized comparison score $\hat{s}_{ij} = 0.9 - \frac{1}{2}(0.2 + 0.2) = 0.7$ increases. Since $\hat{s} = 0.7 > thr_G = 0.6$, the system still come to the correct genuine decision.

Imposter pair - For imposter pairs (i, j) , three situations have to be considered depending on cluster-correspondence of the two samples. The first one refers to the case in which one of the two samples belongs to a cluster with a low local threshold, while the other belongs to one with a large local threshold. In this case, our normalization approach is only marginally changing the comparison score. Therefore, the verification decision is unchanged.

In the second case, both samples belong to clusters with high local thresholds. Consequently, the score is highly reduced and thus, the probability for a false match decreases.

The third case is the most critical, where both samples belong to clusters with low local thresholds. If both samples belong to different clusters, then their embeddings are dissimilar and will result in a low comparison score. Consequently, the risk of a false match is low. If both samples belong to the same cluster, their embeddings are similar and thus, there is a high risk of a false match. However, our method is especially optimized for exactly this (critical) case, since the local thresholds are computed based on intra-cluster performance. Consequently, the false acceptance rate with our normalization is lower or equal than the unnormalized case.

Main error - The main error that can appear with the normalization approach happens at the border of two adjacent clusters with high differences in the local thresholds. Comparing similar samples at the border of these clusters might lead to overcorrections of the scores. However, Figure 4.4 showed that this is rarely the case. Moreover, this can be prevented by a sufficient choice of k , since k determines the number of clusters and a larger number of clusters lead to more fine-grained local thresholds of adjacent clusters.

4.4.2. Experimental Setup

Databases In order to evaluate the face recognition performance of our approach under controlled and unconstrained conditions, we conducted experiments on the public available Adience [EEH14], ColorFeret [Phi+00], and Morph [RT06a] datasets. ColorFeret [Phi+00] consists of 14,126 images from 1,199 different individuals with different poses under controlled conditions. Furthermore, a variety of face poses, facial expressions, and lighting conditions are included in the dataset. The Adience dataset [EEH14] consists of over 26.5k images from over 2.2k different subjects under unconstrained imaging conditions. Morph [RT06a] contains 55,134 images from 13,618 subjects. The ages range from 16 to 77 with a median of 33 years. While Adience contains additional information about gender and age, ColorFeret and Morph also provide labels regarding the subject's ethnicities. The distribution of these attributes in the databases is shown in Table 4.3. In the experiments, this information is used to investigate the face recognition performance for several demographic groups.

Table 4.3.: Attribute distribution of the images in the used databases.

Attribute	Class	Database		
		Adience	ColorFeret	Morph
Gender	Male	51.6%	64.6%	84.6%
	Female	48.4%	35.4%	15.4%
Age	<20	40.0%	1.4%	17.2%
	20-30	33.2%	34.9%	28.1%
	30-40	15.9%	27.9%	28.3%
	40+	10.9%	35.8%	26.4%
Ethnicity	Asian	-	23.3%	0.4%
	Black	-	7.6%	77.2%
	White	-	62.6%	19.2%
	Other	-	6.5%	3.2%

Evaluation metrics In this work, we will report the recognition performances in terms of false non-match rate (FNMR) at fixed false match rates (FMR). As recommended by the European Border Guard Agency Frontex [Fro17], we will use FMR thresholds of 10^{-3} and smaller. To evaluate the amount of demographic bias in the recognition performance, the recognition performance is evaluated within all subgroups and the standard deviation (STD) of these group-specific performances is reported. A low STD refers to a more unbiased attribute performance since the performances of the different attribute classes are similar. In contrast, a high STD refers to a biased attribute performance with strong performance differences between the attribute classes.

Workflow details For the comparison of two samples, both face images get aligned, scaled, and cropped. Then, the preprocessed images are passed into a face recognition model resulting in a face template for each image. The comparison of two embeddings is done using cosine-similarity. In this work, we use FaceNet² [SKP15] and VGGFace³ [PVZ15]. Both models were trained on MS-Celeb-1M [Guo+16]. Moreover, the preprocessing for FaceNet was done based on [Kin09] and for VGGFace, the preprocessing follows the methodology described in [Zha+16]. For all experiment scenarios, subject-disjoint 5-fold cross-validation is utilized and in each iteration, all possible positive and negative

²<https://github.com/davidsandberg/facenet>

³https://github.com/ox-vgg/vgg_face2

face combinations pairs are evaluated.

Baseline approaches We evaluate our fair score normalization approach in comparison with two bias-mitigating works [Ter+20i; Sri+19b] that, just as our solution, act beyond template-generation and thus, are easily-integrable as well. In [Ter+20i], a fair template comparison (FTC) approach is proposed aiming at mitigating ethnic-bias. For our experiments, we trained the model with $\lambda = 0.5$. This choice is based on the recommendation of [Ter+20i]. In [Sri+19b], base normalization and score-level fusion (SLF) strategies are investigated for mitigating bias in face recognition systems. We use their best working approach, namely min-max normalization with a simple sum-fusion rule, as an additional baseline in our experiments combining both utilized face embeddings.

Investigations The investigations of this work are divided into four parts:

1. We first visually demonstrate the need for more individual treatment in face recognition systems. Moreover, we show that our approach is able to treat similar individuals more similarly.
2. We investigate the effect of the individuality parameter k over a wide parameter range since this critically affects the effectiveness of the proposed approach.
3. We investigate the bias-mitigation performance of the unmodified baseline, our normalization approach, and state-of-the-art approaches. Therefore, the intra-class verification performance is investigated for different demographic attributes and the attribute-bias is measured and compared.
4. We finally investigate the overall verification performance to prove that, unlike previous works, our approach enhances the overall performance while mitigating demographic-bias.

4.4.3. Results

Visual demonstration of the need for individuality

Since our approach is based on the idea of individual fairness, we first want to visually demonstrate why this notation is suitable for face recognition. Figure 4.4 shows an t-SNE visualization of the embedding space for the dataset Adience. The t-SNE algorithm maps the high-dimensional embedding space into a two-dimensional space such that similar samples in the high-dimension space lie closely together in two dimensions. Furthermore, each sample is colored based on the local thresholds computed by the proposed approach. Two observations can be made from this figure:

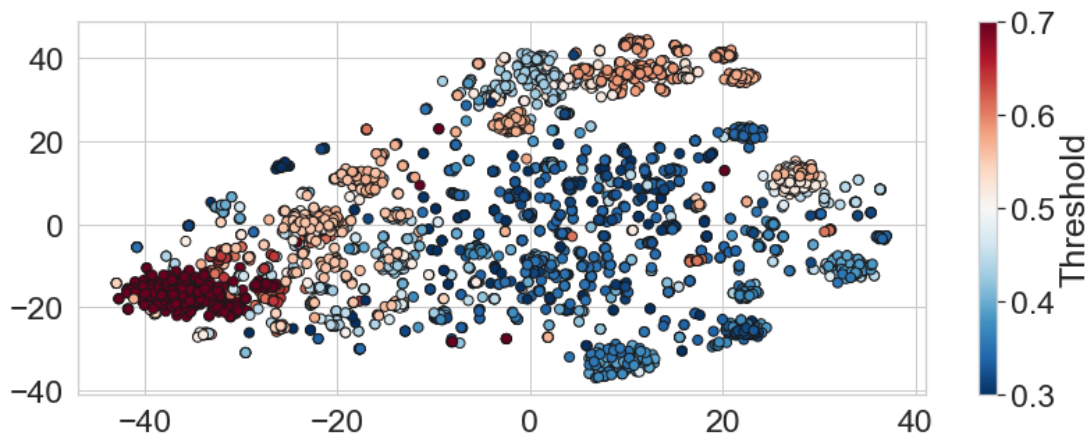


Figure 4.4.: Visualizations of the Adience FaceNet embeddings using t-SNE [MH08]. Each individual is represented as a point and each point is colored based on its optimal local threshold thr_k . The formation of several clusters with similar local thresholds shows that our approach is able to identify similar individuals and to treat them similarly. The large variation of optimal local thresholds (0.3 - 0.7) demonstrates the need for this more individual, and thus fair, treatment.

1. There are several clusters with similar local thresholds in the embedding space. Consequently, our proposed approach is able to identify similar identities and to treat them similarly (through similar local thresholds).
2. The optimal thresholds for each cluster vary significantly from 0.3 to 0.7. This widespread of optimal local thresholds demonstrates the need for a more individual, and thus fair, treatment.

The choice of the individuality parameter k

In this section, we analyse the sensitivity of the individuality parameter k and justify our choice for $k = 100$. Figure 4.5 shows verification performances of the proposed fair normalization over different individuality parameters k on all datasets and both face embeddings. Moreover, the unnormalized baseline is shown. For $k = 1$, the normalization does not change the scores and thus, the same performance is observed with and without the normalization. For $k \geq 1$, the verification performance increases, since our fair score normalization approach leads to more individual treatment of each sample. This can

be observed in all cases (a,b,c,e,f), except for Figure 4.5d. In this case, the clustering algorithm produces clusters of unequal sizes leading to performance degradation. However, this still lies within the standard deviation of the unnormalized case. Moreover, this still leads to a strong bias-mitigation, as we will see in Section 4.4.3. If k is large, the number of samples per cluster decreases. Since these are necessary to determine the local thresholds, the quality of these decreases. This leads to unreliable thresholds and thus, inaccurate recognition performances. For all datasets and both embeddings, this performance drop can be observed for large k . However, individuality parameters around $k \approx 100$ show a generally stable performance. Therefore, we choose $k = 100$ for our experiments.

Analysis of the demographic-bias

Our fair normalization approach aims at mitigating biased recognition decisions of unknown origins. This section analyses this aspect. In Table 4.4, the intra-class recognition performance (in terms of $\text{FNRM}@10^{-3}\text{FMR}$) is shown for several demographic classes with and without our normalization approach. Tables 4.5 and 4.6 use this information to measure the attribute-specific bias in the recognition performances and compares it with previous works. For most attribute classes, the intra-class recognition performance with our fair normalization approach leads to strong enhancements of up to 58%. However, for some classes the recognition performance decreases. This happens when an intra-class recognition performance is much stronger for one class compared to the other classes for this attribute. Since our fair normalization approach aims at enhancing fairness, and thus reduces the performance differences between the different attribute classes, (a) weak classes have to be improved or (b) strong classes have to be adjusted. For instance, the second case happens in ColorFeret for age and ethnicity. The age classes [31-40] and [40+] and white ethnicities perform outstanding well without our normalization and they get adjusted to more closely match the performance of the other attribute classes.

The effectiveness of the proposed normalization approach is shown in Table 4.5 on FaceNet features and in Table 4.6 on VGGFace features. Additionally, a comparison with related works is presented. Here, the bias of an attribute is determined by its standard deviation of the attribute performances. Moreover, the bias reduction rates are shown. Positive values indicate a strong bias-mitigation and vice versa. Please note that the gender-bias on Adience using VGGFace features is already very low and consequently leads to an increase of gender-specific bias on all investigated approaches. SLF [Sri+19b] achieves high bias reduction rates in some cases. However, in 7 out of the 16 cases, it even increases the class-biases. FTC [Ter+20i] also increases the class-biases in many cases. Just the ethnic-bias is consistently reduced. This might relate to the choice of the fairness parameter $\lambda = 0.5$ which is recommended in [Ter+20i] and optimized to the mitigation


of ethnic-bias. For our approach, the biases from various origins are consistently mitigated and bias reduction rates of up to 82.7% are achieved.

The global face recognition performance

This section investigates the overall face recognition performance of our bias-mitigation approach and previous works. Tables 4.7 and 4.8 show the verification performance of FaceNet and VGGFace features on three databases at three decision thresholds. The performance is reported for the unmodified baseline (Base), for our fair normalisation approach (Ours) and previous works (SLF [Sri+19b] and FTC [Ter+20i]). Bias-mitigation often comes at the cost of a decreasing recognition performance. This can be seen for SLF and FTC. For example, the overall recognition performance of SLF on FaceNet features decreases in every case on the Morph dataset. For FTC, the performance decreases in most cases as well. In contrast, our proposed approach significantly enhances the global recognition performance by up to 82.9%, while effectively mitigating bias. Just in one out of 17 cases, the performance slightly decreases due to the failed clustering as discussed in Section 4.4.3.

4.4.4. Interim Conclusion

Despite the progress achieved by current face recognition systems, recent works showed that biometric systems impose a strong bias against subgroups of the population. Consequently, there is an increased need for solutions that increase the fairness of such systems. Previous works focused on learning bias-mitigated face representations. However, these solutions are often hardly-integrable and degrade the overall recognition performance. In this work, we propose a novel fair score normalization approach to mitigate bias from recognition systems. Our unsupervised score normalization approach is easily-integrable into existing systems and significantly enhances the system’s overall recognition performance. Integrating the idea of individual fairness, our solution aims at treating similar individuals similarly. The experiments were conducted on three publicly available datasets captured under various conditions and on two kinds of face embeddings. The results show that the proposed approach significantly reduces demographic-bias, e.g. it mitigates ethnic-bias by 17.4-32.8%. Additionally, it mitigates bias more consistently over demographic domains than related works and strongly enhances the overall recognition performance, e.g. by 16.4-82.90% on the Morph benchmark. In contrast to related works, our method jointly achieves the following points: it (a) does not need additional soft-biometric labels during training or inference time, (b) can be easily integrated into existing face recognition



systems, (c) enhances the total face recognition performance, and (d) leads to a consistent bias-mitigation. Moreover, it is, by design, not limited to face biometrics.

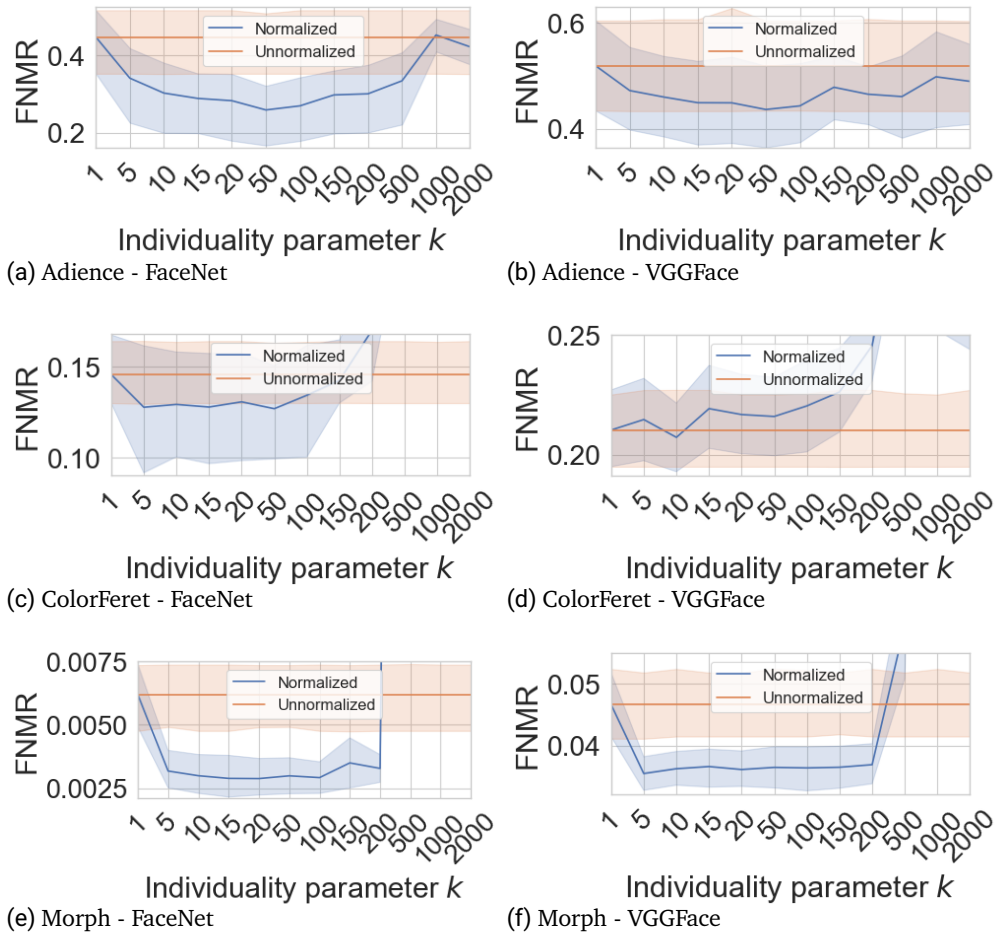


Figure 4.5.: Analysis of the verification performance at an FMR of 10^{-3} based on the individuality parameters k . The proposed normalization approach (blue) is compared to the unnormalized baseline (orange). The analysis includes three datasets and two face embeddings. The shaded areas represent the standard deviation over the 5 cross-validation folds. Individuality parameters around $k \approx 100$ show a generally a stable performance improvement.

Table 4.4.: Intra-class recognition performance of our approach: the performance is shown in terms of FNMR@ 10^{-3} FMR for FaceNet and VGGFace embeddings. The unnormalised (unnorm.) and normalized (norm.) performance within each attribute class is reported with the corresponding performance change (PC). In many cases, the proposed normalization approach enhances fairness by strongly improving the performance of under-performing classes. In other cases, the approach leads to a performance adaptation to minimize the performance differences between the groups, leading to more fair recognition decisions as shown in Tables 4.5 and 4.6.

DB	Attribute	Class	FaceNet			VGGFace		
			Unnorm.	Norm.	PC	Unnorm.	Norm.	PC
Adience	Gender	Male	0.5129	0.2600	49.3%	0.4636	0.4462	3.8%
		Female	0.3837	0.2823	26.4%	0.4703	0.4985	-6.0%
	Age	0-2	0.7764	0.7641	1.6%	0.7861	0.7753	1.4%
		4-6	0.6069	0.5838	3.8%	0.7327	0.7417	-1.2%
		8-12	0.4327	0.3804	12.1%	0.4527	0.4769	-5.3%
		15-20	0.7677	0.4890	36.3%	0.4358	0.4242	2.7%
		25-32	0.2264	0.1540	32.0%	0.3174	0.3049	3.9%
		38-43	0.1766	0.1631	7.6%	0.2670	0.3002	-12.4%
		48-53	0.2253	0.1398	37.9%	0.2859	0.3018	-5.6%
60-100	0.1224	0.1140	6.9%	0.2468	0.2451	0.7%		
ColorFeret	Gender	Male	0.1635	0.1424	12.9%	0.2252	0.2421	-7.5%
		Female	0.2167	0.1891	12.7%	0.2732	0.2704	1.0%
	Age	10-20	0.2118	0.1818	14.2%	0.2912	0.2873	1.3%
		21-30	0.1506	0.1071	28.9%	0.2059	0.2070	-0.5%
		31-40	0.1452	0.1459	-0.5%	0.1842	0.2208	-19.9%
		40+	0.0933	0.1212	-29.9%	0.1701	0.2034	-19.6%
	Ethnicity	Asian	0.3177	0.2553	19.6%	0.3099	0.3170	-2.3%
		Black	0.2489	0.2361	5.1%	0.4120	0.3736	9.3%
		White	0.1089	0.1282	-17.7%	0.2085	0.2228	-6.9%
Other		0.1424	0.1417	0.5%	0.2217	0.2112	4.7%	
Morph	Gender	Male	0.0059	0.0031	47.5%	0.0463	0.0362	21.8%
		Female	0.0364	0.0153	58.0%	0.1220	0.1062	13.0%
	Age	<20	0.0056	0.0034	39.3%	0.0648	0.0585	9.7%
		20-29	0.0039	0.0019	51.3%	0.0461	0.0398	13.7%
		30-39	0.0081	0.0041	49.4%	0.0495	0.0404	18.4%
		40+	0.0137	0.0064	53.3%	0.0586	0.0472	19.5%
	Ethnicity	African	0.0037	0.0036	2.7%	0.0431	0.0389	9.7%
		Asian	1.0000	0.8036	19.6%	1.0000	1.0000	0.0%
		European	0.0069	0.0077	-11.6%	0.0888	0.086	3.2%
Hispanic		0.0062	0.0057	8.1%	0.0396	0.0431	-8.8%	

Table 4.5.: Analysis of the bias reduction of the proposed approach (Ours) in comparison with two previous works (SLF [Sri+19b] and FTC [Ter+20i]) on FaceNet features. The bias is measured in terms of STD of the class-wise FNMRs at an FMR of 10^{-3} . Unlike both previous works, our proposed approach mitigates bias effectively and consistently.

Database	Attribute	Bias (STD)		Bias reduction		
		Baseline	SLF [Sri+19b]	FTC [Ter+20i]	Ours	
Adience	Gender	0.0646	68.5%	-44.9%	82.7%	
	Age	0.2515	11.9%	45.9%	8.9%	
ColorFeret	Gender	0.0266	-8.4%	-85.7%	12.2%	
	Age	0.0420	12.8%	-56.6%	32.5%	
	Ethnicity	0.0833	34.9%	5.9%	32.8%	
Morph	Gender	0.0216	-25.9%	-18.0%	60.0%	
	Age	0.0043	4.3%	-28.4%	56.2%	
	Ethnicity	0.4972	0.4%	24.5%	19.8%	

Table 4.6.: Analysis of the bias reduction of the proposed approach (Ours) in comparison with two previous works (SLF [Sri+19b] and FTC [Ter+20i]) on VGGFace features. The bias is measured in terms of STD of the class-wise FNMRs at an FMR of 10^{-3} . Unlike both previous works, our proposed approach mitigates bias effectively and consistently.

Database	Attribute	Bias (STD)		Bias reduction		
		Baseline	SLF [Sri+19b]	FTC [Ter+20i]	Ours	
Adience	Gender	0.0262	-112.7%	-135.1%	-79.2%	
	Age	0.1935	-0.9%	100.0%	2.0%	
ColorFeret	Gender	0.0142	-21.0%	-81.3%	41.0%	
	Age	0.0339	-47.0%	-237.1%	27.8%	
	Ethnicity	0.0673	25.0%	39.3%	17.4%	
Morph	Gender	0.0503	49.2%	-25.0%	5.8%	
	Age	0.0084	20.4%	-108.7%	1.6%	
	Ethnicity	0.3756	-31.9%	3.8%	20.4%	

Table 4.7.: Investigation of the overall recognition performance of the proposed approach (Ours) in comparison with two previous works (SLF [Sri+19b] and FTC [Ter+20i]) on FaceNet. The FNMR is shown at different FMR thresholds. Base refers to the unmodified FaceNet and VGGFace performance. Even while making the recognition process fairer, in contrast to previous work, our approach consistently improves the global recognition performance.

		Adience		ColorFeret		Morph	
10^{-3} FMR	Base	0.4481		0.1460		0.0062	
	SLF [Sri+19b]	0.4438	1.0%	0.1229	15.8%	0.0095	-53.8%
	FTC [Ter+20i]	0.7109	-58.6%	0.1406	3.7%	0.0081	-30.3%
	Ours	0.2694	39.9%	0.1343	8.0%	0.0029	53.2%
10^{-4} FMR	Base	0.7651		0.3299		0.0219	
	SLF [Sri+19b]	0.6840	10.6%	0.2381	27.8%	0.0318	-45.4%
	FTC [Ter+20i]	0.9160	-19.7%	0.3406	-3.2%	0.0285	-30.1%
	Ours	0.4800	37.3%	0.2517	23.7%	0.0121	44.7%
10^{-5} FMR	Base	0.9324		0.5403		0.0576	
	SLF [Sri+19b]	0.8074	13.4%	0.3658	32.3%	0.0768	-33.3%
	FTC [Ter+20i]	0.9791	-5.0%	0.6009	-11.2%	0.0743	-28.9%
	Ours	0.6813	26.9%	0.3979	26.4%	0.0371	35.6%

Table 4.8.: Investigation of the overall recognition performance of the proposed approach (Ours) in comparison with two previous works (SLF [Sri+19b] and FTC [Ter+20i]) on VGGFace. The FNMR is shown at different FMR thresholds. Base refers to the unmodified FaceNet and VGGFace performance. Even while making the recognition process fairer, in contrast to previous work, our approach consistently improves the global recognition performance.

		Adience		ColorFeret		Morph	
10^{-3} FMR	Base	0.5201		0.2107		0.0465	
	SLF [Sri+19b]	0.4438	14.7%	0.1229	41.7%	0.0095	79.5%
	FTC [Ter+20i]	0.7579	-45.7%	0.4941	-134.5%	0.0681	-46.6%
	Ours	0.4430	14.8%	0.2203	-4.6%	0.0363	21.9%
10^{-4} FMR	Base	0.7404		0.3635		0.1180	
	SLF [Sri+19b]	0.6840	7.6%	0.2381	34.5%	0.0318	73.0%
	FTC [Ter+20i]	0.9780	-32.1%	0.8225	-126.3%	0.1809	-53.3%
	Ours	0.6281	15.2%	0.3474	4.4%	0.0987	16.4%
10^{-5} FMR	Base	0.8782		0.5804		0.2171	
	SLF [Sri+19b]	0.8074	8.1%	0.3658	37.0%	0.0768	64.6%
	FTC [Ter+20i]	0.9976	-13.6%	0.9765	-68.3%	0.3463	-59.5%
	Ours	0.7685	12.5%	0.4778	17.7%	0.0371	82.9%

4.5. Summary

Large-scale face recognition systems are spreading worldwide and are increasingly involved in critical decision-making processes, such as in forensics and law enforcement. Consequently, these systems also have a growing effect on everybody's daily life. However, many current biometric solutions are mainly optimized for maximum recognition accuracy [JNR16] and perform significantly different depend on an individual's demographics [Orc16; AZN18; FPO02; Phi+11; BG18; Gar+16]. This means that, for example, specific demographic groups can be falsely identified as black-listed individuals more frequently than other groups. Consequently, there is an increased need that guarantees fairness for biometric solutions [BG18; GF16; Zem+13] to prevent discriminatory decisions.

From a political perspective, there are several regulations to ensure people the right to non-discrimination, such as Article 7 of the Universal Declaration on Human Rights, Article 14 of the European Convention of Human Rights, and the General Data Protection Regulation (GDPR) [VB17]. Previous works on bias-mitigating face recognition focused on template-level solutions. Due to the difficulty of integrating these approaches into existing face recognition systems, more easily-integrable solutions are needed to ensure non-discrimination in biometric systems. Consequently, in this Chapter, we proposed two bias-mitigating face recognition solutions that operate beyond template-level and thus, are easily-integrable.

In Section 4.3, we proposed Fair Template Comparison (FTC) [Ter+20i], the first bias-mitigating face recognition solution that works on the comparison-level of a biometric system. It replaces the system's similarity function by a fairness-driven model that is trained with a novel penalization term. The proposed penalization term allows to include the notation of individual and group fairness during training that forces the score distributions of different ethnicities to be more similar. The results demonstrate that our FTC approach based on individual fairness is able to effectively reducing ethnic-bias while maintaining a large fraction of the recognition performance.

In Section 4.4, we proposed the main contribution of this chapter, Fair Score Normalization (FSN) [Ter+20f]. FSN is the first bias-mitigating face recognition approach that operates on the score-level of a biometric system. It builds on the notation of individual fairness and thus, aims at treating similar individuals similarly. This is achieved by clustering training samples in the embedding space and computing optimal local thresholds for each cluster. If the comparison score of two samples is calculated, it normalizes this score based on the optimal local thresholds of the clusters that are associated with the samples. This ensures a more individual, unbiased, and fair treatment. The results on three publicly available databases demonstrate that our solution mitigates bias by up to 82.7%. Moreover, it reduces the bias more consistently than existing works and additionally enhances the

overall recognition performance by 53.2% at an FMR of 10^{-3} and by 82.9% at an FMR of 10^{-5} . In contrast to previous works, our proposed FSN solution jointly (a) operates on unlabelled training data, (b) effectively mitigates bias of unknown origins, and (c) strongly improves the overall recognition performance of the system.

5. Enhancing Soft-Biometric Privacy

5.1. Introduction

The face is one of the most used biometric modalities [Dam+18d; Wan+18a]. A typical face recognition system contains feature representations (templates) for each enrolled individual. To verify a subject's identity, a template of this subject's probe is computed and compared against the template of the claimed identity [PKB16]. However, as we demonstrated in Section 3.5, more information than just the person's identity can be deduced from these templates. This includes information about an individual's gender, age, ethnicity, hair style, accessories, sexual orientation and health status [DER16]. Many applications are not permitted by the users to have access to this information. Thus, the stored data should be exclusively used for recognition purposes [MR17; Erk+09]. Consequently, extracting such information without a person's consent is considered a violation of their privacy [Kin13].

In order to prevent this kind of function creep, *soft-biometric privacy* aims at suppressing or hiding privacy-risk information in face biometrics. This is further challenged by simultaneously maintaining a high recognition performance. Previous works proposed privacy-enhancing solutions based on supervised [Mir+18; MR17; OR14] and unsupervised approaches [Ter+19b; Ter+20c]. While unsupervised approaches show a more comprehensive but weaker privacy-enhancement, supervised approaches are limited to the suppression of pre-defined attributes and thus, are vulnerable to unconsidered function creep attacks.

In this chapter, we propose four approaches to enhance soft-biometric privacy in face templates. Section 5.2 provides a summary of related works on soft-biometric privacy including the proposed solutions. Next, these proposed solutions are presented chronologically in the following sections.

- Section 5.3: **Incremental Variable Elimination (IVE)** [Ter+19a] is a supervised approach that incrementally determines and eliminates the highest privacy-risk variables in face templates based on decision-tree ensembles. Contrary to previous works, IVE is able, by design, to suppress binary, categorical, and continuous

attributes.

- Section 5.4: **Similarity-Sensitive Noise Transformations (SSNT)** [Ter+19b] are an unsupervised privacy-enhancing face recognition approaches that inject geometric-inspired noise to templates. This aims at achieving a more comprehensive soft-biometric privacy-enhancement not limited to pre-considered attributes. Moreover, the reduction of the recognition performance is directly controlled by SSNT.
- Section 5.5: **Negative Face Recognition (NFR)** [Ter+20c] is an unsupervised approach to improve soft-biometric privacy for face recognition. While ordinary (positive) face templates contain information of the person’s identity, negative face templates provide random complementary information about this individual. Storing only negative templates in the database, prevents function creep attackers from successfully predicting privacy-sensitive attributes. For verification, the positive template of an individual is compared with the stored negative template of the claimed identity by measuring the dissimilarity.
- Section 5.6: **Privacy-Enhancing Minimum Information Units (PE-MIU)** [Ter+20h] is the main contribution of this chapter. It is a training-free approach to prevent function creep attackers from successfully predicting privacy-sensitive information from face templates. PE-MIU exploits the structural differences between face recognition (use-case) and facial attribute estimation (attack scenario) by creating templates in a mixed representation of minimal information units (MIU). These representations contain patterns of privacy-sensitive attributes in a highly randomized form. Therefore, the estimation of these attributes becomes hard for function creep attacks. During verification, the units of a probe template are assigned to the units of a reference template by solving an optimal best-matching problem. The results demonstrate that on both, maintaining recognition rates and suppressing attribute information, the proposed MIU-based approach consistently outperforms state-of-the-art approaches.

The source-code for each contribution is available under the following link¹.

5.2. Related Work

In the context of face biometrics, privacy has been studied from two perspectives. The first kind focuses on preserving facial characteristics such as gender, age, and expression while

¹<https://github.com/pterhoer/PrivacyPreservingFaceRecognition>

de-identifying face images [Gro+06; JYL15; MPS18; NSM05]. The second kind aims at preventing the estimation of these facial attributes while maintaining its recognition ability [MR17]. In this chapter, we will focus on the latter case. This is known as soft-biometric privacy [Ter+19a; Bor+20]. Solutions on soft-biometric privacy either operate on image- or template-level.

5.2.1. Image-Level Solutions

Since face recognition is based on images of faces, it is a straightforward step to enhance privacy at the image-level. Solutions for this problem are based on image fusion, perturbations, and adversarial learning.

In 2011, Suo et al. [Suo+11] proposed an approach that flips the estimated gender by decomposing the face image and replacing the facial components with similar parts of the opposite gender. This aims at suppressing the gender of the face image. Othman and Ross presented a different approach [OR14] where they proposed a face morphing methodology that iteratively morphs two images and therefore, suppresses gender information at different levels. However, this resulted in morphed images with significant artefacts.

In [Roz+19] and [Roz+16], adversarial images created by using a fast flipping attribute technique showed that it was able to fool their network in predicting binary facial attributes. An incremental flipping approach was proposed by Mirjalili et al. [MR17] with the use of perturbations. In [Chh+18], imperceptible noise was used to suppress k attributes at the same time. However, this noise is trained to suppress attributes from only one specific neural network classifier and consequently, does not generalize to other classifiers.

In [Mir+18; MRR18; MRR19], Mirjalili et al. proposed semi-adversarial networks consisting of a convolutional autoencoder, a gender classifier, and a face matcher. It enhances the soft-biometric privacy on image-level. The autoencoder perturbs the input face image such that it minimizes gender classifier performance while trying to preserve the performance of the face matcher. Training these supervised approaches require a large amount of data with the corresponding privacy-sensitive labels. Moreover, it is limited to the attribute of gender. To overcome this limitation, the authors proposed PrivacyNet [MRR20], a semi-adversarial network extension to suppress multiple attributes. However, this approach is still limited at suppressing pre-defined attributes and thus, it is vulnerable to unseen function creep attacks.

5.2.2. Template-Level Solutions

Recent privacy-enhancing solutions [Ter+19a; Ter+19b; MFV19; Ter+20c; Bor+20; Ter+20h] operate on template-level since most biometric data is stored in templates

rather than images [Dey+14; SRB16]. Moreover, templates offer a less restricted way of encoding information allowing the development of more effective privacy-enhancing approaches. Due to these reasons, the four proposed solutions of this chapter operate on template-level. In contrast to image-level solutions, the following works consider a more critical and challenging scenario of a function creep attacker that knows and adapts to the system’s privacy-mechanism.

In 2019, we proposed an incremental variable elimination (IVE) approach [Ter+19a] to eliminate privacy-risk features from the face templates. Morales et al. [MFV19] proposed SensitiveNet, a neural network that was trained via a modified triplet loss to suppress attribute information. In 2020, Bortolato et al. [Bor+20] proposed PFRNet, an auto-encoder approach that learns privacy-enhancing face representations disentangling identity from attribute information. Since these supervised approaches require privacy-sensitive attribute labels during training, their privacy-protection is limited to the suppression of these pre-defined attributes.

More comprehensive privacy-protection is provided by unsupervised methodologies because these approaches have a more generalized goal of encoding information that does not apply attention mechanisms to single characteristics. In 2019, we proposed similarity-sensitive noise transformations [Ter+19b], more precisely cosine-sensitive noise (CSN) and euclidean-sensitive noise (ESN) transformation. These transformations apply specific noise-injections to the face templates that alter the identity information in a controlled manner while hiding the attribute patterns under noise. In [Ter+20c], we proposed negative face recognition (NFR). While ordinary (positive) face templates contain information of the person’s identity, negative face templates provide random complementary information about this individual. Storing only negative templates in the database, prevents function creep attackers from successfully predicting privacy-sensitive attributes.

While these unsupervised approaches provide a more comprehensive privacy-protection not limited to pre-defined attributes, it is harder to reach high suppression rates while maintaining a high recognition rate as well. In [Ter+20h], we finally proposed a privacy-enhancing solution based on minimum information units (PE-MIU) that overcomes this problem. This training-free approach exploits the structural differences between face recognition (use-case) and the estimation of facial characteristics by function creep attacker (attack-case). In contrast to the attack case which requires exactly one input template, in the use-case of face recognition two templates are given. In this approach, the availability of two templates is used to make the estimation of privacy-sensitive attributes a difficult task. This is achieved by representing the template of an identity in a randomized fashion of template blocks. Due to this kind of representations, function creep attackers can only use minimum information units for their attacks, while for face recognition we can exploit

the second template to align, and thus compare, both representations.

5.2.3. Soft-Biometric Privacy and Cancelable Biometrics

The privacy-issue in biometrics can also be seen from the perspective of cancelable biometrics. Similar to soft-biometric privacy, cancelable biometrics approaches apply one-way functions to transform biometric data [PRC15; MK19; AK20] and store the transformed data [Cas+17]. However, the solutions from both areas target different goals. In cancelable biometrics, the privacy-preservation comes from the computational difficulty to recover the original biometric from the transformed one [PRC15]. The transformed representations aim to achieve irreversibility, revocability, and unlinkability [PS17]. In contrast to this, soft-biometric privacy does not aim at revocability and non-linkability. It aims at suppressing soft-biometric information in biometric data while maintaining a high recognition ability [MRR18; Ter+19b].

5.3. Incremental Variable Elimination

In this section, we propose IVE [Ter+19a], an incremental variable eliminations algorithm that aims at enhancing the soft-biometric privacy of face templates. Our approach is based on decision tree ensembles that allow deriving an importance measure for each privacy-risk variable. In each incrementation step of our solution, the ensemble is trained to predict a sensitive attribute, determine the most important variables of this ensemble, and eliminate these attributes. This allows suppressing sensitive attributes to a high degree.

The challenge of soft-biometric privacy describes a trade-off between maintaining recognition performance and suppressing private attribute estimations. Therefore, we analysed the recognition performance and investigated the soft-biometric attribute estimation performance on the publicly available ColorFeret database [Phi+00]. Unlike previous work, we designed our experiments in the context of an attacker who knows about the used privacy mechanism.

The experiments show that, in many cases, IVE is able to suppress gender and age to a high degree while maintaining a high recognition performance. Especially for function creep attacks with high confidence, IVE shows a significant sensitive attribute suppression.

The main contribution of this section is a privacy-preserving solution that

- i) is able to suppress binary, categorical, and continuous attributes;
- ii) works on the biometric template level; and

-
-
- iii) considers a more challenging attack scenario of an attacker that adapts to the systems privacy mechanism.

So far, previous work proposed several solutions to preserve soft-biometric privacy in face representations. However, all solutions limited their contribution to binary attribute suppression. Moreover, their privacy-preservation approaches were restricted on image level and only consider attacks which do not know about the systems privacy mechanism. In this work, we

- i) present an approach that we used to suppress binary and categorical attributes;
- ii) use them on biometric templates, since most biometric representations are stored in form of templates rather than images [Dey+14; SRB16]; and
- iii) further investigate the privacy performance in a more critical and challenging scenario than previous work, *viz.* in the context of a function creep attacker that adapts his attack to the systems privacy mechanism.

5.3.1. Methodology

Improving soft-biometric privacy aims at suppressing soft-biometric attributes such that function creep attackers would not be able to make reliable estimations of these attributes. In this work, we propose an incremental variable elimination algorithm that aims at suppressing privacy-sensitive attributes in face representations. This approach is based on decision tree ensembles by exploiting the fact that they can be used to derive an importance measure for each variable. By incrementally learning these ensembles and eliminating the high privacy-risk variables, it allows to suppress privacy-sensitive attributes while approximately preserving the recognition ability of the biometric templates. An overview of IVE is illustrated in Figure 5.1. Since this solution is based on tree ensembles, it allows a better generalizability even with less amount of data. This is critical because the main goal is to prevent the storage of a large amount of privacy-sensitive data. Contrary to previous work, the training time of this solution is short and works for binary, categorical, and continuous attributes. In the following, the two building blocks for IVE are introduced, the decision tree construction and the variable importance measure.

Tree construction and splitting

A decision tree represents a tree structure T that produces a random output variable $y \in \mathcal{Y}$ from a random input vector $x \in \mathcal{X}$. This tree consists of internal and terminal nodes.

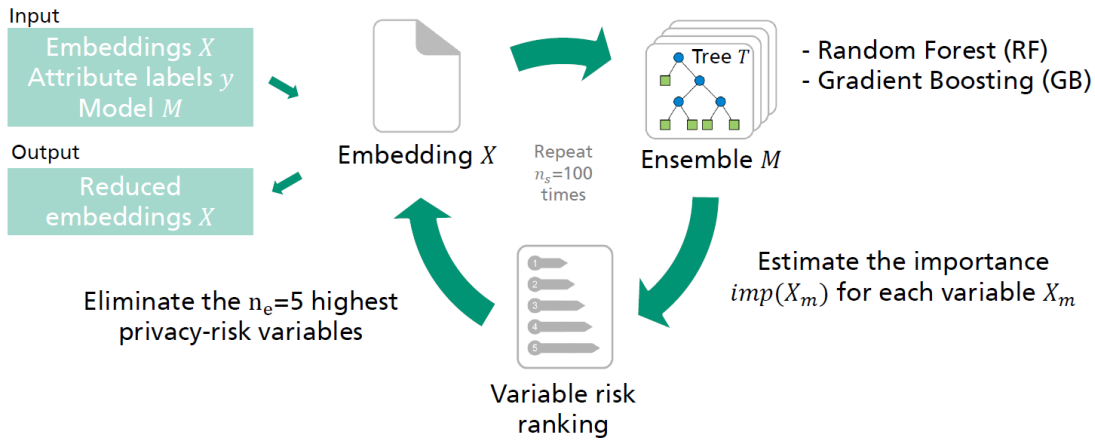


Figure 5.1.: Overview of the proposed incremental variables elimination approach. On the embeddings X a decision tree ensemble is trained to predict the privacy-sensitive attribute. The decision tree ensemble allows to derive importance statements $imp(X_m)$ about each variable m . Then the n_e highest privacy-risk variables are determined and eliminated from X . These steps are repeated n_s times.

Each internal node t is labelled with a binary test on a variable. The branches of this node represent the outcomes of this test, leading to nodes t_L and t_R . Each terminal node represents a predicted class label with the best guess value of the output variable y .

To construct a tree, a recursive procedure over a training set is done. This procedure determines, at each node t , the split s_t for which the partition of the N_t node samples into t_L and t_R maximizes

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R). \quad (5.1)$$

The variables $p_L = \frac{N_{t_L}}{N_t}$ and $p_R = \frac{N_{t_R}}{N_t}$ describe the proportion of samples that traversed the tree to the left and right child nodes, while $i(t)$ is an arbitrary node impurity measure of a node t . The proportion of samples that belongs to class c of all samples that traverse through node t are given by $p_c(t)$ [Lou+13]. In our case, we choose the gini impurity

$$i(t) = I_G(t) = 1 - \sum_c p_c(t)^2, \quad (5.2)$$

as the node impurity measure. We use ensemble methods to reduce the variance, because a single tree typically suffer from the high variance problem.

Variable importance measure

The mentioned node impurity measure $i(t)$ (eq. 5.2) of a node t can be utilized to derive an importance measure for each variable X_m . This importance measure is given by

$$imp(X_m) = \frac{1}{N_T} \sum_{T \in \mathcal{T}} \sum_{t \in T: v(s_t) = X_m} \frac{N_t}{N} \Delta i(s, t), \quad (5.3)$$

and describes the total decrease in node impurity, weighted by the proportion of samples reaching that node and averaged over all trees $T \in \mathcal{T}$ in the ensemble of size N_T . The inner sum goes over all nodes t in the tree T , where the split s_t was performed on the variable X_m . The variable importance measure uses the fact that variables found at the top of the trees contribute to the final prediction decision of a larger fraction of input samples. Therefore, this expected fraction can be utilized to estimate the relative importance of a feature given a decision tree model [Lou+13].

Incremental variable elimination (IVE)

The variable importance measure $imp(X_m)$ (eq. 5.3) offers the opportunity to develop a variable elimination procedure that allows to suppress privacy-sensitive attributes in face representations. The idea is to train a decision tree ensemble \mathcal{M} to predict a soft-biometric attribute y and utilize this learned tree ensemble to estimate the importance of each variable. Due to the tree structure and degree of randomness in the trees, the model will focus more on some features in the upper levels of the trees. Therefore, the importance estimations for variables appearing at the bottom levels are inaccurate. In order to find and eliminate the $n_s \times n_e$ truly most important variables, the procedure of training the model, determining the n_e most important variables, and eliminating these, have to be repeated incrementally. Here, n_s and n_e describes the number of incremental steps and the number of variable eliminations per step.

The IVE approach is explained in Algorithm 2. As an input, it takes the two algorithm parameters n_s and n_e , the training model \mathcal{M} , and the data (X, y) . The algorithm performs n_s steps. In each step, it trains the decision tree ensemble model \mathcal{M} using the training data X and its corresponding label y . Then, it estimates the importance for each variable using Equation 5.3 and determines the n_e most important variables. Here, the function $findHighest(n_e, var_{imp})$ returns the n_e variables X_m with the highest variable importance $imp(X_m)$. Next, function $eliminateVar(X, idx_{cur})$ eliminates the variable in X that are of highest importance for the trained model. The whole algorithm returns a list of eliminated variables var_e that we use to enhance the privacy of the biometric templates. By simply eliminating the variables var_e from an unseen representation, the degree of information

Algorithm 2 - IVE ($n_s, n_e, \mathcal{M}, X, y$)

Input: number of steps n_s , number of eliminations n_e , train model \mathcal{M} , data matrix X , and labels y

Output: List of eliminated variables var_e

```
1:  $var_e \leftarrow$  empty list
2: for  $i \leftarrow 1, n_s$  do
3:    $model \leftarrow \mathcal{M}.train(X, y)$ 
4:    $var_{imp} \leftarrow$  empty list
5:   for all variables  $X_m$  do
6:      $var_{imp} \leftarrow var_{imp} + (X_m, imp(X_m))$ 
7:   end for
8:    $idx_{cur} \leftarrow findHighest(n_e, var_{imp})$ 
9:    $X, y \leftarrow eliminateVar(X, idx_{cur})$ 
10:   $var_e \leftarrow var_e + idx_{cur}$ 
11: end for
12: return  $var_e$ 
```

about the sensitive attribute y can be reduced and thus, it prevents function-creep attackers from reliable estimations.

