FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Explainable Artificial Intelligence – Getting insights from Deep Neural Networks for Interpretable and Fair Face Recognition

Ana Dias Teixeira de Viseu Cardoso

Mestrado Integrado em Bioengenharia

Supervisor: Ana Filipa Sequeira

Co-Supervisor: Pedro Neto

October 10, 2023

Explainable Artificial Intelligence – Getting insights from Deep Neural Networks for Interpretable and Fair Face Recognition

Ana Dias Teixeira de Viseu Cardoso

Mestrado Integrado em Bioengenharia

Approved in public examination by the Jury:

President: Luís Filipe Teixeira

Referee: Nuno Gonçalves

Referee: Ana Filipa Sequeira

October 10, 2023

Abstract

Human faces convey information about gender, age and ethnicity and more abstractly about a subject's emotions and social context. The capacity to identify and authenticate individuals based on their facial features currently represents the most commonly used type of data in biometric systems. Face recognition technology has evolved significantly in recent years, propelled by the proliferation of digital image data and the rise of Artificial Intelligence, which is directly linked with the development of Deep Learning methods. Deep Neural Networks are used on several fronts and achieve impressive results, even when comparing to the ones achieved by humans performing the same tasks.

The remarkable Deep Learning developments carried consequences, namely the transition from understandable models into black box systems. The trade-off between good quantitative results and the fairness and transparency of a model needs to be considered. Explainable Artificial Intelligence focus on the explainability of a model and unveils certain challenges and biases that remain present, particularly racial bias. This is a complex issue with implications on both ethical and social dimensions, transcending the domain of technology. Although racial bias is currently more studied, there is still little information available on the impact on the performance of face recognition algorithms. One of the primary contributors to racial bias is the imbalance in training data, given that many datasets are predominantly composed of images from a specific ethnic group and lack in diversity. Therefore, one of the main efforts to mitigate racial bias includes creating more diverse training databases, as well as developing fair algorithms. As face recognition systems have been adopted as a powerful security tool, racial inequity can translate into social injustices and misidentifications, raising the need for awareness on this topic.

This dissertation delves into the analysis and exploration of racial bias in face recognition systems. The work developed commences with a background and literature review, tracing the evolution of face recognition. The methods adopted focus on the use of race-aware databases and we aim to evaluate if the face recognition model used performs differently with four racial groups (Caucasian, African, Asian and Indian) under the same conditions. To investigate racial bias, especially intra-racial bias, various experiments were performed, starting with an analysis of the effects that image transformations have on a particular race. Moreover, gradient maps were generated for all races in the same layers of the network, allowing an analysis of the regions of interest in the input images. We performed practical experiments on neural network activations to look for a possible connection between human face recognition of subjects from other races and automatic face recognition evaluated on a race-aware dataset. As deep neural networks cannot be evaluated over time, the analysis made focused on how data flows through the network layers in a specific order. We calculated metrics such as mean and standard deviation from the neural network activation values extracted from the network's layers and the results were compared between races. At last, using the neural network feature maps generated from specific layers, we tested the separability of racial groups.

Even though various ideas were pursued, the experiments did not present a clear and straightforward conclusion on racial bias and the reasoning behind it. However, it is mandatory that this topic keeps on being studied and addressed. Moreover, in terms of future work, it may be interesting to focus on some racial bias mitigation techniques and, by adding synthetic bias to the data, measuring its quantitative impact on the tests performed in this dissertation.

Resumo

Os rostos humanos transmitem informações acerca do género, idade e etnia e, de uma forma mais abstrata, sobre emoções e o contexto social de uma pessoa. A capacidade de identificar e autenticar indivíduos com base nas suas características faciais atualmente representa o tipo de dados mais usados em sistemas biométricos. A tecnologia de reconhecimento facial evoluiu significativamente nos últimos anos, impulsionada pela proliferação de dados de imagens digitais e pelo aparecimento da inteligência artificial, que está diretamente ligada com o desenvolvimento de métodos de *Deep Learning*. As *Deep Neural Networks* são usadas em várias frentes e têm vindo a alcançar resultados impressionantes, mesmo quando comparadas com as capacidades humanas ao realizar as mesmas tarefas.

Os notáveis desenvolvimentos na área de *Deep Learning* resultaram em consequências, nomeadamente a transição de modelos compreensíveis e interpretáveis para sistemas opacos (*black-box*). O equilíbrio entre bons resultados quantitativos e a transparência de um modelo precisa de ser considerada aquando do desenvolvimento de um algoritmo. A *Explainable Artificial Intelligence* foca-se na capacidade de explicação e interpretação de um modelo e é responsável por desvendar certos desafios e vieses presentes, nomeadamente o viés racial. Este é um problema complexo com implicações em dimensões éticas e sociais, transcendendo o domínio da tecnologia. Embora o viés racial seja atualmente mais estudado, há ainda pouca informação disponível acerca das implicações desta no desempenho dos algoritmos de reconhecimento facial. Um dos principais contribuintes para o viés racial é o desequilíbrio nos dados presentes em *datasets* de treino, uma vez que muitos *datasets* são predominantemente compostos por dados de um grupo étnico específico e carecem em diversidade. Desta forma, um dos principais esforços dos investigadores para mitigar o viés racial inclui a criação de bases de dados de treino mais diversificadas, bem como o desenvolvimento de algoritmos mais justos. Uma vez que os sistemas de reconhecimento facial têm sido adotados como uma ferramenta útil na área da segurança, a desigualdade racial pode traduzir-se em injustiças sociais e identificações policiais erradas, o que leva à necessidade de consciencialização sobre este tópico.

O objetivo desta dissertação é explorar a análise e estudo do viés racial em sistemas de reconhecimento facial. O trabalho desenvolvido começa com uma revisão de literatura e antecedentes no tópico, traçando a evolução do reconhecimento facial até aos dias de hoje. Os métodos adotados durante este trabalho concentram-se na utilização de bases de dados sensíveis à raça e o objetivo é avaliar se o modelo de reconhecimento facial utilizado se comporta de maneira diferente com algum dos quatro grupos raciais (Caucasiano, Africano, Asiático e Indiano). Para investigar o viés racial, especialmente o viés intra-racial, foram realizadas várias experiências, começando com uma análise dos efeitos que transformações de imagens têm numa raça específica. Além disso, foram gerados mapas de gradiente para todas as raças nas mesmas camadas da rede neuronal, permitindo uma análise das regiões de interesse nas imagens de *input*. Foram também realizadas experiências nas ativações da rede neuronal de forma a procurar uma possível conexão entre o reconhecimento facial humano, com uma possível ativação neuronal díspar devido à etnia dos

sujeitos observados, e o reconhecimento facial por sistemas automáticos, avaliado à luz da discriminação racial. Uma vez que as *Deep Neural Networks* não podem ser avaliadas ao longo do tempo, a análise realizada focou-se na forma como os dados fluem através das camadas da rede numa ordem específica. Métricas como a média e o desvio padrão foram calculadas a partir dos valores extraídos das ativações da rede neural e os resultados foram comparados entre raças. Utilizando mapas de características da rede gerados a partir de camadas específicas, testou-se a separabilidade e classificação de grupos raciais.

Embora várias ideias tenham sido testadas, no geral, as experiências não apresentaram uma conclusão clara e direta acerca do viés racial e a sua razão de ser. No entanto, é imperativo que este tópico continue a ser estudado e abordado em diversas áreas de estudo. Além disso, em termos de trabalho futuro possível, poderá ser interessante focar a atenção na mitigação do viés racial e, adicionando viés sintético aos dados, medir o seu impacto quantitativo nos testes que foram realizados nesta dissertação.

Acknowledgments

My first acknowledgment goes to both my supervisors, Ana Filipa Sequeira and Pedro Neto, who accompanied me through this dissertation. I want to thank you both for being extremely understanding and patient throughout this entire process.

To my parents, there are not enough words that would allow me to express my feelings of gratitude, love, and, above all, pride. You are truly the perfect example of work hard and make your dreams come true. The perfect duo, the ones that never turn their backs on me, letting me learn from my mistakes, but always being ready to pull me out of the dark. For you, the best of me will never be enough. I dedicate this accomplishment entirely to you.

To my brother and sister, thank you for your patience. Our bond became stronger since I started my master's degree and more than ever I can say that family is everything. Siblings are the friends we can't choose, but I got very lucky. To pipa, my beloved dog, the fact that you can make me smile when nothing else can speaks for itself. You are the craziest, funniest dog alive and you made me understand the concept of unconditional love.

To Cuca and Mariana, thank you for all the Ciao Bella dinners, the Camélia lunches, the "alheira" days and all the endless hours of good conversation. "Amigas amigas" that I hope to keep for life.

To Dafi, the one that came back right on time to save me and my degree. You believed more in me than I ever could, and you picked up what I thought was lost. You are the smartest person I know and I am forever grateful for the way you kept somewhat cool when I was asking the dumbest of questions. Thank you for the wake up calls, the long train rides to my house, and the many many times you "worked" just for me to work. Thank you for trying to make me see the good in every bad situation and for always believing that I was going to finish this. Staying delusional, it's the solution. Cheers to you, the next round is on me.

A very special mention for these last months of co-working space at my house, the time spent eating Adelaide's food at the "farm" and all the food in my house in general, being obsessed with playing monopoly and singing peaches. Best office ever.

Thank you to Luís for the constant availability.

To my best friend and soulmate of six years, Gil. Six years of complaining, screaming, being angry, depressed and annoyed at everything and everyone. You were there through the darkest of times and I could never thank you enough for not only being there, but by making me laugh like I never laughed before. I said it and I will say it a million times more: you are truly the only reason I survived FEUP and you are the reason that I am here today. I will forever cherish the memories we made together and keep on telling the craziest university stories that always involve you. All good things come to an end, and in this case, I hope it is the opposite. "Diz o roto ao nu", the blind leading the blind. With so much love and gratitude, I wish you all that you deserve: the very best.

Ana Cardoso

Contents

List of Figures

List of Tables

Abbreviations and Symbols

AI	Artificial Intelligence
ANNs	Artificial Neural Networks
CAM	Class Activation Map
CNNs	Convolutional Neural Networks
DCNNs	Deep Convolutional Neural Networks
DL	Deep Learning
EFR	Effective Receptive Fields
FC	Fully- Connected Layer
FMR	False Matches Rate
fMRI	Functional Magnetic Resonance Imaging
FR	Face Recognition
FRT	Face Recognition Technology
GPU	Graphics Processing Unit
Grad-CAM	Gradient-Weighted CAM
ID	Identity
KNN	K-Nearest Neighbors
ML	Machine Learning
ORE	Other-Race Effect
PCA	Principal Component Analysis
PIN	Personal Identification Number
ReLu	Rectified Linear Unit
RF	Receptive Field
RGB	Red Green Blue
SER	Skewed Error Ratio
SOTA	State-of-the-Art
STD	Standard Deviation
SVM	Support Vector Machine
TSNE	t-Distributed Stochastic Neighbor Embedding
xAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Context

Each human face is unique, providing insight into a person's identity. The human face conveys a diverse amount of information to an observer, including information about gender, age and ethnicity. More than that, it gives insight into the person's emotions and adds social context. The majority of human faces share the same set of features, such as eyes, nose, and mouth, that are roughly similarly arranged. The difference comes from slight and subtle variations in that exact configuration and its form, making each person's face distinctive [17]. The analysis of the human face and facial behavior is an interdisciplinary research area involving psychology, neuroscience, and engineering [13].

Biometric systems are capable of analysing and quantifying human's physical features (e.g. fingerprint, iris, palm print, and face) as well as behavioral traits (e.g. signature, walking, speech patterns, and facial dynamics) [8]. Individual human characteristics are transformed into biometric data and processed by algorithmic systems in order to verify or single out the identity of a person. These systems are part of our daily lives and we use them to unlock our phones or to cross borders at airports [18]. Face recognition (FR) systems use face measurements and allow for identification and authentication [19]. Recently, facial biometrics has been one of the most used methods for biometric data [8]. These types of systems keep improving at an incredibly fast pace, growing in use and application [20]. This rapid evolution means more data is constantly being added, and new users are identified and authenticated. Face recognition, in contrast to other biometric systems, does not necessarily require the active cooperation of the subject and can be performed unobtrusively, making it particularly suitable for surveillance applications [13].

Face recognition algorithms have benefited a lot from the growth of Deep Learning methods. These evolved in a way that led to the systems being capable of outperforming humans in specific tasks [21]. The development of the Artificial Intelligence (AI) field is directly linked with the improvements in Deep Learning and Deep Learning methods [13]. Deep Learning (DL) is a technique especially valuable in nonlinear and large-scale problems that are not suited for manual feature extraction, where Machine Learning would usually be applied [22]. Artificial Neural

Networks (ANNs) apply multiple layers to learn data with multiple levels of feature extraction [23]. These networks, which were initially inspired by the human neural system, consist of an input layer of neurons, one or more hidden layers of neurons and a final output layer. As our brain presents remarkable neural networks, granting connections between different parts of it and making it so that face recognition comes as naturally for us as possible, researchers have shown a significant interest in replicating these aspects through computational models [24].

Neuroscience suggests that local features are detected in earlier visual layers of the visual human cortex and only then progress to more complex patterns in an hierarchical manner. Taking this as inspiration, Convolutional Neural Networks (CNNs) embrace the concepts of receptive fields (RF) and effective receptive fields (EFR). A receptive field is a local region on the output of the previous layer that a neuron is connected to and an effective receptive field represents the area of the original input image that can influence the activation of a network neuron. These concepts converge for the first convolutional layer, but, as the CNN hierarchy progresses they start to differ: the RF is equal to the filter size of the previous layer and the ERF indicates the extent of the input image that modulates the neural activity [25]. Therefore, a parallelism in the computational engines for both human and machine-based face recognition may be a possibility. Both biological and artificial neural networks are nonlinear, with local convolutions executed in cascaded layers of neurons throughout time, and both artificial neural networks and humans learn in multiple steps [26]. Face recognition systems in applied settings such as law enforcement have spurred comparisons between how DCNNs and humans behave when performing the recognition task and researchers keep looking for similarities between the two.

Nonetheless, with the outstanding Deep Learning developments came consequences, the main one being the transformation of the biometric systems into black box systems. Even though researchers keep achieving incredible quantitative results, the main focus has shifted towards improving the models' fairness, transparency and explainability. More than just working on improving quantitative results and the accuracy of a model, it became essential to understand the reasoning behind a prediction and increase the user's trust, which can be one of the main barriers against the proliferation of smart technology through society [27]. Explainable Artificial Intelligence (xAI) is the area of AI that worries about explainability and whether a particular model is transparent and fair, making it possible for a user to understand what is happening when the model is implemented. The aim is to provide the user with an answer to the "why?"- answered by the explanation given by the model's prediction- and "how?"- providing an understanding of the model and the process that leads to the output [28]. Even though the current models are very accurate and their metrics are optimized, certain challenges and biases remain present [2]. We may try to apply reverse engineering to understand how deep networks recognize faces at a conceptual and representational level. However, reverse engineering aims to understand how a complex system like the human brain solves a problem such as face recognition. For that, we need to understand how the model works before applying it [26]. When using DCNNs in an attempt to emulate human facial perception, researchers must choose to either apply a smaller and controlled model or a larger and uncontrolled one. The first one is easier to analyse, but it may be limited in computational power,

performance and accurate predictions. The larger networks are closer to neural systems but may be untraceable.

To perform automatic FR, one of the most important parts of the process lies in choosing large, clean and diverse databases, both for training and testing. However, over time, researchers concluded that the most commonly used datasets needed to be more representative of particular groups of people, especially regarding race. Racial bias was concealed due to the biased benchmarks until more and more people began questioning this issue and its degrading impact. Although racial bias is now more studied and has become one of the goals to tackle in this field, it remains vacant. There is still very little testing done and minimal information available on the topic, making it hard to measure the consequences on the FR algorithm's performance [29].

1.2 Motivation

Face recognition represents a very challenging problem in image analysis and computer vision, and it has been increasingly receiving attention over the years because of its various applications in very different domains. Some of the applications include security and surveillance, identity verification, criminal justice systems, and video indexing (labelling faces in videos). Moreover, traditional security solutions, such as passwords, badges, the traditional ID cards, and PINs, have many limitations (e.g. misplacement or theft), which leads to the increased interest in biometric systems [19].

In a more futuristic light, data-driven marketing is trying to replace conventional strategies, and face recognition marketing has emerged as a significant tactic in the automated marketing systems of today. The customer's facial expressions, eye movements, blinks, pupil dilation and head movements can be measured using a facial recognition system based on artificial intelligence [30].

Face Recognition Technologies (FRT) offer various business opportunities for both users and developers. A Mordor-Intelligence study revealed that the total market value of face recognition was 4.4 billion in 2019, and it is expected that this number will increase to 10.9 billion by 2025 [13]. China became the world leader in FRT, both in use and in development and currently, it is the country with the more eccentric and out-of-the-box use of FR systems [1] [2]. China's loose data privacy laws facilitate this situation. An example of this happened in 2017 when Yum China partnered with the mobile payment service Alipay and came up with a "smile to pay" system. Previously registered users could confirm a payment just by smiling, without having to use a card or a smartphone [31].

Nevertheless, FR models' transparency and fairness decrease as they become more complex and accurate. People lose trust in how they work and become sceptical about their daily use. This issue opens the door for the much needed improvement in the explainability field, changing the focus from improving the model's performance to increasing its transparency. Moreover, it is

[1] https://www.wired.com/story/china-is-the-worlds-biggest-face-recognition-dealer/
[2] https://www.reuters.com/technology/china-drafts-rules-using-facial-recognition-technology-2023-08-08/

known that both FR algorithms and associated databases suffer from bias, namely racial bias [29]. This is an issue, given that specific groups of people are not correctly represented, which makes models' performance discriminatory. Our world is full of diversity, and the technology we develop should keep up with this, including and correctly representing subjects from all ethnicity groups, ages, genders and appearances.

1.3 Objectives

The presence of racial bias is factual and the journey to unveil the reasoning behind this issue has long started.

Our goal with this dissertation is not only to detect, but also to explore the logic for the occurrence of racial bias in face recognition systems. A comprehensive investigation is carried out and we propose to shed some light into the presence, the extent and implications of racial bias.

With this research work, we focus on the analysis of deep neural networks from different points of view, trying to understand the way they perform racial distribution on the latent space, analysing the gradient, identifying bias patterns and conducting evaluations on neural activations. Using neural feature maps extracted from face images, we perform classification by separating the data by race. Moreover, some image transformations aim to try to evaluate the impact that said modifications have on different ethnicity groups.

By pursuing these objectives and trying to get more insights on racial bias, ultimately we hope to contribute to this research area.

1.4 Contributions

Regarding the presented methodologies, the main contributions are the following:

- Extensive and thorough study of the main state-of-the-art approaches in face recognition and racial bias.
- Research on possible motives behind racial bias, with a primary emphasis on intra-racial bias.
- Exploration of the possibility of using the human face recognition process as inspiration in regards to neural activation, especially when it comes to neural activation as a result of processing faces from a distinct race than ours.
- Study of the use of the results from neural network activation to classify different ethnicity groups.

1.5 Document Structure

This dissertation contains a total of 6 chapters. After the introduction, Chapter 2 presents a background on the relevant topics for this work. Chapter 3 consists of a literature review on the subjects that are pertinent to the intended goals. Chapter 4 lays out the methodology and the developed work and Chapter 5 presents the results associated with the methods described as well as a discussion of said results. Finally, Chapter 6 rounds up the work and finishes with suggestions for possible future work.

Chapter 2

Fundamental Concepts

2.1 Human Face Recognition

The human face reveals a lot of information to an observer, and it has been studied as a way to identify people [32]. Human face recognition is the process by which a person judges if they have seen a face before, and it is an impressive process in terms of speed and accuracy. Furthermore, it proceeds via various stages and unravels over time [33]. Initially, incoming visual data is processed based on its immediate perceptual characteristics, forming a structural representation. Subsequently, this representation is further refined into a more abstract, perspective-neutral model of the face, facilitating comparisons with other faces stored in memory [34]. The examination of this cognitive task within the field of psychology dates back to 1980 [32], with potential antecedents preceding that period. A common research topic on face recognition is the difference in performance across individuals and even across distinct groups of people, meaning that a wide range of face recognition skills varies from person to person. A face is considered to be recognized when it is familiar. On the other hand, a previously unseen face may be considered unfamiliar or "unknown". In the human face recognition process, recognition and identification are considered different, and one does not need to identify an individual to recognize them [35]. Notwithstanding the variant face recognition abilities that change from individual to individual, it is known that humans, in general, perform a lot better when recognizing a familiar or well-known face. This continues to be true even when there are some changes in the photometric conditions, such as illumination, or in the person's appearance, such as changes in hair colour, natural ageing or some form of disguise [36].

