Alma Mater Studiorum - Università di Bologna
CAMPUS DI CESENA

Dipartimento di Informatica - Scienza e Ingegneria
Corso di Laurea in Ingegneria e Scienze Informatiche

Video-based Morphing Attack Detection

Elaborato in Visione Artificiale

Relatore
Prof.ssa Annalisa Franco
Co-relatori
Prof. Guido Borghi
Prof. Matteo Ferrara

Presentata da
Fabio Notaro

Anno Accademico 2024/2025

Abstract

Face morphing represents a real threat to facial recognition systems, as it enables the generation of hybrid images of two subjects capable of evading biometric controls, both human and automatic, adopted in critical contexts such as airport gates.

In this thesis first the main morphing attack detection (MAD) generation and detection techniques, both S-MAD (single-image) and D-MAD (differential), are analyzed in detail.

A new and novel paradigm, called V-MAD (Video-based Morphing Attack Detection), which exploits video sequences instead of single static images to improve the effectiveness of detection, is then introduced and explored.

After analyzing various MAD score fusion strategies and integration with image quality metrics, it is described in detail how it was possible to build a proprietary dataset conforming to ICAO standards, characterized by excellent quality and good variability and heterogeneity of the data present, and experiments were conducted on it and on public and semi-public databases.

The results obtained show that V-MAD, even when implemented with simple aggregation techniques, has the potential to be able to outperform traditional D-MAD systems, with further improvement achieved through the use of Face Image Quality Assessment algorithms.

This work therefore sets the stage for a new direction in biometric security research, opening concrete prospects for the adoption of more robust control systems that are resilient to morphing attacks.

”Teme nosce”

Latin locution

Dedicated to my family for their unconditional support and daily affection, which have always been a point of reference in every choice I have made.

Dedicated to friends, long-time and newer, in university and outside, in particular Matteo, Mattia, Nicholas, Giuseppe, Sara and Giulio, for moments of genuine lightness and for making even the most difficult times more bearable with their sincere presence.

Dedicated to Andrea, for his sincere friendship and helpfulness he has always shown me, in and out of university, with whom I have shared both the satisfactions and fears of this journey.

Dedicated to Giacomo and Alex, fellow travelers between university desks with whom I shared laughter, hard work and an untold number of projects always tackled with team spirit and friendship.

Dedicated to my supervisors, Annalisa Franco and Guido Borghi, for their constant availability, attentive guidance and valuable suggestions that have significantly enriched this work.

Dedicated to Dr. Nicolò Di Domenico and Dr. Lorenzo Pellegrini for their generous and knowledgeable support, patience and advice that made a difference not only in this thesis but also in my educational journey.

Dedicated finally to myself, for learning over time to recognize that difficulties are part of the journey and that facing them together with the people around us makes it more human and a little less scary.

Contents

Chapter 1

FACE MORPHING

1.1 Definition

Face morphing is a computational graphics technique of combining two or more images of real human faces to generate an intermediate image, which has features of both original subjects and can be considered as a specific frame of a smooth transition between the input faces[1].

This technique has applications in both artistic and scientific fields, for example in the creation of special effects for movies or in the study of the perception of human facial expressions, yet it can also be misused to bypass the automated biometric security systems that are becoming increasingly popular in many contexts such as banks and airports.

The face morphing process, which will be explored in detail in this chapter, is based on identifying key points (called landmark points) in the faces of the original subjects, usually in the eye, mouth and nose regions, and then building a landmark map.
Successively, through geometric distortion combined with a brightness value blending operation, an intermediate image is generated that shows a smooth transition between the distinctive features of the source images.

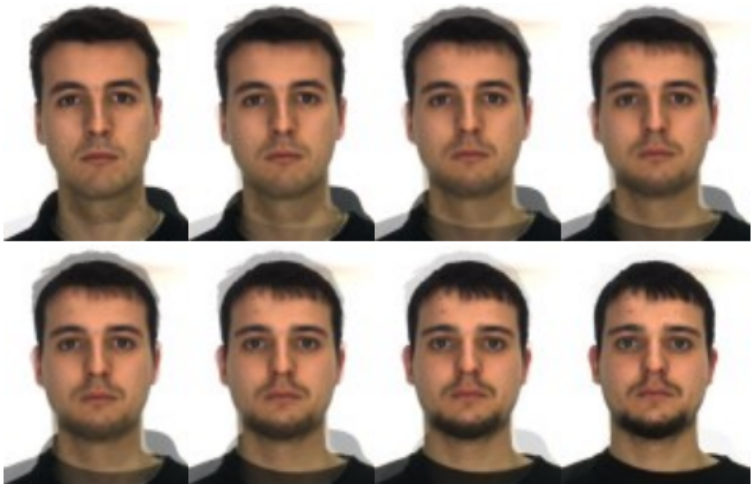


Figure 1.1: transition between frames that gradually fade from one face to another. Source:[1].

Face morphing can be performed manually or by specialized software that simplifies the identification of key facial points and speeds up the intermediate image generation process[2].

This thesis analyzes face morphing primarily from the perspective of its problematic implications related to the world of biometric security, highlighting the risks, consequences and possible countermeasures to be taken.
In particular, since this technique allows the creation of hybrid images of multiple subjects, it can be exploited to bypass biometric recognition systems adopted in sensitive contexts such as banks and airports[3].

Some examples of possible threats include:

- bypass of facial recognition systems through synthetic images that allow unauthorized access to sensitive data or protected facilities
- compromise of the principle of uniqueness between a document and its legitimate holder
- fraud related to forgery and identity theft
- privacy risks $\rightarrow$ arising from the ability of face morphing to generate deceptive images that could damage an individual's reputation
- reduced effectiveness of biometric security measures due to difficulties in distinguishing a genuine face from a manipulated one.

To mitigate these risks, it is essential to promote awareness of the limitations of current biometric technologies and to incentivize research not only to improve facial recognition systems but also to develop advanced algorithms capable of detecting face morphing, an issue that will be explored in more detail in the next chapter.

1.2 A case study

A concrete example to elaborate on the concepts outlined above involves a criminal attempting to circumvent facial recognition systems at an airport to board a flight despite a ban against him. In 2002, the human face was adopted as a biometric feature for automated identity verification in machine readable electronic travel documents (eMRTDs[4]).

In any case, the facial image contained in the document to be considered valid must meet strict criteria established by ISO[5].
However, the method by which the image that will be attached to the ID is provided varies by country and can generally be done in two ways:

- direct capture of the applicant's face using a high-resolution camera at the document issuing office $\rightarrow$ highly secure method given the controlled nature of the capture setup, which therefore prevents deliberate alteration of the image
- or submission of an image printed by the citizen himself, which is then attached to the electronic document $\rightarrow$ more vulnerable method, as it allows face morphing techniques to be applied before delivery.

In the second scenario, the possibility of altering the image before its submission introduces a potential risk of fraud[6].

It is important to note that this attack does not compromise the validity of the document itself, but rather aims to fool the control system, whether human or automated.

In addition, unintentional alterations can also occur that, while not the result of an attempt of face morphing, are nonetheless problematic.
Factors such as aesthetic filters, distortions and resizing can alter the structure and proportions

of the face, compromising the biometric identification process[7].

Various researches[8] have analyzed the impact of alterations on the performance of facial recognition systems, showing that the algorithms currently in use adequately handle minimal variations but are vulnerable in the face of more substantial changes.
This can result in an increase in false rejection rate, i.e. the likelihood that the Automatic Border Control (ABC) system will fail to correctly recognize the captured face compared to the face recorded in the eMRTD.

The studies cited above analyze the possibility of attacking an ABC system[9] exploiting facial images obtained by morphing between faces of different people.
During verification, the image captured in real time (e.g. by an airport camera) is compared with the one stored in the electronic passport: if the latter has been altered by morphing, the document can be used by multiple individuals, allowing malicious individuals to exploit the passport of an accomplice or unsuspecting citizen to unduly pass through controls[10].

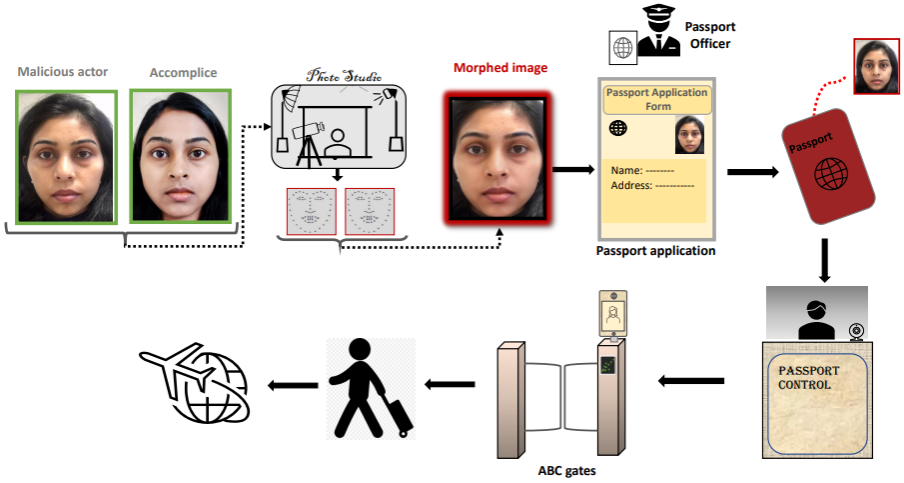


Figure 1.2: example of attack by presentation of a morphed photo during electronic document creation. Source:[10].

Recent studies have investigated both the feasibility of these attacks and the resilience of security systems in the presence of manipulated images[11].
A significant case is the one conducted by the University of Bologna[1], which tested two of the most popular facial recognition software: Neurotechnology VeryLookSDK and Luxand FaceSDK.

This study stands out for its focus on realistic reproduction of the process: in fact, the software settings were calibrated according to the guidelines of FRONTEX[12], European border control agency.
Successively, image pairs of individuals with a slight similarity were selected, avoiding false positives with the default parameters[13].
These images were then combined through a morphing process, described below.

As mentioned, morphing is a technique for gradually transforming one face into another, generating hybrid images of the two subjects.
In the study, this was done using GIMP and its GAP (GIMP Animation Package) plugin, with the goal of creating an artificial face that resembled both starting subjects.
Even though the initial faces had significant differences, it was still possible to obtain a credible image[14].

The morphing process adopted included the following steps:

- manual overlapping of the images based on the position of the eyes
- identification of key points of the face, such as eyes, cheekbones, eyebrows, nose and chin
- automatic generation of a sequence, a transition between the two faces via GAP
- selection of the intermediate frame that was similar for both subjects involved
- possible manual retouching to correct defects or visual artifacts.

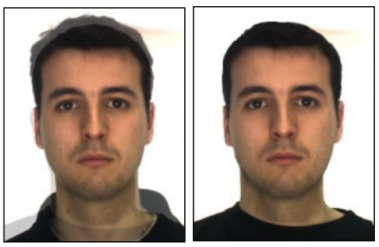


Figure 1.3: example of selected frame before and after manual retouching. Source:[1].

The experiment was conducted on several pairs of men and women, as well as two special cases (a man-woman mix and a combination of three individuals).

The results showed that all tests resulted in a successful attack, demonstrating that facial recognition software cannot detect morphed images with sufficient reliability.

The study's conclusions highlight some key points:

- the possibility of submitting a printed photograph for eMRTD application is a risk factor → it is therefore suggested that only direct image capture at the issuing office be adopted
- the quality of morphed images can be so high as to fool even an experienced human operator
- the blending operation[15] used in morphing processes can generate artifacts near key points of the face, which in some cases reduce the quality of the result.

1.3 Face morping generation

It should be noted that the method described in the previous section represents only one of several ways to generate a morphed image, known as Landmark-based morph generation[16].
In addition to this technique, there are other more advanced and effective ones that exploit Deep Learning[17], specifically neural networks called Generative Adversarial Networks (GANs)[18].

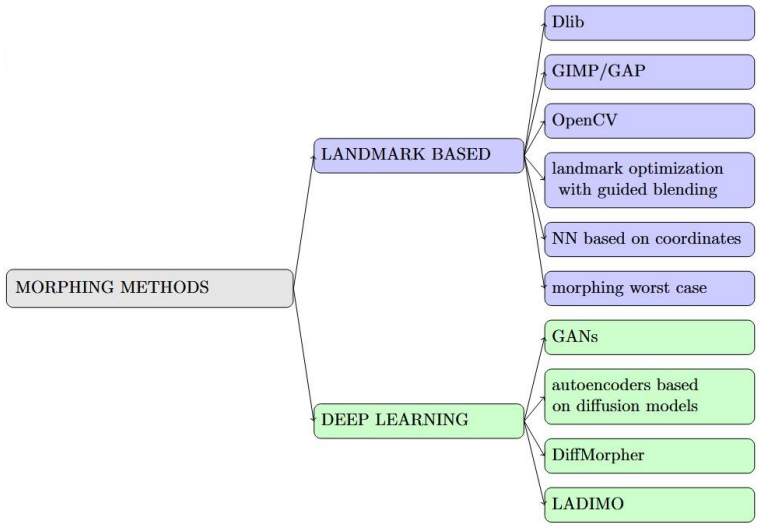


Figure 1.4: naming, classification and subdivision of the main generation techniques for face morphing.

1.3.1 Landmark based morphing

Landmark-based attacks, as mentioned earlier, are based on the detection of key facial features in the nose, eyes and mouth regions.
Currently, there are several software programs that can automatically detect these specific somatic features necessary for morphing two images (e.g. Dlib).
After this detection, the points are overlaid and aligned at an intermediate position.
Among the various techniques used for this operation, it is worth mentioning the Free Form Deformation[19] and the Deformation by Moving Least Squares[20].

The accuracy in identifying and overlapping these points has a direct impact on the quality and credibility of the morphed images generated.
In addition, manual detection, for now, provides more accurate results than automatic localization, which tends to focus exclusively on the central region of the face, neglecting areas such as the forehead and cheekbones.

Once the key points of the two faces have been determined, the morphing process can be mathematically described as follows[21].

Assuming we define I_0 and I_1 as the two starting images and P_0 and P_1 as the respective sets of corresponding points, each frame resulting from morphing is a weighted linear combination of I_0 and I_1:

$$I_\alpha(p) = (1-\alpha) \cdot I_0(\omega_{P_\alpha \to P_0}(p)) + \alpha \cdot I_1(\omega_{P_\alpha \to P_1}(p))$$

where:

- p represents a generic pixel position
- α is the frame weight factor, i.e. the contribution of image I_1 in morphing (e.g. if $\alpha = 0.4$, the morphed image will be composed for 40% by I_1 and for 60% by I_0)
- P_α is the set of corresponding points aligned according to the parameter α
- $\omega_{P_\alpha \to P_\beta}(p)$ is a warp function[22].

During this process, some pixel blocks are replaced, causing misalignment of other pixels, which can generate inconsistent and unrealistic features.
These defects can affect the quality and credibility of the resulting image[23].

The main problems that emerge are:

- an obvious ghost effect around the face, due to the fact that landmarks are placed in the central region of the face, without considering elements such as hair or ears[24]
- more subtle signs of alteration, such as double edges or unnatural iris reflections, resulting from insufficient or inaccurate distribution of facial landmarks.

To correct these defects image enhancement techniques, including image smoothing, image sharpening, edge correction, manual retouching, background substitution and skin color equalization, must be employed.
As mentioned earlier, many libraries and open source software, including GIMP and OpenCV, take advantage of landmark-based morphing.

More recently, as of 2023, there have been further significant improvements in landmark-based morphing techniques.
The following are some of the most relevant contributions:

- landmark optimization and guided blending $\rightarrow$ the paper[76] proposes a new method that improves landmark selection and uses a Graph Convolutional Network type for image fusion $\rightarrow$ the result is a morphing attack that is visually more realistic than the common and more difficult to detect
- neural networks based on coordinates $\rightarrow$ the paper[77] presents a new continuous neural model that avoids discrete interpolation of landmarks, modeling morphing as an implicit function of coordinates $\rightarrow$ this allows smooth deformations and a significant improvement in visual quality, actually producing results that approach those of the most advanced GANs
- morphing worst-case $\rightarrow$ the paper[78] proposes instead a framework that generates morphed images optimized to represent the worst possible cases in an attack (i.e. specially designed, built and optimized to cause the maximum possible recognition error).

1.3.2 Deep Learning based morphing

The second category of technologies that can generate morphed images is those that exploit Deep Learning techniques.
These methods exploit particular neural networks called Generative Adversarial Networks (GANs), which make it possible to create morphed images by combining two input faces within the latent space of the network, that is, from a compressed representation of it.
In general, this technique has some critical issues, including limited similarity to the source faces, as it is complex to maintain their identity.

However, these images do not exhibit the artifacts typical of landmark-based methods, although they may still exhibit distortions typical of GANs[25].
In fact, although GAN-generated morphed images do not require manual post-generation interventions or preliminary landmark-point alignment steps, landmark-based methods continue to ensure superior quality of the produced face.

The GANs work through the interaction between two main components: a generator and a discriminator.
The generator is tasked with producing images that are as similar as possible to those in the training set, while the discriminator trains to distinguish between real and artificially generated images.
During the training process, the generator gradually improves the quality of the images produced to make it increasingly difficult for the discriminator to detect their artificial origin.
In time, this basic architecture has been refined by introducing additional elements, such as encoders, to achieve more advanced performance.

