BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

SECURITY IMPLICATIONS OF DEEPFAKES IN FACE AUTHENTICATION

BEZPEČNOSTNÍ DOPADY DEEPFAKES V OBLASTI ROZPOZNÁVÁNÍ OBLIČEJŮ

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR **Bc. MILAN ŠALKO**
AUTOR PRÁCE

SUPERVISOR **Ing. ANTON FIRC**
VEDOUCÍ PRÁCE

BRNO 2023

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Master's Thesis Assignment

141060

Institut:	Department of Intelligent Systems (UITS)
Student:	**Šalko Milan, Bc.**
Programme:	Information Technology and Artificial Intelligence
Specialization:	Computer Vision
Title:	**Security Implications of Deepfakes in Face Authentication**
Category:	Security
Academic year:	2022/23

Assignment:

1. Study deepfakes and face recognition systems. Focus on the current state of deepfake abilities to spoof face authentication.
2. Get familiar with the currently available tools and techniques for creating facial deepfakes, verify their technical feasibility, and evaluate their availability and complexity.
3. Define possible deepfake spoofing attacks on two commercial face recognition systems. Focus on the usage of face-swap and reenactment deepfakes.
4. Experimentally execute defined spoofing attacks and evaluate the resilience of selected face recognition systems. Use statistical tests to evaluate differences in matching scores.
5. Discuss the feasibility of spoofing face recognition systems using deepfakes and suggest possible methods for protection.

Literature:

- FIRC Anton a MALINKA Kamil. The dawn of a text-dependent society: deepfakes as a threat to speech verification systems. In: *SAC '22: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. New York, NY: Association for Computing Machinery, 2022, s. 1646-1655.
- FIRC Anton, MALINKA Kamil a HANÁČEK Petr. Creation and detection of malicious synthetic media - a preliminary survey on deepfakes. In: *Sborník příspevků z 54. konference EurOpen.CZ, 28.5.-1.6.2022*. Radešín, 2022, s. 125-145. ISBN 978-80-86583-34-1.
- Handbook of Digital Face Manipulation and Detection. (2022). In C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, & C. Busch (Eds.), Advances in Computer Vision and Pattern Recognition. Springer International Publishing. https://doi.org/10.1007/978-3-030-87664-7

Requirements for the semestral defence:
1-3

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Firc Anton, Ing.**
Head of Department:	Hanáček Petr, doc. Dr. Ing.
Beginning of work:	1.11.2022
Submission deadline:	17.5.2023
Approval date:	3.11.2022

Abstract

Deepfakes, media generated by deep learning that are indistinguishable to humans from real ones, have experienced a huge boom in recent years. Several dozen papers have already been written about their ability to fool people. Equally, if not more, serious, may be the problem of the extent to which facial and voice recognition systems are vulnerable to them. The misuse of deepfakes against automated facial recognition systems can threaten many areas of our lives, such as finances and access to buildings. This topic is essentially an unexplored problem. This thesis aims to investigate the technical feasibility of an attack on facial recognition. The experiments described in the thesis show that this attack is not only feasible but moreover, the attacker does not need many resources for the attack. The scope of this problem is also described in the work. The conclusion also describes some proposed solutions to this problem, which may not be difficult to implement at all.

Abstrakt

Deepfakes, médiá generované hlbokým strojovým učením, ktoré sú pre človeka nerozoznateľné od skutočných, zažívajú v posledných rokoch obrovský rozmach. O ich schopnosti oklamať ľudí už bolo napísaných niekoľko desiatok článkov. Rovnako závažný, ak nie závažnejší, môže byť problém, do akej miery sú voči nim zraniteľné systémy rozpoznávania tváre a hlasu. Zneužitie deepfakes proti automatizovaným systémom rozpoznávania tváre môže ohroziť mnohé oblasti nášho života, napríklad financie a prístup do budov. Táto téma je v podstate nepreskúmaným problémom. Cieľom tejto práce je preskúmať technickú realizovateľnosť útoku na rozpoznávanie tváre. Experimenty opísané v práci ukazujú, že tento útok je nielen uskutočniteľný, ale navyše útočník naň nepotrebuje veľa prostriedkov. V práci je opísaný aj rozsah tohto problému. V závere je opísaných aj niekoľko navrhovaných riešení tohto problému, ktoré vôbec nemusia byť náročné na implementáciu.

Keywords

deepfakes, facial recognition, biometrics systems, machine learning, cyber security

Klíčová slova

deepfakes, rozpoznavanie tváre, biometrické systémy, strojové učenie, kyberbezpečnosť

Reference

ŠALKO, Milan. *Security Implications of Deepfakes in Face Authentication.* Brno, 2023. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Anton Firc

Rozšířený abstrakt

Deepfakes, syntetické médiá vytvorené umelou inteligenciou, sa v posledných rokoch stali rozšíreným fenoménom. Rýchlosť, akou sa toto odvetvie výskumu umelej inteligencie rozvíja, vyráža dych. Nástroje na generovanie deepfake dokážu vytvoriť alebo upraviť obsah za zlomok času alebo nákladov oproti tradičnej úprave videa alebo obrazu. Každý teda môže vytvoriť deepfakes v online nástroji bez hlbokých znalostí neurónových sietí. Dôkazom je aj to, ako sa deepfakes rozšírili na sociálnych sieťach, kde sa zvyčajne používajú na zábavu. Alebo to, že sa varovania pred ich možným zneužitím objavujú aj v mainstreamových médiách.

Deepfakes napriek svojej pomerne zlej povesti môžu mať aj množstvo pozitívnych využití, či už ide o správy s umelou inteligenciou, kde syntetický obraz moderátora 24 hodín denne 7 dní v týždni prináša divákom najnovšie udalosti, môžu pomôcť aj pri tvorbe filmov, školských materiálov alebo v zdravotníctve, kde vratia hlas ľuďom, ktorí ho stratili v dôsledku choroby. Tieto spôsoby využitia sú však ešte len v počiatkoch.

Realistickosť deepfakes, ktorá mnohokrát znemožňuje ľuďom rozoznať deepfakes od skutočných medií, prináša nové riziká. V kombinácii s ich širokou dostupnosťou pre všetkých a zlým úmyslom môžu mať veľké množstvo negatívnych využití. Deepfakes sa tak stávajú hrozbou pre jednotlivcov, kde môžu byť použité pri nových útokoch. Príkladom sú nové formy phishingu, rôzne formy spoofingu alebo vytváranie videí a fotografií s cieľom zdiskreditovať danú osobu. Deepfakes však predstavujú hrozbu aj pre spoločnosť ako celok, napríklad ich použitie pri vytváraní fakenews alebo pri všeobecnom podkopávaní dôveryhodnosti verejných orgánov v očiach občanov.

Rovnako závažným, ak nie závažnejším problémom môže byť nielen to, či deepfakes môžu oklamať ľudí, ale aj to, do akej miery sú voči nim zraniteľné systémy rozpoznávania tváre a hlasu. Najmä ak si uvedomíme, že biometrické systémy rozpoznávania tváre sa stali bežnou súčasťou nášho každodenného života a čiastočne vytlačili napríklad prihlasovanie pomocou odtlačkov prstov v mobilných telefónoch. Prihlasovanie do aplikácií a online bankovníctva sú dobrými príkladmi, na ktoré by sa mohol zamerať útočník, ktorý sa rozhodne použiť deepfakes proti týmto systémom.

V poslednom čase bolo vyvinutých niekoľko nástrojov, ktoré sú dostupné na internete, napríklad rôzne voľne dostupné výskumné práce alebo platené nástroje, ktoré poskytujú možnosť vytvoriť pomerne vierohodný deepfake. V súčasnosti neexistuje veľa prác, ktoré by sa zameriavali na overenie technickej realizovateľnosti útoku na systémy rozpoznávania tváre. Cieľom tejto práce je odpovedať na otázky súvisiace s používaním deepfakes a ich použitím proti biometrii tváre. Na základe týchto skutočností a poznatkov získaných v počiatočných fázach tejto práce bolo navrhnutých niekoľko vektorov útoku. Tie by mali odpovedať na základnú otázku vyplývajúcu z tejto práce, a to: Ako sú bežne dostupné komerčné biometrické systémy tváre zraniteľné voči útokom využívajúcim deepfakes?

Prvý vektor útoku ma overiť technickú realizovateľnosť útoku na rozpoznávania tváre pomocou deepfakes. Táto tvorba je nenáročná na zdroje a znalosti takže platí že kvalitný deepfake je možné vytvoriť za použitia telefónu a človek nemusí mať žiadne väčšie znalosti ako sa deepfakes generujú. Ďalej sme sa oboznámili s komerčnými riešeniami na ktorých sme plánovali zisťovať ako veľmi sú biometrické systémy zraniteľné voči deepfakes. Tieto systémy majú využitie v rôznych oblastiach ako je online bankovníctvo alebo overovaní identity v aplikáciách.

Za použitia voľne dostupných generátorov deepfakes a dodržania pár zásad sme vytvorili deepfakes ktoré boli dostatočne kvalitné aby ich systém pre rozpoznávanie považoval za platný vstup. Je nutné dodať že naivný prístup prenesenia tváre na akúkoľvek inú osobu sa

ukázal ako chybný. Pri generovaní je nutné dať si pozor aby aj herec ktorý nám poslúži ako podklad mal podobné rysy tváre. Je to mierne skomplikovanie útoku ale nie neriešiteľný problém. V takýchto prípadoch sa dá ľahko najať herec alebo využiť množstvo videí dostupných na internete. Pri dodržaní týchto zásad systém identifikuje média prezentované útočníkom ako vybranú obeť. Tieto zistenia môžeme využiť pri návrhu ďalších experimentov, ktoré budú mať za ciel odhaliť aký rozsiahly je tento problém. Ďalším zaujímavým zistením je že one-shot systémy si nevedia poradiť so syntézou v prípade že je pri overení nutne otočiť hlavu o 90°, pripadne keď si osoba musí prejsť rukou pred tvárou. Toto zistenie môže slúžiť ako možná ochrana proti tomuto typu útoku.

V nasledujúcom útoku sme overovali tieto zistenia. Nakoľko neexistuje dataset ktorý by slúžil pre kontrolu odolnosti biometrických systémov voči deepfakes. Rozhodli sme sa zobrať už existujúci deepfake dataset z ktorého sme vyberali fotky ktoré spĺňajú ICAO štandard. Potom sme začali s meraním nad samotnými biometriami. Zo získaných dát a za použitia štatistických testov sme zistili že biometrické systémy pre rozoznávanie tváre môžu byť oklamané pomocou deepfakes. Nakoľko deepfakes skóre niesu štatistický podobné s impostor skóre. To znamená, že biometria nevie jednoznačne odmietnuť deepfakes. Tento problem sa zväčší pri skutočnosti, že použitý dataset už je pár rokov starý a deepfakes v ňom sú ďaleko za kvalitou dnešných deepfakes. Z nameraných dát je tiež zrejmé že v pripade kvalitnejších dát sa podobnosť posunie skôr k genuine skóre.

Riešením tohto problému môže byť využitie viacerých fotografií z videa namiesto jednej pre overenie identity osoby. Tiež môže pomôcť ak od užívateľa budeme vyžadovať vykonávanie úkonov s ktorými majú dnešné generátory deepfakes stále problém ako je zamávanie pred tvárou alebo otočenie hlavy.

Security Implications of Deepfakes in Face Authentication

Declaration

I declare that I have prepared this thesis independently under the supervision of Ing. Anton Firc. I have listed all literary sources, publications and other sources from which I have drawn.

.....
Milan Šalko
May 15, 2023

Acknowledgements

I would like to thank my supervisor Ing. Anton Firc, for his professional and personal help in the creation of this master's thesis. Furthermore, I would like to thank Mgr. Kamil Malinka Ph.D. for his valuable advice, without which this work would not have been possible. Thanks also go to my parents and brother for the love and support they give me. Finally, I would like to thank my friends who have been my support during the writing of this thesis.

Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Contents

Chapter 1

Introduction

Deepfakes, synthetic media created by artificial intelligence [37], are a relatively new phenomenon. The quality of these media has passed the point where people cannot reliably distinguish them from the real thing. They have the ability to alter our perception and handling of information considerably. One of the most popular applications of deepfake technology is facial deepfakes. This is evidenced by the number of tools used to create them. Although deepfakes can have many positive applications in the entertainment industry [25] or in medicine [63], the risks should not be underestimated.

There is often talk about the misuse of deepfakes to create fake news and propaganda. Or about their potential for various forms of fraud and blackmail [5]. Undoubtedly, people often cannot distinguish deepfakes images from real ones [40]. However, there is less discussion about the risks that they pose to facial recognition systems. These systems have become part of everyday life, from logging into mobile banking to security controls at airports and other public places. There is currently unclear how much of a problem the deepfakes may be for these systems. And in fact, the whole topic of deepfakes and their risks to automatic facial recognition systems is an unexplored area.

This work aims to investigate how much of a problem the currently freely available tools for creating facial deepfakes are for face recognition systems deployed in common areas of life. Areas, where it makes the most sense for an attacker to use deepfakes to overcome the system will be selected. The first experiment is to determine what steps are required to create a deepfake capable of overcoming a facial recognition system. It also describes the limitations of various deepfakes facial generation systems that can be used in an attack. We will also focus on the technical complexity and describe the knowledge an attacker must have to execute the attack successfully.

The next experiment builds on the first experiment's results and verifies how big a problem deepfakes pose for face recognition systems. This experiment's results also highlight that face recognition systems cannot reliably distinguish deepfakes from valid inputs. In the last experiment, our goal was to verify whether the robustness of the system against deepfakes increases when the system requires a short video for verification as more frames are verified. Several questions were asked during the experiments, which were answered sequentially during the evaluation of the experiments. At the end of the thesis, there is also a discussion highlighting some of the problems associated with the current setup and the quality of the available datasets. The main contributions of the work are:

- This work shows that modern facial deepfakes can pose an important threat to face recognition systems.

- We have also shown that attacking facial biometrics using deepfakes is technically feasible, even with minimal resource requirements.
- Several ways to improve face recognition systems against deepfakes have been described in the thesis.
- It has been shown that it is necessary for datasets to capture state-of-the-art ways of generating facial deepfakes.

In Chapter 2, facial biometric systems and facial recognition applications were discussed. Deepfakes, including the risks and available tools, were described in Chapter 3. In Chapter 4, individual attacks were proposed, questions were also asked and answered in Chapter 5, and findings from experiments were presented. Chapter 6 discusses the results obtained, and the expected direction of future work is also presented.

Chapter 2

Biometric system

Biometrics is the science of determining an individual's identity based on physiological (e.g., fingerprint, face), chemical (e.g., DNA) or behavioural characteristics (e.g., speech, handwriting). This characteristic must be both specific to a particular person and measurable. The importance of biometrics has arisen in need for large-scale identity management systems, where accurately determining an individual's identity is a necessity [48]. A biometric system can be defined as a pattern recognition system that extracts a set of features and compares it with the features in the database based on the biometric data obtained from an individual. Depending on the context of the application, a biometric system can operate either in authentication mode or identification mode [29]. In this chapter, the concept of biometric systems and biometrics will be introduced. It will also describe the metrics used for the performance of biometric systems. The second part of the chapter will introduce face recognition and methods for liveness checking. In the last part, facial recognition tools will be introduced.

2.1 Performence measurement of biometric systems

This section discusses insights from [4]. Biometric systems usually do not compare two identical samples of a user's biometric traits since changes occur with each scan, whether caused by the sensor, the environment, or the user.

Many real and fake attempts are used to measure the biometric performance of a biometric system, and all are stored with a similarity score. A genuine attempt is a single attempt by a user to match his/her own stored template. An impostor attempt is an attempt to match someone else's template.

The distributions of both fake and true attempts are probability density functions that tell us how many attempts fall into a given interval of the matching score. Plotting these probability density functions gives us matching scores distribution graph. Pairs of False Reject Rate (FRR) and False Accept Rate (FAR) values can be calculated by applying a varying score.

2.2 Facial recognition

Facial recognition is a biometric technology that uses an individual's facial characteristics to identify them uniquely. Facial recognition systems can identify people in photographs, videos or in real-time. Facial recognition systems are most often based on digital images.

A face analyzer is a software that identifies or confirms a person's identity using their face. Automatic face recognition can be viewed as a pattern recognition problem where recognition must take place in a high-dimensional space. [19] There are several approaches to shape recognition:

2.2.1 Linear/Nonlinear projection methods

The most known method is the Turk and Pentland method [60], which is based on the Principal Component Analysis (PCA) method. A set of orthogonal eigenvectors resulting from PCA have been called eigenfaces because of their similarity to faces.

The first step is to find the face in the image. Then the software identifies nodal points on the face. The nodal points represent the peripheral points of the facial features, most often including the corners of the eyes or the mouth. The distance between these points is then measured: between the eyes, between the nose and the mouth, the distance and shape of the cheekbones, etc. These distances are then to projected onto the eigenfaces to produce a set of projection weights. These projection weights were used according to the nearest neighbour rule to select the image that is closest in the space of weights. Thanks to this, the closest identity was found, and the person was identified.