Human faces share some standard features, such as eyes, nose and mouth, that are roughly arranged in a similar configuration. How we can differentiate faces comes from small changes in the form and layout of these facial features, and sometimes it is a matter of analysing very small details. Human face recognition varies with three different factors: stimulus factors, subject factors, and photometric conditions [17]. The stimulus factors point to the fact that not all faces are equally recognizable; some have distinctive features or a distinctive configuration, making it easier to differentiate from others. When we consider a face unusual, it results from a judgment

based on what the person considers to be the “usual” face and the average feature [37]. This means that faces are represented in relative terms instead of absolute ones and that we encode a face in terms of its deviation from what we consider to be the prototype face. It has been proven that we identify better computer-based caricatures where the feature values are exaggerated and deviate from the average values, making it easier to perform the task at hand [17]. The interaction between stimulus and subject factors leads to the conclusion that it is unrealistic to expect that all individuals, especially from different backgrounds and groups, share the same "average face". This brings light to the fact that depending on where we grow up, we may have a different idea of what is average and what is considered to be out of place or different, exclusively because of the amount of exposure we have as infants [17] [38]. This can result in the other race effect (ORE), which will be explored in more detail in the following point. When talking about photometric conditions, the main focus is the change in viewpoint and illumination, given that these are also some of the current challenges for algorithms to surpass in order to function in real-world situations [17].

After years of in-depth studies on how humans analyse and process other people’s faces, it can be concluded that different parts of the brain are active during this task and that the part of the brain that is used varies with the type of information conveyed. Static features related to identity and categorical information are most likely processed in a different part of the brain than motion information- that gives insight on social cues- and even emotional details are analysed in another region [17]. The way humans can keep track of an enormous number of individual faces is impressive and far greater than the number of objects from other classes that we are capable of memorizing [39]. As previously mentioned, different people perform distinctly when it comes to processing faces, and some possess better natural skills than others. Researchers discovered a relation between a better performance and a more extensive activation of certain parts of the brain [40].

The neuropsychological data on what was discovered to be a condition named *prosopagnosia* helped to bring light to the possibility that processing faces is special and different from what was seen with other classes of objects. This condition is defined as losing the ability to recognize faces even though the person can still recognize other types of objects [41]. Moreover, there are some cases in which someone who suffers from this condition can identify facial expressions without being able to recognize the person behind those same expressions. It is also essential to highlight that in these cases, recognizing is different than identifying, and the observer may be able to quickly identify the person by their voice or another distinctive clue [42].

Prejudice is a persistent and prevalent theme associated with human cognition, and it can be defined as a state of mind or behavior that criticizes or denigrates others on account of the group to which the individual belongs. There are various forms of prejudice, some more extensive than others, such as prejudice based on gender, age, sexuality, and religion. However, race-related prejudice is the more studied form and is especially relevant to this dissertation [43]. The task of recognizing a face can be influenced by both the race of the observer as well as the race of the individual that we wish to recognize. As these systems are being increasingly applied in the security field, it becomes even more crucial to understand the effects that race plays on the

accuracy results of face recognition systems.

2.1.1 Other-Race Effect

When talking about the average face, it is highly unlikely that it is the same for all different groups of subjects. This means that depending on where we grow up, we may have a different take on what we consider “the average”. This accounts for the other-race effect (ORE), which states that we recognize our own race more accurately than faces from other races [44]. This effect results from the interaction between the stimulus and the subject experience factors, as previously mentioned. As a possible explanation for the other-race effect, the contact hypothesis states that an individual’s experience with a certain race affects the representation of the distinctive features of the own-race faces, making them more detailed and accurate. However, when it comes to other-race faces, they are not as well characterized by those features [17]. The ORE can be explained by both the perceptual expertise model as well as the social cognitive model. The last one is based on the fact that humans tend to perceive individuals in terms of social groups. “In-groups” represent the group of people we consider to be similar to, and “out-groups” represent the opposite [45]. Many studies show that own-race faces are processed configurably, whereas other-race faces are processed with a more feature-based strategy. This effect can be measured in an experimental scenario, where subjects of different races are put to test on their ability to distinguish faces. Given that the other-race effect results in better face recognition capacity of same-race faces and basic level recognition for other-race faces, it is believed that there can be an "expertise training" where subjects are trained to better recognize and distinguish between other-race faces. This training is expected to reduce the implicit racial bias correlated with the improvement in differentiating other-race faces [46].

Implicit racial bias refers to the stereotypes and discriminatory behaviors based on race that start to develop unconsciously during early childhood. This type of prejudice then perpetrates into adulthood and affects different aspects of both the personal and social individual’s life. Even though researchers have tried different techniques to reduce implicit racial bias, by the adult age, it is already highly consolidated and resistant to change. This allowed psychologists to conclude that the best approach is to try and change these behaviors during childhood. Several studies showed that one successful approach involves perceptual individuation training, where the participants learn how to distinguish subjects from different races thus giving them experience with other-race faces [47]. Another way is to classify individuals by race, which is called categorization [45]. The early discrepancy between own- versus other-race faces does not only have perceptual consequences in terms of categorization and face recognition, but also social ones in terms of racial bias. Children associate positively or negatively with their own- versus other-race faces because of their asymmetric exposure during their early years. However, this implicit racial bias against unfamiliar races can be trained and may be malleable. The perceptual-social linkage hypothesis suggests that the appearance of implicit racial bias from an early age results from the tendency to categorize other-race faces and form positive associations with own-race faces and, therefore, familiar categories [48].

Moreover, past research suggests that racial cues can affect a number of different brain areas and their activity. Functional magnetic resonance imaging (fMRI) measures the changes in blood flow that occur with brain activity and some studies based on this imaging technique have shown that there are differences in brain activity due to implicit racial bias [49]. The inability to discern differences among individuals from outside one's own social group can have immediate and significant real-world repercussions. These consequences range from relatively trivial occurrences, such as mistakenly confusing two co-workers of the same ethnic background, to life-altering situations, such as incorrectly identifying an innocent person from a police lineup. These errors can arise from inaccuracies in memory and judgment, or they may originate from the fundamental manner in which we perceive individuals belonging to different social groups. Research demonstrates that out-group deindividuation manifests in the early stages of facial perception, evidenced by reduced neural sensitivity to variations in the facial features of individuals from different racial backgrounds. It was proven that White Americans are more sensitive to perceptual differences between White faces than Black faces, and even though these perceptual mechanisms are not clear, members from out-groups are perceived as multiple instances of the same category rather than distinct individuals [50].

Visual neuroscience offers methodologies to investigate how face-sensitive systems respond to intra-group variability. Extensive research underscores a fundamental aspect of brain processing: its inclination to familiarization when repeatedly exposed to the same stimuli. Neural populations display diminished activity following recurrent exposure to stimuli to which they are attuned, a phenomenon referred to as neural adaptation [50].

2.2 Automatic Face Recognition

Biometrics are biological measurements of physical and behavioral characteristics that make personal identification possible. With the rapid advancements in networking and the problems faced in the security department, there is a massive market for a reliable user authentication technique. This comes hand-in-hand with the need to protect important and personal information, entailing that only an authorized user should be granted access. Instead of relying on traditional security domains, biological traits cannot be misplaced, forgotten or forged. There are several different biometric traits, such as face, fingerprint, hand geometry, iris detection, and voice or speech recognition. The most used characteristic nowadays is the face, as it promotes a non-intrusive approach without capture delay, does not require the cooperation of the user, and can be obtained without violating the personal private space [1] [8]. Furthermore, it does not require the use of expensive sensors, since RGB cameras are sufficient.

Face recognition (FR) has been actively studied since the 1970s [51]. FR uses biometric data, such as face measurements, and it is used for the purpose of identification and verification. This technology can match a human face from a digital image or a video frame against a database full of different faces. It can also be applied in a live scenario, as per the example of security cameras. It is useful in a lot of different areas, such as security and surveillance, finances and

retail, and even daily life and comfort. The most traditional approaches to face recognition based on Machine Learning are centered on either the location or shape of some facial attributes and the spatial relationship between them, such as eyes, eyebrows, and nose, or the overall analysis of the face as a combination of several canonical faces. There are three main tasks that can be distinguished on a FR system: enrollment, authentication and identification. The first step allows the user to associate the biometric data with their identity and to save it properly in a gallery. To authenticate, it is necessary to make a direct comparison between the biometric information from the claimed user and the data saved in the gallery linked to this said person (1:1 comparison)-verification. In the end, for identification, the system takes the biometric data and the information stored in the gallery and tries to make a match (1:N comparison). If there is none, it will return an empty array and no identity [1]. In Fig. 2.1, we can see the various components of a face recognition system. The wide adoption of the FR systems implies that they will be used in a large and diverse population from very different backgrounds and demographics. Consequently, these systems must be capable of managing information from different users in an equal and fair way [46].

From the early 2010s, FR systems have benefited from the exponential growth of Deep Learning methods. These methods are also known for excelling in other Artificial Intelligence (AI) fields. The success associated with AI can be generally accounted for Deep Learning methods suffering a vast improvement, the large availability of databases and the improvements in computational power (GPU cards) [28]. Around 2014, Deep Convolutional Neural Networks (DCNNs) brought the ability to recognize faces "in the wild", e.g. in a real world scenario. Machines started to solve the problem of generalized face recognition. These algorithms are trained with millions of images of thousands of different individuals captured "in the wild", making the results more robust. As the currently available algorithms are expected to be used very diversely, their data must reflect such diversity [52] [53].

However, with the major improvements in performance came a significant increase in model complexity, which led to these systems becoming more opaque. The opaqueness of a ML model is believed to be proportional to its performance. The consequence may be the appearance of hidden biases, privacy issues and lack of transparency. Hence, the focus of researchers has changed from exclusively improving models' performance to using models that are more interpretable and that incorporate fairness. Furthermore, it is essential to develop methods that explain the reasoning behind a prediction, helping to turn a black box system into a transparent one. One of the most important elements when adopting a new technology is consumer trust. Consumers should know what is happening and understand why it is happening. Users should also know that their personal information is secured and what it is being used for [2].

In the current era of machine-based face recognition, it is important for researchers to understand in which circumstances the machine performs better than humans and vice-versa. More than that, it may be relevant to study a possible collaboration between humans and machines when performing face recognition, given that it may result in more accurate and robust results [36].

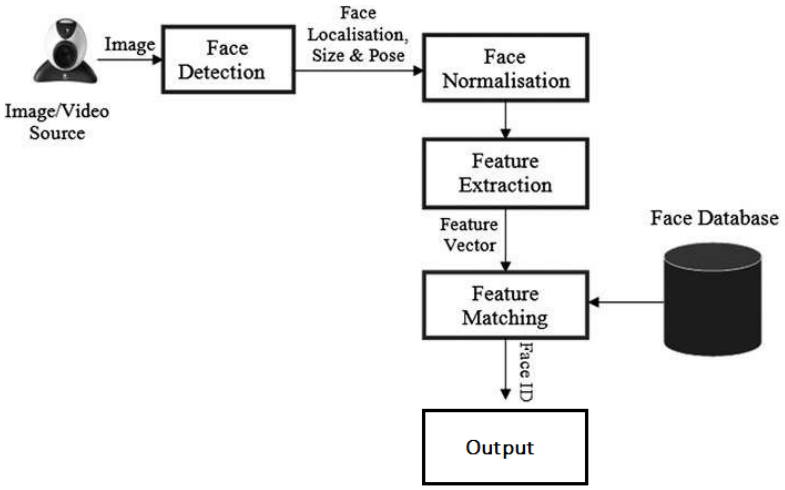


Figure 2.1: Face Recognition system adapted from [1].

2.2.1 Racial Bias in algorithms

Given that humans suffer from the ORE when performing face recognition, it is also expected that face recognition algorithms should be tested on the impact that different races have on the models' performance [46].

Differences in face recognition algorithms as a function of race have been reported since the early 1990's. One of the first studies to examine race bias in an algorithm [54] showed that the model had greater identification accuracy with the race with which it had more experience with. This was tested by training the model with either Asian faces as the minority race and Caucasian faces as the majority race or vice-versa. During the 2000s, researchers studied the effect of race on face recognition systems and concluded again that there were accuracy differences as a function of race [55]. In 2011, researchers continued exploring this topic and concluded that the part of the world where the algorithm is originated is also responsible for race bias. They compared two algorithms, one Western and one East Asian, and concluded that the first one performed better with Caucasian faces and the second one was more accurate for East Asian faces [46]. During the year of 2012, Klare *et al.* [56] tested the effects of race, gender and age and concluded that faces from young, female and black individuals suffered against all other demographic groups. Additionally, it was inferred that training the algorithms with more diverse databases helped to reduce these specific demographic biases, but did not eliminate them.

Nagpal *et al.* [57] presented the idea that upon limited exposure to other races, face recognition algorithms mimic the human inclination towards own-race bias. In this work, the authors also concluded that, similarly to human behavior, Deep Learning Networks have a tendency to focus on selected facial regions for a particular race, which varies across different races.

As stated before, FR systems have been increasingly adopted as a strong and powerful tool for security purposes. Racial inequity can translate into social injustices, as per the example of the

well known case of Robert Williams [58]. This citizen was wrongfully arrested in January 2020 because of a false match performed by a FR system against the suspect's face. A study published by NIST [59] showed that the number of false match rates (FMR) varies between 10 to 100 times across different demographics, and that it is more common that a system performs a false match than a false non-match. Moreover, East African, West African and East Asian subjects have the highest false match rates. On the other hand, Eastern Europeans have the lowest.

2.3 Explainable Artificial Intelligence

Artificial Intelligence (AI) has evolved immensely in recent years, and to a point where it has been incorporated into a wide range of services and products. With the development of AI, systems became able to perform data-driven decisions and even compete with human performance when it comes to certain tasks. There are numerous examples of the applicability of AI in day-to-day tasks, such as the models that run in our smartphones and similar types of technology, and its relevance in areas such as banking and investments, as well as in law enforcement and security. Moreover, it has become quite essential in the medical field, helping doctors to diagnose and detect diseases, as well as helping in the outlining of the treatment process [60, 61].

The classic and simpler ML models used to solve more straightforward and basic problems have evolved into Deep Neural Networks (DNNs) that allow the extraction of valuable information from complex datasets [62]. In general, and after some testing, it was concluded that a deeper network is better at decision-making than a shallower one. However, the deeper the network, the more parameters it has and the more complex its design becomes. As the number of learnable variables increases, the data flow across the different network layers becomes more challenging to examine and understand. Therefore, on numerous occasions, the outcome from these models is quite difficult to interpret and accept by the users, given the black-box nature that became associated with these networks. On the contrary, simpler ML models are easier to comprehend and trust, given that a person, without much context, can try to understand the model by glancing at the chosen parameters without needing another model to provide an explanation. Decision trees and linear classifiers are examples of what is considered an easier-to-understand and more transparent model, and, in general, it is simpler for humans to reason with these models' final decisions. These represent an example of what can be called white-box or glass-box models [2].

Consequentially, it is important to understand the underlying reason behind the decision made by an AI algorithm, and more recently, the importance of this subject has arisen. This leads us to the need for eXplainable AI (xAI) methods in order to answer the issue presented above and, in general, to answer the following question: can it be explained what are the factors affecting a model and the reasoning behind the model's decision? It also becomes clear that it is essential to maintain a good balance between explainability and performance. In theory, the equilibrium between these two concepts is interesting and should be kept in mind, but in real-world applications, it becomes more challenging to set a limit and apply it [63].

As stated before, the ultimate goal for xAI techniques is to produce models with an adequate trade-off between explainability and performance or between interpretability and accuracy. Associated with xAI, there are two main concepts: explainability and interpretability. Even though there is an open discussion on the difference between the two, and some researchers may even consider them interchangeable, in this work, they are presented as separate and are considered to have different definitions [64]. Explainability became one of the main issues surrounding ML models, and it is responsible for bringing light to the decision-making mechanism while building users' trust and focusing on fairness and ethics. It is focused on the model's final decision and tries to present the users with the belief that AI is making a factual and non-biased decision. On the other hand, interpretability discloses the internal properties of the model and focuses on transparency. It is responsible for providing additional information or an explanation to help interpret the system's operation. Ultimately, it aids users with some knowledge to partially "open" the black-box model. By opening up a window into a black-box model, it is possible to expose some security vulnerabilities, create algorithms with human values and help individuals make more informed choices [2].

There are other important concepts related to xAI, represented in Figure 2.2. The relationship between them is also represented. Besides explainability and interpretability, other topics, such as transparency and robustness, are relevant and should be considered when analysing a model. Fairness is another term vastly used when diving into xAI, and it is especially relevant to this work. It is associated with the ability of a model to make unbiased decisions without favouring a specific population against the rest of the groups represented in the input data. Securing and maintaining a model's fairness is a challenge, given that in some databases and algorithms, some groups are represented and treated unfairly [2].

Figure 2.2: Relation between xAI's relevant concepts. [2]

Researchers came up with different ways to tackle the interpretability issue, and divided the various techniques into pre-, in-, and post-model methods, and even categorized them into post

hoc or not. The pre-model method focuses on the input data and the understanding of its distribution, and it is quite relevant to detect biases. Analysing and understanding the data used for training allows for increased confidence in the posterior decision. The in-model technique focuses on the integration of interpretability into the model. However, the centre of attention towards more accurate models has been on post-model interpretability. These techniques aim at generating explanations for a given prediction [28].

Attribution-based xAI methods are included in the post-hoc methods. In order to explain a model's predictions, xAI seeks to assign attributions to each input feature (e.g. pixels in an image input) in a way that makes their contribution clear. One example of an attribution-based method is the gradient-based one, which generates a "saliency map" responsible for indicating the contributions of each variable in the input space [2]. Researchers tried to come up with different explanation visualization techniques, one of them the class activation map (CAM) [65]: a weighted activation map generated for each image that, when generated for a specific class, discriminates image regions employed by the CNN to identify that class. Gradient-weighted CAM (Grad-CAM) [3] is also a visualization technique applicable to CNNs and does not require models to go through retraining or architectural modifications. These result from the combination of the feature maps using the gradient. The CNN layers capture both spatial information as well as high-level semantics, and the final CNN layer represents the optimal composition for extracting relevant data. In a nutshell, an importance score is attributed to each neuron by calculating the mean of the gradients of the logits of the target class concerning the feature activation maps of the final convolutional layer [60]. In Figure 2.3, there is a visual representation of what a gradient-weighted map may look like.

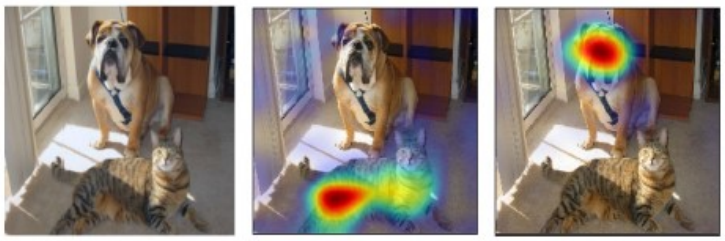


Figure 2.3: Example of a cat and dog Grad-CAM visualization, that was modified from Figure 1 of the Grad-CAM paper [3].

2.4 Machine Learning

2.4.1 Principal Component Analysis

Principal component analysis (PCA) is a popular technique for dimensionality reduction, analysing data containing a high number of dimensions/ features and increasing the interpretability of the data, while preserving the maximum amount of relevant information and enabling the visualization

of multidimensional data. This reduction is accomplished by linearly transforming the data into a new coordinate system. The principal components are the orthogonal axes along which the data exhibits the highest variance. These principal components are characterized by their associated eigenvalues and eigenvectors, e.g. the first principal component corresponds to the eigenvector with the highest eigenvalue. An eigenvector represents the direction in which data varies the most, while the eigenvalue indicates the magnitude of variance [66].

2.4.2 t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three- dimensional map. It works by modeling each high-dimensional object by a two or three-dimensional point in a way that similar objects are represented by nearby points and dissimilar objects by distant ones. It performs a non-linear dimensionality reduction allowing the separation of data that can not be separated by a straight line [67]. In Figure 2.4 there is an example of what a TSNE output plot can look like.

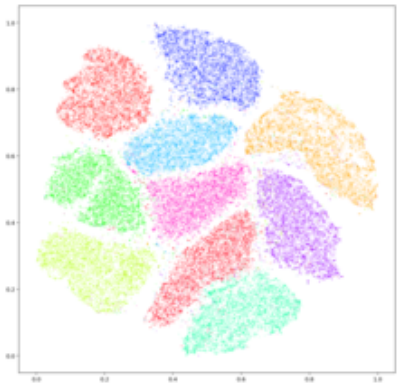


Figure 2.4: Example of a t-distributed stochastic neighbor embeddings on MNIST dataset.

2.4.3 Machine Learning classifiers

In machine learning, a classifier is an algorithm trained to make predictions about the class or category of a given input. The input data is usually represented as a set of features or attributes, and the goal of a classifier is to assign the input data to one of the predefined classes based on these features. These algorithms are often used in supervised learning, where a training dataset with labelled examples is provided. Each example has associated features and a corresponding label that indicates its true class, and the classifier is able to learn from them and consequently make predictions on new data.

2.4.3.1 Support Vector Machine

Support Vector Machine (SVM) is one of the most popular machine learning classifiers. This algorithm aims to find an hyperplane that separates data points of different classes in a high-dimensional space. The data points that are closer to the hyperplane are known as support vectors. The goal is to maximize the margin between classes while minimizing classification errors, which means that class points should be close to each other and as far away as possible to support vectors of each class (Figure 2.5). SVM can be used either for classification or for regression tasks.