5.3.2. Experimental Setup

Database - For the experiments, we utilized the ColorFeret database [Phi+00], because it contains high resolution (512x768 pixels) face images with the corresponding information about identity, gender, and age. The database consists of 14,126 images from 1,199 different individuals with different poses under controlled conditions including a variety of face poses, facial expressions, and lighting conditions. To eliminated variabilities induced by the pose variation, we focused on frontal images and reduced the age categories to four (20, 30, 40, 50 years) to create a balance on the age labels. Around 64% of these samples are of male subjects, while the remaining 36% show female subjects.

Evaluation metrics - Only mention the evaluation metrics and refer to the background chapter Enhancing the soft-biometric privacy is about degrading the attribute estimation performance while preserving the recognition ability. For evaluating the recognition performance, we report our results in the widely used equal error rate (EER) metric. This metric gives the false acceptance rate at a threshold where it equals the false rejection rate. The attribute evaluation part of this work is about classifying gender and age classes. We report our results in terms of correct overall/female/male classification rate

(COCR/CFCR/CMCR). The CFR/CMCR describes the percentage of all correctly classified female/male samples, while the COCR represents the general accuracy.

Workflow details - For preprocessing, we aligned the face images using Dlib [Kin09] and cropped them to 256x256 pixels. Afterwards, the images were reshaped to 112x96 pixels and normalized such that the pixel values are within a range of [-1,1]. These cropped, reshaped, and normalized images were passed to the pretrained SphereFace network [Liu+17] to extract a 512-dimensional embedding. For the experiments, these embeddings were used and normalized using z-score scaling. All results are reported in terms of 10-fold cross-validation. First, the proposed IVE algorithm is trained and used to eliminate variables ($n_s = 100, n_e = 5$). Then, the base estimators are trained and evaluated on the reduced templates. For tuning the hyperparameters of these estimators, 20 steps of Bayesian optimization was applied.

Investigations - In this section, we investigate the proposed incremental variable elimination (IVE) algorithm for the purpose of enhancing soft-biometric privacy. The algorithm is based on a tree ensemble and thus, we evaluated two common decision tree ensemble methods, random forest (RF) and gradient boosting trees (GB). In order to analyse the effect of IVE towards the attribute suppression, we chose 8 widely used classifiers for the attribute prediction.

To evaluate the identity preservation of IVE, the recognition performance is analysed over the number of important variables eliminated. The estimation performance of the soft-biometric attributes gender and age was further analysed over the same number of eliminations. This aims at analysing the attribute suppression ability of IVE. In order to develop a deeper understanding of the effect of IVE on different base estimators, their decision performance is analysed for each attribute class separately. Some applications need high confidence that the current instance is correctly classified to the attribute class of interest. In order to simulate such a scenario, the true positive rate (TPR) of the female and male classes was evaluated at a fixed false positive rate (FPR) of 5%.

5.3.3. Results

In the context of soft-biometrics, privacy preservation describes a complex trade-off between identity preservation and the suppression of soft-biometric attributes. In the beginning, we present a visually aided analysis of the gender separability and the effect of IVE with RF. This visualization was done by utilizing t-distributed stochastic neighbor embedding (t-SNE) with 750 randomly chosen samples and can be seen in Figure 5.2. Without any variable eliminations ($n_s \times n_e = 0$), the female and male clusters are very well separated. With more important variables eliminated, this clear separation partially breaks and with $n_s \times n_e = 500$ out of 512 possible eliminations, the two classes become

randomly distributed.

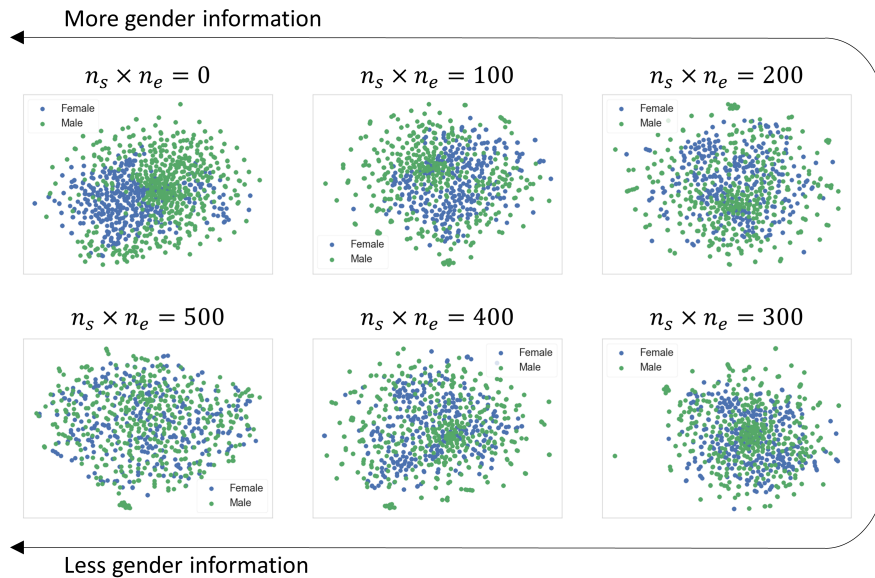


Figure 5.2.: Visualizing the effect of IVE for different numbers of eliminated variables ($n_s \times n_e$). For the visualization, the t-distributed stochastic neighbour embedding (t-SNE) were utilized with 750 randomly chosen samples. The blue and green markers indicate female and male samples.

In order to evaluate the identity preservation ability of IVE, Figure 5.3 shows the recognition performance over the number of eliminated variables $n_e \times n_s$ for two tree ensemble methods. Without IVE, the EER lies around $(3.1 \pm 0.1)\%$. With IVE, the EER grows to around 4% with 300 eliminations. The recognition error grows exponential with further eliminations.

Figure 5.4 shows the total gender decision performance over the number of eliminated variables. Both tree ensembles lead to a very similar base estimator behaviour. This is due to the fact that both ensembles cause similar variable eliminations. In most cases, eliminating more variables lead to a lower gender estimation performance and eliminating the most important 300 to 400 variables lead to a significant performance decrease, because the classifiers can not find reliable indications for a class. Two exceptions are logistic regression and support vector machines with RBF-kernel. These are able to maintain high performance over many eliminations and drops down heavily around 400 eliminations to random behaviour.

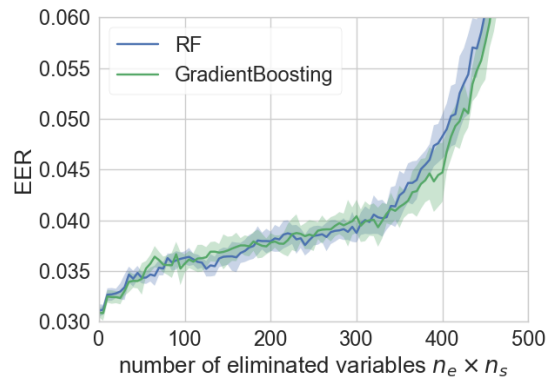


Figure 5.3.: Recognition performance in terms of equal error rate (EER) over the number of eliminated variables. The variable importance for the elimination decisions were determined by the tree ensembles random forest (RF) and gradient boosting trees (GradientBoosting).

In order to develop a deeper understanding of the effect of IVE, Figure 5.5 shows the decision performance divided by each gender. For both ensemble methods, the CMCR drops significantly towards 0% with a growing number of eliminated variables, while the CFCR remains at high values. This indicates that IVE causes the base estimators to always predict the same class.

For many applications, the trustworthiness of the attribute predictions must be reliable as explained in Section 5.3.2. To take this into account in the analysis, Figure 5.6 shows the true positive rate (TPR) of the female and male samples at a fixed false positive rate (FPR) of 5%. It can be seen that the classifications rate drop per gender class is very steep and 200 to 300 variable eliminations are enough to cause a random behaviour for most estimators. Again LogReg and SVM(RBF) are two exceptions which needed 400 variable eliminations in order to cause a random behaviour and thus, an optimal privacy protection for this attribute.

Previous work only considered suppression of binary attributes. In this work, we further analyse the effect of IVE on a categorical attribute such as age classes. Figure 5.7 shows the classification accuracy for the age classes dependent on the number of eliminated variables. For both ensemble methods, two observations can be made: first, the accuracies are considerably lower than in the case of gender estimation. This is probably because in this case, more possible outcomes are available and age estimation from face is generally a harder problem than gender estimation. Second, with a growing number of eliminations

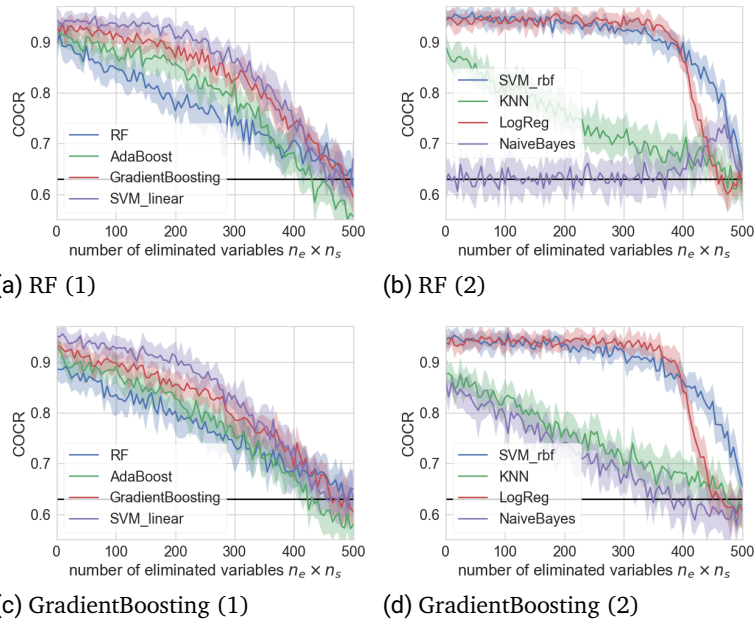


Figure 5.4.: Gender suppression performance in terms of correct overall classification rate (COCR) considering eight base estimators. The black line indicate a random classifier behaviour.

the performance decrease is more flat than in the binary case. This has two reasons: firstly, the overall performance is lower and thus, there is less space for a steeper decrease. Secondly, this might indicate that the age information in the SphereFace representation is more evenly distributed than in the gender case and thus, more variables have to be eliminated to suppress this attribute.

In Figure 5.8 and 5.9, the age estimation performance is shown for the different age classes. Due to similar learned tree structures, the behaviour of the base estimators on the eliminations is similar. Furthermore, high prediction accuracy of young ages (20) can be observed, while adjacent age groups show a worse performance. This is probably because the age group of around 20 years is overrepresented in the database. With a growing number of important variable eliminations, the performance of the age classes 30-50 years will decrease, while the performance of the 20 years class will remain very high. Consequently, IVE forces the classifiers to only predict the majority class.

To summarize the main results, Table 5.1 shows the effect of IVE for a Gradient Boosting

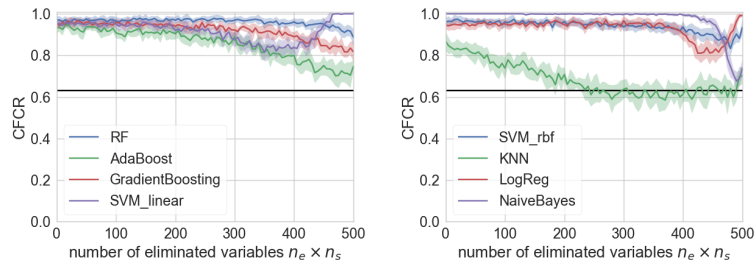
Tree. It shows the recognition performance, as well as the gender and age estimation performance over the number of eliminated variable $n_s \times n_e$. The estimations performance is shown for two widely used classifier, RF and SVM. With more eliminated variables, the estimation performance drops, while the recognition performance is only slightly affected.

Table 5.1.: Overview of IVE performance for Gradient Boosting: the recognition performance [%] is shown over the number of eliminated variables $n_s \times n_e$, as well as the gender and age COCR [%] for two classifiers.

$n_s \times n_e$	Recognition EER	Gender COCR		Age COCR	
		RF	SVM	RF	SVM
0	3.1	89.8	94.8	57.0	68.7
100	3.5	84.9	94.8	56.2	67.6
200	3.8	77.9	94.0	50.1	65.7
300	4.0	73.8	92.3	50.6	62.3
400	4.5	70.5	86.4	47.9	58.8
500	12.3	61.5	64.7	42.9	47.5

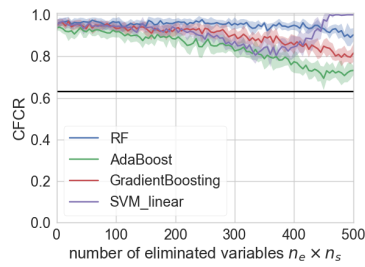
5.3.4. Interim Conclusion

In this section, we successfully present the incremental variable elimination (IVE) algorithm to enhance the privacy of face templates. By incrementally eliminating the most important variables from the face templates, we are able to incrementally suppress sensitive attributes while maintaining the templates recognition ability. We conducted the experiments on a publicly available database in the context of attackers with prior knowledge about the systems privacy mechanism. Comparisons with eight base estimators showed that in many cases, our IVE solution was successfully able to suppress gender and age to a high degree with a neglectable loss in the recognition performance. It was shown that in many cases, IVE forces the base estimators to always predict the same class. Especially investigating estimations at a high confidence level showed a significant performance drop. Unlike previous work, training of IVE does not require a large amount of privacy-sensitive labelled data and is able, by design, to suppress binary, categorical, and continuous attributes. However, this approach is limited to the suppression of pre-defined attributes.

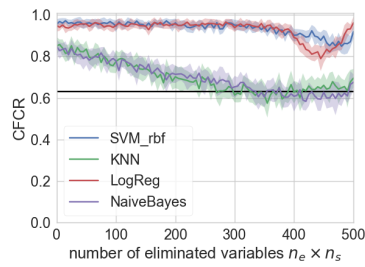


(a) RF-female (1)

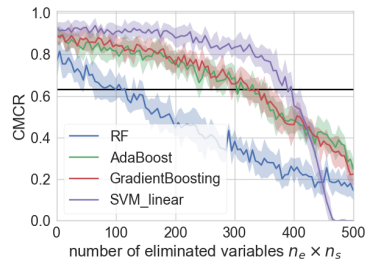
(b) RF-female (2)



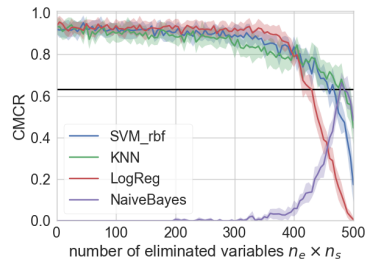
(c) GradientBoosting-female (1)



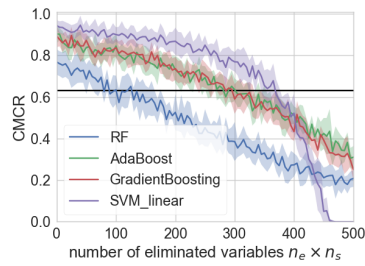
(d) GradientBoosting-female (2)



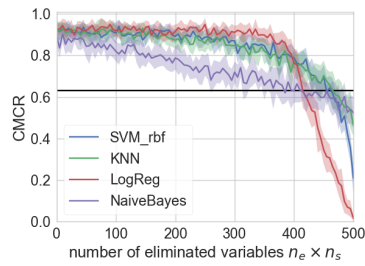
(e) RF-male (1)



(f) RF-male (2)

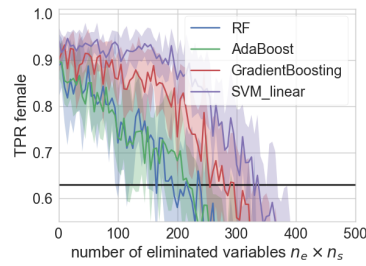


(g) GradientBoosting-male (1)

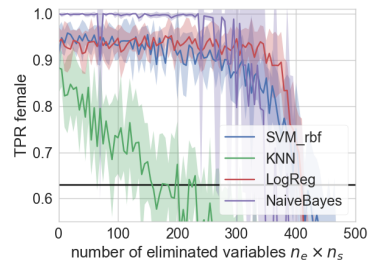


(h) GradientBoosting-male (2)

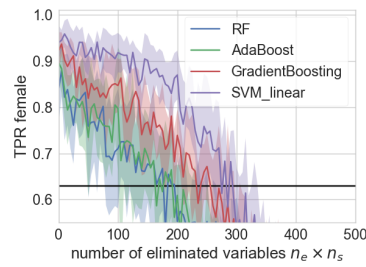
Figure 5.5.: Gender suppression performance per gender considering eight base classifiers.



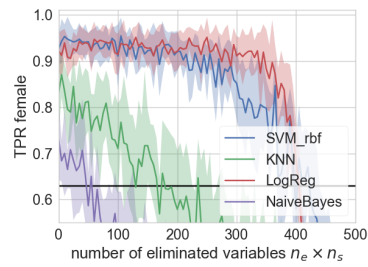
(a) RF-female (1)



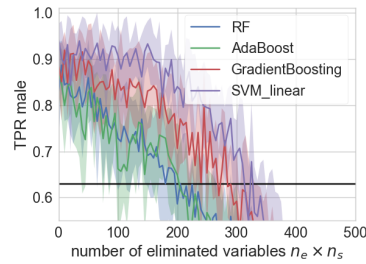
(b) RF-female (2)



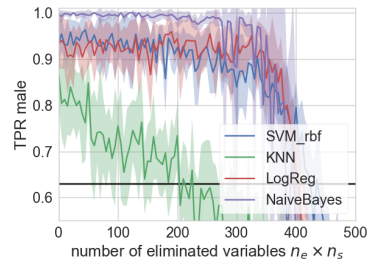
(c) GB-female (1)



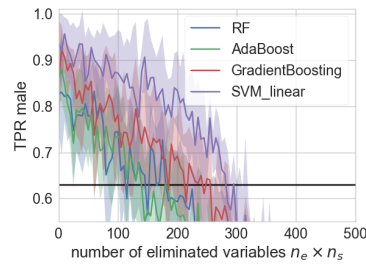
(d) GB-female (2)



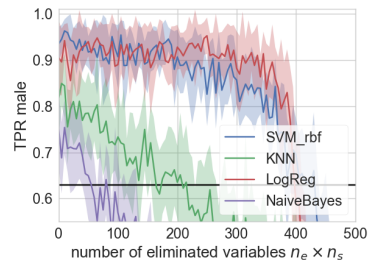
(e) RF-male (1)



(f) RF-male (2)

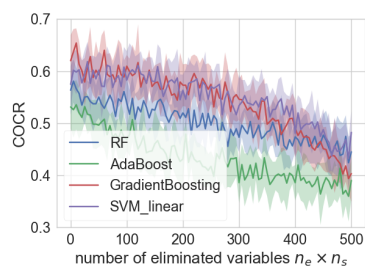


(g) GB-male (1)

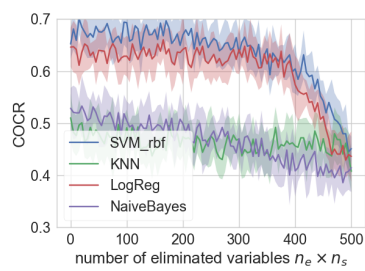


(h) GB-male (2)

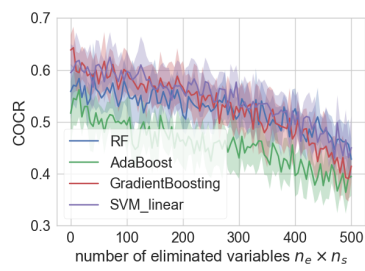
Figure 5.6.: Female and male gender suppression performance for high confident estimations. The results report the true positive rates of the female and male instances at a fixed false positive rate of 5%.



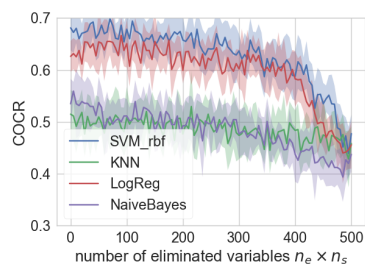
(a) RF (1)



(b) RF (2)

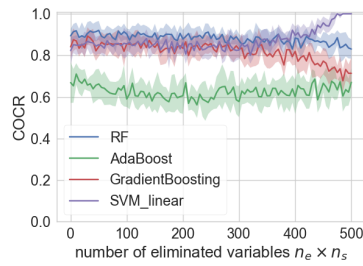


(c) GB (1)

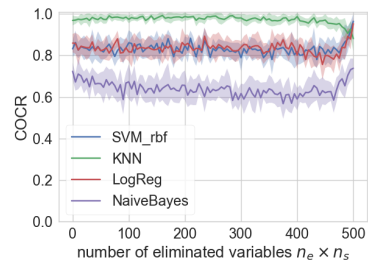


(d) GB (2)

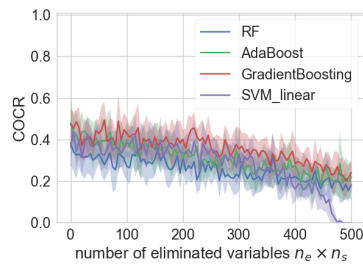
Figure 5.7.: Age class suppression performance over the number of eliminated variables $n_e \times n_s$.



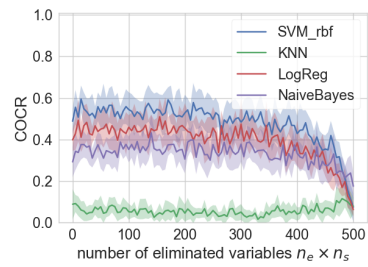
(a) RF-20 (1)



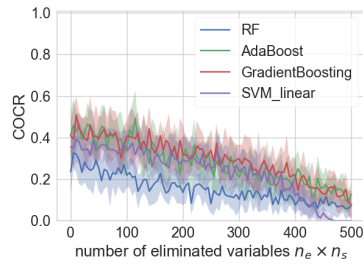
(b) RF-20 (2)



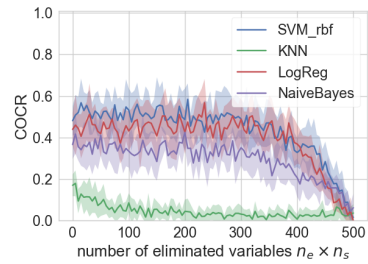
(c) RF-30 (1)



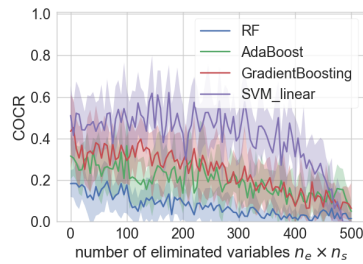
(d) RF-30 (2)



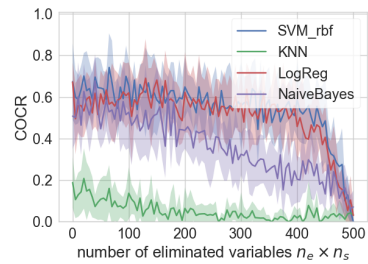
(e) RF-40 (1)



(f) RF-40 (2)

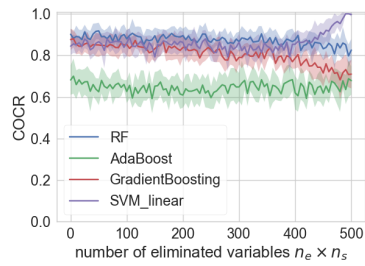


(g) RF-50

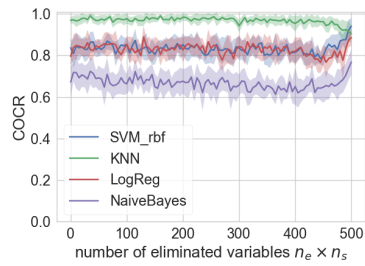


(h) RF-50

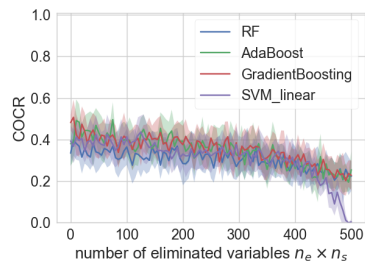
Figure 5.8.: Age suppression performance per age group. IVE was applied based on random forest.



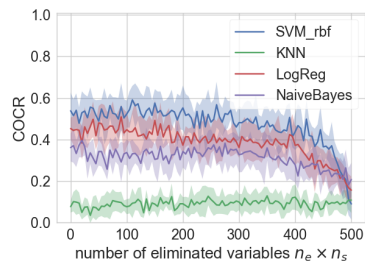
(a) GradientBoosting-20



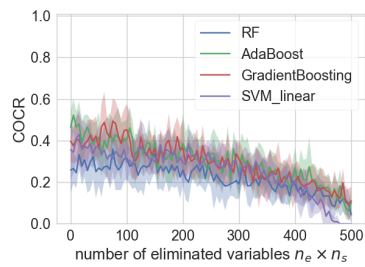
(b) GradientBoosting-20



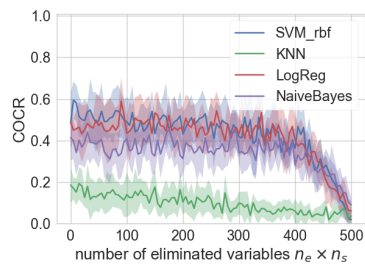
(c) GradientBoosting-30



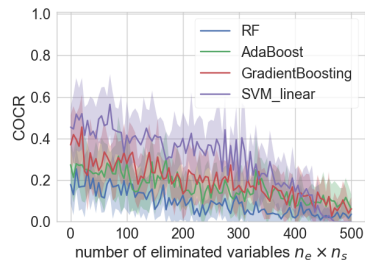
(d) GradientBoosting-30



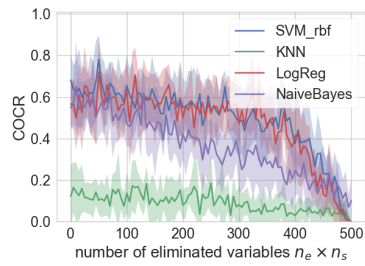
(e) GradientBoosting-40



(f) GradientBoosting-40



(g) GradientBoosting-50



(h) GradientBoosting-50

Figure 5.9.: Age suppression performance per age group. IVE was applied based on gradient boosting trees.

5.4. Similarity-Sensitive Noise Transformations

Previous works on the enhancement of soft-biometric privacy require a large amount of data with privacy-sensitive annotations. However, (a) this annotated data is often not available in large quantities and (b) this limits the privacy-enhancement to pre-defined attributes. In Section 5.3, we proposed an approach is able to enhance the user’s privacy with less annotated data. However, the limitation on pre-defined attributes makes it vulnerable to unknown attacks. Therefore, in this section, we propose and investigate similarity-sensitive noise transformations and dimensionality reduction techniques, to suppress soft-biometric information in face representations. Unlike previous work, in this section, we propose and analyse unsupervised techniques that do not require this information.

The challenge of soft-biometric privacy describes a trade-off between maintaining the recognition performance and suppressing the private attribute estimation. Therefore, we performed a comprehensive investigation on the publicly available ColorFeret database. This includes an analysis of the recognition performance and an investigation of the soft-biometrics attribute estimation performance. For these attributes, we analysed the binary attribute of gender and the continuous attribute of age. While gender is a widely studied soft-biometric attribute, a continuous attribute like age was never investigated in the soft-biometric privacy literature. Further, we investigated scenarios in which an attacker has and has not prior knowledge about the used privacy mechanism. The results show that such an informed attacker is able to make significantly better predictions than an attacker without this prior knowledge. This also holds in the context of noisy face representations. Here, applying similarity-sensitive noise transformations lead to a weaker estimation performance and to less confident predictions. Furthermore, we investigated how the applied methods affect the prediction performance of different kind of estimators. In order to measure the benefits of the privacy-preserving methods, an evaluation metric was proposed which captures the trade-off between the privacy gain and the identity preservation loss. The experiments show that the proposed cosine-sensitive noise transformation has a very promising privacy gain without significantly changing the recognition performance.

In this section, we develop a methodology which works in an unsupervised manner, and thus, requires no prior knowledge about soft-biometric attributes that should be suppressed [Ter+ 19b]. Unlike previous works, we evaluate our approach on binary and continuous attributes and consider scenarios in which the attacker has prior knowledge about the privacy mechanism.

5.4.1. Methodology

Enhancing soft-biometric privacy aims at suppressing soft-biometric attributes so that function creep attackers can not make reliable estimations. In this work, we want to investigate solutions for this problem that can be deployed without the explicit knowledge of the private attributes. Therefore, we propose an unsupervised approach based on similarity-sensitive noise transformations. These proposed transformations add noise to a feature representation $\vec{v} \in \mathbb{R}^n$ so that the privacy of this data is enhanced in an unsupervised manner. The advantage of these noise transformations is that the noise level, and thus the impact of the noise, can be controlled in terms of similarity. Therefore, the effect on the recognition performance can be estimated and limited. In this work, we suggest two variants of similarity-sensitive noise transformations. The euclidean-sensitive noise transformation interprets a representation \vec{v} as a point in a n dimensional space and moves this point to the $n - 1$ dimensional sphere around \vec{v} with radius r . The cosine-sensitive noise transformation interprets the representation \vec{v} as a vector and creates a vector which lies on the $n - 1$ dimensional cone around vector \vec{v} with an fixed cosine similarity of Θ .

Euclidean-sensitive noise - Given a vector $\vec{v} \in \mathbb{R}^n$ with $\|\vec{v}\| = 1$, the euclidean-sensitive noise transformation returns a point \vec{x} which lies on the $n - 1$ dimensional sphere with radius r around \vec{v} . To ensure that these randomly generated points on the sphere follow a uniform distribution, Marsaglia's algorithm [Mar72] is used. Creating a randomized vector \vec{z} in which each component

$$z_i \sim \mathcal{N}(0, 1), \quad (5.4)$$

is gaussian distributed, allows to directly compute a random point

$$\vec{x} = \vec{v} + \frac{r}{\|\vec{z}\|} \vec{z}, \quad (5.5)$$

on the $n - 1$ dimensional sphere around \vec{v} with radius r . A three-dimensional visualization of this sphere is shown in Figure 5.10a. This transformation ensures that the euclidean distance between the untransformed vector \vec{v} and transformed vector \vec{x} remains at r and thus, the recognition loss is restricted.

Cosine-sensitive noise - In this work, we propose the cosine-sensitive noise transformation. This transformation creates a point \vec{x} on a $n - 1$ dimensional cone with angle θ

around vector \vec{v} . A three-dimensional visualization can be seen in Figure 5.10b. In order to sample such a point \vec{x} on this cone, a randomized vector

$$\vec{z} = \frac{\vec{z}_0 + \vec{v}}{\|\vec{z}_0\| + \|\vec{v}\|}, \quad (5.6)$$

has to be created that is uniformly distributed around vector \vec{v} . Here, $\vec{z}_0 \in \mathbb{R}^n$ is a vector with gaussian distributed components ($z_{0,i} \sim \mathcal{N}(0, 1)$). A vector \vec{x}' can then be created by a linear combination of \vec{v} and \vec{z}

$$\vec{x}' = \mu\vec{v} + \lambda\vec{z}. \quad (5.7)$$

The parameters μ and λ can be determined by setting two conditions. The first condition aims at normalizing \vec{x}'

$$\|\vec{x}'\| = 1, \quad (5.8)$$

which leads for an expression for λ

$$\lambda(\mu) = \pm\sqrt{\mu^2 [(\vec{v} \cdot \vec{z})^2 - 1] + 1} - \mu(\vec{v} \cdot \vec{z}). \quad (5.9)$$

The second condition aims at restricting the cosine similarity between \vec{v} and \vec{z} . Fixing this similarity to be Θ

$$\Theta = \cos(\theta) = \cos(\vec{v}, \vec{x}) \quad (5.10)$$

$$= \frac{\vec{v} \cdot \vec{x}}{\|\vec{v}\| \cdot \|\vec{x}\|} = \vec{v} \cdot \vec{x} = \mu + \lambda(\mu) (\vec{v} \cdot \vec{z}), \quad (5.11)$$

controls the cosine similarity between \vec{v} and \vec{x} and it makes the system of equations solvable. Here, θ describes the angle between \vec{v} and \vec{x} , This leads to a solution for parameter μ

$$\mu = \frac{\pm(\vec{v} \cdot \vec{z})\sqrt{\Theta^2[(\vec{v} \cdot \vec{z})^2 - 1] - (\vec{v} \cdot \vec{z})^2 + 1}}{(\vec{v} \cdot \vec{z})^2 - 1} + \Theta. \quad (5.12)$$

In the last step, the length of the vector \vec{x}' is changed randomly, since this does not affect the cosine similarity between \vec{v} and \vec{x}

$$\vec{x} = z \cdot \vec{x}' \quad \text{where } z \sim \mathcal{U}(1, 100). \quad (5.13)$$

Here, the random variable z is drawn from a uniform distribution in the range of 1 to 100, which is an arbitrary choice. The transformation $\vec{v} \rightarrow \vec{x}$ ensures that the angle, and thus the cosine similarity, between these vectors is fixed at θ . As an advantage, this leads to a restricted recognition loss, which can be controlled easily, while perturbing other patterns in the noise-prone face representation.

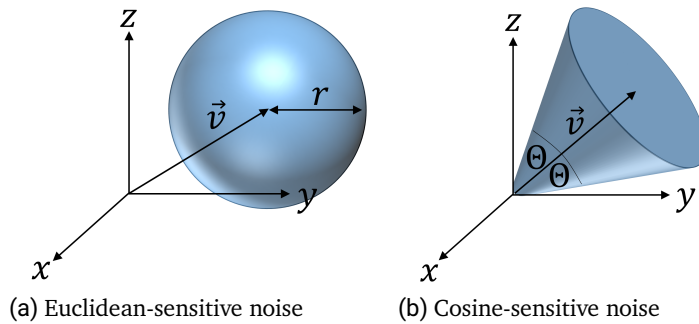


Figure 5.10.: Visualization of the geometric noise sampling principles for a given vector \vec{v} in three dimensions.

5.4.2. Experimental Setup

Database - For the experiments we utilized the ColorFeret database [Phi+00], because it contains high resolution (512x768 pixels) face images with the corresponding information about identity, gender, and age. The database consists of 14,126 images from 1,199 different individuals with different poses under controlled conditions. Further, a variety of face poses, facial expressions, and lighting conditions are included in the dataset. For the experiments, we focused on frontal images. Around 40% of these images are of female subjects, while the age varies from 10 to 70 years. An age distribution of the database can be seen in Figure 5.11.

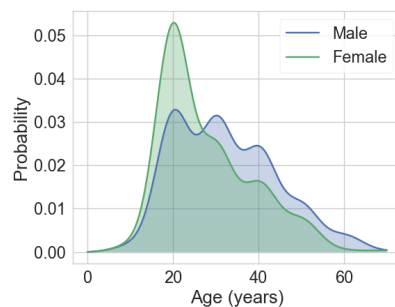


Figure 5.11.: Age distribution of the used database.

Evaluation metrics - In Section 5.4.3, the results and discussions will be divided into four

parts, which all analyse a different aspect of the problem. For each aspect a characteristic evaluation metric is used. For the identity preservation part, the comparison of two biometric template are done using cosine similarity (eq. 5.10) and we report our results in the widely used equal error rate (EER) metric. This metric gives the false acceptance rate at a threshold where it equals the false rejection rate. For the gender classification task, most results are reported in the total gender decision accuracy, which is slightly biased to the majority class of male. In the age estimation part, the mean absolute error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - x_i|, \quad (5.14)$$

is used to evaluate the regression performance. It measures the average of the absolute different between the predicted values x_i and the true ages y_i . In order to evaluate if the privacy enhancing method is beneficial, we propose the privacy-gain identity-loss coefficient (*PIC*).

$$PIC = \frac{AE' - AE}{AE} - \frac{RE' - RE}{RE} \quad (5.15)$$

This value is defined by the attribute prediction errors AE and AE' and the recognition errors RE and RE' before and after the privacy enhancing transformation. It measures the difference of the relative error increase of the attribute estimation and the recognition performance. Positive values indicate that the privacy gain is higher than the loss in the identity preservation performance and thus, measures how beneficial it is to apply the privacy transformation.

Workflow details - For preprocessing, we aligned the face images using Dlib [Kin09] and cropped them to 256x256 pixels. Afterwards, the images were reshaped to 112x96 pixels and normalized such that the pixel values are within a range of [-1,1]. These cropped, reshaped, and normalized image were given to the pretrained SphereFace network [Liu+17] to extract a 512 dimensional embedding.

For the experiment, these embeddings were used and normalized using z-score scaling. After this scaling, the dimensionality reduction is performed, followed by the similarity-sensitive noise transformation. In order to train and test the soft-biometric estimators, a 5-fold cross validation was performed. In each fold, 33% of the data was randomly chosen and used for testing. The rest was used for training. For tuning the hyperparameters of the estimators, 20 steps of Bayesian optimization was applied.

Investigations - In this work, we investigate the effect of the proposed similarity-sensitive noise transformation and dimensionality reduction techniques on unsupervised privacy preservation for face images. For the dimensionality reduction task, we investigate

the effect of employing linear principle component analysis (PCA) [TB99], non-linear kernelized PCA (KernelPCA) [SSM99] and independent component analysis (ICA) [HO00]. For the similarity-sensitive noise transformations, we will analyse the effect of cosine- and euclidean-sensitive noise, described in Section 5.4.1. The investigation of the privacy-enhancing solutions is divided in four subsections analysing different aspects of the problem: (1) recognition performance, (2) gender estimation performance, (3) age estimation performance, (4) trade-off discussion.

(1) In Subsection 5.4.3, the identity preservation performance of the solutions is considered. Improving the privacy causes a decrease in the identity preservation performance. In this part, the influence of the solutions on the recognition performance is investigated. The genuine and imposter matching score distributions are analysed for different noise settings and the effect of the dimensionality reduction methods are investigated in terms of EER. Finally, the two aspects, dimensionality reduction and noise level, are jointly analysed.

(2) The goal of soft-biometric privacy-preserving methods is to prevent attackers from reliably estimate private attributes from biometrics templates. In Subsection 5.4.3, the estimation performance of various classifier and similarity-sensitive noise transformations are analysed for the binary attribute of gender. In order to obtain generalized conclusions, three different kind of binary classifiers are used for the experiments [Ped+11]. For a linear classifier, logistic regression is used, while for learning a non-linear decision boundary, a support vector machine (SVM) is applied with a gaussian kernel. To also consider ensemble methods, the random forest was included in the pool of classifiers. In the experiments, the gender decision performance of the classifiers were evaluated in the context of dimensionality reduction and similarity sensitive noise transformation. In order to understand the effect of these methods on the estimations, the matching score distributions were evaluated, as well as the classification performance for the female and male class respectively. Further, the gender decision performance was investigated in a scenario with and without prior knowledge about the privacy-mechanism of the system. Since some applications require very reliable decisions, the true positive rate (TPR) of the female and male classes was evaluated at a fixed false positive rate (FPR) of 5%. This simulates an application which has to be 95% sure about the decision made by the classifier.

(3) Besides the binary attribute of gender, we also investigate the estimation performance of the continuous attribute of age, which contains more degrees of freedom and thus, changes the properties of the problem. The results for this investigations are shown in Subsection 5.4.3. The MAE of the age prediction was evaluated for different dimensionality reduction methods and noise levels. Further, directed estimation error distributions were calculated, in order to understand the effect of the noise transformations on the regressor

estimations.

(4) Soft-biometric privacy consists of a trade-off between privacy-gain and the recognition loss. In Subsection 5.4.3, an investigation on this trade-off was done. In order to figure out in which cases it is beneficial to employ the privacy-preserving solutions and to measure its value. Therefore, PIC curves are calculated for gender and age to evaluate the dimensionality reduction techniques and the similarity-sensitive noise transformations. This leads to clear application recommendations.

5.4.3. Results

In the context of soft-biometrics, privacy preservation describes a complex trade-off between the recognition performance and the possibility of the unauthorized estimation of soft-biometric attributes. To start, we present a visually aided analyses of the attribute separability and the effect of the similarity-sensitive noise transformations. This will be done in the following.

In Figure 5.12, the SphereFace representations of 750 randomly chosen identities from the ColorFeret database where visualized via t-distributed stochastic neighbour embedding (t-SNE) [MH08]. The colouring was done based on their gender (top row) and based on their age (bottom row). In (a), it can be seen that the data can be clearly separated into female and male samples. Introducing similarity-sensitive noise to the data partially breaks this separability. For cosine-sensitive noise (b), a lot of samples are arranged in a circle, while the rest is clustered in the center. This is probably due to the random nature of the cosine-sensitive noise transformation which also changes the length of the resulting vectors. At this state, it is hard to identify a decision boundary to separate the two classes. In (c), euclidean-sensitive noise was introduced. Even if a separability is still visible, the boundary becomes blurred and a layer is formed in which it is not clear which samples belongs to which gender.

For the same two-dimensional representations, also the age of the identities was made visible in Figure 5.12 (d)-(f). There, red markers indicate persons aged 20 years and younger, while blue markers represent persons that are 50 and more years old. The clusters of the two extreme cases can be distinguished, while the age groups between this cases are found around these clusters. In (e), cosine-sensitive noise was introduced to the data with the already described centering effect. Inside of this centered cluster, the age structure is still noticeable. Introducing euclidean-sensitive noise to the face representations (f) leads to a small blurring of the age boundaries.

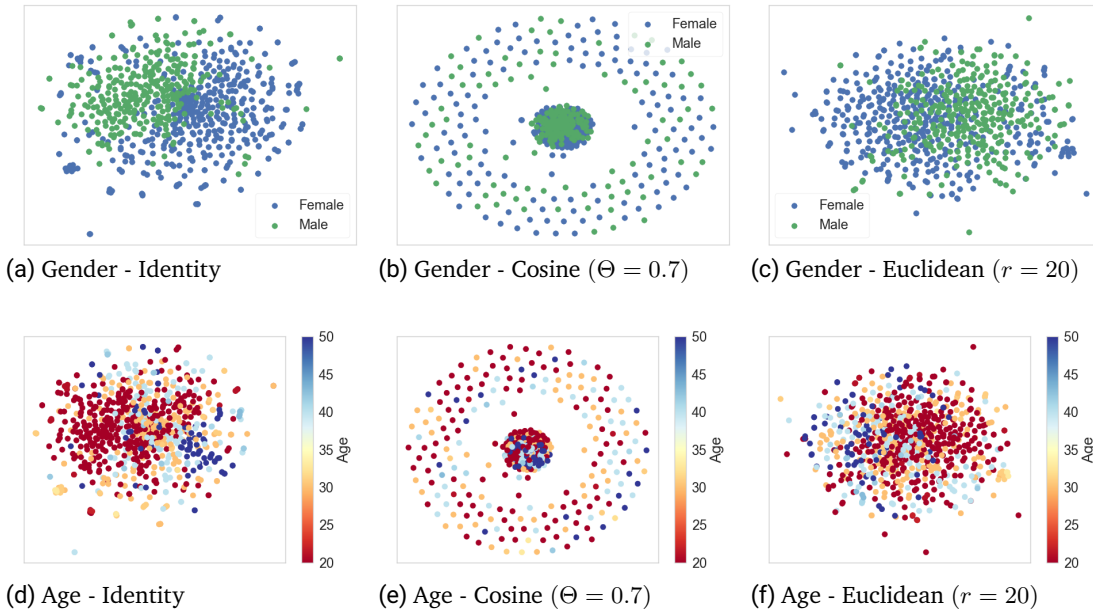


Figure 5.12.: Visualization of gender and age structures of the ColorFeret database with and without the presence of noise. For the plot 750 samples were randomly chosen and visualized using t-distributed stochastic neighbour embedding.

Investigate recognition performance

Soft-biometric privacy must go ahead with an preserved identity performance, in order to ensure the intended functionality of the biometrics system. Therefore, in this subsection, the recognition performance is analysed in the context of the proposed similarity-sensitive noise transformations and dimensionality reduction techniques.

In Figure 5.13, generic matching score distributions are shown for the case without noise (a) and for the cases with noise (b)-(c). The overlap between the genuine and imposter score distributions already indicates that introducing noise leads to a small loss in the recognition performance to enhance privacy.

In Figure 5.14, the effect of the dimensionality reduction methods and the similarity-sensitive noise transformations is shown in more details. In (a), the equal error rate (EER) is shown over the number of dimensions for three different dimensionality reduction methods. For 128 dimensions, the EER is close to the EER of the original SphereFace representation (3.2%). This error rate grows with a decreasing number of dimensions.

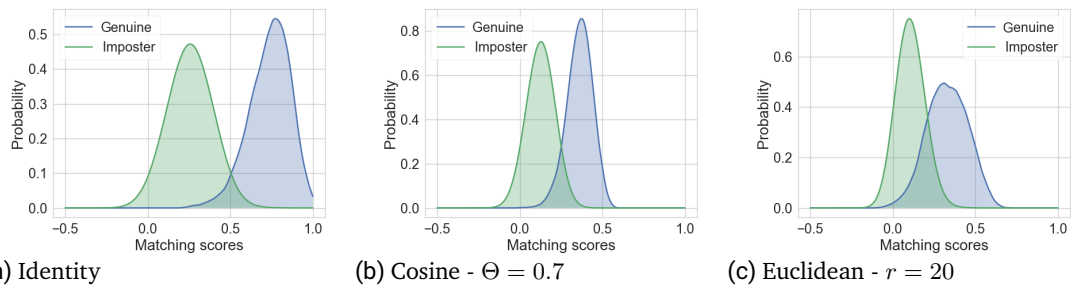


Figure 5.13.: Score distributions for the recognition task. The SphereFace embeddings were used for the comparisons with and without the presence of noise.