LDA (Linear Discriminant Analysis) was introduced as a better alternative to PCA. The advantage is that it provides discrimination between classes, whereas PCA deals with the input data as a whole without paying attention to the underlying structure. Indeed, the main goal of LDA is to find the basis vectors providing the best discrimination between classes while trying to maximize the differences between classes and minimize the differences within a class. [36]

2.2.2 The neural networks

Another approach is the use of neural networks. Facenet can be an example of such an approach. A simple scheme can be seen in Figure 2.1. This neural network takes an image as input and computes a vector of size 128, also called an embedding. An embedding is actually a representation of the most important features of a face. An embedding is a representation of a point in Euclidean space. So we can say that the neural network maps the image into Euclidean space where the distance in space corresponds to the similarity of the face. It is also true that the image of person A will be placed closer to the images of person A than to the images of any other person in the database. [52]

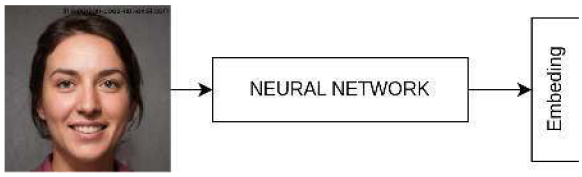

Figure 2.1: FaceNet takes an image of a face as input and outputs the embedding vector.

2.3 Division of face recognition systems

The accuracy of these systems depends on the conditions in which the face was taken. The main elements that affect the quality of the image are the illumination and possible shadows in the face area, the resolution, and the quality of the camera itself.

The following facts come from the source [12]. These systems can be broadly divided into two categories: two-dimensional (2D) and three-dimensional (3D) face recognition. 2D systems collect and process a single two-dimensional image of a face. 3D facial recognition systems use a variety of techniques, such as multi-camera capture, patterned illumination

2D face recognition

2D systems use photos obtained from a regular camera (mobile phone, tablet, access terminal). In this photo, the system recognizes the nodal points of the face such as the mouth eyes, and nose. Based on the distance of these points, the system then creates an internal representation which it compares with the one in the database. The advantage of this technology is a relatively simple setup and also a lower cost.

With these systems, however, we can also use video which contains much more information than in a 2D image. The simplest way to use the additional information in the image is to look at the video as a disordered sequence of 2D images. During testing, each frame of the sequence under test votes independently for a particular identity. A suitable fusion of these votes can be used to obtain a final identity. For example, at the score level where the resulting scores are averaged to get the final decision. However, these approaches do not exploit the temporal information that is also present in the video sequences. [14]

3D face recognition

Systems for 3D face recognition usually use a stereo camera or a scanner to create a three-dimensional representation of the face. The advantage is much more reliable face recognition since 3D systems are able to capture depth in the image. However, there are other ways of capturing faces in 3D, an option is to use multiple 2D cameras and then use deep learning to create a 3D representation. It is also possible to project a pattern onto the face and compute a 3D representation based on the occlusion. The disadvantage of these systems is their high cost. On the other hand, they improve accuracy and reduce false positives. [14]

2.3.1 Facial images resolution

The following findings in this section are based on [13]. International Civil Aviation Organization's standard (ICAO) facial images are currently most commonly used for passport verification and criminal identification. It specifies a minimum resolution which is measured by the number of pixels between the eyes. The standard states that the distance between the eyes must be at least 60 pixels. This distance cannot be achieved by upsampling the photograph or other modifications. The ICAO standard sets out the conditions that must be met:

- The photograph must be free of noise
- The color tone must reflect the real color of the skin.
- The person must be looking directly into the camera with a neutral facial expression.

- All facial features must be visible and not covered by anything
- The background must be uniform, ideally white or light shades of grey with no patterns present
- The light must cover the face evenly and must not create shadows on the face
- The eyes must be visible in case the person wears glasses, and there must be no reflections in the glasses
- In the case of wearing a head covering, the face must be visible from the chin to the forehead.

However, the source images may not always be perfect, which is why there is a nominal resolution of the face. This is based on the assumption that humans are able to reliably recognize a face if it has at least 12 pixels between its tails. For example, older TV broadcasts had a resolution of 480p in sd quality, but even at this low resolution, people are able to recognize individual faces. This fact may be interesting for future face recognition systems.

2.4 Liveness detection

In face recognition, liveness detection is important to prevent spoofing attacks. Live detection is a technique in which an algorithm detects securely whether the source of a biometric sample is from a fake representation or is a real live person. In this section, the most commonly used liveness detection techniques will be discussed. This section discusses the findings from [16].

Blinking based liveness detection

Blink-based vividness detection is one of the most commonly used approaches. The main advantage is the ease of implementation into existing solutions. It also eliminates the need to add additional hardware. Another advantage is naturalness since the average person blinks every 2-4 seconds.

Movement of the eyes based detection

Eye movement analysis was introduced by Hyung-Keun Jee for an embedded facial recognition system. The method detects the eyes in the input images, and then the deviations of each eye region are calculated to determine whether the input face is real or not. The basic assumption is that due to blinking and uncontrolled pupil movements of human eyes, there should be recognizable face changes.

Liveness detection by optical flow

The method analyses the differences and properties of the optical flow generated from 3D objects and 2D planes. The motion of the optical flow field is a combination of four basic motion operations: translation, displacement, rotation, and swing. The authors found that the fourth operation produces differences in the optical flow field. The optical flow field for 2D objects can be represented as a projective transformation. The optical flow allows the derivation of a reference field and thus allows the determination of whether the area

under test is planar or not. For this purpose, the difference between the optical flow fields is calculated. To decide whether a face is a real face or not, this difference is determined by a threshold value

Variable Focusing based analysis

The key approach of the variable focus face liveness detection technique is to exploit the variation of pixel values by focusing pixels taken with a different focus. In the case of real faces, the in-focus regions are bright and the others are blurred due to depth information. In contrast, there is little difference between images taken at different focuses from a printed copy of the face because they are not integral.

2.5 Use cases for facial recognition systems

Facial recognition is one of the most widespread and widely used biometric technologies today. Nowadays, many people regularly come into contact with these systems in various parts of their lives, from security to entertainment, and it is becoming an integral part of our daily lives. This section looks at exactly where facial recognition is currently being used.

2.5.1 Security and survivalence

Facial recognition is widely used in security and surveillance systems to identify potential threats. One of the main advantages of facial recognition technology in security and surveillance is its ability to identify people, even in crowded and busy environments quickly.

Searching for evidence on videotape, on the other hand, is very time-consuming. For example, the process of determining whether a suspected terrorist has visited a location can take thousands of hours of video footage from hundreds of cameras without facial recognition. However, if these recordings are run through a facial recognition system, the search can be reduced to a fraction of the time.

However, the search for people does not have to be limited to criminals or terror suspects. Many people are lost daily for various reasons such as age and mental illness [44].

2.5.2 Banking and finance

With the rapid development of technology, the distance between customers and businesses using online services has shrunk significantly. Therefore, one of the most common use cases for facial recognition is the implementation of Know Your Customer (KYC) checks. KYC is a regulatory requirement for banks, financial services, and insurance companies. It requires financial institutions to verify the identity, eligibility, and risk factors associated with their customers. [42]

This requirement specifically applies to digital banks in terms of preventing false learning and identity fraud. It is also important in recognizing money laundering schemes. In order to verify the identity of a user when creating a new bank account, it is necessary to compare the face from a government-issued document, such as a driving license or passport, with the user's real face that the user takes a picture of using their device [42].

Another use of facial recognition in the financial world can be ATMs that, in addition to the pin code, also capture a photo of the face of the person using the ATM [43].

2.5.3 Access to devices

Facial recognition on mobile phones and laptops is gaining more and more attention due to the large number of applications it can offer. And not just for unlocking but also for accessing applications such as Internet banking apps. Or as a mere replacement for other forms of verification such as fingerprint, pin or pattern entry. The most famous example came in September 2017, when Apple introduced the iPhone X and the Face ID feature [24].

Therefore, another use case is Lenovo's collaboration with NeuroTechnology to create a login system for Lenovo laptops. The goal was to create a system for logging into the operating system where, instead of entering a password or PIN, a photo of the user from the camera would suffice [3].

2.5.4 Age verification

Biometric age verification is another technique that uses unique human characteristics to verify a person's age more accurately. The necessity for this verification lies in the fact that we want to prevent, for example, children from accessing sites with adult content, be it pornography or gambling. It is also necessary to verify the age of the customer when selling certain products.

The use case of such verification is as follows: The user first enters a photo of the document (passport, id card) into the system. The system should verify that the document is valid. Then it prompts the user to take a photo of himself or herself and this photo is compared with the features from the photo on the document. And also the age of the customer is estimated. Age estimation relies on deep neural networks.

An example of the use of age verification is the case of the social network Yubo, which needed to verify the age of its users [7]. Yubo is a youth app and if you are over 50 or under 13, the network should prevent you from creating an account. The procedure was simple, the user entered their age and took a selfie, the neural network estimated the age and if it matched the entered age, the user was verified. However, if it failed to estimate the age the user was asked for a photo of ID or passport.

2.5.5 Elections

The following findings in this section are based on [49]. In recent years, the idea of electronic elections has become more and more common. Although the use of biometrics for voter authentication in elections is not entirely new and a number of states are using them. Examples include Ghana, Afghanistan, India, and Tanzania.

Increasingly, the idea of bringing these elections to smartphones is also emerging. An example is West Virginia which has launched a pilot project for online voting via smartphone in a separate app that serves for identity verification. The voter first uploads a photo of their identification documents and then creates a live photo against which the documents are compared. After facial recognition, a fingerprint is also requested if the device allows it. Another example is Canada which, during the Covid-19 pandemic, introduced parliamentary voting using a parliament-managed smartphone which worked similarly to the previous example.

2.6 Attacks on biometric systems

This section takes insights from [47]. The biometric system is exposed to numerous malicious attacks that can cause various forms of security threats. Malicious attacks on the biometric system pose a security problem and reduce the performance of the system.

An attack can be defined as an attempt by an unauthorized entity to deceive the verifier into believing that the unauthorized entity is a participant. If the attacker is an individual or organization acting with malicious intent to compromise the system. [22]

A generic biometric system can be decomposed into several parts. The phases of such a general system of systems are shown in Figure 2.2. There are a total of eight basic sources of attack on such systems. Some of these attacks can be prevented, for example, by placing the database and some elements in a secure location. Of course, encrypted communication between all parts of the system should also be a matter of course.

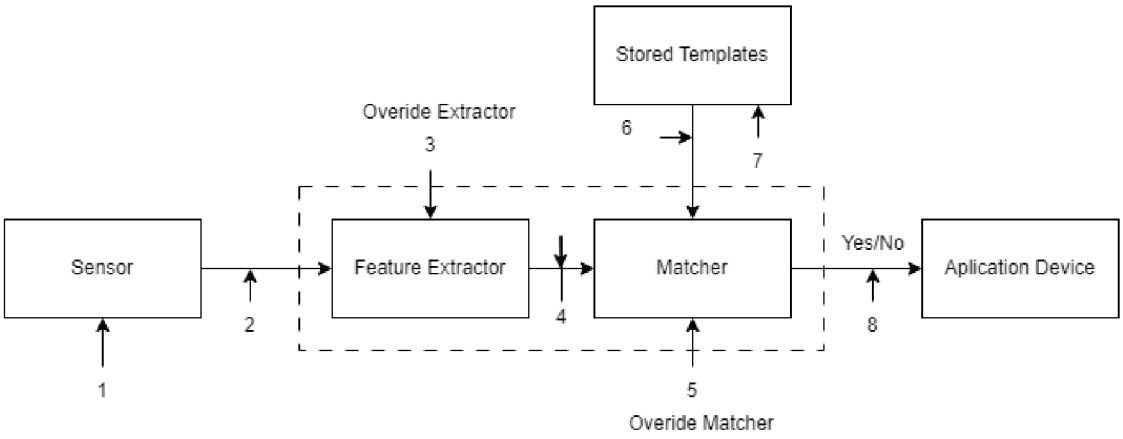

Figure 2.2: Possible attack points in the biometrics-based system.

1. **Fake biometrics on the sensor** – In this type of attack, fake biometric data is used, the features of which are uploaded in the system. Examples are fake fingers, copies of signatures, and face masks.

2. **Resending the old digitally stored biometric signal** – the old recorded signal is inserted into the system, bypassing the sensor. An example is submitting an old copy of a fingerprint image or a speaker's audio signal (replay attack).

3. **Feature extraction override** – a feature extractor could be attacked by an attacker using, for example, a Trojan horse to produce feature sets selected by the attacker

4. **Feature representation interference** – after extracting features from the input signal, these features are replaced by another synthetic feature set (assuming the representation is known). Feature extraction and matcher are usually inseparable, and this method of attack is extremely challenging. However, if the data is transferred to a remote matcher, this threat is real.

5. **Override matcher** – to generate artificially high or low match scores directly

6. **Database-to-matcher data transfer attack** – templates from the stored database are sent to the matcher, an attacker can try to modify these templates before they reach the matcher.

7. **Manipulation of stored templates** – an attacker attempts one or more or more stored templates in the database could be altered, which could lead to the authorization of a fraudulent person.

8. **Changing the final decision** - the attacker can try to change the result and replace it with his own. Even if a real pattern recognition system has excellent performance characteristics, it becomes unusable by simply overwriting the result.

Chapter 3

About deepfakes

Deepfakes are becoming a part of life and are getting more and more convincing. The technology has found applications mainly in the entertainment industry, but it can also find applications in medicine and other fields. Although it is still a relatively new technology its prevalence and the danger of its misuse have already forced some countries and companies to start restricting the technology. Deepfakes are generated using deep learning methods, and their impact may grow in the future.

This section describes deepfakes, how they are generated, their positive uses as well as the threats they pose to society. The first part of the chapter defines the concept of deepfakes and their uses (both positive and negative). Next, the detection and creation of deepfakes is described, and finally, tools for creating deepfakes and datasets are introduced.

3.1 What is a deepfake?

Deepfake is content created by artificial intelligence that is authentic in the eyes of humans. The word deepfake is a combination of the words „deep learning“ and „fake“ and refers primarily to content generated by an artificial neural network, which is a branch of machine learning. Mirsky and Lee [37] define deepfake as „believable media generated by a deep neural network“. They also broadly divide deepfake into three categories: reconstruction, replacement, and synthesis. Deepfake technology uses deep learning algorithms to manipulate or alter existing images, videos or audio recordings.

There are a large number of areas in which deepfakes can create content. There are deepfakes generators for animals, housing, photos of cities, and many other things. The most common form of deepfakes involves the creation and manipulation associated with people, such as face synthesis or face swap. A particularly interesting realm is the generation of deepfake videos which is described in the next chapter. [37]

3.2 Facial manipulation techniques

While “photoshopping” images has long been a mainstay of editing face. The resolution and quality of images produced by neural networks methods have seen great improvement in last years. Before the generative techniques, the number of facial manipulation has been limited due to a lack of advanced editing tools, the need for specialized knowledge, and the laborious and drawn-out process. Nowadays, it’s getting much simpler to digitally create imaginary faces or alter a real person’s face in a video. [59] This section discussed

the approaches to fake image detection as stated by Ruben Tolosana et al. [59]. Facial manipulation techniques can be categorized to 4 main groups by the level of manipulation. These techniques are:

- **Entire face synthesis** - In this category, the objective is to create non-existent realistic faces using GANs or hybrid GANs, which are usually combined with other generative models to improve the training stability. These techniques have received great attention and made great progress since their emergence in 2014. The results of these techniques are on a high level of believability. An example of such a technique can be seen in Figure 3.1.

Figure 3.1: Example of entire face synthesis. Source: thispersondoesnotexist.com

- **Identity swap** this manipulation consists of replacing the face of one person in a video with the face of another person. It is also the most popular face manipulation category nowadays. There are two main approaches: i) using classical computer graphics techniques (FaceSwap) ii) using neural networks. A demonstration of this approach is illustrated in Figure 3.2.

Figure 3.2: Example of face swap. Source: youtu.be/2svOtXaD3gg youtu.be/b5AWhh6MYCg

- **Attribute manipulation** – the main goal of these techniques is modifying some attributes of the face like hairstyle, age, eye color, etc. The majority of those approaches adopt GANs for image-to-image translation. We also know these types of manipulation as face editing or face retouching. Figure 3.3 illustrates a demonstration of this approach.

Figure 3.3: Example of attribute manipulation. Source: ailab.wondershare.com/tools/aging-filter.html

- **Expression swap** known as face reenactment, consists of modifying the facial expression of the person. As in the previous categories, this type of manipulation also use neural networks. This approach is exemplified in Figure 3.4.

Figure 3.4: Example of expression swap. Source: Xpression: Next-Gen Face Swap

3.3 How deepfake are used

Deepfake technology is evolving to such a level that it is difficult for people to distinguish fake videos/images from real ones. This fact is confirmed by Nightingale and Farid in their paper [40]. On the one hand, these technological advances in deepfakes open up new possibilities, especially in the creative sphere. On the other hand, it brings a number of potential threats in the potential case of misuse. In this subsection, examples of positive uses of deepfakes will first be described. In the second part, the threats in case of misuse of the technology will be described.

3.3.1 Positive use of deepfakes

Although there are often negative associations with deepfakes, particularly in terms of their potential to mislead or deceive, these technologies can also have positive uses and can be beneficial to society. Examples can be seen, for example, in medicine, education, but especially in the entertainment industry. The following section describes some positive examples of applications.

Education

Deepfakes can also improve some teaching practices in schools. Especially for teachers, it will give them new ways to present information to students in an interesting way. An example would be the creation of very realistic and engaging historical reconstructions. It also allows creating new content in a relatively cheap and easy way, such as educational videos or educational language recordings [17].

The creative sphere and advertising

Deepfakes can be used to create more compelling and personalized ads. For example, a celebrity who promotes a product or service can be featured in an ad, even if they can't physically appear. A good example is the creation of advertisements, for example, the advertising spot from Slovenska sporitelna. Using deepfake technology from Respeecher, the voice of the famous comedian Julius Satinsky, who died 20 years ago, was brought to life. [55] Another positive use can be the addition of deepfakes to films to cast characters played by deceased actors. Star Wars, for example, is well to use the same technology from Respeecher to create the voice of Darth Vader [25].