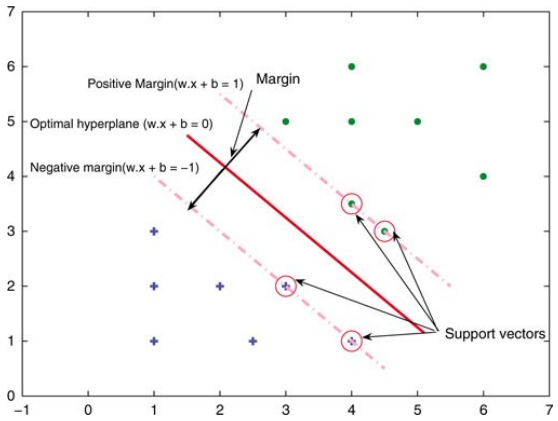


Figure 2.5: Support Vector Machine Hyperplane illustration in a two-dimensional input space based on margin maximization. [4]

2.4.3.2 K-nearest neighbors

This algorithm computes the distance between each training and test samples in the dataset and returns the k-closest training samples. It classifies data points based on the majority class among their k-nearest neighbors in the training data. Visual example in Figure 2.6.

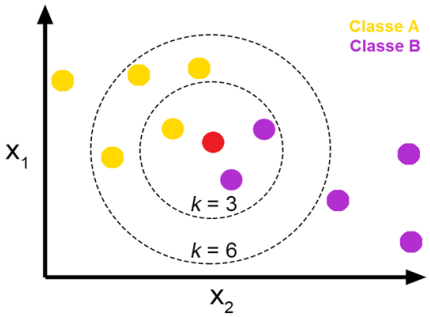


Figure 2.6: K-Nearest Neighbors Algorithm where the predicted class is typically the class most voted of its neighbours [5].

2.4.4 Deep Learning

Deep Learning is a sub-field of machine learning that comprises complex computational models composed of multiple processing layers that learn data with different levels of abstraction. At its core, deep learning involves the use of Artificial Neural Networks (ANNs) that perform tasks that require pattern recognition and data representation. These networks draw inspiration from the human brain regarding its structure and function, mimicking the use of interconnected layers of neurons to model data. These models have shown an outstanding performance in various domains, such as image and speech recognition, and are capable of handling large and complex datasets. Deep learning networks also show an ability to learn features and patterns from raw data [68].

The more recent breakthroughs in these matters come from the availability of powerful hardware (for example, GPUs), larger available datasets, and the development of innovative and more complex neural network architectures [13].

Nonetheless, the Deep Learning field is still evolving and new models and techniques are still emerging, adding to the wide range of neural networks architectures already existent. There are different models tailored for specific tasks and data types, one of them being the Convolutional Neural Networks (CNNs). CNNs are a specific class of Deep Neural Networks that are usually used for tasks that involve grid-like data, such as images and videos. One of the key advantages of CNNs applied to image processing is their ability to learn features in an hierarchical manner, meaning that starting from simple edges and textures the network proceeds to more complex shapes and object details.

The CNN's architecture involves some fundamental blocks that are briefly presented here:

- **Convolutional layer**: the "core" building block of CNNs that extract features from the input data. Each layer suffers a convolution between the input and a filter (or kernel) to obtain a feature map. After this step, these are fed to the next layer as the new input data. The filters are usually smaller than the input image and act as a sliding window over them. In Figure 2.7 it is illustrated an example of a convolution (a) and the equivalent transposed convolution (b) with a 3 x 3 filter kernel applied to a 4 x 4 feature map. The regions that were used to compute the output are coloured green.

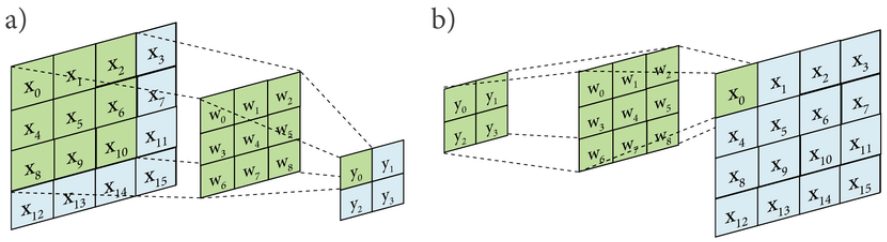


Figure 2.7: Schematic of a discrete convolution (a) and the equivalent transposed operation (b). [6]

- **Pooling layer**: non-linear down-sampling; reduces the dimensionality of the feature map, while maintaining important information. There is Average Pooling and Max-Pooling. The first one receives the number of values in the window of the layer, defined by a specific value and calculates the average value. Max-Pooling outputs the maximum value instead of the average. The representation of both can be found in Figure 2.8.

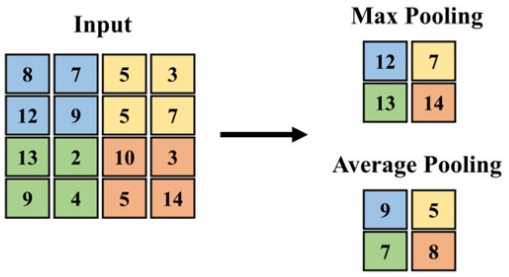


Figure 2.8: Pooling operation done by a Max-Pooling and Average Pooling. Adapted from [7].

- **Rectified Linear Unit (ReLU) layer**: this layer performs the activation function, returning 0 if the input value is less than 0 and the input value itself otherwise. Its purpose is to improve the non-linearity of the image's pixel data. Usually, in a CNN, this layer is applied after a convolutional layer and precedes a max-pooling.

- **Fully Connected Layer (FC)**: it is responsible for the high-level reasoning after applying various convolutional layers and max-pooling layers. In most deep learning models, the last layers are fully connected layers to compile the data extracted by the other layers and compute the output.

In Figure 2.9, there is a schematic representation of a Deep Convolutional Neural Network architecture for an example of image classification. The cat picture is the input image and the represented layers are the ones previously described.

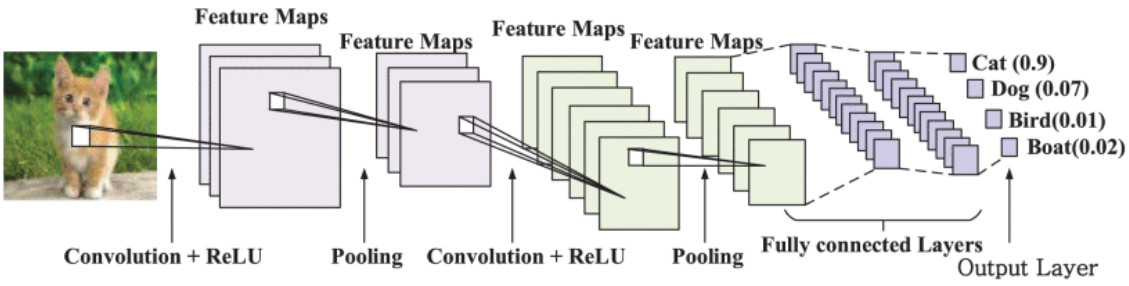


Figure 2.9: An example of a deep convolutional neural network architecture. From [7].

2.5 Summary

Concluding this chapter, the main concepts to retain:

- Neural adaptation in humans shows that familiarization to a certain stimuli can happen, resulting in a diminished neural activity. The more accustomed we are with a certain race makes seeing faces from said race generate a lower neural activity than with faces from other racial groups.
- A face recognition system can perform verification (1:1 comparison) or identification (1:N comparison).
- Racial bias exists in face recognition algorithms, affecting the model's performance in a discriminative way.
- Face recognition systems are currently black-box systems and it is important to focus on increasing the explainability and interpretability of models. One way to help with the interpretation of a model can be the use of gradient maps to visualize the contributions of each variable in the input space.

Chapter 3

State-of-the-Art

3.1 Automatic Face Recognition

The face recognition problematic was summed up by Jafri and Arabnia [19] as follows: "Given both an input face image as well as a database of face images, how can it be verified the identity of the person in the input image?". The history of face recognition dates all the way back to the 1950's, but research on an automatic approach is considered to be from around the 1970's. In the preliminary works, researchers used features based on distances between the more important regions of the face. Around the year of 1990, as a result of the development in hardware and the increasing importance of security applications, more studies around this issue were published [8].

Wang and Deng [12] divided the progress of image-based face recognition techniques into four main conceptual development phases: i) Holistic or appearance-based approaches that use the face region as a whole and use both linear and non-linear methods to map the face into a lower dimensional subspace; the work of Turk and Pentland [69, 70] represents one of the first successful methods developed, known as Eigenfaces. Other approaches use linear subspaces [71], manifold learning [72] and sparse representations [73]. ii) Local-feature based face recognition algorithms that use hand-crafted features to describe the face, as per example local binary patterns and variants [74]. iii) Methods that use learning-based local descriptors and that learn the discriminant image filters [75]. iv) Deep Learning methods, that started to pick up popularity after the great success of AlexNet in the ImageNet competition in 2012 [62]. These methods brought a new perspective to the face recognition problem and lead to the achievement of performances similar to humans on large-scale datasets [76].

Face recognition systems can be divided into two main groups: image-based or video-based methods. The first one tries to recognize a person by the physical appearance of the face, and the second one uses both appearance as well as changes in the dynamics of the face through time [8].

FR systems usually consist of six steps, that are represented in Figure 3.1.

As seen in the output portion of the image (Figure 3.1), face recognition can be either a identification problem or a verification one [8, 77, 78]. Face identification is viewed as a 1:N matching issue, where the query face is placed side-by-side with all the other faces in the database of known

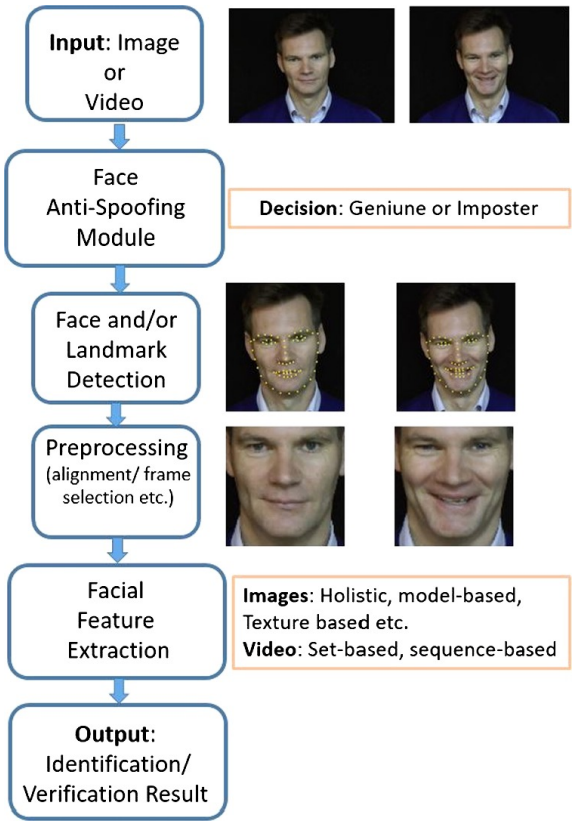


Figure 3.1: Main steps in a Face Recognition system. Illustration from [8].

identities and the decision is made as a result of the comparison between them. On the other hand, face verification is a 1:1 matching problem, where the identity of the face in question is either rejected or confirmed when compared with the data available of the claimed identity on the database [8, 78].

In the work of Ranjan *et al.* [9], FR is sectioned into three modules, where the first is the face detector, the second is responsible for performing facial landmarking and alignment of the detected faces to normalized canonical coordinates, and, finally, the last module implements the actual face recognition step. The system is layed out in Figure 3.2. In the FR module, face anti-spoofing is responsible for recognizing whether the face is real and live or spoofed [79, 80]. Face processing is used to handle variations in, per example, illumination, occlusions, poses and age. In the training step, discriminative deep features are extracted; after, face matching is responsible for feature classification in the testing data.

The typical pipeline of a FR system involves mapping the face after detection and alignment into a feature vector or embedding. Two face images are compared by their relative embeddings and the degree of identity similarity is measured. The embeddings ideally should present a small intra-class and large inter-class variation. To achieve this goal, different solutions opted to train Deep Neural Networks by either directly learning the embedding on the latent space (e.g. using triplet loss) or by learning an identity classification problem (e.g. Softmax loss). Although it

is important to understand face detection and face alignment- both the process as well as the evolution of the applied methods-, this work mainly focus on face recognition systems.

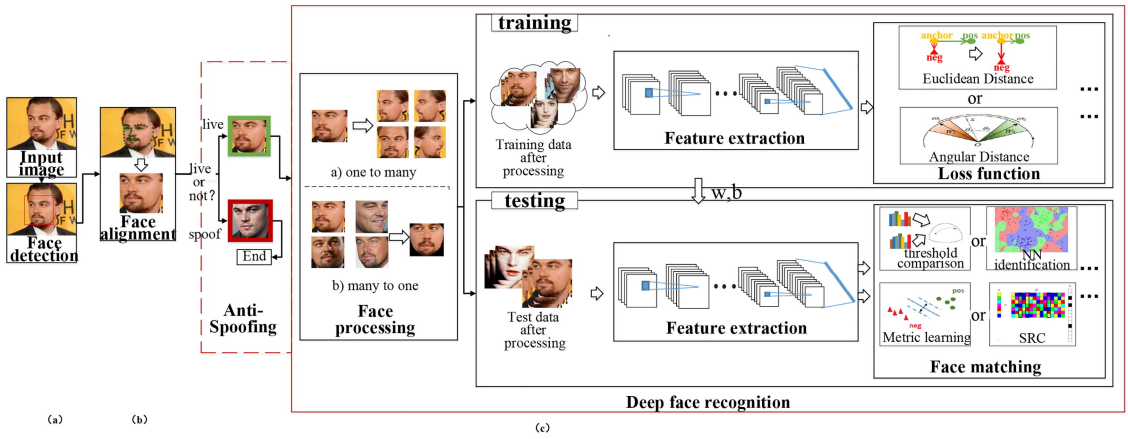


Figure 3.2: Deep FR system with face detector (a) and alignment (b). The last module, FR (c) is the subject of this work. Illustration from [9].

3.1.1 Face detection

Face detection is an essential step in a FR system, and it is responsible for estimating the bounding-box of the face (or faces) in an image or video-frame. This step should be able to deal with changes in pose, illumination and scale, and should also act as indifferent to the background of the image as possible. The cropped image that results from the detection aids the model to find and extract the essential features to make an accurate prediction [8].

The Viola-Jones [81] is a widely used face detector algorithm, specially on frontal face images. It is based on Haar-like features, and while it works in real-time, it presented problems dealing with occlusions and significant pose variations. It was considered state-of-the-art for many years, and it remains relevant to this day when dealing with RGB images. Most early works focused on designing robust features and training effective classifiers. Besides the Haar-like features used in the Viola-Jones case [81], the features could also be pixel-based, and researchers used relations between adjacent pixels to find the boundaries of faces [82]. Moreover, statistics-based features were also used: from spatial histograms (LBP-based) [83], to a combination of histogram oriented gradients and local binary patterns (HOG-LBP) [84], or even edge orientation [85]. Nevertheless, these face detection algorithms relied on hand-crafted features and on the separate optimization of each component involved, making the face detection step less than optimal.

There are other relevant approaches, but recently the focus has been on Deep Learning based methods, that present extremely good results [9]. Some of these methods where initially used for object detection, e.g. the single shot detector (SSD) [86], as it has been considered that face recognition is a more specific case of object recognition, where the object in question is not rigid, but rather variable [19]. Convolutional Neural Networks (CNNs) are specially pertinent, with remarkable successes in image classification and object detection. As CNNs are built to learn invariant

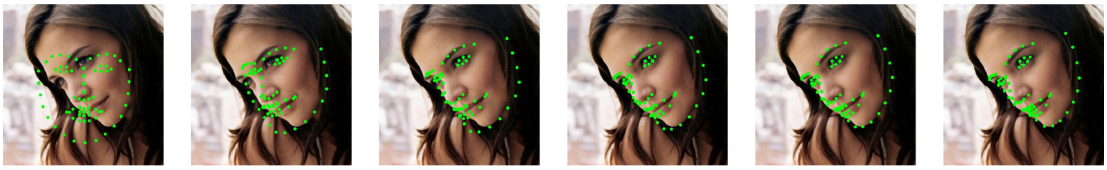
representations of images, they became fit to deal with pose variations, changes in illumination, different angles and occlusions. Ultimately, these models use their layers to extract features without having to manually define where to look. Consequently, these new and powerful algorithms surpassed the performance of the traditional models and quickly became state-of-the-art [8].

3.1.2 Facial Landmarking and Alignment

Following the FR pipeline, after face detection follows the estimation of certain landmarks, such as the corners of the eyes, eyebrows, and mouth. These represent the most relevant points of the face to perform face alignment, which has been proven to be beneficial for face recognition. The aim is to estimate these landmarks in order to align the face into a canonical position [8].

As with face detection, there are various methods that were researched for performing face alignment and facial landmarking. The studies on facial landmarking are summarized in various survey papers [87] [88] [89]. Aiming to evaluate the landmark localization performance, in [88] two different metrics were used: the ground truth based localization error and task-oriented performance. As expected, due to the most recent advances in Deep Learning, the performance of facial landmarking methods improved and evolved [8].

Face alignment consists of locating semantic facial landmarks such as eyes, nose, and mouth, and is essential for tasks like face recognition, face animation and 3D face modeling [90]. The classic face alignment methods, e.g., Active Shape Model (ASM) [91] [92] or Active Appearance Model (AAM) [93] [94] search for landmarks based on the global shape models. The latter uses an appearance model to reconstruct the entire face and estimates its shape by minimizing the texture residual. However, the learned appearance models have limited power to capture complex and subtle face image variations in pose, expression, and illumination, which may not work on unseen faces. Regression-based methods learn a regression function that maps image appearance to the target output [90]. In the work of Cristinacce and Cootes [92], learned regressors are used for individual landmarking. As only local image patches are used for training and there is no exploitation between landmarks, these learned regressors can be considered weak and have trouble handling pose variations. Figure 3.3 illustrates the results of face alignment by cascaded regression [10].



(a) (b) (c) (d) (e) (f)

Figure 3.3: Illustration from [10] with the results of face alignment in different stages of cascaded regression. The shape estimate is initialized and later, iteratively, updated through a cascade of regression trees: (a) is the initial shape estimate and (b) to (f) are estimations at different stages.

The work of Sun *et al.* [95] was pioneer in trying to apply Deep Convolutional Networks to the face alignment task. Afterwards, with the work of Zhang *et al.* [96], deep CNNs started being widely exploited. Methods capable of performing multi-task learning- which can combine face detection and landmark localization with other tasks, such as pose estimation-, became relevant. One approach to multi-task learning is the MTCNN- "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks" [11]-, which uses multitask cascaded convolutional networks. The model is divided into 3 main parts: (1) a fully CNN by the name of Proposal Network (P-Net) that aims to find the bounding box for the detection, (2) a refine network (R-Net), that is another CNN whose goal is to reject a large number of false candidates, and the last stage, stage (3) that is similar to the previous one, but that presents a special focus on identifying face regions with more supervision. The visual illustration of the architecture can be found in Figure 3.4. This model reaches a performance of 99.83% in the Labeled Faces in the Wild (LFW) dataset [97], surpassing the human performance for this same dataset: 97.53%.

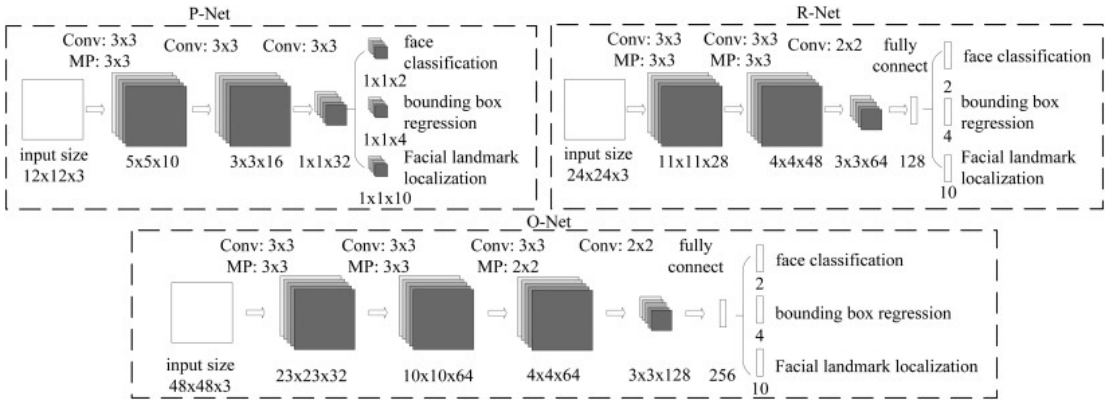


Figure 3.4: Architecture of P-Net, R-Net, and O-Net from the MTCNN. In the image "MP" stands for max pooling and "conv" for convolution. Illustration from [11].

3.1.3 Network architecture and Training loss

Since the very early stages of FR, the mainstream network architectures have always followed those of object classification, as per the example of the well known case of AlexNet [62]. This algorithm achieved the SOTA recognition accuracy during the ImageNet large-scale visual recognition competition (ILSVRC) in the year of 2012, and in summary consists of five convolutional layers and three fully connected layers.

In 2014, VGGNet [98] presented a standard network that consisted of small convolutional filters throughout and doubled the number of features maps after pooling. This work increased the flexibility to learn progressive nonlinear mappings. In contrast with this network of 16-19 weight layers, GoogleNet [99] came up with a 22-layer network, and its main trademark is the improved utilization of computing resources. Both the depth and the width of the network were increased while keeping the computational cost constant.