In addition, since 2023, a new technology has proved exploitable for generating robust morphed images: diffusion models.
These approaches have shown promising results in overcoming the limitations of GAN-based methods, especially in terms of visual fidelity and identity preservation.
The following are the main contributions that have emerged in the recent literature in this area:

- autoencoders based on diffusion models[56] $\rightarrow$ this method exploits an autoencoder built on diffusion models to generate high quality morphed images $\rightarrow$ such images are inter-

polated within the latent space of the model, resulting in realistic and visually consistent morphing → such a system also includes an optimized variant (Fast-DiM) that significantly reduces the computational cost

- controlled semantic interpolation → DiffMorpher[57] uses an interpolation strategy between two images by acting simultaneously on latent space, noise and attention → such an approach allows smooth and natural transitions between faces, generating consistent morphing without the need for landmarks or post-processing

- inversion of biometric templates in latent space with LADIMO[58] → this method generates morphed images from biometric templates (embedding) using a latent diffusion model (LDM) → namely, the system rebuild two original faces from a single morphed image and combines them to create an additional high-fidelity morphed variant with more realistic skin textures and improved identity preservation.

Below, the table 1.1 summarizes the main advantages and disadvantages of the two morphed image generation techniques discussed:

TECHNIQUE	BENEFITS	DISADVANTAGES
LANDMARK BASED	Availability of free tools, high quality of generated image, effectiveness in fooling current facial recognition systems, ease of use through automated procedures[26]	It requires manual intervention, necessity of post-generation processing and requires a certain similarity between the subjects chosen[27]
DEEP LEARNING BASED	Doesn't require manual intervention and greater simplicity	It requires complex training and can generate undesirable geometric distortions and necessity of attention in the selection of subjects based on factors such as age, gender and ethnicity

Table 1.1: advantages and disadvantages comparison of landmark based and deep learning based morphing.

Finally, the image below shows a comparison of the main morphing methods, both landmark-based and deep learning:

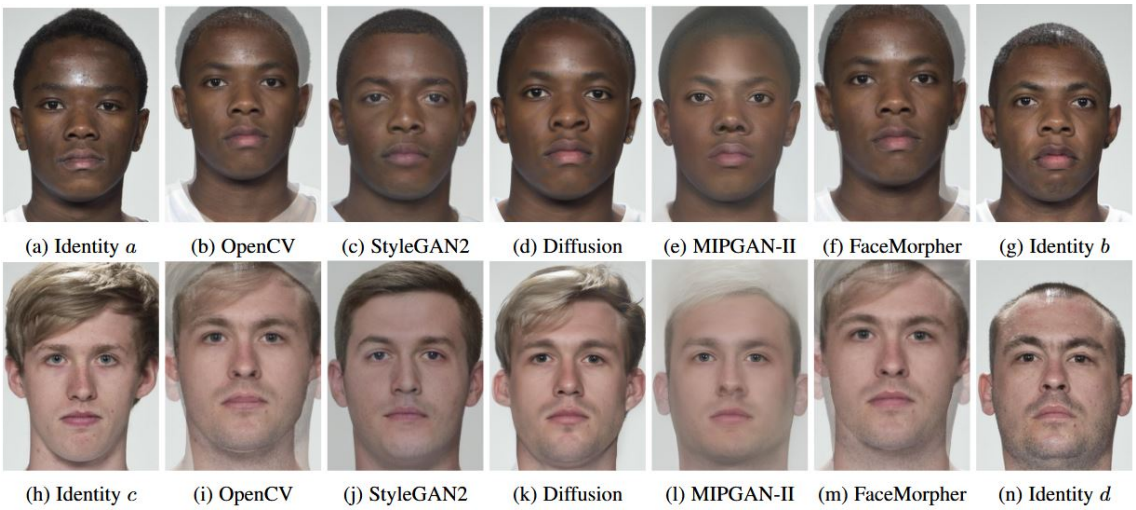


Figure 1.5: comparison between the main morphing generation methods. Source[56].

Chapter 2

FACE MORPHING DETECTION

2.1 Introduction

Morphing Attack Detection (MAD)[28] refers to the process by which we check whether a facial image has been manipulated or altered through morphing techniques.
This method is crucial for attesting the authenticity of images and countering the use of counterfeit photographs for illicit purposes.

As mentioned earlier, the main purpose of a morphing-based attack is to evade facial recognition systems (FRS) present at Automatic Border Control (ABC) checkpoints by presenting an electronic passport (eMRTD) or ID whose portrait has been altered through face morphing techniques.
In this scenario, an attacker could, with the cooperation of an accomplice, circumvent security controls and gain access to areas precluded to him.

However, it is not only border controls that are exposed to this kind of threat: other sectors are also potentially vulnerable, such as health care, finance and law which in fact may experience increased risk from the use of manipulated facial images.

In addition, the accessibility of free software that allows even inexperienced users to generate counterfeit images significantly increases the vulnerability of facial recognition systems.
An additional element that encourages the spread of these practices is the possibility of renewing official documents remotely: in fact, several countries, including the United States, New

Zealand and Ireland, allow passport renewal through the online submission of an updated photograph, which may have been previously altered through morphing techniques.

In response to all these critical issues, several studies have been conducted in recent years[29][30][31], both in academia and industry, some of them funded by government institutions[32], as the European Union.
These researches are aimed at the analysis and development of advanced techniques and algorithms for MAD.
A major contribution in this area has been made by the University of Bologna, which has promoted the SOTAMD project, which represents a benchmark for the evaluation of MAD techniques, thanks to its platform based on a dataset of nearly 6,000 high-quality images, designed to perform comparative analyses under realistic conditions.

2.2 Databases for MAD

Morphing Attack Detection (MAD) techniques benefit from a range of databases, both public and restricted, that support the evaluation of Facial Recognition Systems (FRS) and the effectiveness of MAD algorithms.
Over the past decade, several databases have been introduced to support both Single-image MAD (S-MAD) and Differential MAD (D-MAD), with varying features in terms of image quality, resolution and acquisition pipelines.

One of the earliest morphing-specific databases was developed by the University of Bologna.
It initially contained 14 synthetic morphed images created using landmark-based techniques and has since been expanded to over 80 images, captured in controlled acquisition settings.
Despite its relatively small size, it has been widely adopted in academic research, especially for preliminary evaluations of morphing detection methods.

A significant contribution comes from the Idiap Research Institute, which created a morphing database based on publicly available face datasets such as FRLL and FRGC.
The resulting database includes 400 bona fide and 1200 morphed images, created using both landmark-based and triangulation-based morphing techniques.

It features multiple morphing quality levels and is widely used as a benchmark in the evaluation of MAD systems.

Within the context of the Facial Verification Competition (FVC), multiple MAD evaluation benchmarks have been introduced.
These include dedicated tasks for both S-MAD and D-MAD, each with its own datasets and acquisition protocols.
The FVC platform does not host a single FVC-Morphing Database, but rather organizes periodic competitions with different datasets, such as the SMDD and SYN-MAD datasets, that vary in complexity, image source and morphing techniques.

Similarly, the NIST Face Recognition Vendor Test (FRVT) MORPH, is another evaluation platform that consists of a series of standardized and independent benchmarks.
These benchmarks are based on thousands of real-world images (including passport-style photos) and test MAD systems against a variety of morphing scenarios, such as low-quality, high-quality and automatically generated morphs.
The associated datasets are not publicly accessible but are used under agreement with participating vendors and research institutions.

Despite the increasing number of available datasets, many remain under restricted access due to licensing and privacy concerns.
Nonetheless, publicly available resources, such as the Idiap morphing database and the University of Bologna dataset, remain essential to enable reproducible research in the field of MAD.

The table 2.1 below shows a summary of the databases and benchmarks for morphing detection cited above.

Table 2.1: summary of face morphing detection databases and benchmarks.

DATABASE / BENCHMARK	# BONA FIDE	# MORPHED	NOTES
University of Bologna	–	>80	landmark-based morphing, controlled acquisition, restricted but available for academic research
Idiap-MDB (FRLL + FRGC)	400	1,200	multiple quality levels, morphs generated using landmark and triangulation methods
FVC Benchmarks	varies	varies	multiple benchmark tasks (S-MAD, D-MAD), high-resolution images, standardized evaluation, not a single static DB
MORPH-AID	∼5,000	∼10,000	includes classical and GAN-based morphs, variable lighting, pose and demographic diversity
NIST FRVT / FATE MORPH	–	thousands	ongoing evaluation platform, includes various morph types, datasets not publicly available

2.3 Traditional MAD techniques

As discussed in the previous sections, several studies[33] have pointed out that a human observer, regardless of his level of experience, encounters considerable difficulty in determining whether an image has been altered by morphing techniques.

As a result, it is essential to adopt automatic approaches to deal with the problem of MAD (Morphing Attack Detection)[34].

The current and most common MAD methodologies can be divided into two main categories,

depending on the number of images used in analysis:

- Single-image based MAD (S-MAD)[35] → the detection of the presence of manipulation by morphing is carried out by analyzing and studying only the suspected image
- Differential-image based MAD (D-MAD) → the analysis is based on the study and comparison of a pair of images, where one represents the photo present on an official document (whose authenticity needs to be established), while the other is a reference image acquired in real time and in a trusted environment.

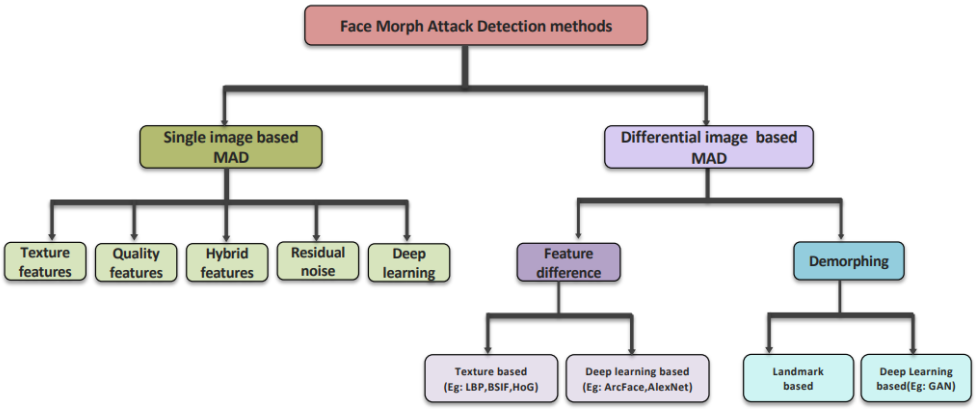


Figure 2.1: classification and subdivision of MAD techniques. Source:[10].

Regardless of the MAD approach adopted, it is essential to consider that facial images may come from different sources, thus presenting variable features, such as resolution and size.
For this reason, before proceeding with the analysis, it is necessary to perform a normalization operation on the images.
From a technical point of view, normalization involves the following operations:

- identifying the position of the center of the eyes and the tip of the nose
- resizing the image so that the distance between the two eyes is 150 pixels
- extracting a sub-image of size 350x400 pixels, centered on the tip of the nose.

2.3.1 S-MAD

S-MAD methodologies are applied in cases where analysis for the detection of alterations must be carried out exclusively on the single image provided to the algorithm, thus without any possibility of comparison with additional reference images.

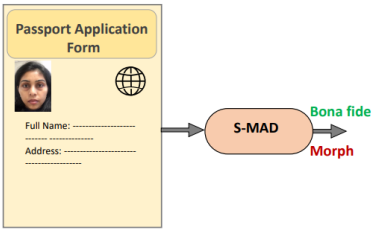


Figure 2.2: in this situation, manipulation detection is done exclusively based on a single image. Source:[10].

These techniques can be further divided into five distinct categories based on the features analyzed:

- S-MAD based on texture features[36] → exploits the extraction and analysis of texture and surface properties of the image → processing algorithms evaluate elements such as gradient distribution, variance and correlation between adjacent pixels to identify any inconsistencies, as previous studies[37] have shown that images altered by morphing exhibit different texture characteristics than authentic images, with JPG compression further emphasizing the anomalies → this methodology is generally fast and accurate
- S-MAD quality-based[38] → employs visual quality parameters, including brightness, sharpness and geometric distortions, to detect potential anomalies → the collected assessments are processed to estimate the level of image degradation and if this exceeds a predefined threshold the algorithm flags a possible manipulation → this approach is generally quick and easy to implement, although it may be less accurate than other methods
- S-MAD based on residual noise[39] → detects any discontinuities in pixels by subtraction between the original image and a noise-filtered version of it, deleting noise such as chromatic aberrations or ghosting → the resulting inconsistencies may reveal alterations not visible to the naked eye
- S-MAD based on deep learning[40] → exploits neural networks previously trained on extended datasets to autonomously detect manipulations in the image → due to the ability to learn complex patterns, this method offers high accuracy in detecting forgeries
- hybrid S-MAD[41] → combines two or more sensing techniques to increase the robustness and reliability of the system → although this strategy improves accuracy, it simultaneously introduces a significant increase in computational cost and complexity.

Compared with D-MAD methodologies, S-MAD techniques are generally more complex and less reliable because they lack a reference image for comparison.
The analysis is therefore based on the assumption that morphing processes leave imperceptible traces in the manipulated image, but identifiable through advanced detection techniques[42].

2.3.2 D-MAD

The goal of the Differential Image Based MAD method, on the other hand, is to determine whether the image under examination has been subject to morphing by comparing it with another photograph of the same individual acquired in a secure environment.
These techniques are particularly useful in the context of border checks described earlier, where the passport image potentially altered by morphing is verified by comparison with a photograph acquired by the ABC system.

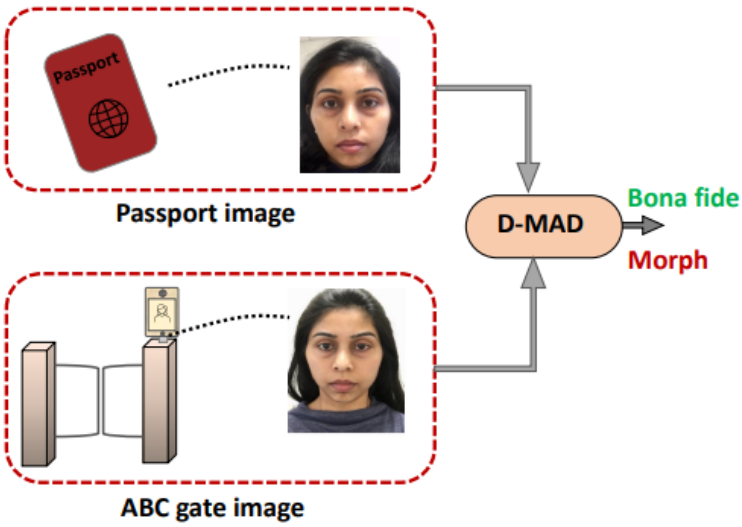


Figure 2.3: in this scenario, you have two images of the subject, which are compared to detect possible morphing attacks. Source:[10].

This approach can also be divided into several categories:

- D-MAD based on feature difference[43] → compares features extracted from the two images (the one presented and the one acquired) to highlight anomalies through several possible techniques (landmark analysis, spectral and texture analysis)
- Demorphing[44] → attempts to reverse morphing process to reconstruct source images, using landmarks or deep learning networks
- Deep Feature Difference Based → uses deep extractors (usually ResNet, ArcFace or MagFace) to obtain embedding, then compares differences and classify with SVM or dedicated networks
- Multispectral MAD[79] → exploits multispectral acquisitions (even up to seven bands) to highlight artifacts not visible in the RGB spectrum
- Fusion Identity with Artifact Analysis[80] → combines identity information (deep embedding) with forensic artifact analysis (texture and lighting) and merges decisions to increase accuracy, especially in cases of similar subjects
- Multimodal LLM based MAD[81] → emerging approach that uses multimodal ChatGPT or Gemini models in zero-shot → has the advantage of providing decisions with natural explanations by exploiting chain-of-thought prompts.

2.3.3 Performance indicators

Numerous studies examining the behavior and performance of the most common Morphing Attack Detection (MAD) techniques are available in academic context.
These studies specifically aim to analyze and evaluate the following aspects:

- level of vulnerability of facial recognition systems
- effectiveness of MAD techniques.

Regarding the first aspect, i.e. vulnerability assessment of the face recognition system (FRS), this focuses on the ability to identify morphed images and verify their association with the original subjects.
In the literature, two main metrics are commonly used to quantify this parameter:

- Mated Morph Presentation Match Rate (MMPMR) measures the fraction of correctly identified morphed images relative to the subjects involved in their generation, according to the formula

$$MMPMR = \frac{1}{M} \cdot \sum_{m=1}^{M} [min(1..N_m) \cdot S_m^n] > \tau$$

where *M* represents the total number of morphed images, N_m the number of subjects involved, S_m^n the degree of similarity between the subject *n* and the morphed image *m*, while τ is the authentication threshold of the FRS $\rightarrow$ this metric assesses the vulnerability of the system by analyzing whether the source subjects are recognized in the same manipulated image

- Fully Mated Morph Presentation Match Rate (FMMPMR) $\rightarrow$ extension of the previous metric, it also takes into account the number of attempts made and the weight of each, according to the formula

$$FMMPMR = \frac{1}{P} \cdot \sum_{M,P} (S1_M^P > \tau) \wedge ... \wedge (Sk_M^P > \tau)$$

where *P* is the total number of authentication attempts with test images, *K* the number of subjects involved, Sk_M^P the degree of compatibility of the subject *k* in an attempt *p* with the morphed image *M* $\rightarrow$ the fact that this formula considers the total number of audits allows a more detailed analysis of FRS security.

As for the evaluation of MAD performance, however, these are measured through metrics standardized by ISO[5].
Since MAD can be modeled as a binary classification problem, the basic metrics are:

- Attack Presentation Classification Match Rate (APCER) $\rightarrow$ measures the percentage of morphing attacks that are incorrectly accepted as authentic images (false negatives)

- Bona Fide Presentation Classification Match Rate (BPCER) $\rightarrow$ evaluates the percentage of authentic images that are misclassified as morphing attacks (false positives).

Algebraically:

$$BPCER(\tau) = \frac{1}{N} \cdot \sum_{i=1}^{N} H(b_i - \tau)$$

$$APCER(\tau) = 1 - \left[\frac{1}{M} \cdot \sum_{i=1}^{M} H(m_i - \tau) \right]$$

where:

- N is the total number of datasets used
- M is the number of models analyzed
- b_i e m_i represent the detection scores
- τ is the decision threshold
- *H(x)* is a step function.