Medicine

Deepfakes can also be used to create artificial medical images for testing purposes, for example, to generate MRI images of brain tumors [8]. Synthetic data generation can also solve the privacy problem. The use of deepfakes could also be used in plastic surgeries to visualize the expected result.

They can be used to create synthetic videos or images to help patients better understand their condition and treatment options, for example, by using 3D models of their disease.

Another use may also be in voice synthesis for patients with amyotrophic lateral sclerosis who lose their voice due to loss of muscle control [63].

Gaming industry

The video gaming industry can also benefit from AI-generated graphics and imagery [9]. Video reenactment can significantly reduce the cost of making a game. Deepfakes could be used to sample facial expressions which would save a lot of resources. Deepfakes could also be used to generate facial movements in real-time based on chat generated text.

3.3.2 Risks associated with deepfakes

The availability and ease of use of deepfake technology also create many risks. Deepfakes can significantly threaten our society, political system, or personal security because they threaten national security by spreading propaganda and interfering in elections. They also hamper citizens' trust in information from the authorities and increase the risks associated with the cybersecurity of people and organizations.

Deepfakes pose a greater threat than „traditional" fake news because they are harder to detect and people are more likely to believe such reports. This technology allows the production of seemingly legitimate news videos that can attack various areas [61].

Politics and propaganda

Putting words in someone's mouth in a video that goes viral is a powerful weapon in today's disinformation wars because videos edited in this way can easily change public opinion [61]. Even more so if it is an election campaign or a war where hybrid methods are also used.

A notable instance is the application of deepfakes during electoral campaigns. In 2018, a Belgian political party produced a fake video of Donald Trump's speech in which the US president calls on the country to follow America's example and withdraw from the Paris climate agreement. This video aimed to start a debate on climate change in Belgian society [15].

Another such example is the videos filmed during the conflict in Ukraine. A video of Ukrainian President Zelensky calling on soldiers to lay down their arms has gone viral on Facebook. Viewers quickly pointed out that Zelensky's fake accent was off. Although the deepfake was not of particularly high quality, it is still dangerous because it pollutes the information ecosystem [10].

The latest examples are the highly realistic looking photos of Pope Francis wearing [28] a puffer coat or the photos of Donald Trump being arrested by the police. Of course, this is a deepfake, but these photos went viral in early 2023 within a few hours. Deepfakes were created by the Midjourney model, which was developed by the company of the same name.

False accusation and deepfake porn

In 2019, a deepfake video that depicted popular Malaysian actor Zulkifli „Zul" Ariffin in a pornographic scene went viral. It was not just the dissemination of fake intimate videos, but in Muslim Malaysia, making, possessing and distributing, or even displaying pornography is a criminal offense under Article 292 of the Penal Code [31]. By creating evidence through deepfakes, it would be possible to accuse the victim of various crimes that the victim did not even commit.

In addition to sexual blackmail, deepfake porn can also be a problem. This has been confirmed by the deepfake detection platform Sensity, which in 2019 came up with a report that 96% of deepfakes on the internet are pornographic and 90% of them are women [45].

The use of deepfake pornographic photos in blackmail or as part of revenge could be a big problem in the future.

Further problems with creating false evidence can be the erosion of trust in the courts and justice. It can also set a precedent that the defense can present any electronic evidence (video, image, recording) as false and should be ignored. [46]

Another threat associated with deepfakes is creating deepfake intimate material, also called deepfake porn. Already in 2017, a profile on Reddit posted deepfake pornography of celebrities. To create this, an attacker will overlay a victim's face onto the body of a pornography actor, making it appear that the victim is engaging in the act.

The use of deepfake pornographic photos in blackmail or as part of revenge may be a big problem in the future. And it will be necessary to develop reliable tools to detect such materials [61].

Identity frauds

As Europol has pointed out, another area in which deepfakes can be abused is identity theft fraud. For example, creating fake documents or verifying identities at banks or offices. It is true that passports or ID cards, for example, contain many other security features, but deepfakes can push the boundary to the point where visual checks will not be enough. [5]

Another threat, particularly to the banking sector, can be ghost fraud, where a fraudster steals the identity of a recently deceased person. For example, an attacker may gain access to their current or savings account or apply for a loan. With deepfake technology and the fact that banking has moved online, attackers can create a very convincing illusion that a real, live person is accessing the account, making the fraud much more believable [58].

Deepfakes can be used for malicious purposes specifically in combination with social engineering, for example, to gain unauthorized access to personal data, such as to banks or to communicate with authorities.

While the idea of such fraud may seem like science fiction to some, there was a case in Saudi Arabia in 2020 where deepfake voice generation technology was used to trick a manager of a bank branch who subsequently authorized the transfer of 35 million USD into fraudulent accounts. So far this is a relatively isolated case but it is clear that such frauds will increase [11].

3.4 Restricting deepfakes

The dangers of deepfakes were discussed in Section 3.3.2. There is no general agreement on how to regulate deepfakes. However, we can already see the first signs. In June 2022, Google banned the training of models to generate deepfakes. For example, when attempting to run a DeepFaceLab notebook, the user is warned that access to the platform may be blocked if they continue. On the other hand, some models such as the First Order Model are still usable [62].

China has also moved to restrict deepfakes. This is not a complete ban on deepfakes, but such content will have to be clearly marked as deepfake and the user must be properly authenticated. Users must give consent if their image is to be used in any deep synthesis technology. Also, the possibility of using deepfakes to create disinformation should be restricted. [30]

Restricting deepfakes for the creation of non-consensual deepfake porn is being sought by the UK in the new Online Safety Act 2022. The Act would punish the sharing of deepfake

porn of a person who has not consented to it [27]. The EU will also add to the restrictions by requiring providers to step up the fight against deepfakes and disinformation [18].

3.5 Deepfake detection

Even before the deepfake, recognizing manipulated media was a very necessary task, but with the advent of the deepfake, the requirements for recognizing manipulated content have increased even more, and it is more than certain that this will continue to be the case in the years to come. It is true that nowadays some techniques for detecting manipulated content are still functional, but it is only a matter of time before the artifacts that make neural networks cannot be detected by conventional methods [21]. This section will discuss the theory of detecting deepfakes in images and videos. Some of the techniques are similar to those already introduced in the context of liveness detection.

3.5.1 Fake image detection

This section describes facts taken from the work of Omkar Salpekar [51]. One approach is image preprocessing, e.g., Gaussian blurring, to remove low-level high-frequency traces of images generated using Generative Adversarial Networks (GAN). This increases the pixel-level statistical similarity between real images and fake images and allows the classifier to learn more important and meaningful features that have better generalization ability.

Also, GAN-based deepfake detection can be viewed as a hypothesis testing problem. A minimum distance between the distributions of legitimate images and the images generated by a particular GAN is defined. The results show that this distance increases when the GAN is less accurate, and thus it is easier to detect deepfakes. For cases where the GAN generates high-resolution images, this method is no longer functional.

The latest approach in deepfakes recognition is the so-called 2-phase learning. The first phase uses the Common Fake Feature Net (CFFN), which is trained as a Siamese network constructed from the ResNet18 CNN and trained with the triple loss for the first few epochs to learn the feature-level differences between fake and real images.

In the second phase, a simple classifier CNN is added to the CFFN at the output layer which takes the output from the CFFN and through several convolutional and linear layers produces an output which is a binary classifier. As a whole, this pair of networks is then trained using a cross-entropy loss function.

3.5.2 Fake video detection

For deepfake video detection, it is not possible to use the procedures described in the previous subsection because important artifacts are lost during video compression [64].

However, we can use other detection methods. The most primitive way is to exploit the fact that individual frames have a high degree of correlation between them. In deepfake videos, we can observe a certain degree of inconsistency between the frames [64].

Other methods take advantage of artifacts in the video that are created, for example, when adding a face to an existing background, which creates inconsistent transitions that can be detected. Another possibility is to look for inconsistencies in the movements of, for example, the generated face relative to the head [64].

The most recent approach is the use of Biological-signals-based methods. Although GANs achieve a high degree of realism they still fail to correctly replicate some common

human facial features. It can be said that these methods try to detect what video makes real rather than focusing on the artifacts generated by GANs. These biological features can be, for example, eye blinks which can be detected using the techniques already mentioned in Section 2.4. Another equally interesting approach is to observe the periodic change of the skin due to the pumping of blood through the blood vessels in the skin [6].

3.6 How deepfakes are made

All the information in this section has been taken from the document *A Survey on Deepfake VideoDetection* [64]. Image creation and editing have been around long before deepfakes themselves. The first photo manipulations can be traced back to the 19th century. A well-known photo is that of General Grant at City Point, this photo is made up of three different photos Another example of a photo might be various shots from World War II. For example the famous photo of the liberation of Berlin where the photo of the soldiers has been edited to add smoke over the city. What makes deepfakes revolutionary is the speed and ease with which the fake can be created.

The basis for creating deepfakes is deep learning. To create a deepfake video of a person, the attacker first needs to train a neural network on the real data of the person. The network will create a realistic image of what the target looks like from many angles and in different lighting. Making the deepfake as believable as possible is also an important composition and the overall realism of the setup. According to the objectives, we can divide the generation of deepfakes faces into the following categories: face swapping, face reenactment, text to video synthesis.

3.6.1 Face swapping

The main goal of face swapping is to transfer the face of person A to the face of person B while the facial expressions and emotions of person B are preserved. The first approaches appeared back in 2017 and it was the work of Korshunov et al. [32], who were able to use CNN to capture the appearance of a person and then create high-quality face swapped photos. The approach wasn't suitable for creating higher quality videos because it didn't work with temporal continuity.

In same year, new method was shown. Olszewski et al. [41] was proposed a new approach for generating videos with a single image and a source video sequence. The neural network generated a per frame deformation of the RGB image using the source video. The first face swapped video was created.

After that, many face-swapping frameworks were proposed. The most famous is DeepFaceLab. In recent years, FaceShifter has been proposed. This model performs complex integration of face attributes. As a result, it generates high fidelity swapped faces.

General process of face swap video generation

The deepfake algorithms used in face swapping usually use an autoencoder. An autoencoder consists of two components: an encoder and a decoder. Elements are first extracted from the image by the encoder and then inserted into the decoder to reconstruct the original image. This was described in more detail in Section 3.7.1.

During the training process, two encoders with equal weights are trained to extract common features in the source and target faces. Then, the extracted features are fed into

the two decoders to reconstruct the faces. When the training process is completed, the latent face generated from face A will be passed to decoder B. Decoder B will attempt to reconstruct face B from the feature given face A. In other words, the face generated by decoder B will have the same expression as face A.

3.6.2 Face reenactment

The following section is based on a survey by Nguyen et al. [39]. In contrast to face replacement technologies, facial reconstruction algorithms attempt to change the expressions of people in videos. Attackers can create videos in which they manipulate someone into doing something they didn't actually do.

Already in 2006, we were able to encounter the first facial reconstruction techniques. The basis was a facial template which was then parametrically modified according to facial expressions. This provided a high degree of realism. However, they lacked temporal coherence.

After the development of deep machine learning and the increase in execution came other techniques. Among the most famous ones is Face2Face which reconstructed the facial features of target and source actors using a non-rigid model. This method also brought a new way of generating mouth movements. Where the best frame from the whole sequence was always searched. Face2Face achieved great results compared to previous solutions but was still not realistic enough.

New generative neural models using space-time architectures have been introduced to improve these problems. The main contribution of this new approach was the design of a novel space-time coding as a conditional input for video synthesis, resulting in synthesized videos with a high degree of space-time continuity. Compared to Face2Face, this approach was also able to transmit head position, gaze direction, or blinking thus solving the main problems of the Face2Face algorithm.

General process of face reenactment video generation

As the former, a low-dimensional representation of the source and target video parameters is obtained using the face reenactment method. With this representation, it is possible to transfer the head position and expression into the parameter space.

The scene lighting and identity parameters are then preserved during reconstruction while the head position and facial expression and gaze parameters are changed. Synthetic images of the target actor are then regenerated based on the modified parameters. These images are then fed as input to a video-to-rendering conversion network, which is then trained to convert the synthesized input to realistic output.

3.6.3 Text to video synthesis

In this category, we can include several models such as audio2video, where from the input audio and a short piece of video a phoneme-pose dictionary is first created from this dictionary and the input audio, individual facial expressions are computed from which an output video is then created, which is then combined with the audio. Another slightly more advanced model is text2audio which first creates an audio recording from the text and input data and then uses this to create a video. The models can also synthesize other motions such as different facial expressions in addition to mouth movements. An example

is the work of Zhang et al. which needs about a minute of training data to create a fairly realistic facial animation, which is already comparable to oneshot face swap models. [65]

Another approach can be to use diffusion models to generate videos. This approach does not yet provide simultaneous sound generation, but the results are still interesting. These models preserve the continuity between frames. It consists of 3 parts: the variational autoencoder (VAE), U-Net, and an optional text encoder. Diffusion is performed in several steps, with each step working with the input latent array to produce another latent array that more closely resembles the input text and all visual information. [26]

3.7 More about neural network and generative adverisal networks

The findings described in the following sections are based on the work of Mirsky and Lee [37]. Neural networks are nonlinear models to generate or output or predict based on the input. They consist of layers of neurons, with each layer connected sequentially by weighted connections also called synapses. These weights describe the concepts learned by the model. To perform a network on input x that has n dimensions, the process of forward-propagation is used, wherein x passes through every layer, and a summary of the neuron's output is obtained using an activation function like Sigmoid or ReLU.

To summarise this process, we consider M a black box and denote its execution as M(x)=y. A dataset of paired samples of the form (x_i, y_i) is obtained to train M, and an objective loss function L is defined. The loss function is used to generate a signal at the output of M, which is back-propagated through M to detect the errors of the individual weights. An optimization algorithm, such as gradient descent (GD), is then used to update the weights for several epochs. The function L measures the error between the input x and the predicted output y_0. As a result, the network learns the function $M(x_i) = y_i$ and can be used to make predictions on unseen data.

One approach employed in certain deepfake models is referred to as one-shot or few-shot learning. This technique permits the adaptation of a pre-existing network's output to a novel dataset X^t that is akin to the original dataset X used for training. Two common approaches to this are passing information about $x_0 \in X_0$ to the inner layers of M during the feed-forward process and performing several additional iterations of training on several samples from X_0.

Generative adversarial networks and autoencoders are the most commonly used approaches to creating deepfakes. Both of these approaches involve training a neural network to learn the underlying patterns and properties of a dataset and then using this knowledge to generate new data. This section describes these networks and also how they are trained.

3.7.1 Autoencoders

The facts were taken from Mirsky and Lee [37]. A decoder encoder consists of at least two networks, an encoder and a decoder. A simple diagram of this type of network can be seen in Figure 3.5. The encoder-decoder has narrower layers towards the center, forcing the network to summarize the observed concepts. The summary of the encoder for input is often described as encoding or embedding. The encoder is a feedforward, fully connected neural network that compresses a variable-length sequence into a latent space representation and encodes the input image as a compressed representation in a reduced dimension.

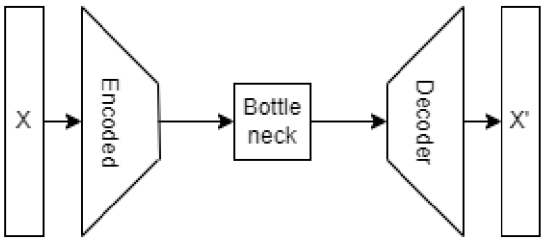

Figure 3.5: The autoencoder architecture.

The decoder is also a forward network like the encoder and has a similar structure to the encoder. It is configured to generate the output sequence from the internal vector representation back to a variable-length sequence. The goal of an autoencoder is to learn a lower-dimensional representation (encoding) for higher-dimensional data, usually to reduce dimensionality, by training the network to capture the most important parts of the input.

3.7.2 Generative adverisal networks

The facts in this section are based on [39]. The objective of the generative adversarial network (GAN) is to create something new based on previous data. This is why GANs are ideal for generating deepfakes. A common GAN model consists of two neural networks: a generator G and a discriminator D. As we can see in Figure 3.6. Given a dataset of real images X. The goal of the generator G is to produce images $G(z)$ similar to the real images X, where z is the signal noise. The key point is that the generator only communicates with the discriminator and cannot access the real data. The goal of the discriminator G is to correctly classify the generated images G and the real images X.

The discriminator D is trained to improve its classification capacity, to maximize $D(X)$, which represents the probability that X is a real image and not a fake image generated by G. On the other hand, G is trained to minimize the probability that its outputs are classified by D as synthetic images, minimizing 1 - D(G(z)). This is a minimax game between two players D and G. If at any time the discriminator cannot notify the distinction between the two generate images and actual images representation is considered as converged.

The training ends when the results are good enough so that generator G is able to produce images that are very similar to real images, while discriminator D is able to distinguish fake images from real images with high accuracy.

3.7.3 Recurrent Neural Networks

An RNN is a form of neural network that is designed to process data that is both sequential and of varying lengths. After computing the input x (i-1), the network can preserve its internal state and leverage it to process the next input x (i) and subsequent ones. In the production of deepfakes, Recurrent Neural Networks (RNNs) are frequently utilized to process audio and, on occasion, video. RephraseLong Short-Term Memory (LSTM) and Gate Recurrent Units (GRU) are more developed versions of RNNs.