In 2016, ResNet [23] became known for proposing layers that learn a residual mapping with reference to the layer inputs, easing the training needed for deeper networks. ResNet, short for Residual Network, was introduced to try and solve the vanishing gradient problem: when there are multiple layers in a network, as the gradient is backpropagated to earlier layers, the repeated multiplication process makes the gradient infinitely small. There are many variants of the ResNet architecture as per the ResNet-18, ResNet-34, ResNet-50, and ResNet-100 [5]. By 2017, SENet [100] surpassed the previous SOTA recognition accuracy and won the ImageNet large-scale visual recognition competition (ILSVRC). The novelty laid in a "Squeeze-and-Excitation" (SE) block, that could be integrated with other architectures, such as ResNet, improving them. This SE block adaptively recalibrates feature responses by modelling interdependencies between channels. In Figure 3.5, it can be found a chronological representation of the most influential architectures in deep FR that were previously described.

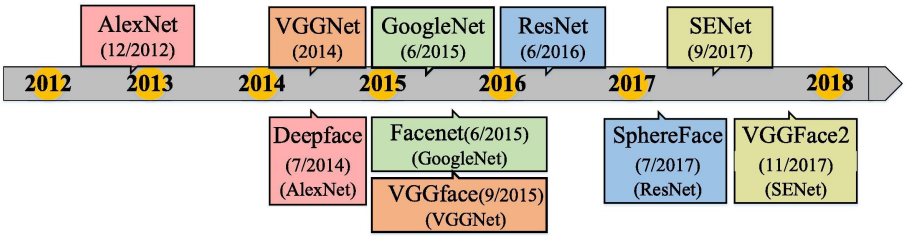


Figure 3.5: The top row shows the typical network architectures, and the bottom row presents the well-known FR algorithms that use the above architectures. The algorithms that use the same architecture have the same color block. Illustration from [12].

A schematic representation (Figure 3.6) of the previously presented networks shows the network architecture for AlexNet, VGGNet, GoogleNet, ResNet and SENet.

One way to improve the generalized performance of face recognition is to include as many identities as possible in the training set. For example, both Facebook and Google have a deep FR system trained by a gigantic number of IDs: around 10^6- 10^7. However, these very complex datasets are not accessible for the general public, which means that researchers have to look for other ways to make deep features more discriminative. Therefore, in order to significantly improve FR methods, there have been great efforts to develop different loss functions that can enhance the discriminative power [101]. The solutions to train Deep Neural Networks can be either by directly learning the embedding or by learning an identity classification problem. The former is associated with Triplet loss and the latter with Softmax loss [15].

After the development of AlexNet [62], that used cross-entropy based softmax loss for feature learning, both DeepFace [76] as well as Deep ID [102] adopted this same method. However, researchers realized that this approach to calculate loss is not sufficient to learn discriminative features. As a result, the possibility to come up with a novel loss function became the focus of the research in FR.

Before the year of 2017, Euclidean-distance-based loss [103] played a very important role in loss functions. It is a metric learning method, that embeds images into Euclidean space, and tries

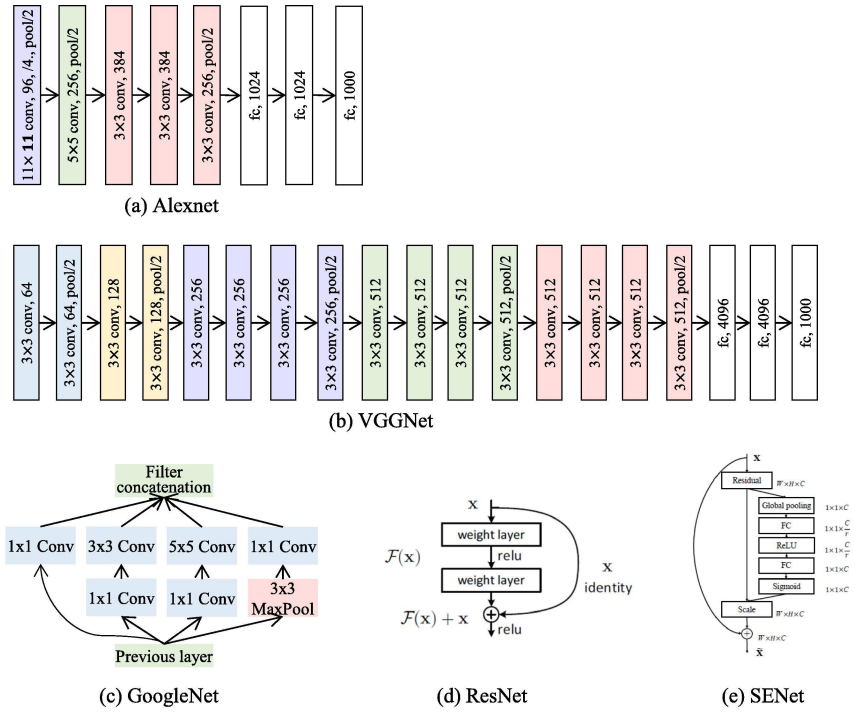


Figure 3.6: Different network architectures. Illustration from [12].

to reduce intra-variance and enlarge inter-variance [12]. The contrastive loss and triplet loss are the loss functions usually associated. The first one requires face images pairs, pulling together the positive ones and pushing apart the negative pairs, considering the absolute distances of the matching and non-matching pairs. The contrastive loss function can be calculated by applying the following equation, where $y_{ij} = 1$ means x_i and x_j are matching samples and $y_{ij} = 0$ means non-matching, $f(.)$ represents the feature embedding, and ε^+ and ε^- control the margins of the matching and non-matching pairs respectively [12]:

$$\mathscr{L} = y_{ij} max(0, ||f(x_i) - f(x_j)||_2 - \varepsilon^+) + (1 - y_{ij}) max(0, \varepsilon^- - ||f(x_i) - f(x_j)||_2) \tag{3.1}$$

The work behind DeepID2 [104] combined softmax for face identification as well as contrastive loss for face verification. Moreover, joint Bayesian (JB) was applied as a way of obtaining a robust embedding space. Both DeepID2+ [105] as well as DeepID3 [106] were extensions of DeepID2 [104], and represented a introduction to VGGNet and GoogleNet. However, the main problem with using contrastive loss is that the margin parameters are very difficult to choose.

On the contrary, triplet loss considers the relative difference of the distance between matching and non-matching pairs. Google proposed FaceNet [53]: a GoogleNet trained in a large private dataset that achieved a performance of 99.63%. It requires face triplets, minimizing the distance between an anchor and a positive sample of the same identity and maximizing the distance between the anchor and a negative sample of a different identity. FaceNet used $||f(x_i^a) - f(x_i^p)||_2^2 + \alpha <$

$-||f(x_i^a) - f(x_i^n)||_2^2$, with x_i^a, x_i^p, and x_i^n as the anchor, positive and negative samples respectively; α is a margin. Due to training instability when using triplet loss, researchers kept looking for other alternatives, such as center loss and variants. Center loss learned a center for each class and penalized the distances between deep features and corresponding class centers [12].

Therefore, around the year of 2017, angular/ cosine-margin-based loss became popular, as well as feature and weight normalization. The softmax loss, that uses $(W_1 - W_2)x + b_1 - b_2 = 0$ as the decision boundary, with x as the feature vector, W_i as the weights, and b_i as bias, was reformulated into a large-margin softmax loss (L-Softmax) [107]. In L-Softmax $b_1 = b_2 = 0$, which means that the decision boundaries for classes 1 and 2 become $||x||(||W_1||cos(m\theta_1) - ||W_2||cos(\theta_2)) = 0$ and $||x||(||W_1||\ ||W_2||cos(\theta_1) - cos(m\theta_2)) = 0$, respectively, where m is a positive integer that introduces an angular margin, and θ_i is the angle between W_i and x. Fixed on L-Softmax [107], A-Softmax [108] loss normalizes the weight W to a greater extent, in a way that makes the normalized vector to be able to lie on an hypersphere. Consequently, the discriminative face features can be learned on the hypersphere manifold with an angular margin. SphereNet [109] was introduced as a deep hyperspherical convolution network that adopts an angular-margin-based loss. However, to try to overcome the optimization issue in both L-Softmax as well as A-Softmax, which incorporate the angular margin in a multiplicative manner, ArcFace [101] and CosFace [110] use an additive angular/cosine margin. The general angular margin penalty-based loss (L_{AML}) is defined as follows [15]:

$$L_{AML} = \frac{1}{N}\sum_{i\in\mathbb{N}} -\log \frac{e^{s\left(\cos(m_1\theta_{y_i}+m_2)-m_3\right)}}{e^{s\left(\cos(m_1\theta_{y_i}+m_2)-m_3\right)} + \sum_{j=1,j\neq y_i}^{c} e^{s\left(\cos(\theta_j)\right)}}, \tag{3.2}$$

where m_1,m_2 and m_3 are the margin penalty parameters for SphereFace [108], ArcFace [101] and CosFace [110], respectively. CosFace proposed $m_1 = 1$, $m_2 = 0$, and $m_3 = \alpha(0 < \alpha < 1 - \cos(\frac{\pi}{4}))$, making its decision boundary $\cos(\theta_{y_i}) - \cos(\theta_j) - m_3 = 0$. Later, ArcFace setted up $m_1 = 1$, $m_2 = \alpha$, and $m_3 = 0(0 < \alpha < 1.0)$. Therefore, its decision boundary is $\cos(\theta_{y_i} + m_2) - \cos(\theta_j) = 0$.

Both the work of FairLoss [111] and AdaptiveFace [112] proposed further adjusted margins in order to address the problem of unbalanced data. Even though angular/cosine-margin-based loss adds discriminative constraints on a hypershpere manifold and achieves good results on a clean dataset, it still shows some vulnerability to noise and performs worse than center loss or softmax in a high-noise region [113].

As stated previously, during 2017 some works focused on trying to improve model performance by normalizing the features and weights in loss functions. Normface [114] explained the need for this normalization, both analytically as well as geometrically.

In Table 3.1, we can see the difference in accuracy of various methods evaluated on the LFW dataset (described in more detail in the following section), with information on the loss function as well.

Table 3.1: The accuracy of different methods evaluated on the LFW dataset. Adapted from [12]

Method	Public. Time	Loss	Architecture	Number of Networks	Training Set	Accuracy±Std (%)
DeepFace [76]	2014	softmax	AlexNet	3	Facebook (4.4M, 4K)	97.35 ± 0.25
DeepID3 [106]	2015	contrastive loss	VGGNet-10	50	CelebFaces+ (0.2M, 10K)	99.53± 0.10
FaceNet [53]	2015	triplet loss	GoogleNet-24	1	Google (500M, 10M)	99.63 ± 0.09
VGGface [115]	2015	triplet loss	VGGNet-16	1	VGGface (2.6M, 2.6K)	98.95
L- softmax [107]	2016	L- softmax	VGGNet-18	1	CASIA-WebFace (0.49M, 10k)	98.71
L2- softmax [116]	2017	L2- softmax	ResNet-101	1	MS- Celeb- 1M (3.7 M, 58K)	99.78
SphereFace [108]	2017	A- softmax	ResNet-64	1	CASIA-WebFace (0.49M, 10k)	99.42
CosFace [110]	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49M, 10k)	99.33
ArcFace [101]	2018	arcface	ResNet-100	1	MS- Celeb- 1M (3.8 M, 85K)	99.83

3.2 Databases

Early facial recognition research relied on relatively small databases that were compiled in carefully controlled lab settings, as per the example of ORL [117]: one of the first image-based databases that contained 400 images from 10 subjects. Some samples taken from the ORL can be seen in Figure 3.7. Alongside this, one of the first ever video-based face databases was released in 1997, and included 70 videos from 40 subjects.

Figure 3.7: Samples from the Olivetti Research Laboratory database. Illustration from [13].

Through the years, databases have become progressively larger with millions of images or videos captured under uncontrolled conditions. The development of more complex databases facilitates the research in FR and aids the field's evolution, given that some of the more simple databases have became saturated, e.g. Labeled Faces in the Wild (LFW) [97]. LFW was first introduced in 2007 and marks the beginning of FR performed on images under unconstrained conditions (Figure 3.8).

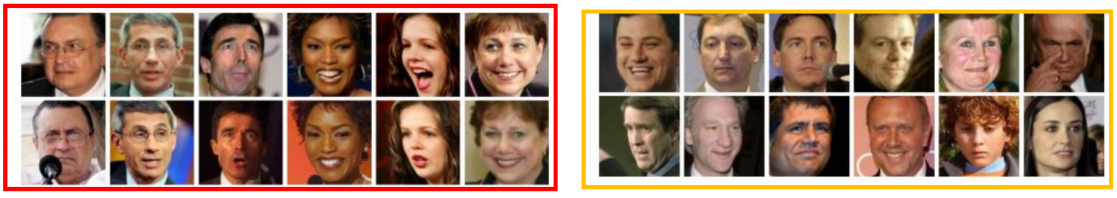


Figure 3.8: Samples from the Labeled Faces in the Wild database. Illustration from [13].

Especially in the early works, deep FR algorithms were trained on private training datasets. Internet giants such as Google and Facebook presented FaceNet [53], a model trained on 200 M images of 3 M subjects; and DeepFace [76], a model trained on 4 M images of 4 K subjects,

respectively. Even though there were some relevant results from both of these, researchers are not able to reproduce or even compare their methods without access to a public training dataset. With this issue in mind, CASIA-Webface [118] offered the first widely-used public training dataset for deep model training. It comprises of 0.5 M images of 10 K celebrities gathered from the web, and it is a good resource for academic comparisons, given its moderate size and easy usage. However, when it comes to advanced deep learning algorithms, it is not sufficient in terms of data and ID size. Celeb-1 M [119], VGGface2 [120] and MegaFace [121] are three datasets with over 1 M images that became relevant for large-scale training. There are some samples taken from the VGGFace2 dataset represented in Figure 3.9. In Table 3.2, the main image-based databases used to perform FR are presented.

Figure 3.9: Samples from the VGGFace2 database. Illustration from [13].

When speaking about face databases, it came to researcher's attention that it may be pertinent to look for a good trade-off between the depth and the breadth of the dataset. While the depth of a database addresses a wide range intra-class variation, such as changes in lighting, age, and pose, the breadth ensures that the trained model covers a sufficiently variable appearance of various subjects. VGGface2 [120] includes a large-scale training dataset with depth, containing a limited number of subjects, but many images for each one of them. On the other hand, MS-Celeb-1 M [119] and MegaFace [121] represent good examples of large-scale training datasets with breadth, covering many subjects, but with limited images per subject. The work of Cao *et al.* [120] revealed that it may be beneficial to use a model that was first trained on MS-Celeb-1 M [119] and after on VGGface2 [120], which is representative of focusing first on breadth and then depth.

Another important factor to a large, clean and meaningful database is the data noise that results from the data source and collecting strategies. In large-scale datasets, it is expected that label noise may be present. In the research work followed by Wang *et al.* [113] the noise percentage increases along the data scale, which means that more data reflects in more noise. Moreover, label flip noise affects the performance of a model, especially if the model uses A-softmax [108] for the loss. One of the approaches to solve this issue was presented by Deng *et al.* [101] and starts by cleaning the noise found in MS-Celeb-1 M [119] before making the new clean dataset public (MS1MV2). Another approach [128] shifted the focus from noise clearance to adding more unlabeled data.

Table 3.2: The most commonly used FR datasets. Adapted from [12] and [8]

Datasets	Publish Time	#photos	#subjects	# of photos per subject	KeyFeatures
ORL [117]	1994	400	10	-	all pictures are frontal
Indian Face [122]	2002	440	40	-	images taken with a bright homogeneous background and subjects in an upright, frontal position.
LFW [97]	2007	13233	5749	-	un-posed photos, mainly frontal views.
CASIAWebFace [123]	2014	494,414	10,575	2/46.8/804	celebrity
CelebFaces+ [102]	2014	202,599	10,177	19.9	private
Facebook [76]	2014	4.4M	4K	800/1100/1200	private
VGGFace [115]	2015	2.6M	2,622	1,000	depth; celebrity; annotation with bounding boxes and coarse pose
Google [53]	2015	>500M	>10M	50	private
MS-Celeb-1M (Challenge 1) [119]	2016	10M3.8M (clean)	100,00085K (clean)	100	breadth; central part of long tail; celebrity; knowledge base
MS-Celeb-1M (Challenge 2) [119]	2016	1.5M (baseset) 1K (novelset)	20K (baseset) 1K (novelset)	1/-/100	low-shot learning; tailed data; celebrity
MegaFace [124]	2016	4.7M	672,057	3/7/2469	breadth; the whole long tail; commonalty
CFPW [125]	2016	7000	500		frontal-profile images of celebrities
VGGFace2 [120]	2017	3.31M	9,131	87/362.6/843	depth; head part of longtail; crosspose ,age and ethnicity; celebrity
MS-Celeb-1M (Challenge 3) [119]	2018	4M (MSv1c) 2.8M (Asian-Celeb)	80K (MSv1c) 100K (Asian-Celeb)	-	breadth; central part of long tail; celebrity
IMDB-Face [113]	2018	1.7M	59K	28.8	celebrity
CPLFW [126]	2018	11652	3968	-	celebrity
MillionCelebs [127]	2020	18.8M	636K	29.5	celebrity

3.2.1 Data bias

As per the example of CASIA-WebFace [118] and MS-Celeb-1 M [119], large-scale training datasets are usually a result of images obtained from websites and Google Images. Most of the images consist of celebrities in formal settings- smiling, with make-up, young and well put together-, given that there is easy access to them on-line. These types of databases are very different from the ones that try to include images captured in the daily life, e.g. MegaFace [121]. The present biases can be attributed to outside factors in the data collection process, such as cameras, lighting, and different backgrounds.

Dataset biases negatively impact cross-dataset generalization, which means that the performance of a model that was trained on a certain dataset drops significantly when switched to another one [129].

When it comes to demographic bias, especially related to race/ethnicity, gender and age, it is consensual that it is a universal and a very urgent issue that needs to be solved. It is very common that in databases used for training and testing Deep Learning models, the male, White and middle-aged subject appears more frequently. The model replicates and may even amplify the biases, which influences heavily its performance and accuracy results when applied to other demographic groups [12]. Some previous works on this matter [130] [131] showed that the female, Black, and younger subgroup of the mainly-used databases are usually more difficult to recognize when applying a FR system. Phillips *et al.* [132] presented a work where they evaluated FR algorithms on the FRVT 2006 [133] images and concluded that these performed better on natives. Another similar study [56] collected mug shots of White, Black and Hispanic subjects and concluded that the Black cohorts are more difficult to recognize. Moreover, the commonly-used databases for deep FR, as LFW [97], do not include significant racial diversity. This issue is visually represented in Table 3.3.

Table 3.3: Statistical demographic information of the most commonly-used training and testing datasets. From [12].

Train/ Test	Database	Race(%)				Gender(%)	
		Caucasian	Asian	Indian	African	Female	Male
train	CASIA-WebFace [118]	84.5	2.6	1.6	11.3	41.1	58.9
	VGGFace2 [120]	74.2	6	4	15.8	40.7	59.3
	MS-Celeb-1M [119]	76.3	6.6	2.6	14.5	-	-
	BUPT-Balancedface [29]	25	25	25	25	-	-
	BUPT-Globalface [29]	38	31	18	13	-	-
test	LFW [97]	69.9	13.2	2.9	14	25.8	74.2
	IJB-A [134]	66	9.8	7.2	17	-	-
	RFW [14]	25.9	21.8	26.1	26.2	-	-

Therefore, as a way to try and tackle racial bias, an issue that has not been thoroughly studied yet, and prove that the SOTA algorithms work unequally with different races, Wang *et. al* [14] presented a Racial Faces in the Wild (RFW) dataset. Some examples of images from the RFW dataset can be found in Figure 3.10.

Figure 3.10: In rows, from the top to bottom: Caucasian, Indian, Asian, African. Illustration from [14].

It is a race-balanced training database, where all four represented races- Caucasian, African, Asian, and Indian-, have the same number of images (each one has 25% of all the images in the dataset). The images were collected from the MS-Celeb-1 M [119] database, and the "Nationality" attribute was used to collect the images for Asians and Indians. When it comes to Caucasians and Africans, the Face++ API was used to estimate the subject's race, and the images with a low confidence estimation score were manually checked [14]. In Table 3.4 the different values for face verification accuracies on the RFW database are presented. These results confirm the existence of racial bias in recognition APIs and FR algorithms.

Table 3.4: Racial bias in FR algorithms. Face verification accuracies (%) on the RFW database. Adapted from [12].

Model	LFW	RFW			
		Caucasian	Indian	Asian	African
Microsoft	98.22	87.60	82.83	79.67	75.83
Face++	97.03	93.90	88.55	92.47	87.50
Center-loss [135]	98.75	87.18	81.92	79.32	78.00
Sphereface [108]	99.27	90.80	87.02	82.95	82.28
Arcface [101]	99.40	92.15	88.00	83.98	84.93
VGGface2 [120]	99.30	89.90	86.13	84.93	83.38

The presence of such bias may result in mistreatment of certain demographic groups, exposing them to a higher risk of fraud or making their access to services more difficult. This led to the FR

community to agree that addressing these data bias and enhancing fairness in FR systems is urgent and necessary. Consequently, collecting balanced data to train a fair model or even trying to design a unbiased algorithm should be the way to go [12].

3.3 Summary

This chapter presents a literature review of both face recognition algorithms as well as databases associated with face recognition. There are different network architectures and training loss methods presented and proposed through the years. Regarding databases, the focus is the introduction of a race-aware dataset: RFW.