In (b), the recognition performance is shown in the context of cosine-sensitive noise. For the reduction to 128 dimensions, the EER is plotted over the noise level, where $\Theta = 1$ indicates a representation vector that is just randomly changed in length. Further, the EER for the original representation (Identity) is plotted. For all cases, the recognition error grows with an increasing noise level. Here, PCA and KernelPCA show a very similar behaviour, probably because neural network representations tend to be entangled, which makes it easier for dimensionality reduction. ICA shows a promising EER without noise, but the independent component structure turned out to be more sensitive to the noise than the other methods.

Introducing noise on the original representations shows to have the least impact in terms of EER, because the variations induced by the noise can spread in more dimensions. In (c), the same was done for euclidean-sensitive noise transformations. Here, at a radius of $r = 0$, all methods perform nearly equally good, since the noise at this point does not change anything. Again, with a growing noise level, the EER grows as well, while the original representation is most robust against the noise in terms of EER. It is noticeable that the recognition performance of ICA drops heavily when euclidean-sensitive noise is introduced. This is because the ICA fails to converge during training when this kind of noise is applied.

In order to investigate the influence of similarity-sensitive noise transformations and dimensionality reduction techniques together on the recognition performance, Figure 5.15 shows two parameter space plots for the two similarity-sensitive noise methods on PCA. Here, red areas indicate a low EER, while a high EER is represented by blue areas. Higher dimensions can tolerate a greater noise level without losing much performance, while in lower dimensions, the applied noise has a strong effect on the recognition rates.

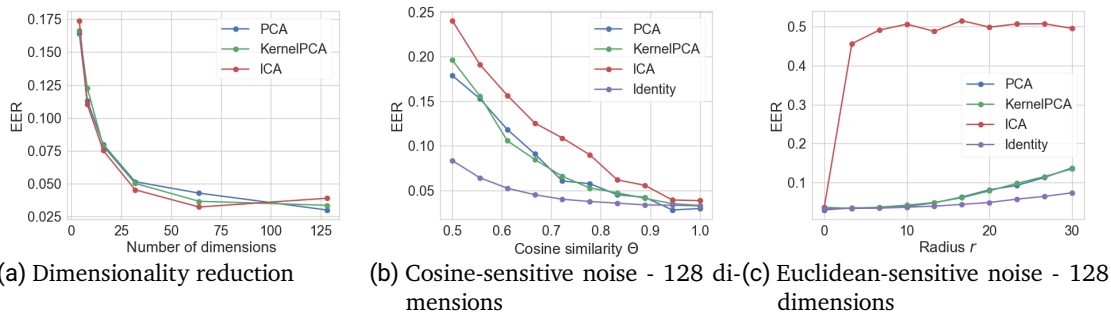


Figure 5.14.: Recognition performance for different dimensionality reduction techniques and similarity-sensitive noise transformations. The "Identity" label refers to the original SphereFace representation.

Summary - recognition performance: In summary of the recognition performance evaluation, the face representations can be reduced to 64 dimensions without substantially losing too much discriminability. However, lower dimensional representations tend to be more susceptible to introduced noise. Therefore, higher dimensions are favoured. In terms of similarity-sensitive noise, noise values in the range of $\Theta \in [0.7, 1]$ for cosine and $r \in [0, 20]$ for euclidean are reasonable. This limits the EER increase to a maximum of 50%, which still leads to a relative low absolute EER.

Investigate binary gender estimation

The main essence of soft-biometric privacy is to suppress the possibility of an unauthorized estimation of soft-biometric attributes from a biometric template. One of the most discussed attributes in related literature is gender. In this subsection, the influence of dimensionality reduction and similarity-sensitive noise transformations on achieving this goal is analysed.

In Figure 5.16, the gender estimation performance of three dimensionality reduction methods is shown for three binary classifiers. Even without the presence of noise, the classification performance is decreasing with less number of dimensions. However, this effect is insignificantly small and is generally independent of the used classifier and the unsupervised dimensionality reduction method.

In order to analyse the classifier behaviour in the context of similarity-sensitive noise, Figure 5.17 shows the score distributions for female and male samples that are learned from three different classifiers. In all cases, it can be observed that the overlap between the classes grows when noise is introduced. Further, the peaks of the two classes come

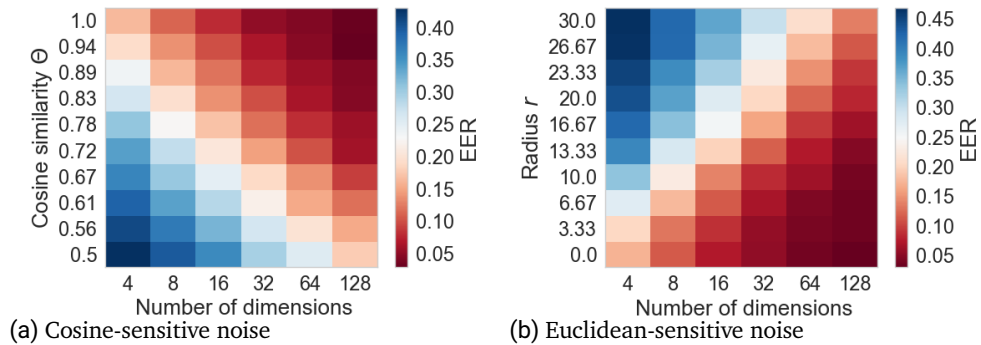


Figure 5.15.: Recognition performance for different dimensions and sampling values for PCA and cosine-/euclidean-sensitive noise transformations.

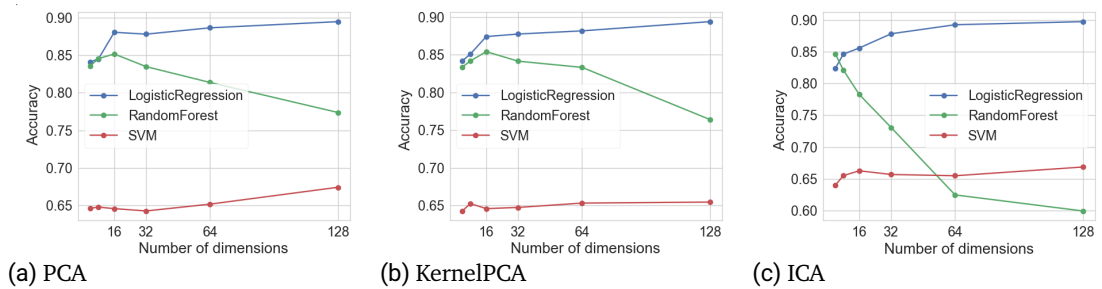


Figure 5.16.: Evaluation of the gender estimation performance for different dimensionality reduction scenarios.

closer to each other in the noise-prone scenarios.

To investigate the effect of the similarity-sensitive noise in two different knowledge level attacks, Figure 5.18 presents the gender decision performance for cosine- and euclidean-sensitive noise and different dimensionality reduction methods. Further, the scenario of an informed and uninformed attacker was simulated. The solid lines represent a scenario of an attack without knowledge about the privacy/noise mechanism (only the classifier testing was done on the noise-prone data), while the dashed lines represent an attacking scenario in which the attacker has prior knowledge about the privacy mechanism (classifier training and testing was done on noise-prone data).

In the top row (a)-(c) of Figure 5.18, the effect of cosine-sensitive noise can be observed.

It can be seen that at the minimum noise level ($\Theta = 1$), high gender accuracies of over 90% can be achieved. The best performance was reached by the logistic regression approach. The entangled feature maps of the original representation favours linear approaches, which are more robust to outliers.

In all cases with and without dimensionality reduction, an increasing noise level resulting in a lower gender estimation performance is observed. For PCA 128 and ICA 128, this loss is even bigger than in the original representations, since the influence of the noise is stronger in lower dimensions. A very significant observation is that, in nearly all cases, an informed attacker is able to make much better predictions than an attacker without prior knowledge. This is significant since all previous works exclusively consider the later case.

In general, it can be seen that cosine-sensitive noise transformation is an promising tool for preserving privacy. Even at the minimum noise level, where just the length of the vectors are randomly changed, the loss of the gender decision performance is significant. This especially holds for classifiers like random forest or support vector machines which are very dependent on the absolute values of the features.

In the bottom row (d)-(f) of Figure 5.18, the effect of the euclidean-sensitive noise is observed. At the minimum noise level, equivalent to the noiseless representation, all classifiers show a very high gender decision performance. In (d), SVM shows the best gender decision performance with close to 95% accuracy. However, it also shows the most significant drop in accuracy when the noise level is increased. In (d) and (e), it can be observed that all classifiers loose performance while increasing the noise level. Considering the effect of different types of attackers, again an informed attacker is able to achieve higher estimation performances in all cases. In (f), the same effect as in Section 5.4.3 can be seen. When introducing euclidean-sensitive noise, ICA fails to converge. Consequently, the data representations become more meaningless and the performance drops.

In order to understand the behaviour of the classifiers more deeply in the context of similarity-sensitive noise and attacking scenarios with and without prior privacy knowledge, Figure 5.19 shows the performance of female and male samples. To be more precise, it shows the probability to correctly classify a given female/male sample. In the top row (a)-(d), the gender decision performances for cosine-sensitive noise transformations are shown. For the original representation (a)-(b), for the logistic regression classifier both the correct female and male classification rate (CFCR/CMCR) drops with growing noise level. For random forest the CMCR remains the same for different noise levels. However, the CFCR drops significantly. An even more extreme behaviour can be observed at the SVM performance. Due to its sensitivity to the absolute feature values, it achieves a very high CMCR over 95% while the CFCR is close to 20%.

In the bottom row (e)-(h) of Figure 5.19, the gender decision performance for euclidean-

sensitive noise is shown. At the minimum noise level ($r = 0$), the correct classification performance is high for both female and male samples. For logistic regression the CFCR and the CMCR decreases with higher noise level as expected. For random forest, the CMCR remains high, however, the CFCR drops significantly when the noise is introduced. Regardless of these results, the behaviour of the SVM is quite unusual. In the scenario of an uninformed attacker, higher noise levels lead to states in which the SVM estimates all samples as males. Compared to the case of an informed attacker, where both correct classification performances decrease, the RBF-kernel SVM probably learned a spherical decision boundary with the female class in the center. Consequently, adding euclidean-sensitive noise shifts the samples in the outer male regions. In the dimensionality reduction cases, the influence of the noise is stronger on the gender decision performance. Therefore, the change of the CFCR and the CMCR is stronger.

For many applications, the trustworthiness of the soft-biometric predictions must be reliable. Therefore, in Figure 5.20, the true positive rate (TPR) of the female and male classes are shown at a fixed false positive rate (FPR) of 5%. This simulates an application in which the estimation must be 95% sure that the sample has a female or male origin. In the top row (a)-(d), the results are shown for cosine-sensitive noise for the original representation and for the PCA representation reduced to 128 dimensions. For logistic regression, at this confident level the TPR drops severely with increasing noise level. A stronger performance drop is seen for random forest. For both, informed and uninformed attack scenarios, the TPR loss is high. However, the TPR of the attacker with prior knowledge is significantly higher than in the uninformed scenario. Same for SVM, a performance decrease with growing noise level can be observed. Nevertheless, in the uninformed attacker scenario, the performance for the confident female estimation is very low, while the performance for the confident male estimation is relatively high. In an informed attacking scenario, this behaviour is better compensated.

The confidence performance results for the euclidean-sensitive noise are shown in the bottom row (e)-(h). Increasing the noise level leads to a reduction in the female/male TPR for all cases using the original face representations. This TPR reduction is significantly higher in the PCA reduced representations.

Figure 5.21 shows the parameter space of the number of dimensions over the noise levels in order to analyse the effect of dimensionality reduction together with the similarity-sensitive noise. Here, dark red areas represent high accuracies, while dark blue regions indicate a very low gender decision accuracy. For every plot, PCA was used as the dimensionality reduction method. In (a)-(b), the parameter space is shown for logistic regression. It is noticed that a low number of dimensions and a high noise level lead to a weak gender estimation performance. Further, the noise level is more important for the performance reduction. In (c)-(d), the parameter space is shown for random forest. The best gender

decision performance can be found at the minimum noise level in lower dimensions, while the lowest performance can be found at high noise values. The parameter space plots for euclidean-sensitive noise transformation are shown in (e)-(h). In (e)-(f), these are given for logistic regression. As before, a low number of dimensions and a high noise level lead to a weak gender estimation performance and vice versa. Comparing an uninformed attack scenario (e) with an informed attacker (f), the estimations in the second case have a higher minimum boundary and therefore, leads to better gender estimations. In (g)-(f), the parameter space plots are given for random forest. Again, the greatest reduction in the gender decision performance is given at low dimensions and high noise levels.

Summary - gender estimation: This subsection evaluated the gender decision performance in the context of soft-biometric privacy preservation. It was shown that logistic regression leads to the highly accurate predictions, because the linear classifiers work well with entangled embeddings like the used SphereFace representations. This high accuracy can be even achieved in very low dimensions, since the used dimensionality reductions methods works in an unsupervised manner. A high correlation between the noise level and the loss in the gender decision performance can be observed when introducing noise to the face representations. This lead to a higher privacy level. A significant factor for the performance loss appears to be the prior knowledge about the privacy-mechanism. It is demonstrated that an informed attacker can make significantly more accurate predictions than an uninformed attacker. Finally, the reliability of the gender estimation performance is investigated. In a scenario in which the classifier is set to be 95% sure that a true estimate is correct, increasing the noise level leads to a strong drop in the estimation performance.

Investigate continuous age estimation

Previous works evaluated their privacy-preserving solutions solely on binary attributes. However, continuous attributes might behave different in these solutions. Therefore, in this subsection, we discuss the influence of dimensionality reduction and similarity-sensitive noise on the continuous attribute of age.

In Figure 5.22, the age estimation performance is analysed. In the top row (a)-(c), the results for three different regressors are shown in the context of cosine-sensitive noise. Generally, for increased noise level the MAE is increasing too. Furthermore, it is very conspicuous that the errors for the scenario of an informed attacker is significantly lower than the errors in an uninformed attacking scenario. Errors are generally higher in the cases with dimensionality reduction for the same noise levels than in the cases without dimensionality reduction. The SVM showed the most stable results over the evaluation settings and noise levels.

In the bottom row (d)-(f) of Figure 5.22, the age estimation performance is discussed for the euclidean-sensitive noise. Again, an increased noise-level leads to higher MAE. The MAE in the context of dimensionality reduction showed to become higher than in the context without dimensionality reduction. The noise-free representations ($r = 0$) have the lowest MAE of about 6 (years). Comparing the two noise sampling methods shows that, in general, cosine-sensitive noise leads to higher MAE.

In order to discuss the effect of the noise on the age estimation in more details, Figure 5.23 shows the directed age estimation error distributions for the three regressors in two different attacking scenarios. Negative x-values indicate that the regressor underestimates the age, while positive x-values indicate that it overestimates the ages. In the top row (a)-(c), the scenario of an uninformed attacker (without prior knowledge) is shown, while the bottom row (d)-(e) shows the scenario on attacker with prior knowledge. For ridge regression, introducing similarity-sensitive noise to the data leads to more uniformly distributed curves. Therefore, the prediction errors are high. For random forest, introducing noise leads to distribution shifts to the left and thus, the regressor will suffer on underestimation. The distributions for SVM show a remarkable behaviour. When noise is introduced to the data, the directed error distributions splits in four peaks. This happens because, in this case, the SVM only make predictions in the narrow range of 29 to 36 years, while the other regressors produces outputs in a range of 20 to 50 years. Therefore, the similarity-sensitive noise leads to oversimplification.

To jointly investigate the noise-influence and the dimensionality reduction on the age estimation performance, Figure 5.24 shows the parameter space plots for PCA and ridge regression. Red areas represent a high MAE and blue areas indicate low MAE. In (a) and (b), the age estimation performance is shown for cosine-sensitive noise with and without prior knowledge, while (c) and (d) show the same for euclidean-sensitive noise. In all cases, a high number of dimensions and a high noise level leads to higher MAE. Further, a clear difference between the precision of an informed and an uninformed attacker is observed, since the MAE magnitudes in the uninformed attacking scenario are much higher.

Summary - age evaluation: In this subsection, the age estimation performance was investigated in the context of privacy preservation. Support vector regression showed to produce the most accurate predictions in general. To enhance the privacy, noise transformations were introduced to the face representations, which lead to higher estimation errors and thus, a higher privacy protection level. Especially cosine-sensitive noise showed a significantly higher MAE increase than the euclidean-sensitive noise. Comparing the two attacking scenarios, an informed attacker will produces estimations with less errors. Finally, we investigated the directed estimation error distributions to analyse the effect of the noise on the regressors predictions. For a linear regressor like ridge regression, the

predictions become more uniformly distributed. For ensemble methods like random forest, a distribution shift is observed, which leads to an underestimation trend. Introducing noise to the support vector regressor leads to predictions in very narrow range of 7 years and thus, the noise pushes the regressor to oversimplification.

When is it beneficial to introduce noise?

So far, we have discussed the effect of the proposed similarity-sensitive noise transformations and the dimensionality reduction techniques on recognition, gender estimation and, age estimation performance. In this subsection, we gather this information together and investigate the conditions where the privacy gain and the recognition loss trade-off is beneficial. In section 5.4.2, we introduced the metric *PIC* to measure the gain in privacy versus the loss in recognition performance.

In Figure 5.25, this trade-off is analysed for the gender attribute. The top row (a)-(c) shows the results for cosine-sensitive noise with and without a PCA dimensionality reduction to 128 dimensions under both attacking scenarios. In (a) and (b), the same is demonstrated for the original face representations. The proposed cosine-sensitive noise shows a high benefit (in terms of PIC) over the whole evaluated parameter range. This even holds in the case of an attack who has prior knowledge about the privacy mechanism.

In (c) and (d) of Figure 5.25, the same evaluation was done on PCA reduced representations. Here, the same high PIC values can be observed. However, at $\Theta \approx 0.7$ the trade-off between privacy gain and recognition loss cancel each other and it becomes unfavourable.

The bottom row (e)-(h) of Figure 5.25 show the results for euclidean-sensitive noise. In (e) and (f), there is a benefit when logistic regression or Random Forest are used. However, if SVM is used the PIC is around zero. For the reduced representations (g)-(h), the most noise parameter values are beneficial. One exception is random forest. In this case, for a higher noise level the PIC becomes negative.

In Figure 5.26, the privacy gain identity preservation loss trade-off is shown for the age attribute under two different attacking scenarios. The top row (a)-(d) shows the PIC behaviour for the cosine-sensitive noise. For the original representations (a)-(b), $\Theta \geq 0.7$ is always beneficial, even in the case in which an attack knows about the privacy mechanism. For the reduced representations (c)-(d), positive PIC coefficients can be obtained for $\Theta \geq 0.9$ for all evaluated regressors. The bottom row (e)-(h) shows the PIC behaviour for the euclidean-sensitive noise. Here, the privacy gain recognition loss trade-off is mostly negative and thus, it is not recommended to employ this strategy in a biometric system.

Summary - PIC evaluation: In summary of the PIC evaluation, the trade-off between the privacy gain and the identity preservation loss was investigated for the binary attribute of gender and the continuous attribute of age. It turned out that the PIC results for the

dimensionality reduction methods were very diverse. For some estimators, they proved to be very beneficial, but since the influence of the noise on lower dimensional data is higher, it is unfavourable in general. While the euclidean-sensitive noise showed more positive results for suppressing gender, it turned out to be not useful for continuous attributes like age. However, the proposed cosine-similarity noise showed some promising results. For gender estimation as well as for age estimation, high PICs between 1 and 4 can be observed. Even in the case of an informed attacker, who has prior knowledge about the privacy mechanism, cosine similarity in the range of $\Theta \in [0.7, 1]$ showed always a highly beneficial trade-off between the privacy gain and the recognition loss.

5.4.4. Interim Conclusion

In this work, the proposed similarity-sensitive noise transformations and dimensionality reduction techniques were successfully evaluated for the task of enhancing privacy. Unlike previous work, these solutions can be employed without sensitive user information and further, offer a privacy enhancement for more than one privacy-sensitive attribute, even unknown ones. The similarity-sensitive noise transformations inject geometrical-inspired noise to a face representation and can be controlled in terms of template similarity. This has the advantage that the loss in recognition performance can be restricted with regard to the applied noise level.

Soft-biometric privacy is determined by a trade-off between preserving identity and suppressing the possibility of an unauthorized estimation of private attributes. Therefore, we conducted a comprehensive investigation on a publicly available database. This included an analysis of the recognition performance, as well as the performance for estimating the soft-biometric attributes gender and age.

We evaluated two scenarios of attackers, with and without prior knowledge about the privacy mechanism, and showed that an informed attacker is able to perform significantly better predictions. We found out that higher noise levels lead to lower prediction accuracies. Further, if highly reliable estimations are required, the proposed similarity-sensitive noise leads to a clear drop in estimation performance. In order to better understand how the noise transformations affect different kinds of estimators, the prediction behaviour of three kinds of classifiers and regressors were analysed. For instance, we showed that introducing noise to the face representations leads to oversimplified SVM predictions. Finally, we defined the *PIC* measure, which evaluates the trade-off between the privacy gain and the loss of identity preservation. Evaluating the proposed cosine-sensitive noise transformation leads to high *PICs* and thus, to a clear deployment recommendation. This even holds in scenarios in which the attacker knows about the utilized privacy mechanism. However, strong privacy-enhancements can not be achieved without a strong degradation



of the recognition performance.

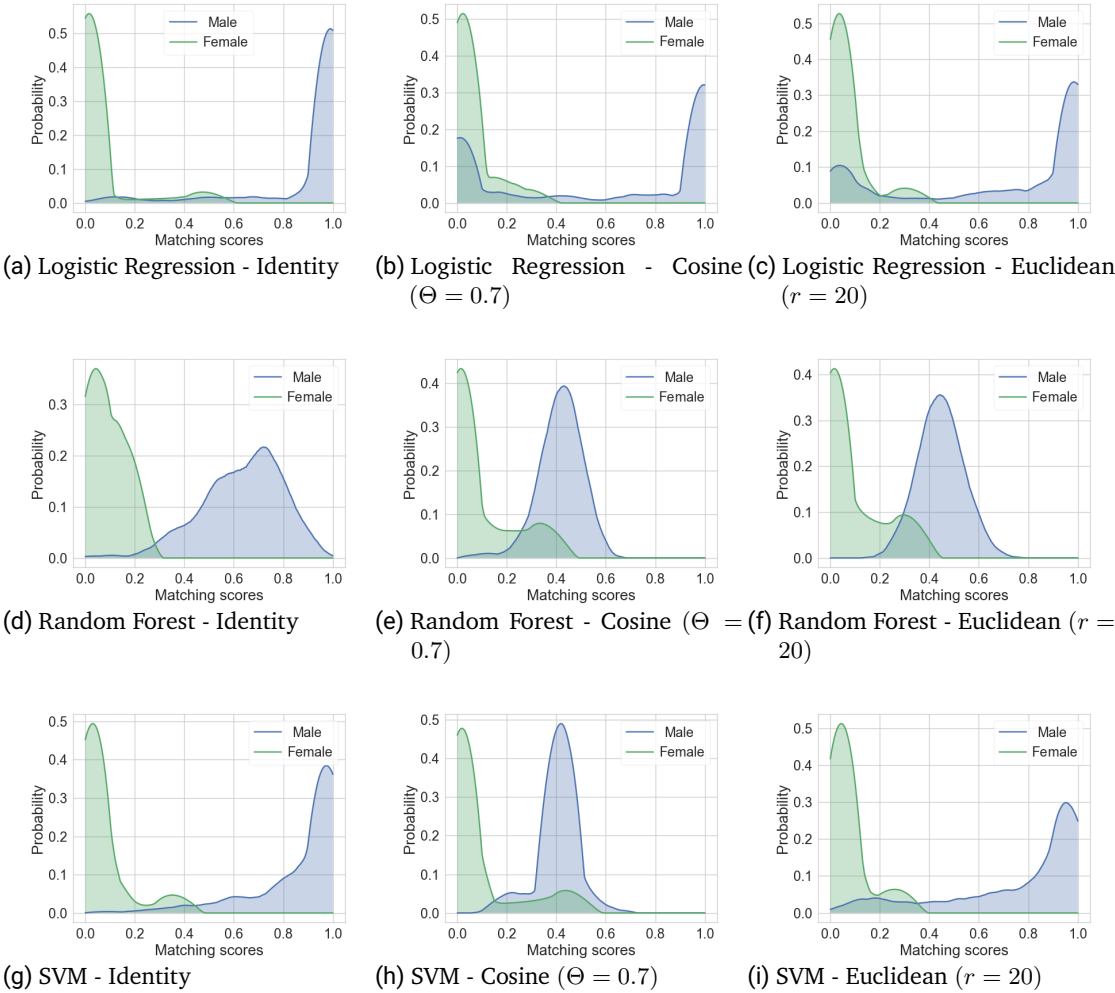


Figure 5.17.: Score distributions for both gender classes. The classifiers were trained on the transformed data, simulating an attacking scenario in which prior knowledge about the privacy mechanism is available.

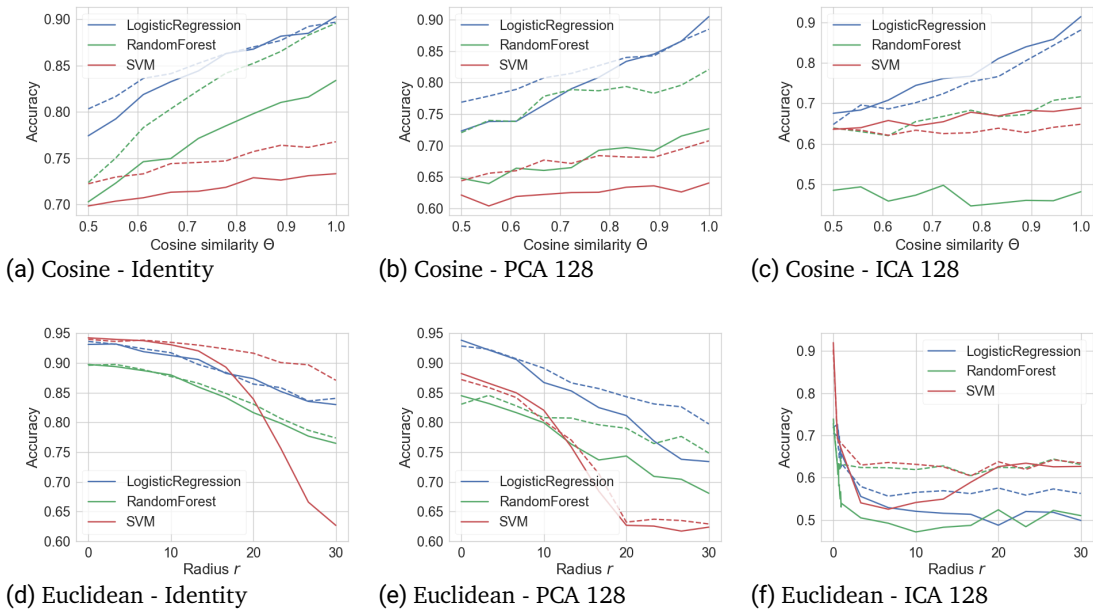


Figure 5.18.: Gender estimation performance on three different representations and two attacking scenarios. The top row (a)-(c) presents the effect of the cosine-related noise, while the bottom row (d)-(f) presents the effect of euclidean-related noise. The performance for uninformed attackers are shown in solid lines, while attackers with prior knowledge about the privacy mechanism are illustrated in dashed lines.

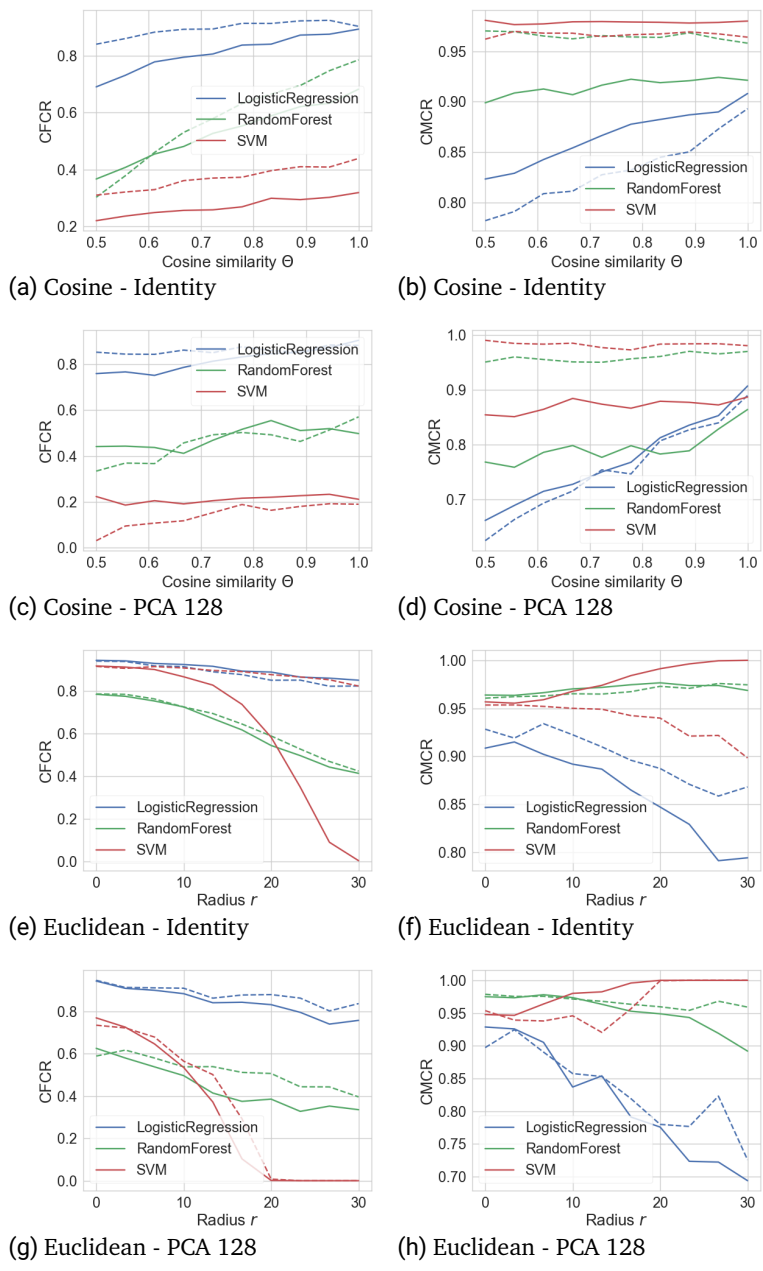


Figure 5.19.: Gender estimation performance in terms of correct female/male classification rate (CFMR/CMCR). Dashed lines indicate an attack with prior knowledge about the privacy mechanism.

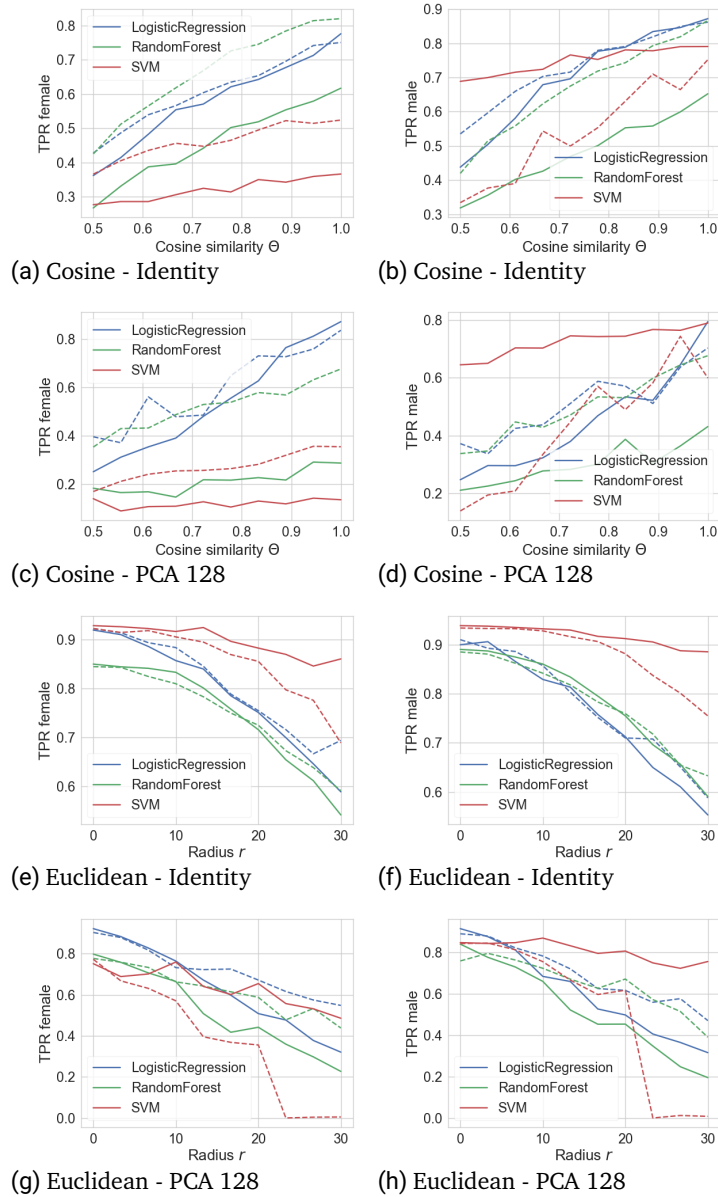


Figure 5.20.: Gender estimation performance for the gender classes female and male in a reliability scenario. Reported are the true positive rate (TPR) at a fixed false positive rate (FPR) of 5%. The dashed lines indicate the reliable classification performance of an informed attacker.

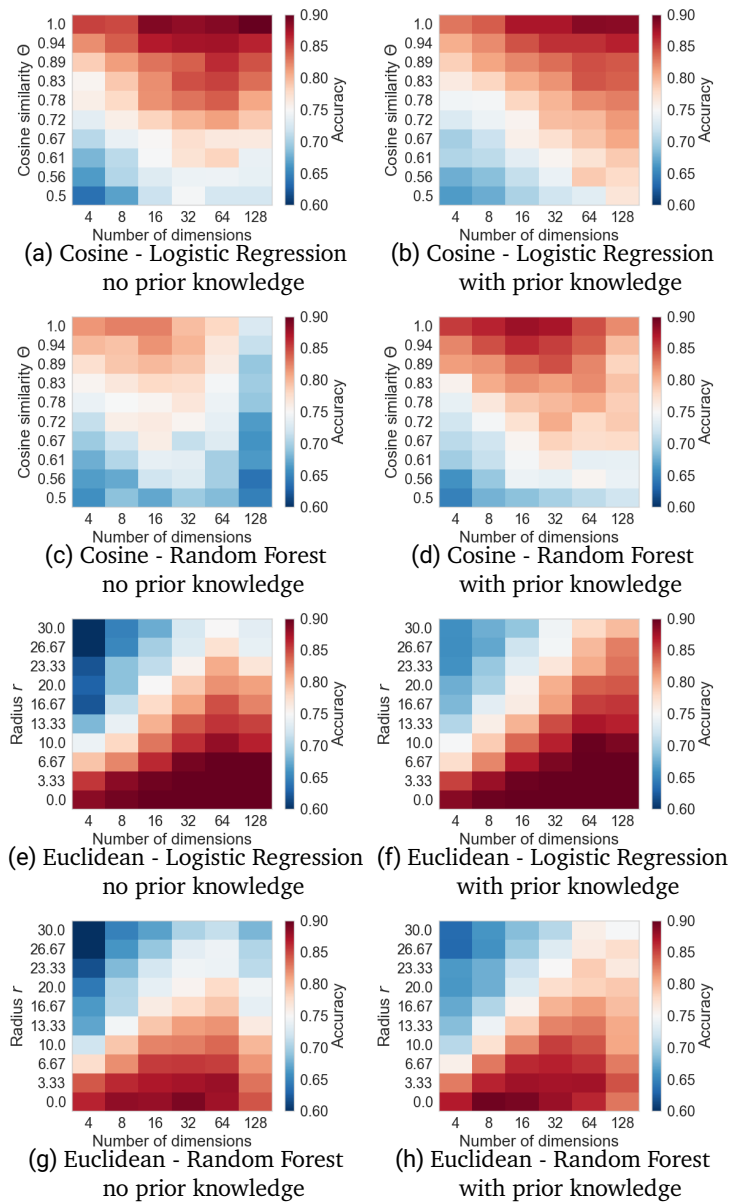


Figure 5.21.: Joint investigation of PCA dimensionality reduction and similarity-sensitive noise transformation for two different classifier and two attacking scenarios.

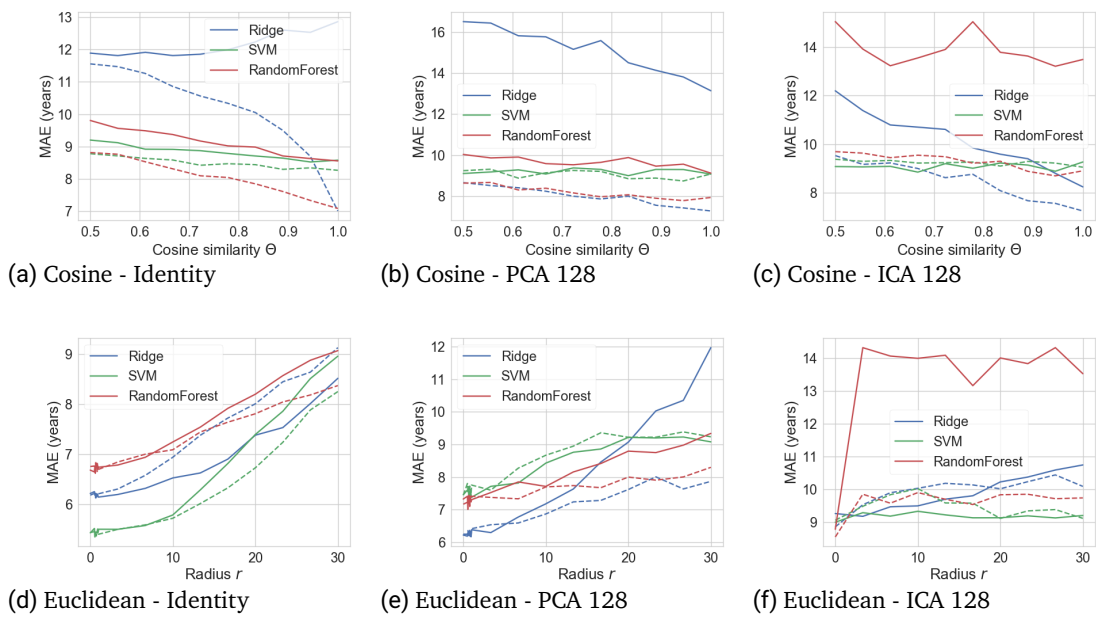


Figure 5.22.: Analysis of the age estimation performance in the context of dimensional-reduction and similarity-sensitive noise. The dashed lines indicate the performance of an attacker which exploits the knowledge about the systems privacy mechanism.

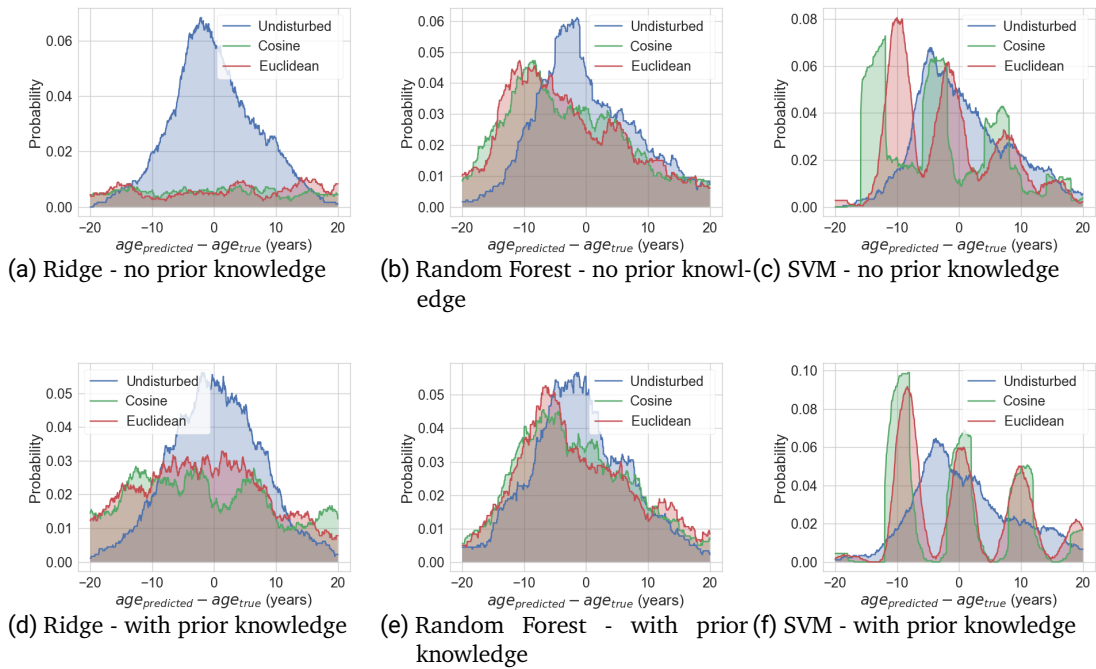


Figure 5.23.: Directed age estimation error distributions for three different estimators in the context of similarity-sensitive noise (Cosine: $\Theta = 0.7$, Euclidean: $r = 20$) and without (Undistributed). The top row (a)-(c) refers to an attacker without prior knowledge about the privacy mechanism, while the bottom row (d)-(f) refers to an attacking scenario in which an attacker is exploiting this information.

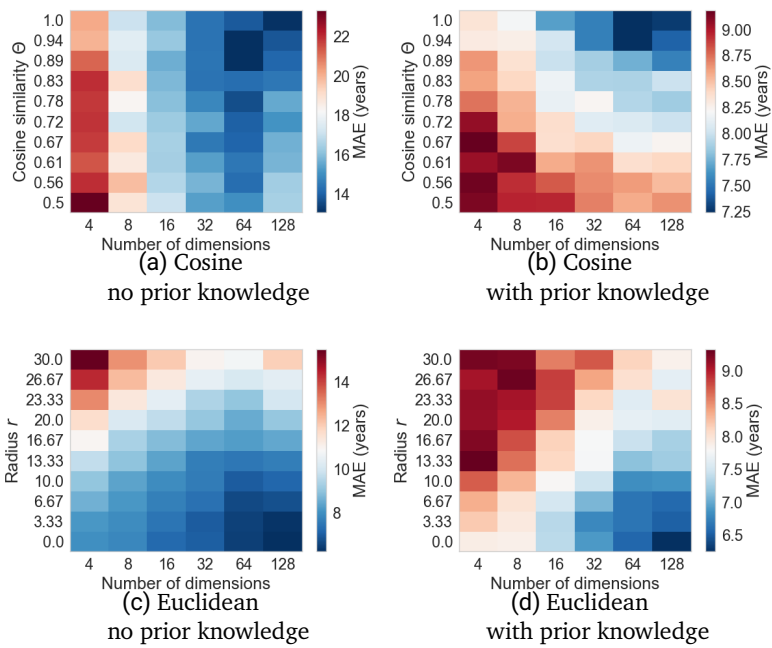


Figure 5.24.: Joint investigation of dimensionality reduction with PCA and similarity-sensitive noise transformations for the task of age estimation. In (a) and (c), uninformed attacker do the ridge regression, while in (b) and (d), the attacker knows about the privacy mechanism.

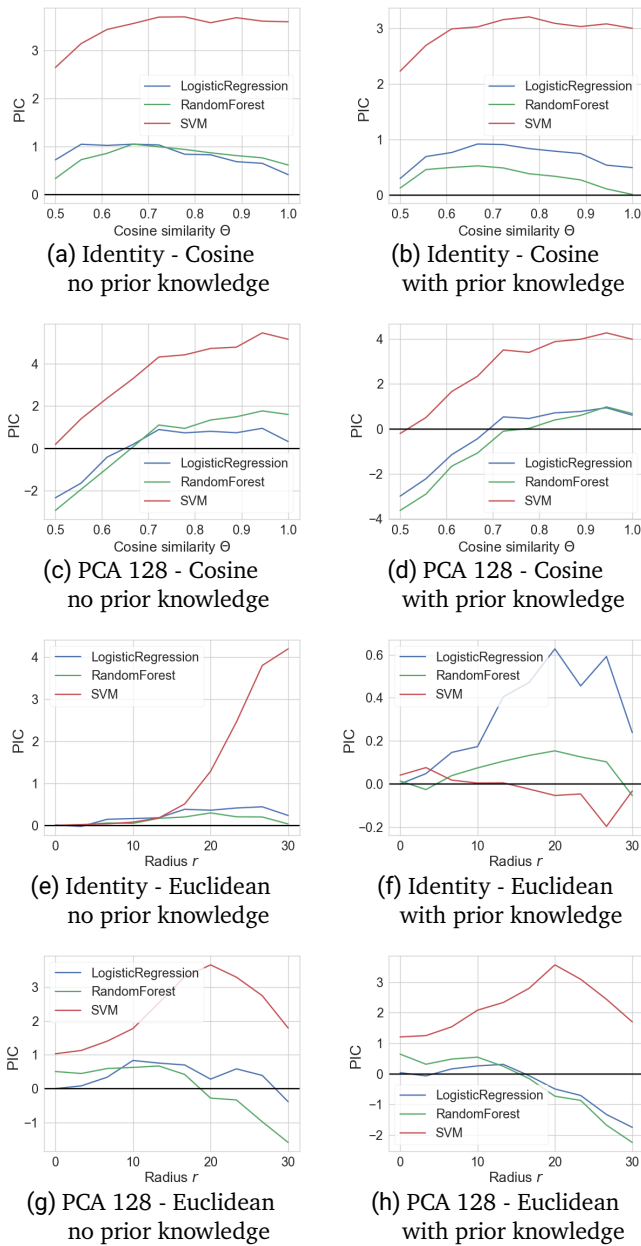


Figure 5.25.: PIC curves for the binary attribute of gender. Cosine-sensitive (a)-(d) and euclidean-sensitive (e)-(h) noise transformations are considered in different contexts of attacking scenarios and representations.