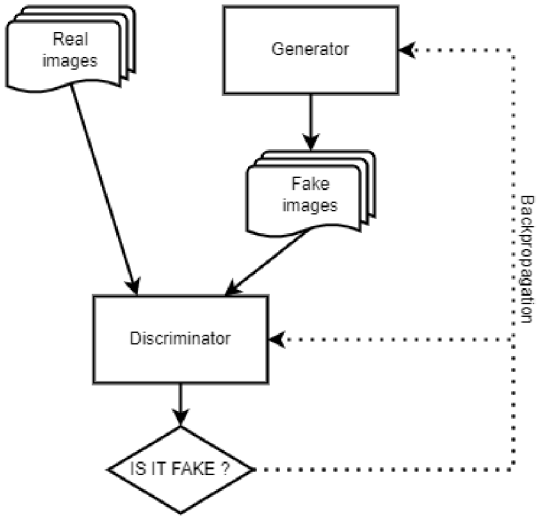

Figure 3.6: The GAN architecture.

3.7.4 Generalization

The discoveries outlined in the subsequent sections stem from the research conducted by Mirsky and Lee as referenced in [37]. The deepfake model can be trained to work only with a specific set of source or target identities. It can be difficult to create and train an identity-independent model because of the correlations that the model learned between s and t during training. We know three primary types in terms of generalization:

- **one-to-one** – The model uses specific identity A as input to create deepfakes of specific identity B.
- **many-to-one** – The model can use any identity to create a deepfake identity B.
- **many-to-many** – Can use any identity to create a deepfake of any identity.

3.8 Avalible tools

This section will describe the tools for creating facial deepfakes. The ones that are available for regular users have been selected. Emphasis will be placed on output quality, orderliness on hardware and time. The quality will be evaluated subjectively, focusing on the number of artifacts and the output image's resolution. The quality of the results and possible problems, such as artifacts or distortions, will also be described. Most of the tools mentioned are freely available on the Internet.

3.8.1 deepswap.ai

deepswap.ai [1] is an online web application for generating deepfakes. After purchasing a subscription, the user has credits for which he can generate deepfake videos. The authors do not give more information about the technology used, except that they use „AI algorithms“. It certainly involves the use of some generative neural network.

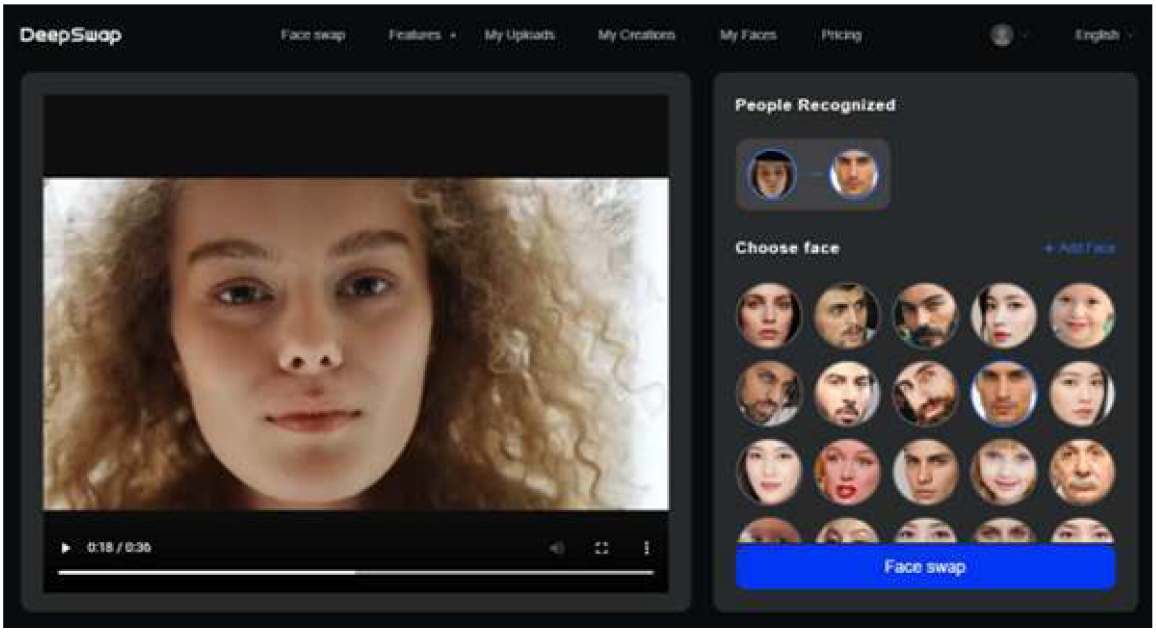

Figure 3.7: User interface of deepswap.ai. Source: https://markets.businessinsider.com/news/stocks/deepswap-ai-launches-next-generation-online-deepfake-software-1032145082

The user can create a deepfake quite easily in a friendly user interface that guides the user on what to do, and the user interface can be seen in Figure 3.7. As a first step, the user must upload a video in which the selected person will be face swapped. The application will analyze this video, which takes a few seconds. In the next step, the user uploads a photo of the person he wants to face swap into the video. Then he has to wait a few minutes for the video to be generated, which will appear on the „My Creation“ page.

Figure 3.8: Example of output from deepswap.ai On left Elon Musk as Joe Biden, right Will Smith as Barack Obama.

One of the advantages of using the application is certainly the simplicity and intuitiveness of the user interface. Even less technically skilled users will be able to use the application. Deepfake can be created in just a few minutes. The user needs only basic knowledge of working with the computer (uploading a video, clicking a few buttons).

As a web application, the user does not need powerful hardware and can generate videos on any device with internet access.

The result is quite impressive quality, as shown in Figure 3.8, especially considering the time and ease of use. The generator managed to transfer most of the attributes from the original video, for example, the winking or the head tilt. However, there is occasional flickering in the areas where the face swapped face blends with the background.

3.8.2 Xpression

Xpression [2] is a virtual camera for mobile phones. Which allows users to instantly transform into anyone or anything with a face, using a single photo and in realtime. The app is freely available in the app store and play store. The user interface is intuitive and simple, allowing instant use without any setup.

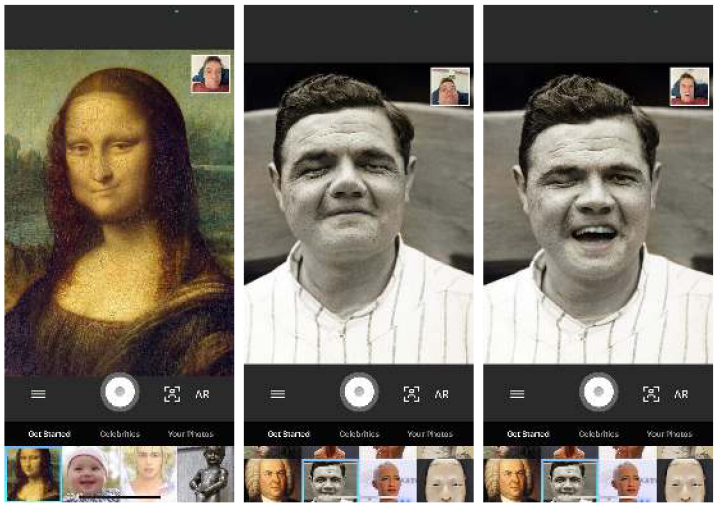

Figure 3.9: Examples of outputs from Xpression App.

The advantage is the real-time processing and intuitiveness of use. It also provides video reenactment, but the results looked rather unsatisfactory due to the number of artifacts the system generates. The output of this application can be seen in Figure 3.9.

3.8.3 First order motion

The first-order motion model is a framework created by Aliaksandr Siarohin [54]. The model does not need any annotation of the animations in the control video or any prior information about the particular animated object. It was trained on videos of the same categories (faces, body movements). Its advantage is that it can animate any object according to the control video using a representation consisting of a set of learned key points together with their local affine transformations. The generator network models combine the appearance obtained from the source image and motion derived from the control video. The implementation contains several models specially trained for different categories. The whole code is published on GitHub and anyone can try it on Google Colab.

The startup itself is not difficult, although it is not completely intuitive. After loading the project, the user is presented with a dialog box for uploading the source data, which is the control video and a photo of the person to be reenacted.

Figure 3.10: Examples of outputs from First Order Motion. [54]

The results are quite good, but there are some problems, for example with head movements with quite significant deformations of the face. The disadvantage may be the hardware complexity, which can be quite easily bypassed with Google Colab[1], although for longer videos the memory may not be enough even here. Another problem may be the resolution which is only 256x256 pixels.

3.8.4 GHOST

Alexander Groshev et al. [23] proposed a new solution for image-to-image and image-to-video face swapping. They called it GHOST (Generative High-fidelity One Shot Transfer).

This solution took as baseline FaceShifter (image-to-image) architecture to which they added a couple of enhancements which include a new eye-based loss function, face mask smooth algorithm, a new face swap pipeline for image-to-video face transfer, a new stabilization technique to reduce face jittering on adjacent frames. They have also added a super-resolution stage which allows the solution to generate high-quality images and videos. They also added an attribute encode that extracts more face attributes (such as pose, expression, and image color) from the face, which adds more detailed information to the generator for the output image/video. The whole code is published on GitHub and anyone can try it on Google Colab[2].

Launching and creating is not extremely difficult as in the previous tool, but the user has to solve a few minor problems with libraries and file paths. Otherwise, the creation of the deepfake itself is again a matter of a few dozen seconds.

The results of this tool can be seen in Figure 3.11. Another positive from an attacker's perspective is that the network can generate data in Full HD. The improvement in the quality of eye movement is also noticeable. Occasionally there is a slight flickering in the result. Still, it is not as pronounced as in other presented devices, which is probably due to the stabilization and blending presented in the paper.

[1] https://colab.research.google.com/github/AliaksandrSiarohin/first-order-model/blob/master/demo.ipynb

[2] https://colab.research.google.com/drive/1B-2JoRxZZwrY2eK_E7TB5VYcae3EjQ1f

Figure 3.11: Examples of outputs from GHOST. [23]

3.9 Datasets

This subsection will describe reference datasets for working with deepfakes. These datasets are mainly intended for training deepfakes recognition systems. No dataset has been standardized to evaluate the robustness of individual face recognition systems against deepfakes. However, they usually do not contain enough data to compute impostor and genuine scores.

The datasets described in this section can be divided into two subcategories according to the way the media is generated, face replacement datasets and face reconstruction datasets.

3.9.1 Face swapped dataset

The first category of datasets are datasets created by the face swapped method. This method consists of taking person A's face and pasting it on the face of person B in the target video. Section 3.6.1 discussed this method in more detail.

- **Celeb-DF** - This dataset was selected for experiments and is described in greater detail in Section 3.9.3.
- **DeepFakesDataset** - The videos in this dataset are diverse real-world samples in terms of resolution, compression, lighting, aspect ratio, frame rate, motion, pose, cosmetics, occlusion, content, and context, up to 142 videos, 32 minutes, and 17 GB in total. Synthetic videos are compared to their original counterparts where possible. The dataset is publicly available for academic purposes. [20]

3.9.2 Facial Reenactment

The second part of the dataset media was created using face reenactment. In this method, the attributes of the facial expression of person A are computed, and these are transferred to the facial expression of person B. This method has been described in more detail in Section 3.6.2.

- **FaceForensics** - The dataset contains 977 downloaded YouTube videos, 1000 original extracted sequences that contain an unmasked face that can be easily tracked, as well as their manipulated versions using four methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures.[50]

- **KoDF** is a large dataset of synthesized and real-world videos targeting Korean subjects, which is used for the deepfake detection problem. The dataset includes 62,166 real videos and 175,776 fake videos from 403 subjects. The fake videos are created using 6 different techniques: FaceSwap, DeepFaceLab, FSGAN, FOMM, ATFHP, and Wav2Lip. [33]

3.9.3 Celeb-DF

This subsection is based on a paper that was published with the dataset [35]. The Celebrity Deepfake (Celeb-DF) dataset is a large-scale challenging dataset for deepfake forensics. It consists of 590 real videos and 5639 deepfake videos (containing around 2 million frames). The real videos are collected from Youtube. The average video length is 13 seconds with a framerate of 30 frames per second. The dataset consists of 59 identities of real persons, of which 56.8% are male and 43.2% are female. In addition, the real videos exhibit a large range of changes in aspects such as the subjects' face sizes (in pixels), orientations, lighting conditions, and backgrounds

Figure 3.12: Examples of deepfakes from Celeb-DF dataset. Source:[35]

Deepfake videos were generated by swapping faces between the 59 identities. These faces are at a resolution of 256x256 pixels. The creators of the dataset focused specifically on removing free artifacts such as flickering at the edges of the face swapped face or skin color inconsistencies arising at the transition of the face swapped face and the background. Examples of these deepfakes from the dataset can be seen in Figure 3.12.

We decided to use this dataset in our work as it achieves quite good quality deepfakes and simultaneously contains several original videos of persons, so it is possible to count all the necessary scores from the dataset.

Chapter 4

Experiment design

In Chapter 2, face recognition technology was introduced. Then, in Chapter 3, deepfake technology and the threats it poses were introduced. This chapter will evaluate the different use cases where facial recognition is used. These use cases will be categorized according to how easy it would be to launch an attack and whether it is worthwhile. Also, the model of the attacker will be presented, and his abilities will be evaluated. Next, the unification of the problem of multiple use cases into a single application will be presented. In the last section, individual attacks will be described.

4.1 Existing works

In this work, we focus on the areas of deepfake creation as well as their detection using common face recognition tools. Previous works, such as the work of Shahroz Tariq et al. [57] have attacked commercial recognition APIs from Microsoft and Amazon. These APIs have been shown to be deceived by deepfakes in up to 78% of cases. The experiments showed that some deepfake generation methods pose a greater threat to recognition systems than others and that each system responds differently to attacks.

Further work by P. Korshunov et al. [32] described that the VGG and Facenet models are not able to distinguish GAN-generated effectively and face swapped faces from real ones.

Other work by Changjiang Li [34], which looked at deepfakes attacks, pointed out that facial authentication systems are biased against white males. Subsequently, it was found that female and non-white identities are more effective at bypassing authentication systems.

This master thesis differs from the other's work because it attacks real deployed systems for identity verification. It also aims to find out whether and how difficult it is to perform such an attack on common biometric facial recognition systems.

4.2 Suitable use cases

In Section 2.5, we have given many examples where we can use face recognition. We have also divided these systems into three categories, as seen in Figure 4.1. The first category is the use cases that do not allow the exploitation of deepfakes in attacking the given systems. In this category, we can include systems whose design does not allow the use of 2D deepfakes, an example being 3D facial recognition systems. Another aspect that makes their use impossible or significantly more difficult is the presence of security forces or

security services that would quickly detect such actions. Although airport security checks are often mentioned in facial recognition, the use of deepfakes is very difficult and unlikely. This is certainly true in general for all cases where there is some sort of control present that might notice suspicious activity during facial recognition.

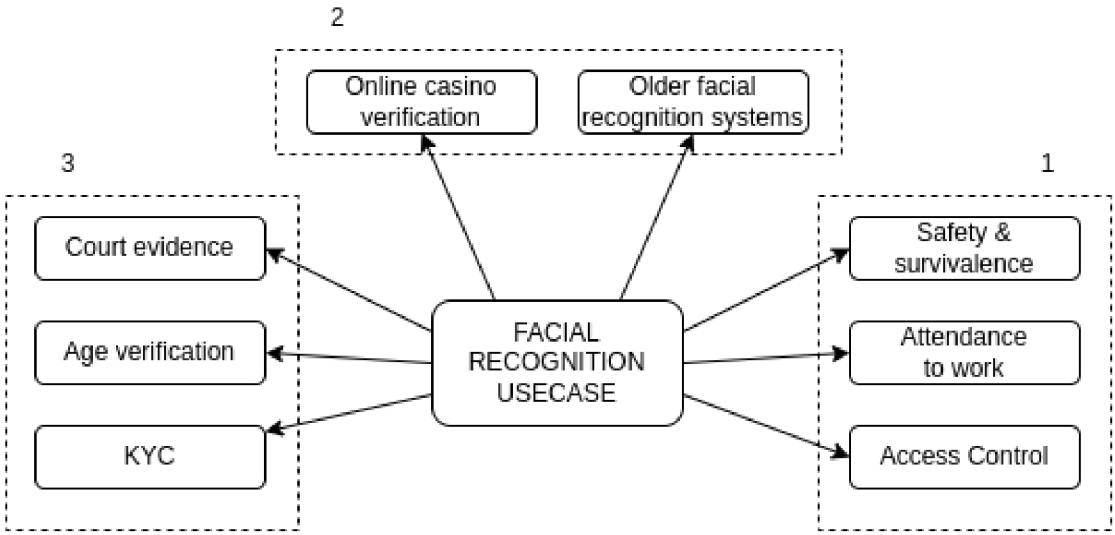

Figure 4.1: Examples of the use of facial recognition are divided into 3 categories: 1. It is difficult to attack them with deepfakes. 2. It is not worth it to use deepfakes to attack. 3. Appropriate use of deepfakes when attacking.

In addition, there are cases where it does not make sense to attack a given system using deepfakes. These are use cases where only a single face photo is required or outdated facial recognition that can be fooled by a presentation attack. As an example, consider a betting company where a user has to upload a photo of his ID and a photo of his face to verify his account when registering. For such a system, there is no need to create a deepfake and a suitable photo of the victim is sufficient.