Chapter 4

Methodology

4.1 Developed Strategy

This chapter will detail the work developed towards accomplishing the established objectives. All the work carried out and portrayed in this dissertation was, from the start, very exploratory and based on an investigative approach. Accordingly, the various practical tests performed were designed as the experiment work progressed and their results led us towards certain directions.

We started by defining the model to be studied and present the motivation behind this choice. We further explored the model properties and the databases used to benchmark said model. Additionally, the model was benchmarked on a database designed to assess race bias.

Afterwards, different tests were conducted on the images of the chosen datasets, aiming to assess the influence of certain alterations (e.g. changes in contrast and illumination, rotation and grayscaling) had in the method's performance and, therefore, test which alterations affected more a particular race. By performing incremental changes to the illumination of an image, per example, the goal was to evaluate if one of the four racial groups felt these modifications more prominently than the others.

In order to try and understand what the model considered to be the most relevant spatial areas used to construct a face embedding in each image for each race, mean gradient maps were extracted in three specific layers of the network- first layer, mid-layer and last layer. The idea behind this step was to enable comparisons between the regions of interest on images from the same layer, but belonging to different ethnicity groups. Differences in patterns may also indicate different levels of neural activation, which allowed the transition for the following steps.

The next part of the work focused on the idea that there could be a parallelism between the human face recognition process and an automatic FR system. Regarding how different brain regions are activated through time with different intensities, the goal was to look for similarities in the FR model's network pipeline and how the data flows throughout the layers in a specific order. The values for all the layers' neural activations were retrieved and these results were analysed in order to look for a sequential layout similar to the neural activations in the human brain. Metrics such as the mean and standard deviation were calculated for all the network's layers and analysed

to try to evaluate differences across racial groups. Furthermore, it allows us to study if the bias is already identifiable with these statistics, e.g. a lower performance means less discriminative power, which can be caused by a lower STD value in a specific group.

At this point, there was an interest to test the influence that the network architecture used had in the results, especially given that the chosen algorithm was based on a ResNet-100 trained on the MS1MV2 dataset, and some of the images on RFW can be found in both databases (data leakage). Therefore, we further explored three instances of a smaller network (ResNet-34), trained on three different databases [136]. The ResNet-34 trained on MS1MV2 was retained as a way to perform comparisons with ResNet-100 on the same data, ensuring that we would only be studying the impact of the architecture change. We replicated previous experiments on these networks.

Finally, the last section of this dissertation aims to show that the bias on face recognition is related to intra-race samples and not inter-race. As such, we show that there is a progressive separability of ethnicity groups across the layers of the network.

4.2 Algorithm Selection

The first step was to select the algorithm that would be used throughout the entire duration of the practical experiments. As previously highlighted in Chapter 3, there were various options for good FR-performing models that could have been chosen.

The Additive Angular Margin Loss is intended to improve the discriminative power of a model and add stabilization to the training process, however, as stated before, the main challenge in loss functions such as ArcFace [101], CosFace [110] and SphereFace [108] is selecting the ideal margin penalty value. In these three methods, the margin was selected through trial and error and the authors lay on the assumption that the samples are equally distributed on the geodesic space around the class centers. Given that this assumption could not be held with largely different intra-class variations between the samples in the training dataset, Boutros *et al.* proposed ElasticFace [15]. ElasticFace brings a looser single margin value by deploying a random margin drawn from a normal distribution, aiming to improve face recognition accuracy by targeting enhanced intra-class compactness and inter-class discrepancy.

For the present work, it needs to be highlighted that, even though ArcFace could be considered as a possible algorithm option, the results for the pre-trained model weights are not available in PyTorch [137], meaning that there is no checkpoint associated for easy use of the pre-trained model. This is a relevant factor to consider, given that there is a preference on using this deep learning library to implement the model. ElasticFace's repository on GitHub [1] includes checkpoints for all the available models, making testing easier by having official pre-trained versions. Moreover, when comparing the results from ElasticFace with fixed margin penalty methods and recent state-of-the-art, the former enhanced the face recognition accuracy and increased performance on seven out of the nine benchmarks used [15].

[1] https://github.com/fdbtrs/ElasticFace

ElasticFace random margin penalty can also be integrated into the angular margin-based softmax losses, e.g. CosFace and ArcFace, and it is extended over the angular margin by deploying random values from a Gaussian distribution. The probability density function of a normal distribution is defined as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{4.1}$$

where μ is the mean and σ is its standard deviation. To demonstrate the elastic margin, the authors integrated the randomized margin penalty in ArcFace (ElasticFace-Arc) and CosFace (ElasticFace-Cos).

ElasticFace-Arc (L_{EArc}) can be defined as:

$$L_{EArc} = \frac{1}{N}\sum_{i\in N} -\log \frac{e^{s\left(\cos\left(\theta_{y_i}+E(m,\sigma)\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+E(m,\sigma)\right)\right)} + \sum_{j=1,j\neq y_i}^{c} e^{s\left(\cos\left(\theta_j\right)\right)}} \tag{4.2}$$

ElasticFace-Cos (L_{EArc}) can be defined as:

$$L_{ECos} = \frac{1}{N}\sum_{i\in N} -\log \frac{e^{s\left(\cos\left(\theta_{y_i}\right)-E(m,\sigma)\right)}}{e^{s\left(\cos\left(\theta_{y_i}\right)-E(m,\sigma)\right)} + \sum_{j=1,j\neq y_i}^{c} e^{s\left(\cos\left(\theta_j\right)\right)}}, \tag{4.3}$$

where $E(m,\sigma)$ is a normal function that returns a random value from a Gaussian distribution with mean m and standard deviation σ. From the above equations, the decision boundaries for these two methods can be concluded to be: $\cos(\theta_{y_i} + E(m,\sigma)) - \cos(\theta_j) = 0$ for ElasticFace-Arc and $\cos(\theta_{y_i}) - \cos(\theta_j) - E(m,\sigma) = 0$ for ElasticFace-Cos. It is worth noting that when σ is 0 the ElasticFace-Arc is equivalent to ArcFace and the same happens for the ElasticFace-Cos and CosFace.

In Figure 4.1, there is an illustration of the decision boundaries of ArcFace, ElasticFace-Arc, CosFace and ElasticFace-Cos.

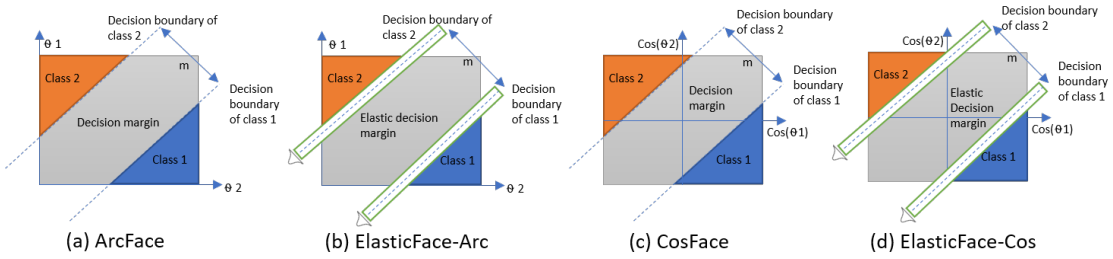


Figure 4.1: Illustration taken from the ElasticFace paper [15]. Decision boundary of (a) ArcFace, (b) ElasticFace-Arc, (c) CosFace, and (d) ElasticFace-Cos for binary classification.

Given that the goal was not the algorithm itself but rather to use it to get insights on the influence of race, the ElasticFace algorithm used had been trained already and it was only implemented. Essentially, it was only used as a tool for the performed tasks. The algorithm selection proceeded

to a careful analysis of the work already developed and published on the author's GitHub [2]. The training settings in the original work were mostly unchanged: the network architecture used (in the beginning stages of this work) was the ResNet-100 [23], the scale parameter s was set to 64, and the mini-batch size to 512. Moreover, all the models were trained with the Stochastic Gradient Descent (SGD) optimizer with a learning rate of $1e^{-1}$. The momentum was set to 0.9 and the weight decay to $5e^{-4}$. Random horizontal flipping with a probability of 0.5 for data augmentation was used during training. The images used have 112 x 112 x 3 size and produce 512- d feature embeddings. Additionally, these are aligned and cropped with the Multi-task Cascaded Convolutional Network (MTCNN) [11] that was presented in Chapter 3 and all of the images' pixels are normalized to values between -1 and 1.

The dataset used to train the original ElasticFace model was the MS1MV2 [101], which is a refined version of the MS-Celeb-1M [119] and contains 5.8M images of 85K identities. Most of the recent works in the FR area [101], [138] [139] [140] have been trained on this dataset, meaning that by following the trend, direct comparisons between ElasticFace and state-of-the-art were enabled. There are nine benchmarks of diverse nature used in the work of Boutros *et al.* [15] and they make comparisons on FR accuracy possible. The benchmarks are Labeled Faces in the Wild (LFW) [97], AgeDB-30 [141], Cross-age LFW (CALFW) [142], Cross-Pose LFW (CPLFW) [126], Celebreties in Frontal-Profile in the Wild (CFP-FP) [125], IARPA Janus Benchmark-B (IJB-B) [143], IARPA Janus Benchmark-C (IJB-C) [144], MegaFace [121] and MegaFace (R) [101]. The verification accuracy was based on the performance results of LFW, AgeDB-30, CALFW, CPLFW and CFP-FP.

The results from the mentioned benchmarks are layed-out in Table 4.1, both from some state-of-the-art works as well as the results from ElasticFace.

Table 4.1: Achieved results on the LFW, AgeDB-30, CALFW, CPLFW, and CFP-FP benchmarks. ElasticFace outer-performs 7 out of the 9 benchmarks, scoring very closely to the SOTA on LFW and CALFW. The top performances are bold and are noted with rank numbers from 1 to 3. Table adapted from [15].

Method	Training Dataset	LFW	AgeDB-30	CALFW	CPLFW	CFP-FP
		Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)
ArcFace [101] (CVPR2019)	MS1MV2 [7], [4]	99.82(3)	98.15	95.45	92.08	98.27
CosFace [110] (CVPR2018)	private	99.73	-	-	-	-
GroupFace [145] (CVPR2020)	clean MS1M [7], [4]	**99.85(1)**	98.28(3)	**96.20(1)**	93.17	98.63
CurricularFace [139] (CVPR2020)	MS1MV2 [7], [4]	99.80	98.32(2)	**96.20(1)**	93.13	98.37
MagFace [140] (CVPR2021)	MS1MV2 [7], [4]	99.83(2)	98.17	96.15	92.87	98.46
Partial-FC-ArcFace [138] (ICCVW2021)	MS1MV2 [7], [4]	99.83(2)	98.20	96.18(2)	93.00	98.45
Partial-FC-CosFace [138] (ICCVW2021)	MS1MV2 [7], [4]	99.83(2)	98.03	**96.20(1)**	93.10	98.51
ElasticFace-Arc [15]	MS1MV2 [7], [4]	99.80	**98.35(1)**	96.17(3)	93.27(2)	98.67(2)
ElasticFace-Cos [15]	MS1MV2 [7], [4]	99.82(3)	98.27	96.03	93.17	98.61(3)
ElasticFace-Arc+ [15]	MS1MV2 [7], [4]	99.82(3)	**98.35(1)**	96.17(3)	**93.28 (1)**	98.60
ElasticFace-Cos+ [15]	MS1MV2 [7], [4]	99.80	98.28(3)	96.18(2)	93.23(3)	**98.73(1)**

During the practical experiments, the specific model used from the four available (ElasticFace-Arc, ElasticFace-Cos, ElasticFace-Arc+ and ElasticFace-Cos+) was the ElasticFace-Arc and for setting up the model the parameters that needed to be changed were the loss function and the

[2] https://github.com/fdbtrs/ElasticFace

configured output, that needed to be the path for the pre-trained model weights. The file with the weights was downloaded directly from the ElasticFace's GitHub [3].

4.3 Databases

Initially, the datasets used for evaluation of the pre-trained model were the same as the ones already described as benchmarks- LFW, AgeDB-30, CALFW, CPLFW and CFP-FP. These are of an heterogeneous nature, representing some known vulnerabilities of face recognition.

The AgeDB-30 benchmark includes images with large age gaps, making its intra-user variation very large. AgeDB contains 16,488 images of celebrities and every one of those images has annotations with respect to identity, age and gender. There is a total of 568 subjects and the average of images per subject is 29. The frontal-to-profile face verification benchmark (CFP-FP) also shows a large intra-user variation. Both Cross-Age LFW (CALFW) and Cross-Pose LFW (CPLFW) are renovations from the Labeled Faces in the Wild (LFW), aiming to establish databases more complex to evaluate the performance of real world conditions on face recognition. The former considers specially age gaps and the latter pose variations. In Figure 4.2 it can be found some examples of images taken from these 4 benchmarks.

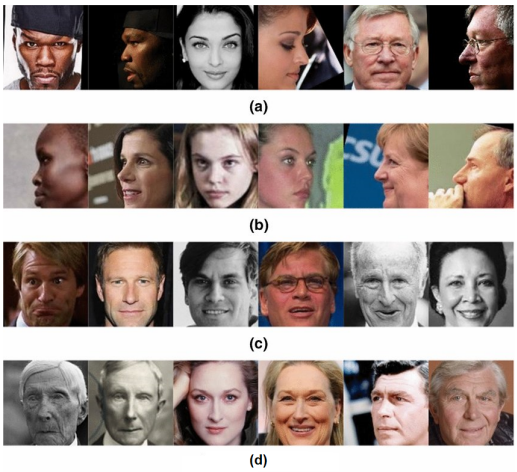


Figure 4.2: Examples of images from (a) CFP-FP (b) CPLFW (c) CALFW (d) AgeDB-30. Image from [16].

As the main focus of this work is to analyse the effects that race plays on a FR model's performance, it is important to use a racially-balanced dataset. Racial Faces in the Wild (RFW) [14] is a testing database that was developed for studying racial bias in face recognition, as has been previously mentioned in Chapter 3. There are four subsets: Caucasian, Asian, Indian, and African, each one of them with between 2400-3000 individuals and 9688 to 10415 images, which are combined into 6000 image pairs of genuine and impostor samples for face verification (Table 4.2 shows the number of subjects and images per race).

[3] https://github.com/fdbtrs/ElasticFace

Table 4.2: Number of identities and images in RFW. Table from the original RFW website (http://www.whdeng.cn/RFW/testing.html).

RFW	# Subjects	# Images
Caucasian	2959	10196
Indian	2984	10308
Asian	2492	9688
African	2995	10415

In the RFW paper [14], the authors present the average face for each race from their database of images. We replicated this task and in Figure 4.3, the image on the left, composed of two columns, belongs to the RFW original paper [14], and represents the mean faces from the four races, in rows from top to bottom: Caucasian, Indian, Asian and African. The right portion of the same Figure, shows the results we obtained, with the mean faces from each race occupying the equivalent row from the image on the left. These average faces were obtained by summing the pixels of all the 12000 images from each race subset and calculating the average value. It is important to mention that this dataset is not balanced with respect to gender, so there are more male samples across all races [4], which justifies the image results for the average face of each race (images on the right of Figure 4.3) being primarily masculine. This database was essential throughout the practical work and it will be mentioned in the following sections when presenting the various tests made.

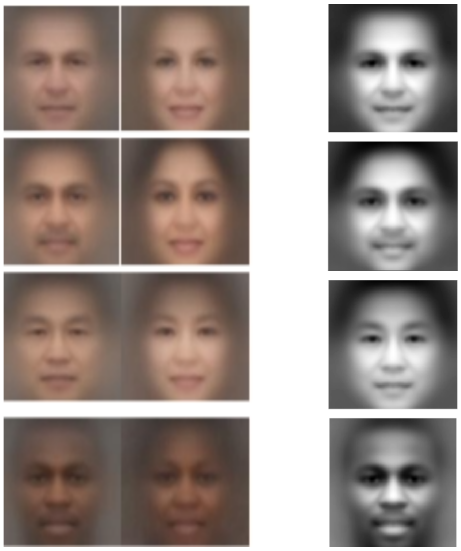


Figure 4.3: On the left, average faces for each race taken from an adapted image from [14]. On the right, average faces resulting from the evaluation process performed on ElasticFace with the RFW database. From the top row to the bottom row: Caucasian, Indian, Asian and African.

Alongside with this database, Wang *et al.* [29] also presented four training databases to try and bring social awareness to the presence of racial bias on training data, allowing the study of facial bias and fair performance. From the four, two of them were especially relevant for the work

[4] http://www.whdeng.cn/RFW/index.html

presented further on this dissertation: BUPT-Globalface and BUPT-Balancedface. Globalface consists of 2M images from 38K celebrities and its racial distribution is 38% Caucasian, 18% Indian, 31% Asian and 13% African. This distribution is approximately the same as that of the world's population at the time of its development. Balancedface is formed with an equal number of identities from each one of the four races (each represents 25% of the whole database), meaning that it is balanced.

4.4 Image transformation

The practical experiments started on the basis of simple transformations applied to the input images to see the effects that these modifications had on the model's performance. For this, a Python library was used- Albumentations [5]. This library is compatible with PyTorch and of simple implementation and application. It acts at the pixel level and includes a vast amount of possible transformations.

Initially, the datasets tested were the benchmarks used for verification on ElasticFace-Arc - LFW, AgeDB-30, CALFW, CPLFW and CFP-FP. The transformations were a 180º rotation, obtaining the grayscaled and the negative images, and changing the brightness and contrast of the images. These last two modifications are colour augmentation functions and work by altering the pixel values from an image. Changing the brightness results in a darker or lighter image compared to the original: when the value is set to zero it refers to a completely black image and one to a completely white image. On the other hand, contrast is the degree of separation between the darkest and brightest areas of an image.

We applied changes both in brightness and contrast at the same time. The parameters were set for various values as a way of testing different ranges; moreover, all the transformations were applied to all of the images.

After these initial experiments, the same process with the same transformations was performed on RFW to try and test if a specific ethnicity could be more affected by the applied modifications. These alterations may be especially relevant when dealing with over or under-exposed images and when testing the degree of influence that a change in brightness or contrast may have in a specific race.

4.5 Grad-CAM Visualizations

As it was previously stated in Chapter 3, explainability is a field of AI that gained a lot of attention and researchers keep trying to implement methods that focus on increasing the level of transparency and interpretability of DL models. Visual explanation techniques are attribution-based methods that help to understand a model's prediction by assigning attributions to each input feature. The "saliency maps" are gradient-based and indicate the contribution of each input variable,

[5] https://github.com/albumentations-team/albumentations/

and in this case these variables are pixels from the input images. The goal of this set of experiments was to visually try to analyse the way that the model processes images from different racial groups and observe the regions of interest highlighted in the maps. Throughout the network the maps also change, given that firstly the model focuses on a more broader context and it increases its attention to details as the pipeline progresses.

The visualization technique used was the Gradient-weighted Cam (Grad-CAM [3]), where the final map results from the combination of feature maps that use the gradient. The premise was to select three layers from distinct points on the network's pipeline and generate the corresponding mean gradient map. The three more logical points to attain the Grad-CAM output images were the first convolutional layer, an intermediate layer and the second batch normalization layer, corresponding with the beginning, middle and end of the network, respectively. We obtained these maps for a subset of each ethnicity group on the RFW dataset. In the end of this process, there were three maps for each race, allowing a direct comparison between them. The library used for generating the maps was MONAI [6] which is known from being implemented in the Medical Imaging field.

The generated maps help to pinpoint the regions that were relevant to the network's decision and help to understand where the model focuses its attention, e.g the object or features of interest. Furthermore, the Grad-CAM should also aid to visualize neural activation, highlighting the parts of the input image that contributed the most for the produced embedding. Again, to get insights on the influence of race on a FR algorithm, the comparison of the maps from the same layers for each race could help bring light to the main distinctions between them in the model's perspective, and possibly show different visual patterns.

As previously mentioned, the algorithm selected, ElasticFace-Arc, uses a ResNet-100 backbone. This network was trained on the MS1MV2 database. Aforesaid, there are some issues associated with the use of this dataset combined with RFW, e.g. data leakage, as there are some images that are simultaneously on both databases. Therefore, the idea to only change the type of network would allow to test for the impact of this alteration alone, as the new network implemented would also be trained on the MS1MV2. Thus, a ResNet-34 already trained on MS1MV2 was implemented. The gradient maps were generated following the same conditions as before, with the only difference being the network architecture. Furthermore, this experiment of changing the network architecture was extended to another two databases already mentioned: BUPT-Globalface and BUPT-Balancedface. The ResNet-34 had already been trained on both of them as well [136] and again the goal was to only implement and analyse the results. These two databases were particularly important and used for the remaining tests given that the Globalface reflects pre-existent bias mirroring the distribution of the four ethnic groups in the world, and the Balancedface is equalized.

[6] https://monai.io/

4.6 Neural Activation in FR systems

In the very beginning of this dissertation, we introduced the notion that researchers have long shown an interest in attempting to come up with a parallelism between the human face recognition process and an automatic FR system, especially when it comes to taking inspiration from the human biological approach and applying it on artificial models.

The human face recognition task develops over time, even if this period of time is incredibly small, and different regions of the brain are activated throughout, with the neurons of these same regions following the activation process. Additionally, there is the hypothesis presented in Chapter 2 in the "Other-race effect" section, that the human brain shows higher neural activation when seeing faces from other races, meaning that observing a face from a different race than ours or different from the race we grew up surrounded by, may induce more neurons to be activated.