Another significant metric is the Detection Equal-Error-Rate (EER), which represents the point at which BPCER and APCER coincide.
For its visualization and identification, the Detection Error Trade-off (DET) curve which illustrates the trend of BPCER versus APCER, is commonly used in the present thesis:

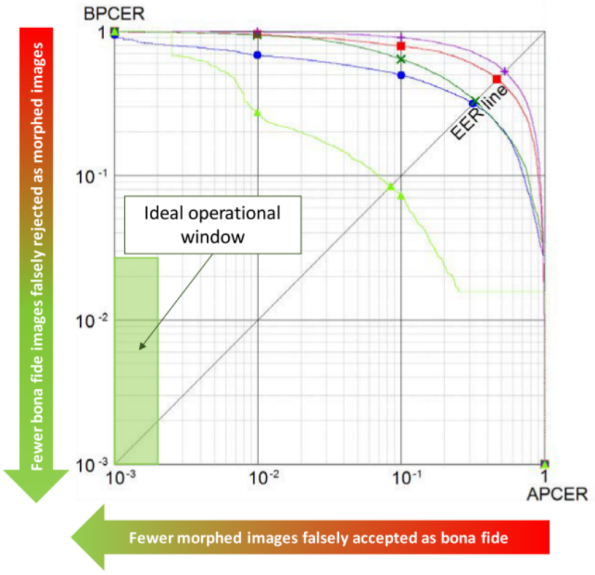


Figure 2.4: EER graph, a lower value indicates a system that is more effective in recognizing morphing attacks. The allowable region is defined by the FRONTEX operational criteria.

In addition to EER, other significant metrics include:

- $BPCER_{0.1} \rightarrow$ value of BPCER when APCER is 10%
- $BPCER_{0.05} \rightarrow$ value of BPCER when APCER is 5%
- $BPCER_{0.01} \rightarrow$ BPCER value when APCER is equal to 1%.

These metrics are critical as FRONTEX has defined minimum operational thresholds that MAD systems must meet to be deemed reliable and functional.
The scientific literature has conducted multiple studies on these metrics, exploiting experimental datasets.
Among the most significant results we find that:

- the gender of the subject affects the likelihood of deception → morphing attacks are 10% more effective in female faces
- the quality of manual edits impacts the success of the attack
- the recognition software reacts differently depending on the morphing algorithm used
- the quality of morphing is a critical factor in the performance of S-MAD and D-MAD
- S-MAD is generally more problematic than D-MAD due to the scarcity of information available
- no current technique yet meets the operational requirements of FRONTEX[45].

The studies also highlight some open challenges, including:

- lack of large-scale public datasets due to legal and privacy constraints
- difficulties in generalizing MAD techniques
- uneven criteria in selecting subjects for morphing
- problems in recognizing twins, look-alikes and individuals with similar traits
- effects of the print-and-scan (P&S) stage, which reduces digital details useful for detecting morphing.

A promising study[29] proposes the automated generation of morphed images to improve the training of neural networks.
The results show that the inclusion of simulated P&S images significantly increases the performance of MAD systems.

2.3.4 More recent approaches

Recent research[46] are focusing attention on overcoming the limitations and constraints imposed by current approaches to MAD.
These obstacles include the limited availability of training data, both in terms of quantity and variety, as well as privacy restrictions that obstruct data sharing and exchange.

A particularly promising line of research explores the use of incremental training for MAD, applied to data held by different research groups.
This approach would avoid direct data sharing, replacing it with transfer of the trained model.
The main problem this methodology aims to solve is the rigidity of current privacy regulations, which impose stringent constraints on the collection, dissemination and sharing of biometric data, particularly facial images.
In this context, progress in developing new algorithms for MAD is obstructed by the fact that each laboratory must independently build, train, test and evaluate its own models using only the data at its disposal.
These restrictions adversely affect the reproducibility and generalization ability of the models, which show a sharp drop in performance when exposed to never seen before data.

Actually, two Deep Learning techniques are emerging as possible solutions to these issues:

- Continual Learning[47] $\rightarrow$ described in detail in the current section
- Federated Learning[48].

Continual Learning focuses on the ability of models to learn new information incrementally, without losing knowledge already acquired[49].
This is a crucial challenge in artificial intelligence, as many models tend to forget prior information when exposed to new data.
This methodology, which is still being studied for its application to MAD, requires that during the training phase each dataset is presented only once.
This promotes incremental learning, making it easier to adapt to new data without having to continuously store or reprocess it.

This feature mitigates the phenomenon known as catastrophic forgetting[50], that is the loss

of previous information as the model is updated with new data.

This happens because most models are optimized to adapt to the latest data by overwriting or deleting previously learned data.

Luckily, the Continual Learning paradigm envisions that training data are not always available at all times, thus facilitating the development of systems that can continuously learn from new information without necessarily having to access historical data.

The typical workflow of this approach is illustrated in the picture below:

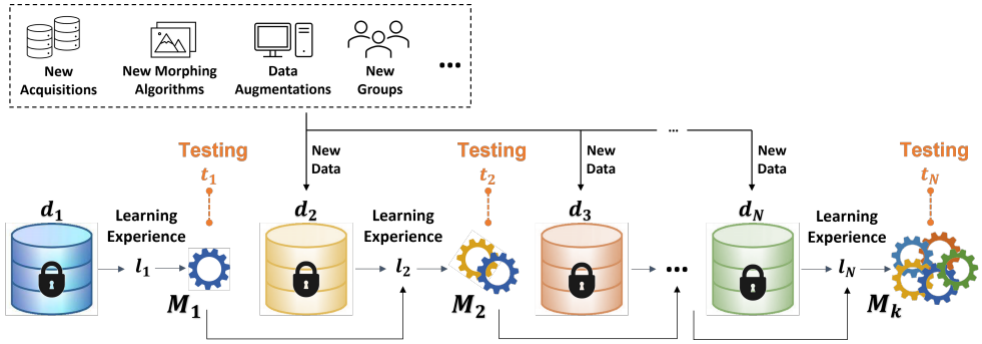


Figure 2.5: the model is trained incrementally on distinct datasets ($d_1..d_N$) and evaluated at different stages of validation ($t_1..t_N$). This training process can be extended without the need to retrain the model from the beginning or transfer data from the previous model. Source:[46].

This strategy allows a research team to train its model internally on its own data and later share it with other groups, allowing them to continue training on their own datasets without the need to transfer the actual data.
This methodology could stimulate the creation of new algorithms for MAD or improve existing ones.

A significant example is provided by the study[46], in which experiments with different levels of Continual Learning were conducted to analyze the behavior and performance of the models in different scenarios:

- Naive $\rightarrow$ traditional classifier training, in which data are always available and the phenomenon of catastrophic forgetting is not counteracted
- Fine tuning $\rightarrow$ the model is initially trained as in the previous case, but then the intermediate part of the neural network is frozen, while the final part (devoted to classification) continues the learning process
- Continual Learning $\rightarrow$ implemented through methods such as Learning Without Forgetting (in which the previous model is also kept available during training) and Elastic Weight Consolidation (which reduces the reliability of the weights attached to past data).

The analyses conducted comparing evaluation metrics in these scenarios showed that Continual Learning could be an effective solution for MAD.
However, further studies are needed to fully explore its potential and possible limitations.

Chapter 3

V-MAD

As mentioned in the previous chapter, current morphing detection techniques mainly focus only on a single image or a pair of them.
In contrast, the innovative alternative technique explored in this thesis is known as V-MAD[51](Video-based Morphing Attack Detection) and is based on video sequences captured by cameras and facial recognition tools used, for example, for airport and border control:

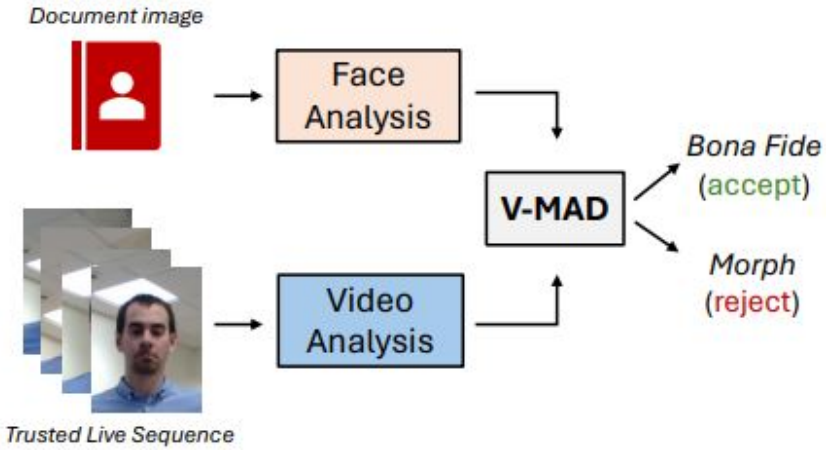


Figure 3.1: the alternative technique known as V-MAD would make it possible to discover whether or not the image on a travel document has undergone a morphing process by comparing it with a video, i.e. a sequence of frames.

Indeed, in the study[51], the authors suggest that the availability of more frames may be an opportunity to improve and harden systems for MAD, in part because of the possibility of discarding suboptimal frames, such as caused by adverse lighting conditions or face occlusions.

3.1 Mathematical formulation of the problem

The identity verification process that takes place at boarding controls involves comparing the photo d in the eMRTD (the electronic travel document) with a video sequence $F = (f_1, .., f_n)$ consisting of n frames captured in real time (e.g. from a camera).
So a V-MAD system $V(d, F)$ should analyze the entire sequence F and compare it with the presented image d to return as a result a single value representing the morphing probability of that image.

While there are currently no further studies of this approach in the literature, the study[51] attempts to investigate the adaptation of D-MAD techniques to the V-MAD task as described in the figure below:

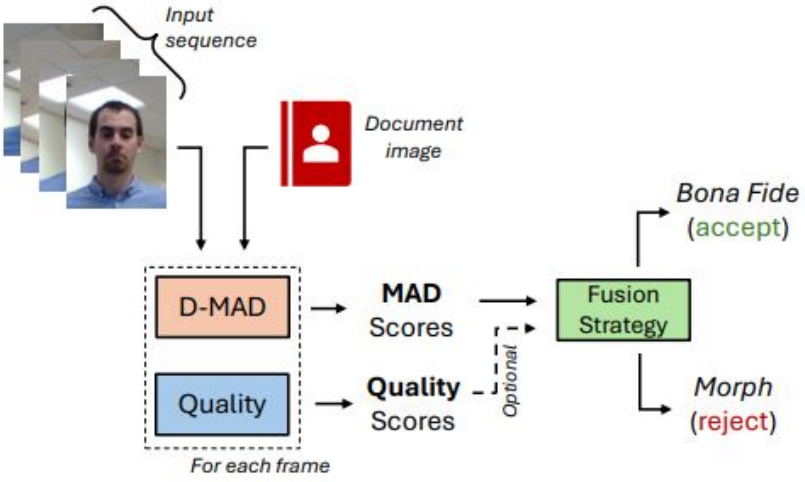


Figure 3.2: a first practical implementation of V-MAD might involve the analysis of each frame of the input video sequence by a quality control tool and a D-MAD model, which also receives the document image as input. Both the output produced by the MAD model and the quality scores would then have to be combined with some kind of aggregation strategy to produce in output the morphing probability.

Hence, in the proposed implementation of V-MAD we have a D-MAD system D capable of computing a morphing score $D(d, f_i)$, understood as the probability that the image d present on the travel document is forged based on its comparison with the frame f_i.
This computation can be repeated for each frame of the sequence F, effectively producing a sequence of scores: $S(d, F) = (D(d, f_i)$ for $i = 1..n)$.

In its simplest formulation, a V-MAD system will combine through a f function the sequence of computed D-MAD scores $S(d, F)$: $V(d, F) = f(S(d, F))$, where $f : R \to R$.

3.2 Aggregation of D-MAD scores

As mentioned in the previous section, an appropriate f function must be chosen to merge the computed scores for each individual frame.

Between the simplest possibilities are worth mentioning:

- the average D-MAD score of the sequence $F \to V(d, F) = \frac{1}{n} \sum_{i=1}^{n} D(d, f_i)$
- the median of the D-MAD scores of the sequence
- a voting system based on D-MAD scores calculated $\to V(d, F) = \frac{1}{n} \sum_{i=1}^{n} m(D(d, f_i))$, where $m(D(d, f_i)) = 1$ if $D(d, f_i) >$ threshold and 0 otherwise.

3.3 Inclusion of aspects of image quality

As anticipated at the beginning of this chapter, important contributions to the V-MAD task can be made by the exploitation and analysis of metrics and scores inherent the quality of each frame composing F.
In this slightly more complex conception, the input to be received by the V-MAD system will consist of two sets of scores, the D-MAD scores of individual frames and the scores related to the quality of those frames: $V(d, F) = f(S_D(d, F), S_Q(F))$.

As for the quality of the image contained on the travel document, this should also be considered, but it is possible to ignore this aspect since images on official documents must pass strict quality controls.

Also in this case, an appropriate f function must then be found to combine the two types of scores, such as:

- a weighted average, where the overall V-MAD score is given by the average of the D-MAD scores of each frame weighted by the corresponding quality score $\rightarrow V(d, F) = \sum_{i=1}^{n} D(d, f_i)Q(f_i)$
- best quality, in which the final V-MAD score is given by the D-MAD score obtained by the frame that achieved the highest quality score $\rightarrow V(d, F) = D(d, f_k)$ with $k = argmax\ Q(F)$.

The study[51] also reports possible algorithms that can be used to obtain the required quality scores, known as FIQA algorithms (Face Image Quality Assessment):

- MagFace[52] proposes loss functions that learn feature embeddings capable of measuring the quality of the face represented $\rightarrow$ has been shown that the magnitude of such embeddings increases consistently for faces having high probability of being recognized
- CR-FIQA[53] estimate the quality of the face represented by learning to predict its classifiability
- SER-FIQ[54] produces quality scores through an unsupervised methodology that relies on the robustness of image embeddings learned through deep learning techniques $\rightarrow$ more in detail, this approach studies the variability of embeddings generated by a subnetwork of a facial model to estimate the robustness of the sample representation and consequently its quality.

In addition to these algorithms, it is worth mentioning that there are also other measures, formally defined in the ISO OFIQ standard[55], which are very accurate and can be used to evaluate image quality, some of which have a significant impact for MAD, such as:

- illumination uniformity $\rightarrow$ measures the illumination difference between the left and right sides of the face by calculating the intersection of the normalized brightness histograms calculated on the left and right regions of the face
- defocus $\rightarrow$ analyzes the sharpness level of the image and returns a score computed as the difference between the face region and the same region obtained after passing an average filter
- pose $\rightarrow$ focuses on analyzing the position and angle of the head.

The following figure shows the scores obtained from both the FIQA algorithms and the standards above:

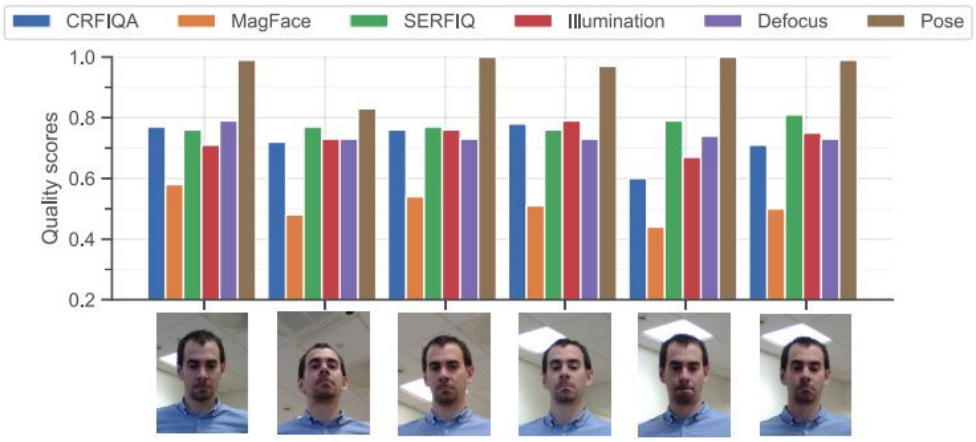


Figure 3.3: remember, however, that while the first three algorithms analyze overall image quality, the last three indicators focus only on limited and certain features.

3.4 Brief overview of how ArcFace works

It is necessary to make a small digression on ArcFace[71], about what it is and how it works, since understanding this tool is essential to understand how it was possible to perform some key operations of face verification and morphing detection (also in the case of V-MAD).
Additionally, it is important to understand how ArcFace works beacuse it has been used in the face detection and embedding extraction phase on many different datasets studied in the present thesis, but this will be explained later on in the following chapters.

ArcFace is in fact now a very useful tool in the area of the so-called deep face recognition, in which deep convolutional neural networks (DCNNs) are used to map face images into a discriminative embedding space.
In such a space, the goal is that features of faces of the same identity are close together (low intra-class distance) and features of faces of different identities are far apart (high inter-class distance).
ArcFace, which is often improperly called a face detector, is actually a loss function, developed to specifically optimize this goal by improving the discriminative ability of face embeddings.

That is, contrary to what one might think, ArcFace is not a neural network per se, but rather is a loss function designed to effectively train an existing generic DCNN (such as ResNet50 or ResNet100) for the task of face recognition.

In fact, the typical structure used with ArcFace consists of a DCNN (ResNet50 and ResNet100 are used in its presentation paper)[45].
Such a network is responsible for taking as input a face image (typically appropriately preprocessed) and passing it through a series of convolutional and pooling layers to produce a feature vector of size 512.

The embedding size is due to the particular terminating structure of the network, consisting of a batch normalization layer, a dropout layer, a fully connected layer and again a batch normalization layer.