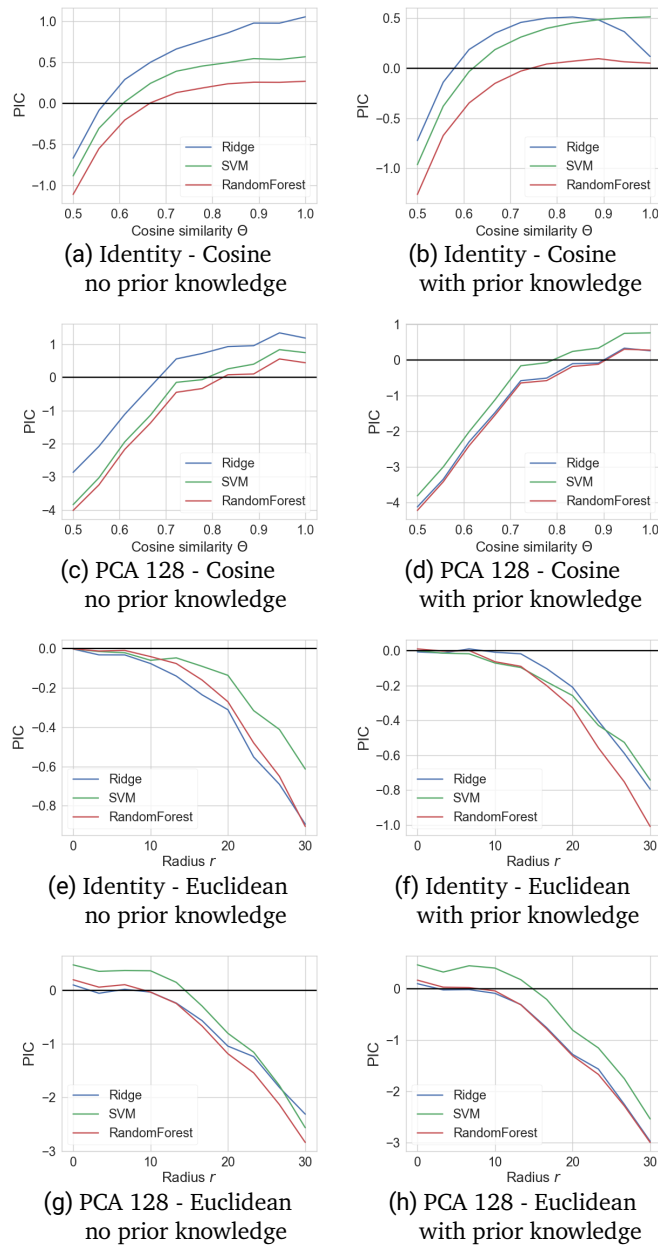


Figure 5.26.: PIC value curves for the continuous attribute of age. Cosine-sensitive (a)-(d) and euclidean-sensitive (e)-(h) noise transformations are considered in different contexts of attacking scenarios and representations.

5.5. Negative Face Recognition

Previous works proposed privacy-enhancing solutions that (a) are limited to the suppression of single pre-defined attributes [Mir+18; MR17; OR14] or (b) provide a more comprehensive privacy-enhancement [Ter+19b] as shown in Section 5.4. While solutions from (a) are vulnerable to unconsidered function creep attacks, approach from (b) show either a weak suppression or recognition performance.

In this section, we propose negative face recognition (NFR) [Ter+20c], a novel unsupervised face recognition approach that performs comparisons of face templates in a complementary (negative) fashion. While ordinary positive templates describe individuals how they actually are, negative templates stores only random complementary information about the individual. This suppresses privacy-sensitive information in the template and thus, prevents function creep attackers from easily extracting this information. In order to forecast and guarantee a certain recognition performance, we provide a theoretical reasoning of our solution and further demonstrate its correctness empirically.

Soft-biometric privacy is challenged by maintaining a high recognition performance while achieving a high suppression performance for privacy-sensitive attributes. Therefore, we analyse both aspects on two publicly available databases under controlled and uncontrolled circumstances. The evaluation of the attribute suppression performance is done on three soft-biometric attributes: gender, age, and race. Unlike most of the previous works, we design our experiments in the context of a function creep attacker who knows and adapts to the used privacy mechanism.

The experiments show that our proposed approach is able to reach 2-4 times higher suppression rates than previous works under different attack mechanisms and attributes while maintaining significantly higher recognition performances. In the uncontrolled scenario, our solution fully retains the recognition performance while reaching suppression rates of up to 36%.

5.5.1. Methodology

Enhancing the soft-biometric privacy aims at preventing function creep attackers from reliably estimating privacy-risk characteristics. This problem is further challenged by simultaneously maintaining a high recognition ability. To solve these issues, we propose negative face recognition. While in usual face recognition systems, the used templates describe the properties of an individual, our negative templates only contain complementary information and thus, describe properties that a person does not have. We store only negative (reference) templates in a database and comparing it with positive (probe) templates by calculating their dissimilarity. Due to the complementary nature of the

compared templates, a high dissimilarity indicates that the templates belong to the same subjects and vice versa. Since the stored negative templates only contain some random complementary information, it prevents function creep attackers from successfully deducing privacy-sensitive information. Further, it allows a more generalized soft-biometric privacy-protection that is, unlike previous works, not limited to the suppression of a pre-defined characteristic. It is further a promising candidate for template protection, as shown in a similar approach [Zha+18] for iris. It provides noninvertability, revocability, and nonlinkability, which are the key properties of template protection. However, the template protection applicability is out of the scope of this work.

Since the idea of this work is to store only random complementary information of an individual in the database (in form of negative templates), the next section describes the enrolment process. This is followed by a section of the adapted verification process because the template comparison within our approach is dealing with complementary template versions.

Enrolment phase

In the enrolment phase, given a face image I , the corresponding face embedding x is extracted from I . Then, this embedding is transformed in the negative domain resulting in a negative template t_- , which is stored in the database. The generation of a negative template t_- from a face embedding x is done in three steps: First, the face embedding x is enlarged to get a higher-dimensional version v . Second, v is discretized to create a positive template t_+ , and third, a negative template t_- is generated from its positive complement by replacing each feature entry with a random value that does not match the original entry.

Embedding enlargement In the first step, the given face embedding x is transformed into a higher-dimensional space while maintaining its recognition ability. Therefore, a face recognition model, called enlargement-network, is trained to take the used face embedding x as an input and outputs the higher-dimensional face embedding v of size L . The network and its training are described in Section 5.5.2. The enlargement step is necessary because (a) the genuine/imposter decision is based on the dissimilarity between a positive and a negative template and (b) the negative template generation is based on a randomized process. If a positive and a negative template belong to different subjects, but are of low dimensionality, there is a higher chance that the negative template is very dissimilar from the positive one. For increased dimensionalities, the positive and the negative templates share more similar feature entries and thus, increases the similarity. In terms of positive-negative template comparison, a high similarity indicates an imposter

comparison. Consequently, high dimensional templates are needed for negative face recognition to reduce the recognition errors from the randomization process.

Embedding discretization In the second step, the enlarged embedding v is feature-wise discretized into k bins. The k bins were chosen beforehand on the enlarged training data using a quantile strategy that divides each feature range into k bins such that every bin contains an approximately equal number of samples. Following this binning ranges, each feature entry of v is replaced with the value $l \in \mathcal{K} = \{1 \dots, k\}$ of its corresponding bin. This results in a discretized positive template $t_+ \in \mathcal{K}^L$. Discrete features are required for the feature-wise computation of complementary feature sets that is needed in the next step of the negative template generation.

Negative template generation The third step replaces each feature entry of the positive template t_+ with a random value from the complementary feature set. This results in a negative template that contains facial properties that the person does not possess and thus, it is hard to estimate the soft-biometrics of that individual. Given a positive template $t_+ \in \mathcal{K}^L$, a negatively associated template t_- is generated feature-wise. This is done by replacing each feature entry of t_+ with a randomly chosen value from \mathcal{K} that does not match the original entry. To be precise, for each component i the negative representation $t_-^{(i)} \in \mathcal{K} \setminus \{t_+^{(i)}\}$ is given by a randomly chosen value from the complement set $\mathcal{K} \setminus \{t_+^{(i)}\}$. This results in the negative template t_- . In the last step of the enrolment, the negative template t_- is stored in the database associated with the enrolled identity.

Verification phase

In the verification phase, an individual claims an identity and the negative (reference) template of the claimed identity is loaded from the database. This negative template is then compared with the positive (probe) template from the captured individual. In order to verify a person's identity with our negative face recognition approach, (1) the positive probe template and the negative reference template have to be allocated and (2) the templates are compared against each other to determine a comparison score. This comparison score is used to make a verification decision.

Template allocation Given an input face image from an individual, first, its embeddings (Section 5.5.1) have to be computed and second, discretized (Section 5.5.1) to obtain the positive (probe) template t_+ . The negative (reference) template t_- is loaded from the database. The positive and the negative template can then be compared.

Positive-negative template comparison In order to compute a comparison score between the positive and the negative template t_+ and t_- , we utilize a normalized hamming-like distance

$$NHD(t_+, t_-) = 1 - \frac{1}{|t_+|} \sum_{i=1}^{|t_+|} \delta(t_+^{(i)}, t_-^{(i)}). \quad (5.16)$$

The delta function $\delta(a, b)$ returns 1 if a equals b and 0 otherwise. The size of the templates is defined by $|t_+| = L$. The *NHD* measures the dissimilarity of t_+ and t_- and, due to the complementing nature of positive and negative templates, it can be directly utilized as a comparison score. Since the positive template defines properties of the corresponding individual, while the negative template describes properties that the individual does not contain, a larger (*NHD*) distance represents a higher probability that the templates originate from the same subject and vice versa.

In the case that a negative template t_-^A was generated from the positive template t_+^A . By the construction of negative templates, t_+^A and t_-^A have the maximum dissimilarity. The bigger the difference between the positive templates of the probe and the reference, the lower is the dissimilarity. This is because only feature-level errors in the positive domain can produce features in the negative domain that collide with the corresponding features in the positive domain. As long as the positive templates from probe and ref are reasonably similar, the positive-negative template comparison is highly dissimilar, indicating a genuine pair. This explains the recognition performance of the negative templates.

About the gain of privacy

Our negative face recognition approach makes a face recognition system less vulnerable to function creep attacks in cases where attackers get access to the stored data. Since only negative templates are stored, the information about an individual is limited to the deeply-encoded description of complementary nature. This enables our solution to offer a more comprehensive privacy-protection that is not limited to single pre-defined attributes.

However, in the case of function creep attackers getting access to multiple negative templates that were created from the same positive templates, a statistical analysis might enable a reconstruction of the positive template. Consequently, in this special case, a reconstruction and thus, a reliable privacy-sensitive attribute estimation might be possible. To prevent this attack strategy, we recommend the use of differently-trained enlargement-networks for different databases. This prevents the generation of negative templates from the same positive template and thus, circumvent this statistical analysis-based attack strategy even in the case of attackers getting access to multiple negative face databases.

Theoretical foundation

Since our approach is based on a randomized process in the template generation, we provide a statistical reasoning for the negative-positive comparison performance. Given a theoretical or empirical score distribution of genuine and imposter scores, this allows to predict the negative face recognition performance including the probabilities of falsely rejected and falsely accepted subjects. Consequently, optimal hyperparameters can be chosen, as well as large-scale experiments can be simulated, without the computational costs for sophisticated experiments.

Given two positive templates $t_+^A, t_+^B \in \mathcal{K}^L$ with a distance of,

$$HD(t_+^A, t_+^B) = \sum_{i=1}^{|t_+|} \delta(t_+^{(i)}, t_+^{(i)}) = D, \quad (5.17)$$

then, the probability of this distance in the negative domain,

$$HD(t_-^A, t_-^B) = D' \quad \text{with} \quad D' \in [L - D, L] \quad (5.18)$$

follows a Bernoulli distribution and is given by,

$$Pr [D'|D] = \binom{D}{\mu(D')} \left(\frac{k-2}{k-1}\right)^{\mu(D')} \left(\frac{1}{k-1}\right)^{D-\mu(D')}. \quad (5.19)$$

The Bernoulli distribution can be assumed, since only entries of equal values contribute to the distance and the state of this entries is given by a fixed probability. The number of bins in this equation is described by $k = |\mathcal{K}|$ and $\mu(D')$ is given by

$$\mu(D') = D' - (L - D), \quad (5.20)$$

the number of entries that have to be flipped to the same entry in order to achieve the determined distance of $D' - D$. Based on our negative template generation principle (Section 5.5.1), colliding bin labels t_+^A and t_+^B in the positive domain, will not collide in the negative domain and thus, will contribute to the distance. In order to achieve a distance of $HD(t_-^A, t_-^B) = D'$ in the negative domain, $\mu(D')$ bin labels have to be flipped such that they will contribute to the distance calculation. The probability for such a single flip is given by $\frac{k-2}{k-1}$ and thus, the collision probability is given by $\frac{1}{k-1}$.

Given two positive templates t_+^A, t_+^B with a distance of $HD(t_+^A, t_+^B) = D$, then Equation 5.19 gives the probability for the two templates to have a hamming distance of D' if one of the templates is in the negative domain.

5.5.2. Experimental Setup

Databases

In order to evaluate and compare our approach under both controlled and uncontrolled conditions, we conducted experiments on the public available ColorFeret [Phi+00] and Adience [EEH14] databases. ColorFeret [Phi+00] consists of 14,126 images from 1,199 different individuals with different poses under controlled conditions. A variety of face poses, facial expressions, and lighting conditions are included in the dataset. For each face in the database, information about the person’s gender, age, and race (black, white, asian, and others) are given. We categorized the age labels into four classes (20-29, 30-39, 40-49, and 50+ years) to create age-balanced dataset. This allows an effective training of the function creep estimators. The Adience dataset [EEH14] consists of 26,580 images from over 2,284 different subjects under uncontrolled imaging conditions. Adience contains additional information about the individual’s gender and age. The age labels come from human investigations and are divided into eight age classes (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+). We choose these databases because they were captured under controlled and uncontrolled conditions and provide information of soft-biometric attributes and information about the identities. The soft-biometric information is only used during the evaluation. Using the soft-biometric and the identity information allows to deeply investigate the privacy-enhancing technologies by analysing the recognition performance, as well as the suppression performance of privacy-sensitive attributes.

Evaluation metrics

Enhancing soft-biometric privacy describes a trade-off between the desired degradation of the attribute estimation performance by function creep attackers and the desired preservation of the recognition ability. In this work, we report our verification performances in terms of false non-match rate (FNMR) at fixed false match rates (FMR). We also report the equal error rate (EER), which equals the FMR at the threshold where $FMR = 1 - FNMR$. This acts as a single-value indicator of the verification performance. In order to evaluate the attribute suppression performance, we report our results in terms of attribute classification accuracy on balanced test labels and in terms of attribute suppression rates. The suppression rate

$$sr = \frac{acc_{org} - acc_{mod}}{acc_{org}} \quad (5.21)$$

describes the reduction of the attribute-prediction accuracy of the unmodified (original) templates acc_{org} in comparison to the accuracy of the templates acc_{mod} with privacy-

enhancement. A higher suppression rate indicates an advanced privacy-improvement.

Basic face recognition model

The proposed negative face recognition approach builds on arbitrary face embeddings. In this work, we utilize the widely used FaceNet model² [SKP15], which was pretrained on MS-Celeb-1M [Guo+16]. In order to extract an embedding of a face image, the image is aligned, scaled, and cropped as described in [KS14] and then passed into the model. The output of this network is a 128-dimensional embedding. The comparison of two such embeddings is performed using cosine-similarity.

Enlargement-network training

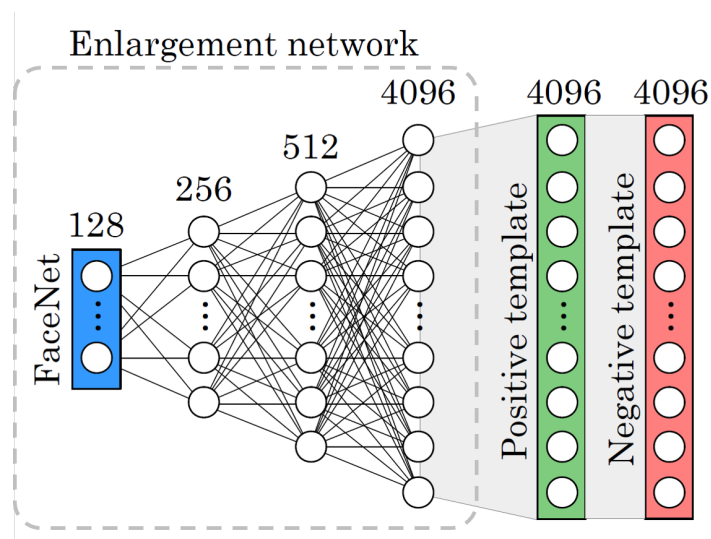


Figure 5.27.: Enlargement-network and positive/negative template generation: the enlargement-network structure is shown without softmax layer. Given a face embedding (FaceNet) a larger representation of this embedding is computed. A positive template is created by discretisation and replacing each feature entry with an item its complementary set, a negative template is generated.

²<https://github.com/davidsandberg/facenet>

Our approach requires high-dimensional face embeddings, in order to create discriminative negative templates. As described in Section 5.5.1, an enlargement-network is used to expand the low-dimensional face embeddings to size L . This network has an input size of 128, corresponding to the used face embeddings, and an output size of $L = 4096$. It consists of three layers with 256, 512, and 4096 neurons, and is shown in Figure 5.27. The first layers are activated by a ReLU function, while the forth layer holds a tanh activation, such that the output-features are within the range of $[-1, 1]$. To train the network, a softmax layer is added to classify the identities in the test set with a binary cross-entropy loss. The training is done with an AdaDelta optimizer (learning rate $lr = 0.5$) over 50 epochs of training. Dropout ($p = 0.5$) [Sri+14] and Batchnormalization [IS15] is applied on every layer. After the training the softmax layer is removed.

Function creep attacks

For the evaluation of the attribute suppression, we simulate the critical scenario of a function creep attacker that adapts to the system’s privacy mechanism. The adaptation step is done by training (function creep) classifiers on the transformed (normalized and scaled) templates to predict the privacy-sensitive attributes. These classifiers include random forest (RF), support vector machines (SVM), k-nearest neighbors (kNN), and logistic regression (LogReg). The hyperparameters of these classifiers are fine-tuned with Bayesian optimization.

Baseline approaches

In Section 5.2, we mentioned that many privacy-enhancing methods were proposed that manipulate the face images itself using supervised approaches. However, most biometric systems store face templates instead of images [Dey+14; SRB16] and furthermore, supervised approaches are vulnerable to attacks on attributes that were unconsidered during training. Therefore we proposed an unsupervised privacy-enhancing approach working on template-level and compare it against two state-of-the-art solutions with the same working principles. In this work, we use similarity-sensitive noise transformations [Ter+19b] as baselines. More precisely, we compare our proposed NFR approach against cosine-sensitive noise (CSN) and euclidean-sensitive noise (ESN).

We calibrate the hyperparameters of these baselines in such a way that they reach similar verification EER performances. By doing so it is possible to fairly compare these methods in terms of suppression rates. For all experiment scenarios, subject-disjoint 5-fold cross-validation is utilized to use all the data available for independent testing and training. Therefore, we divide the database in five folds of approximately equal sizes such that the

identities of one fold do not appear in other folds. The validation is performed in five rounds. In each round, one fold is used for testing while the others are used from training. The performance over all folds is reported as the average performance and its standard deviation.

Investigations

The investigations of this work are divided in five parts:

1. We show the need for a privacy-enhancing technology by demonstrating that there is a significant leakage of privacy-sensitive information from face templates on both databases.
2. We analyse the face verification performance of our privacy-enhancing solution to check to which degree the recognition ability is maintained and compare it with previous works.
3. We investigate the attribute suppression performance of our solution and the baselines in the critical scenario of a function creep attacker that adapts to the system's privacy mechanism. This evaluates the soft-biometric privacy-protection.
4. We analyse the parameter space of our solution to provide a deeper understanding of our solution.
5. Lastly, we provide an empirical validation of the theoretical reasoning and validate its correctness.

5.5.3. Results

Analysis of the function creep performance

Table 5.2 shows the attribute prediction performance of three privacy-sensitive attributes in a scenario without privacy-preservation. The performance of four function creep classifiers is shown under controlled (ColorFeret) and uncontrolled (Adience) circumstances using the original and positive embeddings. Especially gender and race can be determined with very high accuracies. This holds for the original embeddings as well as the (high dimensional and discrete) positive embeddings. The table demonstrates that there is a significant information leakage of privacy-sensitive information from face templates and thus, a great need for privacy-enhancing technologies.

Face verification performance

In Table 5.3 and 5.4, the recognition performance of the baseline approach is shown in comparison to state-of-the-art [Ter+19b] and our approach. In order to make a fair comparison of the attribute suppression analysis, the hyperparameters of both unsupervised state-of-the-art approaches, CSN and ESN, are calibrated such that it matches the EER of our approach for $k = 3$ and $k = 4$ bins. In Table 5.3, the recognition performance is shown for the ColorFeret database. While the EER of templates without privacy-enhancement is about 2%, our approach with $k = 3$ ($k = 4$) bins leads to an EER around 3% (4%). Even if CSN and ESN are calibrated to have a comparable EER, their FNMR for low FMR is significantly higher than our approach. In Table 5.4, the recognition performance is shown for the Adience database. It is observed that the recognition performance for our approaches ($k = 3$ and $k = 4$) is very close to the original performance, while the CSN and ESN show a strongly degraded performance. While CSN and ESN are based on noise injections that leads to a partial identity loss, our approach is based on a complementary representations, which keeps the identity information, but transforms it irreversibly.

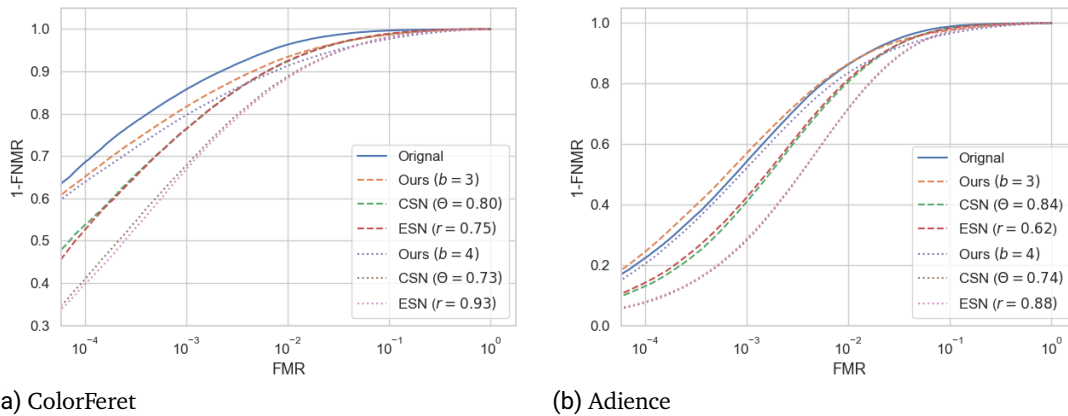


Figure 5.28.: Face recognition performance comparing the performance of the original templates, our approach and related work. Our solution is able to maintain the verification performance to a higher degree than previous works.

To get a more detailed look in the recognition performances over a wider range of decision thresholds, Figure 5.28 shows ROC curves on both datasets. In Figure 5.28a, the performance is shown under controlled face image capture conditions, while in Figure 5.28b the same is shown under uncontrolled conditions. In both cases, it can be observed

that recognition performance is very close to the performance of the original representations, while the CSN and ESN shows a strongly degraded performance. Especially under uncontrolled conditions (Figure 5.28b) the performance even surpasses the performance of the original representations by a small amount due to its error correction ability. This demonstrates, in contrast to previous work, that our solution is can maintain identity information to a large degree.

Privacy-sensitive attribute suppression

In order to compare the soft-biometrics privacy-enhancement, Table 5.5 shows the suppression rates for four classifiers on three privacy-sensitive attributes. The attribute suppression performances of our approach are shown and compared with state-of-the-art approaches (CSN, ESB) [Ter+19b] calibrated to the same verification EER. In [Ter+19b], CSN showed significantly better performance than ESN, especially in suppressing attribute prediction performance for SVM and LogReg. However, CSN transforms each feature vector to a random length $r \in [1, 100]$, which makes it hard to handle for classifiers such as SVM and LogReg. This is not the case in our experiments since we simulated a committed function creep attacker that does not only train on transformed data but also rescales the feature vectors to unit-length. This prevents classifiers, such as SVM and LogReg, from unstable estimations. On both databases, our solution achieves relatively high suppression rates on all classifiers and all attributes. Generally, our privacy-enhancement approach leads to 2-4 times higher suppression rates compared to previous work under different attack mechanisms and attributes.

Investigation of the parameter space

In the following, the parameter space is analysed to increase the understanding of our solutions behaviour. More precisely, the two parameters of our solution, the template size L and the number of bins k , are varied and for every parameter combination the face verification performance (in terms of EER) and the attribute prediction performance from different function creep estimators are shown. Figure 5.29 and 5.30 show the results for $k = 3, 4$ on ColorFeret. Figure 5.31 and 5.32 show the same on Adience. In these Figures the number of bins k and the embedding sizes L are analysed in the ranges of $k = [3, 4]$ and $L = [64, 4096]$. All these scenarios show that a bigger embedding size L leads to a lower face verification error. This observation agrees with the nature of positive-negative template comparisons. Higher dimensional templates reduce the effect of random collisions for the positive-negative template comparison. In lower dimensions, a random collision for an imposter comparison has a high impact on the resulting comparison score.

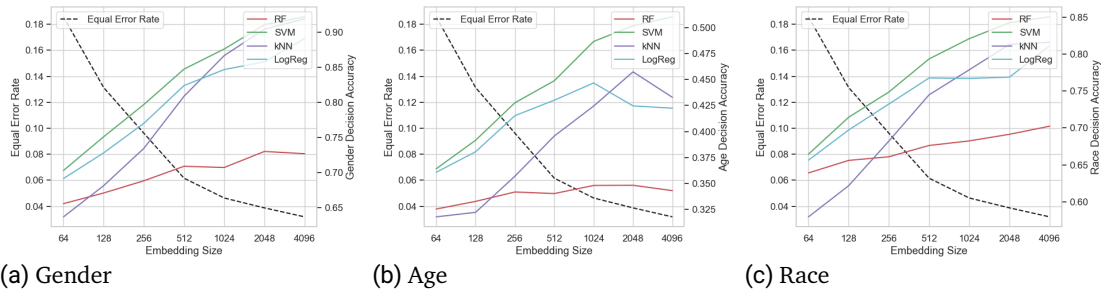


Figure 5.29.: On ColorFeret the face verification EER and the attribute estimation performances of function creep estimators are shown for different embedding sizes and a fixed bin size of $k = 3$. The estimation performance is analysed for the attributes of gender, age, and race.

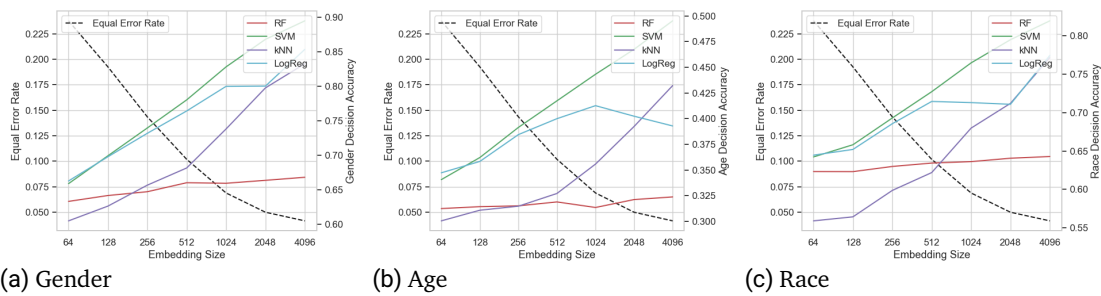


Figure 5.30.: On ColorFeret the face verification EER and the attribute estimation performances of function creep estimators are shown for different embedding sizes and a fixed bin size of $k = 4$. The estimation performance is analysed for the attributes of gender, age, and race.

Towards the bin sizes k , it can be observed that $k = 3$ has a lower face verification error than $k = 4$, but also higher prediction accuracies from all function creep estimators. Higher k leads to more variabilities, which affects verification as well as the estimation of privacy-sensitive attributes. These observations hold for both datasets and all function creep estimators. Consequently, parameter k and L have to be chosen to accomplish the desired trade-off between attribute suppression and verification performance.

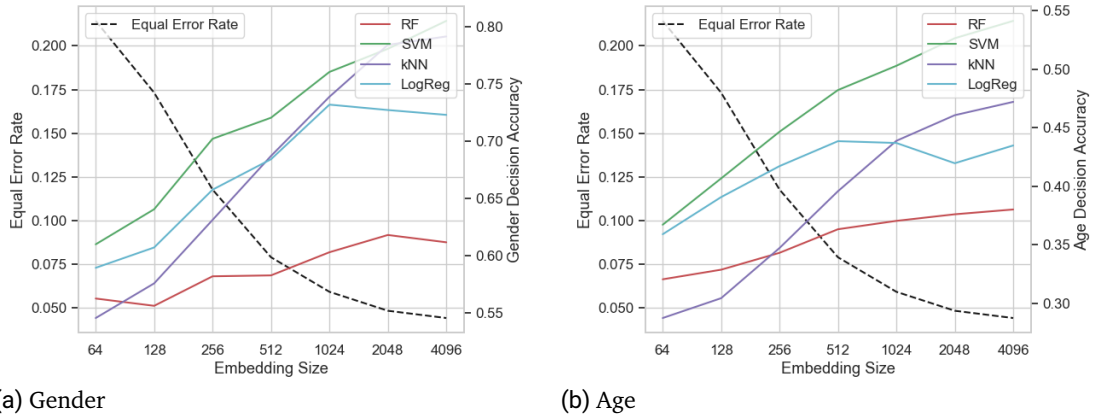


Figure 5.31.: On Adience the face verification EER and the attribute estimation performances of function creep estimators are shown for different embedding sizes and a fixed bin size of $k = 3$. The estimation performance is analysed for the attributes of gender and age.

Theoretical Reasoning Analysis

In Section 5.5.1, a theoretical reasoning for our negative face recognition approach was developed. Here, we want to prove its correctness by empirically predicting the score distributions of our approach and comparing it with the achieved scores distributions. For each comparison score in the positive domain, Equation 5.19 is used to calculate the most probable score in the negative-positive domain. Repeating this process with every score in the distribution results in the score distributions in Figure 5.33. This figure shows the distributions of the genuine and imposter scores of our proposed approach with $k = 3$, as well as its theoretically predicted distribution. It can be seen that on both databases, the predicted distributions accurately correspond to the empirical score distributions. This validates our theoretical considerations from Section 5.5.1.

5.5.4. Interim Conclusion

In this section, we successfully proposed negative face recognition, a privacy-enhancing face recognition solution operating on the template-level. It prevents function creep attackers from successfully predicting privacy-sensitive information from stored face templates. Our novel solution is based on the comparison of positive probe templates with

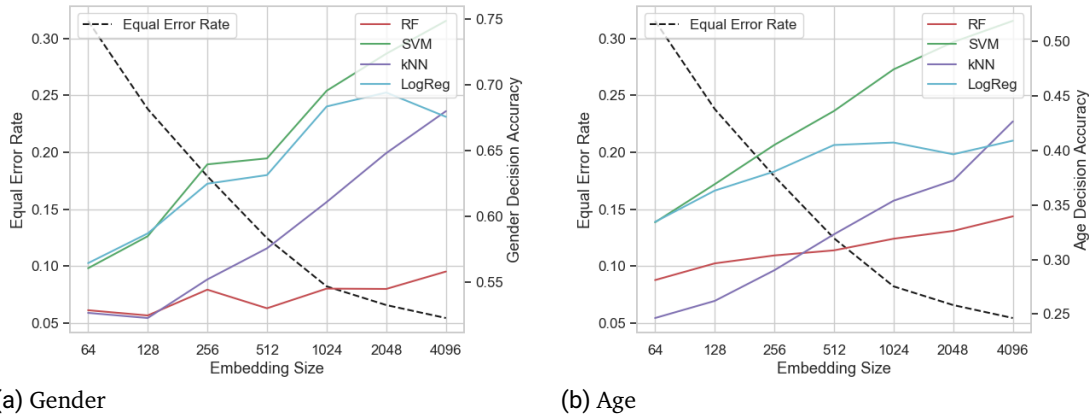


Figure 5.32.: On Adience the face verification EER and the attribute estimation performances of function creep estimators are shown for different embedding sizes and a fixed bin size of $k = 4$. The estimation performance is analysed for the attributes gender and age.

negative reference templates. While positive templates contain the facial properties of an individual, negative templates contain random complementary information, i.e. properties that the face does not have. Since only negative templates are stored in the database, a reliable function creep estimation of privacy-sensitive information is prevented. To guarantee a certain recognition performance, we further provided a theoretical foundation of our solution and proved its correctness empirically. The experiments were conducted on two publicly available databases and on three privacy-sensitive attributes. In the experiments, we simulated function creep attackers that know about the system’s privacy mechanism and adapt their attacks based on it. The experiments demonstrated the effectiveness of our approach under both, controlled and uncontrolled image capturing conditions. Our proposed unsupervised solution significantly outperforms comparable approaches from previous work, while maintaining a significantly higher recognition performance. In the uncontrolled scenario, negative face recognition fully retains the recognition performance while achieving suppression rates of up to 36%. Our solution is characterized by the fact that it prevents the accumulation of privacy-sensitive information during the training and offers more comprehensive privacy-protection. Unlike previous works, negative face recognition is not limited to the suppression of single attributes.

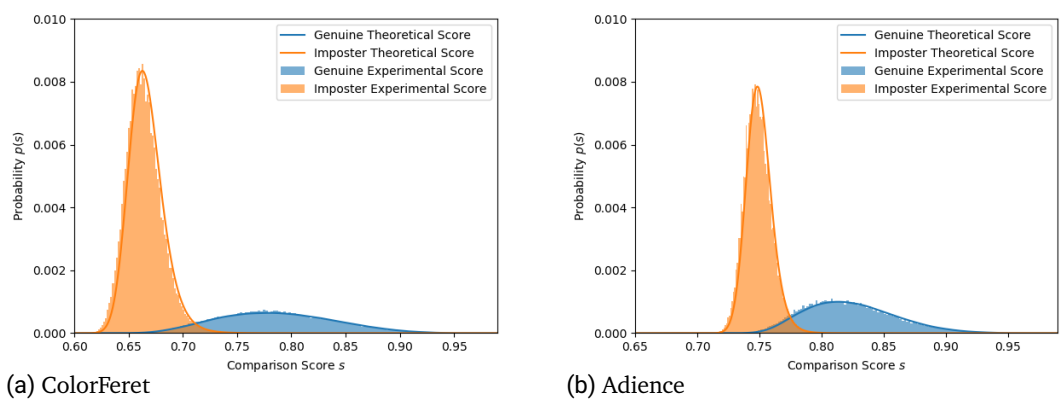


Figure 5.33.: Validation of the theoretical reasoning: score distribution of the empirical data versus the theoretical predictions. The distributions show the genuine and imposter scores for $k = 3$ bins. The theoretical score predictions are done with Equation 5.19 on the positive score distributions. The theoretical predictions accurately match the experimental scores.

Table 5.2.: Attribute prediction performance on original and positive templates (without privacy-enhancement). The prediction accuracies of four function creep estimators are shown on two databases. A function creep attacker would be able to predict the soft-biometric attributes with high accuracies demonstrating the need for privacy-enhancement.

Dataset	Attribute	RF	SVM	kNN	LogReg	
Original	ColorFerret	Gender	93.37% \pm 1.22%	96.61% \pm 1.02%	97.30% \pm 0.39%	95.74% \pm 1.14%
		Age	49.17% \pm 2.40%	57.40% \pm 2.63%	47.71% \pm 2.06%	57.12% \pm 2.91%
		Race	82.21% \pm 1.07%	88.73% \pm 1.17%	85.10% \pm 2.58%	88.03% \pm 1.46%
Adience	Gender	82.78% \pm 1.79%	84.16% \pm 2.34%	84.91% \pm 2.45	82.43% \pm 2.78%	
	Age	53.27% \pm 4.08%	60.36% \pm 4.10%	51.31% \pm 3.08%	58.26% \pm 4.93%	
Positive (b=3)	Colorferret	Gender	90.05% \pm 2.08%	96.92% \pm 0.86%	94.89% \pm 0.69%	93.45% \pm 0.86%
		Age	44.47% \pm 2.78%	53.51% \pm 1.69%	45.99% \pm 1.69%	46.95% \pm 1.19%
		Race	79.73% \pm 0.84%	87.52% \pm 1.59%	84.09% \pm 2.33%	84.46% \pm 1.92%
Adience	Gender	77.73% \pm 1.82%	87.88% \pm 2.72%	82.89% \pm 2.64%	80.48% \pm 2.62%	
	Age	48.48% \pm 2.15%	59.51% \pm 3.60%	48.90% \pm 3.03%	49.06% \pm 3.25%	
Positive (b=4)	Colorferret	Gender	90.38% \pm 2.51%	97.12% \pm 0.59%	95.57% \pm 0.62%	93.79% \pm 0.72%
		Age	45.46% \pm 1.33%	54.54% \pm 1.23%	47.29% \pm 1.07%	48.20% \pm 1.06%
		Race	80.46% \pm 0.88%	87.66% \pm 1.82%	83.93% \pm 2.05%	85.10% \pm 1.53%
Adience	Gender	76.71% \pm 2.01%	87.38% \pm 1.89%	82.95% \pm 1.68%	79.43% \pm 1.75%	
	Age	50.43% \pm 2.62%	60.74% \pm 2.86%	51.10% \pm 2.66%	51.70% \pm 2.00%	

Table 5.3.: Face recognition performance on ColorFeret. The original face recognition performance is compared against three privacy-enhancing approaches, our proposed negative face recognition approach, cosine-sensitive noise (CSN), and euclidean-sensitive noise (ESN).

	FNMR@ 10^{-2} FMR	FNMR@ 10^{-3} FMR	EER
Original	3.65% \pm 0.95%	14.22% \pm 3.49%	1.97% \pm 0.21%
Ours ($k = 3$)	6.50% \pm 1.10%	18.32% \pm 4.23%	3.18% \pm 0.20%
CSN ($\Theta = 0.80$)	7.61% \pm 1.01%	23.54% \pm 4.42%	3.25% \pm 0.19%
ESN ($r = 0.75$)	7.47% \pm 1.12%	23.55% \pm 3.93%	3.21% \pm 0.25%
Ours ($k = 4$)	8.65% \pm 1.22%	20.26% \pm 2.81%	4.15% \pm 0.40%
CSN ($\Theta = 0.73$)	11.18% \pm 1.24%	32.16% \pm 5.03%	4.20% \pm 0.23%
ESN ($r = 0.93$)	11.56% \pm 0.99%	33.01% \pm 4.54%	4.24% \pm 0.19%

Table 5.4.: Face recognition performance on Adience. The original face recognition performance is compared against three privacy-enhancing approaches, our proposed negative face recognition approach, cosine-sensitive noise (CSN), and euclidean-sensitive noise (ESN).

	FNMR@ 10^{-2} FMR	FNMR@ 10^{-3} FMR	EER
Original	13.68% \pm 5.24%	45.71% \pm 6.88%	3.83% \pm 0.72%
Ours ($k = 3$)	13.42% \pm 4.79%	43.14% \pm 9.14%	4.43% \pm 0.80%
CSN ($\Theta = 0.84$)	19.36% \pm 6.30%	59.04% \pm 6.90%	4.48% \pm 0.75%
ESN ($r = 0.62$)	18.60% \pm 5.76%	57.56% \pm 6.59%	4.49% \pm 0.72%
Ours ($k = 4$)	16.35% \pm 4.20%	47.93% \pm 8.69%	5.44% \pm 0.78%
CSN ($\Theta = 0.74$)	28.29% \pm 7.49%	71.79% \pm 6.59%	5.57% \pm 0.80%
ESN ($r = 0.88$)	28.16% \pm 6.91%	71.27% \pm 5.73%	5.49% \pm 0.70%

Table 5.5.: Attribute suppression performance on the ColorFerret and Audience databases. The gender, age, and race suppression rates are shown for four function creep estimators. The highest suppression rates are highlighted.

	Gender suppression rate				Age suppression rate				Race suppression rate			
	RF	SVM	kNN	LogReg	RF	SVM	kNN	LogReg	RF	SVM	kNN	LogReg
Ours ($k = 3$)	22.2%	4.5%	5.5%	6.9%	30.2%	11.1%	9.3%	26.1%	14.6%	4.1%	4.1%	7.8%
CSN ($\Theta = 0.80$)	7.3%	3.6%	1.5%	4.1%	10.6%	6.0%	2.4%	6.7%	6.1%	2.3%	0.8%	2.5%
ESN ($r = 0.75$)	7.8%	3.8%	1.6%	3.9%	9.7%	5.9%	1.2%	7.4%	5.9%	2.3%	0.4%	2.8%
Ours ($k = 4$)	28.4%	7.4%	14.4%	10.9%	34.2%	13.7%	9.4%	31.2%	21.8%	7.6%	9.3%	12.0%
CSN ($\Theta = 0.73$)	11.0%	5.0%	2.6%	5.3%	12.1%	8.6%	4.2%	9.6%	8.5%	3.7%	1.6%	4.3%
ESN ($r = 0.93$)	10.3%	5.4%	3.5%	5.8%	13.5%	8.1%	4.8%	8.8%	8.7%	3.6%	2.6%	4.0%
Ours ($k = 3$)	26.1%	4.3%	6.8%	12.3%	28.7%	10.3%	8.0%	25.4%	-	-	-	-
CSN ($\Theta = 0.84$)	8.8%	2.8%	2.5%	4.9%	10.3%	4.9%	5.7%	5.8%	-	-	-	-
ESN ($r = 0.62$)	7.0%	4.3%	2.5%	5.2%	10.1%	4.6%	6.5%	5.0%	-	-	-	-
Ours ($k = 4$)	32.6%	11.0%	19.9%	18.0%	36.3%	14.1%	16.9%	29.8%	-	-	-	-
CSN ($\Theta = 0.74$)	14.6%	6.2%	6.2%	7.6%	16.6%	9.1%	12.2%	9.8%	-	-	-	-
ESN ($r = 0.88$)	14.1%	7.2%	6.5%	7.7%	15.1%	8.3%	11.9%	8.4%	-	-	-	-

5.6. PE-MIU: Privacy-Enhancement via Minimum Information Units

Previous works proposed privacy-enhancing solutions based on supervised [Mir+18; MR17; OR14] and unsupervised approaches [Ter+19b; Ter+20c]. While unsupervised approaches show a more comprehensive but weaker privacy-enhancement, supervised approaches are limited to the suppression of pre-defined attributes and thus, are vulnerable to unconsidered function creep attacks.

In this section, we propose PE-MIU [Ter+20h], a privacy-enhancing face recognition approach based on minimum information units. PE-MIU is a novel, training-free, and privacy-enhancing face recognition approach that works on the biometric template-level. Exploiting the structural differences between face recognition (use-case) and the estimation of facial attributes (attack scenario), our approach divides face templates into small blocks of minimal information units and randomly changes their positions in the templates. Since the information of privacy-sensitive attributes is usually distributed across the template, this approach significantly reduces the chance of function creep attackers to successfully estimate privacy-sensitive information from the modified face templates. To compare two modified templates, and thus verify if these belong to the same identity, we introduce an optimal assignment protocol. In this protocol, the minimal information units of both templates are assigned based on their optimal matching. This assignment is used to align and compare the templates.

The experiments were conducted on three publicly available databases in the context of function creep attackers who know and adapt to the used privacy mechanism. To put the results in a broad perspective, we compare our proposed solution against five state-of-the-art approaches that try to suppress the attribute gender on template-level. The experiments show that PE-MIU outperforms all other approaches in terms of suppressing privacy-risk attributes and maintaining recognition performance. It is able to reach significantly higher gender suppression rates than previous works in all investigated cases, and, at the same time, reaches a face recognition performance close to the unmodified face recognition system.