This hypothesis was the starting point for what may be considered the focal aspect of the experimental work on this dissertation: the interest in pursuing an investigation that possibly shows a parallelism between human neural activation and deep neural networks, when both perform the task of face recognition with focus on the influence of race and racial bias. Therefore, and given that a deep neural network can not be evaluated through time, as it is very difficult to analyse which or how many neurons are activated in a very specific point in time, the approach to try and perform a comparison was to focus on the structural development of models: the progressive activation of consecutive layers. Thus, even though networks do not work in a chronological order, their architecture has a premeditated sequence and the model follows a pipeline.

4.6.1 Layer neural activation

Starting for acquiring the activation values for each layer, from the first convolutional layer to one of the last linear ones (in total, there were six layers used for the experiments), the focus was to obtain some informative statistics: the mean and the standard deviation of the neural activation in each layer. These statistical values were calculated for all four subsets on RFW. The use of the mean in this case aims to allow a inter-racial analysis, given that the mean values can give insights into the separability of the ethnicity groups. On the other hand, the standard deviation values extracted may help to study intra-racial bias: higher values might indicate high variance or discriminative power inside classes. Globally, the aim of this task was to see if there was some kind of pattern in the results and if there was anything notoriously different for any of the races.

Additionally, using exclusively the final network layer with 12000 images and 512 features, the mean value for these features was calculated, leaving the output with the dimension of 12000 x 1. Afterwards, with this output, it was calculated the standard deviation. Again, the process was repeated for all four races. For this task, the motivation was to compare this approach with the simple extraction of the STD values. This experiment did not lead to any conclusions, so it was not included in the Results Chapter 5.

Using PCA to perform a dimension reduction in the last layer, the same metrics described in the beginning of this subsection were also retracted. Before performing the PCA, the input

variables had to be normalized, so that they were scale independent and stable, and each point is transformed by subtracting the mean of the data and dividing it by the data's standard deviation. With this dimension reduction of the neural activation output of the linear layer, the goal was to use only one dimension for these new metric calculations, while providing a more insightful dimension reduction strategy when compared with the average. Therefore, instead of 512 different values, the PCA uses a representative value of all these features, while keeping the maximum amount of information possible. Again, as the experiment did not lead to any relevant conclusions, the results were not included as well.

As in the prior section, all the steps were repeated and the activation values were also obtained with the ResNet-34 trained on the MS1MV2, BUPT-Globalface and Balancedface.

4.6.2 Race classification

RFW was proposed as a race-aware database and it was expected that the model would be able to separate and detect differences between races, having a good inter-racial performance. Like humans, that show an ability to easily distinguish subjects from their own race, but have a harder time setting apart subjects from another race, automatic FR models can also show liabilities in this task. This dissertation had a special focus on intra-racial bias analysis. For a model to be able to distinguish between subjects from the same race, the distance between their embeddings needs to be a compromise: low enough that they do not increase intra-ethnicity errors, but still distant so that the embeddings are not misjudged as being equal, e.g. same person.

To test the ability of a ML classifier to correctly assign the race of a subject just by analysing an image's neural activation matrix in a specific layer of the network, these activation matrices were extracted from the beginning, middle and end of the network of the ElasticFace tested on RFW. The two ML classifiers used were the K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). For all the layers, for the KNN, different values of K were tested out: 5, 7 and 9. Similarly, in all the layers, the parameters for SVM were the same: the kernel value, the c and gamma were set as the default values of the function (c=1, kernel= 'rbf' and gamma='scale').

The values for the neural activations of the Caucasian subset in one of the final layers of the network (from the linear portion) were extracted. Given that it was a layer from the end of the network, after all the applied convolutions, the output activations corresponded to the final 512 features, putting the output dimensions at 12000 images per 512 features (12000, 512). The activation values were concatenated into a vector. Afterwards, the same process was applied to the African subset of the database and its activation values were also extracted and saved in another vector. The same goes for the remaining two races: Asian and Indian. Simultaneously, as each vector with the results from the images of one race was saved, there was another vector, of the same length, that was created. As it was necessary to apply a true label to each activation matrix (each row in the first described vectors), this new vector would hold the indexes of the future classes. Four different ethnicity groups meant that there should be four indexes/classes:

- Class 0: Caucasian

- Class 1: African
- Class 2: Asian
- Class 3: Indian

In the end, there were 4 pairs of 2 vectors: for each race, one of them containing the results of the neural activations of the desired layer and the other vector would be of the same length, but would only have the correspondent index to that specific race. For example, the first row of the vector containing the activation results for Caucasians would correspond to a 0 in the same row of the other vector.

Afterwards, as the goal was to join all the activation results into one big vector before performing the classification, the four sub-vectors with the activation matrices were concatenated as well as the other four holding the respective indexes. With 12000 input images per race, concatenating the four sub-vectors meant that the final one presented 48000 images per 512 features. After making sure that these two final vectors were properly shuffled, and keeping in mind that the rows should be swapped in pairs as a way of assuring that a certain item in the first vector would maintain its true label even after shuffling, these were splitted into train and test data. It is common that the training subset represents 80% of the entire data, leaving the remaining 20% for testing. Therefore, 38400 images were used as the training data. Following a normalization of the training data, the classification with the two chosen methods- KNN and SVM-, was performed. Besides the results for the test and training accuracy, confusion matrices and classification reports were obtained.

Going up the network, the activation matrices outputted for each image enlarge in dimension. The same process for obtaining the activation matrices and the two final vectors for classification was implemented, however as it was necessary that the input data for these ML classifiers had only two dimensions maximum, it was required to apply a type of dimension reduction. The middle layer used had, after concatenating and shuffling the vectors with the four races, the following dimensions: (48000, 256, 14, 14). Therefore, firstly a simple mean performed in each one of the 256 channels was carried out, resulting in a vector of 48000 per 256. Besides the calculation of the mean, there was another approach tested, to try and possibly improve the classification results and allow a comparison: a reshape of the matrix that would result in the product of the three last dimensions (48000, 256 x 14 x 14). After these steps, the classification proceeded in the same way as previously explained: a split between training and testing data, followed by a normalization of the training data and the classification itself.

At last, race classification was carried out in the first layer. Again, it was expected that in the first convolutional layer, before performing all the convolutions and max-poolings, the dimensions of the matrices would be greater, as there is a lot more information/ features being extracted. Given that each subset has 12000 images, which is a significant amount of data to be processed, when trying to perform the extraction of the activation matrices for the first layer there was a clear increase in the computational memory required to perform the same task. These hardware

setbacks meant that as the embeddings were too large with (12000, 64, 112, 112), it was necessary to calculate the mean of the two last dimensions right after acquiring each activation matrix. This was performed in each one of the 12000 iterations, and before appending the results from each iteration to a vector. This approach allowed the stacking of the four race sub-vectors into a single one, as described for the other layers. As a result of the mean applied right after the extraction of the activation matrix, the final vector with all the data was two-dimensional already. Without the need for another reshape or extra processing, the KNN and SVM were applied.

4.7 Summary

This chapter presented the various steps of the work and the reasoning behind them. In the Developed Strategy section, there is a workflow layout that introduces each method. Besides image transformations and the generation of activation maps with Grad-CAM, the rest of the work was based on neural network activation and the information that these values can convey in relation to race. Focusing on intra-racial bias analysis, the classification task aims to look for a progressive separability between racial groups across the network pipeline. The direction of the practical experiments was adapted throughout time and some tasks that were performed to reach a specific goal or conclusion did not present the desired information.

Chapter 5

Results and Discussion

This chapter aims to present the most relevant results regarding the various stages of the executed work, namely the image transformations, the gradient maps, the network layer activation and the classification process with the activation matrices. Additionally, these results are discussed and associated limitations are described.

5.1 Image Transformation

We have computed the accuracy values for all the benchmark databases mentioned in the ElasticFace paper. Moreover, histograms with the positive and negative distances were generated and can be found in Figure 5.1. The positive and negative, in green and red, respectively, resulted from the verification process performed by the model. As introduced in Chapter 2, a verification is a 1:1 comparison, where two faces are directly compared. In ElasticFace-Arc, the model considers a positive verification if there is a match: the euclidean distance is computed for two embeddings (the vectors containing facial features from two images) and this value must be lower than an established threshold. On the contrary, a negative example corresponds to two embeddings that are further away than an established distance, meaning that the features vectored are not similar, e.g. the two images do not belong to the same person.

Looking at the histograms of the five datasets (Figure 5.1), there is a convergence zone in all of them besides LFW, where the green and red portions are very well separated. The overlay in the remaining datasets is represented by a darker red and shows the instances where the model had difficulties in clearly telling a positive from a negative verification.

After analysing these graphs of the distribution of the datasets, some alterations were performed similarly in all of them: an inversion of the image (a 180º rotation), grayscaling and applying the negative, and testing changes in brightness and contrast on the original images. The accuracy results for all the modifications can be seen in Table 5.1, where the first column shows the results for the accuracy of the original/unchanged images, allowing for a direct comparison with the results from the transformations. Moreover, as mentioned during the last chapter (Chapter 4), the brightness and contrast can be changed simultaneously, with the variables being independent.

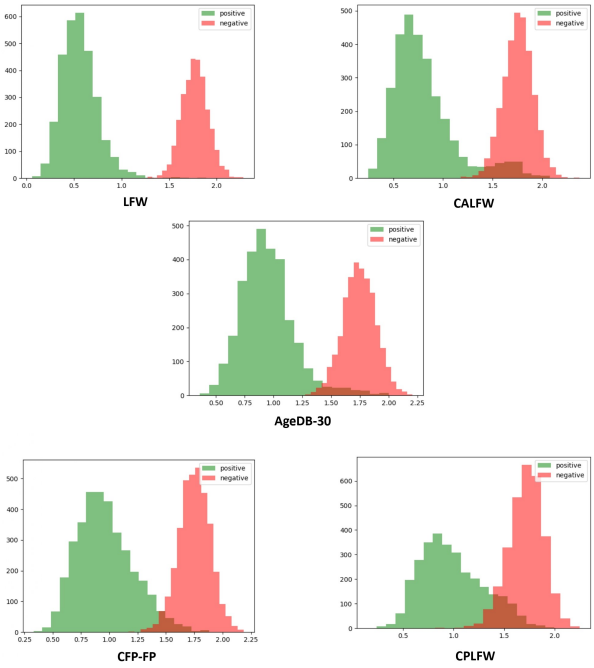


Figure 5.1: Histograms representing the positive and negative verifications performed on LFW, CALFW, AgeDB-30, CFP-FP and CPLFW.

After applying a small (0.2) and later on a big (0.8) change to both variables, there was also an interest to see the consequences of applying only one of the operations, meaning that only the value of either the brightness or contrast would be different from 0. Two columns on Table 5.1 under the brightness and contrast portion have the results from the experiments first described (columns 1 and 3), where the same value is applied for both variables. The other two remaining columns under this same section of the table present the results for only one of the variables being tested: in the second column, brightness was set to 0 and the value of contrast was increased and in the last column this same logic was implemented in the other way around. It is important to highlight the fact that as it was concluded that applying a change in the 0.2 order did not altered the final performance in a considerable amount, for these last experiments, where only one of the variables was changed, it was only used the value 0.8.

Table 5.1: Accuracy results of the verification for the image transformations performed on LFW, CALFW, AgeDB-30, CFP-FP and CPLFW.

					Brightness and Contrast			
	Normal	Rotation	GrayScaled	Negative	0.2	brightness= 0 , contrast=0.8	0.8	brightness= 0.8 , contrast=0
LFW	0.998	0.725	0.998	0.813	0.998	0.998	0.845	0.919
CALFW	0.960	0.620	0.957	0.703	0.961	0.960	0.821	0.893
AgeDB-30	0.984	**0.507**	0.975	0.658	0.982	0.976	0.820	0.888
CFP-FP	0.986	0.534	0.971	0.666	0.985	0.978	0.823	0.884
CPLFW	**0.932**	0.541	**0.915**	**0.648**	**0.929**	**0.925**	**0.773**	**0.835**

After applying the transformations the same histograms were generated, to visualize the impact on the verification process. As an example of these histograms, Figure 5.2 presents two plots from the CPLFW database where it is notorious the overlap between positive and negative.

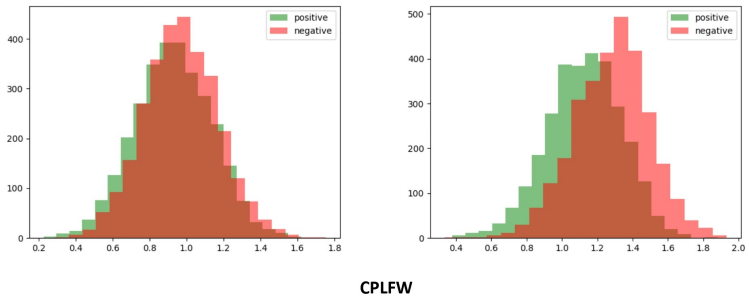


Figure 5.2: Histograms from CPLFW: on the left, after applying rotation and on the right, after applying the negative.

Analysing these accuracy values and knowing the individual details about each one of the databases (reported in the previous Chapter 4 in the databases section 4.3), a debate on possible motives behind the results may become adequate. After performing all transformations, the best accuracy all around belongs to LFW. A small increase in the same degree in both brightness and contrast, corresponding to the 0.2 column, shows that the impact on the model's performance is minimal. On the other hand, when increasing the brightness and contrast for a higher value (0.8), all databases suffer a considerable drop in performance. It can be concluded that a change in brightness impacts more the model's performance than a change in contrast, even when talking about an alteration within the same range. The last four databases are especially impacted by an increase in the image brightness and contrast, which is relevant to notice when considering their constitution, that is more complex and heterogeneous than LFW. CPLFW was the database that presented the worst results in accuracy, excluding only the rotation experiment.

As presented in Section 4.4, the performed alterations were experimental and their main goal was to see if one of the datasets was especially affected when comparing to the others. Again, as the focus of this dissertation is investigating racial bias, the same exact transformations were performed on a race-conscious database (RFW). The goal was to obtain the results from these modifications applied on RFW and analysing them, looking for the possibility that one of the races may be more affected by the alterations than the others and, therefore, indicate the presence of racial bias.

The histograms of the four ethnicities (Figure 5.3) show a visual representation of the verification process performed and the positive and negative examples.

Afterwards, the image modifications were applied with the same values for the brightness and contrast tests. The accuracy results can be found in Table 5.2. Again, the independency between

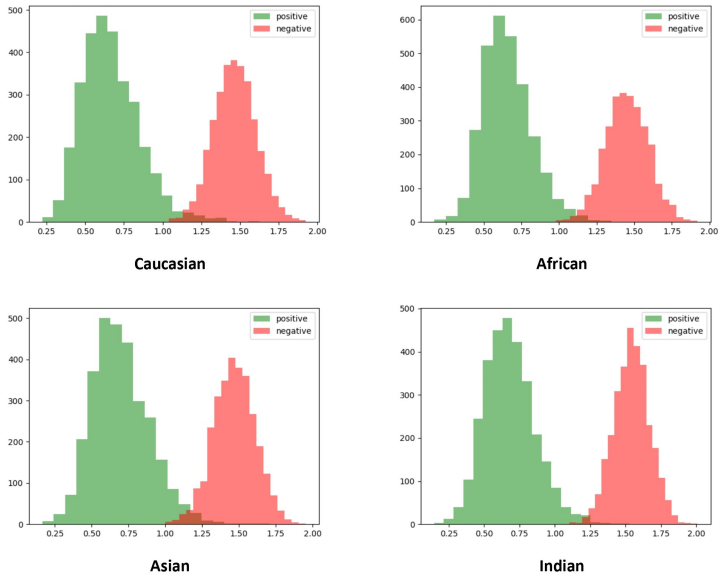


Figure 5.3: Histograms representing the positive and negative verifications performed on RFW: Caucasian, African, Asian and Indian.

contrast and brightness was tested, aiming to verify if altering the value of only one had a bigger impact on a particular race.

Table 5.2: Accuracy results for the image transformations performed in RFW: African, Asian, Caucasian and Indian.

					Brightness and Contrast			
	Normal	Rotation	Grayscaled	Negative	0.2	brightness= 0 , contrast=0.8	0.8	brightness= 0.8 , contrast=0
African	0.993	**0.534**	0.983	**0.612**	0.989	0.990	0.781	**0.797**
Asian	**0.988**	0.579	**0.967**	0.623	**0.985**	**0.973**	**0.776**	0.841
Caucasian	0.995	0.593	0.985	0.669	0.994	0.989	0.799	0.860
Indian	0.990	0.589	0.978	0.616	0.989	0.980	0.790	0.848

As expected, the Caucasian subset was the least affected by any of the transformations. Performing a rotation of the face, as with the databases tested previously, had an extensive effect on the accuracy all around. Moreover, the results showed that applying a small change in brightness and contrast simultaneously (0.2) did not have a relevant impact on accuracy. The same happened when altering only the contrast value using 0.8. For the implementation of the same modification in both variables (0.8), the accuracy values were, as expected, more affected. Looking at the last column, where there was an increase of illumination without changing the degree of separation between the dark and light pixels, the accuracy result of the African subset was the lowest out of all the races. This was specially relevant given that in the other colour augmentation tests, this subset was the second with the highest accuracy, following Caucasian. This result points to the fact that this race may be more affected by an increase in brightness than the others, being more

susceptible to light changes.

5.2 Grad-CAM Visualizations

The mean gradient maps were generated as a way of visualizing the model's regions of interest in the input images. As described, three layers were selected for the generation process and these should represent different moments on the network, analysing the evolution of the features of interest. Therefore, the chosen layers were the first convolutional layer, one of the middle layers and the second batch normalization layer. In theory, the heat-map of the first convolutional layer should display the most accurate visual explanation of the classified object, e.g. the mean face for each race. The middle layers tend to capture textures at a higher level of abstraction, and the final layers tend to focus on semantic information. The produced heat-maps were overlayed over the original images so there was a better visual understanding of the results. The colour map used can be interpreted in the following way: the more vivid/ darker the colours on the map, the larger is the corresponding absolute value of activation.

The gradient maps are constituted from top to bottom and from left to right in the following way:

1. The first row represents the mean gradient maps from the first convolutional layer where (a) Caucasian (b) African (c) Asian (d) Indian

2. The second row presents the mean gradient maps outputted from an intermediate layer with (e) Caucasian (f) African (g) Asian (h) Indian

3. The last row shows the mean Grad-CAM visualizations for the second batch normalization layer, after all the convolutions with (i) Caucasian (j) African (k) Asian (l) Indian.

The first maps were generated with the ResNet-100 from ElasticFace (trained on MS1MV2) and evaluated on RFW. These can be found in Figure 5.4.

Regarding the first row of Figure 5.4, it portrays the idea presented above, displaying the most accurate visual representation of the objects of interest. Given that these maps were generated by calculating the mean of the attributions for each layer, the faces in the first row can be interpreted as a portrayal of the mean faces for each race, with mean features. Moreover, in the first convolutional layer the amount of information was bigger, with more features, so it was expected that the output gradient map would be the most similar to the input images. Following the network architecture, the amount of features reduces as we descend, until it reaches the output layer with 512 features.

As it was previously introduced on Chapter 4, the network associated with the ElasticFace model (ResNet-100) was switched for a ResNet-34 also trained on MS1MV2. Again, this portion of the practical work was inspired by the desire to assess what would be the impact on the results by only changing the network architecture. The resulting Grad-CAM output images can be seen in Figure 5.5.

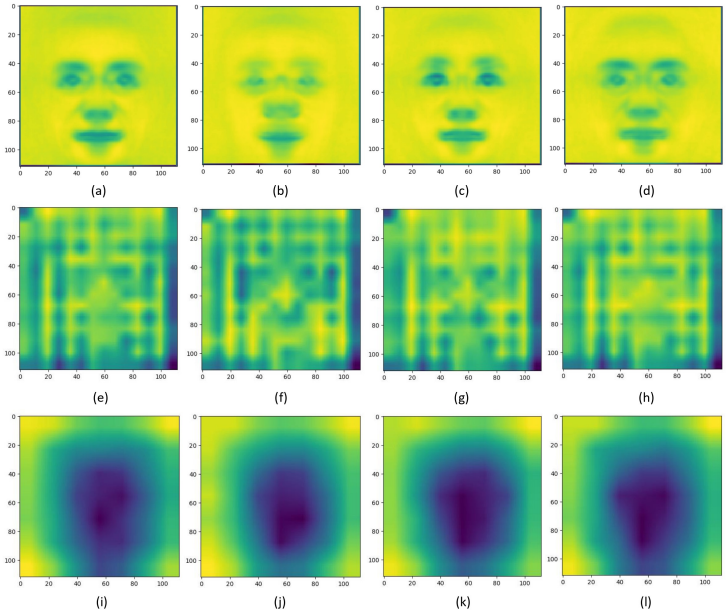


Figure 5.4: Grad-CAMs generated with a ResNet-100 network trained on MS1MV2.