As anticipated, the loss function is crucial in guiding the training of DCNN.
Traditionally, the softmax function is used to train simple classifiers for face recognition, but it has the disadvantage that it can learn separable features for closed-set classification problems (identities seen during training), but it is not sufficiently discriminative for open-set face recognition (identities not seen during training) and its size increases linearly with the number of identities.

To address these limitations, recent methods have incorporated an edge into the softmax loss function.
ArcFace builds on this idea and proposes a so-called Additive Angular Margin Loss.

In more detail, the operation of ArcFace is based on L2 normalization of both the embedding features (x_i) and the weights (W_j) of the last fully-connected layer of the network.
Following this normalization, the scalar product between feature x_i and weight W_j becomes equal to the cosine of the angle θ_j between them, multiplied by a scaling factor s (the fixed norm of features):

$$W_j^T x_i = \|W_j\| \|x_i\| \cos\theta_j = 1 \cdot s \cdot \cos\theta_j = s\cos\theta_j$$

This normalization therefore pushes the embedding features to be distributed over a hypersphere of radius s.

The softmax cost function modified by ArcFace can be formally defined as:

$$L_3 = -\log \frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{N} e^{s\cos\theta_j}}$$

where:

- $x_i \in R^d$ is the feature (embedding) of the i-th sample
- y_i is the label of the class to which sample i belongs
- $W_j \in R^d$ is the j-th column of the weight matrix W, representing the center of class j
- N is the number of classes (identity)
- θ_j is the angle between feature x_i and the center of class W_j
- θ_{y_i} is the angle between feature x_i and the center of its corrected class W_{y_i}
- m is the Additive Angular Margin
- s is the feature scale factor, typically set to 64.

The crucial part of ArcFace is the addition of the angular margin m to the correct class angle (θ_{y_i}), thus obtaining $\cos(\theta_{y_i} + m)$.
This simple addition of the additive angular margin has a clear geometric interpretation in that it directly optimizes the margin over the geodesic distance (distance measured along the curved surface of the sphere between two points and not along a straight line in three-dimensional space) in the normalized hypersphere (in other words it makes it more difficult for the network to correctly classify a sample of class y_i, by forcing the feature x_i to be even closer, i.e. at a smaller angle, to the center of the correct class W_{y_i} than would be required by traditional softmax).
This increases intra-class compactness and inter-class discrepancy.

In fact, the idea of adding a margin to the loss function is not novel, but compared with other loss with margin ArcFace has a constant linear angular margin along the entire angular range, which corresponds exactly to the geodesic distance.
It has subsequently been shown that this leads to more stable training and superior performance.

It is important to emphasize again that ArcFace, being a loss function, is not directly concerned

with detecting faces within a photo.
The detection of faces is one of many preprocessing steps that occur before the image is given as input to the DCNN trained with ArcFace.
In particular, in the original paper the authors mention using RetinaFace to detect faces through their five distinctive facial points, which are then used to generate normalized face crops with size 112x112.
It is precisely on such crops that the DCNN trained with ArcFace is applied to extract the embedding.
Then, the typical process of using ArcFace is as follows:

- an image (which may contain one or more faces) is fed to a face detector such as RetinaFace for example
- such a detector identifies the position of each face and key facial points (eyes, nose and mouth)
- using these points each detected face is aligned (normalized for pose and size) and cropped
- the cropped and normalized face image is given as input to the DCNN trained with ArcFace loss
- finally the DCNN produces the embedding vector for that face.

The resulting embedding vector is the result of the mapping of the face image by the DCNN.
During training with ArcFace this mapping is optimized so that the resulting vectors in the embedding space have the desired properties for face recognition, that is:

- intraclass compactness $\rightarrow$ the vectors of faces of the same identity are driven to be close together (small angle/low geodesic distance in the hypersphere)
- interclass discrepancy $\rightarrow$ the vectors of faces of different identities are driven to be far apart (wide angle/high geodesic distance in the hypersphere).

Embedding, being normalized, is on a hypersphere of radius s.
The similarity between two embedding vectors of different faces can therefore easily often be measured (which will be done for multiple purposes in our GazeWay dataset) by the cosine similarity, which is directly related to the angle between the vectors on the hypersphere: a high cosine similarity (small angle) indicates that the faces are probably of the same identity, while

a low cosine similarity (large angle) indicates that they are probably of different identities.
The embedding vector represents the discriminative features of the face extracted from the DCNN.
It does not represent pixels or direct image information in the traditional sense, but rather an abstract encoding of the facial features that were deemed important by the model to distinguish different identities during ArcFace-guided training.
The discriminative nature of the model is confirmed by its ability to achieve high performance on facial recognition benchmarks.

In addition, the original paper explores the ability of the ArcFace-trained model to address the reverse problem of mapping embeddings to facial images.
Without training additional generators or discriminators (which is done in GANs), the pre-trained ArcFace model can generate identity-preserving facial images simply by using network gradients and statistical priors stored in Batch Normalization (BN) layers.
It demonstrates that the ArcFace model is not only discriminative but also has a remarkable generative capability, encoding substantial information about the distribution of faces:

Figure 3.4: examples of face generation showing how ArcFace can rebuild realistic faces from model parameters alone.

Chapter 4

DATASETS FOR V-MAD

As anticipated in previous chapters, the idea of V-MAD is very innovative and very recent, so it is still very difficult to find datasets suitable for this task.
In order to overcome this obstacle, this chapter provides a detailed description of the properties that the structure of a usable dataset for V-MAD should have and what are two probable datasets designed for other tasks but which could still be suitable for morphing detection by video.

4.1 Ideal structure of a dataset for V-MAD

Although, as mentioned in the introduction of this chapter, as yet there are no datasets dedicated to V-MAD, the following are some features that an ideal dataset of this type should inevitably possess.
The purpose of the following descriptions is twofold: not only to provide guidelines for the dataset we have collected, but also to help identify datasets created for other purposes that could be adapted to face morphing via video.

An ideal dataset for training and evaluating video-based face morphing techniques should indeed be carefully designed to reflect both the natural variability of the subjects and the actual conditions of use.
The desirable features that such a dataset should possess are therefore outlined below:

- ICAO-compliant frontal photographs → each subject should be represented by at least one frontal photograph that complies with ICAO standards, i.e. possessing all the quality requirements imposed by the International Civil Aviation Organization → this implies

adherence to stringent requirements relating to uniform illumination neutrality of facial expression, facial position and image quality, so as to realistically simulate the images used in official documents

- multiple videos per subject → for each individual the dataset should include several videos recorded in different sessions, so as to reflect some natural variations of the subject over time such as different facial poses and movements, different lighting conditions (natural, artificial, backlighting), slight changes in appearance (haircut, beard, glasses, clothing, makeup) and different backgrounds and environmental contexts

- consistent and accurate labeling → all photographs and video frames related to the same subject should share a unique identifier, so as to allow proper association during the morphing process and to facilitate evaluation of the robustness of the detection techniques

- high diversity and variability of the sample → the dataset should include a large number of subjects with significant variation in age, sex, ethnicity and somatic features, so that models trained on the dataset can generalize effectively to real populations without presenting bias

- accessibility and reusability → the dataset should be readily available and usable by the scientific community, preferably under a license that allows use for research purposes → it is also desirable that data be provided in a standard format (e.g. images in PNG or JPG and video in MP4 or already split into frames) accompanied by structured metadata files (e.g. CSV or JSON) containing labels, demographic information and details of capture sessions or already labeled at the filename level

- quality and resolution → images and videos should have sufficient resolution to allow proper extraction of facial biometric features without excessive compression artifacts.

Meeting the above requirements would allow not only realistic simulation of morphing attack scenarios but also validation of countermeasures under diverse and complex conditions, reflecting the challenges of real-world biometric applications.

4.2 ChokePoint Dataset

As anticipated, one of the datasets that possesses a structure similar to that ideal for the V-MAD task (but originally designed for something else) is called ChokePoint[59] and includes a set of photos and videos proposed specifically for research in the field of facial recognition in real-world, non-cooperative scenarios.
This dataset is now a significant resource for evaluating facial recognition algorithms in video surveillance contexts.

4.2.1 ChokePoint dataset structure

The dataset was acquired in a controlled but realistic environment, simulating a video surveillance scenario in a corridor, with the purpose of filming individuals passing through a bottleneck (hence the name chokepoint, bottleneck precisely) where face shots are most likely to be obtained.
The acquisition setup included:

- 3 different corridors, named Portal 1, Portal 2 and Portal 3
- 4 or 5 cameras for each corridor, arranged to capture the subjects from different angles
- 48 video sequences, each depicting almost all the subjects, for a total of more than 64000 video frames
- 29 subjects traversing the portals in natural conditions, simulating realistic behavior
- video captured at 30 FPS with a resolution of 704x576 pixels.

Figure 4.1: example of camera setup and placement in a gateway.

4.2.2 Labeling and annotations

Images and video frames in the dataset are accompanied by detailed annotations that facilitate their use for face recognition and quality assessment.
In particular, each frame is labeled with:

- ID of the depicted subject → a unique identifier associates the face with the person appearing in the frame
- location of the face → for each detected face the corresponding bounding box is also provided
- facial alignment → the images have already been processed to obtain aligned faces, facilitating the feature extraction phase
- camera and portal IDs → each image is associated with the video source from which it was extracted
- temporal information → each frame is labeled with its temporal position within the video sequence, so as to preserve the chronological order of acquisition.

Annotations make it easy to construct face tracks, i.e. temporal sequences of faces belonging to the same individual, which are crucial for video-based techniques and automatic selection of high-quality frames.

4.2.3 Content and features

ChokePoint has some basic features that make it particularly interesting for scientific research, such as:

- variations in pose, illumination and face quality between consecutive frames
- presence of blur, partial occlusions and expression variations
- heterogeneity of subjects' trajectories in corridors
- a context similar to the real one, but with controlled conditions for experimental validation.

Figure 4.2: examples of heterogeneity in both the backgrounds and lighting conditions of different video sequences.

4.2.4 Utility for the V-MAD task

The ChokePoint dataset is quite suitable for generic face recognition task since:

- offers multi-angle video sequences from which it is possible to analyze how face quality changes over time
- allows to easily exploit face quality assessment techniques (extensively described in the previous chapter) and to select the most suitable frames
- the realistic but controlled environment allows one to evaluate the effectiveness of the system in scenarios and applications not too different from real ones.

Unfortunately, however, it cannot be used in tests regarding specifically V-MAD because it has really important defects: it does not include ICAO-compliant frontal photos of subjects and also has a very limited number of subjects (only 29).

4.3 PASC Dataset

An additional dataset that has a very similar structure to that identified as ideal for V-MAD turns out to be the so-called Point-and-Shoot Face Recognition Challenge (PASC) dataset[60], developed in 2013 to address facial recognition challenges in realistic scenarios using digital cameras.

4.3.1 Structure of the dataset

The PASC dataset includes:

- 9376 images of 293 subjects
- 2802 videos of the same 293 subjects.

The images are balanced and heterogeneous with respect to several factors such as:

- distance from the camera → images acquired from both near and far
- use of different cameras
- both frontal and non-frontal views
- different locations → acquisitions were made in nine different environments.

4.3.2 Content and features

The PASC dataset also has some key features that make it particularly interesting for research, many of which it also shares with the ChokePoint dataset:

- natural variations in pose, illumination and image quality
- uncontrolled conditions to reflect realistic scenarios
- detailed annotations and labeling.

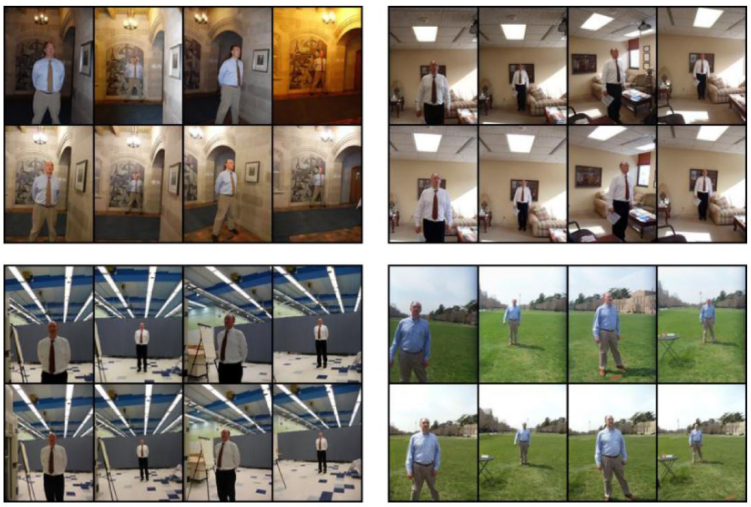


Figure 4.3: examples of images from the PaSC dataset showing variations in pose and distance from the camera.

4.3.3 Utility for the V-MAD task

The PASC dataset does not deviate much from the ideal structure for V-MAD identified earlier, as it has:

- high diversity and variability of data
- realistic scenarios
- complete and accurate annotations and labeling.

Figure 4.4: examples of frames extracted from videos in the PaSC dataset highlighting variations in lighting, angle and background.

However, exactly as the ChokePoint, the PASC dataset also has major flaws and defects that make it unusable for the V-MAD task: it does not include ICAO-compliant frontal photos of the subjects depicted in the videos and it is extremely complicated to obtain access to the dataset.

4.4 Comparison of ChokpePoint, PASC and other datasets explored for V-MAD

The following table 4.1 summarizes and reports relevant features and notes regarding the analyzed datasets, including ChokePoint and PASC, for the V-MAD task.
Defects that make them unattractive for this task are also highlighted:

FEATURE	CHOKE POINT [59]	PASC [60]	MBGC V2 [61]	YOU TUBE FACES [62]	IJB-A [63]	UMD FACES [64]
NUMBER OF SUBJECTS	29	293	?	1595	500	8277 in photos + 3100 in videos
NUMBER OF IMAGES	64204	9376	?	0	5712	367888
VIDEO NUMBER	48	2802	?	3425	2085	22000
ICAO COMPLIANT PHOTOS	no	no	no	no	no	no
MUTABLE VIDEO CONDITIONS	yes	yes	yes	no	yes	yes
NOTES	limited number of sub-jects	difficult to down-load	non-trivial use	very com-pli-cated to obtain	absence of frontal images	non-ICAO com-pliant photos, down-load un-avail-able

Table 4.1: comparison between the features of the different dataset analyzed.

As the table shows, there are currently no datasets in the literature that are totally suitable for the V-MAD task, particularly because no dataset simultaneously offers an adequate number of subjects, their frontal photos meeting ICAO standards and their videos under fairly changing conditions.

To solve this problem, a considerable part of the work of this thesis involved the collection and creation of an internal dataset at the University of Bologna, the obtaining and features of which are described in detail in the following chapter.

4.5 GazeWay: our dataset

To make up for the lack of a dataset that met the ideal requirements for the V-MAD task (which were identified and reported in the previous section), a substantial part of the work in this thesis involved the collection, organization, structuring, verification and labeling of an internal dataset at the University of Bologna, hereafter called Gazeway.

In more detail, the dataset is composed by a total of 325 morphed photos and, for each of the 65 subjects involved:

- a pair of ICAO compliant 3024x5376 frontal photos
- a pair of non-ICAO compliant 3024x5376 frontal photos
- a total of 6 video sequences at 30 FPS, decomposed into frames of size 1280x720, acquired in 2 different environments and for each of these two environments taking three different poses (looking frontally toward the camera, looking around or looking around with some accessory partially occluding the face).

Figure 4.5: ICAO compliant (on the left) and non-ICAO compliant (on the right) photos of subject ID065.

(a) frontal gaze (b) looking around (c) looking around with occlusion

Figure 4.6: video frame of subject ID065 in sequence/environment 01.

(a) frontal gaze (b) looking around (c) looking around with occlusion

Figure 4.7: video frame of subject ID065 in sequence/environment 02.

The fact of having filmed the same subject in two different environments allows for some variability in terms of brightness, background and shadows to be introduced into the dataset. In fact, the first environment (labeled in the dataset as sequence 01) represents the entrance to an office that has a mixture of both artificial and natural lighting (given by the combination of the office lights and a window located to the left of the faces), while on the other hand, the second environment (labeled in the dataset as sequence 02) sees the subject moving into an underground parking lot, so the image is much darker and the available lighting is solely artificial.

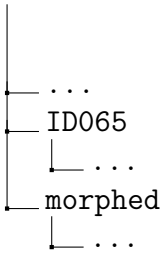
High variability, as mentioned, was introduced by asking the subject to move toward the camera in three different modes:

- with a fixed gaze toward it
- looking around, taking care to rotate the head both vertically and horizontally
- looking around but with an accessory (randomly chosen from time to time between scarf, cap or both) partially occluding the face.

4.5.1 Structure of the dataset

This section shows the structure and organization of the dataset at the level of individual folders and files:

```
/
├── ID001
│   ├── ICAO_photo
│   │   └── img.jpg
│   ├── non_ICAO_frontal_photo
│   │   ├── img1.jpg
│   │   └── img2.jpg
│   └── video_sequences
│       ├── sequence_01
│       │   ├── frontal_gaze
│       │   │   ├── depth
│       │   │   └── rgb
│       │   ├── looking_around
│       │   │   └── ...
│       │   └── looking_around_with_occlusion
│       │       └── ...
│       └── sequence_02
│           └── ...
└── ID002
    └── ...
```



As can be seen, therefore, as far as video sequences are concerned, images are available both in the classic RGB and depth format, in which the value of individual pixels is an indicator of the detection distance.

In total, the weight of the dataset is around 103GB.

4.5.2 Acquisition equipment

Two different cameras were used to capture the photos and videos that make up the dataset:

- Nokia Lumia 930 → for the frontal photos (both ICAO and non-ICAO)
- Intel D435i → for the video sequences.

The following subsections provide the technical specifications of both cameras used.