5.6.1. Methodology

Enhancing soft-biometric privacy aims at preventing function creep attackers from successfully predicting privacy-risk characteristics. This task is further challenged by simultaneously maintaining a high recognition ability. With our PE-MIU approach, we exploit the structural differences between a face recognition scenario and the scenario of a function

creep attacker. While the function creep attacker aims at the predicting privacy-sensitive information from *one template*, in face recognition *two templates* are compared to decided if they belong to the same identity or not. In this section, we propose a training-free approach for privacy-preserving face recognition, PE-MIU. Our PE-MIU approach divides face templates into small blocks and randomly changes their positions. These blocks are noted as minimum information units (MIU). Consequently, it is hard to reliably predict these characteristics. For the purpose of recognition, two templates are given and their relation to each other can be used to find corresponding MIUs. In the first step, the optimal assignment between the MIU's per template are calculated, to verify if two MIU-based templates belong to each other. In the second step, this assignment is used to align the templates and further compute their comparison score. This idea is illustrated in Figure 5.34 and detailed in the rest of this section.

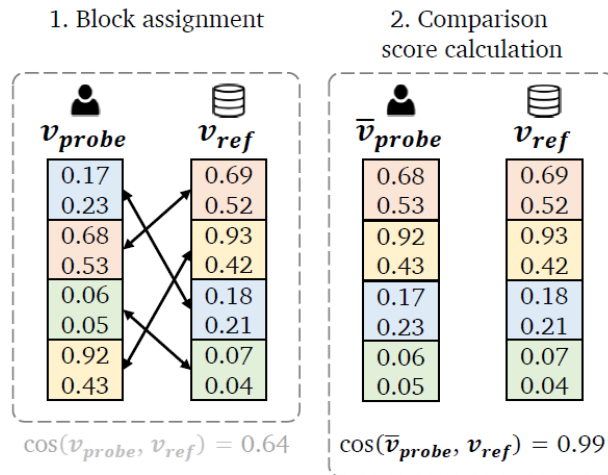


Figure 5.34.: Illustration of the comparison of two MIU templates. In the first step, an optimal assignment of the MIU blocks per template are computed. In the second step, this assignment is used to align both templates such that they can be compared with a standard similarity function.

Enrolment phase

In the enrolment phase, given a face image I , the corresponding MIU template v is computed and stored in the database. The computation of the MIU-based template is described in Algorithm 3. Given a face image I , the corresponding face embedding $x \in \mathcal{R}^L$

is extracted (*createEmbedding*) from I , where L is the size of the embedding. This face embedding x is divided (*divideMIU*) into L/s MIU blocks of size s . Then, the positions of these units are exchanged randomly (*shuffle*) resulting in a face template v where every entry consists of a feature block. This MIU template is then stored in the database. The process of dividing the embedding into MIU blocks and shuffling the block positions is illustrated in Figure 5.35.

Algorithm 3 - ComputeMIUTemplate(I, s)

Input: Face image I , bin size $s = 16$

Output: Face template v to be stored in the database

- 1: $x \leftarrow \text{createEmbedding}(I)$
 - 2: $v_{org} \leftarrow \text{divideMIU}(x, s)$
 - 3: $v \leftarrow \text{shuffle}(v_{org})$
 - 4: **return** v
-

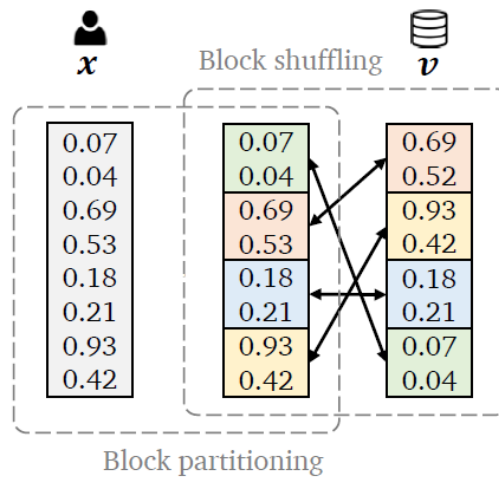


Figure 5.35.: Illustration of the key parts during enrolment: first, the face embedding x is divided into equally-sized blocks, the MIUs. Second, the position of these MIU blocks are randomly shuffled.

Verification phase

In the verification phase, an MIU reference template v_{ref} stored in the database is compared with an MIU probe template v_{probe} from a captured individual. The verification is done in two steps: first, the MIU-blocks of v_{probe} and v_{ref} have to be assigned such that there is an optimal pair-wise matching between the blocks of both templates. Second, the probe template v_{probe} is aligned to v_{ref} such that the matched MIUs are at the same entries of the templates. The aligned probe template \hat{v}_{probe} is then compared to v_{ref} using a similarity metric. In our case, as will be explained later, we use the cosine similarity metric as it was recommended to be used with the original templates in our experiments.

MIU-Block assignment In order to compare a probe embedding x with a block-wise reference template v_{ref} , an MIU representation of x have to be computed and the optimal matching of the MIU-blocks of each template has to be found. Similarly to the block partitioning during the enrolment (see Figure 5.35), the probe face embedding x is divided into MIU of size s , resulting in v_{probe} . Then, the best matching between the two MIU templates is computed. In graph theory, this problem is known as weighted bipartite matching problem [RT12] and is equivalent to the following optimization

$$\min_{\chi} \sum_{i,j} C_{i,j} \chi_{i,j}. \quad (5.22)$$

Applied to our problem, the cost matrix $C_{i,j}$ describes the euclidean distance between then MIU i and j and $\chi_{i,j}$ is the resulting binary assignment matrix with $\chi_{i,j} = 1$ if and only if the i^{th} probe MIU is assigned to the j^{th} reference MIU. This problem can be solved via the Hungarian [KY55] or the Ford-Fulkerson [FF10] algorithm.

The best matching task can be formulated and solved as a minimum cost maximum-network-flow problem [BDM12]. Therefore, an acyclic graph is constructed as shown in Figure 5.36. The edge weights from source q to the nodes (MIU blocks) of v_{probe} are set to 1. The same applies for the weights of the edges from the v_{ref} nodes to the sink z . The weights for the edges connecting the blocks between v_{probe} and v_{ref} are determined by its euclidean distances resulting in the cost matrix C . The optimal assignment of these MIUs is then defined by the maximum flow from source q to sink z .

Aligned comparison score After the optimal block assignment is found, the order of the blocks of the probe template are chosen such that the matched blocks are at the same positions. This results in an aligned probe template \bar{v}_{probe} . After the MIU-blocks of the probe and the reference templates \bar{v}_{probe} and v_{ref} are aligned, the comparison score

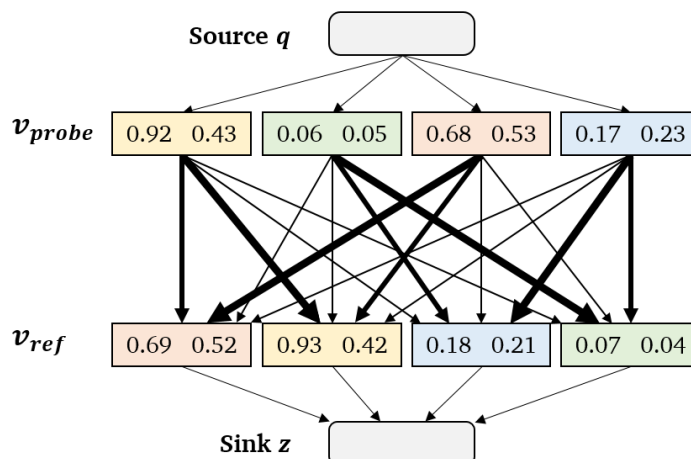


Figure 5.36.: The MIU-block assignment can be solved as a maximum-flow problem where the maximum flow from source q to sink z is defined by the optimal matchings. The weights between the MIU-blocks of v_{probe} and v_{ref} are given by their euclidean distances. Broader edges represent higher weight values.

$cs(\bar{v}_{probe}, v_{ref})$ of these templates is computed. In this work, we use cosine similarity for this comparison score calculation. However, this metric should be chosen according to the utilized face embeddings.

Summary of the verification phase To summarize the verification phase using MIU-templates, Algorithm 4 describes how a comparison score between a probe face image I of an identity and a reference template v_{ref} (stored in the database) of the claimed identity is computed. First, a MIU-based template v_{probe} is computed from image i using Algorithm 3 (*computeMIUTemplate*). Then, the probe template v_{probe} is aligned to v_{ref} (*align*) as described in Section 5.6.1. Finally, the aligned probe template \bar{v}_{probe} is compared with v_{ref} using a pre-defined comparison score metric $cs(\bar{v}_{probe}, v_{ref})$ such as cosine similarity. This comparison score is then used to determine if the person belongs to the claimed identity.

Algorithm 4 - Compare(I_{probe}, v_{ref})

Input: Face image I_{probe} , claimed-identity template v_{ref}

Output: Comparison score $score$

- 1: $v_{probe} \leftarrow \text{computeMIUTemplate}(I, s = 16)$
 - 2: $\bar{v}_{probe} \leftarrow \text{align}(v_{probe}, v_{ref})$
 - 3: $score \leftarrow cs(\bar{v}_{probe}, v_{ref})$
 - 4: **return** $score$
-

Properties of block-wise representations

The soft-biometric privacy-protection of the proposed method lies in the randomized nature of the MIU representation. Due to the fact that the previous order of the MIU-blocks is unknown and can only be reconstructed with an unmodified face embedding of the same identity, function creep attackers can only use the set of the minimal information units for their attacks.

Soft-biometric privacy usually describes a trade-off between suppressing privacy-sensitive attributes and maintaining the recognition ability of its templates. In this work, this trade-off is determined by the size of the MIU blocks s . Higher MIU sizes result in a weaker privacy-protection, due the fact that higher block sizes contain more attribute information. However, higher MIU sizes also leads to less misassigned MIU-blocks and thus, it leads to a lower recognition errors as well. For this work, we choose an MIU size of $s = 16$ to balance these two points. The effect of changing this parameter is investigated in Section 5.6.3.

The key part of verifying a persons identity with the proposed method is the MIU-block assignment. As indicated in Figure 5.34, the comparison of two not-aligned MIU templates results in a weak recognition performance. The block assignment, needed for the computation of the aligned MIU templates, is done via the Hungarian algorithm [KY55], since it provides stable and optimal assignments. This method scales with $\mathcal{O}(n^3)$, where $n = L/s$ is the number of MIU-blocks per template. Consequently, higher privacy-protection (smaller s) comes at the cost of higher computation times. However, this can be mapped to a complexity of $\mathcal{O}(n^2 \log n)$ by using the approach presented from Ramshaw and Tarjan [RT12].

5.6.2. Experimental Setup

Database

We conduct experiments on the publicly available ColorFeret [Phi+00] and Adience [EEH14] and Labeled Faces in the Wild (LFW) [Hua+07] databases to evaluate and compare our solution to related works. ColorFeret [Phi+00] consists of 14,126 images from 1,199 different individuals with different poses under controlled conditions. The Adience dataset [EEH14] consists of 26,580 images from over 2,284 different subjects under uncontrolled imaging conditions. Labeled Faces in the Wild (LFW) [Hua+07] provides 13,233 face images from 5749 identities. The databases cover a wide range of variations in illumination, focus, blurriness, pose, and occlusions. Moreover, the databases include information about the identities and their gender. This allows to deeply investigate the privacy-preservation techniques of the attribute gender, as well as their recognition performances.

Evaluation metrics

Preserving soft-biometric privacy is challenged by a trade-off between the desired degradation of the attribute estimation performance by function creep attackers and the desired preservation of the recognition ability. In the experiments, we report the verification performances in terms of false non-match rate (FNMR) at fixed false match rates (FMR). We further report the equal error rate (EER), which equals the FMR at the threshold where $FMR = 1 - FNMR$. Both verification performance measures are defined in the ISO standard [06]. In order to evaluate the attribute suppression performance, we report the results in terms of balanced attribute classification accuracy, since this allows an unbiased performance measure on testing data with unbalanced attribute information. This balanced accuracy is equivalent to the standard accuracy with class-balanced sample weights. A value of 50% is the best possible case for a privacy-preserving methodology and the worst outcome for a function creep attacker. In order to evaluate if the privacy enhancing method is beneficial, we use the privacy-gain identity-loss coefficient (PIC) defined in Section 5.4 and reported in recent works [Ter+19b; Bor+20]. The PIC is defined as

$$PIC = \frac{AE' - AE}{AE} - \frac{RE' - RE}{RE}. \quad (5.23)$$

The value is defined by attribute prediction errors AE' and AE and the verification errors RE' and RE with and without the privacy-preserving methodology. Positive values indicate that the privacy gain is higher than the loss in the identity preservation

performance. Since it measures how beneficial it is to apply the privacy transformation, a higher PIC coefficients indicates a better privacy-enhancing technique.

Face recognition model

In this work, our block-assignment-based approach builds on arbitrary face embeddings of certain dimensions. In the experiments, we utilize the widely used FaceNet model³ [SKP15] pretrained on MS-Celeb-1M [Guo+16]. In order to extract an embedding of a face image, the image is aligned, scaled, and cropped as described in [KS14]. The preprocessed face image is then passed into the face recognition model to obtain a 128-dimensional face embeddings. The comparison of two such embeddings is performed using cosine-similarity.

Function creep attacks

In this work, we consider two kinds of function creep attacks, the standard attack (S-ATK) and the advanced attack (A-ATK). We decided to introduce A-ATK due to the limited effectiveness of S-ATK on our proposed approach. Both attacks evaluate the attribute suppression performance and simulate the critical scenario of a function creep attacker that knows the systems privacy mechanism and adapts to it.

For the S-ATK, the adaptation is done by training (function creep) classifiers on the privacy-enhanced templates to predict the privacy-sensitive attributes. Before the training of these classifiers, the transformed templates are further normalized and scaled to unit-length. The utilized classifiers include random forest (RF), support vector machines (SVM), k-nearest neighbours (kNN), and logistic regression (LR). The hyperparameters of these classifiers are fine-tuned with Bayesian optimization.

During the experiments, we realized that these naive function creep attacks (S-ATKs) show only a very limited effect on our proposed approach, meaning that the classification performance with optimized function creep classifiers show a close to random behaviour. Therefore, we additionally considered more challenging attack classifier approaches for our proposed solution that is directly customised to achieve the highest classification accuracies. The most successful kind of attacks were the ones that learn to predict the gender for each MIU-block separately. During prediction, each of these blocks of a face template is classified separately and the predicted scores per classed are fused with a mean-fusion-rule [Ter+19c; Ter+18a]. In this work, we refer to this attack as A-ATK.

For the evaluation, we consider function creep attacks to the privacy-sensitive attribute gender as done in previous works [Mir+18; MRR18; MRR19; Ter+19b; Ter+20c;

³<https://github.com/davidsandberg/facenet>

Ter+19a; Ter+19b; MFV19; Bor+20]. The reason for this choice is that gender information can be estimated from face templates with very high accuracies [Ter+19d; Ter+19c]. Moreover, it requires only a binary decision, which makes it an easy target for function creep attackers and a challenge for privacy-preserving methodologies.

Baseline approaches

To evaluate our proposed training-free and template-based solution in a broad setting, we compare it against 5 recent template-based privacy-preserving face recognition approaches. These include the two supervised solutions PFRNet [Bor+20] and IVE [Ter+19a] and three unsupervised solutions NFR [Ter+20c], CSN [Ter+19b], and ESN [Ter+19b]. PFRNet [Bor+20] aims at learning a feature representation that disentangle identity from gender. The original network was optimized for an embedding size of 512. In our evaluation setting an embedding size of 128 is used. Consequently, the network was adapted such that the encoder consists of two layer with size 128 and 100+28 dimensions and the decoder consists of two layer with 128 dimensions as this adaptation showed the best privacy-preserving performance while maintaining high verification rates. As proposed in Section 5.3, IVE [Ter+19a] incrementally eliminate the most privacy-risk features from a face template to suppress the attribute information. CSN and ESN [Ter+19b] are based on geometric-inspired noise-injections that alter the inherent identity information in a controlled manner as introduced in Section 5.4. In contrast to mentioned approaches, NFR [Ter+20c] stores only complementary information about an individual in a face template and during deployment, it compares the probe template with a reference template in the complementary domain by calculating its dissimilarity. This solution was introduced in Section 5.5.

In order to make the experiments as comparable as possible, we calibrated the hyperparameters of these baselines in such a way that they reach a similar verification EER performance if possible. For all experiment scenarios, the same subject-exclusive 5-fold cross-validation setup is utilized. This includes training the function creep classifiers, as well as training the baseline approach for privacy-enhancing face recognition. The setup is shown in Table 5.6. For the three utilized databases, it provides details about each fold properties. It should be noted that for LFW, the gender distribution is unbalanced. For this reason, we choose the balanced attribute classification accuracy as described in Section 5.6.2.

Investigations

The investigations of this work are divided in four parts:

-
-
- A. We analyse the face verification performance of our privacy-enhancing solution in comparison to previous works.
 - B. We investigate the attribute prediction performance in a qualitative and quantitative manner.
 - 1) The qualitative investigation provides a qualitatively aided analysis of the gender separability of the original and the MIU-based templates. This is done by providing a visual understanding of the proposed approach.
 - 2) The quantitative investigation analyses the attribute prediction performance of the original template, on our solution, and on state-of-the-art. This is done in the critical scenario of a function creep attacker that adapts to the systems privacy mechanism using the attack scenarios S-ATK and A-ATK.
 - C. We analyse the parameter space of our solution to provide a deeper understanding of the influence of the MIU-block size on several aspects of our methodology.
 - D. Lastly, we focus on the strongest attack for each method and summarize the methods recognition ability, as well as the privacy-protection in a joint manner. This includes reporting the privacy-gain identity-loss coefficients (PIC) to measure and compare the usefulness of the studied approaches in the context of the most successful function creep attacks.

5.6.3. Results

Face verification performance

In Figure 5.37, the face verification performance is shown on three databases. The performance of the original FaceNet embeddings is shown along the performance of six privacy-enhancing approaches including our proposed approach. It can be seen that all approaches show a degraded face verification performance compared to the original embeddings. This is shown in every privacy-enhancing work [Ter+19b; Ter+19a; MRR19; MFV19], since soft-biometric privacy defines a trade-off between maintaining identity information and suppressing privacy-sensitive attributes. For lower FMR, NFR [Ter+20c] is an exception of this trade-off. Since in the NFR approach the comparison score is computed by the dissimilarity between the positive probe and the negative reference template, it is more robust to embeddings with more intra-class variations. In total, our proposed approach shows the most similar verification performance to the original templates. This can be noticed by both, the ROC curves and the EER values. The recognition performance is mostly maintained, due to nearly error-free MIU assignments.

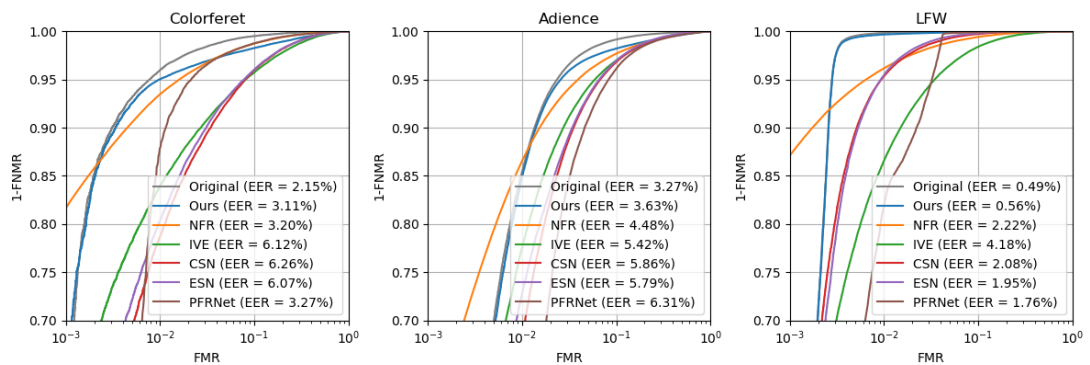


Figure 5.37.: Face recognition performance of our MIU-based solution in comparison with five state-of-the-art approaches on three databases. In addition, the face verification performance of the unmodified (original) face templates are shown.

Attribute suppression performance

Qualitative Analysis In order to provide a visual understanding of the proposed approach towards the suppression of gender characteristics, Figure 5.38 represents a 2D-visualization of 1000 randomly chosen identity-embeddings. The visualization was done by utilizing t-distributed stochastic neighbour embeddings (t-SNE) [MH08]. Female samples are characterized by orange dots, while male samples are represented by blue points. The visualizations are provided on three databases for the unmodified (original) templates (a, d, g), for the MIU-based templates of our PE-MIU approach (b, e, h), and for individual MIU-blocks individually (c, f, i). A clear separation is observed in the visualizations of the unmodified images (a, d, g) indicating that the attribute gender can be correctly predicted to a high degree. In contrast, the plots visualizing our approach (b, e, h) show highly randomized patterns indicating that it is hard to reliably estimate the correct attribute. In order to show that the same applies for individual MIUs separately, Figure 5.38 (e, f, i) show the same visualization per MIU. Similarly as for our full approach (e, f, i), no pattern between the different gender classes is easily observable.

Quantitative Analysis To deeply understand the privacy-enhancement of our solution along with previous works, Table 5.7 shows the balanced accuracies for four optimized function creep classifiers on the three databases. The gender prediction performance is shown for the unmodified (original) templates, for the our PE-MIU approach (Ours), and

for five state-of-the-art solutions. Moreover, the results for a highly challenging attack methodology (A-ATK), directly designed to maximize successful attacks on our approach, is shown as Ours*.

The gender decision accuracies of the original FaceNet embeddings show high values, demonstrating the need for privacy-enhancing technologies. The state-of-the-art privacy-preserving face recognition approaches lead to degraded estimation performances. However, the gender decision accuracies, and thus the resulting attribute suppression, varies a lot depending on the utilized database and function creep classifier. Generally, the highest privacy-improvement is observed for our proposed approach. The function creep classifiers achieve correct classification performances close to a random decision behaviour of 50% in most cases. One exception is the scenario where the KNN estimator was used on the Adience database. Here, PFRNet reaches a slightly more randomized behaviour (54.53%) than our proposed approach (45.45%). However, this comes at the cost of a lower verification performance, e.g in terms of EER where the original templates achieve an EER of 3.27%, our PE-MIU approach reaches an EER of 3.63%, and PFRNet reaches an EER of 6.31%.

In the last row of Table 5.7, the suppression performance of our PE-MIU approach (Ours*) in the context of an highly advanced and adapted attack methodology (A-ATK) is shown. It demonstrates a strong gender suppression performance can be achieved even in this more critical and challenging attack scenario.

Parameter Analysis

The block size s of PE-MIU is the key to determine the privacy trade-off between reaching high attribute suppression rates and maintaining a high recognition performance. Therefore, this parameter is investigated in this Section on the three databases ColorFeret, Adience, and LFW. Figure 5.39 analyses the influence of the block size on the two aspects of the mentioned trade-off. High block sizes lead to lower recognition EER, since the number of possible wrongly-assigned MIU-blocks is lower. At the same time, high block sizes contain more patterns that allow function creep classifiers to successfully predict privacy-risk attributes. Figure 5.39 shows that a block size $s = 16$ represents a good balance between both aspects of the soft-biometric privacy trade-off.

The block size s also determines the computational complexity of our proposed MIU-based privacy-preserving face recognition approach. Table 5.8 shows the average computation time needed for the different MIU-steps. All computational efficiency analyses are based on using a personal computer with an Intel(R) Core(TM) i7-7700 processor. During

enrolment, the MIU-template must be generated. This can be implemented efficiently⁴ and thus, can be performed in the order of a few microseconds per template. During the verification phase, the MIU-blocks are assigned⁵ and then, the aligned templates are compared⁶. The biggest part of the computation time is needed for the MIU-block assignment. The average comparison time of previous works [Ter+19b; Bor+20] is around $0.10ms$ using the same CPU. Consequently, the strong privacy-enhancement and recognition performance of our proposed approach comes at the cost of higher comparison times.

In order to analyse the susceptibility to errors, Figure 5.40 shows the average ratio of misassigned blocks per genuine pair comparison. These statistics are shown for different block sizes s and for the three utilized datasets. For small block sizes (e.g. $s = 4$), the MIU-blocks contain few information for a reliable assignment. In this case, around 50% of the blocks are incorrectly assigned, explaining the relatively low recognition performance for $s = 4$ in Figure 5.39. On the other hand, for large block sizes (e.g. $s = 32, 64$), the ratio of misassigned blocks is close to zero (0.5% on ColorFeret, 0% on Adience and LFW, when $s = 64$) and thus, the templates are perfectly aligned in nearly all cases. The perfect alignment leads to low recognition errors. However, it also leads to higher gender decision accuracies of the function creep estimators, as it is demonstrated in Figure 5.39. In this work, beside analysing different block sizes, we decided to use a block size of $s = 16$, since it provides a suitable trade-off between maintaining the recognition ability and achieving a high privacy-enhancement, as supported by the information presented in Figure 5.40. For $s = 16$, the ratio of misassigned blocks varies between 1-5% on the different databases (5% on ColorFeret containing profile face images). This shows that in most cases, two genuine MIU-templates are close to perfectly aligned. At this block size, genuine MIU-blocks that are very similar can be wrongly assigned. However, since these are very similar to each other, the aligned MIU-templates are similar as well and thus, these misassigned blocks have only a minor impact on genuine comparisons.

Summary and Usability

Soft-biometric privacy is challenged by maintaining a high recognition performance and degrading the prediction performance of privacy-sensitive attributes. In Figure 5.41, both aspects can be observed simultaneously under the critical scenario of the most successful individual function creep attack. This is shown for each of the three databases. The x-axis represents the recognition error in terms of EER while the y-axis shows the balanced

⁴In this work, we implemented this part with Numpy [WCV11].

⁵For the block assignment, the Hungarian algorithm implementation from SciPy [Vir+20] was used.

⁶The comparison calculation with cosine similarity was computed with Scikit-learn [Ped+11].

gender prediction error. Consequently, a highly successful privacy-enhancing solution can be found in the top left corner. Moreover, the PIC coefficient is calculated and represented by the radius of the shaded area around a marker. Since PIC measures the advantages of applying the privacy-preserving methodology (see Section 5.6.2, a bigger shaded area represents a high usefulness of applying a solution. As demonstrated our proposed solution achieves the lowest recognition error on all scenarios. Moreover, it also leads to the highest gender prediction errors in most cases and to the highest PIC coefficients in all cases. Consequently, the PIC values (represented as the shaded areas) indicate that our proposed approach is significantly more effective than previous work.

5.6.4. Interim Conclusion

In this section, we proposed PE-MIU, a training-free and privacy-preserving face recognition approach based on minimum information units (MIUs). Our solution exploits the structural differences between the different setups of face recognition (use-case) and facial attribute estimation (attack scenario). This is achieved by dividing a face template into several MIU-blocks and randomly changing their position in the template. This kind of randomized representations changes the pattern of its attributes for each template. Consequently, it is hard for function creep attackers to predict these privacy-sensitive attributes. The experiments were conducted on three publicly-available databases comparing our solution to five state-of-the-art approaches. In the experiments, we simulated function creep attackers that know about the systems privacy mechanism and adapt their attacks based on it. The results show that our novel face recognition approach is able to consistently reach low attribute prediction rates in all investigates scenarios, outperforming all state-of-the-art approaches in most cases. Simultaneously, our solution maintains its recognition ability to a significantly higher degree than previous work. Consequently, unlike previous work, the proposed methodology is characterized by its ability to maintain a high recognition performance while reaching high attribute suppression rates which are not limited to the suppression of predefined attribute.

Table 5.6.: Properties of the used cross-validation setup for the three utilized datasets, ColorFeret, Adience, and LFW. The number of samples, identities, and the percentage of female individuals are reported per training and testing fold.

Dataset	Testing fold					Training fold					
	0	1	2	3	4	0	1	2	3	4	
ColorFeret	# samples	2160	2162	2159	2159	2159	8639	8637	8640	8640	8640
	# identities	189	191	190	190	190	761	759	760	760	760
	Ratio of Female	63.6%	68.2%	58.6%	60.5%	67.9%	63.8%	62.6%	65.0%	64.6%	62.7%
Adience	# samples	3868	3868	3868	3868	3867	15471	15471	15471	15471	15472
	# identities	456	456	457	457	456	1826	1826	1825	1825	1826
	Ratio of Female	55.5%	47.2%	46.2%	45.9%	47.0%	46.6%	48.7%	48.9%	49.0%	48.7%
LFW	# samples	2629	2629	2628	2628	2628	10513	10513	10514	10514	10514
	# identities	1137	1145	1146	1146	1147	4584	4576	4575	4575	4574
	Ratio of Female	20.7%	23.1%	26.0%	23.0%	19.8%	23.0%	22.4%	21.7%	22.4%	23.2%



Figure 5.38.: Visualization of different face representations of the three databases, ColorFeret (5.38a - 5.38c), Adience (5.38d-5.38f), and LFW (5.38g-5.38i). For the visualizations, 500 female and 500 male identities were chosen randomly and their templates are reduced to two dimensions using t-distributed stochastic neighbour embedding (t-SNE) [MH08]. The first row (a, d, g) shows the t-SNE plots for the original facenet embeddings. The second row (b, e, h) shows the same plots for templates modified by our PE-MIU approach. The last row represents the t-SNE plots created from each block of the templates. The lower separability introduced by PE-MIU in comparison to the unmodified templates is demonstrated by the increasingly overlapping samples.

Table 5.7.: Balanced gender decision accuracies on the three databases ColorFeret, Adience, and LFW. The gender prediction performance is determined by four function creep classifiers on the unmodified (original) templates, on our MILU-based templates, and the templates created by previous works using S-ATK. The results showing the most randomized accuracies per classifier are highlighted. Ours* represents the highly challenging attack methodology (A-ATK) that was specifically designed to maximize successful predictions on modified MILU-based templates as explained in Section 5.6.2.

Method	ColorFeret				Adience				LFW			
	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM
Original	97.55%	96.90%	91.62%	97.62%	84.37%	87.32%	84.52%	89.81%	85.99%	86.02%	66.61%	89.50%
IYE [Ter+19a]	89.57%	79.35%	75.28%	95.34%	77.53%	63.72%	76.37%	84.48%	65.76%	62.12%	64.78%	74.76%
ESN [Ter+19b]	96.01%	91.37%	83.17%	95.60%	84.11%	78.21%	80.89%	88.23%	77.16%	71.20%	56.02%	84.02%
CSN [Ter+19b]	92.06%	89.33%	82.88%	86.40%	80.03%	75.28%	74.09%	62.15%	69.41%	66.46%	55.56%	82.56%
NFR [Ter+20c]	91.95%	89.13%	72.68%	92.24%	79.15%	72.30%	61.18%	80.51%	87.86%	76.27%	77.55%	85.07%
PRFNet [Bor+20]	81.17%	84.93%	65.95%	82.00%	54.53%	58.25%	52.70%	64.63%	82.16%	78.97%	54.43%	83.38%
Ours	67.64%	63.32%	51.87%	68.71%	45.45%	44.71%	50.63%	49.94%	55.38%	52.69%	50.23%	63.54%
Ours*	69.73%	66.01%	54.97%	70.51%	66.75%	56.12%	66.36%	70.90%	59.98%	53.84%	53.41%	67.93%

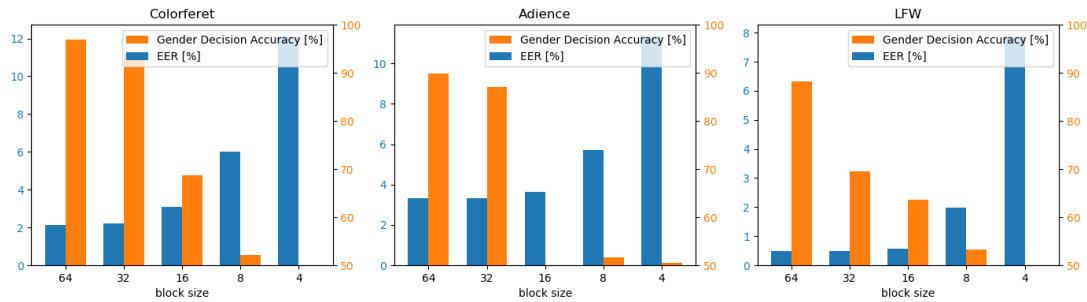


Figure 5.39.: Investigation of the block size: the gender decision performance and the recognition EER over the different block sizes are shown for the three databases. The gender decision accuracy comes from the most successful function creep estimator in Table 5.7, the SVM.

Table 5.8.: Average computational time (in $m.s$) of the different MIU steps for different block sizes s . The values represent the computational time on for an Intel(R) Core(TM) i7-7700 CPU with 3.60 GHz on 128-dimensional templates. The template generation refers to the enrolment phase, while the other steps refer to the verification phase.

Timings [$m.s$]	Block size s				
	4	8	16	32	64
MIU-template generation	0.0039	0.0022	0.0014	0.0010	0.0008
MIU-block assignment	7.10	2.23	0.67	0.33	0.06
Comparison score calculation	0.10	0.10	0.10	0.10	0.10
Complete MIU-verification	7.20	2.33	0.77	0.43	0.16

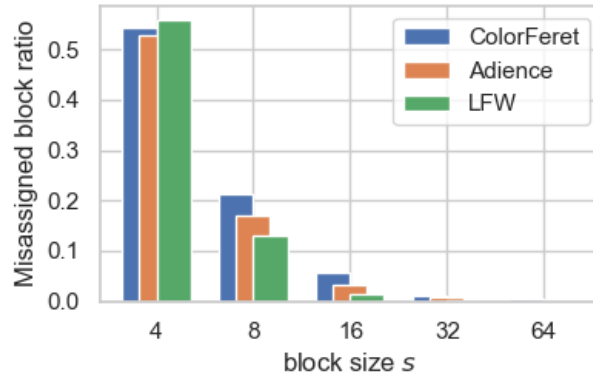


Figure 5.40.: Analysis of misassigned MIU-blocks. The average ratio of misassigned blocks per genuine pair comparison is shown for different block sizes on the used databases. Higher block sizes reduce the possibility of misassignments.

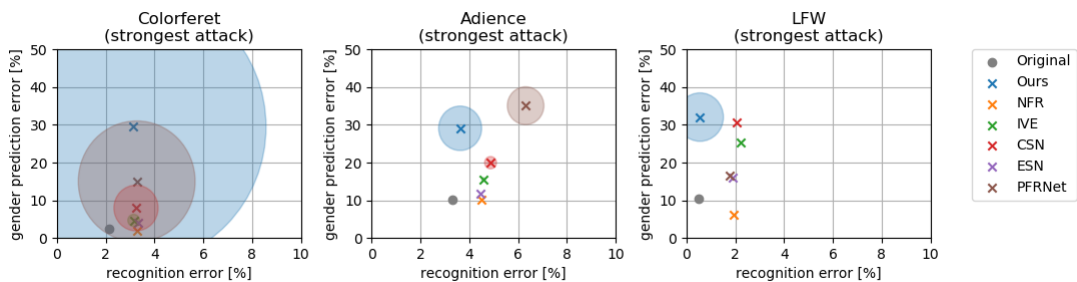


Figure 5.41.: Joint analysis of the privacy-risk attribute suppression performance (in terms of balanced accuracies) and the verification performance (in terms of EER). The performance of the unmodified (original) templates (grey dot) is shown in comparison with our MIU-based approach (ours) and five approaches from previous work. The prediction errors refer to the individually most successful attack classifier. To visually encode the usefulness in this challenging attack scenario, PIC values are calculated and represented as shaded areas around the method markers. Negative PIC values are neglected in the plot.

5.7. Summary

The face is a widely used biometric modality [Dam+18d; Wan+18a] that does not require an active user-participation [Dam+18a; Tri17]. A typical face recognition system contains feature representations (templates) for each individual enrolled. Comparing two templates allows to verify a claimed identity or to identify an unknown subject [PKB16]. However, recent work showed that more information than just the person's identity can be deduced from these templates [DER16]. With the use of soft-biometric estimators, information about gender, age, ethnicity, sexual orientation or the health status can be obtained [DER16; WK18]. Since in many applications the users do not permit to have access to this information, this shows a major invasion of privacy. In many systems, the stored data should be exclusively used for recognition purposes [MR17] and extracting such information without a person's consent can be considered as a violation of their privacy [Kin13]. Soft-biometric privacy-enhancing technologies aim at hiding or suppressing privacy-sensitive information in face templates to prevent function creep.

Soft-biometric privacy is challenged by maintaining a high recognition performance while effectively suppressing privacy-sensitive attributes. Unlike previous works, in this chapter, we (a) proposed mainly unsupervised privacy-enhancing face recognition solutions that (b) are not restricted to the suppression of pre-defined attributes, and (c) operates on template-level. Moreover, (d) we investigate the privacy performance in a more critical and challenging scenario of a function creep attacker that adapts his attacks to the systems privacy mechanism.

In Section 5.3, we proposed IVE [Ter+19a], an Incremental Variable Elimination that determines and eliminates the highest privacy-risk variables in face templates. Although it was able, by design, to suppress binary, categorical, and continuous attributes, similar to previous works, it is restricted to the suppression of pre-defined attributes.

In Section 5.4, we proposed Similarity-Sensitive Noise Transformations [Ter+19b]. These unsupervised privacy-enhancing face recognition approaches inject geometric-inspired noise to biometric templates. Despite that it offers a more comprehensive privacy-enhancement, higher suppression performance came at the cost of a low recognition ability.

In Section 5.5, we proposed negative face recognition [Ter+20c]. This unsupervised and privacy-enhancing face recognition approach introduces negative templates that, in contrast to ordinary (positive) face templates, contain only random complementary information about the individual. Consequently, storing only negative templates in the database, prevents function creep attackers from reliably predicting privacy-sensitive attributes. Although this approach preserves the recognition performance to a high degree, it still leaves space for improvements concerning the attribute suppression.

Finally, we proposed the main contribution of this chapter in Section 5.6, privacy-enhancement based on minimum information units (PE-MIU) [Ter+20h]. This training-free and privacy-enhancing face recognition approach prevents function creep attackers from successfully predicting privacy-sensitive information from face templates. PE-MIU exploits the structural differences between face recognition (use-case) and facial attribute estimation (attack scenario). It creates templates consisting of minimum information units (MIUs) in a random order. This makes the estimation of privacy-sensitive attributes a hard task for function creep attacks. During verification, the MIUs of a probe template are assigned to the MIUs of a reference template by solving an optimal best-matching problem. This allows an alignment of both templates and thus, a meaningful comparison. The results demonstrate that on both, maintaining recognition rates and suppressing attribute information, PE-MIU consistently reaches high performances and outreaches state-of-the-art solutions.

6. Conclusion and Future Work

This thesis has been guided by five research questions as defined in Section 1.1. Detailed responses to these questions were given in the previous Chapters 3, 4, and 5. This chapter summarizes the conclusions of this work and presents an outlook for future research.

6.1. Conclusion

Face recognition systems have a growing effect on everybody's daily life including critical decision-making processes. The wide-spread of these systems is based on the advances in extracting deeply-learned templates of face images that provide a strong identity-discriminability. However, the success of these templates comes at the cost of two major concerns caused by soft-biometric attributes.

Bias concerns - The performance of current biometric solutions are often strongly dependent on the user's soft-biometric attributes. This lead to remarkable differences in the recognition performance for different individuals and thus, to strong discriminatory effects.

Privacy concerns - The deeply-learned template of an individual is extracted to verify a person's identity. However, also privacy-sensitive information is encoded in such a template. For many applications, these templates should be used for recognition only. This raises major privacy issues since this information can be extracted without authorization.

This work aims at mitigating soft-biometric driven bias and privacy concerns from face recognition systems to enhance the reliability, trust, and dissemination of these systems. The mitigation of these concerns is guided by a set of five unsolved research questions. These are designed to first understand the influence of specific soft-biometric attributes on these concerns and then, to use the gained knowledge for the development of effective mitigation mechanisms. Unlike previous works, the proposed solutions are easily-integrable into existing systems and aim for comprehensive mitigation that is not limited to pre-defined attributes.

The thesis is divided into three parts focusing on (1) the investigation of soft-biometric driven bias and privacy concerns, (2) the mitigation of soft-biometric bias, and (3) the mitigation of soft-biometric privacy issues in face recognition.

Investigation of soft-biometric driven concerns The first part of this work (Chapter 3) aims at investigating soft-biometric attributes and their influence on soft-biometric privacy and bias issues in face recognition. The investigations demonstrate the need for more generalized solutions beyond the mitigation of demographic attributes only. Moreover, the findings guided the development of the solutions for mitigating these concerns.

In order to analyse soft-biometric driven concerns in face recognition, these soft-biometric attributes have to be reliably estimated. Answering RQ1, a novel reliability measure is proposed to quantify the confidence of the model's prediction. Utilizing multiple stochastic forward passes through dropout-reduced neural networks, the centrality and dispersion of these predictions are used to derive a prediction confidence. The methodology was shown to be highly successful for the estimation of soft-biometric attributes. Moreover, it creates the basis of the investigations on the soft-biometric driven bias and privacy concerns (RQ2 and RQ4).

To analyse the influence of soft-biometric attributes on the behaviour of face recognition systems, as stated in RQ2, the proposed reliability measure is used to create the MAAD-Face database. The proposed MAAD-Face annotations database consists of 123.9Mio high-quality attribute annotations of 47 different binary attributes for 3.3Mio face images. Consequently, it provides 15 and 137 times more attribute labels than related databases, such as CelebA and LFW, and further provides annotations of higher quality. These characteristics make MAAD-Face highly suitable for a comprehensive analysis of face recognition bias.

The next contribution of this work was an investigation of the influence of soft-biometric attributes on the performance of face recognition systems. This aims to partially answer RQ2 "How do specific soft-biometric attributes affect the behaviour of face recognition systems?". The investigation is built on the MAAD-Face database and investigates the influence of 47 attributes on the verification performance of two popular face recognition models. To prevent misleading statements about the effect of an attribute on the performance differences, control group based validity values are introduced to decide if unbalanced test data causes the performance differences and the correlations between annotations are analysed to emphasize if an attribute bias might originate from correlating annotations. The results demonstrate that also many non-demographic attributes strongly affect the recognition performance, such as accessories, hairstyles and colors, face shapes, or facial anomalies.

To fully analyse the behaviour of soft-biometric attributes on face recognition systems, also the soft-biometric bias in face quality assessment is investigated. Face quality assessment aims at estimating the utility of a face image for the purpose of recognition and plays a major role in the enrolment of face images. The analysis focused on the correlation between face recognition bias and bias in face quality assessment. It was shown that current face quality assessment solutions have to deal with the same bias-related issues than in face recognition models caused by unintended bias-transfers during the training phases.

Lastly, RQ4 is answered by analysing what (soft-biometric) information is stored in biometric face templates. This aims to support the development of privacy-enhancing face recognition technologies. The question is answered by investigating the predictability of 113 attributes from face templates at different difficulty-levels with the help of the reliability measure of RQ1. Understandable statements about the stored attribute information are derived by categorizing each attribute into one of three predictability classes. The results show that up to 74 attributes can be accurately predicted from face templates demonstrating the need for privacy-enhancing solutions in face recognition. Despite that face templates are learned to be robust to non-permanent factors, the results demonstrate that especially these attributes are easily-predictable. This includes information about age, hairstyles, haircolors, beards, and accessories, such as make-up, lipstick, and glasses. The results show that much more information is stored in biometric templates than reported in previous works demonstrating the need for more generalized solutions to enhance the soft-biometric privacy in face recognition systems.

This part of this work answered RQ 2 and 4 (What soft-biometric attributes are stored in biometric face templates and how do these affect the behaviour of face recognition systems?). Unlike previous works that focused their investigations on demographic attributes, it was shown that also a large number of non-demographic attribute are stored in face templates and that these significantly affect the behaviour of face recognition systems. This demonstrate the strong need for face recognition solutions mitigating soft-biometric bias and privacy concerns beyond demographics.

Mitigation of soft-biometric bias in face recognition The second part (Chapter 4) of this work deals with the mitigation of soft-biometric bias in face recognition. Previous works developed solutions for this problem that mitigate demographic bias and require computationally heavy database replacements when integrated into existing systems. Answering RQ3, in this thesis, two solutions are proposed that operate on the comparison- and the score-level of a recognition system and thus, can be easily-integrated in an existing system. Moreover, the proposed solutions are not limited to the mitigation of

demographic-bias.

The first contribution for bias-mitigation is a supervised fair template comparator that integrates different notations of fairness at the comparison-level of the system by replacing the deployed similarity function with a fairness-driven similarity estimator. A fairness-term is integrated into the loss function and forces the score distributions of different groups (e.g. ethnicities) to be similar. The solution achieved bias reduction rates between 15.35% and 52.67% while only marginally affecting the recognition performance.

The second contribution for mitigating bias in face recognition is an unsupervised fair score normalization approach. The proposed solution integrates the notation of individual fairness at the score-level of the system and thus, aims at treating similar individuals similarly. This is achieved by clustering training samples in the embedding space and computing optimal local thresholds for each cluster. For calculating the comparison score of two samples, it normalizes this score based on the optimal local thresholds of the sample-associated clusters. This ensures a more individual, unbiased, and fair treatment. The results on three publicly available databases demonstrate that the proposed solution mitigates bias by up to 82.7%. Moreover, it reduces the bias more consistently than existing works and enhances the overall recognition performance by 53.2% at an FMR of 10^{-3} and by 82.9% at an FMR of 10^{-5} . In contrast to previous works, the proposed fair score normalization solution jointly (a) operates on unlabelled training data, (b) effectively mitigates bias of unknown origins, and (c) strongly improves the overall recognition performance of the system.