Examples of use cases that are suitable to attack with deepfakes are those where there is no one to control the recognition process. At the same time, the verification input must be more complex than a single frame. Such inputs are various KYC systems, such as online banking, where an attacker could get access to the funds in the account after a successful attack. It can also be media creation for false accusations. In these examples, the attacker has the advantage of being able to prepare the means of attack without attracting attention. He can also have multiple attempts at a successful attack. These use cases have an additional advantage in the way the attacker inserts data into the system:

- The attacker will play a deepfake to the sensor (camera). This is a technically less complex attack that will be exploited by a naive attacker.
- The attacker will bypass the sensor and insert the data directly into the feature extractor. This attack is already technically more difficult because it assumes that the attacker has the knowledge to hijack the camera and insert his own video into the system

4.3 Attacker model

We assume that an attacker is a person or organization whose goal is to get into a system secured by a facial recognition system. We also assume that the attacker has not had the opportunity to influence the development of such a system (e.g., by modifying the training dataset or otherwise changing the algorithm). However, the attacker can influence the inputs to the system that will be evaluated (e.g., a presentation attack). It is also likely that the attacker is able to obtain materials to create deepfakes from freely available sources such as social network profiles, etc. Or he may have accessed these materials by some other unspecified means. We also assume that the attacker has access to tools for creating deep fakes, whether freely available or paid. We further assume that the attacker will use all of his knowledge and effort to get into the system in question.

However, we must further divide the attacker into two examples based on his knowledge:

4.3.1 Naive attacker

The first type of attacker is the so-called naive attacker. This attacker has no deeper knowledge of the system or the creation of deepfakes. Suppose such an attacker tries to insert a deepfake into the system that he creates in online, without caring about the overall quality (poor image quality, many artifacts, poorly chosen background for the face swap, etc.). Such an attacker also has almost no knowledge of the system he is attacking. And tries to insert inputs into the system to break in. Thus, we can say that he treats the recognition system as a black box.

4.3.2 An attacker with knowledge

The second type of attacker is someone who has much more knowledge than the example of the other attacker. This person understands how facial recognition works. Also, assume that this person has done research on how the system responds to various inputs, for example, by running and testing a demo application (or the whole system). She has also probably researched the available tools and will use the ones that give the best results. When creating a deepfake, this type of attacker will be careful about various artifacts generated from the neural network as well as the correct choice of inputs (e.g., for a face swap, it will choose as an actor a person who has the same hair color or body build as the victim it is targeting).

4.4 Unification of the problem

As mentioned in Section 2.2, in face recognition, key points are found on the face and then a vector is computed from the distance between these points. This distance vector computed from the input photograph of the face is then compared to the pattern vector stored in the system. Furthermore, use cases described in Section 4.2 can be simplified into a problem in which the vector computed from the input photograph is compared to the vector stored in the system. For this reason, only this comparison is tested as it is the essence of all the identified usecases allowing deepfake spoofing attacks.

The knowledge from Section 2.3 will be used for an experiment that will simulate video as input. The video can be viewed as a set of frames, where each individual frame votes for or against the identity. The continuity of the image inputs is not considered because the biometric systems themselves do not allow it.

4.5 Selecting photos for a dataset

Since creating a new deepfake dataset would be too time and resource-consuming, I decided to use the existing Celeb-DF deepfake dataset. The advantage of the Celeb-DF dataset is that it provides multiple videos of a real person.

Since the dataset contains common video recordings of interviews downloaded from the Internet, the person in the video is not always looking directly into the camera. It is also quite often the case that the person is not properly illuminated or has their eyes closed or does not have a neutral facial expression. Examples of these inappropriate shots can be seen in Figure 4.2.

Figure 4.2: Examples of lower quality images where people are not looking directly into the camera or have facial expressions that do not match ICAO requirements.

When creating the dataset it will be necessary to select images that are at least close to ICAO standards, we have described this standard in Section 2.3.1. It will also be necessary to select images that are spaced some time apart to avoid identical or very similar images in the sequence.

4.6 Design of experiments

In the previous chapters, the topic of facial recognition and its wide use in today's world, whether it is for banks, insurance, or logging into other applications and services, has been discussed. However, with the advent of deepfakes and the significant shift in their creation come to the risks associated with their misuse in automatic facial recognition. The question, therefore, arises as to how robust these biometric facial recognition systems are against these types of attacks.

Therefore, in the first experiment, we will focus on the technical feasibility of such an attack using the available resources, be it the commercially available ones or open-source solutions that an attacker has the possibility to access. It will also describe the necessary conditions that an attacker must satisfy in order to trick a face recognition system.

The second experiment will investigate the feasibility of an attack against specific facial recognition solutions. The experiment will incorporate the findings from the previous sections. The last experiment we have designed will look at simulating an attack where biometrics requires multiple images as input for identity verification.

4.6.1 Technical feasibility of deepfake attack

This experiment aims to investigate the technical difficulty of creating a deepfake. To create a deepfake we will use the tools described in Section 3.8, these are *deepswap.ai, First order*

motion, GHOST. All these tools are available online and do not require any additional hardware. All of these systems are one-shot, which speeds up the creation of deepfakes considerably. And the last two mentioned are freely available as demos for scientific papers.

The next part of the experiment will be to get acquainted with commercial solutions of biometric systems for face recognition: IFace from Innovatrics, and Megametcher from Neurotechnology. These biometrics were chosen because they are relatively widespread solutions for facial recognition. The third biometric is a solution from Regula in the form of a demo application on their website. Also, the goal will be to find a suitable solution for dataset evaluation.

The goal of this section will be to see how easy it is to use these tools. This experiment should answer the following research questions:

1. How difficult is it to create a video deepfake?
2. How difficult is it to create a video deepfake that can fool a facial recognition system?

After familiarizing ourselves with the tools, we will try to create a few deepfakes as an experiment. We will then embed these deepfakes into selected biometric face recognition systems, and obtain scores from them.

In case the deepfake does not achieve the required score we will try to develop a procedure to improve these deepfakes. The goal of this experiment will be to create a deepfake that breaks such a system. A simple sketch of the experiment can be seen in Figure 4.3. Another benefit of this experiment will be the findings that can be used in future experiments or papers.

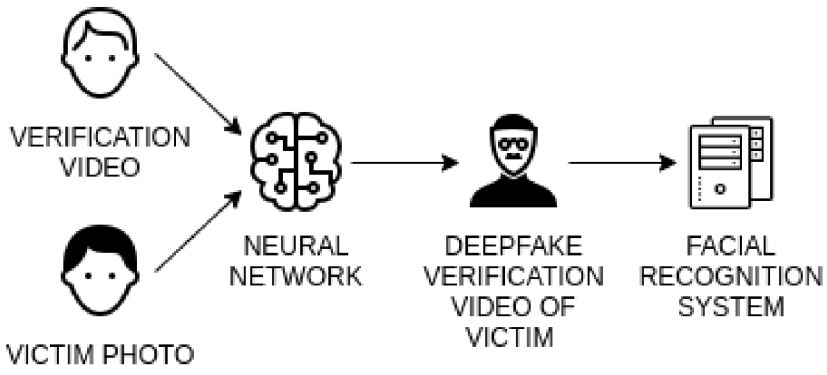

Figure 4.3: Chart of the attack, first a photo of the victim will be obtained and a driving video created. Then a deepfake will be created using the selected models. This deepfake will finally be inserted into the facial recognition system.

4.6.2 Comparing two face images

From the previous experiment, we know that using deepfakes it is possible to make a facial recognition system recognize an attacker as a victim. We have also found that such an attack is technically feasible and the attacker does not even need many resources for it. Therefore, the goal of the second experiment is to determine how susceptible biometric systems are to being fooled by deepfakes. As in the previous experiment, the selected

biometrics will again be Megamatcher or Neurotechnology and Innovatrics' IFace. First, we will need to sample the videos from the Celeb-DF dataset, this procedure was described in Section 4.5. Images from this dataset will be fed into the biometrics. For each person of the 58 people in the dataset, 3000 comparisons will be made to calculate individual scores. The score of each person will be calculated and then for the whole dataset. The dataset can be divided into three parts to calculate the individual scores:

- Genuine score - Two real photos of the person will be inserted into the system from which the output will be calculated.
- Impostor score -To get a score, two photos will be entered into the system with different people in them.
- Deepfake score - A real photo of the person and a deepfake of the person will be entered into the system to get the score.

Finally, the resulting scores will be evaluated and the resulting statistical similarity will be calculated between each part. The experiment should further answer the following questions:

- How easy is it to fool selected facial recognition systems using deepfake deception?
- Is the difference in obtained matching scores statistically significant?

4.6.3 Comparing a sequence of frames with an image

The final design of the experiment will investigate whether the face recognition system becomes more robust against deepfakes if a sequence of frames is selected from a video instead of a single photo. This assumption is based on the fact that in deepfake videos there are various artifacts that may or may not appear in all frames.

The goal is to verify if this makes the face recognition system more robust to deepfakes when processing a sequence of frames. The assumption in this experiment is that if the biometric system has to process more frames from the video, the probability of processing frames that are of lower quality increases, and thus the overall score decreases. The selected biometric systems are Megamatcher or Neurotechnology and IFace from Innovatrics as in previous experiments. However, since these systems do not allow video processing, a sequence of photos will be fed into the system during the experiment. One photo of a person will be inserted each time as input and this will be compared with the sequence. The video sequence will have 5 frames that are at least one second of recording apart so that small facial shifts may occur the deepfake generator may not capture that and thus may generate artifacts. For all 58 identities, 100 sequence comparisons will be performed. The resulting score will be computed as the average of the individual scores from a given sequence, where each photo will affect the result with its score. The method of averaging the resulting frames to evaluate the video has been described in Section 2.3. The experiment should give us an answer to the following questions.

- Will the system improve its robustness to deepfakes if it processes a sequence of photos instead of a single photo?

4.7 Evaluation of experiments

To find out more about the measured results we will use several metrics. The first is the true-match-cheater-match graph is a graphical representation of the distribution of the different early classes and shows us how well the system can distinguish between these classes, an example can be seen in Figure 4.4 on the left. On the right side, we can see two individual measures:

- The False Non-Match Rate (FNMR) measures the system's susceptibility to misidentification or misauthorizations of the actual user. The FNMR is expressed as the percentage of times an input or authentication attempt is made where the user's face is incorrectly rejected (false negative) because the similarity score is below a prescribed threshold [53].
- The False Match Rate (FMR) is a measure of the susceptibility of a system to misidentification or authorization of an unauthorized user. It is measured by the number of false positive identifications or authorizations divided by the total number of identification attempts [53].

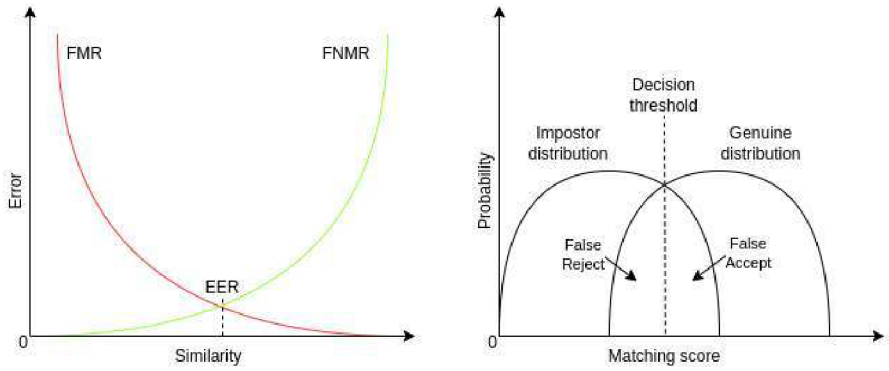

Figure 4.4: FMR, FNMR and EER on the left and Matching scores distribution graph on the right.

In Figure 4.4 we can also see the EER value, where the point on the FMR curve equals the FNMR. This point indicates the threshold value at which the false acceptance and rejection rates are equal. A device with lower EER is regarded to be more accurate [53].

To evaluate the statistical difference between deepfake, impostor, and genuine scores. We chose to use the student's t-test for statistical hypotheses. This test is used to determine whether two subsets of data are significantly different from each other. Before evaluation, it is necessary to satisfy the condition that the data have a normal distribution. Two types of t-tests are known: the independent samples t-test and the paired samples t-test.

The first mentioned t-test is used when the groups being compared are independent of each other. This means that the data in one group of samples have no effect on the data in the other group. An example might be two classes of pupils where the observed fact will be a 100-meter race, the results of each class being independent of the other. The paired t-test is appropriate to use when we have an educational program and we do a test on the same sample of students first before and then after taking this course.

The paired t-test is used when the two groups being compared are not mutually independent, which means that there is some relationship between the data in one group and the data in the other group.

The result of this test is a t-value which tells us the degree of difference between the two means as well as a p-value which tells us the probability of observation between the two means. If the p-value is less than a pre-established significance level then it is judged that there is a significant difference between the two groups.

Chapter 5

Realisation of experiments

This chapter describes the execution of the individual experiments together with their evaluation as described in Chapter 4. The results of each experiment will also be described here. The individual experiments are described in the order they were described in the previous chapter because of their logical continuity.

5.1 Tested biometric facial recognition systems

This section describes how to work with individual biometric face recognition systems. These biometrics will then be used in the following experiments.

5.1.1 Megamatcher

MegaMatcher [38] is designed for developers of large-scale AFIS and multi-biometric systems, available as a software development kit that enables the development of large-scale products for the identification of one or more biometric fingerprint, iris, face, voice, or palmprint for Microsoft Windows, Linux, macOS, iOS, and Android platforms. This technology ensures high reliability and speed of biometric identification even when using large databases.

For input photos, it is required that the face in the photo has a distance at least 64 pixels between the eyes. Also, the recognition engine requires the face rotation to be a maximum of 180 degrees, and the head tilt to be up to 15 degrees (a maximum of 25 degrees if multiple views of the same face covering different nod angles were used during registration). Head rotation in the horizontal axis should not exceed 90 degrees.

The Megamatcher tool is available on the Internet as a demo application, where the user, after registering, receives login data to the system, which can be used for further work. In addition to face matching tools, fingerprint or iris matching tools are also available.

The menu itself is quite intuitive as we can see the application in Figure 5.1. First of all, it is necessary to upload the profile of the person to be compared. This is done via the „Enroll“ tab where we can insert the already mentioned additional biometric data. After assigning the identifier, the system will tell us that a new profile has been registered. Next, for verification, we can proceed to the „Identify“ tab, where the user uploads the photo he wants to compare with those stored in the system. After the evaluation, the user receives a notification that will take him to the evaluation itself.

As a next step, we've embedded several pairs of real photos of people into Megamatcher. First to check how the system works and also to get an estimate of the first results, as seen

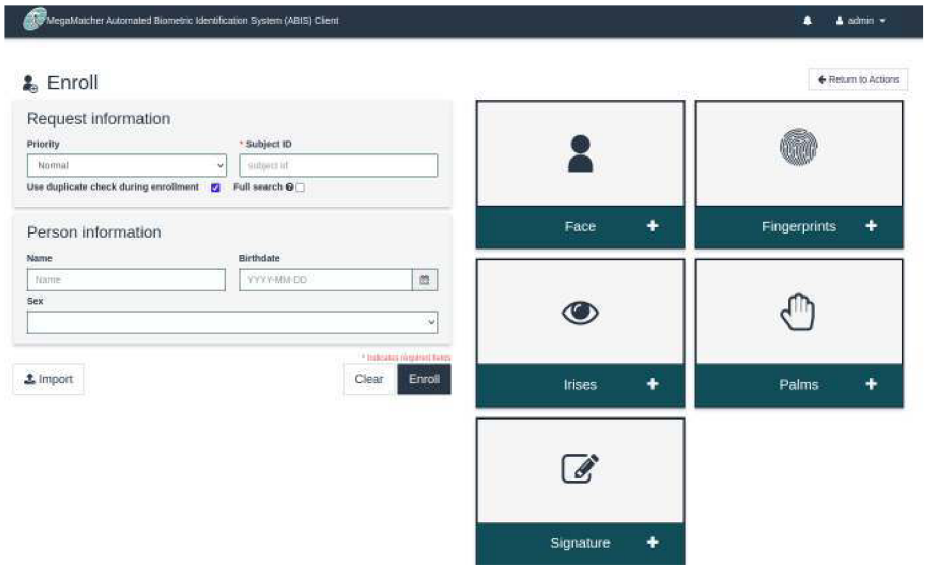

Figure 5.1: Megamatcher user interface

Attempt	Score achieved
1. attempt	101
2. attempt	116
3. attempt	91
4. attempt	114
5. attempt	110

Table 5.1: The score achieved when comparing the same person in Megamatcher. From the information provided by the manufacturer, it is not possible to interpret exactly what the score means.

in Table 5.1. At this stage, we have not yet inserted deepfake data and just wanted to verify the functionality of the system.

5.1.2 IFace

IFace SDK 3.0 [56] is a facial biometric technology. Features include real-time identification and authentication (1:1 matching), multi-face tracking, and person analytics, including age and gender profiling. The technology is based on deep neural networks and provides verification capabilities, both from a still image and from video footage in all standard formats.

For input photos, it is required that the face in the photo has a distance at least 120 pixels between the eyes and a minimum resolution 600x600 px. The image should not be compressed or resized. Also, it should not have too strong a backlight or sidelight. Also it should not be overexposed or underexposed. The system also provides two types of active liveness:

- Smile liveness - challenges the user to smile
- Eye-gaze liveness - challenges the user to follow a dot on the screen with their eyes

We used the IFace 3.6 as a terminal application that was part of the code included with the product. This application took two parameters which were input images, to be compared. The application can run in two modes: accurate mode and fast mode. The system returns two types of scores according to the manufacturer's documentation they are:

- **fast** - some partially covered faces or faces with sunglasses may be overlooked. Also, faces printed on ID cards may not be recognized. However, the speed performance of face detection is much better than using other modes.
- **accurate** - partially obscured, blurred profile faces or faces with sunglasses are detected. The CPU speed of this face detection is slower compared to the other modes.