Nokia Lumia 930

The camera on the Nokia Lumia 930 smartphone model, which turns out to be a rather old smartphone, but still has a high-performance camera, was exploited for capturing the frontal photos.
The following table 4.2 summarizes some of the relevant technical features of the camera it incorporates:

RESOLUTION	20 MP
OPTICAL STABILIZATION	yes
FLASH	dual-LED
VIDEO RECORDING	1080p at 30 FPS
FRONT CAMERA	1.2 MP and 720p video

Table 4.2: technical features of the camera into Nokia Lumia 930 smartphone.

Intel D435i

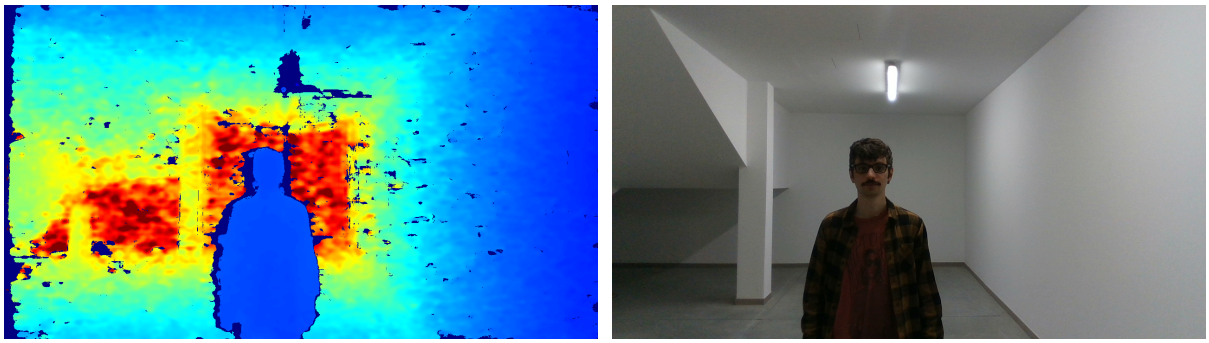
The Intel RealSense D435i is a stereo depth camera very common in robotics, deep learning, arctic vision and augmented reality applications:

Some relevant features of this camera are shown in the table 4.3 below:

MODEL NAME	Intel RealSense Depth Camera D435i
TYPE OF DEPTH	stereo depth
DEPTH RESOLUTION	up to 1280x720 and 30 FPS
DEPTH FREQUENCY	up to 90 FPS (at lower resolutions, like 640x360)
FIELD OF VIEW (FOV)	85.2° horizontal, 58° vertical, 94° diagonal
OPERATING RANGE	0.1m to 10m
IMU INTEGRATED	yes
RGB RESOLUTION	up to 1920x1080 and 30 FPS
CONNECTION INTERFACE	USB 3.1 Gen 1 Type-C
DIMENSIONS	90mm x 25mm x 25mm
WEIGHT	72 grams
API SUPPORTED	Intel RealSense SDK 2.0 (C++, Python, etc.)
POWER SUPPLY	via USB
ASSEMBLY	standard tripod thread, mounting clip

Table 4.3: technical features of the Intel RealSense Depth Camera D435i.

As mentioned, the camera described in the present section turns out to be a depth camera, so it is capable of filming the same scene and encoding it in both the classical RGB format and the depth format, in which each pixel corresponds to a distance value (pixels tending to blue represent regions closest to the camera, while pixels tending to red represent regions furthest from the camera):



(a) depth (b) RGB

Figure 4.8: encoding of the same pixel in the two formats.

4.5.3 ICAO compliance verification

As stated in the previous sections, the dataset contains three frontal photos for each subject: one ICAO compliant and two non-ICAO compliant.
To check the compliance of a photo against the ISO requirements expressed in ISO/IEC 19794-5[65] a combination of three tools was exploited:

- BioLab-ICAO-Check Tool[66] → software developed by the University of Bologna
- icaonet[67] → based on deep learning techniques
- BioGaze → uses deep learning and computer vision techniques to verify where the subject is looking.

The tool that was used the most was BioLab-ICAO-Check Tool, which takes a photo as input and performs the tests reported in table 4.4:

TEST CODE	DESCRIPTION
1	Eye Location Accuracy
2	Blurred
3	Looking Away
4	Ink Marked/Creased
5	Unnatural Skin Tone
6	Too Dark/Light
7	Washed Out
8	Pixelation
9	Hair Across Eyes
10	Eyes Closed
11	Varied Background
12	Roll/Pitch/Yaw Greater than a predefined threshold
13	Flash Reflection on Skin
14	Red Eyes
15	Shadows Behind Head
16	Shadows Across Face
17	Dark Tinted Lenses
18	Flash Reflection on Lenses
19	Frames too Heavy
20	Frame Covering Eyes
21	Hat/Cap
22	Veil over Face
23	Mouth Open
24	Presence of Other Faces or Toys too Close to Face

Table 4.4: tests verified by the BioLab-ICAO-Check tool.

The algorithm returns as output a string showing the scores obtained from the photo provided as input against the above tests.

Each test is considered passed if it scored greater than or equal to a given minimum threshold given in the table 4.5 below:

TEST CODE	THRESHOLD
2	4
3	64
4	99
5	81
6	70
7	56
8	10
9	75
10	100
11	99
12	100
13	77
14	39
15	96
16	86
17	28
18	43
19	33
20	35
21	62
22	66
23	100
24	86

Table 4.5: minimum acceptance thresholds related to each test performed by the BioLab-Icao-Check tool.

The paper[66] produced by professors and researchers at the University of Bologna contains a chapter describing in more detail how photo compliance verification algorithms work.
The following bulleted list briefly summarizes the algorithms used to identify relevant facial features and to perform the verifications reported above:

- face position → the bounding box enclosing the face is obtained by combining algorithms proposed in [68] and [69]
- eye position → starting from the rectangle containing the face, the coordinates of the centers of the two pupils are detected (in case of partial or total occlusions, such as hair or dark glasses covering the eye, a probabilistic estimate of its position is returned)
- position of the nose → based on the rectangle containing the face and the coordinates of the eyes, first a rectangular region is preselected in which it is highly probable to locate the nose and then rigid template matching is exploited to precisely locate the position of the nose
- position of the mouth
- presence of glasses
- segmentation → precise knowledge of the shape and position of the components of the face (such as face, hair, clothing and background) is obtained by means of a multiclassifier that studies the colors and textures of the different regions that make up the image
- ICAO verification 2 (blurred) → the level of blurring of the facial region is measured by calculating a metric known as TSI (Top Sharpening Index)
- ICAO verification 3 (looking away) → this method is based on the assumption that if the pupils are not are exactly in the center of the eyes then the subject is looking not ahead → it is also assumed that the subject is not looking ahead if the ICAO 12 (Roll/Pitch/Yaw Greater than a predefined threshold) test is not passed → otherwise, the score is calculated inversely with respect to the distance between the center of the eyes and the position of the pupils
- verification ICAO 4 (ink marked/creased) → different studies have shown that the YCbCr space is particularly useful for discovering ink marks and creases, even in the background

- verification ICAO 5 (unnnatural skin tone) → again, bringing the photo into the YCbCr color space helps to separate natural skin tones from possible alterations → the score in this case is computed as the percentage of face pixels having a natural color tone
- verification ICAO 6 (too dark/light) → the color histogram of the image is analyzed since usually the color histogram of an image having normal illumination presents an approximately uniform and scattered distribution over all possible gray levels, while if the photo is too dark or too light the distribution will be concentrated only in the lower or upper part of the histogram
- verification ICAO 7 (washed out) → washed out images are detectable since they present, in the image converted to grayscale, a reduced dynamic range compared to the normal
- verification ICAO 8 (pixelation) → such an effect can be revealed by a combination of Prewitt's algorithms and the Hough transform, which can show abundant presence of perfectly vertical and/or horizontal edges in small regions → the score of this test is inversely proportional to the number of perfectly vertical and horizontal lines detected →

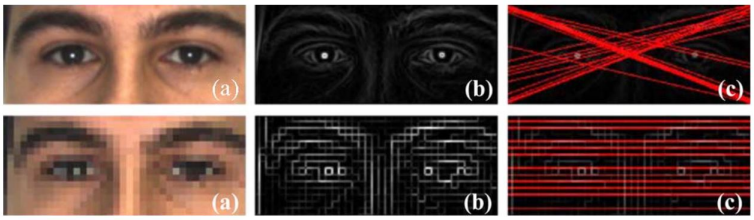


Figure 4.9: straight lines identified in a conforming (top) and nonconforming (bottom) photo.

- verification ICAO 9 (hair across eyes) → the score for this test is inversely proportional to the percentage of pixels near the eye regions classified as hair in the segmentation step
- verification ICAO 10 (eyes closed) → this test is evaluated by computing the percentage of pixels in the eye region that possess a natural and acceptable color for a sclera (white part of the eye)
- verification ICAO 11 (varied background) → the score of this test is inversely proportional to the number of pixels that correspond to edge and that are detected in the region labeled as background in the segmentation step

- verification ICAO 12 (roll/pitch/yaw greater than 8°) → this test is composed of three submodules, each of which estimates the rotation of the face in one of three dimensions → the first module estimates the roll angle by computing the deviation between the x-axis and the line connecting the two pupils → the second module estimates the pitch angle by studying the distance between the center of the eyes and the tip of the nose, the distance between the center of the eyes and the top of the mouth and the distance between the tip of the nose and the top of the mouth → finally the third module estimates the yaw angle and all these three estimates are merged together to produce a score between 0 and 100
- verification ICAO 13 (flash reflection on skin) → the flash reflection on skin is very evident in the images obtained as a difference between the saturation and red-color channels
- verification ICAO 14 (red eyes) → to check for the presence of red eyes, it is sufficient to identify the iris region and compare the color of that region with a threshold to classify its color as natural or red
- verification ICAO 15 (shadows behind head) → analysis of some training images revealed that the generic X channel of the generic XYZ color space highlights the presence of shadows behind the head well if the background is uniform, so the score for this test is inversely proportional to the ratio of the sum of the channels of the first component of the color space to the number of pixels that make up the background
- verification ICAO 16 (shadows across face) → in this other case, the presence of shadows across the face are instead well evident in the generic Z channel of the generic XYZ color space, so the score for this test is computed as the percentage of pixels belonging to the face region that are not 0 in the version of the image binarized according to an appropriate threshold on the Z channel
- verification ICAO 17 (dark tinted lenses) → be M and B respectively a rectangular mask covering the eye region and the binarized image of the differences between the generic Y and Z channels of the generic XYZ color space, then the score of this test is obtained by multiplying the probability that the subject is wearing glasses with the percentage of nonoccluded pixels (i.e. equal to 0 in B) in the M region
- verification ICAO 18 (flash reflection on lenses) → the score of this test is computed by multiplying the probability that the subject is wearing glasses with a score obtained by a

special algorithm based on the color segmentation of eye regions →

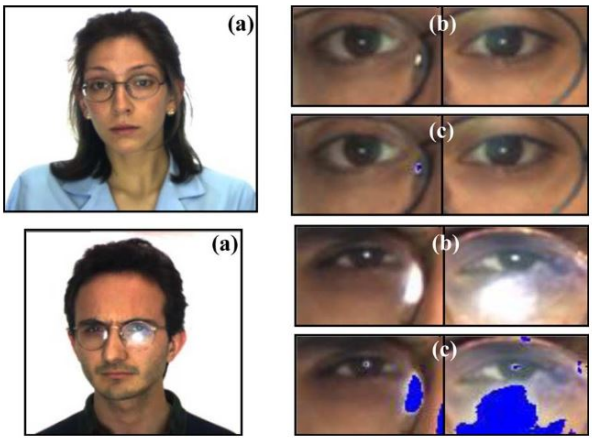


Figure 4.10: results produced by the RGB color segmentation algorithm on the eye regions in a conforming (top) and nonconforming (bottom) image.

- verification ICAO 19 (frames too heavy) → this method exploits the verifications performed for test 5 (unnatural skin tone), in that the score is computed by multiplying the probability that the subject wears glasses with the percentage of pixels having an unnatural tone in the rectangular face region enclosing both eyes
- verification ICAO 20 (frames covering eyes) → similarly to verification 8 (pixelation), this test also exploits Hough and Prewitt's transfromate methods to compute the product between the probability of the subject wearing glasses and the number of non-horizontal lines detected
- verification ICAO 21 (hat/cap) → given the rectangular region of the top of the head and the binarized B channel derived from the color-corrected image, the score for this test is proportional to the percentage of nonoccluded pixels (i.e. having value 0 in B) in that region
- verification ICAO 22 (veil over face) → is quite similar to the method proposed in the previous point but the region does not cover the top of the head but the bottom →

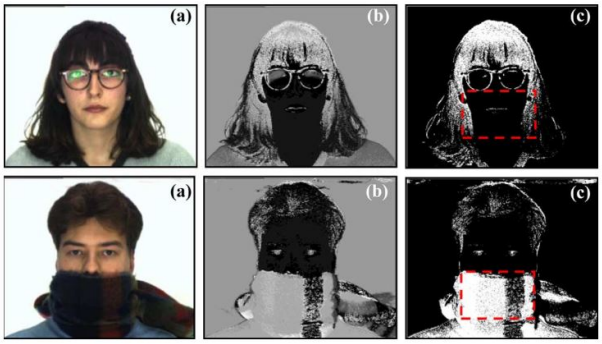


Figure 4.11: results produced by the algorithm on a conforming (top) and nonconforming (bottom) image.

- verification ICAO 23 (mouth open) → the score for this test is computed as a weighted sum of the height of the mouth and the presence of teeth
- verification ICAO 24 (presence of other faces or toys too close to face) → the idea behind this test is that the presence of objects close to the face is revealed by regions in the background characterized by colors that differ from the background itself, whereby the image is partitioned into regions having homogeneous colors, so that the score is inversely proportional to the number of regions found in the background area →

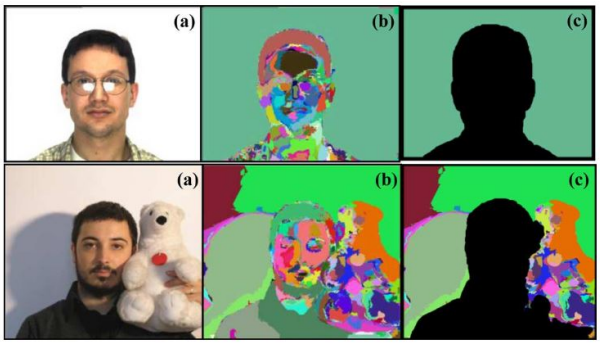


Figure 4.12: results produced by the algorithm on a conforming (top) and nonconforming (bottom) image.

4.5.4 Naming convention

Regarding the naming convention used for the images in the dataset, it respects the format established in the table 4.6 below:

FIELD	DESCRIPTION	VALUES
xxx	subject identifier	001–999
gg	gender	GM → male, GF → female
eee	ethnicity/apperance	EEA → European/American EAF → African EIA → Indian-Asian EAS → East-Asian EME → Middle Eastern
s	acquisition site	B → Belgium (LPA) L → Belgium (Leuven) O → Norway P → Portugal R → Greece U → Germany I → Italy
aaa	age	A00 → unknown, A01–A99 years
ttt	glasses	T00 → none, T10 → glasses
li	type	LG → image from the gate LP → image from the passport LI → image ICAO compliant LV → video from the gate
ddd	device	D00 → Unknown D01 → COGNITEC D02 → IDEMIA D03 → VISIONBOX D04 → Intel D435i D05 → Nokia Lumia 930
ss	session number	S01–S99
cc	camera	C1–C9
inn	image/video number	I001–I999
fff	image/video format	F00 → digital, F99 → unknown

Table 4.6: naming convention format.

Therefore, considering for example the image below:

it will be labeled as 065–GM–EEA–I–A28–T10–LV–D04–S02–C1–I000403–F00, since:

- 065 → the id of the subject represented is 065
- GM → the subject represented is male
- EEA → the subject represented has European/American ethnicity
- I → the frame was acquired in Italy
- A28 → the depicted subject is 28 years old
- T10 → the subject depicted in the frame is wearing glasses
- LV → the frame belongs to a video sequence collected at a gate
- D04 → the frame was acquired from device 04, i.e. from an Intel D435i camera
- S02 → the frame belongs to the second video acquisition session
- C01 → the frame was acquired from the first (and only) camera
- I000403 → the id of that frame within the video sequence to which it belongs is 000403
- F00 → the frame is in digital format.

4.5.5 Face detection and embedding extraction

Following the collection of the dataset, several operations were performed, many of them related to the search for similarity of the subjects present and this was because of a twofold reason:

- the similarity between subjects in the dataset will be exploited to figure out which subjects to merge to generate morphed images → this is disussed in the next sections
- it is interesting to observe how recognition ability varies for both genuine and impostor comparisons → by genuine comparisons is meant the study of differences between the ICAO compliant image of a subject and the frame of a video in which he appears, while impostor comparisons mean the study of the differences between the ICAO compliant image of a subject and the frame of a video in which a different subject appears or a morphed image and the videos of the two original subjects→ this will be discussed in the next chapter.

4.5.6 Face detection using ArcFace

As explained in the section dedicated to ArcFace, the first operation performed on the dataset was to use the implementation of ArcFace included in InsightFace framework to locate all the faces in the dataset (included in both ICAO and non-ICAO compliant video frames and photos), create the cropped images (cropped only on the faces) and from them extract the 512-dimensional embeddings that will be used both for morphing generation and genuine/impostor comparison:

(a) original (b) cropped

Figure 4.13: example of cropping effected by ArcFace from a video frame.

From an implementation point of view, below is the Python script that was used to extract the 512-dimensional embedding given the image using InsightFace, i.e. ArcFace having RetinaFace as DCNN.

Specifically, it detects the face within an image using InsightFace and extracts the corresponding embedding, which is saved in an .npy file.