Mitigation of soft-biometric privacy concerns in face recognition The third part (Chapter 5) of this thesis aims at enhancing the soft-biometric privacy in face recognition systems. Previous works developed solutions for this problem that focus on the suppression of specific demographic attributes and are further hardly-integrable into existing systems. Answering RQ5, four easily-integrable solutions are proposed that aim at suppressing privacy-risk information of various origins from face templates while maintaining a high recognition ability. The proposed solutions either manipulate existing face templates directly or change the template-representation including inference-process for verification.

Incremental variable elimination identifies and eliminates privacy-risk variables from face templates to reduce the encoded privacy-sensitive information. The approach is based on a decision tree ensemble that allows deriving a variable importance measure. This measure is used to incrementally find and delete variables that allow predicting sensitive attributes. In contrast to previous works, this approach is, by design, able to suppress binary, categorical, and continuous attributes.

Similarity-sensitive noise transformations inject geometric-inspired noise to face tem-

plates to enhance soft-biometric privacy in an unsupervised manner. This aims at achieving a more comprehensive soft-biometric privacy-enhancement than previous works that is not limited to pre-considered attributes. Moreover, the degradation of the recognition performance can be directly controlled by the transformation parameters.

Negative face recognition is a proposed unsupervised recognition approach that stores negative templates of the users containing only information that the individuals do not have. Storing only negative templates in the database, prevents function creep attackers from successfully leaking privacy-risk information. For verification, the positive template of an individual is compared with the stored negative template of the claimed identity by measuring their dissimilarity. Even in unconstrained scenarios, negative face recognition fully retains the recognition performance while achieving suppression rates of up to 36% outperforming related solutions.

PE-MIU is a proposed solution to enhance the soft-biometric privacy in face recognition based on minimum information units. This training-free approach exploits the structural differences between face recognition and facial attribute estimation by creating templates in a mixed representation of minimal information units. These representations contain patterns of privacy-sensitive attributes in a highly randomized form and thus, the estimation of these attributes becomes hard for function creep attacks. During verification, these units of a probe template are assigned to the units of a reference template by solving an optimal best-matching problem. This allows our approach maintaining a high recognition ability. The results demonstrate that PE-MIU is able to consistently reach higher suppression rates in all investigated scenarios, outperforming all state-of-the-art approaches in most cases. Simultaneously, PE-MIU maintains its recognition ability to a significantly higher degree than state-of-the-art solutions. Unlike previous works, the proposed solution offers a strong and comprehensive privacy-enhancement without the need for training or modification of the deployed face recognition model.

In this thesis, I investigated soft-biometric driven bias and privacy concerns in face recognition systems and proposed solutions for their mitigation. Previous investigations on these concerns focused on demographics only. The analysis of this work demonstrated that soft-biometric attributes beyond demographics affect these concerns and thus, solutions are needed that consider soft-biometric attributes in general. However, current solutions for mitigating these concerns are limited to single attributes and are further hard to integrate into existing systems. Therefore, in this work, I proposed several highly-effective solutions for mitigating these concerns that are not limited to single attributes and are easily-integrable into existing systems. This aims at enhancing the reliability, trust, and dissemination of these systems. In addition, the proposed solutions of this work are not

limited to face biometrics.

6.2. Future work

This thesis analysed soft-biometric driven bias and privacy concerns in face recognition systems and proposed several solutions to mitigate these. The findings of the investigations not only affected the design of the proposed solutions but might also have strong implications on the development of future works. Moreover, the concepts of the proposed solutions might be successfully transferable to different domains such as the concept shift to different biometric modalities.

Implications of the investigations In this work, the performed investigations on soft-biometric driven bias and privacy concerns in face recognition demonstrated the research gaps that future works have to address. It was shown that also a wide range of non-demographic attributes is causing these concerns. Therefore, future works have to propose more generalized solutions that are not limited to pre-defined (demographic) attributes.

Moreover, it is shown that the influence of soft-biometric attributes is not limited to the performance of face recognition systems. Rather, it is shown that soft-biometric attributes affect the general behaviour of face recognition systems including the assessment of face image quality. The investigations of this work demonstrate that current face recognition solutions possess similar bias-related issues than for face verification. Consequently, future work might work on making the face image quality assessment, and thus the enrolment process in general, fairer.

Future work on bias mitigation in face recognition This work focuses on easily-integrable solutions for mitigating soft-biometric driven bias in face recognition systems. The high integrability was achieved by the development of solutions beyond the template-level. Representation-learning approaches operate closer to the origin of the bias and thus, might mitigate bias to a higher degree. Moreover, mitigating bias on the representation-level also reduces the bias in the face quality assessment since the use of face recognition models as a basis for face quality assessment leads to an unintended bias-transfer as shown in this thesis. Current representation-learning approaches for mitigating bias in face recognition focus on the mitigation of demographic attributes. To extend the generalizability of these approaches, future works might include the notation of individual fairness into the network training. The proposed fair score normalization approach demonstrated the effectiveness of this fairness notation to achieve bias mitigation not limited to pre-defined attributes.

Future work on the enhancement of soft-biometric privacy The proposed privacy-enhancing approach based on minimum information units (PE-MIU) demonstrates that it is able to maintain the face recognition performance while comprehensively suppressing privacy-risk information in face templates even in the critical scenario of a function creep attacker that knows and adapts to the systems privacy mechanism. This success comes at the cost of a long comparison time. Future works might elaborate on this issue, for instance by providing a faster approximative algorithm to solve the optimal best matching problem of the minimum information units.

To make privacy-enhancing methodologies more easily comparable, privacy benchmarks and evaluation protocols are needed for soft-biometric privacy in face recognition. This might additionally accelerate the development of such solutions since less time is spent on re-implementing existing approaches leaving more time for the privacy-enhancement.

Joint mitigation of bias and privacy concerns Although the same soft-biometric attributes causing the bias and privacy concerns in face recognition, current solutions either focus on the mitigation of bias or privacy concerns separately. Consequently, many solutions can not be applied simultaneously and thus, does not allow joint mitigation of both concerns. Future work might explore synergy effects of solutions for both concerns and merge these concepts to develop a coherent framework that jointly mitigates soft-biometric driven bias and privacy concerns in face recognition.

A. Appendix

Section 3.4 investigates bias in face quality assessment solutions answering RQ2. In this investigation, one of the most important quality assessment solutions is SER-FIQ. SER-FIQ is an unsupervised face quality assessment concept that uses the robustness of a face representation as its quality indicator. In Section 3.4, this is only shortly summarized since it does not fit into the general theme of this thesis. To enable a full understanding of the results on bias in face quality assessment, in the following, the proposed face quality assessment concept is introduced in more details [Ter+20g].

SER-FIQ: Stochastic Embedding Robustness for Face Image Quality

Introduction

Face images are one of the most utilized biometric modalities [Wan+18a] due to its high level of public acceptance and since it does not require an active user-participation [Tri17]. Under controlled conditions, current face recognition systems are able to achieve highly accurate performances [GNH18]. However, some of the most relevant face recognition systems work under unconstrained environments and thus, have to deal with large variabilities that leads to significant degradation of the recognition accuracies [GNH18]. These variabilities include image acquisition conditions (such as illumination, background, blurriness, and low resolution), factors of the face (such as pose, occlusions and expressions) [11; 15] and biases of the deployed face recognition system. Since these variabilities lead to significantly degraded recognition performances, the ability to deal with these factors needs to be addressed [Her+19].

The performance of biometric recognition is driven by the quality of its samples [BJ18]. Biometric sample quality is defined as the utility of a sample for the purpose of recognition [Her+19; Phi+13; Gao+07; BJ18]. The automatic prediction of face quality (prior to matching) is beneficial for many applications. It leads to a more robust enrolment for face recognition systems. In negative identification systems, it prevents an attacker from

getting access to a system by providing a low quality face image. Furthermore, it enables quality-based fusion approaches when multiple images [DSN14] (e.g. from surveillance videos) or multiple biometric modalities are given.

Current solutions for face quality assessment require training data with quality labels coming from human perception or are derived from comparison scores. Such a quality measure is generally poorly defined. Humans may not know the best characteristics for the utilized face recognition system. On the other hand, automatic labelling based on comparison scores represents the relative performance of two samples and thus, one low-quality sample might negatively affect the quality labels of the other one.

In this work, we propose a novel unsupervised face quality assessment concept by investigating the robustness of stochastic embeddings. Our solution measures the quality of an image based on its robustness in the embedding space. Using the variations of embeddings extracted from random subnetworks of the utilized face recognition model, the representation robustness of the sample and thus, its quality is determined. Figure A.1 illustrates the working principle.

We evaluated the experiments on three publicly available databases in a cross-database evaluation setting. The comparison of our approach was done on two face recognition systems against six state-of-the-art solutions: three no-reference image quality metrics, two recent face quality assessment algorithms from previous work, and one commercial off-the-shelf (COTS) face quality assessment product from industry.

The results show that the proposed solution is able to outperform all state-of-the-art solutions in most investigated scenarios. While every baseline approach shows performance instabilities in at least two scenarios, our solution shows a consistently stable performance. When using the deployed face recognition model for the proposed face quality assessment methodology, our approach outperforms all baseline by a large margin. Contrarily to previous definitions of face quality assessment [BJ18; 11; 15; Her+19] that states the face quality as a utility measure of a face image for an *arbitrary* face recognition model, our results show that it is highly beneficial to estimate the sample quality with regard to a specific (the deployed) face recognition model.

Related Work

Several standards have been proposed for insure face image quality by constraining the capture requirements, such as ISO/IEC 19794-5 [11] and ICAO 9303 [15]. In these standards, quality is divided into *image-based* qualities (such as pose, expression, illumination, occlusion) and *subject-based* quality measures (such as accessories). These mentioned standards influenced many face quality assessment approaches that have been proposed in the recent years. While the first solutions to face quality assessment focused on analytic

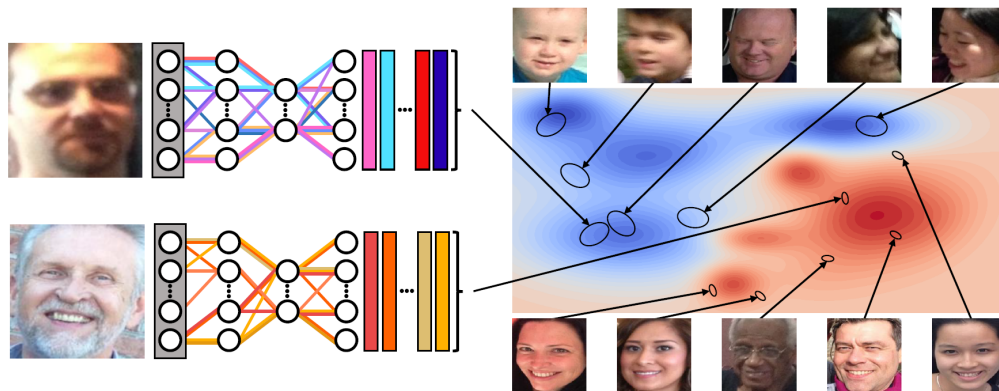


Figure A.1.: Visualization of the proposed unsupervised face quality assessment concept. We propose using the robustness of an image representation as a quality clue. Our approach defines this robustness based on the embedding variations of random subnetworks of a given face recognition model. An image that produces small variations in the stochastic embeddings (bottom left), demonstrates high robustness (red areas on the right) and thus, high image quality. Contrary, an image that produces high variations in the stochastic embeddings (top left) coming from random subnetworks, indicates a low robustness (blue areas on the right). Therefore, it is considered as low quality.

image quality factors, current solutions make use of the advances in supervised learning.

Approaches based on analytic image quality factors define quality metrics for facial asymmetries [Gao+07; Fer+12], propose vertical edge density as a quality metric to capture pose variations [Was+17], or measured in terms of luminance distortion in comparison to a known reference image [ZG17]. However, these approaches have to consider every possible factor manually, and since humans may not know the best characteristics for face recognition systems, more current research focus on learning-based approaches.

The transition to learning-based approaches include works that combine different analytical quality metrics with traditional machine learning approaches [Phi+13; Aba+14; HSM06; AHB12; DVS14].

End-to-end learning approaches for face quality assessment were first presented in 2011. Aggarwal et al. [Agg+11] proposed an approach for predicting the face recognition performance using a multi-dimensional scaling approach to map space characterization features to genuine scores. In [Won+11], a patch-based probabilistic image quality approach was designed that works on 2D discrete cosine transform features and trains a

Gaussian model on each patch. In 2015, a rank-based learning approach was proposed by Chen et al. [Che+15]. They define a linear quality assessment function with polynomial kernels and train weights based on a ranking loss. In [KLR15], face images assessment was performed based on objective and relative face image qualities. While the objective quality metric refers to objective visual quality in terms of pose, alignment, blurriness, and brightness, the relative quality metric represents the degree of mismatch between training face images and a test face image. Best-Rowden and Jain [BJ18] proposed an automatic face quality prediction approach in 2018. They proposed two methods for quality assessment of face images based on (a) human assessments of face image quality and (b) quality values from similarity scores. Their approach is based on support vector machines applied to deeply learned representations. In 2019, Hernandez-Ortega et al. proposed FaceQnet [Her+19]. This solution fine-tunes a face recognition neural network to predict face qualities in a regression task. Beside image quality estimation for face recognition, quality estimation has been also developed to predict soft-biometric decision reliability based on the investigated image [Ter+19d].

All previous face image quality assessment solutions require training data with artificial or manually labelled quality values. Human labelled data might transfer human bias into the quality predictions and does not take into account the potential biases of the biometric system. Moreover, humans might not know the best quality factors for a specific face recognition system. Artificially labelled quality values are created by investigating the relative performance of a face recognition system (represented by comparison scores). Consequently, the score might be heavily biased by low-quality samples.

The solution presented in this paper is based on our hypothesis that representation robustness is better suited as a quality metric, since it provides a measure for the quality of a single sample independently of others and avoids the use of misleading quality labels for training. This metric can intrinsically capture image acquisition conditions and factors of the face that are relevant for the used face recognition system. Furthermore, it is not affected by human bias, but takes into account the bias and the decision patterns of the used face embeddings.

Methodology

Face quality assessment aims at estimating the suitability of a face image for face recognition. The quality of a face image should indicate its expected recognition performance. In this work, we based our face image quality definition on the relative robustness of deeply learned embeddings of that image. Calculating the variations of embeddings coming from random subnetworks of a face recognition model, our solution defines the magnitude of these variations as a robustness measure, and thus, image quality. An illustration of this

methodology is shown in Figure A.2.

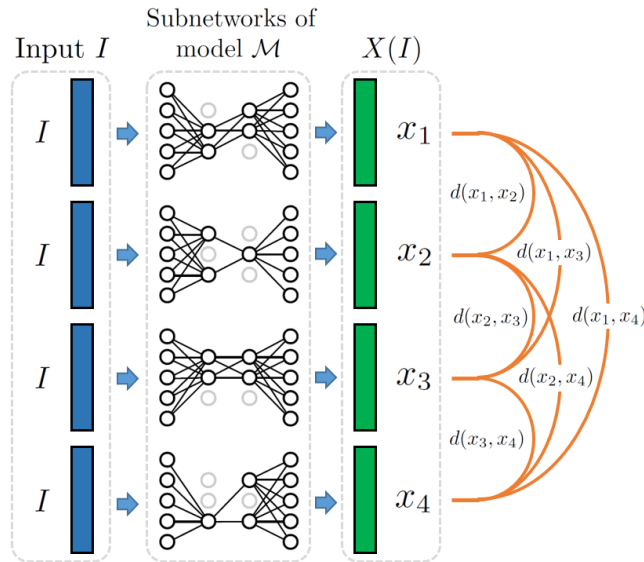


Figure A.2.: Illustration of the proposed methodology: an input I is forwarded to different random subnetworks of the used face recognition model \mathcal{M} . Each subnetwork produces a different stochastic embedding x_s . The variations between these embeddings are calculated using pairwise-distances and define the quality of I .

Sample-quality estimation More formally, our proposed solution predicts the face quality $Q(I)$ of a given face image I using a face recognition model \mathcal{M} . The face recognition model has to be trained with dropout and aims at extracting embeddings that are well identity-separated. To make a robustness-based quality estimation of I , $m = 100$ stochastic embeddings are generated from the model \mathcal{M} using stochastic forward passes with different dropout patterns. The choice for m is defined by the trade-off between time complexity and stability of the quality measure as described in Section A. Each stochastic forward pass applies a different dropout pattern (during prediction) producing a different subnetwork of \mathcal{M} . Each of these subnetworks generates different stochastic face embeddings x_s . These stochastic embeddings are collected in a set $X(I) = \{x_s\}_{s \in \{1, 2, \dots, m\}}$. We

define the face quality

$$q(X(I)) = 2 \sigma \left(- \frac{2}{m^2} \sum_{i < j} d(x_i, x_j) \right), \quad (\text{A.1})$$

of image I as the sigmoid of the negative mean euclidean distance $d(x_i, x_j)$ between all stochastic embeddings pairs $(x_i, x_j) \in X \times X$. The sigmoid function $\sigma(\cdot)$ ensures that $q \in [0, 1]$. Since Gal et al. [GG16] proofed that applying dropout repetitively on a network approximates the uncertainty of a Gaussian process [Ras06], the euclidean distance is a suitable choice for $d(x_i, x_j)$. A greater variation in the stochastic embedding set X indicate a low robustness of the representation and thus, a lower sample quality q . Lower variations in X indicate high robustness in the embedding space and is considered as a high sample quality q . The quality prediction strategy is summarized in Algorithm 5.

Algorithm 5 Stochastic Embedding Robustness - SER($I, \mathcal{M}, m = 100$)

Require: preprocessed input image I , NN-model \mathcal{M}

Ensure: quality value Q for input image I

- 1: $X \leftarrow$ empty list
 - 2: **for** $i \leftarrow 1, \dots, m$ **do**
 - 3: $x_i \leftarrow \mathcal{M}.pred(I, dropout = True)$
 - 4: $X = X.add(x_i)$
 - 5: **end for**
 - 6: $Q \leftarrow q(X)$
 - 7: **return** Q
-

Properties The aim of SER-FIQ is to estimate the face image quality from the perspective of utilisation in recognition tasks, which might be different than estimating the notion of image quality. An image that produces relatively stable identity-related embeddings despite various variations (here caused by dropout) is an image with high utilisation in a recognition task, given that the recognition network training aims at being robust against intra-identity variations.

Face recognition algorithms are trained with the aim of learning robust representations to increase inter-identity separability and decrease intra-identity separability. Assuming that a face recognition network is trained with dropout and the quality of a sample correlates with its embedding robustness, different subnetworks can be created from the basic model so that they possess different dropout patterns. The agreement between the subnetworks can be used to estimate the embedding robustness, and thus the quality.

If the m subnetworks produce similar outputs (high agreement), the variations over these random subnetworks (the stochastic embedding set X) are low. Consequently, the robustness of this embedding, and thus the quality of the sample, is high. Conversely, if the m subnetworks produce dissimilar representations (low agreement), the variations over the random subnetworks are high. Therefore, the robustness in the embedding space is low and the quality of the sample can be considered low as well.

Our approach has only one parameter m , the number of stochastic forward passes. This parameter can be interpreted as the number of steps in a Monte-Carlo simulation and controls the stability of the quality predictions. A higher m leads to more stable quality estimates. Since the computational time $t = \mathcal{O}(m^2)$ of our method grows quadratically with m , it should not be chosen too high. However, our method can compensate for this issue and can easily run in real-time, since it is highly parallelizable and the computational effort can be greatly reduced by repeating the stochastic forward passes only through the last layer(s) of the network.

In contrast to previous work, our solution does not require quality labels for training. Furthermore, if the deployed face recognition system was trained with dropout, the same network can be used for determining the embedding robustness and therefore, the sample quality. By doing so the training phase can be completely avoided and the quality predictions further captures the decision patterns and bias of the utilized face recognition model. Therefore, we highly recommend utilizing the deployed face recognition model for the quality assessment task.

Experimental Setup

Databases The face quality assessment experiments were conducted on three publicly available databases chosen to have variation in quality and to prove the generalization of our approach on multiple databases. The ColorFeret database [Phi+00] consists of 14,126 high-resolution face images from 1,199 different individuals. The data possess a variety of face poses and facial expressions under well-controlled conditions. The Adience dataset [EEH14] consists of 26,580 images from over 2,284 different subjects under unconstrained imaging conditions. Labeled Faces in the Wild (LFW) [Hua+07] contains 13,233 face images from 5749 identities. For both datasets, large variations in illumination, location, focus, blurriness, pose, and occlusion are included.

Evaluation metrics To evaluate the face quality assessment performance, we follow the methodology by Grother et al. [GT07] using error versus reject curves. These curves show a verification error-rate over the fraction of unconsidered face images. Based on the

predicted quality values, these unconsidered images are these with the lowest predicted quality and the error rate is calculated on the remaining images. Error versus reject curves indicates good quality estimation when the verification error decreases consistently when increasing the ratio of unconsidered images. In contrast to error versus quality-threshold curves, this process allows to fairly compare different algorithms for face quality assessment, since it is independent of the range of quality predictions. The curve was adapted in the approved ISO working item [20] and used in the literature [BJ18; TG15; GNH19a].

The face verification error rates within the error versus reject curves are reported in terms of false non-match rate (FNMR) at fixed false match rate (FMR) and as equal error rate (EER). The EER equals the FMR at the threshold where $FMR = 1 - FNMR$ and is well known as a single-value indicator of the verification performance. These error rates are specified for biometric verification evaluation in the international standard [16]. In our experiment, we report the face verification performance on three operating points to cover a wider range of potential applications. The face recognition performance is reported in terms of EER and FNMR at a FMR threshold of 0.01. The FNMR is also reported at 0.001 FMR threshold as recommended by the best practice guidelines for automated border control of Frontex [Fro17].

Face recognition networks To get face embedding from a given face image, the image is aligned, scaled, and cropped. The preprocessed image is passed to a face recognition models to extract the embeddings. In this work, we use two face recognition models, FaceNet [SKP15] and ArcFace [Den+19]. For FaceNet, the image is aligned, scaled, and cropped as described in [KS14]. To extract the embeddings, a pretrained model¹ was used. For ArcFace, the image preprocessing was done as described in [Guo+18] and a pretrained model² provided by the authors of ArcFace is used. Both models were trained on the MS1M database [Guo+16]. The output size is 128 for FaceNet and 512 for ArcFace. The identity verification is performed by comparing two embeddings using cosine-similarity.

On-top model preparation To apply our quality assessment methodology, a recognition model that was trained with dropout [Sri+14] is needed. Otherwise, a model containing dropout need to added on the top of the existing model. The direct way to apply our approach is to take a pretrained recognition model and repeat the stochastic forward passes only in the last layer(s) during prediction. This is even expected to reach a better

¹<https://github.com/davidsandberg/facenet>

²<https://github.com/deepinsight/insightface>

performance than training a custom network, because the verification decision, as well as the quality estimation decision, is done in a shared embedding space.

To demonstrate that our solution can be applied to any arbitrary face recognition system, in our experiments we show both approaches: (a) training a small custom network on top of the deployed face recognition system, which we will refer to as *SER-FIQ (on-top model)*, and (b) using the deployed model for the quality assessment, which we will refer to as *SER-FIQ (same model)*.

The structure of *SER-FIQ (on-top model)* was optimized such that its produced embeddings achieve a similar EER on ColorFeret as that of the FaceNet embeddings. It consists of five layers with $n_{emb}/128/512/n_{emb}/n_{ids}$ dimensions. The two intermediate layers have 128 and 512 dimensions. The last layer has the dimension equal to the number of training identities n_{ids} and is only needed during training. All layers contain dropout [Sri+14] with the recommended dropout probability $p_d = 0.5$ and a tanh activation. The training of the small custom network is done using the AdaDelta optimizer [Zei12] with a batchsize of 1024 over 100 epochs. Since the size of the in- and output layers (blue and green) of the networks differs dependent on the used face embeddings, a learning rate of $\alpha_{FN} = 10^{-1}$ was chosen for FaceNet and $\alpha_{AF} = 10^{-4}$ for the higher dimensional ArcFace embeddings. As the loss function, we used a simple binary cross-entropy loss on the classification of the training identities.

Investigations To investigate the generalization of face quality assessment performance, we conduct the experiments in a cross-database setting. The training is done on ColorFeret to make the models learn variations in a controlled environment. The testing is done on two unconstrained datasets, Adience and LFW. The embeddings used for the experiments are from the widely used FaceNet (2015) and recently published ArcFace (2019) models.

To put the experiments in a meaningful setting, we evaluated our approach in comparison to six baseline solutions. Three of these baselines are well-known no-reference image quality metrics from the computer vision community: Brisque [MMB12], Niqe [MSB13], Piqe [Ven+15]. The other three baselines are state-of-the-art face quality assessment approaches from academia and industry. COTS [Neu19] is an off the shelf industry product from Neurotechnology. We further compare our method with the two recent approaches from academia: the face quality assessment approach presented by Best-Rowden and Jain [BJ18] (2018) and FaceQnet [Her+19] (2019). Training the solution presented by Best-Rowden was done on ColorFeret following the procedure described in [BJ18]. The generated labels come from cosine similarity scores using the same embeddings as in the evaluation scenario. For all other baselines, pretrained models are utilized.

Our proposed methodology is presented in two settings, the *SER-FIQ (on-top model)* and

the *SER-FIQ (same model)*. *SER-FIQ (on-top model)* demonstrates that our unsupervised method can be applied to any face recognition system. *SER-FIQ (same model)* make use of the deployed face recognition model for quality assessment, to show the effect of capture its decision patterns for face quality assessment. In the latter case, we apply the stochastic forward passes only between the last two layers of the deployed face recognition network.

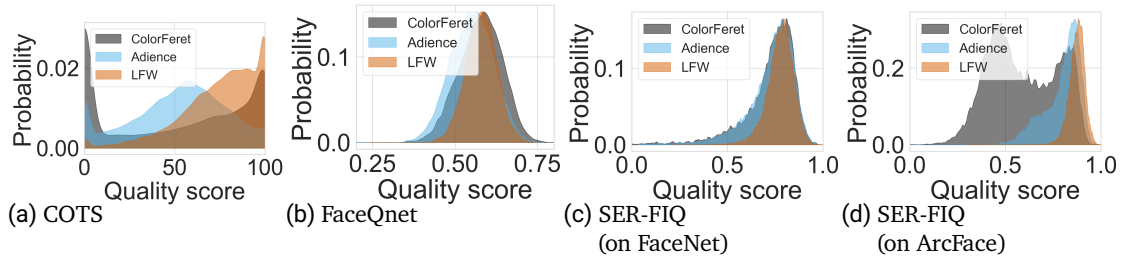


Figure A.3.: Face quality distributions of the used databases: Adience, LFW, and ColorFeret. The quality predictions were done using the pretrained models FaceQnet [Her+19], COTS [Neu19], and the proposed SER-FIQ (same model) based on FaceNet and ArcFace.

Database face quality rating To justify the choices of the used databases, Figure A.3 shows the face quality distributions of the databases using quality estimates from four pretrained face quality assessment models. ColorFeret was captured under well-controlled conditions and generally shows very high qualities. However, it contains non-frontal head poses and for COTS and SER-FIQ (on FaceNet) (Figure A.3a) this is considered as low image quality. Because of these controlled variations, we choose ColorFeret as the training database. Adience and LFW are unconstrained databases and for all quality measures, most face images are far away from perfect quality conditions. For this reason, we choose these databases for testing.

Results

The experiments are evaluated at three different operation points to investigate the face quality assessment performance over a wider spectrum of potential applications. Following the best practice guidelines for automated border control of the European Border and Coast Guard Agency Frontex [Fro17], Figure A.4 shows the face quality assessment performance at a FMR of 0.001. Figure A.6 presents the same at a FMR of 0.01 and Figure A.7 shows

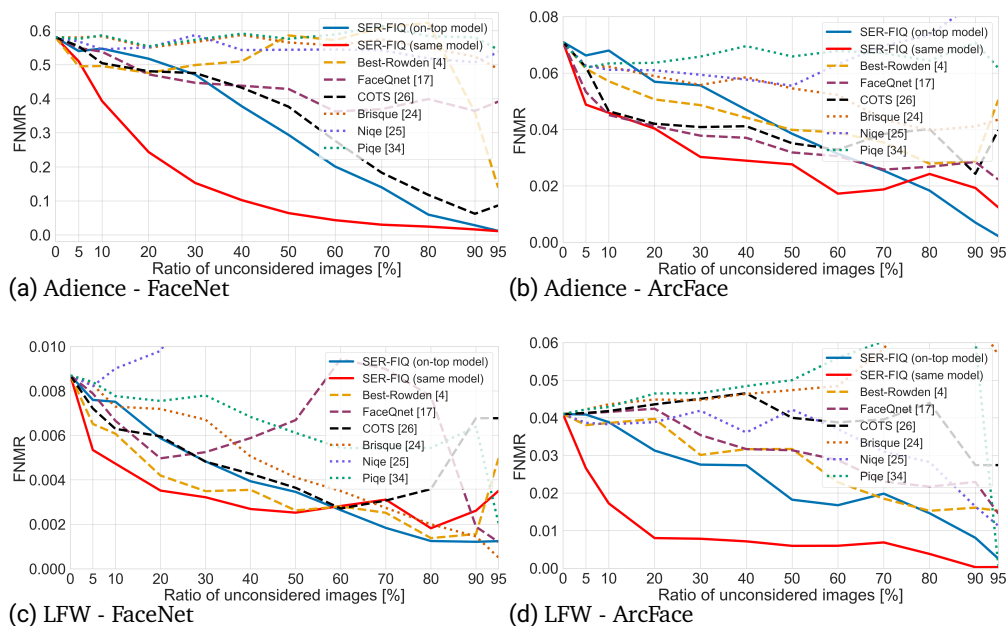


Figure A.4.: Face verification performance for the predicted face quality values. The curves show the effectiveness of rejecting low-quality face images in terms of FNMR at a threshold of 0.001 FMR. Figure A.4a and A.4b show the results for FaceNet and ArcFace embeddings on Adience. Figure A.4c and A.4d show the same on LFW.

the face quality assessment performance at the widely-used EER. Moreover, Figure A.5 shows sample images with their corresponding quality predictions. Since the statements about each tested face quality assessment approach are very similar over all experiments, we will make a discussion over each approach separately.

No-reference image quality approaches To understand the importance of different image quality measures for the task of face quality assessment, we evaluated three no-reference quality metrics Brisque [MMB12], Niqe [MSB13], Pique [Ven+15] (all represented as dotted lines). While in some evaluation scenarios the verification error decrease when the proportion of neglected images (low quality) is increased, in most cases they lead to an increased verification error. This demonstrates that image quality alone is not suitable for generalized face quality estimation. Factors of the face (such as pose,

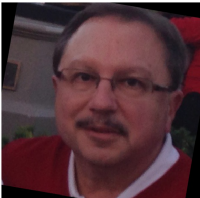
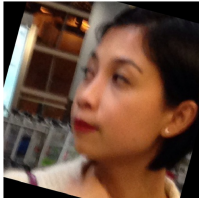

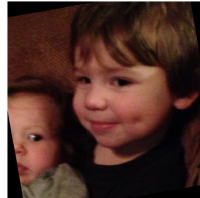
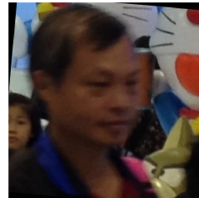
				
SER-FIQ: 0.86	SER-FIQ: 0.57	SER-FIQ: 0.08	SER-FIQ: 0.03	SER-FIQ: 0.10
Best-Rowden: 0.74	Best-Rowden: 0.82	Best-Rowden: 0.58	Best-Rowden: 0.79	Best-Rowden: 0.56
FaceQnet: 0.94	FaceQnet: 0.71	FaceQnet: 0.72	FaceQnet: 0.73	FaceQnet: 0.70
COTS: 0.74	COTS: 0.00	COTS: 0.18	COTS: 0.07	COTS: 0.35

Figure A.5.: Sample face images from Adience with the corresponding quality predictions from four face quality assessment methods. SER-FIQ refers to our same model approach based on ArcFace.

occlusions, and expressions) and model biases are not covered by these algorithms and might play an important role for face quality assessment.

Best-Rowden The proposed approach from Best-Rowden and Jain [BJ18] works well in most scenarios and reaches a top-rank performance in some minor cases (e.g. LFW with FaceNet features). However, it shows instabilities that can lead to highly wrong quality predictions. This can be observed well on the Adience dataset using FaceNet embeddings, see Figure A.4a and A.6a. These mispredictions might be explained by the ColorFerret training data that does not contain all important quality factors for a given face embedding. On the other hand, these quality factors are generally unknown and thus, training data should never be considered to be covering all factors.

FaceQnet FaceQnet [Her+19], proposed by Hernandez-Ortega et al., shows a suitable face quality assessment behaviour in most cases. In comparison with other face quality assessment approaches, it only shows a mediocre performance. Although FaceQnet was trained on labels coming from the same FaceNet embeddings as in our evaluation setting, it often fails in predicting well-suited quality labels on these embeddings, e.g. in Figure A.4c on LFW. Also on Adience (e.g. Figure A.6a and A.7a), the performance plot shows a U-shape that demonstrates that the algorithm can not distinguish well between medium and higher quality face images. Since the method is trained on the same features, these FaceNet-related instabilities might result from overfitting.

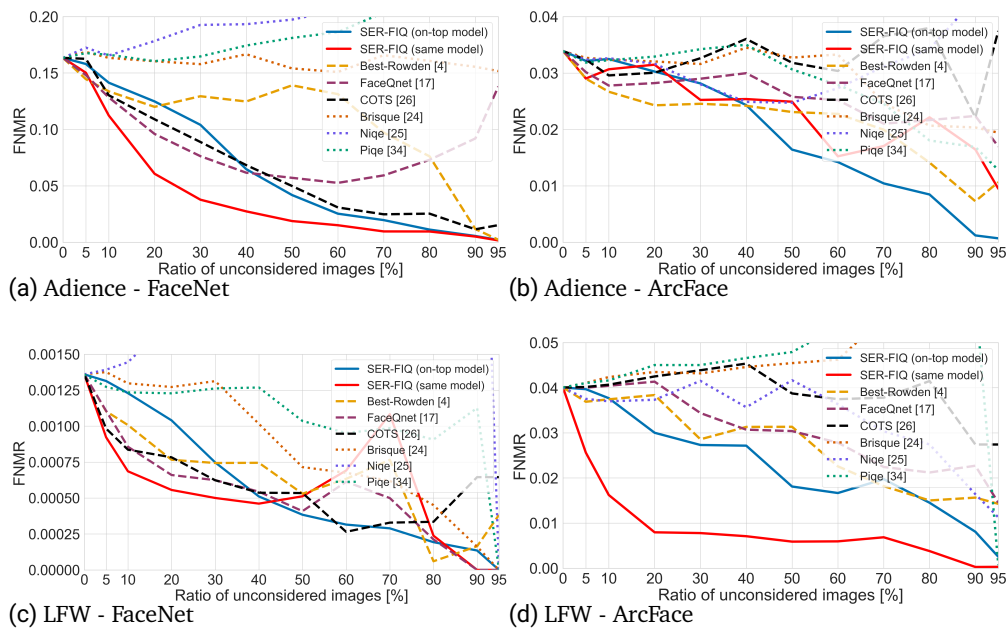


Figure A.6.: Face verification performance for the predicted face quality values. The curves show the effectiveness of rejecting low-quality face images in terms of FNMR at a threshold of 0.01 FMR. Figure A.6a and A.6b show the results for FaceNet and ArcFace embeddings on Adience. Figure A.6c and A.6d show the same on LFW.

COTS The industry baseline COTS [Neu19] from Neurotechnology generally shows a good face quality assessment when the used face recognition system is based on FaceNet features. Specifically on LFW (see Figure A.4c, A.6c, and A.7c) a small U-shape can be observed similar to FaceQnet. While it shows a good performance using FaceNet embeddings, the face quality predictions using the more recent ArcFace embeddings are of no significance (see Figure A.4b, A.4d, A.6b, A.6d, A.7b, and A.7d). Here, rejecting face images with low predicted face quality does not improve the face recognition performance. Since no information about the inner workflow is given, it can be assumed that their method is optimized to more traditional face embeddings, such as FaceNet. More recent embeddings, such as ArcFace, are probably intrinsically robust to the quality factors that COTS is trained on.

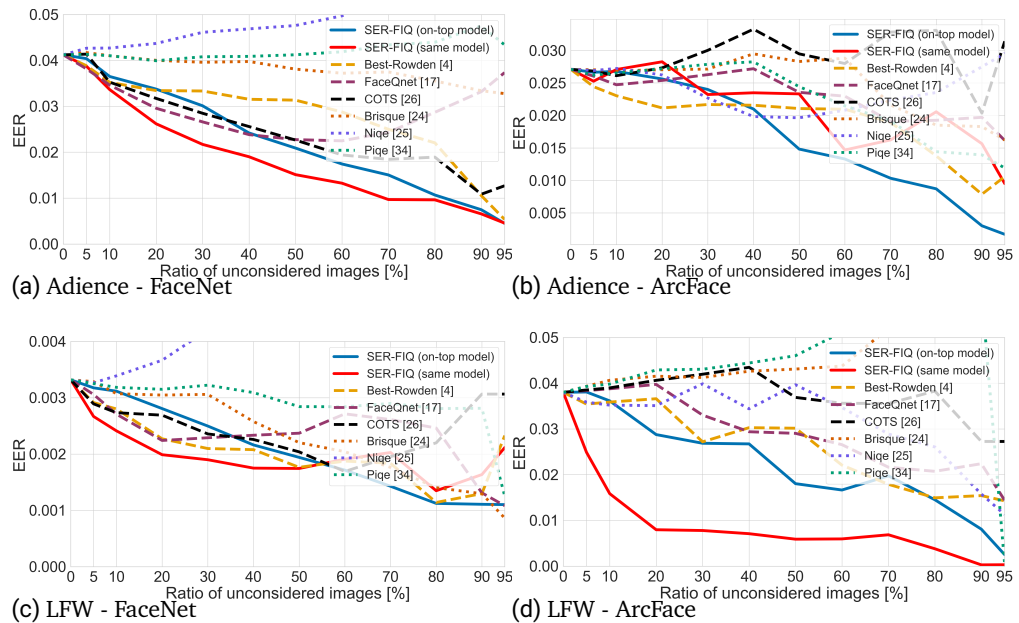


Figure A.7.: The face verification performance given as EER for the predicted face quality values. The curves show the effectiveness of rejecting low-quality face images in terms of EER. Figure A.7a and A.7b show the results for FaceNet and ArcFace embeddings on Adience. Figure A.7c and A.7d show the some on LFW.

SER-FIQ (on-top model) On the contrary to the discussed supervised methods, our proposed unsupervised solution that builds on training a small custom face recognition network shows a stable performance in all investigated scenarios (Figure A.4, A.6, and A.7). Furthermore, our solution is always close to the top performance and outperforms all baseline approaches in the majority of the scenarios, e.g. in Figure A.4a, A.4d, A.6a, A.6b, A.6d, A.7a, A.7b, and A.7d. Our method proved to be particularly effective in combination with recent ArcFace embeddings (see Figures A.6b, A.6d, A.7b, and A.7d). The unsupervised nature of our solution seems to be a more accurate and more stable strategy.

SER-FIQ (same model) Our method that avoids training by utilizing the deployed face recognition systems is build on the hypotheses that face quality assessment should aim at estimating the sample quality of a *specific* face recognition model. This way it adapts

to the models' decision patterns and can predict the suitability of face sample more accurately. The effect of this adaptation can be seen clearly in nearly all evaluated cases (see Figure A.4, A.6, and A.7). It outperforms all baseline approaches by a large margin and demonstrates an even stronger performance at small FMR (see Figures A.4a, A.4b, A.4c, and A.4d at the Frontex recommended FMR of 0.001). This demonstrates the benefit of focusing on the face quality assessment to a specific (the deployed) face recognition model.

Conclusion

Face quality assessment aims at predicting the suitability of face images for face recognition systems. Previous works provided supervised models for this task based on inaccurate quality labels with only limited consideration of the decision patterns of the deployed face recognition system. In this work, we solved these two gaps by proposing a novel unsupervised face quality assessment methodology that is based on a face recognition model trained with dropout. Measuring the embeddings variations generated from random subnetworks of the face recognition model, the representation robustness of a sample and thus, the sample's quality is determined. To evaluate a generalized face quality assessment performance, the experiments were conducted using three publicly available databases in a cross-database evaluation setting. We compared our solution on two different face embeddings against six state-of-the-art approaches from academia and industry. The results showed that our proposed approach outperformed all other approaches in the majority of the investigated scenarios. It was the only solution that showed a consistently stable performance. By using the deployed face recognition model for verification and the proposed quality assessment methodology, we avoided the training phase completely and further outperformed all baseline approaches by a large margin. Our approach is characterized by high parallelizability, its easy integration into existing face recognition systems, and it is not limited to face biometrics.

B. Publications and Talks

This work is partially based on the following publications and talks.

B.1. Publications

First-Author Publications

- [Ter+20a] Philipp Terhörst, Daniel Fährmann, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Beyond Identity: What Information Is Stored in Biometric Face Templates?” In: *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, USA, Sept 28 - Oct 1, 2020*. IEEE, 2020.
- [Ter+20b] Philipp Terhörst, Marco Huber, Naser Damer, Peter Rot, Florian Kirchbuchner, Vitomir Struc, and Arjan Kuijper. “Privacy Evaluation Protocols for the Evaluation of Soft-Biometric Privacy-Enhancing Technologies”. In: *2020 International Conference of the Biometrics Special Interest Group, BIOSIG 2020, Darmstadt, Germany, September 16-18, 2020*. IEEE, 2020.
- [Ter+20c] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Face Quality Estimation and Its Correlation to Demographic and Non-Demographic Bias in Face Recognition”. In: *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, USA, Sept 28 - Oct 1, 2020*. IEEE, 2020.
- [Ter+20d] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Post-comparison mitigation of demographic bias in face recognition using fair score normalization”. In: *Pattern Recognition Letters* 140 (2020), pp. 332–338. ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2020.11.007>. URL: <http://www.sciencedirect.com/science/article/pii/S0167865520304128>.

-
- [Ter+20e] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 5650–5659. DOI: 10.1109/CVPR42600.2020.00569. URL: <https://doi.org/10.1109/CVPR42600.2020.00569>.
- [Ter+20f] Philipp Terhörst, Kevin Riehl, Naser Damer, Peter Rot, Blaz Bortolato, Florian Kirchbuchner, Vitomir Struc, and Arjan Kuijper. “PE-MIU: A Training-Free Privacy-Enhancing Face Recognition Approach Based on Minimum Information Units”. In: *IEEE Access* (2020).
- [Ter+20g] Philipp Terhörst, Mai Ly Tran, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Comparison-Level Mitigation of Ethnic Bias in Face Recognition”. In: *2020 International Workshop on Biometrics and Forensics, IWBF 2020, Porto, Portugal, April 29-30, 2020*. IEEE, 2020.
- [Ter+19a] Philipp Terhörst, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Suppressing Gender and Age in Face Templates Using Incremental Variable Elimination”. In: *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*. IEEE, 2019, pp. 1–8. DOI: 10.1109/ICB45273.2019.8987331. URL: <https://doi.org/10.1109/ICB45273.2019.8987331>.
- [Ter+19b] Philipp Terhörst, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations”. In: *Appl. Intell.* 49.8 (2019), pp. 3043–3060. DOI: 10.1007/s10489-019-01432-5. URL: <https://doi.org/10.1007/s10489-019-01432-5>.
- [Ter+19c] Philipp Terhörst, Marco Huber, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Multi-Algorithmic Fusion for Reliable Age and Gender Estimation from Face Images”. In: *22st International Conference on Information Fusion, FUSION 2019, Ottawa, Canada, July 2-5, 2019*. IEEE, 2019.
- [Ter+19d] Philipp Terhörst, Marco Huber, Jan Niklas Kolf, Ines Zelch, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Reliable Age and Gender Estimation from Face Images: Stating the Confidence of Model Predictions”. In: *10th IEEE International Conference on Biometrics Theory, Applications and*

-
- Systems, BTAS 2019, Tampa, Florida, USA, September 23-26, 2019*. IEEE, 2019.
- [Ter+18a] Philipp Terhörst, Naser Damer, Andreas Braun, and Arjan Kuijper. “Deep and Multi-Algorithmic Gender Classification of Single Fingerprint Minutiae”. In: *21st International Conference on Information Fusion, FUSION 2018, Cambridge, UK, July 10-13, 2018*. IEEE, 2018, pp. 2113–2120. DOI: 10.23919/ICIF.2018.8455803. URL: <https://doi.org/10.23919/ICIF.2018.8455803>.
- [Ter+18b] Philipp Terhörst, Naser Damer, Andreas Braun, and Arjan Kuijper. “Minutiae-Based Gender Estimation for Full and Partial Fingerprints of Arbitrary Size and Shape”. In: *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part I*. Ed. by C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler. Vol. 11361. Lecture Notes in Computer Science. Springer, 2018, pp. 171–186. DOI: 10.1007/978-3-030-20887-5_11. URL: https://doi.org/10.1007/978-3-030-20887-5%5C_11.
- [Ter+18c] Philipp Terhörst, Naser Damer, Andreas Braun, and Arjan Kuijper. “What can a single minutia tell about gender?” In: *2018 International Workshop on Biometrics and Forensics, IWBF 2018, Sassari, Italy, June 7-8, 2018*. IEEE, 2018, pp. 1–7. DOI: 10.1109/IWBF.2018.8401554. URL: <https://doi.org/10.1109/IWBF.2018.8401554>.