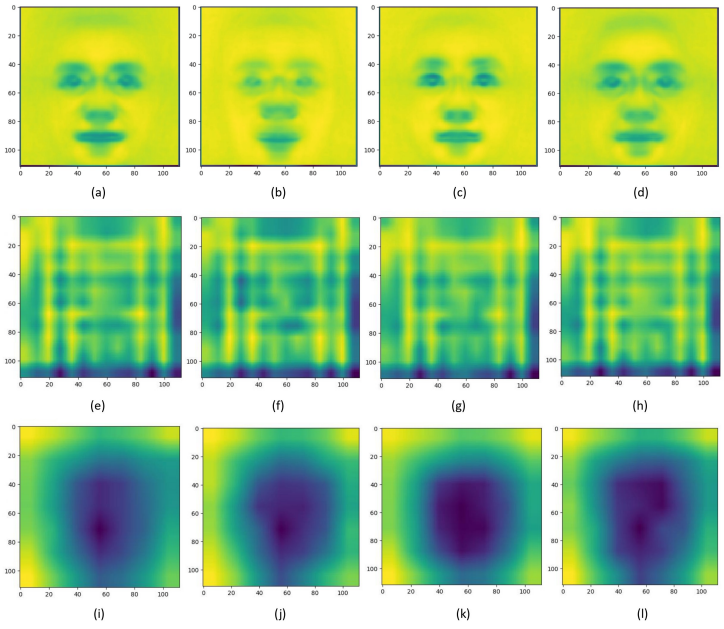


Figure 5.5: Grad-CAMs generated with a ResNet-34 network trained on MS1MV2.

Observing the Grad-CAM images from both networks there is not a notorious difference, meaning that in general the results seem to be pretty similar. Nonetheless, in the final layer, it is possible to observe that using the ResNet-34 the blue regions are a little less intense, pointing to lower values of activation when compared to the results from Figure 5.4 per example.

Extending the experiments, besides testing exclusively switching the backbone, there was also an interest in using ResNet-34 trained on two race-aware databases: BUPT-Globalface and BUPT-Balancedface. Figures 5.6 and 5.7 show the results for each one of them respectively.

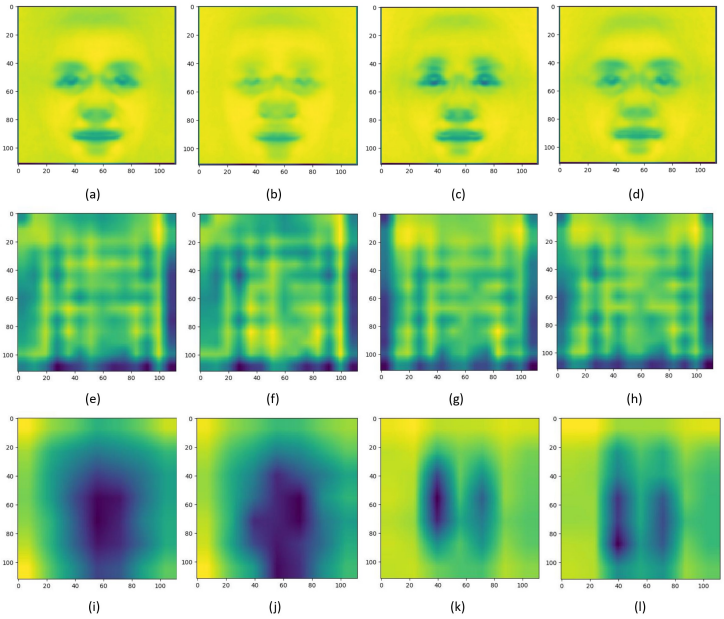


Figure 5.6: Grad-CAMs generated with a ResNet-34 network trained on BUPT-Globalface.

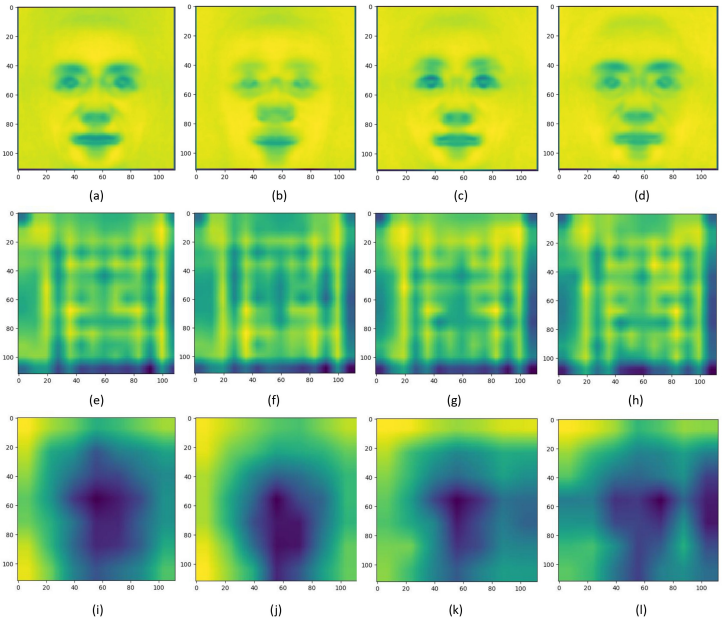


Figure 5.7: Grad-CAMs generated with a ResNet-34 network trained on BUPT-Balancedface.

Analysing the images from Figure 5.6, the last row presents the biggest differences. The last

layer map for both Asian and Indian (third and fourth columns) present very light demarcated regions and less disperse when compared to the previous maps.

Lastly, the results from the ResNet-34 Balanced (Figure 5.7) show that the last layer has less vivid demarcated regions when compared to Figure 5.4, per example.

5.3 Neural Activation in a FR system

5.3.1 Layer activation

Regarding the neural activation in different layers of the network, two important metrics were calculated: mean and standard deviation. Once more, the goal was to see an evolution throughout the layers, while being able to analyse if race played a part on these values. The six layers used start with the Parameterized ReLU, before the first convolutional layer, and end in the linear portion of the architecture. These follow the network's pipeline from start to end (left to right on Table 5.3). The results from this table correspond to the activation mean of the 12000 images from each subset.

Table 5.3: Mean values for neural activation from ResNet-100 trained on MS1MV2.

	Pre-Relu	**Layer1**	**Layer2**	**Layer3**	**Layer4**	**Linear**
Caucasian	0.00978	0.01899	0.03450	0.02325	0.00797	0.01348
African	0.00998	0.01897	0.03454	0.02395	0.00836	0.01343
Asian	0.00996	0.01928	0.03402	0.02317	0.00863	0.01334
Indian	0.00988	0.01966	0.03466	0.02334	0.00846	0.01352

In Table 5.4, the values for the standard deviation of each layer in each race are represented.

Table 5.4: Standard deviation (std) values from ResNet-100 activation trained on MS1MV2.

	Pre-Relu	**Layer1**	**Layer2**	**Layer3**	**Layer4**	**Linear**
Caucasian	0.00889	0.01062	0.01611	0.01043	0.00430	0.00511
African	0.00942	0.01074	0.01617	0.01089	0.00465	0.00511
Asian	0.00919	0.0109	0.01589	0.01054	0.00489	0.00508
Indian	0.00908	0.01097	0.01612	0.01051	0.00466	0.00514

Once more, the network was switched for a ResNet-34 trained on the same database than the one used in ResNet-100, in order to evaluate the impact of the network itself. Tables 5.5 and 5.6 show the mean and the standard deviation values, respectively.

Table 5.5: Mean values for neural activation from ResNet-34 trained on MS1MV2.

	Pre-Relu	**Layer1**	**Layer2**	**Layer3**	**Layer4**	**Linear**
Caucasian	0.00297	0.00563	0.00514	0.00414	0.00157	0.00200
African	0.00299	0.00562	0.00507	0.00416	0.00173	0.00207
Asian	0.00303	0.00569	0.00512	0.00410	0.00169	0.00202
Indian	0.00303	0.00581	0.00519	0.00415	0.00169	0.00208

Table 5.6: Standard deviation (std) values from ResNet-34 activation trained on MS1MV2.

	Pre-Relu	**Layer1**	**Layer2**	**Layer3**	**Layer4**	**Linear**
Caucasian	0.00519	0.00619	0.00489	0.00374	0.00165	0.00152
African	0.00520	0.00624	0.00487	0.00377	0.00189	0.00157
Asian	0.00535	0.00631	0.0049	0.00371	0.00187	0.00153
Indian	0.00533	0.00639	0.00494	0.00376	0.00184	0.00158

Later on, with the switch to ResNet-34 trained on either Globalface or Balancedface, there could be a possible analysis of the impact of race distribution. The following two tables (Table 5.7 and Table 5.8) show the results for the mean and std for Globalface and the following two for the same statistics but with Balancedface (Table 5.9 and Table 5.10).

Table 5.7: Mean values for neural activation from ResNet-34 trained on Globalface.

	Pre-Relu	**Layer1**	**Layer2**	**Layer3**	**Layer4**	**Linear**
Caucasian	0.00671	0.01142	0.0114	0.0095	0.00452	0.00675
African	0.00647	0.01135	0.01122	0.00944	0.00496	0.00733
Asian	0.00677	0.01151	0.01138	0.00943	0.00453	0.00713
Indian	0.00674	0.01177	0.01156	0.00955	0.00461	0.00718

Table 5.8: Standard deviation (std) values from ResNet-34 activation trained on Globalface.

	Pre-Relu	**Layer1**	**Layer2**	**Layer3**	**Layer4**	**Linear**
Caucasian	0.00918	0.01120	0.01000	0.00795	0.00439	0.00513
African	0.00954	0.01122	0.00986	0.00795	0.00494	0.00557
Asian	0.00945	0.01140	0.01003	0.00790	0.00453	0.00542
Indian	0.00933	0.01157	0.01016	0.00802	0.00452	0.00545

Table 5.9: Mean values for neural activation from ResNet-34 trained on Balancedface.

	Pre-Relu	**Layer1**	**Layer2**	**Layer3**	**Layer4**	**Linear**
Caucasian	0.01039	0.01699	0.01658	0.01409	0.00697	0.01391
African	0.00985	0.01681	0.01639	0.01399	0.00732	0.0147
Asian	0.01049	0.01711	0.01659	0.01392	0.00695	0.01395
Indian	0.01047	0.01747	0.01685	0.01414	0.00697	0.01442

Table 5.10: Standard deviation (std) values from ResNet-34 activation trained on Balancedface.

	Pre-Relu	**Layer1**	**Layer2**	**Layer3**	**Layer4**	**Linear**
Caucasian	0.01175	0.01564	0.01435	0.01154	0.00647	0.01055
African	0.01123	0.01558	0.01421	0.01151	0.00698	0.01115
Asian	0.01204	0.01587	0.01443	0.01142	0.00657	0.01061
Indian	0.01196	0.01608	0.01461	0.01164	0.00644	0.01093

To draw meaningful conclusions on these results, the mean and std values were examined across different layers and racial groups, and even different networks trained on different databases (with variable race distributions). Looking at the tables presented, there is not a clear pattern that may lead to a conclusion on the implication of race on layer activation. Moreover, there is not a discrepancy or outlier that helps to reach an interpretation on the behavior of the overall network behavior and potentially identify the source of bias. As it was previously pointed out, the use of the mean was directed towards the analysis of the racial separation (inter-racial). On the other hand, the calculation of the standard deviation in this case aimed to study the separation in the same ethnicity, e.g. an intra-racial analysis. The focus of this work was on this intra-racial analysis, however as there is not a prominent pattern in the std tables that may indicate the presence of bias, there is not a direct conclusion on this topic.

In Table 5.11, the values for the model's performance using each one of the presented networks evaluated on RFW are layed out. Moreover, the last two rows show the values for the standard deviation (std) and the skewed error ratio (SER) between the four races in each network. The skewed error ratio is calculated using the following expression:

$$SER = \frac{100 - \min(acc)}{100 - \max(acc)} \tag{5.1}$$

The standard deviation aims to evaluate the variance between the accuracy values. The lower std values for the two networks trained on MS1MV2, suggest that the system's performance is more consistent and it is not significantly affected by the racial group. On the other hand, the higher std value for the ResNet-34 trained on Globalface may indicate the opposite. As for the SER values, this variable measures the relative difference in error rates between different groups. In an ideal system, fair and unbiased, it would be expected for the SER results to be closer to 1 for all racial subsets.

Table 5.11: Accuracy results in percentage (%) for each one of the networks used with the respective training dataset, standard deviation (std) and skewed error ratio (ser).

	ResNet-100 MS1MV2	**ResNet-34 GLOBAL**	**ResNet-34 BALANCED**	**ResNet-34 MS1MV2**
Caucasian	99.52	97.67	96.60	99.12
African	99.33	93.87	93.37	98.05
Asian	98.80	94.15	94.03	97.27
Indian	99.03	95.52	94.50	98.10
STD	0.32	1.73	1.40	0.76
SER	2.48	2.63	1.95	3.10

5.3.2 Race classification

Regarding the classification process described in the previous chapter, the 2D separation between classes was obtained by applying TSNE to each one of the layers used. This method was employed for visual purposes only, so that there was a representation of the four classes and their interaction and dispersion in a 2D space.

The four classes correspond to the four subsets in RFW:

Class 0: Caucasian **Class 1:** African **Class 2:** Asian **Class 3:** Indian

There are 3 Figures: Figure 5.8, Figure 5.9 and Figure 5.10 that correspond to the class distribution on the first, middle and last layer, respectively.

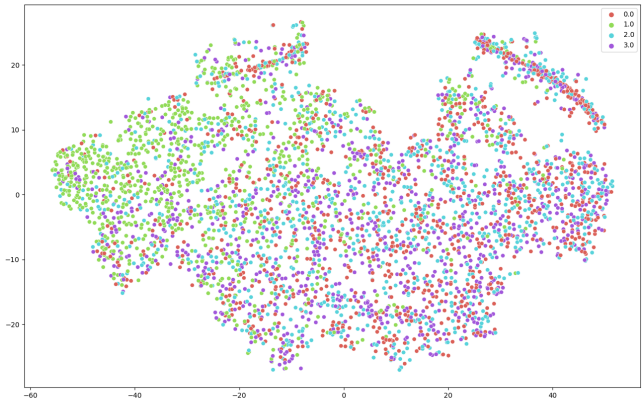


Figure 5.8: Visual representation of class distribution on the first layer of the network. Red- Caucasian ; Green- African ; Blue- Asian; Purple- Indian

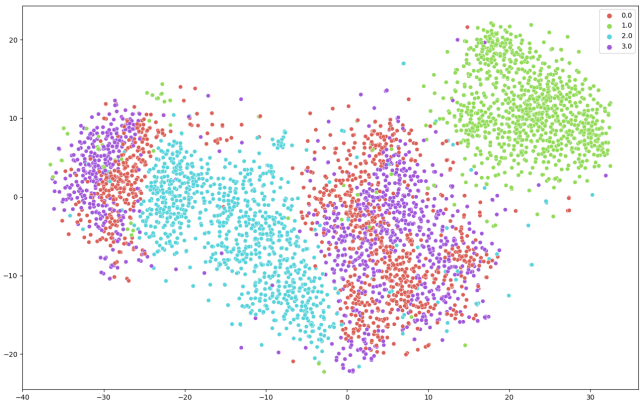


Figure 5.9: Visual representation of class distribution on the middle layer of the network. Red: Caucasian ; Green: African ; Blue: Asian; Purple: Indian

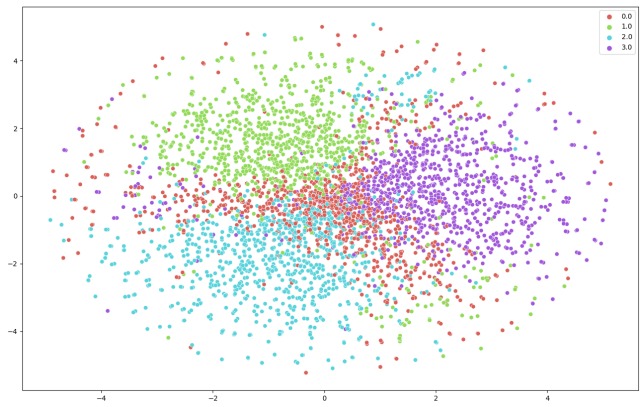


Figure 5.10: Visual representation of class distribution on the final layer of the network. Red: Caucasian ; Green: African ; Blue: Asian; Purple: Indian

These images show that, in the beginning, the class distribution was very random and disperse (Figure 5.8); in Figure 5.9 the classes start to form clusters and become more organized. Lastly, the last plot (Figure 5.10) shows that the majority of the instances in the African, Asian and Indian classes formed a cluster pretty well separated from the others. On the other hand, class 0, represented in red, is more disperse and in the center of the other classes.

These 2D plots of the race classification process help to visually understand that in the last layer the classes follow a pattern and are aggregated in a way that allows for separation between them. On the contrary, the initial layer presents a mixture of instances from the 4 classes, clearly more difficult to classify and separate.

5.3.2.1 Last layer

As it was already pointed out in the last chapter, in the last layer of the network (in the linear portion) the amount of information is the smallest out of all the layers. As the network tapers as we go through the layers, the amount of features diminishes. Therefore, in the last layer, the activation matrices extracted had only 512 features.

Following the methodology from chapter 4, both KNN and SVM were performed.The results for both the training and the test accuracy are shown in Table 5.12.

Table 5.12: Accuracy results for KNN and SVM in the last layer.

	KNN			**SVM**
	K			-
	5	7	9	**-**
Training accuracy	0.988	0.978	0.967	0.996
Test accuracy	0.957	0.943	0.935	0.982

Comparing the KNN performance results, for all values of K, with the SVM, it can be concluded that this last classifier works best at separating the activation matrices into classes.

A confusion matrix was generated in order to evaluate the performance of the classification algorithms, providing a detailed breakdown of the model's predictions versus the actual outcomes. In the first diagonal of the confusion matrices are the correct predictions made by the algorithm. The confusion matrix that follows (Figure 5.11) is the one generated as a result of KNN (k=5) and it is used as a visual example taken from the options generated for this layer, in order to comprehend this tool.

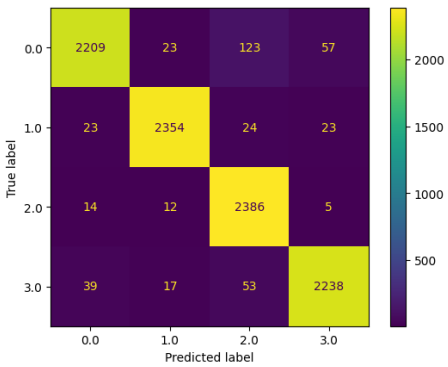


Figure 5.11: Confusion Matrix for the KNN with k=5 performed in the last layer.

As it can be seen in Figure 5.11, the highest number of incorrect predictions was 123 and reflects the number of subjects that were predicted as Asians instead of the true label of the class 0 (Caucasians).

5.3.2.2 Middle layer

Regarding the middle layer, it has already been presented the issue faced with the matrix dimensions. Two options were established in order to approach this issue: the use of the mean and a reshape of the matrices. The results correspondent with these two options for performing KNN are presented in Table 5.13. For SVM the method used was only the mean, so there is only one column with the accuracy values for this classifier.

Table 5.13: Accuracy results for KNN (with both the mean and reshape methods) and SVM in the intermediate layer (with just the mean).

	KNN (Mean)			**KNN (Reshape)**			**SVM**
	K			**K**			-
	5	7	9	5	7	9	**-**
Training accuracy	0.920	0.912	0.908	0.950	0.944	0.941	0.982
Test accuracy	0.867	0.873	0.877	0.912	0.916	0.919	0.958

The best accuracy value was also obtained with SVM, as in the last layer. Moreover, as it was expected, the performance values in general are lower than in the linear layer, as a result of the increase in features that are extracted, which makes the classification process more complex. Again, a confusion matrix was extracted (Figure 5.12).

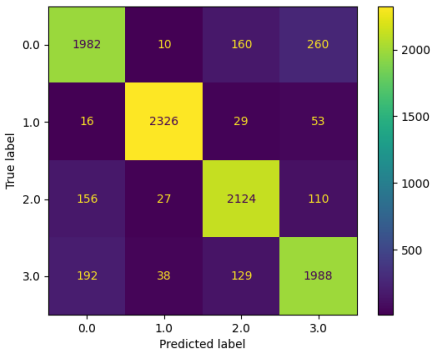


Figure 5.12: Confusion Matrix for KNN with k=9 performed in the middle layer using a reshape of the activation matrices.

Analysing the confusion matrix above, the highest value for a false prediction corresponds to the case in which the true label was Caucasian and the model predicted the label as Indian. As it was the case for Figure 5.11, where the model had the highest value for false predictions associated with the Caucasian as the true label, it can be concluded that even though the model can successfully distinguish between races, the races more easily misclassified in these two layers were a result of possible similarities between the activation matrices from Asians and Indians when compared with Caucasians.

5.3.2.3 Initial Layer

Lastly, the classification was performed on the initial layer of the network. In the precedent section, Chapter 4, it was justified the need to perform the mean for each activation matrix extracted before even compiling the data from all the races into one training vector.

In Table 5.14, the results for the training and testing accuracies with KNN and SVM are layed out.

Table 5.14: Accuracy results for KNN and SVM in the initial layer.

	KNN			**SVM**
	K			-
	5	7	9	**-**
Training accuracy	0.628	0.588	0.564	0.486
Test accuracy	0.460	0.453	0.460	0.480

These results are very low when compared to the other two layers, resulting from the high level of feature abstraction, making it harder to distinguish between races at the beginning of the network. Moreover, the dimension reduction performed in the beginning stages could also be a playing factor in these results, given that there could be a loss of a lot of relevant information.