The image of the cropped face is then saved as well:

```python
import numpy as np
import cv2
from insightface.app import FaceAnalysis
import os
import gc

def extract_and_save_descriptor(image_path):
    app = FaceAnalysis()
    app.prepare(ctx_id=-1)

    img = cv2.imread(image_path)
    if img is None:
        print(f"Image '{image_path}' cannot be loaded. Skipping...")
        return

    if len(img.shape) == 2 or img.shape[2] != 3:
        print(f"Image '{image_path}' must be in color (RGB). Skipping...")
        return

    img_rgb = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)
    faces = app.get(img_rgb)
    if not faces:
        print(f"No face detected in '{image_path}'. Skipping...")
        return

    descriptor = faces[0].embedding
    print(f"Descriptor dimensions for '{image_path}': {descriptor.shape}")

    output_path = os.path.splitext(image_path)[0] + ".npy"
    np.save(output_path, descriptor)

    bbox = faces[0].bbox
    x1, y1, x2, y2 = map(int, bbox)

```

```
35      cropped_face = img[y1:y2, x1:x2]
36      cropped_image_path = os.path.splitext(image_path)[0] + "_cropped_face.jpg"
37      cv2.imwrite(cropped_image_path, cropped_face)
38
39      os.remove(image_path)
```

To analyze the provided code in more detail, it is useful to underline that the insightface.app.FaceAnalysis module of the InsightFace framework implements a complete pipeline for facial analysis, including face detection and description using pre-trained deep learning models:

- as far as face detection is concerned, this is done through the RetinaFace model, which is able to simultaneously regress the face bounding box (coordinates x_1, y_1, x_2, y_2), a confidence estimate and five specific landmarks (the two eyes, the nose and the two extreme points of the mouth) $\rightarrow$ thanks to the method app.get(image) it is therefore possible to obtain a list of the faces detected in the input image, in which each element contains the bounding box of the face (bbox), the coordinates of the 5 previously described landmarks (kps) and the embedding

- as for the embedding, it is generated by ArcFace, which is capable of projecting the face image into a R^{512} L2-normalized vector space, where the angular distance between vectors reflects facial similarity, with the ultimate goal of increasing separability between different classes and improving intra-class discriminability.

Of particular interest is also to observe in 3D space the distribution of embeddings referring to the ICAO-compliant frontal photos of all 65 subjects composing the GazeWay dataset.
Clearly, since such embeddings are arranged in 512-dimensional space, dimensionality reduction techniques, such as t-SNE, must be used to observe them more conveniently[72], which yes make it easier to visualize the arrangement of embeddings in space but clearly also result in a considerable loss of information, so the following image is only meant to show the distribution of embeddings, the distances are not true:

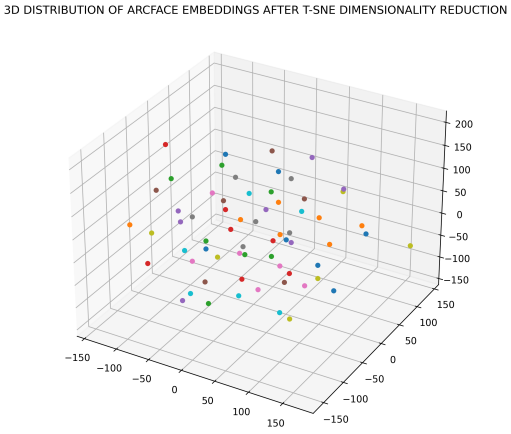


Figure 4.14: distribution in 3D space of embeddings related to ICAO-compliant frontal photos of all subjects in the dataset.

Equally interesting may be to observe the spatial arrangement (again in a 3D space obtained by dimensionality reduction with t-SNE) of all embeddings related to all photos and frames of the same subject in the dataset (frames, ICAO and non-ICAO compliant):

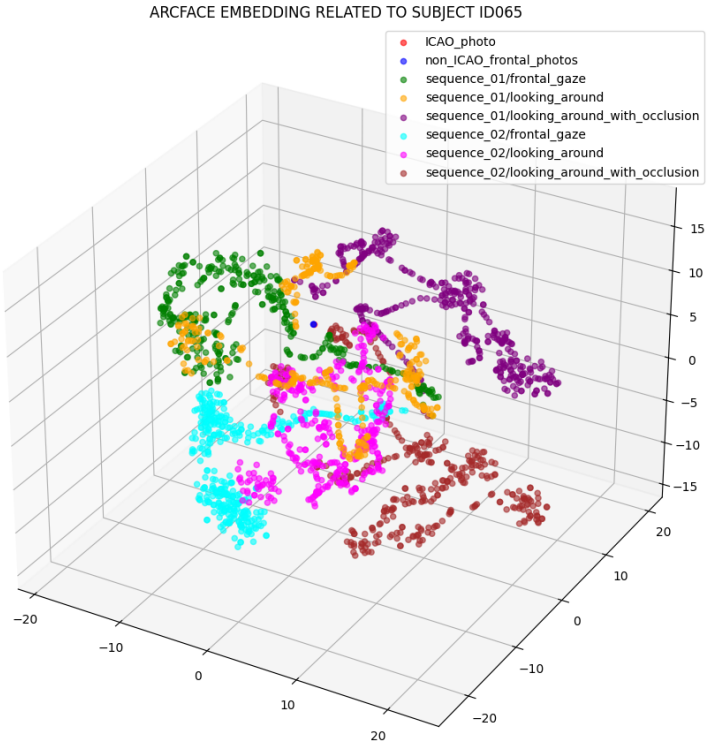


Figure 4.15: distribution in 3D space of embeddings related to all photos and frames of subject ID065.

4.5.7 Morphing

This section explains how were generated the 325 morphed photos included into the GazeWay dataset.

Morphing criteria

After exploiting, as described in the previous section, ArcFace for the computation of the embedding of every single image included in the dataset (ICAO, non-ICAO and video frame), the cosine similarity between all the ICAO fronatal photos was exploited to decide which subjects combine in order to generate a morphed image: therefore, for each of the 65 subjects in the dataset, the cosine similarity with the other subjects was computed and the pairs with the high-

est similarity were stored and it was from the ICAO compliant frontal photos of these pairs that the morphed images were generated.

Morphing generation

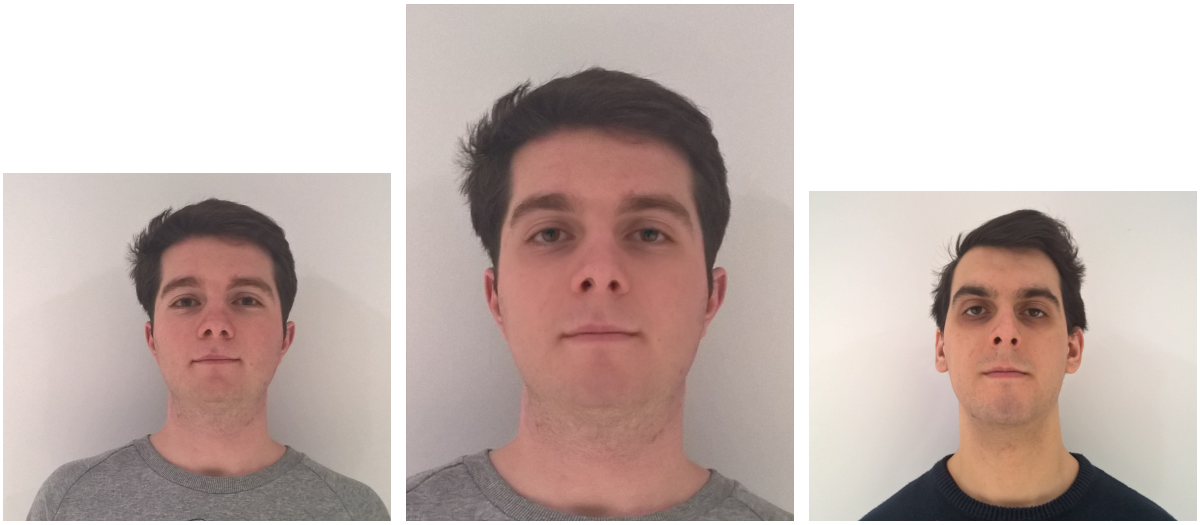
After the identification of the best pairs to combine, as mentioned in the first chapter of this thesis, there were a variety of methods and techniques for producing morphed images.
In the case of the dataset images, in particular, they were generated by combining texture blending and shape warping techniques, looking for the best and most realistic combinations of the weighted factors α_B and α_W, and finally including in the process also automatic (not manual) retouching and color equalization operations to increase the realism of the produced results.
In more detail, it was also mentioned in the first chapter of this thesis that the typical face morphing process is about combining the identities of two different subjects in a single image by applying the two transformations of shape warping and texture blending but using a single weight called morphing factor α for both such transformations.
Instead, as suggested by the study[73], morphed images in our GazeWay dataset do not exploit a single factor for both tasformations but two separate and distinct weight factors: α_B for blending and α_W for warping.
In more detail, the process of generating morphed images can be summarized in the bulleted list below:

- starting from the two ICAO compliant images $I0$ and $I1$ identification of relevant face points
- application of the warping function to align the landmarks of one image to the other $\rightarrow$ as mentioned the degree of warping is controlled by the warping factor α_W
- application of image blending, which is nothing but a weighted average of the original pixel intensity of the two source images $\rightarrow$ again the degree of blending is controlled by the blending factor α_B
- the most realistic morphed image was selected by trying various combinations of α_W and α_B in [0, 0.1, 0.2, 0.3, 0.4, 0.5].

Following this pipeline, as anticipated, in order to make the resulting morphed photo even more realistic, a simple automatic retouching procedure was adopted: this procedure focused on replacing the background region surrounding the face region with the region of one of the two

original photos and automatically equalizing the skin color.

Below is an example showing the good quality of the results produced:



(a) ICAO photo of subject 025 (b) morphed photos (c) ICAO photo of subject 051

Figure 4.16: example of starting and morphed photos of subjects 025 and 051.

4.5.8 Statistics of subjects present

Since this is a dataset acquired in a university and educational context it would have been easy to unintentionally introduce biases within the dataset, such as subjects around 20 years of age, predominantly male and European/American ethnicity.
However, efforts were made to introduce as much variability as possible, as also evidenced by the following statistics and infographics, which show yes some trends and features more common than others but in a nonetheless accentuated and limited form:

- total number of images (frontal + video frames + morphed) $\rightarrow$ 331377
- average number of frames in each video $\rightarrow$ 848
- average number of images per subject $\rightarrow$ 5094
- average age of subjects $\rightarrow$ 30.4 years.

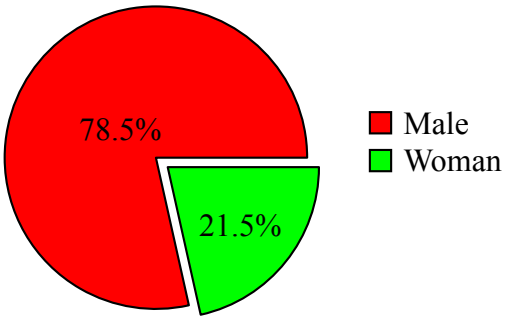


Figure 4.17: percentage of the gender distribution in the dataset.

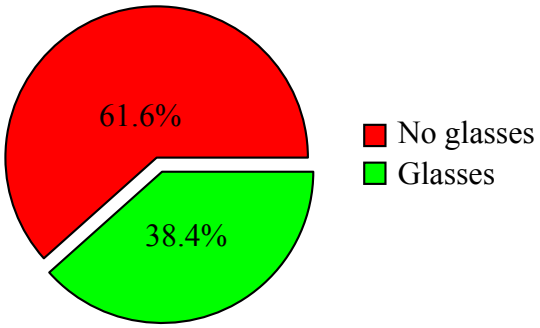


Figure 4.18: percentage of the distribution of subjects with and without glasses in the dataset.

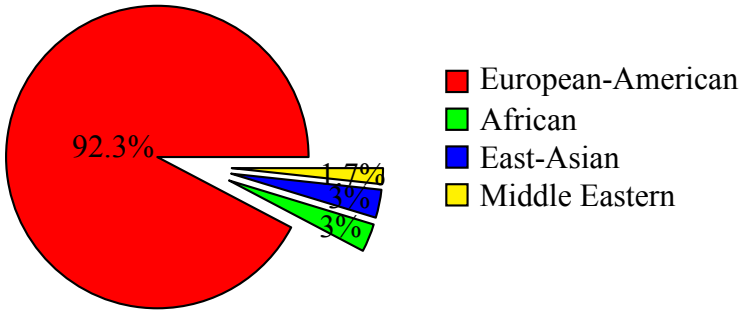
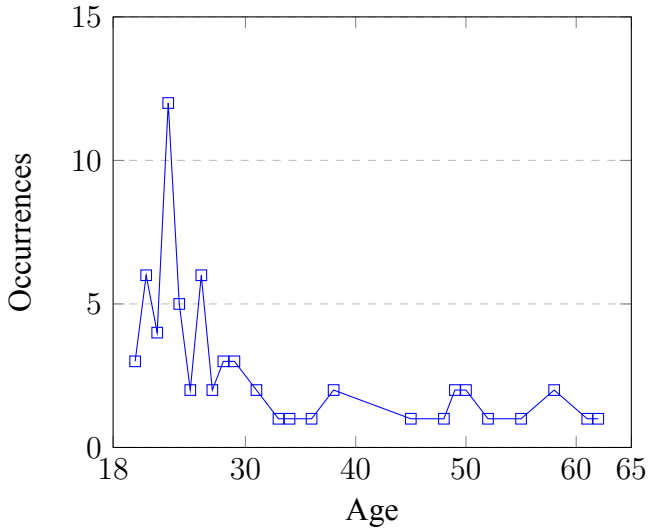


Figure 4.19: percentage of the ethnicity distribution of the subjects in the dataset.

Finally, the graph below shows the age distribution of the subjects in the dataset:



Last peculiar feature that deserves to be highlighted turns out to be the presence of family relationships between subjects in the dataset, in fact:

- subjects ID029 and ID050 are siblings
- subject ID039 is father of subject ID043.

Such relationships will be of particular interest in the similarity search and morphing phases of the present subjects.

Chapter 5

EXPERIMENTS CONDUCTED ON THE GAZEWAY DATASET

Experiments conducted on the Gazeway dataset can be classified into two categories:

- facial recognition → this task does not exploit morphed photos in any way but merely studies scores related to cosine distance from genuine (the frontal photo of a subject with all frames in which the same subject appears) and impostor (the frontal photo of a subject with videos of a few different but similar subjects) comparisons
- video-based morphing detection → for this other task the morphed pictures generated and included in the dataset were instead exploited, since the scores were computed no longer by cosine distance but by an SVM classifier with RBF kernel, which is also in this case asked to analyze genuine (the frontal photo of a subject with all frames in which the same subject appears, exactly as in the previous case) and impostor (every morphed photo with videos of the two subjects that was mixed for producing it) comparisons.

The bulleted list that follows provides a brief overview of the operations performed in both the tasks:

- face detection → using ArcFace it was possible to detect all faces within all images (ICAO, non-ICAO and video frames) that make up the dataset and encode them using 512-dimensional embeddings
- computation of scores (using cosine distance or the SVM classifier) for genuine and impostor comparisons

- drawing of the resulting DET curves, so as to evaluate the performance of a simple binary classifier asked to distinguish a genuine comparison from an impostor
- study of the variation in recognition and morphing detection as a function of the distance between the camera and the detected face.

It should be pointed out, however, that actually the following operations were performed not only on our Gazeway dataset, but had been preliminarily performed on the ChokePoint dataset as well, offering unsatisfactory results, however, which are briefly described below.

5.1 Face recognition using cosine distance

Once all 512-dimensional embeddings were obtained, it was possible to study the cosine distances separating the frontal photo of each subject with all the video frames in which he appears. More in detail, as expressed before:

- in the so-called genuine comparisons the cosine distance between the frontal photo of each subject with the frames of the videos in which he appears
- on the other hand in the so-called impostor comparisons the cosine distance between the ICAO compliant frontal photo of each subject and the videos in which subjects similar to him but not himself appear $\rightarrow$ to identify the subjects most similar to each other it was sufficient to calculate the cosine similarity between the ICAO compliant frontal photos of each subject and identify the pairs with highest similarity.

The mathematical formulas related to the calculation of cosine similarity and cosine distance between two vectors of dimension 512 are given below:

$$\text{cosine\_similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} = \frac{\sum_{i=1}^{512} A_i B_i}{\sqrt{\sum_{i=1}^{512} A_i^2} \cdot \sqrt{\sum_{i=1}^{512} B_i^2}}$$

$$\text{cosine\_distance}(\mathbf{A}, \mathbf{B}) = 1 - \text{cosine\_similarity}(\mathbf{A}, \mathbf{B}) = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} = 1 - \frac{\sum_{i=1}^{512} A_i B_i}{\sqrt{\sum_{i=1}^{512} A_i^2} \cdot \sqrt{\sum_{i=1}^{512} B_i^2}}$$

As far as genuine comparisons were concerned, it was therefore possible to produce a database (in the form of a text file) that included all the computed cosine distances, of which a very small section is given in the table 5.1 below for illustrative purposes:

SUBJECT	SEQUENCE	POSE	FRAME	DISTANCE
ID001	01	frontal_gaze	I000000	0.562880
ID001	01	frontal_gaze	I000001	0.483114
...	...	...	...	...
ID065	02	looking_around_with_occlusion	I000377	0.940997

Table 5.1: fragment of the table containing all the cosine distances that emerged from the genuine comparisons.