Co-Author Publications

- [Bor+20] Blaz Bortolato, Marija Ivanovska, Peter Rot, Janez Krizaj, Philipp Terhörst, Naser Damer, Peter Peer, and Vitomir Struc. “Learning privacy-enhancing face representations through feature disentanglement”. In: *15th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2020, Buenos Aires, Argentina, Nov 16-22, 2020*. IEEE, 2020.
- [Bou+19] Fadi Boutros, Naser Damer, Philipp Terhörst, Florian Kirchbuchner, and Arjan Kuijper. “Exploring the Channels of Multiple Color Spaces for Age and Gender Estimation from Face Images”. In: *22st International Conference on Information Fusion, FUSION 2019, Ottawa, Canada, July 2-5, 2019*. IEEE, 2019.

-
- [Dam+19] Naser Damer, Alexandra Mosegui Saladie, Steffen Zienert, Yaza Wainakh, Philipp Terhörst, Florian Kirchbuchner, and Arjan Kuijper. “To Detect or not to Detect: The Right Faces to Morph”. In: *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*. IEEE, 2019, pp. 1–8. DOI: 10.1109/ICB45273.2019.8987316. URL: <https://doi.org/10.1109/ICB45273.2019.8987316>.
- [Dam+18a] Naser Damer, Viola Boller, Yaza Wainakh, Fadi Boutros, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “Detecting Face Morphing Attacks by Analyzing the Directed Distances of Facial Landmarks Shifts”. In: *Pattern Recognition - 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings*. Ed. by Thomas Brox, Andrés Bruhn, and Mario Fritz. Vol. 11269. Lecture Notes in Computer Science. Springer, 2018, pp. 518–534. DOI: 10.1007/978-3-030-12939-2_36. URL: https://doi.org/10.1007/978-3-030-12939-2%5C_36.
- [Dam+18b] Naser Damer, Fadi Boutros, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “P-Score: Performance Aligned Normalization and an Evaluation in Score-Level Multi-Biometric Fusion”. In: *26th European Signal Processing Conference, EUSIPCO 2018, Roma, Italy, September 3-7, 2018*. IEEE, 2018, pp. 1402–1406. DOI: 10.23919/EUSIPCO.2018.8553553. URL: <https://doi.org/10.23919/EUSIPCO.2018.8553553>.
- [Dam+18c] Naser Damer, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “Fingerprint and Iris Multi-Biometric Data Indexing and Retrieval”. In: *21st International Conference on Information Fusion, FUSION 2018, Cambridge, UK, July 10-13, 2018*. IEEE, 2018, pp. 2083–2090. DOI: 10.23919/ICIF.2018.8455390. URL: <https://doi.org/10.23919/ICIF.2018.8455390>.
- [Dam+18d] Naser Damer, Yaza Wainakh, Viola Boller, Sven von den Berken, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “CrazyFaces: Unassisted Circumvention of Watchlist Face Identification”. In: *9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018*. IEEE, 2018, pp. 1–9. DOI: 10.1109/BTAS.2018.8698557. URL: <https://doi.org/10.1109/BTAS.2018.8698557>.
- [Dam+17a] Naser Damer, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “Efficient, Accurate, and Rotation-Invariant Iris Code”. In: *IEEE Signal Process. Lett.*

24.8 (2017), pp. 1233–1237. DOI: 10.1109/LSP.2017.2719282. URL: <https://doi.org/10.1109/LSP.2017.2719282>.

[Dam+17b] Naser Damer, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “General borda count for multi-biometric retrieval”. In: *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*. IEEE, 2017, pp. 420–428. DOI: 10.1109/BTAS.2017.8272726. URL: <https://doi.org/10.1109/BTAS.2017.8272726>.

[Dam+17c] Naser Damer, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “Indexing of Single and Multi-instance Iris Data Based on LSH-Forest and Rotation Invariant Representation”. In: *Computer Analysis of Images and Patterns - 17th International Conference, CAIP 2017, Ystad, Sweden, August 22-24, 2017, Proceedings, Part II*. Ed. by Michael Felsberg, Anders Heyden, and Norbert Krüger. Vol. 10425. Lecture Notes in Computer Science. Springer, 2017, pp. 190–201. DOI: 10.1007/978-3-319-64698-5_17. URL: https://doi.org/10.1007/978-3-319-64698-5_17.

Under Review

[Med+21] Blaz Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter Schreier, Arun Ross, Peter Peer, and Vitomir Struc. “Privacy-Enhancing Face Biometrics: A Comprehensive Survey”. In: *IEEE Trans. Inf. Forensics Secur.* (2021). Under review.

[Ter+21a] Philipp Terhörst, Andre Boller, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “MiDeCon: Unsupervised and Accurate Fingerprint and Minutia Quality Assessment based on Minutia Detection Confidence”. In: *2021 IEEE International Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*. Under review. IEEE, 2021.

[Ter+21b] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales, Julian Fierrez, and Arjan Kuijper. “A Comprehensive Study on Face Recognition Biases Beyond Demographics”. In: *IEEE-Transactions of Technology and Society* (2021). Under review. URL: <https://arxiv.org/abs/2012.01030>.

[Ter+20a] Philipp Terhörst, Daniel Fährmann, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “On Soft-Biometric Information Stored in Biometric Face Embeddings”. In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2020). Under review.

-
- [Ter+20b] Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “MAAD-Face: A Massively Annotated Attribute Dataset of Face Images”. In: *IEEE Trans. Inf. Forensics Secur.* (2020). Under review. URL: <https://arxiv.org/abs/2012.01030>.
- [Ter+20c] Philipp Terhörst, Marco Huber, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Unsupervised Enhancement of Soft-biometric Privacy with Negative Face Recognition”. In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2020). Under review. URL: <https://arxiv.org/abs/2002.09181>.

For an updated publication list, please check my Google Scholar profile: https://scholar.google.de/citations?hl=en&user=4iERqCYAAAAJ&view_op=list_works&sortby=pubdate

B.2. Invited Talks

1. Philipp Terhörst: *What information is stored in face templates? And how does it relate to fairness.* EAB Seminar on "Demographic fairness in biometric systems", European Association for Biometrics, online, 09 March 2021.
2. Philipp Terhörst: *Unsupervised Estimation of Face Image Quality.* German TeleTrust Biometrics Working Group, European Association for Biometrics, online, 14 September 2020.
3. Philipp Terhörst: *Soft-Biometric Privacy: Attacks and Circumventions.* 10th Norwegian Biometrics Laboratory Annual Workshop, Norwegian Biometrics Laboratory at NTNU in Gjøvik, Norway, 4 March 2020.

C. Supervising Activities

The following list summarizes the supervising activities of the author. The results of these works were partially used in the thesis.

Bachelor Thesis

1. André Boller, Prof. Dr. Arjan Kuijper (supervising professor), and Philipp Terhörst (supervisor). Towards unsupervised fingerprint image quality assessment. TU Darmstadt, 2021 (expected graduation time).
2. Jonas Henry Grebe, Prof. Dr. Arjan Kuijper (supervising professor), and Philipp Terhörst (supervisor). Anomaly-based Face Search. TU Darmstadt, 2020.
3. Paul Frederik Franz Ludwig Wochner, Prof. Dr. Thomas Walther (supervising professor), and Philipp Terhörst (supervisor). How Do Demographic Soft-Biometric Attributes Affect Kinship Verification? TU Darmstadt, 2019.

Master Thesis

1. Marco Huber, Prof. Dr. Arjan Kuijper (supervising professor), and Philipp Terhörst (supervisor). Explainable Face Image Quality Assessment. TU Darmstadt, 2021 (expected graduation time).
2. Mai Ly Tran, Prof. Dr. Arjan Kuijper (supervising professor), and Philipp Terhörst (supervisor). Mitigating Ethnic Bias in Face Recognition Models through Fair Template Comparison. TU Darmstadt, 2019.
3. Daniel Fährmann, Prof. Dr. Arjan Kuijper (supervising professor), and Philipp Terhörst (supervisor). Enhancing the privacy of face recognition and its representations. TU Darmstadt, 2019.

Miscellaneous

The author also supervised several student projects from lectures and interships, such as Bachelorpraktikum or Praktikum Visual Computing.

D. Curriculum Vitae

Personal Data

Name Philipp Terhörst
Birth date 13.02.1992

Professional Experience

12/2017 – today Research Scientist at Fraunhofer IGD in Darmstadt, Germany
05/2017 – 11/2017 Student Associate at Fraunhofer IGD in Darmstadt, Germany

Education

Academic Qualifications

10/2014 – 06/2017 M.Sc. Physics at Technische Universität of Darmstadt, Germany
10/2011 – 10/2014 B.Sc. Physics at Technische Universität of Darmstadt, Germany

Awards

2020 IJCB 2020 Qualcomm PC Chairs Choice Best Student Paper Award
International Joint Conference on Biometrics (IJCB) 2020
"Beyond Identity: What Information Is Stored in Biometric Face
Templates?" (IEEE IJCB 2020)

2020 IJCB 2020 Audience's Choice Day 1 Presentation Award
International Joint Conference on Biometrics (IJCB) 2020
"Beyond Identity: What Information Is Stored in Biometric Face
Templates?" (IEEE IJCB 2020)

2020 Best Paper Award - Impact on Society
Fraunhofer IGD and the Visual Computing Groups of TU Darmstadt

	"Detecting Face Morphing Attacks by Analyzing the Directed Distances of Facial Landmarks Shifts" (GCPR 2018)
2020	EAB Biometrics Industry Award 2020 European Association for Biometrics "Mitigating Soft-Biometrics Driven Privacy and Bias Concerns in Face Recognition Systems"
2020	EAB Biometrics Best Speaker Award 2020 European Association for Biometrics "Mitigating Soft-Biometrics Driven Privacy and Bias Concerns in Face Recognition Systems"
2020	Honorary Mention for Most Captivating Poster Presentation 2020 International Workshop on Biometrics and Forensics "Comparison-Level Mitigation of Ethnic Bias in Face Recognition" (IEEE IWBF 2020)
2019	Best Paper Award - Impact on Society Fraunhofer IGD and the Visual Computing Groups of TU Darmstadt "CrazyFaces: Unassisted Circumvention of Watchlist Face Identification" (IEEE BTAS 2018)
2018	Best Paper Award Nomination - Impact on Business Fraunhofer IGD and the Visual Computing Groups of TU Darmstadt "Efficient, Accurate, and Rotation-Invariant Iris Code" (IEEE SPL 2017)
2017	Best Thesis Award Fraunhofer IGD and the Visual Computing Groups of TU Darmstadt "Indexing of Multi-biometric Databases: Fast and Accurate Biometric Search"

Miscellaneous Experience

Reviewer

- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
- IEEE IEEE Transactions on Image Processing (TIP)
- Pattern Recognition (PR)
- IEEE Access

-
- IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)
 - IEEE International Conference of the Biometrics Special Interest Group (BIOSIG)
 - IAPR/IEEE International Workshop on Biometrics and Forensics (IWBF)
 - IAPR International Conference On Biometrics (ICB)

Teaching Activities

2018–2021 Ambient Intelligence at Technical University of Darmstadt
Focusing on Machine Learning and Identification in AmI Systems

Trainings

01/2019 – 12/2020 Participant of the Software Campus Project
EIT ICT Labs Germany GmbH
A management program of the Federal Ministry of Education and
Research (BMBF)

11/2020 Design thinking, Online
Carl Zeiss AG

10/2020 Communicate Convincingly, Online
Rohde & Schwarz GmbH & Co KG

08/2020 Change management, Ditzingen, Germany
Trumpf Group

05/2019 Welcome on stage, Darmstadt, Germany
Software AG

04/2019 Work smart - not hard! Self-management training, Munich, Germany
Rohde & Schwarz GmbH & Co KG

02/2019 Form and lead high performance teams, Stuttgart, Germany
Holtzbrinck Publishing Group

10/2018 Professional presentations, Darmstadt, Germany Fraunhofer IGD

10/2018 Communicate winningly and effectively, Darmstadt, Germany Fraun-
hofer IGD

08/2018 Project management, Darmstadt, Germany Fraunhofer IGD

Memberships

- German Association for Pattern Recognition (DAGM)
- International Association for Pattern Recognition (IAPR)
- European Association for Computer Graphics (EG)
- Institute of Electrical and Electronics Engineers (IEEE)
- The Computer Vision Foundation (CVF)

Bibliography

- [06] *Information technology - Biometric performance testing and reporting - Part 1: Principles and framework*. Standard. International Organization for Standardization, 2006.
- [11] *Information technology – Biometric data interchange formats – Part 5: Face image data*. Standard. International Organization for Standardization, Nov. 2011.
- [12] *Information Technology Vocabulary Part 37: Biometrics*. Standard. International Organization for Standardization, 2012.
- [15] *Machine Readable Travel Documents*. Standard. International Civil Aviation Organization, 2015.
- [16] *ISO/IEC 19795-1:2006 Information technology — Biometric performance testing and reporting*. Standard. International Organization for Standardization, 2016.
- [20] *ISO/IEC AWI 24357: Performance evaluation of face image quality algorithms*. Standard. International Organization for Standardization, 2020.
- [AB20] Vitor Albiero and Kevin W. Bowyer. “Is Face Recognition Sexist? No, Gendered Hairstyles and Biology Are”. In: *CoRR* abs/2008.06989 (2020). arXiv: 2008.06989. URL: <https://arxiv.org/abs/2008.06989>.
- [Aba+14] Ayman Abaza, Mary Ann F. Harrison, Thirimachos Bourlai, and Arun Ross. “Design and evaluation of photometric image quality measures for effective face recognition”. In: *IET Biom.* 3.4 (2014), pp. 314–324. DOI: 10.1049/iet-bmt.2014.0022. URL: <https://doi.org/10.1049/iet-bmt.2014.0022>.

-
- [Agg+08] Gaurav Aggarwal, Nalini K. Ratha, Ruud M. Bolle, and Rama Chellappa. “Multi-biometric cohort analysis for biometric fusion”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*. IEEE, 2008, pp. 5224–5227. DOI: 10.1109/ICASSP.2008.4518837. URL: <https://doi.org/10.1109/ICASSP.2008.4518837>.
- [Agg+11] Gaurav Aggarwal, Soma Biswas, Patrick J. Flynn, and Kevin W. Bowyer. “Predicting performance of face recognition systems: An image characterization approach”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2011, Colorado Springs, CO, USA, 20-25 June, 2011*. IEEE Computer Society, 2011, pp. 52–59. DOI: 10.1109/CVPRW.2011.5981784. URL: <https://doi.org/10.1109/CVPRW.2011.5981784>.
- [AHB12] Ayman Abaza, Mary Ann F. Harrison, and Thirimachos Bourlai. “Quality metrics for practical face recognition”. In: *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012, Tsukuba, Japan, November 11-15, 2012*. IEEE Computer Society, 2012, pp. 3103–3107. URL: <http://ieeexplore.ieee.org/document/6460821/>.
- [AHP06] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. “Face Description with Local Binary Patterns: Application to Face Recognition”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 28.12 (2006), pp. 2037–2041. DOI: 10.1109/TPAMI.2006.244. URL: <https://doi.org/10.1109/TPAMI.2006.244>.
- [AK20] Hiba Basim Alwan and Ku Ruhana Ku-Mahamud. “Cancellable Face Biometrics Template Using AlexNet”. In: *Applied Computing to Support Industry: Innovation and Technology*. Ed. by Mohammed I. Khalaf, Dhiya Al-Jumeily, and Alexei Lisitsa. Cham: Springer International Publishing, 2020, pp. 336–348.
- [AZB20] Vitor Albiero, Kai Zhang, and Kevin W. Bowyer. “How Does Gender Balance In Training Data Affect Face Recognition Accuracy?” In: *CoRR* abs/2002.02934 (2020). arXiv: 2002.02934. URL: <https://arxiv.org/abs/2002.02934>.
- [AZN18] Mohsan S. Alvi, Andrew Zisserman, and Christoffer Nellåker. “Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings”. In: *Computer Vision - ECCV 2018 Workshops - Munich*,

-
- Germany, September 8-14, 2018, *Proceedings, Part I*. Ed. by Laura Leal-Taixé and Stefan Roth. Vol. 11129. Lecture Notes in Computer Science. Springer, 2018, pp. 556–572. DOI: 10.1007/978-3-030-11009-3_34. URL: https://doi.org/10.1007/978-3-030-11009-3%5C_34.
- [Bal+20] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. “Towards causal benchmarking of bias in face analysis algorithms”. In: *CoRR abs/2007.06570* (2020). arXiv: 2007.06570. URL: <https://arxiv.org/abs/2007.06570>.
- [BDM12] Rainer Burkard, Mauro Dell’Amico, and Silvano Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, 2012. DOI: 10.1137/1.9781611972238. eprint: <https://epubs.siam.org/doi/pdf/10.1137/1.9781611972238>. URL: <https://epubs.siam.org/doi/abs/10.1137/1.9781611972238>.
- [Ber+17] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. “A Convex Framework for Fair Regression”. In: *CoRR abs/1706.02409* (2017). arXiv: 1706.02409. URL: <http://arxiv.org/abs/1706.02409>.
- [BG18] Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91. URL: <http://proceedings.mlr.press/v81/buolamwini18a.html>.
- [BHK97] Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 19.7 (1997), pp. 711–720. DOI: 10.1109/34.598228. URL: <https://doi.org/10.1109/34.598228>.
- [BJ18] Lacey Best-Rowden and Anil K. Jain. “Learning Face Image Quality From Human Assessments”. In: *IEEE Trans. Inf. Forensics Secur.* 13.12 (2018), pp. 3064–3077. DOI: 10.1109/TIFS.2018.2799585. URL: <https://doi.org/10.1109/TIFS.2018.2799585>.
- [Bor+20] Blaz Bortolato, Marija Ivanovska, Peter Rot, Janez Krizaj, Philipp Terhörst, Naser Damer, Peter Peer, and Vitomir Struc. “Learning privacy-enhancing face representations through feature disentanglement”. In: *15th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2020, Buenos Aires, Argentina, May 18-22, 2020*. IEEE, 2020.

-
- [Bou+19] Fadi Boutros, Naser Damer, Philipp Terhörst, Florian Kirchbuchner, and Arjan Kuijper. “Exploring the Channels of Multiple Color Spaces for Age and Gender Estimation from Face Images”. In: *22th International Conference on Information Fusion, FUSION 2019, Ottawa, ON, Canada, July 2-5, 2019*. IEEE, 2019, pp. 1–8. URL: <https://ieeexplore.ieee.org/document/9011383>.
- [BPJ98] Ruud Bolle, Sharath Pankanti, and Anil K. Jain. *Biometrics, Personal Identification in Networked Society: Personal Identification in Networked Society*. USA: Kluwer Academic Publishers, 1998. ISBN: 0792383451.
- [Cao+10] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. “Face recognition with learning-based descriptor”. In: *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*. IEEE Computer Society, 2010, pp. 2707–2714. DOI: 10.1109/CVPR.2010.5539992. URL: <https://doi.org/10.1109/CVPR.2010.5539992>.
- [Cao+18] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. “VGGFace2: A Dataset for Recognising Faces across Pose and Age”. In: *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi’an, China, May 15-19, 2018*. IEEE Computer Society, 2018, pp. 67–74. DOI: 10.1109/FG.2018.00020. URL: <https://doi.org/10.1109/FG.2018.00020>.
- [Cas+17] Aniello Castiglione, Kim-Kwang Raymond Choo, Michele Nappi, and Fabio Narducci. “Biometrics in the Cloud: Challenges and Research Opportunities”. In: *IEEE Cloud Comput.* 4.4 (2017), pp. 12–17. DOI: 10.1109/MCC.2017.3791012. URL: <https://doi.org/10.1109/MCC.2017.3791012>.
- [Cav+19] Jacqueline G. Cavazos, P. Jonathon Phillips, Carlos Domingo Castillo, and Alice J. O’Toole. “Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?” In: *CoRR* abs/1912.07398 (2019). arXiv: 1912.07398. URL: <http://arxiv.org/abs/1912.07398>.
- [CBF05] Kyong Jin Chang, Kevin W. Bowyer, and Patrick J. Flynn. “Effects on facial expression in 3D face recognition”. In: *Biometric Technology for Human Identification II*. Ed. by Anil K. Jain and Nalini K. Ratha. Vol. 5779. International Society for Optics and Photonics. SPIE, 2005, pp. 132–143. DOI: 10.1117/12.604171. URL: <https://doi.org/10.1117/12.604171>.

-
- [CBF06] Kyong I. Chang, Kevin W. Bowyer, and Patrick J. Flynn. “Multiple Nose Region Matching for 3D Face Recognition under Varying Facial Expression”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 28.10 (2006), pp. 1695–1700. DOI: 10.1109/TPAMI.2006.210. URL: <https://doi.org/10.1109/TPAMI.2006.210>.
- [Cha+15] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. “PCANet: A Simple Deep Learning Baseline for Image Classification?” In: *IEEE Trans. Image Process.* 24.12 (2015), pp. 5017–5032. DOI: 10.1109/TIP.2015.2475625. URL: <https://doi.org/10.1109/TIP.2015.2475625>.
- [Che+13] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. “Blessing of Dimensionality: High-Dimensional Feature and Its Efficient Compression for Face Verification”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. IEEE Computer Society, 2013, pp. 3025–3032. DOI: 10.1109/CVPR.2013.389. URL: <https://doi.org/10.1109/CVPR.2013.389>.
- [Che+15] Jiansheng Chen, Yu Deng, Gaocheng Bai, and Guangda Su. “Face Image Quality Assessment Based on Learning to Rank”. In: *IEEE Signal Process. Lett.* 22.1 (2015), pp. 90–94. DOI: 10.1109/LSP.2014.2347419. URL: <https://doi.org/10.1109/LSP.2014.2347419>.
- [Che+16] Jun-Cheng Chen, Amit Kumar, Rajeev Ranjan, Vishal M. Patel, Azadeh Alavi, and Rama Chellappa. “A cascaded convolutional neural network for age estimation of unconstrained faces”. In: *8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016*. IEEE, 2016, pp. 1–8. DOI: 10.1109/BTAS.2016.7791154. URL: <https://doi.org/10.1109/BTAS.2016.7791154>.
- [Chh+18] Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta. “Anonymizing k Facial Attributes via Adversarial Perturbations”. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. Ed. by Jérôme Lang. ijcai.org, 2018, pp. 656–662. DOI: 10.24963/ijcai.2018/91. URL: <https://doi.org/10.24963/ijcai.2018/91>.
- [Coo+19] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. “Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial

-
- Systems”. In: *IEEE Trans. Biom. Behav. Identity Sci.* 1.1 (2019), pp. 32–41. DOI: 10.1109/TBIOM.2019.2897801. URL: <https://doi.org/10.1109/TBIOM.2019.2897801>.
- [Dam+17a] Naser Damer, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “Efficient, Accurate, and Rotation-Invariant Iris Code”. In: *IEEE Signal Process. Lett.* 24.8 (2017), pp. 1233–1237. DOI: 10.1109/LSP.2017.2719282. URL: <https://doi.org/10.1109/LSP.2017.2719282>.
- [Dam+17b] Naser Damer, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “General borda count for multi-biometric retrieval”. In: *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*. IEEE, 2017, pp. 420–428. DOI: 10.1109/BTAS.2017.8272726. URL: <https://doi.org/10.1109/BTAS.2017.8272726>.
- [Dam+17c] Naser Damer, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “Indexing of Single and Multi-instance Iris Data Based on LSH-Forest and Rotation Invariant Representation”. In: *Computer Analysis of Images and Patterns - 17th International Conference, CAIP 2017, Ystad, Sweden, August 22-24, 2017, Proceedings, Part II*. Ed. by Michael Felsberg, Anders Heyden, and Norbert Krüger. Vol. 10425. Lecture Notes in Computer Science. Springer, 2017, pp. 190–201. DOI: 10.1007/978-3-319-64698-5_17. URL: https://doi.org/10.1007/978-3-319-64698-5_17.
- [Dam+18a] Naser Damer, Viola Boller, Yaza Wainakh, Fadi Boutros, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “Detecting Face Morphing Attacks by Analyzing the Directed Distances of Facial Landmarks Shifts”. In: *GCPR*. Vol. 11269. Lecture Notes in Computer Science. Springer, 2018, pp. 518–534.
- [Dam+18b] Naser Damer, Fadi Boutros, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “P-Score: Performance Aligned Normalization and an Evaluation in Score-Level Multi-Biometric Fusion”. In: *26th European Signal Processing Conference, EUSIPCO 2018, Roma, Italy, September 3-7, 2018*. IEEE, 2018, pp. 1402–1406. DOI: 10.23919/EUSIPCO.2018.8553553. URL: <https://doi.org/10.23919/EUSIPCO.2018.8553553>.
- [Dam+18c] Naser Damer, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “Fingerprint and Iris Multi-Biometric Data Indexing and Retrieval”. In: *21st International Conference on Information Fusion, FUSION 2018, Cambridge, UK, July 10-13, 2018*. IEEE, 2018, pp. 2083–2090. DOI: 10.23919/ICIF.

-
- 2018.8455390. URL: <https://doi.org/10.23919/ICIF.2018.8455390>.
- [Dam+18d] Naser Damer, Yaza Wainakh, Viola Boller, Sven von den Berken, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. “CrazyFaces: Unassisted Circumvention of Watchlist Face Identification”. In: *9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018*. IEEE, 2018, pp. 1–9. DOI: 10.1109/BTAS.2018.8698557. URL: <https://doi.org/10.1109/BTAS.2018.8698557>.
- [Dam+19] Naser Damer, Alexandra Mosegui Saladie, Steffen Zienert, Yaza Wainakh, Philipp Terhörst, Florian Kirchbuchner, and Arjan Kuijper. “To Detect or not to Detect: The Right Faces to Morph”. In: *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*. IEEE, 2019, pp. 1–8. DOI: 10.1109/ICB45273.2019.8987316. URL: <https://doi.org/10.1109/ICB45273.2019.8987316>.
- [Dam18] Naser Damer. “Application-driven Advances in Multi-biometric Fusion”. PhD thesis. Darmstadt University of Technology, Germany, 2018. URL: <http://tuprints.ulb.tu-darmstadt.de/7324/>.
- [Dan+11] Antitza Dantcheva, Carmelo Velardo, Angela D’Angelo, and Jean-Luc Dugeley. “Bag of soft biometrics for person identification - New trends and challenges”. In: *Multim. Tools Appl.* 51.2 (2011), pp. 739–777. DOI: 10.1007/s11042-010-0635-7. URL: <https://doi.org/10.1007/s11042-010-0635-7>.
- [DDB18] Abhijit Das, Antitza Dantcheva, and Francois Bremond. “Mitigating Bias in Gender, Age and Ethnicity Classification: A Multi-task Convolution Neural Network Approach”. In: *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part I*. Ed. by Laura Leal-Taixé and Stefan Roth. Vol. 11129. Lecture Notes in Computer Science. Springer, 2018, pp. 573–585. DOI: 10.1007/978-3-030-11009-3_35. URL: https://doi.org/10.1007/978-3-030-11009-3_35.
- [Den+19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

-
-
- [DER16] Antitza Dantcheva, Petros Elia, and Arun Ross. “What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics”. In: *IEEE Trans. Inf. Forensics Secur.* 11.3 (2016), pp. 441–467. DOI: 10.1109/TIFS.2015.2480381. URL: <https://doi.org/10.1109/TIFS.2015.2480381>.
- [Dey+14] Subhadeep Dey, Sujit Barman, Ramesh Kumar Bhukya, Rohan Kumar Das, Haris B. C., S. R. M. Prasanna, and Rohit Sinha. “Speech biometric based attendance system”. In: *Twentieth National Conference on Communications, Kanpur*. IEEE, 2014.
- [DHG12] Weihong Deng, Jiani Hu, and Jun Guo. “Extended SRC: Undersampled Face Recognition via Intra-class Variant Dictionary”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 34.9 (2012), pp. 1864–1870. DOI: 10.1109/TPAMI.2012.30. URL: <https://doi.org/10.1109/TPAMI.2012.30>.
- [DHG18] Weihong Deng, Jiani Hu, and Jun Guo. “Face Recognition via Collaborative Representation: Its Discriminant Nature and Superposed Representation”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.10 (2018), pp. 2513–2521. DOI: 10.1109/TPAMI.2017.2757923. URL: <https://doi.org/10.1109/TPAMI.2017.2757923>.
- [DHG19] Weihong Deng, Jiani Hu, and Jun Guo. “Compressive Binary Patterns: Designing a Robust Binary Face Descriptor with Random-Field Eigenfilters”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.3 (2019), pp. 758–767. DOI: 10.1109/TPAMI.2018.2800008. URL: <https://doi.org/10.1109/TPAMI.2018.2800008>.
- [DNJ18] Debayan Deb, Neeta Nain, and Anil K. Jain. “Longitudinal Study of Child Face Recognition”. In: *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*. IEEE, 2018, pp. 225–232. DOI: 10.1109/ICB2018.2018.00042. URL: <https://doi.org/10.1109/ICB2018.2018.00042>.
- [Dro+20] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. “Demographic Bias in Biometrics: A Survey on an Emerging Challenge”. In: *CoRR abs/2003.02488* (2020). arXiv: 2003.02488. URL: <https://arxiv.org/abs/2003.02488>.
- [DSN14] Naser Damer, Timotheos Samartzidis, and Alexander Nouak. “Personalized Face Reference from Video: Key-Face Selection and Feature-Level Fusion”. In: *Face and Facial Expression Recognition from Real World Videos - International Workshop, FFER@ICPR 2014, Stockholm, Sweden, August 24, 2014, Revised Selected Papers*. Ed. by Qiang Ji, Thomas B. Moeslund, Gang

-
- Hua, and Kamal Nasrollahi. Vol. 8912. Lecture Notes in Computer Science. Springer, 2014, pp. 85–98. DOI: 10.1007/978-3-319-13737-7_8. URL: https://doi.org/10.1007/978-3-319-13737-7%5C_8.
- [DVS14] Abhishek Dutta, Raymond N. J. Veldhuis, and Luuk J. Spreeuwens. “A Bayesian model for predicting face recognition performance using image quality”. In: *IEEE International Joint Conference on Biometrics, Clearwater, IJCB 2014, FL, USA, September 29 - October 2, 2014*. IEEE, 2014, pp. 1–8. DOI: 10.1109/BTAS.2014.6996248. URL: <https://doi.org/10.1109/BTAS.2014.6996248>.
- [Dwo+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. “Fairness Through Awareness”. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS ’12. Cambridge, Massachusetts: ACM, 2012, pp. 214–226. ISBN: 978-1-4503-1115-1. DOI: 10.1145/2090236.2090255. URL: <http://doi.acm.org/10.1145/2090236.2090255>.
- [EEH14] Eran Eidinger, Roeen Enbar, and Tal Hassner. “Age and Gender Estimation of Unfiltered Faces”. In: *IEEE Trans. Inf. Forensics Secur.* 9.12 (2014), pp. 2170–2179. DOI: 10.1109/TIFS.2014.2359646. URL: <https://doi.org/10.1109/TIFS.2014.2359646>.
- [Erk+09] Zekeriya Erkin, Martin Franz, Jorge Guajardo, Stefan Katzenbeisser, Inald Legendijk, and Tomas Toft. “Privacy-Preserving Face Recognition”. In: *Privacy Enhancing Technologies, 9th International Symposium, PETS 2009, Seattle, WA, USA, August 5-7, 2009. Proceedings*. Ed. by Ian Goldberg and Mikhail J. Atallah. Vol. 5672. Lecture Notes in Computer Science. Springer, 2009, pp. 235–253. DOI: 10.1007/978-3-642-03168-7_14. URL: https://doi.org/10.1007/978-3-642-03168-7%5C_14.
- [Fan+20] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Demographic Bias in Presentation Attack Detection of Iris Recognition Systems”. In: *CoRR abs/2003.03151* (2020). arXiv: 2003.03151. URL: <https://arxiv.org/abs/2003.03151>.
- [Fer+12] Matteo Ferrara, Annalisa Franco, Dario Maio, and Davide Maltoni. “Face Image Conformance to ISO/ICAO Standards in Machine Readable Travel Documents”. In: *IEEE Trans. Inf. Forensics Secur.* 7.4 (2012), pp. 1204–1213. DOI: 10.1109/TIFS.2012.2198643. URL: <https://doi.org/10.1109/TIFS.2012.2198643>.

-
- [FF10] Lester Randolph Ford and Delbert Ray Fulkerson. *Flows in Networks*. USA: Princeton University Press, 2010. ISBN: 0691146675.
- [FPO02] Nicholas Furl, P. Jonathon Phillips, and Alice J. O’Toole. “Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis”. In: *Cognitive Science* 26 (2002), pp. 797–815.
- [Fro17] Frontex. “Best Practice Technical Guidelines for Automated Border Control (ABC) Systems”. In: (2017). URL: https://frontex.europa.eu/assets/Publications/Research/Best_Practice_Technical_Guidelines_ABC.pdf.
- [Gao+07] Xiufeng Gao, Stan Z. Li, Rong Liu, and Peiren Zhang. “Standardization of Face Image Sample Quality”. In: *Advances in Biometrics*. Ed. by Seong-Whan Lee and Stan Z. Li. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 242–251. ISBN: 978-3-540-74549-5.
- [Gar+16] C. Garvie, Georgetown University. Center on Privacy, Technology, and Georgetown University. Law Center Center on Privacy & Technology. *The Perpetual Line-up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016. URL: <https://books.google.de/books?id=uAYjngAACAAJ>.
- [GB03] Ralph Gross and Vladimir Brajovic. “An Image Preprocessing Algorithm for Illumination Invariant Face Recognition”. In: *Audio-and Video-Based Biometric Person Authentication, 4th International Conference, AVBPA 2003, Guildford, UK, June 9-11, 2003 Proceedings*. Ed. by Josef Kittler and Mark S. Nixon. Vol. 2688. Lecture Notes in Computer Science. Springer, 2003, pp. 10–18. DOI: 10.1007/3-540-44887-X_2. URL: https://doi.org/10.1007/3-540-44887-X%5C_2.
- [Ge+19] Shiming Ge, Shengwei Zhao, Chenyu Li, and Jia Li. “Low-Resolution Face Recognition in the Wild via Selective Knowledge Distillation”. In: *IEEE Trans. Image Process.* 28.4 (2019), pp. 2051–2062. DOI: 10.1109/TIP.2018.2883743. URL: <https://doi.org/10.1109/TIP.2018.2883743>.
- [Gei+20] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. “Garbage in, garbage out?: do machine learning application papers in social computing report where human-labeled training data comes from?” In: *FAT* ’20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*. Ed. by Mireille Hildebrandt, Carlos Castillo, Elisa Celis, Salvatore Ruggieri, Linnet Taylor,

-
- and Gabriela Zanfir-Fortuna. ACM, 2020, pp. 325–336. DOI: 10.1145/3351095.3372862. URL: <https://doi.org/10.1145/3351095.3372862>.
- [GF16] Bryce Goodman and Seth Flaxman. *EU regulations on algorithmic decision-making and a "right to explanation"*. cite arxiv:1606.08813Comment: presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, 2016. URL: <http://arxiv.org/abs/1606.08813>.
- [GG16] Yarín Gal and Zoubin Ghahramani. “Dropout As a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML’16. New York, NY, USA: JMLR.org, 2016, pp. 1050–1059. URL: <http://dl.acm.org/citation.cfm?id=3045390.3045502>.
- [GLJ19] Sixue Gong, Xiaoming Liu, and Anil K. Jain. “DebFace: De-biasing Face Recognition”. In: *CoRR* abs/1911.08080 (2019). arXiv: 1911.08080. URL: <http://arxiv.org/abs/1911.08080>.
- [GLS90] Beatrice A. Golomb, David T. Lawrence, and Terrence J. Sejnowski. “SEXNET: A Neural Network Identifies Sex From Human Faces”. In: *Advances in Neural Information Processing Systems 3, [NIPS Conference, Denver, Colorado, USA, November 26-29, 1990]*. Ed. by Richard Lippmann, John E. Moody, and David S. Touretzky. Morgan Kaufmann, 1990, pp. 572–579. URL: <http://papers.nips.cc/paper/405-sexnet-a-neural-network-identifies-sex-from-human-faces>.
- [GM13] Guodong Guo and Guowang Mu. “Joint estimation of age, gender and ethnicity: CCA vs. PLS”. In: *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013, Shanghai, China, 22-26 April, 2013*. IEEE Computer Society, 2013, pp. 1–6. DOI: 10.1109/FG.2013.6553737. URL: <https://doi.org/10.1109/FG.2013.6553737>.
- [GNH18] Patrick Grother, Mei Ngan, and Kayee Hanaoka. “Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification”. In: *NIST Interagency/Internal Report (NISTIR) (2018)*. URL: <https://www.nist.gov/publications/ongoing-face-recognition-vendor-test-frvt-part-2-identification>.

-
- [GNH19a] Patrick Grother, Mei Ngan, and Kayee Hanaoka. “Face Recognition Vendor Test - Face Recognition Quality Assessment Concept and Goals”. In: NIST, 2019.
- [GNH19b] Patrick Grother, Mei Ngan, and Kayee Hanaoka. “Face Recognition Vendor Test Part 3: Demographic Effects”. In: *NIST Interagency/Internal Report (NISTIR) - 8280* (2019). URL: <https://www.nist.gov/publications/face-recognition-vendor-test-part-3-demographic-effects>.
- [Gro+06] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. “Model-Based Face De-Identification”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2006, New York, NY, USA, 17-22 June, 2006*. IEEE Computer Society, 2006, p. 161. DOI: 10.1109/CVPRW.2006.125. URL: <https://doi.org/10.1109/CVPRW.2006.125>.
- [GT07] Patrick Grother and Elham Tabassi. “Performance of Biometric Quality Measures”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 29.4 (2007), pp. 531–543. DOI: 10.1109/TPAMI.2007.1019. URL: <https://doi.org/10.1109/TPAMI.2007.1019>.
- [Guo+16] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. “MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9907. Lecture Notes in Computer Science. Springer, 2016, pp. 87–102. DOI: 10.1007/978-3-319-46487-9_6. URL: https://doi.org/10.1007/978-3-319-46487-9_6.
- [Guo+17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks”. In: *CoRR* abs/1706.04599 (2017).
- [Guo+18] Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. “Stacked Dense U-Nets with Dual Transformers for Robust Face Alignment”. In: *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 44. URL: <http://bmvc2018.org/contents/papers/0051.pdf>.
- [GZS07] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. “Automatic Age Estimation Based on Facial Aging Patterns”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 29.12 (2007), pp. 2234–2240. DOI: 10.1109/TPAMI.2007.70733. URL: <https://doi.org/10.1109/TPAMI.2007.70733>.

-
- [Han+15] Hu Han, Charles Otto, Xiaoming Liu, and Anil K. Jain. “Demographic Estimation from Face Images: Human vs. Machine Performance”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.6 (2015), pp. 1148–1161. DOI: 10.1109/TPAMI.2014.2362759. URL: <https://doi.org/10.1109/TPAMI.2014.2362759>.
- [Han+18] Hu Han, Anil K. Jain, Fang Wang, Shiguang Shan, and Xilin Chen. “Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.11 (2018), pp. 2597–2609. DOI: 10.1109/TPAMI.2017.2738004. URL: <https://doi.org/10.1109/TPAMI.2017.2738004>.
- [Has+15] Tal Hassner, Shai Harel, Eran Paz, and Roei Enbar. “Effective face frontalization in unconstrained images”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 4295–4304. DOI: 10.1109/CVPR.2015.7299058. URL: <https://doi.org/10.1109/CVPR.2015.7299058>.
- [He+05] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and HongJiang Zhang. “Face Recognition Using Laplacianfaces”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 27.3 (2005), pp. 328–340. DOI: 10.1109/TPAMI.2005.55. URL: <https://doi.org/10.1109/TPAMI.2005.55>.
- [Her+19] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. “FaceQnet: Quality Assessment for Face Recognition based on Deep Learning”. In: *IEEE International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*. June 2019.
- [Hil+18] Matthew Q. Hill, Connor J. Parde, Carlos Domingo Castillo, Y. Ivette Colon, Rajeev Ranjan, Jun-Cheng Chen, Volker Blanz, and Alice J. O’Toole. “Deep Convolutional Neural Networks in the Face of Caricature: Identity and Image Revealed”. In: *CoRR abs/1812.10902* (2018). arXiv: 1812.10902. URL: <http://arxiv.org/abs/1812.10902>.
- [HJ14] Hu Han and Anil K. Jain. “Age, Gender and Race Estimation from Unconstrained Face Images”. In: *MSU Technical Report*. 2014.
- [HO00] Aapo Hyvärinen and Erkki Oja. “Independent component analysis: algorithms and applications”. In: *Neural Networks* 13.4-5 (2000), pp. 411–430. DOI: 10.1016/S0893-6080(00)00026-5. URL: [https://doi.org/10.1016/S0893-6080\(00\)00026-5](https://doi.org/10.1016/S0893-6080(00)00026-5).

-
- [HOJ13] Hu Han, Charles Otto, and Anil K. Jain. “Age estimation from face images: Human vs. machine performance”. In: *International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain*. Ed. by Julian Fierrez, Ajay Kumar, Mayank Vatsa, Raymond N. J. Veldhuis, and Javier Ortega-Garcia. IEEE, 2013, pp. 1–8. DOI: 10.1109/ICB.2013.6613022. URL: <https://doi.org/10.1109/ICB.2013.6613022>.
- [HSM06] Rein-Lien Vincent Hsu, Jidnya Shah, and Brian Martin. “Quality Assessment of Facial Images”. In: *2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*. Sept. 2006, pp. 1–6. DOI: 10.1109/BCC.2006.4341617.
- [HSV19a] John J. Howard, Yevgeniy B. Sirotin, and Arun R. Vemury. “The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance”. In: *10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, USA, September, 2019*. IEEE, 2019.
- [HSV19b] John J. Howard, Yevgeniy B. Sirotin, and Arun R. Vemury. “The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance”. In: *10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, FL, USA, September 23-26, 2019*. IEEE, 2019, pp. 1–8. DOI: 10.1109/BTAS46853.2019.9186002. URL: <https://doi.org/10.1109/BTAS46853.2019.9186002>.
- [HT19] Isabelle Hupont and Carles Fernández Tena. “DemogPairs: Quantifying the Impact of Demographic Imbalance in Deep Face Recognition”. In: *14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, Lille, France, May 14-18, 2019*. IEEE, 2019, pp. 1–7. DOI: 10.1109/FG.2019.8756625. URL: <https://doi.org/10.1109/FG.2019.8756625>.
- [Hua+07] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. rep. 07-49. University of Massachusetts, Amherst, Oct. 2007.
- [Hua+18] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. “Deep Imbalanced Learning for Face Recognition and Attribute Prediction”. In: *CoRR abs/1806.00194* (2018). arXiv: 1806.00194. URL: <http://arxiv.org/abs/1806.00194>.

-
- [Hue+15] Ivan Huerta, Carles Fernández, Carlos Segura, Javier Hernando, and Andrea Prati. “A Deep Analysis on Age Estimation”. In: *Pattern Recognition Letters* 68 (2015). DOI: 10.1016/j.patrec.2015.06.006.
- [HW79] J. A. Hartigan and M. A. Wong. “Algorithm AS 136: A K-Means Clustering Algorithm”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), pp. 100–108. ISSN: 00359254, 14679876.
- [IS15] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by Francis R. Bach and David M. Blei. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, 2015, pp. 448–456. URL: <http://proceedings.mlr.press/v37/ioffe15.html>.
- [JBS15] Karl Ricanek Jr., Shivani Bhardwaj, and Michael Sodomsky. “A Review of Face Recognition against Longitudinal Child Faces”. In: *BIOSIG 2015 - Proceedings of the 14th International Conference of the Biometrics Special Interest Group, 9.-11. September 2015, Darmstadt, Germany*. Ed. by Arslan Brömme, Christoph Busch, Christian Rathgeb, and Andreas Uhl. Vol. P-245. LNI. GI, 2015, pp. 15–26. URL: <https://dl.gi.de/20.500.12116/2272>.
- [JDN04a] Anil K. Jain, Sarat C. Dass, and Karthik Nandakumar. “Can soft biometric traits assist user recognition?” In: *Biometric Technology for Human Identification*. Ed. by Anil K. Jain and Nalini K. Ratha. Vol. 5404. International Society for Optics and Photonics. SPIE, 2004, pp. 561–572. DOI: 10.1117/12.542890. URL: <https://doi.org/10.1117/12.542890>.
- [JDN04b] Anil K. Jain, Sarat C. Dass, and Karthik Nandakumar. “Soft Biometric Traits for Personal Recognition Systems”. In: *Biometric Authentication, First International Conference, ICBA 2004, Hong Kong, China, July 15-17, 2004, Proceedings*. Ed. by David Zhang and Anil K. Jain. Vol. 3072. Lecture Notes in Computer Science. Springer, 2004, pp. 731–738. DOI: 10.1007/978-3-540-25948-0_99. URL: https://doi.org/10.1007/978-3-540-25948-0%5C_99.
- [JFR10] Anil K. Jain, Patrick Flynn, and Arun A. Ross. *Handbook of Biometrics*. 1st. Springer Publishing Company, Incorporated, 2010. ISBN: 1441943757.