As with the previous biometric system, we decided to obtain the early images from several pairs of images. The results can be seen in Table 5.2.

Attempt	**Fast mode score** [%]	**Accurate mode score** [%]
1. attempt	89.677	75.317
2. attempt	88.508	81.547
3. attempt	85.731	79.887
4. attempt	87.816	79.179
5. attempt	83.211	68.436

Table 5.2: The score achieved when comparing the same person in IFace. The resulting score is in percentage.

5.1.3 Regula

The Regula Face SDK is a multi-platform biometric authentication solution that confirms a person's identity using a comprehensive set of technologies. The system is also available as a web demo where 2 photos can be uploaded, and it will be verified whether it is the same person or not.

Apart from general conditions such as good lighting and focus, we were unable to ascertain the exact requirements of the system. Regula provides an online demo of the Regula Face SDK Web API directly on their website, which returns a matching score after uploading 2 photos. It is not written whether the demo itself differs in performance from the delivered product.

The Regula is available online as a demo directly on the manufacturer's website. Among other demonstrations of the use of their products, we can find here also a comparison of the face. Two photos are inserted into the application via the browser, and the application starts the comparison itself. On the right, an evaluation is then displayed with the scores achieved with a color indication of whether the person has been accepted or not. As this is a demo on top of being available as a web and it is relatively easy to fool, this biometrics is only included in the first test.

As with the other biometrics, we decided to insert several pairs of images of the same persons into the system.

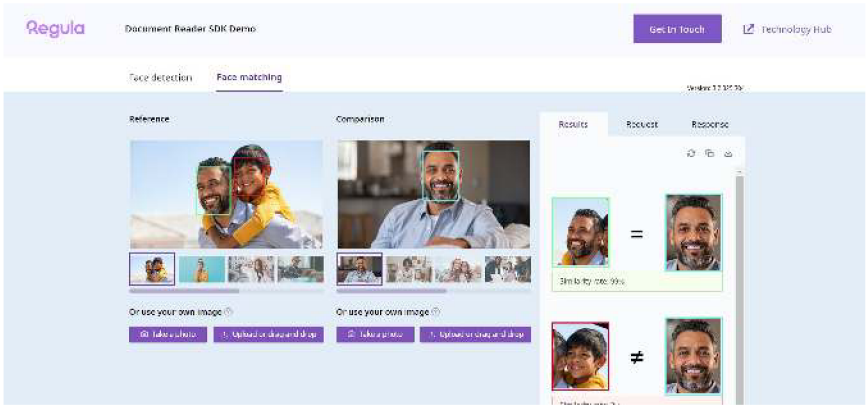

Figure 5.2: Preview of the Regula demo application.

Attempt	Score achieved [%]
1. attempt	99
2. attempt	99
3. attempt	98
4. attempt	99
5. attempt	99

Table 5.3: The score achieved when comparing the same person in Regula. The resulting score is in percentage.

5.2 Experiment 1:Exploring basic concepts

The design of the experiment was described in Section 4.6.1, The aim of the experiment is to verify the technical feasibility of the attack on facial recognition systems. The tools to be tested will be Megamatcher, IFace, and Regula, which were described in Section 5.1. Finally, the findings of the experiment will be described, as well as the answers to the questions that were posed during its design.

5.2.1 GHOST

The tool as such has been described in Section 3.8.4. This tool is available as a demo to the paper on Google Colab and is freely accessible to everyone. Since it is a one-shot system, after running all the blocks, the user has to upload a background video and a single photo to be face swapped into the system.

The advantage is that since it is a face swaped system you can upload any video of a person where the face is visible. Which expands the potential uses. The tool also allows to generate face swaped photos.

It took us about 10 minutes to get acquainted with the system as such. After running all the code blocks, we uploaded the required video and victim image to the Google Colab notebook and started the generation. With a video length of 9 seconds with a framerate of 30 frames per second, the generation itself took about 40 seconds. After the inference is finished, the user has to run the video preview block where the result is shown to him.

The resulting video quality is quite good, considering the minimal amount of artifacts in the image and the faithful blendig of colors with the background. Especially considering

the fact that it is a one-shot system and thus no long training was needed. Also, the connection of the face to the background is quite smooth and there are usually no artifacts in the image. The system can also handle more dynamic scenes. Or slight tilts of the face.

Figure 5.3: From right to left: good result, hand movement in front of the face, winking.

On the other hand, the system has a problem with synthesis when an object such as a hand passes in front of the face swapped face, as we can see in Figure 5.3. At that moment, the whole texture of the face with face swapping is deformed. This system also can't cope with a head rotation of more than about 60 degrees. Once this angle is exceeded, the system loses all nodal points on the shape and the texture of the face swapped face remains just somewhere in the image.

Evaluation of model outputs in individual biometrics

In the next step, we tried to embed the generated images into the biometrics to verify their quality. We always compared a pair of photos of a real person vs a deepfake person. We can see the comparison results for different face recognition systems in Table 5.4.

Although GHOST suffers from some problems, it is the highest quality tool described in this work as shown by the relatively high biometric scores described in the table. Regula's face matching system scored the deepfake as a given person in all tests although the quality of this system is quite questionable as we will see with other deepfake generating tools. In the case of Megamatcher, surprisingly the system managed to outperform three out of five attempts, and the higher two also scored quite high. In the case of IFace, the face was not detected once at all, and the values are just below those we measured in the first part of the experiment.

Sample	**Megamatcher**	**IFace fast/acc [%]**	**Regula [%]**
Attempt 1	87	57.912 / 68.168	99
Attempt 2	103	71.647 / 79.050	99
Attempt 3	104	No faces detected	99
Attempt 4	110	83.921 / 68.992	98
Attempt 5	85	70.205 / 68.654	99

Table 5.4: Scores achieved by deepfakes generated by GHOST over individual biometrics.

5.2.2 First order motion

First Order Motion was introduced in Section 3.8.3. Like GHOST, the creators of First Order Motion have also made their work fully available as a demo on Google Colab. The system requires a photo and a guiding video as input. In this case, the model does not transfer the face from the photo to the video as it did in GHOST, but based on the common key points from the video and the photo, it moves the photo. This approach allows generating not only deepfakes of faces but also of people's movements from one photo.

Evaluation

Creating deepfakes is again quite simple and it takes no more than 10 minutes to start creating. The problem may be the outdated versions, but this problem can be solved with basic Python experience. In the footer of this page, I attach a corrected version of this demo[1]. It took us about 10 minutes to load the model and run it. Creating a new video with generation took us about a minute. The video is again 9 seconds with a framerate of 30 frames per second.

Figure 5.4: Example of quite good output from the model.

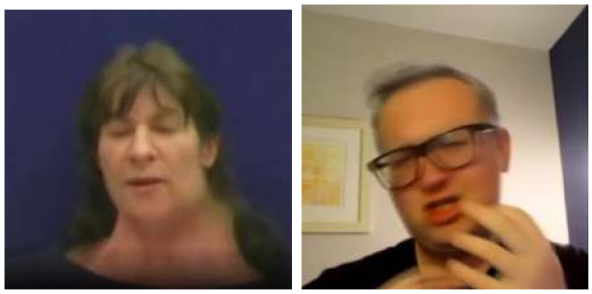

Figure 5.5: Artifacts caused by movement of the hand in front of the face.

As we can see in Figure 5.4, the model performs reasonably well if the video from which the motion is generated does not have significant transformations of the key points compared to the key points in the image. On the other hand, the system cannot cope absolutely in cases when for example a hand passes in front of the face, then the synthesis fails completely. A similar problem is when the head is turned and some key points start to disappear (Figure 5.5), this is manifested by a strange stretching of the texture of the

[1] https://colab.research.google.com/drive/18DMmsQeBzlasjM_GC-BmWBmArYsgtQHU

Figure 5.6: Example when the system couldn't cope with head rotation in the driving video.

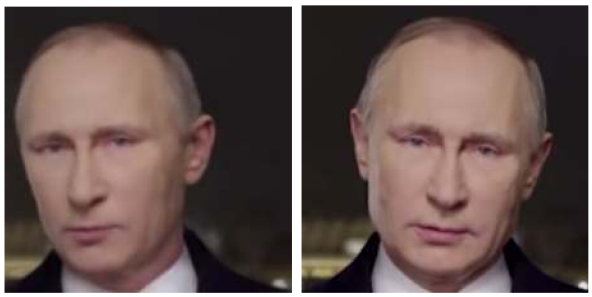

Figure 5.7: The first photo entry that was able to fool the Megamatcher, was created using First Order Motion. The second picture input was inserted into the model as a base from which the deepfake video was created.

main image, which can be seen in Figure 5.6 . The disadvantage of this tool is also the relatively low resolution of the output images/video, which may impair the evaluation of the deepfake as a true one.

In Table 5.5 the results of this tool are visible where the low quality of the output video/images is also visible. IFace evaluated all deepfakes as fake and even in 2 cases it was not even able to identify the face in the image. This is also true for Megamatcher which did not match the face in two inputs. Interestingly though, Megamatcher was once fooled by one input the one we can see in Figure 5.7. This is probably due to an image transformation during generation which was not so pronounced as to distort the image significantly.

Sample	**Megamatcher**	**IFace fast/acc [%]**	**Regula [%]**
Attempt 1	102	NM	99
Attempt 2	86	NM	98
Attempt 3	NM	39.323 / 62.692	92
Attempt 4	91	70.710 / 74.395	98
Attempt 5	NM	43.294 / 61.381	78

Table 5.5: Output of individual biometrics for First Order Motion. As we can see in the first two attempts IFace and Megamatcher were not even able to identify the face (NM label), which suggests low quality deepfakes.

5.2.3 deepswap.ai

We tested deepswap.ai as a paid service on their site. The disadvantage is that there are only a limited number of tokens available for which it is possible to generate deepfakes. This is also the main disadvantage of this tool, the tool itself is more suitable as an entertainment app and not a tool that would be used by a potential attacker.

The generation itself is quite simple, first, you need to upload a photo of the face to be face swapped and then the video itself. After the generation is complete, the user is notified that the result is ready. And this result is displayed on the next page. The overall quality of the deepfakes is good, although there are occasional artifacts in the video. But what we can notice in Table 5.6 these deepfakes don't actually get through any biometrics. Even in the case of Megamatcher, it was not even possible to match the face from which the deepfake was created. The exception is the tool from Regula, which can't handle even poorer quality deepfakes.

Table 5.6: Scores achieved by deepfakes generated by Deepswap.ai over individual biometrics.

Sample	**Megamatcher**	**IFace fast/acc [%]**	**Regula [%]**
Attempt 1	92	62.936 / 71.240	99
Attempt 2	Not matched	30.527 / 34.562	91

5.2.4 Conclusion of the experiment

The experiment discussed the possibility of fooling facial recognition systems using deepfakes. As shown in the experiment, this attack is technically feasible and by using tools for generating deepfakes it is possible to overcome facial recognition systems. But more importantly, such an attack is not at all resource intensive in terms of time or knowledge of the attacker. And it is executable using freely available tools on the Internet such as demos for scientific papers, which are also online, so it is possible to create a good deepfake using a smartphone without installing additional software. Another interesting thing we noticed in the experiment is the fact that current models for generating deepfakes cannot cope with cases where another object moves in front of the face. Also, these systems have a problem if the user is required to turn his head more than about 60 degrees. In this case, artifacts are produced that would be easily detectable. Requiring these simple stops from the user can strengthen the robustness of these systems and at the same time is not prohibitive to the performance required for verification.

Furthermore, we were also able to answer the following questions that were asked during the design of the experiment:

How difficult is it to create a deepfake?

As we have shown in the experiment, generating a deepfake is currently easy and requires almost no knowledge to generate a deepfake. An attacker only needs a mobile phone and access to the Internet to create a good deepfake. We have also shown that it is possible to generate a good deepfake if a few rules are followed, such as appropriate background selection and no object moving in front of the face swapped face. It is also advisable that the actor whose face will be face swapped to the victim's face looks at least a little similar.

How difficult is it to create a video deepfake that can fool a facial recognition system?

When generating deepfakes, it is necessary for the attacker to be careful to choose a suitable background on which the face will be face swapped. It is also preferable to choose a model that faces swaps the face because it can better preserve the quality of the performance. Under these conditions, creating deepfakes that deceive biometrics is not difficult at all.

Furthermore, questions have arisen from this experiment. If it is possible to overcome facial recognition systems, how big is this problem? We will try to answer this question in the next experiment.

5.3 Experiment 2: Robustness of face recognition systems

In a previous experiment, we showed that it is possible to edge face recognition systems using deepfakes. This experiment, the design of which is described in Section 4.6.2, aims to find out how big a problem it is to fool face recognition systems using deepfakes. This section will describe how the experiment will be performed. Next, the results will be evaluated, and the statistical similarity between the different data sets will be determined. Finally, the questions that were asked during the design of the experiment will be answered.

5.3.1 Tested biometric facial recognition systems

Compared to the previous experiment, we decided to exclude the third tested system, Regula, which is absolutely unable to distinguish between deepfakes and real photos. Therefore, the tested biometric facial recognition systems will be Megamatcher from Neurotechnology and IFace from Innovatrics. The systems were introduced at the beginning of this chapter. These systems have the ability to be run via the command line which allows them to be automated for a greater number of comparisons. To automate these biometrics, a script has been created that runs single-unit evaluations and adds the results to a spreadsheet. For each identity, 3000 comparisons were made for all three types of scores: Genuine score, Deepfake score, and Impostor score.

Megamatcher

From the developer's website, we downloaded a trial package with SDK which contains examples of programs one of them is a program that receives two photos as input and runs a comparison of the two files. After the evaluation, we get the scores back.

IFace

For IFace, we used the SDK directly from the company, this package also contains example programs, one of them is a program for comparing two photos. The system returns the pair of scores that were described in the previous experiment.

5.3.2 Obtained results

After performing all the comparisons for both face recognition systems, we were able to compute all three required scores: genuine score, deepfake score and impostor score. From these scores, distribution function plots were then computed. In the following figures, we

can see the distribution for IFace accurate mode in Figure 5.8, IFace fast mode in Figure 5.9, and Megamatcher in Figure 5.10.

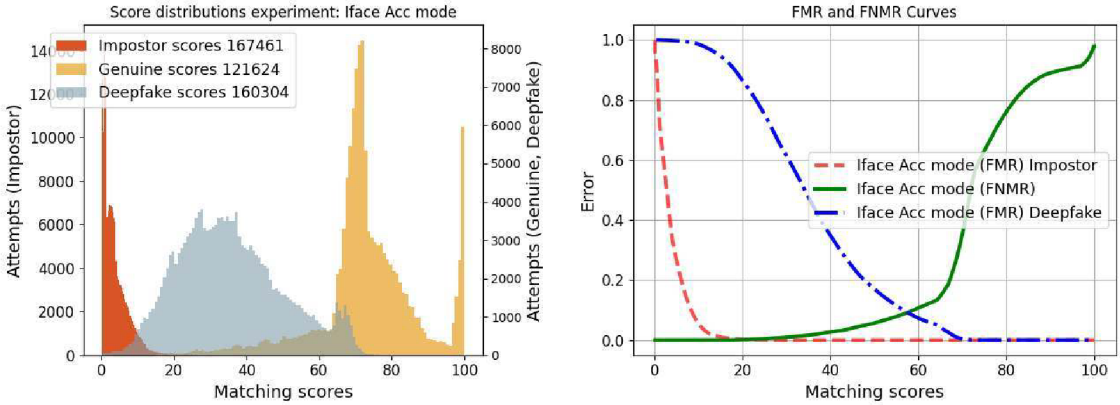

Figure 5.8: Matching scores distribution graph on left and on right FMR / FNMR graphs for the IFace accuracy mode.

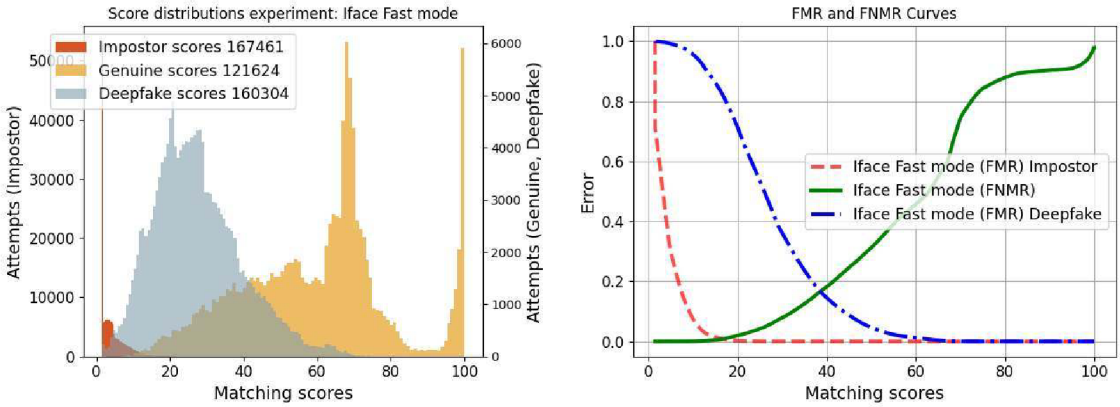

Figure 5.9: Matching scores distribution graph on left and on right FMR / FNMR graphs for the IFace fast mode.

In the charts, we can see some overlap between deepfakes and genuine scores. This confirms that some deepfakes can achieve a sufficient score to be accepted by the system as genuine. However, it should be added that the dataset used is quite outdated (published in 2020), but new types of models to generate deepfakes could achieve even better results.

We can also observe that a few deepfakes were evaluated as impostors. And also their overlap with the impostor set is not so significant. The statistical similarity of the different types of scores will need to be confirmed or refuted by a statistical test, which will be described in the next section of the experiment.