From the previous database it was possible to compute aggregate measures of cosine distances per video, such as mean, maximum and minimum.
To do this, it was sufficient to group the rows according to the triple (SUBJECT, SEQUENCE, POSE), resulting in a new database, similar to the previous one but having a slightly different structure, the contents of which are given in the table 5.2 below:

SUBJECT	SEQUENCE	POSE	AVG	MAX	MIN
ID001	01	frontal_gaze	0.455	0.99264	0.262245
ID001	01	looking_around	0.52690	0.906124	0.32197
...	...	...	...	...	
ID065	02	looking_around_with_occlusion	0.677	0.949603	0.312631

Table 5.2: table fragment containing all cosine distances that emerged from genuine comparisons in aggregate form by video.

Before moving on to the production of the two databases equivalent to above but for impostor comparisons, it was preliminarily necessary to find, for each subject in the dataset, the 6 subjects most similar to him on the basis of the cosine similarity of the embeddings representing ICAO-compliant frontal photos (exactly 6 since for genuine comparisons each ICAO-compliant photo was also compared to 6 video sequences).
Once all the similarity matches between subjects were identified, the cosine distances of the impostor comparisons were identified by comparing each subject's frontal photo with frames of a video sequence in which each of the 6 subjects found to be most similar to him appears, resulting in a database with a structure equivalent to the table 5.3 below:

SUBJECT	SIMILAR SUBJECT	SEQUENCE	POSE	FRAME	DISTANCE
ID001	ID053	02	looking_around	I000000	0.964173
ID001	ID053	02	looking_around	I000001	0.933918
...	...	...	...	...	
ID001	ID048	02	looking_around	I000001	0.809084
...	...	...	...	...	
ID065	ID063	01	frontal_gaze	I000400	0.994516

Table 5.3: fragment of the table containing all the cosine distances that emerged from the impostor comparisons.

From the database above and in a manner entirely analogous to the genuine comparisons, it was then possible to produce the dataset for aggregate cosine distances, obtained by grouping the rows according to tuple (SUBJECT, SIMILAR SUBJECT, SEQUENCE, POSE), yielding the following table 5.4:

SUBJECT	SIMILAR SUBJECT	SEQUENCE	POSE	AVG	MAX	MIN
ID001	ID053	02	looking_around	0.92	1	0.79
ID001	ID048	02	looking_around	0.914	1	0.781
...	...	...	...	...		
ID065	ID063	01	frontal_gaze	0.945	1	0.828

Table 5.4: table fragment containing all aggregate cosine distances that emerged from impostor comparisons.

5.2 Experiments on video-based MAD

It should be emphasized that the operations described in the previous section did not exploit nor use in any way the morphed photos contained in the dataset.
So, as anticipated in the introduction of this chapter, it was necessary to repeat the same steps that composed the piepline described in the previous chapter, but using the morphed images for impostor comparisons and not videos of subjects similar but different to ICAO photos.
In this sense it can be said that we have moved from a face recognition task to a true morphing detection task.
An additional important difference, besides the just mentioned use of morphed photos, concerns the computation of scores: whereas before we used a score based on cosine distance, in this other case each score was produced by a classifier, proposed in the paper[74] and used also in paper[51], trained specifically for the MAD task, which therefore for each photo is able to compute a score ranging between 0 (bona fide image) and 1 (counterfeit image, as morphed).

In more detail, the morphing score produced by the classifier is estimated from deep facial representations extracted from neural networks pre-trained for facial recognition.
Yet be $\mathbf{x}_r \in R^d$ the feature vector (embedding) extracted from the reference image (e.g. an ICAO photo or a morphed photo) and be $\mathbf{x}_p \in R^d$ the vector extracted from the probe image (e.g. a live capture).
The difference vector is calculated as $d = x_r - x_p$.
The vector d thus represents the deviation between the two facial representations and it is then used as input to a supervised classifier, usually an SVM with RBF kernel, which produces a score $s \in [0, 1]$ interpreted as the probability that the reference image is morphed.

Formally, let $f : R^d \rightarrow [0, 1]$ be the discriminant function learned by the classifier, then $s = f(d) = f(x_r - x_p)$ where, as anticipated:

- $s \approx 0$ indicates a bona fide (no morphing) classification
- $s \approx 1$ indicates instead a morphing classification and thus a probable attack.

For training such a classifier, (x_r, x_p) pairs labeled as genuine or morphed were used, constructed from realistic databases that also include post-processed images (such as JPEG2000 or print-scan).
The proposed approach proved to be robust against different morphing and post-processing variants, achieving an EER of less than 3% in the best configuration.

Thanks to the classifier was thus possible to obtain morphing scores for genuine comparisons (always comparing each video frame of the dataset with the ICAO-compliant frontal photo of the subject depicted in the frame), obtaining in fact a database having a structure quite similar to the one reported in the previous chapter obtained on the face recognition task, reported in the following table 5.5:

SUBJECT	**SEQUENCE**	**POSE**	**FRAME**	**SCORE**
ID001	01	frontal_gaze	I000000	0.986929
ID001	01	frontal_gaze	I000001	0.969387
...	...	...	...	...
ID065	02	looking_around_with_occlusion	I000409	0.997254

Table 5.5: database fragment containing all scores computed for genuine comparisons.

Then, following the same pipeline as described in the previous chapter for the facial recognition task, these genuine scores were aggregated by video, so that for each video, the average, maximum and minimum morphing scores were available.
The table 5.6 belows shows the structure of this aggregated database:

SUBJECT	SEQUENCE	POSE	AVG	MAX	MIN
ID001	01	frontal_gaze	0.758	1	0.00643
ID001	01	looking_around	0.8947	1	0.25764
...	...	...	...	...	
ID065	02	looking_around_with_occlusion	0.865	1	0.00012

Table 5.6: database fragment containing all scores obtained from genuine comparisons in aggregated form by video.

The same procedure was also followed for impostor comparisons.
That is, the classifier was exploited to compute the morphing score for each morphed image in the GazeWay dataset with all frames from all videos of both subjects that were mixed to create the mentioned above morphed image.
Such scores are reported in table 5.7 below:

SUBJECT_1	SUBJECT_2	IN_FRAME	SEQUENCE	POSE	FRAME	SCORE
ID001	ID017	ID001	01	frontal_gaze	I000000	0.99
ID001	ID017	ID001	01	frontal_gaze	I000001	0.99
...	...	...	...	...	...	
ID001	ID017	ID017	01	frontal_gaze	I000000	0.99
...	...	...	...	...	...	
ID001	ID020	ID001	01	frontal_gaze	I000000	0.99
...	...	...	...	...	...	
ID002	ID022	ID002	01	frontal_gaze	I000000	0.98
...	...	...	...	...	...	
ID065	ID048	ID065	01	looking_ar...	I000409	1

Table 5.7: database fragment containing all computed scores for impostor comparisons.

From the database above it was then possible to derive the one with aggregated scores by video, obtaining a databases similar to table 5.8:

ID1	ID2	ID_IN_FRAME	SEQUENCE	POSE	AVG	MAX	MIN
001	017	001	01	frontal_gaze	0.99	1	0.83
001	017	001	01	looking_around	0.99	1	0.97
...	...	...	...	...	...		
001	017	017	01	looking_around	0.99	1	0.97
...	...	...	...	...	...		
001	020	001	01	frontal_gaze	0.99	1	0.83
...	...	...	...	...	...		
002	022	002	01	frontal_gaze	0.99	1	0.88
...	...	...	...	...	...		
065	048	065	02	looking_arou...	0.99	1	0.98

Table 5.8: database fragment containing all aggregated morphing scores computed for impostor comparisons.

5.2.1 MAP matrix

After computing the MAD scores, it was possibile to expolit them to determine the so called Morphing Attack Potential matrix[75], an innovative methodology introduced to quantify the risk and vulnerability posed by specific morphing attacks against Facial Recognition Systems (FRS).
In contrast to traditional attack assessment criteria that focus on the attacker's experience or knowledge of the system, MAP provides a consistent methodology for assessing the threat posed by morphed images in relation to the number of attempts/comparisons analyzed.

As explained in the first chapters of this thesis, in fact, a manipulated image generated by merging samples of two subjects can be detected as fraudulent when compared with probe images of the subjects that contributed to its creation.

More in detail, the MAP methodology overcomes the limitations of traditionally existing indicators such as MMPMR and FMMPMR by extending the vulnerability assessment to include two critical factors:

- a variable number of attempts $\rightarrow$ e.g. by comparing the same potentially morphed image with multiple probe images acquired at an automatic border control (ABC) gate for each of the two subjects who created the combined image
- aggregate evaluation of multiple Facial Recognition Systems (FRS).

This allows MAP to simulate a more realistic border control scenario, providing a more complete indicator thorough quantification of the attack potential of a given dataset of morphed images.

The MAP is defined as a matrix of size $m \times n$, where m represents the number of probe images (attempts) available for each contributing subject and n represents instead the number of FRSs considered for evaluation.
So, a generic element of this matrix MAP[r, c] reports the percentage of morphed images that manage to achieve with a certain confidence set by a threshold a robust match with both contributing subjects.
Mathematically, the condition for a morphed image M to be included in the computation of MAP[r, c] is given by:

$$\forall\, M \in dataset:\ C_{\text{MAP}[r,c]} = \begin{cases} +1 & \text{if } fmc(M, P_1, F, r) \geq c \wedge fmc(M, P_2, F, r) \geq c \\ +0 & \text{otherwise} \end{cases}$$

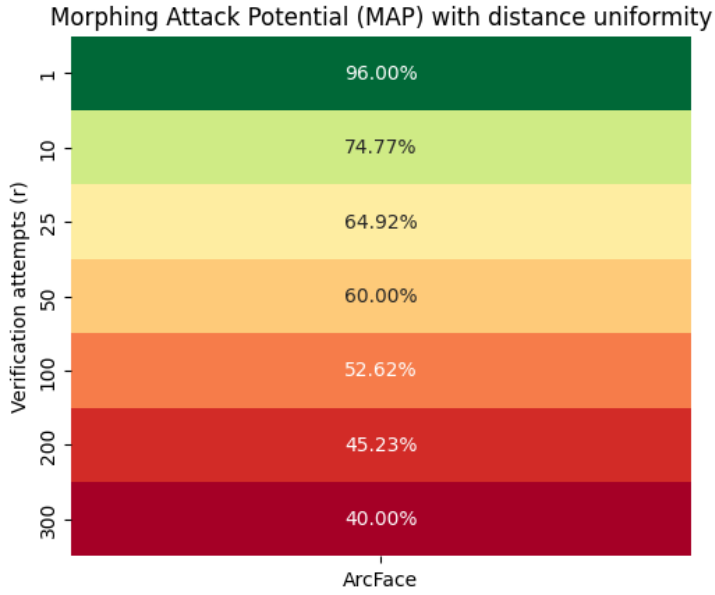
This formula specifies that the condition for including a morphed image M in the computation of MAP[r,c] is that the number of FRSs F for which at least r of the probe images of subject P_1 have been successfully verified (the comparison score yields an outcome above the c threshold) and the same condition must also hold for subject P_2, where P_1 and P_2 are the subject that have been mixed in order to create the morphed image M.
In simple terms, then, the generic MAP[r,c] cell indicates the percentage of morphed images in the dataset that succeed above the verification threshold for both original subjects with at least r probe verification images for each subject and in at least c FRS.

In the context of our dataset, it is necessary to specify that:

- the MAP matrices produced will have a single column since only ArcFace was used as FRS
- the c threshold to be exceeded for the comparison between the morphed photo and the probe image to be verified was set to 0.49320, as suggested by the paper[75], as this

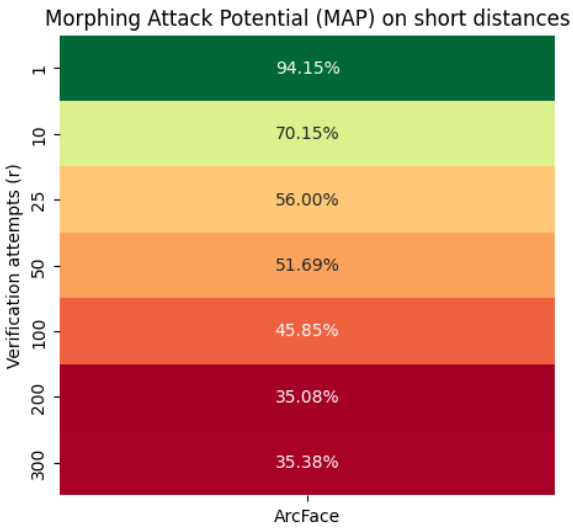
value was experimentally identified as the value that should guarantee a False Match Rate (FMR) of 0.1%, i.e. only 1 false positive per 1000 comparisons.

As an example, the following image shows the MAP matrix produced by considering the entire GazeWay dataset but maintaining a uniform number of distance-based comparisons, i.e. the generic r row includes about the same number of comparisons using frames both in which the original subjects appear at short distances, medium distances and long distances:

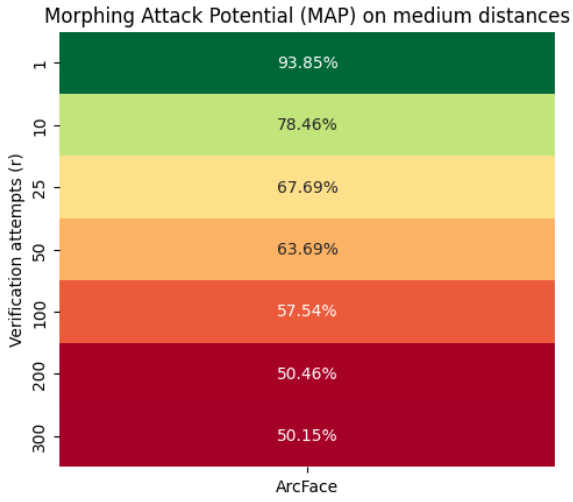


The MAP matrix can also help us to anticipate a concept that will also emerge in the next chapter, namely that the error rate increases with distance.

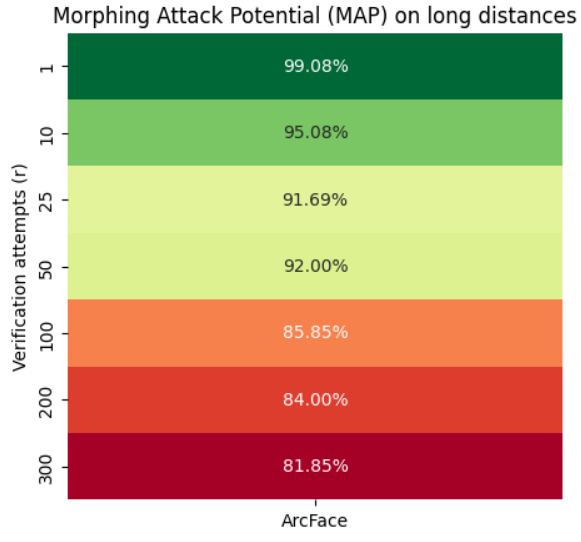
In fact, by computing the MAP matrices also considering the different distance ranges, we also observe here how performance degrades if the distance is high:



(a) MAP over short distance



(b) MAP over medium distance



(c) MAP over long distance

Figure 5.1: comparison of MAP considering the different distances.

In fact, as it is noted, as the number of r comparisons increases, MAPs referred to short and medium distances are quite robust having all in all low percentages of morphed photos that manage to be confused with probe photos, while in long-distance MAP even with a high number of r comparisons this percentage is quite high, indicating that even with so many comparisons available, almost all morphed photos are easily confused with the probes of the two original subjects if they were acquired at long distances.

Chapter 6

RESULTS ON FACE RECOGNITION AND VIDEO-BASED MAD

In this chapter are reported the results obtained in both the task of face recognition and video-based MAD.
Special attention will be paid to the design and study of DET curves (which, as previously said, report error trends as false positives and false negatives vary) and interesting considerations about the variation of results as a function of the distance at which the face of the subject was detected.

6.1 Design of testing protocol

This section describes in more detail how genuine and impostor comparisons were defined in both the face recognition and V-MAD cases and how many there comparisons were done in total.
For what concerns facial recognition:

- genuine comparisons were made by computing the cosine distance between the embedding of the ICAO-compliant frontal photo of each of the 65 subjects and all embeddings obtained from the frames of all video sequences of that subject $\rightarrow$ in total the genuine comparisons for facial recognition turn out to be 144644
- impostor comparisons, on the other hand, were carried out by computing the cosine distance between the embedding of the ICAO-compliant frontal photo of each of the 65 subjects and the embeddings of the frames of 6 videos belonging, instead, to subjects most

similar to him (found by studying the cosine distances of the embeddings obtained from the 65 ICAO-compliant photos) → in total 163624 comparisons were carried out

While, for what concerns the V-MAD task:

- genuine comparisons were made by passing the embedding of the ICAO-compliant frontal photo of each of the 65 subjects and each embedding obtained from the frames of all video sequences of that subject to a SVM classifier with RBF kernel → in total the genuine comparisons for V-MAD turn out to be the same number as facial recognition task, as 163624
- impostor comparisons, on the other hand, were carried out in this case by passing to the classifier the embedding of each of the 325 morphed photos and for each one the embeddings of the frames of all videos belonging to the two subjects that were mixed to create the morphed image → so in this case the total of comparisons were higher, 1655701.

6.2 Results for face recognition

6.2.1 Drawing of DET curves

As anticipated in the initial chapters of this thesis, visualizing the DET curve can be a visual aid in identifying the EER and understanding the trend of BPCER as a function of APCER.
In order for them to be exploited by a simple binary classifier, however, the cosine distances reported in the previous section, whether aggregated or not, must all be converted to scores using the following formula:

$$\text{score} = 1 - \frac{\text{cosine\_distance}}{2}$$

With this conversion, it is possible to map the distances to the outputs of a binary classifier, in that high distances are mapped to scores close to 0 while small distances are mapped to scores close to 1.
It is important to note that this conversion is valid since the specific implementation of cosine distance used (scipy.distance.cosine) returns a value in the range [0, 2].
Other more popular implementations, on the other hand, return the cosine distance in the range [-1, 1] so they would need a different conversion to obtain scores between 0 and 1.
The following two tables 6.1 and 6.2 show fragments of the two genuine and impostor databases in which distances were converted to scores:

SUBJECT	SEQUENCE	POSE	AVG	MAX	MIN
ID001	01	frontal_gaze	0.772	0.50368	0.868877
ID001	01	looking_around	0.73654	0.546937	0.83901
...	...	...	...	...	
ID065	02	looking_around_with_occlusion	0.661	0.525198	0.843684

Table 6.1: table fragment containing all the scores that emerged from the genuine comparisons in aggregate form by video.