-
- [Jia+18] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. “To Trust Or Not To Trust A Classifier”. In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 5542–5553. URL: <http://papers.nips.cc/paper/7798-to-trust-or-not-to-trust-a-classifier.pdf>.
- [JNR16] Anil K. Jain, Karthik Nandakumar, and Arun Ross. “50 years of biometric research: Accomplishments, challenges, and opportunities”. In: *Pattern Recognition Letters 79* (2016), pp. 80–105.
- [JRP04] Anil K. Jain, Arun Ross, and Salil Prabhakar. “An introduction to biometric recognition”. In: *IEEE Trans. Circuits Syst. Video Technol.* 14.1 (2004), pp. 4–20. DOI: 10.1109/TCSVT.2003.818349. URL: <https://doi.org/10.1109/TCSVT.2003.818349>.
- [JYL15] Amin Jourabloo, Xi Yin, and Xiaoming Liu. “Attribute preserved face de-identification”. In: *International Conference on Biometrics, ICB 2015, Phuket, Thailand, 19-22 May, 2015*. IEEE, 2015, pp. 278–285. DOI: 10.1109/ICB.2015.7139096. URL: <https://doi.org/10.1109/ICB.2015.7139096>.
- [KB14] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR abs/1412.6980* (2014).
- [Kin09] Davis E. King. “Dlib-ml: A Machine Learning Toolkit”. In: *J. Mach. Learn. Res.* 10 (Dec. 2009), pp. 1755–1758. ISSN: 1532-4435. URL: <http://dl.acm.org/citation.cfm?id=1577069.1755843>.
- [Kin13] Els J. Kindt. “Biometric Data, Data Protection and the Right to Privacy”. In: *Privacy and Data Protection Issues of Biometric Applications: A Comparative Legal Analysis*. Dordrecht: Springer Netherlands, 2013. DOI: 10.1007/978-94-007-7522-0_3. URL: https://doi.org/10.1007/978-94-007-7522-0_3.
- [KL15] Volodymyr Kuleshov and Percy S Liang. “Calibrated Structured Prediction”. In: *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015, pp. 3474–3482. URL: <http://papers.nips.cc/paper/5658-calibrated-structured-prediction.pdf>.

-
-
- [KL99] Young H Kwon and Niels da Vitoria Lobo. “Age Classification from Facial Images”. In: *Comput. Vis. Image Underst.* 74.1 (1999), pp. 1–21. ISSN: 1077-3142. DOI: 10.1006/cviu.1997.0549. URL: <http://dx.doi.org/10.1006/cviu.1997.0549>.
- [Kla+12] Brendan Klare, Mark James Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. “Face Recognition Performance: Role of Demographic Information”. In: *IEEE Trans. Inf. Forensics Secur.* 7.6 (2012), pp. 1789–1801. DOI: 10.1109/TIFS.2012.2214212. URL: <https://doi.org/10.1109/TIFS.2012.2214212>.
- [KLR15] Hyungil Kim, Seung-Ho Lee, and Yong Man Ro. “Face image assessment learned with objective and relative face image qualities for improved face recognition”. In: *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*. IEEE, 2015, pp. 4027–4031. DOI: 10.1109/ICIP.2015.7351562. URL: <https://doi.org/10.1109/ICIP.2015.7351562>.
- [KND15] John D. Kelleher, Brian Mac Namee, and Aoife D’Arcy. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. The MIT Press, 2015. ISBN: 0262029448.
- [Kor+19] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. “Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2019.
- [Kri+20] K. S. Krishnapriya, Vitor Albiero, Kushal Vangara, Michael C. King, and Kevin W. Bowyer. “Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone”. In: *IEEE Transactions on Technology and Society* 1.1 (2020), pp. 8–20.
- [KS14] Vahid Kazemi and Josephine Sullivan. “One millisecond face alignment with an ensemble of regression trees”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1867–1874.
- [Kum+09] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. “Attribute and simile classifiers for face verification”. In: *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*. IEEE Computer Society, 2009, pp. 365–372. DOI: 10.1109/ICCV.2009.5459250. URL: <https://doi.org/10.1109/ICCV.2009.5459250>.

-
- [Kum+11] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. “Describable Visual Attributes for Face Verification and Image Search”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 33.10 (2011), pp. 1962–1977. DOI: 10.1109/TPAMI.2011.48. URL: <https://doi.org/10.1109/TPAMI.2011.48>.
- [KY55] Harold W. Kuhn and Bryn Yaw. “The Hungarian method for the assignment problem”. In: *Naval Res. Logist. Quart* (1955), pp. 83–97.
- [LH15] Gil Levi and Tal Hassner. “Age and Gender Classification Using Convolutional Neural Networks”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*. 2015. URL: https://osnathassner.github.io/talhassner/projects/cnn_agegender%7D.
- [Lia+19] Jian Liang, Yuren Cao, Chenbin Zhang, Shiyu Chang, Kun Bai, and Zenglin Xu. “Additive Adversarial Learning for Unbiased Authentication”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.
- [Liu+15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. “Deep Learning Face Attributes in the Wild”. In: *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015.
- [Liu+17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. “SphereFace: Deep Hypersphere Embedding for Face Recognition”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 6738–6746. DOI: 10.1109/CVPR.2017.713. URL: <https://doi.org/10.1109/CVPR.2017.713>.
- [Liu+18] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. “Label-Sensitive Deep Metric Learning for Facial Age Estimation”. In: *IEEE Trans. Inf. Forensics Secur.* 13.2 (2018), pp. 292–305. DOI: 10.1109/TIFS.2017.2746062. URL: <https://doi.org/10.1109/TIFS.2017.2746062>.
- [LMZ06] Xiaoxing Li, Greg Mori, and Hao Zhang. “Expression-Invariant Face Recognition with Expression Classification”. In: *Third Canadian Conference on Computer and Robot Vision (CRV 2006), 7-9 June 2006, Quebec City, Canada*. IEEE Computer Society, 2006, p. 77. DOI: 10.1109/CRV.2006.34. URL: <https://doi.org/10.1109/CRV.2006.34>.

-
- [Lou+13] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. “Understanding variable importances in forests of randomized trees”. In: *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013. URL: <http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf>.
- [LPL14] Zhen Lei, Matti Pietikäinen, and Stan Z. Li. “Learning Discriminant Face Descriptor”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 36.2 (2014), pp. 289–302. DOI: 10.1109/TPAMI.2013.112. URL: <https://doi.org/10.1109/TPAMI.2013.112>.
- [LW02] Chengjun Liu and Harry Wechsler. “Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition”. In: *IEEE Trans. Image Process.* 11.4 (2002), pp. 467–476. DOI: 10.1109/TIP.2002.999679. URL: <https://doi.org/10.1109/TIP.2002.999679>.
- [MAP16] Jordi Mansanet, Alberto Albiol, and Roberto Paredes. “Local Deep Neural Networks for gender recognition”. In: *Pattern Recognition Letters* 70 (2016), pp. 80–86. ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2015.11.015>. URL: <http://www.sciencedirect.com/science/article/pii/S0167865515004018>.
- [Mar72] George Marsaglia. “Choosing a Point from the Surface of a Sphere”. In: *Ann. Math. Statist.* (1972). DOI: 10.1214/aoms/1177692644. URL: <https://doi.org/10.1214/aoms/1177692644>.
- [Mas+16] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut, and Gérard G. Medioni. “Do We Really Need to Collect Millions of Faces for Effective Face Recognition?” In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9909. Lecture Notes in Computer Science. Springer, 2016, pp. 579–596. DOI: 10.1007/978-3-319-46454-1_35. URL: https://doi.org/10.1007/978-3-319-46454-1_35.
- [Mas+18] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. “Deep Face Recognition: A Survey”. In: *31st SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2018, Paraná, Brazil, October 29 - Nov. 1, 2018*. IEEE Computer Society, 2018, pp. 471–478. DOI: 10.1109/SIBGRAPI.2018.00067. URL: <https://doi.org/10.1109/SIBGRAPI.2018.00067>.

-
- [Med+21] Blaz Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter Schreier, Arun Ross, Peter Peer, and Vitomir Struc. “Privacy–Enhancing Face Biometrics: A Comprehensive Survey”. In: *IEEE Trans. Inf. Forensics Secur.* (2021). Under review.
- [MFV19] Aythami Morales, Julian Fierrez, and Rubén Vera-Rodríguez. “SensitiveNets: Learning Agnostic Representations with Application to Face Recognition”. In: *CoRR abs/1902.00334* (2019). arXiv: 1902.00334. URL: <http://arxiv.org/abs/1902.00334>.
- [MH08] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* (2008).
- [Mir+18] Vahid Mirjalili, Sebastian Raschka, Anoop M. Namboodiri, and Arun Ross. “Semi-adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images”. In: *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*. IEEE, 2018, pp. 82–89. DOI: 10.1109/ICB2018.2018.00023. URL: <https://doi.org/10.1109/ICB2018.2018.00023>.
- [MK19] Manisha and Nitin Kumar. “Cancelable Biometrics: a comprehensive survey”. In: *Artificial Intelligence Review* (2019). DOI: 10.1007/s10462-019-09767-8. URL: <https://doi.org/10.1007/s10462-019-09767-8>.
- [MMB12] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. “No-Reference Image Quality Assessment in the Spatial Domain”. In: *IEEE Trans. Image Process.* 21.12 (2012), pp. 4695–4708. DOI: 10.1109/TIP.2012.2214050. URL: <https://doi.org/10.1109/TIP.2012.2214050>.
- [MPS18] Blaz Meden, Peter Peer, and Vitomir Struc. “Selective Face Deidentification with End-to-End Perceptual Loss Learning”. In: *IEEE International Work Conference on Bioinspired Intelligence, IWOBI 2018, San Carlos, Alajuela, Costa Rica, July 18-20, 2018*. IEEE, 2018, pp. 1–7. DOI: 10.1109/IWOBI.2018.8464214. URL: <https://doi.org/10.1109/IWOBI.2018.8464214>.
- [MR17] Vahid Mirjalili and Arun Ross. “Soft biometric privacy: Retaining biometric utility of face images while perturbing gender”. In: *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*. IEEE, 2017, pp. 564–573. DOI: 10.1109/BTAS.2017.8272743. URL: <https://doi.org/10.1109/BTAS.2017.8272743>.

-
- [MRR18] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. “Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers”. In: *9th IEEE International Conference on Biometrics Theory, Applications and Systems* (2018).
- [MRR19] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. “FlowSAN: Privacy-Enhancing Semi-Adversarial Networks to Confound Arbitrary Face-Based Gender Classifiers”. In: *IEEE Access* 7 (2019), pp. 99735–99745. DOI: 10.1109/ACCESS.2019.2924619. URL: <https://doi.org/10.1109/ACCESS.2019.2924619>.
- [MRR20] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. “PrivacyNet: Semi-Adversarial Networks for Multi-attribute Face Privacy”. In: *CoRR* abs/2001.00561 (2020). arXiv: 2001.00561. URL: <http://arxiv.org/abs/2001.00561>.
- [MSB13] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. “Making a “Completely Blind” Image Quality Analyzer”. In: *IEEE Signal Process. Lett.* 20.3 (2013), pp. 209–212. DOI: 10.1109/LSP.2012.2227726. URL: <https://doi.org/10.1109/LSP.2012.2227726>.
- [Mur13] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press, 2013. ISBN: 9780262018029 0262018020. URL: https://www.amazon.com/Machine-Learning-Probabilistic-Perspective-Computation/dp/0262018020/ref=sr_1_2?ie=UTF8&qid=1336857747&sr=8-2.
- [MWP98] Baback Moghaddam, Wasiuddin Wahid, and Alex Pentland. “Beyond Eigenfaces: Probabilistic Matching for Face Recognition”. In: *3rd International Conference on Face & Gesture Recognition (FG '98), April 14-16, 1998, Nara, Japan*. IEEE Computer Society, 1998, pp. 30–35.
- [MYM18] Dana Michalski, Sau Yee Yiu, and Chris Malec. “The Impact of Age and Threshold Variation on Facial Recognition Algorithm Performance Using Images of Children”. In: *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*. IEEE, 2018, pp. 217–224. DOI: 10.1109/ICB2018.2018.00041. URL: <https://doi.org/10.1109/ICB2018.2018.00041>.
- [Neu19] Neurotechnology. “Neurotec Biometric SDK 11.1”. In: 2019.

-
- [NH08] Ingo Naumann and Giles Hogben. “Privacy features of European eID card specifications”. In: *Network Security* 2008.8 (2008), pp. 9–13. ISSN: 1353-4858. DOI: [https://doi.org/10.1016/S1353-4858\(08\)70097-7](https://doi.org/10.1016/S1353-4858(08)70097-7). URL: <http://www.sciencedirect.com/science/article/pii/S1353485808700977>.
- [NH10] Vinod Nair and Geoffrey E. Hinton. “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML’10. Haifa, Israel: Omnipress, 2010, pp. 807–814. ISBN: 9781605589077.
- [Nix+15] Mark S. Nixon, Paulo L. Correia, Kamal Nasrollahi, Thomas B. Moeslund, Abdenour Hadid, and Massimo Tistarelli. “On soft biometrics”. In: *Pattern Recognition Letters* 68 (2015). Special Issue on “Soft Biometrics”, pp. 218–230. ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2015.08.006>. URL: <http://www.sciencedirect.com/science/article/pii/S0167865515002615>.
- [NSM05] Elaine M. Newton, Latanya Sweeney, and Bradley Malin. “Preserving Privacy by De-Identifying Face Images”. In: *IEEE Trans. on Knowl. and Data Eng.* (2005).
- [NYC14] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images.” In: *CoRR* abs/1412.1897 (2014). URL: <http://dblp.uni-trier.de/db/journals/corr/corr1412.html#NguyenYC14>.
- [ÖAE16] Gökhan Özbulak, Yusuf Aytar, and Hazim Kemal Ekenel. “How Transferable Are CNN-Based Features for Age and Gender Classification?” In: *2016 International Conference of the Biometrics Special Interest Group, BIOSIG 2016, Darmstadt, Germany, September 21-23, 2016*. Ed. by Arslan Brömme, Christoph Busch, Christian Rathgeb, and Andreas Uhl. Vol. P-260. LNI. GI, 2016, pp. 39–50. DOI: [10.1109/BIOSIG.2016.7736925](https://doi.org/10.1109/BIOSIG.2016.7736925). URL: <https://doi.org/10.1109/BIOSIG.2016.7736925>.
- [OR14] Asem A. Othman and Arun Ross. “Privacy of Facial Soft Biometrics: Suppressing Gender But Retaining Identity”. In: *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*. Ed. by Lourdes Agapito, Michael M. Bronstein, and Carsten Rother. Vol. 8926. Lecture Notes in Computer Science. Springer, 2014, pp. 682–696. DOI: [10.1007/978-3-319-16181-5_52](https://doi.org/10.1007/978-3-319-16181-5_52). URL: https://doi.org/10.1007/978-3-319-16181-5_52.

-
- [Orc16] Mike Orcutt. “Are Face Recognition Systems Accurate? Depends on Your Race”. In: *MIT Technology Review 2016* (2016).
- [Ort+09] Marcos Ortega, Linda Brodo, Manuele Bicego, and Massimo Tistarelli. “Measuring changes in face appearance through aging”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2009, Miami, FL, USA, 20-25 June, 2009*. IEEE Computer Society, 2009, pp. 107–113. DOI: 10.1109/CVPRW.2009.5204302. URL: <https://doi.org/10.1109/CVPRW.2009.5204302>.
- [OTo+18] Alice J. O’Toole, Carlos D. Castillo, Connor J. Parde, Matthew Q. Hill, and Rama Chellappa. “Face Space Representations in Deep Convolutional Neural Networks”. In: *Trends in Cognitive Sciences* 22.9 (2018), pp. 794–809. ISSN: 1364-6613. DOI: <https://doi.org/10.1016/j.tics.2018.06.006>. URL: <http://www.sciencedirect.com/science/article/pii/S1364661318301463>.
- [Par+17] Connor J. Parde, Carlos Domingo Castillo, Matthew Q. Hill, Y. Ivette Colon, Swami Sankaranarayanan, Jun-Cheng Chen, and Alice J. O’Toole. “Face and Image Representation in Deep CNN Features”. In: *12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017*. IEEE Computer Society, 2017, pp. 673–680. DOI: 10.1109/FG.2017.85. URL: <https://doi.org/10.1109/FG.2017.85>.
- [Par+19] Connor J. Parde, Ying Hu, Carlos Domingo Castillo, Swami Sankaranarayanan, and Alice J. O’Toole. “Social Trait Information in Deep Convolutional Neural Networks Trained for Face Identification”. In: *Cognitive Science* 43.6 (2019). DOI: 10.1111/cogs.12729. URL: <https://doi.org/10.1111/cogs.12729>.
- [Ped+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *J. Mach. Learn. Res.* 12 (2011), pp. 2825–2830. URL: <http://dl.acm.org/citation.cfm?id=2078195>.
- [Phi+00] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. “The FERET Evaluation Methodology for Face-Recognition Algorithms”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 22.10 (2000), pp. 1090–1104.

-
- DOI: 10.1109/34.879790. URL: <https://doi.org/10.1109/34.879790>.
- [Phi+11] P. Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J. O’Toole. “An Other-race Effect for Face Recognition Algorithms”. In: *ACM Trans. Appl. Percept.* 8.2 (Feb. 2011), 14:1–14:11. ISSN: 1544-3558. DOI: 10.1145/1870076.1870082. URL: <http://doi.acm.org/10.1145/1870076.1870082>.
- [Phi+13] P. Jonathon Phillips, J. Ross Beveridge, David S. Bolme, Bruce A. Draper, Geof H. Givens, Yui Man Lui, Su Cheng, Mohammad Nayeem Teli, and Hao Zhang. “On the existence of face quality measures”. In: *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, Arlington, VA, USA, September 29 - October 2, 2013*. IEEE, 2013, pp. 1–8. DOI: 10.1109/BTAS.2013.6712715. URL: <https://doi.org/10.1109/BTAS.2013.6712715>.
- [Phi+18] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. O’Toole. “Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms”. In: *Proceedings of the National Academy of Sciences* 115.24 (2018), pp. 6171–6176. ISSN: 0027-8424. DOI: 10.1073/pnas.1721355115. eprint: <https://www.pnas.org/content/115/24/6171.full.pdf>. URL: <https://www.pnas.org/content/115/24/6171>.
- [PKB16] Hemprasad Patil, Ashwin Kothari, and Kishor Bhurchandi. “Expression invariant face recognition using semidecimated DWT, Patch-LDSMT, feature and score level fusion”. In: *Applied Intelligence* (2016).
- [PRC15] Vishal M. Patel, Nalini K. Ratha, and Rama Chellappa. “Cancelable Biometrics: A review”. In: *IEEE Signal Process. Mag.* 32.5 (2015), pp. 54–65. DOI: 10.1109/MSP.2015.2434151. URL: <https://doi.org/10.1109/MSP.2015.2434151>.
- [PS17] P. Punithavathi and Geetha Subbiah. “Can cancellable biometrics preserve privacy?” In: *Biometric Technology Today* 2017.7 (2017), pp. 8–11. ISSN: 0969-4765. DOI: [https://doi.org/10.1016/S0969-4765\(17\)30138-8](https://doi.org/10.1016/S0969-4765(17)30138-8). URL: <http://www.sciencedirect.com/science/article/pii/S0969476517301388>.

-
- [PVZ15] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. “Deep Face Recognition”. In: *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*. Ed. by Xianghua Xie, Mark W. Jones, and Gary K. L. Tam. BMVA Press, 2015, pp. 41.1–41.12. DOI: 10.5244/C.29.41. URL: <https://doi.org/10.5244/C.29.41>.
- [Ras06] Carl Edward Rasmussen. “Gaussian processes for machine learning”. In: MIT Press, 2006.
- [RB17] Ramachandra Raghavendra and Christoph Busch. “Presentation Attack Detection Methods for Face Recognition Systems: A Comprehensive Survey”. In: *ACM Comput. Surv.* 50.1 (2017), 8:1–8:37. DOI: 10.1145/3038924. URL: <https://doi.org/10.1145/3038924>.
- [Rho56] H.T.F. Rhodes. *Alphonse Bertillon, Father of Scientific Detection*. Abelard-Schuman, 1956. URL: <https://books.google.de/books?id=7BmAAAAAMAAJ>.
- [Rob+20] Joseph P. Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. “Face Recognition: Too Bias, or Not Too Bias?” In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. IEEE, 2020, pp. 1–10. DOI: 10.1109/CVPRW50498.2020.00008. URL: <https://doi.org/10.1109/CVPRW50498.2020.00008>.
- [Rod+17] Pau Rodríguez, Guillem Cucurull, Josep M. Gonfaus, F. Xavier Roca, and Jordi González. “Age and gender recognition in the wild with deep attention”. In: *Pattern Recognition* 72 (2017), pp. 563–571.
- [Roz+16] Andras Rozsa, Manuel Günther, Ethan M. Rudd, and Terrance E. Boult. “Are facial attributes adversarially robust?” In: *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*. IEEE, 2016, pp. 3121–3127. DOI: 10.1109/ICPR.2016.7900114. URL: <https://doi.org/10.1109/ICPR.2016.7900114>.
- [Roz+19] Andras Rozsa, Manuel Günther, Ethan M. Rudd, and Terrance E. Boult. “Facial attributes: Accuracy and adversarial robustness”. In: *Pattern Recognit. Lett.* 124 (2019), pp. 100–108. DOI: 10.1016/j.patrec.2017.10.024. URL: <https://doi.org/10.1016/j.patrec.2017.10.024>.

-
- [RT06a] Karl Ricanek and Tamirat Tesafaye. “MORPH: A Longitudinal Image Database of Normal Adult Age-Progression”. In: *Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2006), 10-12 April 2006, Southampton, UK*. IEEE Computer Society, 2006, pp. 341–345. URL: <https://doi.org/10.1109/FGR.2006.78>.
- [RT06b] Karl Ricanek and Tamirat Tesafaye. “MORPH: a longitudinal image database of normal adult age-progression”. In: *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. 2006, pp. 341–345.
- [RT12] Lyle Ramshaw and Robert Endre Tarjan. “A Weight-Scaling Algorithm for Min-Cost Imperfect Matchings in Bipartite Graphs”. In: *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012*. IEEE Computer Society, 2012, pp. 581–590. DOI: 10.1109/FOCS.2012.9. URL: <https://doi.org/10.1109/FOCS.2012.9>.
- [RTV18] Rasmus Rothe, Radu Timofte, and Luc Van Gool. “Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks”. In: *International Journal of Computer Vision* 126.2 (2018), pp. 144–157. ISSN: 1573-1405. DOI: 10.1007/s11263-016-0940-3. URL: <https://doi.org/10.1007/s11263-016-0940-3>.
- [San+16] Swami Sankaranarayanan, Azadeh Alavi, Carlos Domingo Castillo, and Rama Chellappa. “Triplet probabilistic embedding for face verification and clustering”. In: *8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016*. IEEE, 2016, pp. 1–8. DOI: 10.1109/BTAS.2016.7791205. URL: <https://doi.org/10.1109/BTAS.2016.7791205>.
- [Ser+20] Ignacio Serna, Aythami Morales, Julian Fierrez, Manuel Cebrián, Nick Obradovich, and Iyad Rahwan. “Algorithmic Discrimination: Formulation and Exploration in Deep Learning-based Face Biometrics”. In: *Proceedings of the Workshop on Artificial Intelligence Safety, co-located with 34th AAAI Conference on Artificial Intelligence, SafeAI@AAAI 2020, New York City, NY, USA, February 7, 2020*. Ed. by Huáscar Espinoza, José Hernández-Orallo, Xin Cynthia Chen, Seán S. Óhéigeartaigh, Xiaowei Huang, Mauricio Castillo-Effen, Richard Mallah, and John McDermid. Vol. 2560. CEUR Workshop Proceedings. CEUR-WS.org, 2020, pp. 146–152. URL: <http://ceur-ws.org/Vol-2560/paper10.pdf>.

-
- [Sha12] Caifeng Shan. “Learning local binary patterns for gender classification on real-world face images”. In: *Pattern Recognition Letters* 33 (2012), pp. 431–437.
- [She+17] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L. Yuille. “Deep Regression Forests for Age Estimation”. In: *CoRR* abs/1712.07195 (2017). arXiv: 1712.07195. URL: <http://arxiv.org/abs/1712.07195>.
- [SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A unified embedding for face recognition and clustering”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682. URL: <https://doi.org/10.1109/CVPR.2015.7298682>.
- [SP11] Andrew W. Senior and Sharathchandra Pankanti. “Privacy Protection and Face Recognition”. In: *Handbook of Face Recognition, 2nd Edition*. Ed. by Stan Z. Li and Anil K. Jain. Springer, 2011, pp. 671–691. DOI: 10.1007/978-0-85729-932-1_27. URL: https://doi.org/10.1007/978-0-85729-932-1_27.
- [SRB16] Martin Stokkenes, Raghavendra Ramachandra, and Christoph Busch. “Biometric Authentication Protocols on Smartphones: An Overview”. In: *Proc. of the 9th International Conference on Security of Information and Networks*. Newark, NJ, USA: ACM, 2016.
- [Sri+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: <http://jmlr.org/papers/v15/srivastava14a.html>.
- [Sri+19a] Nisha Srinivas, Matthew Hivner, Kevin Gay, Harleen Atwal, Micheal King, and Karl Ricanek. “Exploring Automatic Face Recognition on Match Performance and Gender Bias for Children”. In: *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 2019, pp. 107–115.
- [Sri+19b] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S. Bolme, and Michael King. “Face Recognition Algorithm Bias: Performance Differences on Images of Children and Adults”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019.

-
- [SSM99] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. “Advances in Kernel Methods”. In: ed. by Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola. Cambridge, MA, USA: MIT Press, 1999. Chap. Kernel Principal Component Analysis.
- [SSW09] Ahmad-Reza Sadeghi, Thomas Schneider, and Immo Wehrenberg. “Efficient Privacy-Preserving Face Recognition”. In: *Information, Security and Cryptology - ICISC 2009, 12th International Conference, Seoul, Korea, December 2-4, 2009, Revised Selected Papers*. Ed. by Dong Hoon Lee and Seokhie Hong. Vol. 5984. Lecture Notes in Computer Science. Springer, 2009, pp. 229–244. DOI: 10.1007/978-3-642-14423-3_16. URL: [https://doi.org/10.1007/978-3-642-14423-3_16](https://doi.org/10.1007/978-3-642-14423-3%5C_16).
- [Suo+11] Jin-Li Suo, Liang Lin, Shiguang Shan, Xilin Chen, and Wen Gao. “High-Resolution Face Fusion for Gender Conversion”. In: *IEEE Trans. Syst. Man Cybern. Part A* 41.2 (2011), pp. 226–237. DOI: 10.1109/TSMCA.2010.2064304. URL: <https://doi.org/10.1109/TSMCA.2010.2064304>.
- [Tai+14] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf. “DeepFace: Closing the Gap to Human-Level Performance in Face Verification”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 1701–1708. DOI: 10.1109/CVPR.2014.220. URL: <https://doi.org/10.1109/CVPR.2014.220>.
- [TB99] Michael E. Tipping and Chris M. Bishop. “Probabilistic Principal Component Analysis”. In: *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* (1999).
- [Ter+18a] Philipp Terhörst, Naser Damer, Andreas Braun, and Arjan Kuijper. “Deep and Multi-Algorithmic Gender Classification of Single Fingerprint Minutiae”. In: *21st International Conference on Information Fusion, FUSION 2018, Cambridge, UK, July 10-13, 2018*. IEEE, 2018, pp. 2113–2120. DOI: 10.23919/ICIF.2018.8455803. URL: <https://doi.org/10.23919/ICIF.2018.8455803>.
- [Ter+18b] Philipp Terhörst, Naser Damer, Andreas Braun, and Arjan Kuijper. “Minutiae-Based Gender Estimation for Full and Partial Fingerprints of Arbitrary Size and Shape”. In: *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part I*. Ed. by C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler. Vol. 11361. Lecture Notes in Computer Science. Springer,

-
- 2018, pp. 171–186. DOI: 10.1007/978-3-030-20887-5_11. URL: https://doi.org/10.1007/978-3-030-20887-5%5C_11.
- [Ter+18c] Philipp Terhörst, Naser Damer, Andreas Braun, and Arjan Kuijper. “What can a single minutia tell about gender?” In: *2018 International Workshop on Biometrics and Forensics, IWBF 2018, Sassari, Italy, June 7-8, 2018*. IEEE, 2018, pp. 1–7. DOI: 10.1109/IWBF.2018.8401554. URL: <https://doi.org/10.1109/IWBF.2018.8401554>.
- [Ter+19a] Philipp Terhörst, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Suppressing Gender and Age in Face Templates Using Incremental Variable Elimination”. In: *International Conference on Biometrics, ICB 2019, 4-7 June, 2019, Crete, Greece*. IEEE, 2019.
- [Ter+19b] Philipp Terhörst, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Un-supervised privacy-enhancement of face representations using similarity-sensitive noise transformations”. In: *Applied Intelligence* (Feb. 2019). ISSN: 1573-7497. DOI: 10.1007/s10489-019-01432-5. URL: <https://doi.org/10.1007/s10489-019-01432-5>.
- [Ter+19c] Philipp Terhörst, Marco Huber, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Multi-algorithmic fusion for reliable age and gender estimation from face images”. In: *22st International Conference on Information Fusion, FUSION 2019, Ottawa, Canada, July 2-5, 2019*. IEEE, 2019.
- [Ter+19d] Philipp Terhörst, Marco Huber, Jan Niklas Kolf, Ines Zelch, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Reliable Age and Gender Estimation from Face Images: Stating the Confidence of Model Predictions”. In: *10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, Florida, USA, September 23-26, 2019*. IEEE, 2019.
- [Ter+20a] Philipp Terhörst, Daniel Fährmann, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Beyond Identity: What Information Is Stored in Biometric Face Templates?” In: *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, USA, Sept 28 - Oct 1, 2020*. IEEE, 2020.
- [Ter+20b] Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “MAAD-Face: A Massively Annotated Attribute Dataset for Face Images”. In: *CoRR abs/2012.01030* (2020). arXiv: 2012.01030. URL: <https://arxiv.org/abs/2012.01030>.

-
- [Ter+20c] Philipp Terhörst, Marco Huber, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. *Unsupervised Enhancement of Soft-biometric Privacy with Negative Face Recognition*. 2020. arXiv: 2002.09181 [cs.CV].
- [Ter+20d] Philipp Terhörst, Marco Huber, Naser Damer, Peter Rot, Florian Kirchbuchner, Vitomir Struc, and Arjan Kuijper. “Privacy Evaluation Protocols for the Evaluation of Soft-Biometric Privacy-Enhancing Technologies”. In: *2020 International Conference of the Biometrics Special Interest Group, BIOSIG 2020, Darmstadt, Germany, September 16-18, 2020*. IEEE, 2020.
- [Ter+20e] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Face Quality Estimation and Its Correlation to Demographic and Non-Demographic Bias in Face Recognition”. In: *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, USA, Sept 28 - Oct 1, 2020*. IEEE, 2020.
- [Ter+20f] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. *Post-comparison mitigation of demographic bias in face recognition using fair score normalization*. 2020. DOI: <https://doi.org/10.1016/j.patrec.2020.11.007>. URL: <http://www.sciencedirect.com/science/article/pii/S0167865520304128>.
- [Ter+20g] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, Washington, USA, June 14-19, 2020*. Computer Vision Foundation / IEEE, 2020.
- [Ter+20h] Philipp Terhörst, Kevin Riehl, Naser Damer, Peter Rot, Blaz Bortolato, Florian Kirchbuchner, Vitomir Struc, and Arjan Kuijper. “PE-MIU: A Training-Free Privacy-Enhancing Face Recognition Approach Based on Minimum Information Units”. In: *IEEE Access* 8 (2020), pp. 93635–93647. DOI: 10.1109/ACCESS.2020.2994960. URL: <https://doi.org/10.1109/ACCESS.2020.2994960>.
- [Ter+20i] Philipp Terhörst, Mai Ly Tran, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Comparison-Level Mitigation of Ethnic Bias in Face Recognition”. In: *2020 International Workshop on Biometrics and Forensics, IWBF 2020, Porto, Portugal, April 29-30, 2020*. IEEE, 2020.

-
- [Ter+21a] Philipp Terhörst, Andre Boller, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “MiDeCon: Unsupervised and Accurate Fingerprint and Minutia Quality Assessment based on Minutia Detection Confidence”. In: *2021 IEEE International Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*. Under review. IEEE, 2021.
- [Ter+21b] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales, Julian Fierrez, and Arjan Kuijper. “A Comprehensive Study on Face Recognition Biases Beyond Demographics”. In: *CoRR abs/2012.01030* (2021).
- [TG15] Elham Tabassi and Patrick Grother. “Biometric Sample Quality”. In: *Encyclopedia of Biometrics*. Springer US, 2015.
- [TP91] Matthew Turk and Alex Pentland. “Eigenfaces for recognition”. In: *J. Cognitive Neuroscience* 3.1 (1991), pp. 71–86. doi: <http://dx.doi.org/10.1162/jocn.1991.3.1.71>.
- [Tri17] Bipin Kumar Tripathi. “On the complex domain deep machine learning for face recognition”. In: *Appl. Intell.* 47.2 (2017), pp. 382–396. doi: 10.1007/s10489-017-0902-7. url: <https://doi.org/10.1007/s10489-017-0902-7>.
- [Ull+12] Ihsan Ullah, Muhammad Awais Hussain, Ghulam Muhammad, Hatim A. Aboalsamh, George Bebis, and Anwar Majid Mirza. “Gender recognition from face images with local WLD descriptor”. In: *2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)* (2012), pp. 417–420.
- [VB17] Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. 1st. Springer Publishing Company, Incorporated, 2017. ISBN: 3319579584, 9783319579580.
- [Ven+15] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, S. S. Channappayya, and S. S. Medasani. “Blind image quality evaluation using perception based features”. In: *2015 Twenty First National Conference on Communications (NCC)*. Feb. 2015, pp. 1–6. doi: 10.1109/NCC.2015.7084843.
- [Ver+19] Ruben Vera-Rodriguez, Marta Blazquez, Aythami Morales, Ester Gonzalez-Sosa, Joao C. Neves, and Hugo Proenca. “FaceGenderID: Exploiting Gender Information in DCNNs Face Recognition Systems”. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation

-
- / IEEE, 2019, pp. 2254–2260. DOI: 10.1109/CVPRW.2019.00278. URL: http://openaccess.thecvf.com/content%5C_CVPRW%5C_2019/html/BEFA/Vera-Rodriguez%5C_FaceGenderID%5C_Exploiting%5C_Gender%5C_Information%5C_in%5C_DCNNs%5C_Face%5C_Recognition%5C_Systems%5C_CVPRW%5C_2019%5C_paper.html.
- [Vir+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* (2020). DOI: <https://doi.org/10.1038/s41592-019-0686-2>.
- [Wan+18a] Chang-Peng Wang, Wei Wei, Jiang-She Zhang, and Hou-Bing Song. “Robust face recognition via discriminative and common hybrid dictionary learning”. In: *Applied Intelligence* (2018). URL: <https://doi.org/10.1007/s10489-017-0956-6>.
- [Wan+18b] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. “CosFace: Large Margin Cosine Loss for Deep Face Recognition”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 5265–5274. DOI: 10.1109/CVPR.2018.00552. URL: http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Wang%5C_CosFace%5C_Large%5C_Margin%5C_CVPR%5C_2018%5C_paper.html.
- [Wan+19] Pingyu Wang, Fei Su, Zhicheng Zhao, Yandong Guo, Yanyun Zhao, and Bojin Zhuang. “Deep class-skewed learning for face recognition”. In: *Neurocomputing* 363 (2019), pp. 35–45. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2019.04.085>. URL: <http://www.sciencedirect.com/science/article/pii/S092523121930832X>.
- [Was+17] Pankaj Shivdayal Wasnik, Kiran B. Raja, Ramachandra Raghavendra, and Christoph Busch. “Assessing face image quality for smartphone based face

-
- recognition system”. In: *5th International Workshop on Biometrics and Forensics, IWBF 2017, Coventry, United Kingdom, April 4-5, 2017*. IEEE, 2017, pp. 1–6. DOI: 10.1109/IWBF.2017.7935089. URL: <https://doi.org/10.1109/IWBF.2017.7935089>.
- [Way01] James L. Wayman. “Fundamentals of Biometric Authentication Technologies”. In: *Int. J. Image Graph.* 1 (2001), pp. 93–113.
- [WCV11] Stéfan van der Walt, S. Chris Colbert, and Gaël Varoquaux. “The NumPy Array: A Structure for Efficient Numerical Computation”. In: *Comput. Sci. Eng.* 13.2 (2011), pp. 22–30. DOI: 10.1109/MCSE.2011.37. URL: <https://doi.org/10.1109/MCSE.2011.37>.
- [WD18] Mei Wang and Weihong Deng. “Deep Face Recognition: A Survey”. In: *CoRR abs/1804.06655* (2018). arXiv: 1804.06655. URL: <http://arxiv.org/abs/1804.06655>.
- [WD19] Mei Wang and Weihong Deng. “Mitigate Bias in Face Recognition using Skewness-Aware Reinforcement Learning”. In: *CoRR abs/1911.10692* (2019). arXiv: 1911.10692. URL: <http://arxiv.org/abs/1911.10692>.
- [Wen+16] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. “A Discriminative Feature Learning Approach for Deep Face Recognition”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9911. Lecture Notes in Computer Science. Springer, 2016, pp. 499–515. DOI: 10.1007/978-3-319-46478-7_31. URL: https://doi.org/10.1007/978-3-319-46478-7_31.
- [WG13] Lingyun Wen and Guodong Guo. “A computational approach to body mass index prediction from face images”. In: *Image Vis. Comput.* 31.5 (2013), pp. 392–400. DOI: 10.1016/j.imavis.2013.03.001. URL: <https://doi.org/10.1016/j.imavis.2013.03.001>.
- [WGK15] Xiaolong Wang, Rui Guo, and Chandra Kambhampettu. “Deeply-Learned Feature for Age Estimation”. In: *2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, Waikoloa, HI, USA, January 5-9, 2015*. IEEE Computer Society, 2015, pp. 534–541. DOI: 10.1109/WACV.2015.77. URL: <https://doi.org/10.1109/WACV.2015.77>.

-
- [WK18] Yilun Wang and Michal Kosinski. “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images”. In: *J. of Personality and Social Psychology* (2018).
- [WKW15] Jos van de Wolfshaar, Mahir Faik Karaaba, and Marco A. Wiering. “Deep Convolutional Neural Networks and Support Vector Machines for Gender Recognition”. In: *IEEE Symposium Series on Computational Intelligence, SSCI 2015, Cape Town, South Africa, December 7-10, 2015*. IEEE, 2015, pp. 188–195. DOI: 10.1109/SSCI.2015.37. URL: <https://doi.org/10.1109/SSCI.2015.37>.
- [Won+11] Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, and Brian C. Lovell. “Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2011, Colorado Springs, CO, USA, 20-25 June, 2011*. IEEE Computer Society, 2011, pp. 74–81. DOI: 10.1109/CVPRW.2011.5981881. URL: <https://doi.org/10.1109/CVPRW.2011.5981881>.
- [Wri+09] John Wright, Allen Y. Yang, Arvind Ganesh, Shankar S. Sastry, and Yi Ma. “Robust Face Recognition via Sparse Representation”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 31.2 (2009), pp. 210–227. DOI: 10.1109/TPAMI.2008.79. URL: <https://doi.org/10.1109/TPAMI.2008.79>.
- [Wu+18] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. “A Light CNN for Deep Face Representation With Noisy Labels”. In: *IEEE Trans. Information Forensics and Security* 13.11 (2018), pp. 2884–2896. DOI: 10.1109/TIFS.2018.2833032. URL: <https://doi.org/10.1109/TIFS.2018.2833032>.
- [Yan+07a] Shuicheng Yan, Dong Xu, Benyu Zhang, HongJiang Zhang, Qiang Yang, and Stephen Lin. “Graph Embedding and Extensions: A General Framework for Dimensionality Reduction”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 29.1 (2007), pp. 40–51. DOI: 10.1109/TPAMI.2007.250598. URL: <https://doi.org/10.1109/TPAMI.2007.250598>.
- [Yan+07b] Jian Yang, David Zhang, Jing-Yu Yang, and Ben Niu. “Globally Maximizing, Locally Minimizing: Unsupervised Discriminant Projection with Applications to Face and Palm Biometrics”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 29.4 (2007), pp. 650–664. DOI: 10.1109/TPAMI.2007.1008. URL: <https://doi.org/10.1109/TPAMI.2007.1008>.

-
- [Yi+14] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. “Learning Face Representation from Scratch”. In: *CoRR* abs/1411.7923 (2014). arXiv: 1411.7923. URL: <http://arxiv.org/abs/1411.7923>.
- [Yin+19] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. “Feature Transfer Learning for Face Recognition With Under-Represented Data”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 5704–5713.
- [YLL14] Dong Yi, Zhen Lei, and Stan Z. Li. “Age Estimation by Multi-scale Convolutional Network”. In: *Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part III*. Ed. by Daniel Cremers, Ian D. Reid, Hideo Saito, and Ming-Hsuan Yang. Vol. 9005. Lecture Notes in Computer Science. Springer, 2014, pp. 144–158. DOI: 10.1007/978-3-319-16811-1_10. URL: https://doi.org/10.1007/978-3-319-16811-1%5C_10.
- [ZBB18] Soumaya Zaghbani, Noureddine Boujneh, and Med Salim Bouhlel. “Age estimation using deep learning”. In: *Computers & Electrical Engineering* 68 (2018), pp. 337–347. ISSN: 0045-7906. DOI: <https://doi.org/10.1016/j.compeleceng.2018.04.012>. URL: <http://www.sciencedirect.com/science/article/pii/S0045790617334298>.
- [Zei+18] Chris Zeinstra, Didier Meuwly, Raymond Veldhuis, and Luuk Spreeuwers. “Mind the Gap: A Practical Framework for Classifiers in a Forensic Context”. In: *BTAS*. IEEE, 2018.
- [Zei12] Matthew D. Zeiler. “ADADELTA: An Adaptive Learning Rate Method”. In: *CoRR* abs/1212.5701 (2012). arXiv: 1212.5701. URL: <http://arxiv.org/abs/1212.5701>.
- [Zem+13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. “Learning Fair Representations”. In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 325–333. URL: <http://proceedings.mlr.press/v28/zemel13.html>.
- [ZG17] Fatema Tuz Zohra and Marina L. Gavrilova. “Adaptive Face Recognition Based on Image Quality”. In: (2017), pp. 218–221. DOI: 10.1109/CW.2017.35. URL: <https://doi.org/10.1109/CW.2017.35>.

-
- [Zha+05] Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. “Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A Novel Non-Statistical Model for Face Representation and Recognition”. In: *10th IEEE International Conference on Computer Vision (ICCV 2005), 17-20 October 2005, Beijing, China*. IEEE Computer Society, 2005, pp. 786–791. DOI: 10.1109/ICCV.2005.147. URL: <https://doi.org/10.1109/ICCV.2005.147>.
- [Zha+16] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. “Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks”. In: *CoRR abs/1604.02878* (2016).
- [Zha+17] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. “Range Loss for Deep Face Recognition with Long-Tailed Training Data”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 5419–5428. DOI: 10.1109/ICCV.2017.578. URL: <https://doi.org/10.1109/ICCV.2017.578>.
- [Zha+18] Dongdong Zhao, Wenjian Luo, Ran Liu, and Lihua Yue. “Negative Iris Recognition”. In: *IEEE Trans. Dependable Sec. Comput.* 15.1 (2018), pp. 112–125. DOI: 10.1109/TDSC.2015.2507133. URL: <https://doi.org/10.1109/TDSC.2015.2507133>.
- [ZSL16a] Yang Zhong, Josephine Sullivan, and Haibo Li. “Face attribute prediction using off-the-shelf CNN features”. In: *International Conference on Biometrics, ICB 2016, Halmstad, Sweden, June 13-16, 2016*. IEEE, 2016, pp. 1–7. DOI: 10.1109/ICB.2016.7550092. URL: <https://doi.org/10.1109/ICB.2016.7550092>.
- [ZSL16b] Yang Zhong, Josephine Sullivan, and Haibo Li. “Leveraging mid-level deep representations for predicting face attributes in the wild”. In: *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*. IEEE, 2016, pp. 3239–3243. DOI: 10.1109/ICIP.2016.7532958. URL: <https://doi.org/10.1109/ICIP.2016.7532958>.
- [ZYF11] Lei Zhang, Meng Yang, and Xiangchu Feng. “Sparse representation or collaborative representation: Which helps face recognition?” In: *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*. Ed. by Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool. IEEE Computer Society, 2011, pp. 471–478.

DOI: 10.1109/ICCV.2011.6126277. URL: <https://doi.org/10.1109/ICCV.2011.6126277>.

- [ZZ10] Yin Zhang and Zhi-Hua Zhou. “Cost-Sensitive Face Recognition”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 32.10 (Oct. 2010), pp. 1758–1769. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2009.195. URL: <https://doi.org/10.1109/TPAMI.2009.195>.