5.4 Summary

Closing this chapter, the results that should be highlighted are:

- In the image transformations, the increase in brightness affects the African subset more than the other racial groups.
- The gradient maps obtained for the four networks (ResNet-100 trained on MS1MV2 and ResNet-34 trained on MS1MV2, Globalface and Balancedface) show that in the first layer they represent an accurate visual representation of the classified object, the mid-layer shows levels of higher abstraction and the last layer even more. The bigger differences in the maps across networks was in the last layers.
- The extraction of the mean and standard deviation of the neural network activation values did not lead to any relevant conclusions.
- Analysing the results of the classification performance, there is a progressive level of separability of races across the layers of the network.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

In this dissertation, the aim was to investigate, analyze and try to comprehend the presence and implications of racial bias in face recognition systems. This topic is a complex and critical issue within Deep Neural Networks and through some experiments and presented methodologies some conclusions were reached:

1. Existence of racial bias
    - Even though this issue is already well established and its existence is undeniable, the work performed led to the observation and confirmation on how this type of bias can be manifested in the form of differential accuracy values, error rates and metrics, and system performance across racial groups. The analysis of the image transformation results allowed to detect specific modifications that were more impactful on one of the races, while enabling to verify that distinct databases with different content perform differently.
2. Interpretability
    - Grad-CAM visuals were generated for 3 layers of different networks and it is visible that there are variances in the output, especially in the linear layer where the most detail-oriented features are.
3. Disparities among racial groups (inter-racial and intra-racial)
    - The analysis of standardized metrics such as standard deviation, mean, skewed error ratios and confusion matrices allowed to gain an insight on the concepts of inter-racial and intra-racial bias. However, when it came to look for patterns on the way these metrics appear in different races, there was not a straightforward conclusion and the results of the experiments performed were ultimately inconclusive. The focus was trying to analyse intra-racial bias using standard deviation and the results of the euclidean distances calculated, but a conclusion was not achieved.

4. Race classification using network activation matrices

    - The results of the experiments for race classification using activation matrices from 3 layers of the network allowed to conclude that it is possible for classification models to separate races by their neural activation matrices. However, in the first layer of the network, where there is more information and the features are more abstract, this task is not very successful. The performance values presented for the middle layer and final layer agreed with the idea that as we go through the network pipeline it becomes easier to focus on more specific characteristics, helping to set races apart.

5. Call for fairness and impact on society

    - As it was mentioned during this thesis, racial bias is a complex and important topic that needs attention from researchers in different areas of study. The implications of the issue extend beyond automatic systems and can lead to potential human rights violations, privacy issues and perpetuation of inequities. In face recognition technology, there is the urgent need for a fairness-aware solution, implementing bias mitigation and following with the ethical guidelines.

6.2 Future Work

Again, as this research topic is very complex and still very vacant, future studies should be carried out to gain more insight on it. There are some practical tasks that could be of interest for future testing, namely:

- The race classification task was performed by using the totality of the activation matrices, with all the activation values. In the future, it could be interesting to try to use some specific metrics from each matrix to perform the classification: the use of the maximum activation value, the minimum, and a vector with the mean, STD, maximum and minimum. This could lead to some different results for the performances and possibly help to reach more conclusions.

- In order to tackle the mitigation of racial bias, there could be more experiments on some bias mitigation techniques: evaluation of pre-existent algorithms, data collection strategies and pre-processing methods.

- Focus more on the possibility of a collaboration between different areas of science (e.g. computer science and neuroscience), encouraging interdisciplinary approaches and including some practical experiments that involve both of them.

- Try to force racial bias in the experiments and then measure the quantitative impact of it. This could be implemented in the neural activation experiments made by adding some bias on purpose and see the effects in the results.

References

[1] M. Hassaballah and Saleh Aly. Face recognition: challenges, achievements and future directions. *IET Computer Vision*, 9(4):614–626, 2015. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-cvi.2014.0084, arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-cvi.2014.0084, doi:https://doi.org/10.1049/iet-cvi.2014.0084.

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023. URL: https://www.sciencedirect.com/science/article/pii/S1566253523001148, doi:https://doi.org/10.1016/j.inffus.2023.101805.

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[4] Mathias M. Adankon and Mohamed Cheriet. *Support Vector Machine*, pages 1303–1308. Springer US, Boston, MA, 2009. URL: https://doi.org/10.1007/978-0-387-73003-5_299, doi:10.1007/978-0-387-73003-5_299.

[5] Atsu Alagah Komlavi, Kadri Chaibou, and Harouna Naroua. Comparative study of machine learning algorithms for face recognition. *Revue Africaine de Recherche en Informatique et Mathématiques Appliquées*, 2022.

[6] Lukas Mosser, Olivier Dubrule, and Martin Blunt. Stochastic reconstruction of an oolitic limestone by generative adversarial networks. *Transport in Porous Media*, 125, 10 2018. doi:10.1007/s11242-018-1039-9.

[7] Huo Yingge, Imran Ali, and Kang-Yoon Lee. Deep neural networks on chip - a survey. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 589–592, 2020. doi:10.1109/BigComp48618.2020.00016.

[8] Murat Taskiran, Nihan Kahraman, and Cigdem Eroglu Erdem. Face recognition: Past, present and future (a review). *Digital Signal Processing*, 106:102809, 2020. URL: https://www.sciencedirect.com/science/article/pii/S1051200420301548, doi:https://doi.org/10.1016/j.dsp.2020.102809.

[9] Rajeev Ranjan, Swami Sankaranarayanan, Ankan Bansal, Navaneeth Bodla, Jun-Cheng Chen, Vishal M Patel, Carlos D Castillo, and Rama Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, 35(1):66–83, 2018.

[10] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[11] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[12] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

[13] Insaf Adjabi, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. Past, present, and future of face recognition: A review. *Electronics*, 9(8), 2020. URL: https://www.mdpi.com/2079-9292/9/8/1188, doi:10.3390/electronics9081188.

[14] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 692–702, 2019.

[15] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1578–1587, 2022.

[16] Yoanna Martínez-Díaz, Miguel Nicolás-Díaz, Heydi Vazquez, Luis Luévano García, Leonardo Chang, Miguel Gonzalez-Mendoza, and Luis Sucar. Benchmarking lightweight face architectures on specific face recognition scenarios. *Artificial Intelligence Review*, 54, 12 2021. doi:10.1007/s10462-021-09974-2.

[17] Alice J Otoole. *Handbook of Face Recognition || Psychological and Neural Perspectives on Human Face Recognition*. Chapter, 2005.

[18] Ana Valdivia, Júlia Corbera Serrajòrdia, and Aneta Swianiewicz. There is an elephant in the room: Towards a critique on the use of fairness in biometrics. *AI and Ethics*, pages 1–16, 2022.

[19] Rabia Jafri and Hamid Arabnia. A survey of face recognition techniques. *JIPS*, 5:41–68, 06 2009. doi:10.3745/JIPS.2009.5.2.041.

[20] Jayson Killoran, Yuanyuan Gina Cui, Andrew Park, Patrick van Esch, and Jan Kietzmann. Can behavioral biometrics make everyone happy? *Business Horizons*, 2023.

[21] Adam Bohr and Kaveh Memarzadeh. The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare*, pages 25–60. Elsevier, 2020.

[22] Pramila P Shinde and Seema Shah. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pages 1–6. IEEE, 2018.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Anders Krogh. What are artificial neural networks? *Nature biotechnology*, 26(2):195–197, 2008.

[25] Hung Le and Ali Borji. What are the receptive, effective receptive, and projective fields of neurons in convolutional neural networks? *CoRR*, abs/1705.07049, 2017. URL: http://arxiv.org/abs/1705.07049, arXiv:1705.07049.

[26] Alice J O'Toole and Carlos D Castillo. Face recognition by humans and machines: three fundamental advances from deep learning. *Annual Review of Vision Science*, 7:543–570, 2021.

[27] Muhammad Atif Butt, Adnan Qayyum, Hassan Ali, Ala Al-Fuqaha, and Junaid Qadir. Towards secure private and trustworthy human-centric embedded machine learning: An emotion-aware facial recognition case study. *Computers & Security*, 125:103058, 2023. URL: https://www.sciencedirect.com/science/article/pii/S0167404822004503, doi:https://doi.org/10.1016/j.cose.2022.103058.

[28] Pedro C Neto, Tiago Gonçalves, João Ribeiro Pinto, Wilson Silva, Ana F Sequeira, Arun Ross, and Jaime S Cardoso. Explainable biometrics in the age of deep learning. *arXiv preprint arXiv:2208.09500*, 2022.

[29] Mei Wang, Yaobin Zhang, and Weihong Deng. Meta balanced network for fair face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):8433–8448, 2021.

[30] Gautam Srivastava and Surajit Bag. Modern-day marketing concepts based on face recognition and neuro-marketing: a review and future research directions. *Benchmarking: An International Journal*, 2023.

[31] Vera Lúcia Raposo. (do not) remember my face: uses of facial recognition technology in light of the general data protection regulation. *Information & Communications Technology Law*, 32(1):45–63, 2023.

[32] Vicki Bruce and Andy Young. Understanding face recognition. *British Journal of Psychology*, 77(3):305–327, 1986. URL: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8295.1986.tb02199.x, arXiv:https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8295.1986.tb02199.x, doi:https://doi.org/10.1111/j.2044-8295.1986.tb02199.x.

[33] Katharina Dobs, Leyla Isik, Dimitrios Pantazis, and Nancy Kanwisher. How face perception unfolds over time. *Nature Communications*, 10(1):1258, 2019. URL: https://doi.org/10.1038/s41467-019-09239-1, doi:10.1038/s41467-019-09239-1.

[34] Tobias Brosch, Eyal Bar-David, and Elizabeth A Phelps. Implicit race bias decreases the similarity of neural representations of black and white faces. *Psychological science*, 24(2):160–166, 2013.

[35] Vaidehi Natu and Alice J O'Toole. The neural processing of familiar and unfamiliar faces: A review and synopsis. *British journal of psychology*, 102(4):726–747, 2011.

[36] Jacqueline G Cavazos, Géraldine Jeckeln, Ying Hu, and Alice J O'Toole. Strategies of face recognition by humans and machines. *Deep Learning-Based Face Analytics*, pages 361–379, 2021.

[37] Tim Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, 43(2):161–204, 1991.

[38] Roy S Malpass and Jerome Kravitz. Recognition for faces of own and other race. *Journal of personality and social psychology*, 13(4):330, 1969.

[39] Rob Jenkins, AJ Dowsett, and AM Burton. How many faces do people know? *Proceedings of the Royal Society B*, 285(1888):20181319, 2018.

[40] Daniel B. Elbich and Suzanne Scherf. Beyond the ffa: Brain-behavior correspondences in face recognition abilities. *NeuroImage*, 147:409–422, 2017. URL: https://www.sciencedirect.com/science/article/pii/S1053811916307637, doi:https://doi.org/10.1016/j.neuroimage.2016.12.042.

[41] JANOS KURUCZ and GABRIEL FELDMAR. Prosopo-affective agnosia as a symptom of cerebral organic disease. *Journal of the American Geriatrics Society*, 27(5):225–230, 1979.

[42] Tirta Susilo and Bradley Duchaine. Advances in developmental prosopagnosia research. *Current opinion in neurobiology*, 23(3):423–429, 2013.

[43] Adam Chekroud, Jim Everett, Holly Bridge, and Miles Hewstone. A review of neuroimaging studies of race-related prejudice: does amygdala response reflect threat? *Frontiers in Human Neuroscience*, 8, 2014. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2014.00179, doi:10.3389/fnhum.2014.00179.

[44] Gizelle Anzures, Paul C Quinn, Olivier Pascalis, Alan M Slater, James W Tanaka, and Kang Lee. Developmental origins of the other-race effect. *Current directions in psychological science*, 22(3):173–178, 2013.

[45] Sophie Lebrecht, Lara J Pierce, Michael J Tarr, and James W Tanaka. Perceptual other-race training reduces implicit racial bias. *PloS one*, 4(1):e4215, 2009.

[46] Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111, 2020.

[47] Wen Xiao, Genyue Fu, Paul Quinn, Jinliang Qin, James Tanaka, Olivier Pascalis, and Kang Lee. Individuation training with other-race faces reduces preschoolers' implicit racial bias: A link between perceptual and social representation of faces in children. *Developmental Science*, 18, 10 2014. doi:10.1111/desc.12241.

[48] Miao K. Qian, Paul C. Quinn, Gail D. Heyman, Olivier Pascalis, Genyue Fu, and Kang Lee. A long-term effect of perceptual individuation training on reducing implicit racial bias in preschool children. *Child Development*, 90(3):e290–e305, 2019. URL: https://srcd.onlinelibrary.wiley.com/doi/abs/10.1111/cdev.12971, arXiv:https://srcd.onlinelibrary.wiley.com/doi/pdf/10.1111/cdev.12971, doi:https://doi.org/10.1111/cdev.12971.

[49] Keith B Senholzi, Brendan E Depue, Joshua Correll, Marie T Banich, and Tiffany A Ito. Brain activation underlying threat detection to targets of different races. *Social neuroscience*, 10(6):651–662, 2015.

[50] Brent L Hughes, Nicholas P Camp, Jesse Gomez, Vaidehi S Natu, Kalanit Grill-Spector, and Jennifer L Eberhardt. Neural adaptation to faces reveals racial outgroup homogeneity effects in early perception. *Proceedings of the National Academy of Sciences*, 116(29):14532–14537, 2019.

[51] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science Tech Report*, 2016.

[52] Alice J O'Toole, Carlos D Castillo, Connor J Parde, Matthew Q Hill, and Rama Chellappa. Face space representations in deep convolutional neural networks. *Trends in cognitive sciences*, 22(9):794–809, 2018.

[53] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[54] ALICE J O'TOOLE, Kenneth Deffenbacher, Hervé Abdi, and James C Bartlett. Simulating the 'other-race effect'as a problem in perceptual learning. *Connection Science*, 3(2):163–178, 1991.

[55] 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. *Cognitive science*, 26(6):797–815, 2002.

[56] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on information forensics and security*, 7(6):1789–1801, 2012.

[57] Shruti Nagpal, Maneet Singh, Richa Singh, and Mayank Vatsa. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019.

[58] Kashmir Hill. Wrongfully accused by an algorithm. In *Ethics of Data and Analytics*, pages 138–142. Auerbach Publications, 2022.

[59] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019.

[60] Kumar Abhishek and Deeksha Kamath. Attribution-based xai methods in computer vision: A review. *arXiv preprint arXiv:2211.14736*, 2022.

[61] Pedro C. Neto, Sara P. Oliveira, Diana Montezuma, João Fraga, Ana Monteiro, Liliana Ribeiro, Sofia Gonçalves, Isabel M. Pinto, and Jaime S. Cardoso. imil4path: A semi-supervised interpretable approach for colorectal whole-slide images. *Cancers*, 14(10), 2022. URL: https://www.mdpi.com/2072-6694/14/10/2489, doi:10.3390/cancers14102489.

[62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[63] Barnaby Crook, Maximilian Schlüter, and Timo Speith. Revisiting the performance-explainability trade-off in explainable artificial intelligence (xai). *arXiv preprint arXiv:2307.14239*, 2023.

[64] Adrian Erasmus, Tyler DP Brunet, and Eyal Fisher. What is interpretability? *Philosophy & Technology*, 34(4):833–862, 2021.

[65] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[66] Lindsay I Smith. A tutorial on principal components analysis. *Technical Report OUCS-2002-12*, 2002.

[67] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[68] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[69] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition*, pages 586–587. IEEE Computer Society, 1991.

[70] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[71] Weihong Deng, Jiani Hu, Jiwen Lu, and Jun Guo. Transform-invariant pca: A unified approach to fully automatic facealignment, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1275–1284, 2013.

[72] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):328–340, 2005.

[73] Weihong Deng, Jiani Hu, and Jun Guo. Face recognition via collaborative representation: Its discriminant nature and superposed representation. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2513–2521, 2017.

[74] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006.

[75] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 2707–2714. IEEE, 2010.

[76] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[77] Md Tahmid Hasan Fuad, Awal Ahmed Fime, Delowar Sikder, Md Akil Raihan Iftee, Jakaria Rabbi, Mabrook S Al-Rakhami, Abdu Gumaei, Ovishake Sen, Mohtasim Fuad, and Md Nazrul Islam. Recent advances in deep learning techniques for face recognition. *IEEE Access*, 9:99112–99142, 2021.

[78] Ibrahim Mahmood Rashid Al-Bakri, Muhammad Imran Ahmad, Mohd Nazrin Md Isa, and Mustafa Zuhaer Nayef Al-Dabagh. A review paper on face recognition techniques. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1951–1956, 2023. doi:10.1109/ICACCS57279.2023.10112907.

[79] Pedro C. Neto, Ana F. Sequeira, and Jaime S. Cardoso. Myope models - are face presentation attack detection models short-sighted? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 390–399, January 2022.

[80] Eduarda Caldeira, Pedro C. Neto, Tiago Gonçalves, Naser Damer, Ana F. Sequeira, and Jaime S. Cardoso. Unveiling the two-faced truth: Disentangling morphed identities for face morphing detection, 2023. arXiv:2306.03002.

[81] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57:137–154, 2004.

[82] Shumeet Baluja, Mehran Sahami, and Henry A Rowley. Efficient face orientation discrimination. In *2004 International Conference on Image Processing, 2004. ICIP'04.*, volume 1, pages 589–592. IEEE, 2004.

[83] Hongming Zhang, Wen Gao, Xilin Chen, and Debin Zhao. Object detection using spatial histogram features. *Image and Vision Computing*, 24(4):327–341, 2006.

[84] Xiaoyu Wang, Tony X Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *2009 IEEE 12th international conference on computer vision*, pages 32–39. IEEE, 2009.

[85] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[86] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[87] Oya Çeliktutan, Sezer Ulukaya, and Bülent Sankur. A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1):1–27, 2013.

[88] Dhananjay Rathod, A Vinay, S Shylaja, and S Natarajan. Facial landmark localization-a literature survey. *Int J Current Eng Technol*, 4(3):1901–1907, 2014.

[89] Esra Nur Sandıkcı, Çiğdem Eroğlu Erdem, and Sezer Ulukaya. A comparison of facial landmark detection methods. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2018.

[90] Feng Liu, Qijun Zhao, Xiaoming Liu, and Dan Zeng. Joint face alignment and 3d face reconstruction with application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):664–678, 2018.

[91] Timothy F Cootes, Christopher J Taylor, and Andreas Lanitis. Active shape models: Evaluation of a multi-resolution method for improving image search. In *BMVC*, volume 1, pages 327–336. Citeseer, 1994.

[92] David Cristinacce and Timothy F Cootes. Boosted regression active shape models. In *BMVC*, volume 2, pages 880–889. Citeseer, 2007.

[93] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. In *Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5*, pages 484–498. Springer, 1998.

[94] Iain Matthews and Simon Baker. Active appearance models revisited. *International journal of computer vision*, 60:135–164, 2004.

[95] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[96] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 94–108. Springer, 2014.

[97] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[98] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[99] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[100] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[101] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[102] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014.

[103] Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15, 2002.

[104] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. *Advances in neural information processing systems*, 27, 2014.

[105] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015.

[106] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[107] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

[108] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[109] Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, and Le Song. Deep hyperspherical learning. *Advances in neural information processing systems*, 30, 2017.

[110] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[111] Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10052–10061, 2019.

[112] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11947–11956, 2019.

[113] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018.

[114] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017.

[115] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

[116] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[117] Ferdinando S Samaria and Andy C Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE workshop on applications of computer vision*, pages 138–142. IEEE, 1994.

[118] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[119] 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 14*, pages 87–102. Springer, 2016.

[120] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[121] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016.

[122] Vidit Jain. The indian face database. *http://vis-www. cs. umass. edu/vidit/IndianFaceDatabase/*, 2002.

[123] Yi Dong, Lei Zhen, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 5, 2014.

[124] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7044–7053, 2017.

[125] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

[126] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7), 2018.

[127] Yaobin Zhang, Weihong Deng, Mei Wang, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. Global-local gcn: Large-scale label noise cleansing for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7731–7740, 2020.

[128] Xiaohang Zhan, Ziwei Liu, Junjie Yan, Dahua Lin, and Chen Change Loy. Consensus-driven propagation in massive unlabeled data for face recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 568–583, 2018.

[129] P Jonathon Phillips. A cross benchmark assessment of a deep convolutional neural network for face recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 705–710. IEEE, 2017.

[130] Isabelle Hupont and Carles Fernández. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019.

[131] Ignacio Serna, Aythami Morales, Julian Fierrez, and Nick Obradovich. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence*, 305:103682, 2022.

[132] P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):1–11, 2011.

[133] J Ross Beveridge, Geof H Givens, P Jonathon Phillips, Bruce A Draper, and Yui Man Lui. Focus on quality, predicting frvt 2006 performance. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8. IEEE, 2008.

[134] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1931–1939, 2015.

[135] 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 14*, pages 499–515. Springer, 2016.

[136] Pedro C Neto, Eduarda Caldeira, Jaime S Cardoso, and Ana F Sequeira. Compressed models decompress race biases: What quantized models forget for fair face recognition. *arXiv preprint arXiv:2308.11840*, 2023.

[137] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[138] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021.

[139] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[140] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14234, 2021.

[141] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.

[142] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[143] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus

benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98, 2017.

[144] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.

[145] Yonghyun Kim, Wonpyo Park, Myung-Cheol Roh, and Jongju Shin. Groupface: Learning latent groups and constructing group-based representations for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5621–5630, 2020.