The interesting thing then is the difference between fast and accuracy mods in IFace where the ratio of detected deepfakes improved. This confirms the claims made in the technical documentation that the accuracy mod gives more accurate results even at the cost of higher CPU overhead.

In the case of the Megamatcher charts, we can observe that there is also an overlap between the deepfakes score and the genuine score. Another interesting thing is that the

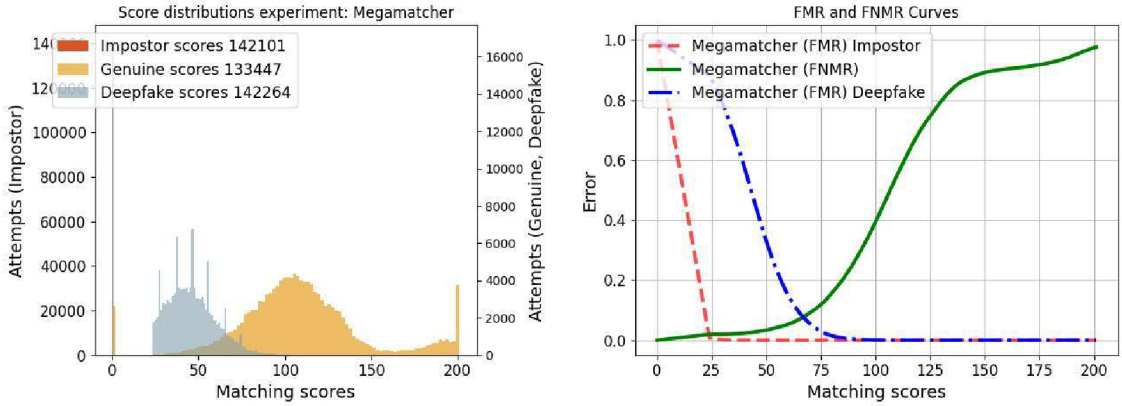

Figure 5.10: Matching scores distribution graph on left and on right FMR / FNMR graphs for the Megamatcher.

impostor score is set to zero for Megamatcher, which we can observe also for deepfakes that have been scored lower than 25 points.

5.3.3 Statistical similarity of results

The last step in the experiment is to verify the statistical similarity between the deepfake score and the genuine score. And in turn, whether there is a statistical difference between deepfake score and genuine score. We verify this using the student's t-test that we introduced in Section 4.7. For all biometrics, we obtained similar results so we will only give one procedure as an example. The results of all tests will be described in Table 5.7.

First, we will verify the statistical similarity of impostor scores and deepfakes scores so we can form the following hypothesis:

$$H_0 : m_D = m_I$$

This means the mean m_D of the population from which the deepfake measurements come is equal to the mean m_I of the population from which the impostor measurements come. We also have the alternative hypothesis:

$$H_A : m_D \neq m_I$$

This means the mean m_D of the population from which the deepfake measurements come is not equal to the mean m_I of the population from which the impostor measurements come. This means that the two sets are statistically different.

After computing the t-test, we get $p - value = 0.0$, with significance level $\alpha = 0.05$. Since $p - value <= \alpha$ we reject the original hypothesis H_0 and accept the hypothesis H_A. So it holds that the set of impostor scores is statistically distinct from the set of impostor scores.

The next step is to verify the statistical similarity of the deepfake score and the genuine score, we get the following hypothesis:

$$H_0 : m_D = m_G$$

This means the mean m_D of the population from which the deepfake measurements come is equal to the mean m_G of the population from which the impostor measurements come. We also have the opposite hypothesis:

$$H_A : m_D \neq m_G$$

This means the mean m_D of the population from which the deepfake measurements come is not equal to the mean m_G of the population from which the impostor measurements come. This means that the two sets are statistically different.

After computing the t-test, we get $p - value = 0.0$, with significance level $\alpha = 0.05$. Since $p - value <= \alpha$ we reject the original hypothesis H_0 and accept the hypothesis H_A. So it holds that the set of impostor scores is statistically different from the set of genuine scores. The results for all sets can be seen in Table 5.7

Score comparison	**Megamatcher**		**IFace fast**		**IFace accurate**	
	statistic	**p-value**	**statistic**	**p-value**	**statistic**	**p-value**
DF vs. IM	770,87	0,0	869,65	0.0	501,39	0.0
DF vs. G	600,63	0,0	710,58	0.0	748,93	0.0
IM vs. G	1077,99	0,0	1752,34	0.0	955,07	0.0

Table 5.7: All sets of scores are statistically independent since the p-value is smaller than the significance level at $\alpha = 0.05$. DF denotes Deepfake, IM denotes Impostor and G denotes Genuine.

5.3.4 Conclusion of the experiment

As we can see in Table 5.7, for all groups of scores we have shown that they are statistically independent of each other. Interestingly, the deepfake score is different from the impostor score. This shows that the tested face recognition systems are not able to determine deepfake as an impostor input. However, we also show that deepfakes are statistically different from genuine scores as well. However, this does not match our hypothesis from the first experiment. This may be due to the ability of systems to realistically counter deepfakes but is more likely due to the older dataset. The distribution plots show that with a poorly set acceptance threshold, it would be possible to fool the system even with a part of the deepfakes from the dataset. We also got answers to the following questions:

How easy is it to fool selected facial recognition systems using deepfake deception?

The measurement results show that facial recognition systems cannot completely reject deepfake as impostor input. However, the experiment also showed that some comparisons achieved scores that would be sufficient to accept deepfakes as valid input, which is particularly the case for Megamatcher. This is evidenced by the EER, which is higher for deepfake scores than for impostor scores.

Is the difference in obtained matching scores statistically significant ?

Using Student's t-test, we have shown in Table 5.7 that the difference between the impostor score and deepfake score is statistically significant. We have also shown that the difference between the deepfake dataset and the genuine score is statistically significant as well. The

reason why there is not more similarity between deepfakes score and genuine score may be due to the use of older dataset.

5.4 Experiment 3: Image comparison against image sequences

The last experiment is to check whether the ability of the systems to resist deepfakes is improved if several photos from a video are compared instead of one. The design of the experiment was described in Section 4.6.3. As in the previous experiment, the biometrics tested will be Megamatcher and IFace.

Compared to the previous experiment, instead of a single photo, a sequence of five photos from a video is compared with a single image of a real person. An evaluation will then be run over each of these photos, and the final score will be as the average of all the scores from all the photos tested. 80 measurements will be taken over each identity. Each of these measurements will consist of a comparison of 5 photos from the video that both meet the ICAO standard and are at least a second of recording apart.

The purpose of selecting frames from the record with a difference of a second is to prevent several very similar frames from appearing in the sequence and being rated with the same early value. From these averaged scores all necessary scores are then calculated: genuine score, deepfake score, and impostor score, which are further used to calculate the distribution graph and also the FMR and FNMR curves.

5.4.1 Obtained results

After performing all the comparisons and averaging the image sequences, we obtained the scores which, as in the previous experiment, we use to compute the graph of the distribution function and also to plot the FMR and FNMR graphs for each face recognition system.

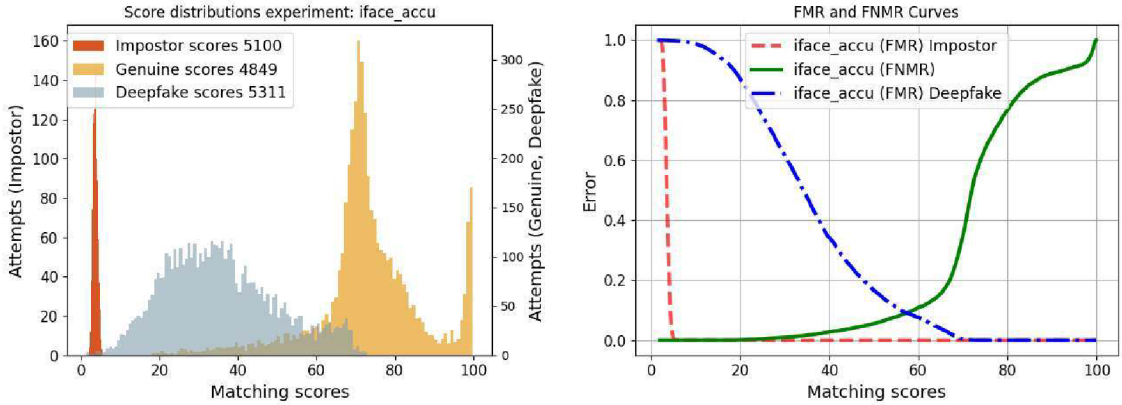

Figure 5.11: Matching scores distribution graph on left and on right FMR / FNMR graphs for the IFace accurate.

From Figures 5.11 and 5.12 we see that the EER between the impostor score and the genuine score has moved closer to zero. Interestingly, there was no significant change in EER between deepfakes and genuine scores. This means that even multiple frames did not help the system detect the deepfake. Thus, the classifier did not become more robust to deepfakes as we expected. The same is true for the graph in Figure 5.13 from the

Megamatcher measurements, where the EER did not change compared to the results in Figure 5.7 from the second experiment.

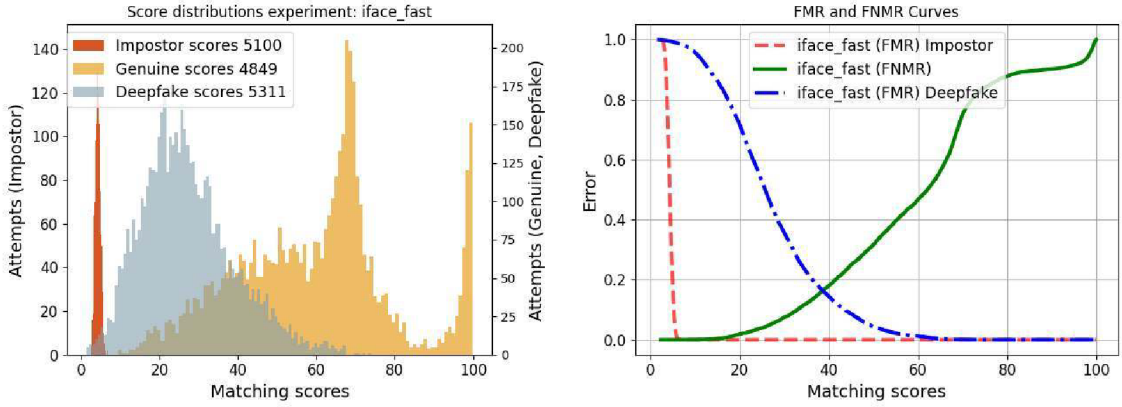

Figure 5.12: Matching scores distribution graph on left and on right FMR / FNMR graphs for the IFace fast.

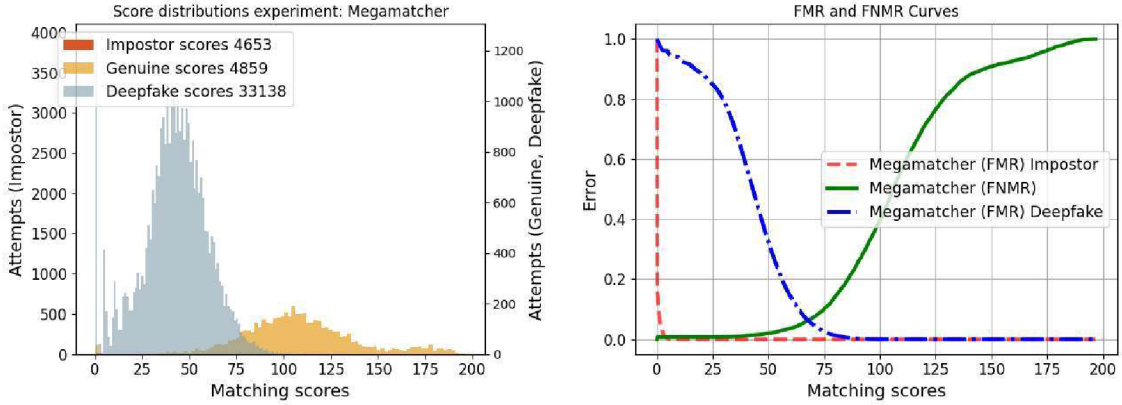

Figure 5.13: Matching scores distribution graph on left and on right FMR / FNMR graphs for the Megamatcher.

5.4.2 Statistical tests

As in the previous experiment, we will use statistical tests to determine the statistical similarity between the deepfake score and the genuine score and the statistical similarity between the impostor score and the genuine score. We also use the Student's t-test introduced in Section 4.7, and we test the hypotheses with significance level $\alpha = 0.05$. The results of these tests can be seen in Table 5.8.

These tests show that, as in the previous experiment, there is no statistical similarity between deepfakes and impostor scores. This again shows that the system cannot unambiguously identify a deepfake as an input that it should reject. On the other hand, it is also shown that there is no statistical similarity between genuine scores and deepfake scores.

Score comparison	Megamatcher		IFace fast		IFace accurate	
	statistic	**p-value**	**statistic**	**p-value**	**statistic**	**p-value**
DF vs. IM	159.9808	0,0	138.9438	0.0	161.01784	0.0
DF vs. G	-216.9981	0,0	-101.2333	0.0	-136.2099	0.0
IM vs. G	242.8665	0,0	199.1738	0.0	369.163	0.0

Table 5.8: Results of statistical tests between individual scores. DF denotes Deepfake, IM denotes Impostor and G denotes Genuine. DF denotes Deepfake, IM denotes Impostor and G denotes Genuine.

5.4.3 Conclusion of the experiment

The experiment showed that the robustness of deepfakes does not improve even when the system has to compare multiple deepfake images against a single real image. This can be seen in the FMR/FNMR plots where the EER value shifted between impostor and genuine but remained the same between deepfake and genuine. Also in this experiment, it was confirmed that the tested systems cannot reliably identify deepfake as an impostor input.

Will the system improve its robustness to deepfakes if it processes a sequence of photos instead of a single photo?

The robustness of the system is not increased. This is evidenced by the fact that the deepfake EER has not moved closer to zero in contrast to the impostor EER. In statistical tests, it was shown that the statistical similarity of impostor early and deepfake early did not change.

Chapter 6

Discusion

With better models and tools for creating deepfakes, there are increasing requirements for updating deepfakes datasets that no longer capture the current state. It will therefore be necessary in the near future to create better datasets based on state-of-the-art techniques for generating facial deepfakes. The next step should also be to build a dataset suitable for testing facial biometrics, and that meets or at least comes close to the requirements for photographs used in face recognition, such as appropriate lighting, looking the subject directly into the camera, and so on, as these facts are not taken into account in the current datasets.

From the dataset quality point of view, it is necessary to mention that deepfakes in them should also consider the similarity between persons when face swapping. An attacker who wants to trick the system will likely look for a person that resembles the victim as closely as possible as a basis for the face swapped face. This was exactly confirmed in our work, where the best results were achieved by deepfakes where the person whose face was replaced was similar to the victim.

We assume that the reason why it could not be clearly demonstrated that face recognition systems are vulnerable to a greater extent is precisely the quality of the dataset used. This dataset no longer captures the current capabilities of deepfakes tools since Celeb-DF is older. It should also be added that it was not created as a dataset for testing the robustness of biometric facial recognition systems against deepfakes. This fact is most evident in the quality of the videos, which are obtained from freely available sources on the Internet. These videos absolutely do not meet the requirements for identity verification based on ICAO standards. The difference in quality can be seen in Figure 6.1.

Despite this fact, the results clearly show the inability of the systems to reject deepfakes unambiguously. In the first experiment, we clearly showed that such an attack can be performed and it is highly likely that with a better dataset, the number of deepfakes that would pass through these systems would increase.

Future work should therefore focus on determining the robustness of biometric facial recognition systems to modern deepfake techniques. There is also scope to test these systems on other datasets to which the author of this thesis did not have access.

In this work, we also described the problems that current tools for creating deepfakes have. The first is the need to ensure that no object, such as a hand, passes in front of the face on which the victim's face will be face swapped. In case such an object appears there, the resulting image will be significantly degraded and various artifacts will start to appear in the image, such as facial features in the media. This was also shown in Section 5.2.2, where

Figure 6.1: Comparison of the quality of deepfakes in the Celeb-DF dataset with the quality of current tools (GHOST).

we can see that suddenly in the image the person's eye distance changes or for example, the hand disappears under the mask.

A similar problem occurs when the person in the video turns their head so dramatically that some of the nodal points of the face are lost. In this case, the face swap will incorrectly overlay the floating face or even make the mask disappear and show the underlying face. In systems that create an image using face reantacment, a number of artifacts are created, such as different stretching of a part of the face, or a change in facial features.

Another interesting fact is that the robustness of the systems does not improve even if multiple images are compared. This can be seen by comparing the graphs from IFace in an accurate mode in Figure 6.2, but also applies to IFace fast mode and Megamatcher. In the first experiment, where a single photo was compared to a single photo, the EER for deepfakes is at 55%, and the EER for impostors is at 20%. In case the resulting score is calculated from multiple photos the EER is reduced to 6%, but for deepfakes, there is no change, which clearly shows that the robustness against deepfakes is not changed.

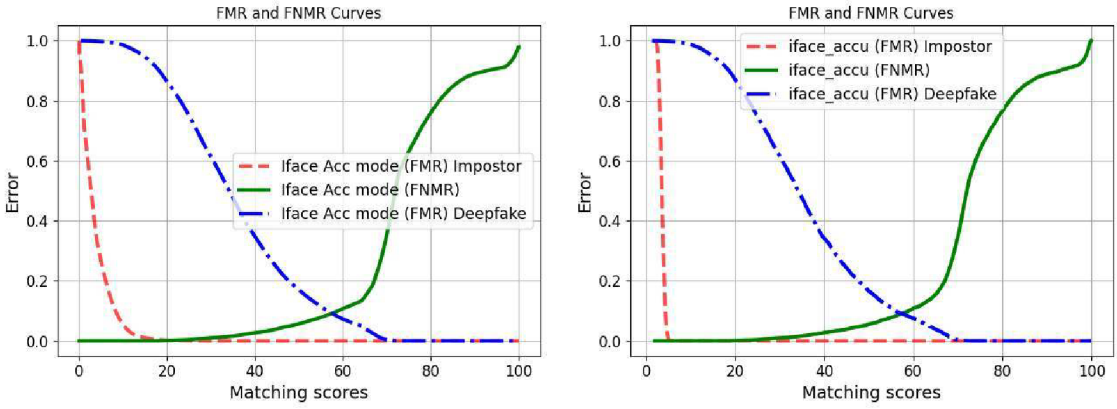

Figure 6.2: FMR / FNMR graphs for the IFace accurate mode. On the left to compare a single photo and on the right compare a sequence of photos. The EER for the impostor score decreased for the image sequence, but the deepfakes score remained the same.

These findings can improve facial recognition systems and help detect deep fakes that could attack the biometric systems in question. Simply turning the head at an angle that no longer shows all the key points of the face. This would, for example, make it completely impossible to use one-off deepfake models because the model cannot generate the side of the face from a single 2D image taken from the front. Another safeguard could be, for example, waving the hands in front of the face during verification. These suggestions can significantly improve the robustness of the systems and are also not implementation specific.

Chapter 7

Conclusion

In this work, we have clearly shown that deepfakes of faces are a threat to automatic face recognition systems because deepfakes can fool these systems. We first verified this kind of attack's technical feasibility and resource requirements.

Our experiments showed that an attacker only needs a single photo to create a conceivable deepfake, which makes the attack much easier. We have also shown that little computational power is needed to perform this attack since a potential attacker can use freely available code samples from scientific papers on the Internet and run them in the phone's browser. These tools make it possible to create a deepfake of sufficient quality to fool facial recognition systems. This clearly demonstrates the technical feasibility of such an attack using current freely available tools.

However, current systems for creating deepfakes of faces also suffer from certain limitations. A significant loss of quality occurs when the face is rotated to an angle where some key points of the face are not visible, then either the face swapped face completely disappears or the face is stretched.

Another problem for deepfake generators is frames where part of the face is covered, for example by a hand. It depends on how large the overlap is, but the resulting errors range from the disappearance of the hand behind the face swapped face to the appearance of various artifacts such as various flickers or poor alignment of the face with the background.

If a biometric system required these authentication activities, the robustness of biometric systems to deepfakes would be significantly increased, at least until the tools for generating deepfakes are improved again.

Through further experiments, we were able to show that the tested face recognition systems are not fully capable of distinguishing deepfakes from real photos. The selected systems tested were Megamatcher from Neurotechnology and IFace from Innovatrics. The results showed that the tested systems could not clearly reject deepfakes, which already indicates a potential problem. We demonstrated this with statistical tests that showed that the set of deepfakes results was statistically independent of the results of the imposters. We also found that the deepfakes score is not statistically similar to the actual score, which could be due to the use of the older Celeb-DF dataset that no longer actively captures the capabilities of the deepfakes generation tools. The videos in this dataset are often not of sufficient quality and also often do not meet ICAO standards.

We also tested whether the robustness of face recognition systems is improved when video is required for verification instead of a single image. The experiment showed that while the EER for imposters decreased, making the system more resilient to imposters, the EER for deepfakes remained the same as in the second experiment. That is, the system

did not become more resilient to deepfake attacks even after processing multiple frames of video. This contradicts the claims of some works that recommend viewing the video as a set of frames as a way to evaluate the video and calculate a score from each of these frames, which is then averaged over the entire video.

Further work should therefore focus on the creation of a new dataset that will reflect the capabilities of modern tools for creating facial deepfakes and at the same time the videos that will be in this dataset will meet certain quality requirements such as resolution, proper lighting or direct looking into the camera.

With the results of all three experiments, we can conclude that facial recognition systems are vulnerable to deepfakes since we have successfully attacked these systems. We have also shown that these systems cannot reliably reject deepfakes and with a better dataset, there is a chance that they would even be able to accept deepfakes as valid input, which would be statistically measurable. To increase resilience, I propose the following measures:

- Require the user to cover part of their face with their hand during authentication.
- Requiring the user to rotate the head by such an angle that some of the face's nodal points disappear so that if it is a deepfake, it is not possible to bind a face swap to the underlying face.

Bibliography

[1] *Ai Face Swap Online app for video, Photo & GIF.* Available at: https://www.deepswap.ai/.

[2] *Xpression camera.* Available at: https://xpressioncamera.com/.

[3] *Case study: Lenovo™ and Verilook - Neurotechnology.* Neurotechnology, 2007. Available at: https://neurotechnology.com/download/CaseStudy_Lenovo_and_VeriLook.pdf.

[4] *Biometric performance evaluation - precise biometrics.* Precise Biometrics AB, 2014. Available at: http://precisebiometrics.com/wp-content/uploads/2014/11/White-Paper-Understanding-Biometric-Performance-Evaluation.pdf.

[5] *Facing reality? law enforcement and the challenge of deepfakes.* Apr 2022. Available at: https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

[6] *Intel introduces real-time Deepfake Detector.* Intel, 2022. Available at: https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html#gs.m9pr0j.

[7] *Yubo's New Age Verification Feature Helps Keep You Safe.* Apr 2023. Available at: https://www.yubo.live/blog/yubos-new-age-verification-feature-helps-keep-you-safe.

[8] 97108. *Generating MRI images of brain tumors with Gans.* Medium, Oct 2019. Available at: https://971080.medium.com/generating-mri-images-of-brain-tumors-with-gans-8cddedbabbe6.

[9] Agostini, D. D. *Why Deepfakes will make you play video games instead of Movies.* Predict, May 2020. Available at: https://medium.com/predict/why-deepfakes-will-make-you-play-video-games-instead-of-movies-99ee5c2d7c9e.

[10] Allyn, B. *Deepfake video of Zelenskyy could be 'tip of the iceberg' in Info War, experts warn.* NPR, Mar 2022. Available at: https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

[11] Anderson, M. *Deepfaked voice enabled $35 million bank heist in 2020.* Oct 2021. Available at: https://www.unite.ai/deepfaked-voice-enabled-35-million-bank-heist-in-2020/.

[12] Aratek. *Reading your face: How does facial recognition work?* Available at: https://www.aratek.co/news/how-does-facial-recognition-work.

[13] BioPass ID. *Learn the requirements of the ICAO standard (ISO 19794-5).* Available at: https://www.biopassid.com/post/norma-icao.

[14] Biswas, S. and Chellappa, R. Face Recognition from Still Images and Video. In: *Encyclopedia of Cryptography and Security.* 2011, p. 437–444. Available at: https://doi.org/10.1007/978-1-4419-5906-5_739.

[15] Burchard, H. v. d. *Belgian Socialist Party circulates 'Deep fake' donald trump video.* POLITICO, May 2018. Available at: https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/.

[16] Chakraborty, S. and Das, D. An Overview of Face Liveness Detection. *CoRR.* 2014, abs/1405.2227. Available at: http://arxiv.org/abs/1405.2227.

[17] Chesney, R. and Citron, D. K. Deep Fakes: A looming challenge for privacy, democracy, and national security. *SSRN Electronic Journal.* 2018. DOI: 10.2139/ssrn.3213954.

[18] Collins, K. *Big Tech faces fines for Deepfakes, bots, fake accounts under new EU rules.* CNET, Jun 2022. Available at: https://www.cnet.com/news/politics/eu-strengthens-disinformation-rules-to-target-deepfakes-bots-fake-accounts/.

[19] Dalvi, J., Bafna, S., Bagaria, D. and Virnodkar, S. *A Survey on Face Recognition Systems.* January 2022.

[20] Demir, I. and Ciftci, U. A. Where Do Deep Fakes Look? Synthetic Face Detection via Gaze Tracking. In: *ACM Symposium on Eye Tracking Research and Applications.* New York, NY, USA: Association for Computing Machinery, 2021. ETRA '21 Full Papers. DOI: 10.1145/3448017.3457387. ISBN 9781450383448. Available at: https://doi.org/10.1145/3448017.3457387.

[21] Firc, A., Malinka, K. and Hanáček, P. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon.* 2023, vol. 9, no. 4, p. e15090. DOI: https://doi.org/10.1016/j.heliyon.2023.e15090. ISSN 2405-8440. Available at: https://www.sciencedirect.com/science/article/pii/S2405844023022971.

[22] Grassi, P. A. and Fenton, J. L. *Digital identity guidelines - NIST.* NIST, Jun 2017. Available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63a.pdf.

[23] Groshev, A., Maltseva, A., Chesakov, D., Kuznetsov, A. and Dimitrov, D. Ghost—a new face swap approach for image and video domains. *IEEE Access.* 2022, vol. 10, p. 83452–83462. DOI: 10.1109/access.2022.3196668.

[24] Hassan, G. and Elgazzar, K. The case of face recognition on mobile devices. *2016 IEEE Wireless Communications and Networking Conference.* 2016. DOI: 10.1109/wcnc.2016.7564975.

[25] Hellerman, J. *How is 'Star wars' using Respeecher to revitalize voices?* No Film School, Jun 2022. Available at: https://nofilmschool.com/respeecher-star-wars.

[26] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M. et al. *Video Diffusion Models*. 2022.

[27] Holt, K. *UK aims to ban non-consensual deepfake porn in online safety bill*. Nov 2022. Available at: https://www.engadget.com/deepfake-porn-uk-ban-online-safety-bill-171007700.html.

[28] Huang, K. *Why pope Francis is the star of a.i.-generated photos*. The New York Times, Apr 2023. Available at: https://www.nytimes.com/2023/04/08/technology/ai-photos-pope-francis.html.

[29] Jain, A. K., Hong, L. and Pankanti, S. Biometric identification. *Commun. ACM*. 2000, vol. 43, p. 90–98.

[30] Kharpal, A. *China is about to get tougher on deepfakes in an unprecedented way. here's what the rules mean*. CNBC, Dec 2022. Available at: https://www.cnbc.com/2022/12/23/china-is-bringing-in-first-of-its-kind-regulation-on-deepfakes.html.

[31] Khidhir, S. *Malaysian actor in „Porn" video blames Deepfake*. Dec 2019. Available at: https://theaseanpost.com/article/malaysian-actor-porn-video-blames-deepfake.

[32] Korshunov, P. and Marcel, S. *DeepFakes: a New Threat to Face Recognition? Assessment and Detection*. 2018.

[33] Kwon, P., You, J., Nam, G., Park, S. and Chae, G. KoDF: A Large-Scale Korean DeepFake Detection Dataset. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. October 2021, p. 10744–10753.

[34] Li, C., Wang, L., Ji, S., Zhang, X., Xi, Z. et al. *Seeing is Living? Rethinking the Security of Facial Liveness Verification in the Deepfake Era*. 2022.

[35] Li, Y. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In: *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*. 2020.

[36] Martinez, A. and Kak, A. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2001, vol. 23, no. 2, p. 228–233. DOI: 10.1109/34.908974.

[37] Mirsky, Y. and Lee, W. The Creation and Detection of Deepfakes: A Survey. *CoRR*. 2020, abs/2004.11138. Available at: https://arxiv.org/abs/2004.11138.

[38] Neurotechnology. *Megamatcher SDK*. Neurotechnology, Jan 2023. Available at: https://www.neurotechnology.com/megamatcher.html.

[39] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh The, T. et al. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*. Elsevier. 2022, vol. 223, p. 103525.

[40] Nightingale, S. J. and Farid, H. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*. 2022, vol. 119, no. 8. DOI: 10.1073/pnas.2120481119.

[41] Olszewski, K., Li, Z., Yang, C., Zhou, Y., Yu, R. et al. *Realistic dynamic facial textures from a single image using Gans.* Jan 1970. Available at: https://openaccess.thecvf.com/content_iccv_2017/html/Olszewski_Realistic_Dynamic_Facial_ICCV_2017_paper.html.

[42] Onfido. *Revolut onboards 12 percent more users with Onfido.* Nov 2022. Available at: https://onfido.com/customer/revolut/.

[43] Ovsiannikov, K. *Face recognition system for ATM Security.* Oct 2022. Available at: https://atmeye.com/blog/face-recognition-atm/.

[44] Parmar, D. N. and Mehta, B. B. *Face Recognition Methods & Applications.* 2014.

[45] Person and Raymond, S. *Deepfake anyone? Ai Synthetic Media Tech enters perilous phase.* Thomson Reuters, Dec 2021. Available at: https://www.reuters.com/technology/deepfake-anyone-ai-synthetic-media-tech-enters-perilous-phase-2021-12-13/.

[46] Petkauskas, V. *An unintended consequence: Can deepfakes kill video evidence?* Available at: https://cybernews.com/privacy/an-unintended-consequence-can-deepfakes-kill-video-evidence/.

[47] Ratha, N. K., Connell, J. H. and Bolle, R. M. *An analysis of minutiae matching strength.* Springer Berlin Heidelberg, Jan 1970. Available at: https://link.springer.com/chapter/10.1007/3-540-45344-X_32.

[48] Rathgeb, C., Tolosana, R., Vera Rodriguez, R. and Busch, C., ed. *Handbook of Digital Face Manipulation and Detection.* 1.th ed. Springer International Publishing, 2022. ISBN 978-3-030-87663-0. Available at: https://doi.org/10.1007/978-3-030-87664-7.

[49] Revathy, G., Bhavana Raj, K., Kumar, A., Adibatti, S., Dahiya, P. et al. Investigation of e-voting system using face recognition using convolutional neural network (CNN). *Theoretical Computer Science.* 2022, vol. 925, p. 61–67. DOI: 10.1016/j.tcs.2022.05.005.

[50] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. et al. FaceForensics++: Learning to Detect Manipulated Facial Images. In: *International Conference on Computer Vision (ICCV).* 2019.

[51] Salpekar, O. DeepFake Image Detection. In:. 2020.

[52] Schroff, F., Kalenichenko, D. and Philbin, J. In:. IEEE, June 2015. DOI: 10.1109/cvpr.2015.7298682. Available at: https://doi.org/10.1109%2Fcvpr.2015.7298682.

[53] Schuckers, M. E. *Computational methods in biometric authentication.* Springer, 2010.

[54] Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E. and Sebe, N. First Order Motion Model for Image Animation. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché Buc, F. d', Fox, E. et al., ed. *Advances in Neural Information Processing Systems.* Curran Associates, Inc., 2019, vol. 32. Available at:

https://proceedings.neurips.cc/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf.

[55] Slovenská sporiteľňa, a. *3 kroky, vďaka ktorým sa Ukrajinskému Startupu Podarilo Oživiť Hlas Jula Satinského*. Nov 2022. Available at: https://www.slsp.sk/sk/aktuality/2022/11/11/ako-ukrajinsky-startup-ozivil-hlas-jula-satinskeho.

[56] s.r.o, I. *Face functions*. Innovatrics s.r.o, Jan 2021. Available at: https://developers.innovatrics.com/digital-onboarding/docs/functionalities/face/.

[57] Tariq, S., Jeon, S. and Woo, S. S. Am I a Real or Fake Celebrity? Measuring Commercial Face Recognition Web APIs under Deepfake Impersonation Attack. *CoRR*. 2021, abs/2103.00847. Available at: https://arxiv.org/abs/2103.00847.

[58] Team, E. *Why deepfake fraud losses should scare financial institutions*. Finextra, Nov 2022. Available at: https://www.finextra.com/blogposting/23223/why-deepfake-fraud-losses-should-scare-financial-institutions.

[59] Tolosana, R., Vera-Rodríguez, R., Fiérrez, J., Morales, A. and Ortega-Garcia, J. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. *CoRR*. 2020, abs/2001.00179. Available at: http://arxiv.org/abs/2001.00179.

[60] Turk, M. and Pentland, A. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*. january 1991, vol. 3, no. 1, p. 71–86. DOI: 10.1162/jocn.1991.3.1.71. ISSN 0898-929X. Available at: https://doi.org/10.1162/jocn.1991.3.1.71.

[61] Westerlund, M. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*. Ottawa: Talent First Network. 11/2019 2019, vol. 9, p. 40–53. DOI: http://doi.org/10.22215/timreview/1282. ISSN 1927-0321. Available at: timreview.ca/article/1282.

[62] Wiggers, K. *Google bans deepfake-generating AI from colab*. Jun 2022. Available at: https://techcrunch.com/2022/06/01/2328459/.

[63] Younglawyerssection. *Deepfakes: A balancing act*. Sep 2019. Available at: https://cbaatthebar.chicagobar.org/2019/09/24/deepfakes-a-balancing-act/.

[64] Yu, P., Xia, Z., Fei, J. and Lu, Y. A survey on Deepfake Video detection. *IET Biometrics*. 2021, vol. 10, no. 6, p. 607–624. DOI: 10.1049/bme2.12031.

[65] Zhang, S., Yuan, J., Liao, M. and Zhang, L. *Text2Video: Text-driven Talking-head Video Synthesis with Personalized Phoneme-Pose Dictionary*. 2022.

Appendix A

Attached media to work

The attached storage media contains files related to the thesis:

- **outputs/** – contains scores obtained from experiments 2 and 3.
- **src/** – contains the Python scripts used to obtain the results from the IFace and Megamatcher biometric systems
- **README.md** – readme file