SUBJECT	SIMILAR SUBJECT	SEQUENCE	POSE	AVG	MAX	MIN
ID001	ID053	02	looking_around	0.53	0.485	0.6
ID001	ID048	02	looking_around	0.542	0.48	0.609
...	...	...	...	...		
ID065	ID063	01	frontal_gaze	0.527	0.458	0.585

Table 6.2: fragment of the table containing all the aggregate scores that emerged from the impostor comparisons

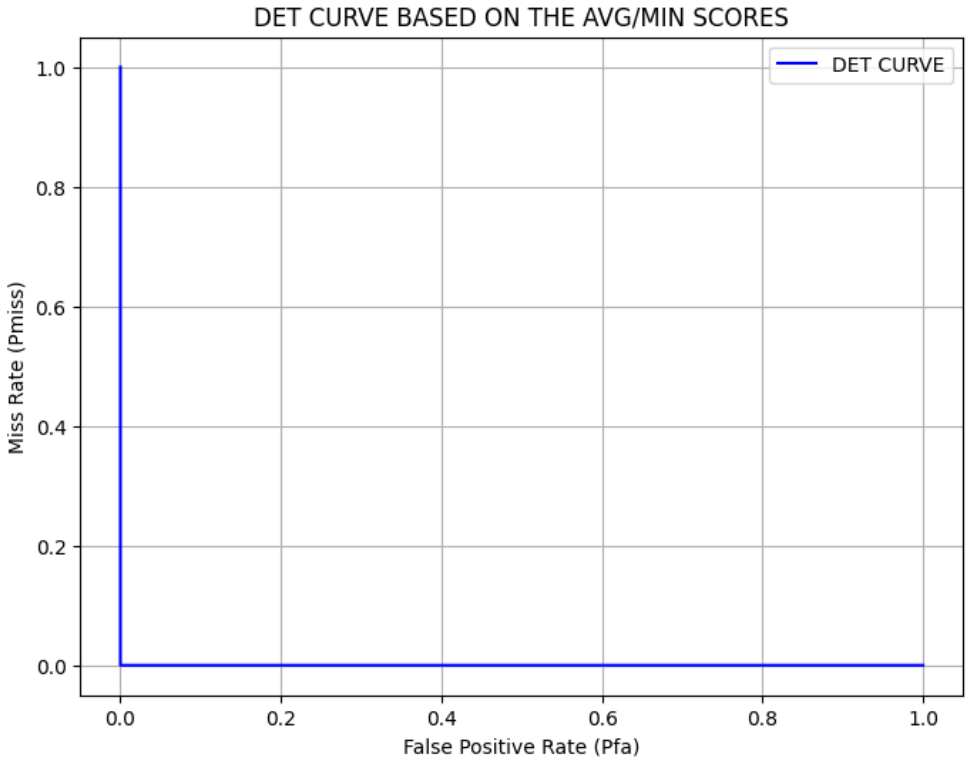
Having converted the distances to scores in both databases, it was possible to draw the DET curves and calculate the Equal Error Rate.

In particular, three DET curves each were produced for both datasets (thus 6 DET curves and six EERs in total), changing from time to time which aggregation strategy consider between average, maximum and minimum of the scores per video.

The DET curves related to the ChokePoint dataset follow.

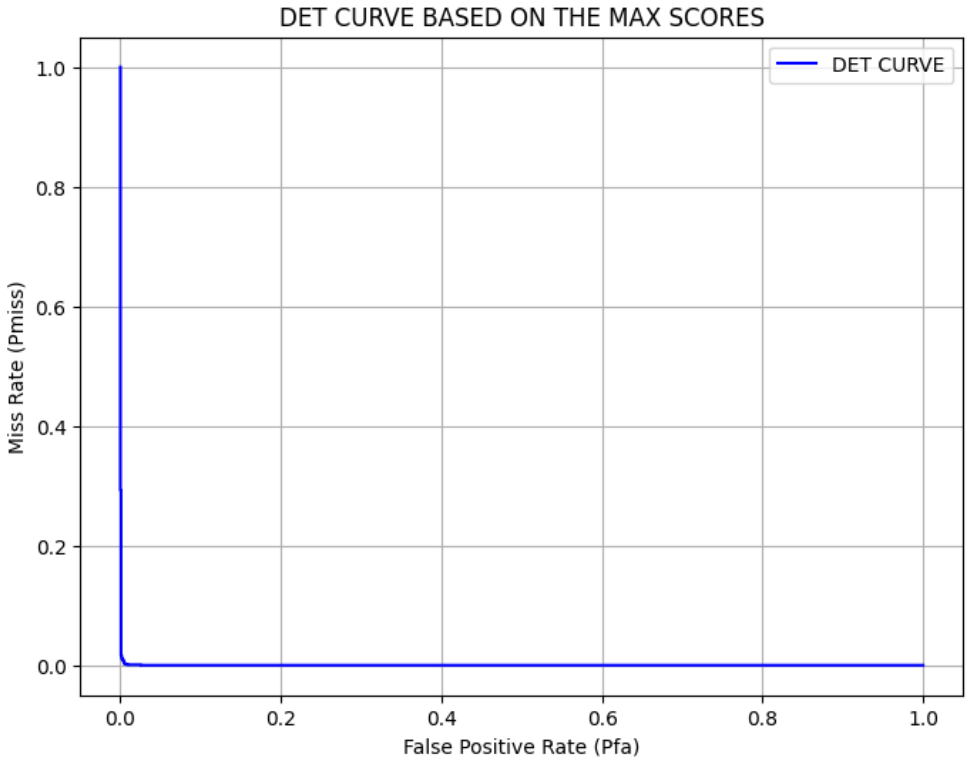
The curve below represents both the DET curve obtained using the mean and the minimum of the distances converted to scores.

In fact, both coincide and show an EER (error rate when APCER and BPCER are equivalent) of 0:

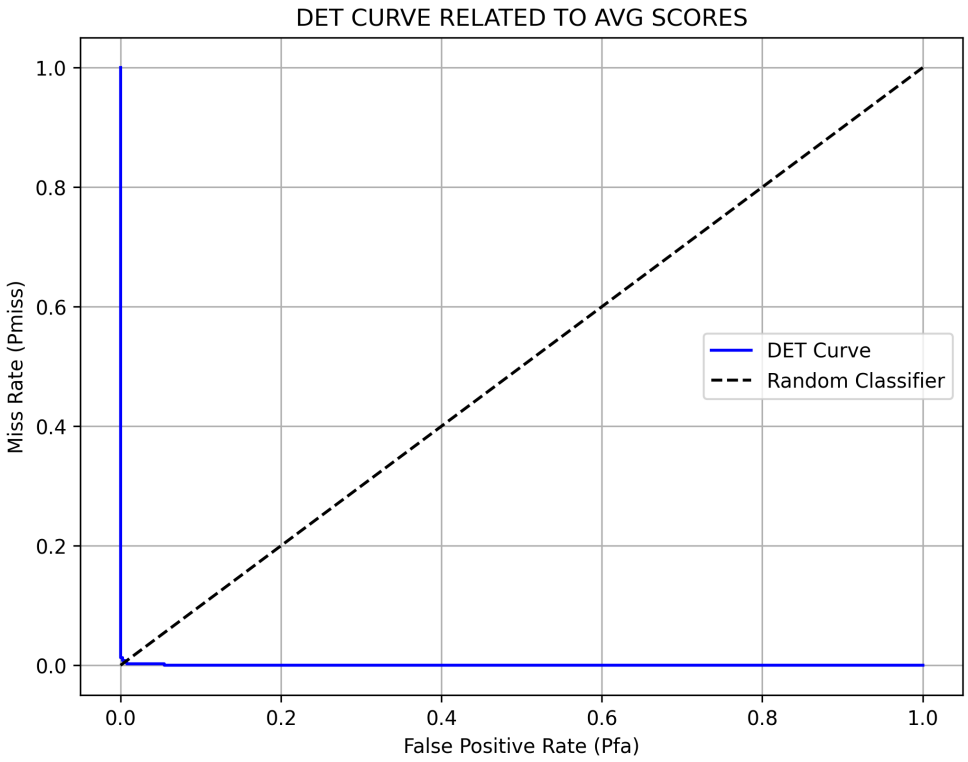


Then follows the DET curve, again calculated on the Chokepoint dataset but using the scores inherent in the maximum distances.

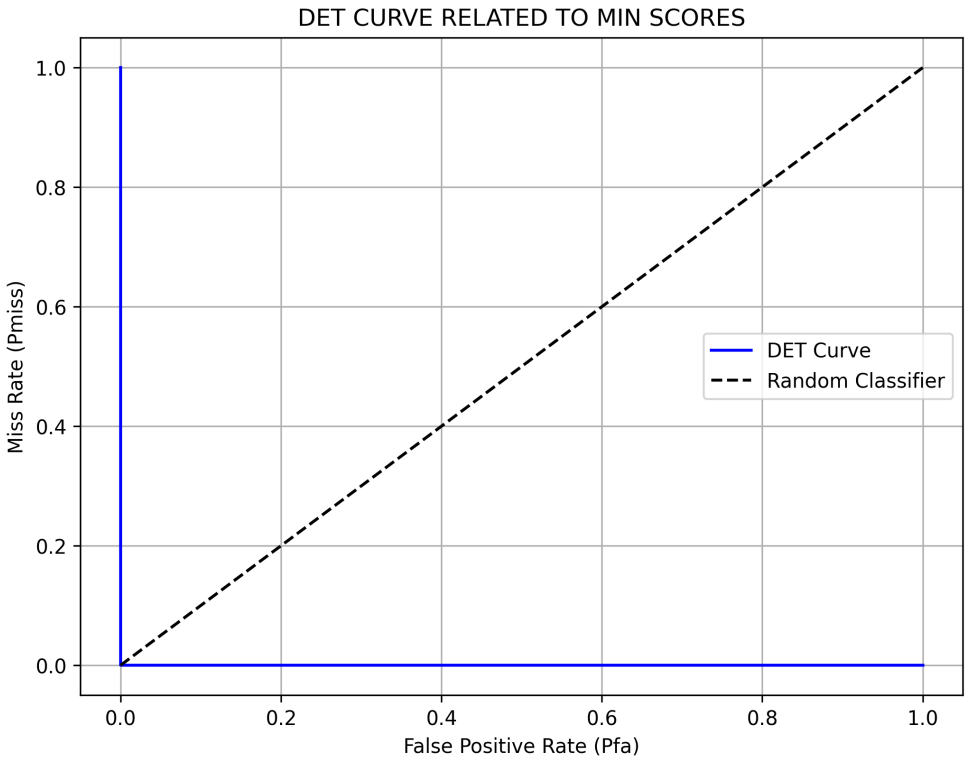
In this other case we see an EER that is always minimal, equal to 0.004 precisely:



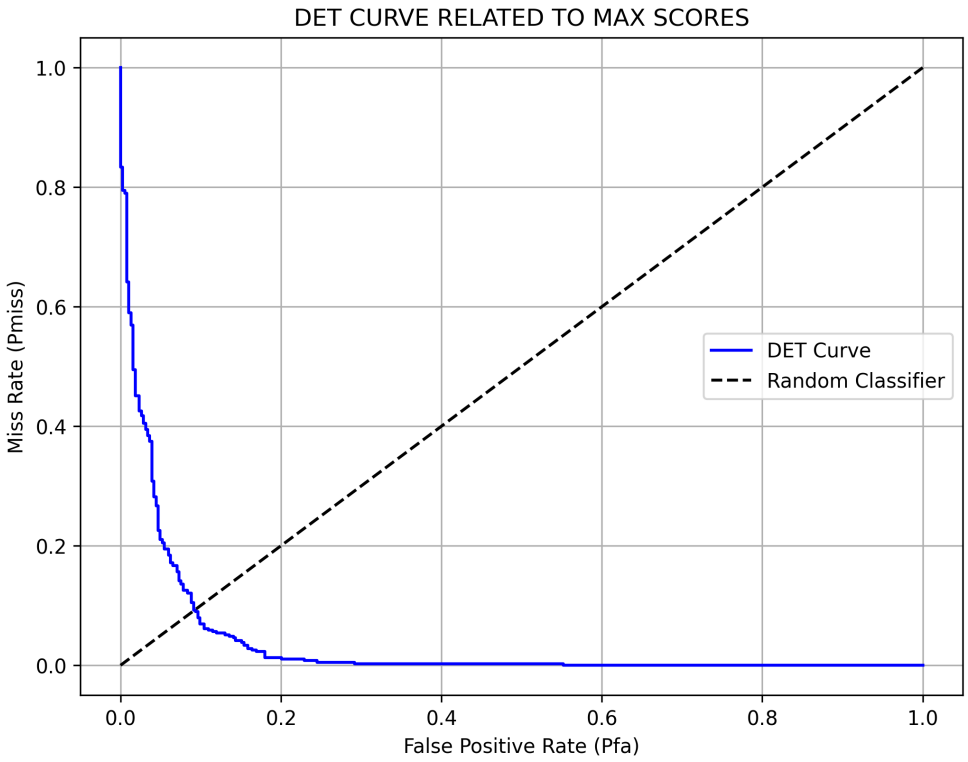
Instead, let us now look at the DET curves related to the scores obtained in the GazeWay dataset, starting with the one obtained from the scores derived from the distance averages, which shows an EER of 0.005 (slightly higher than the same case in the ChokePoint dataset):



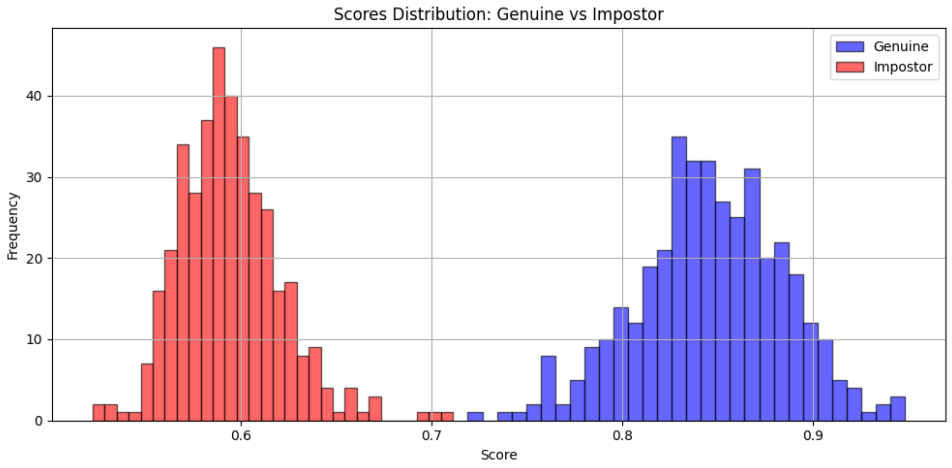
We then go on to show the DET curve obtained using the scores derived from the minimum distances which, however, is completely equivalent to that obtained on the Chokepoint dataset (and the EER is also the same, i.e. equal to 0):



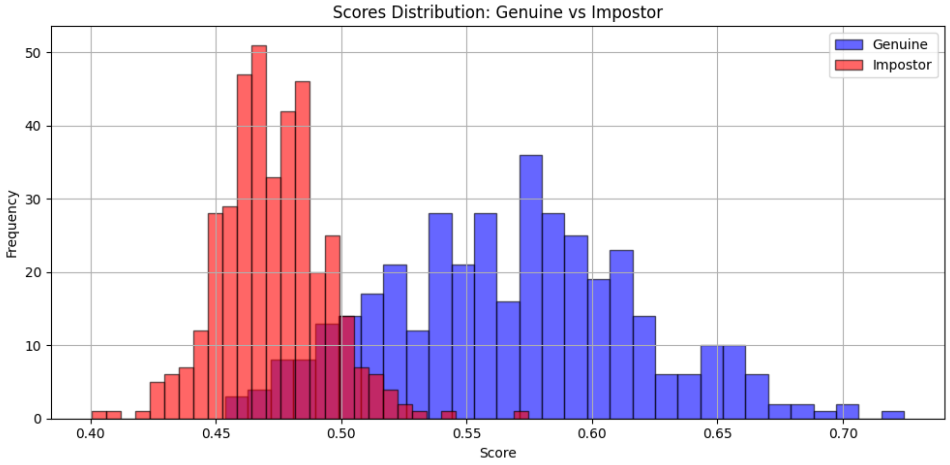
Finally we show the DET curve obtained on the GazeWay dataset using the scores obtained from the maximum distances, which shows slightly less trivial behavior and in fact an EER of 0.091:



The DET curves shown, particularly those from the ChokePoint dataset but also those representing the minimum and average in the GazeWay dataset, exhibit such a distinct pattern and achieve such low EER values primarily because the datasets are relatively simple.
This simplicity allows the classifier to effortlessly distinguish between genuine and impostor scores, as the two distributions are perfectly separable, as illustrated in the figure below.
However, it is also important to note that ArcFace contributes significantly to this performance due to its robustness to variations in pose and lighting, having been trained specifically to handle high variability scenarios commonly encountered in video data.



(a) distribution of scores using scores derived from minimum distances



(b) distribution of scores using scores derived from maximum distances

The subimage (a) shows the distribution of the scores obtained from the minimum distances: as can be seen there is no overlap between the genuine and impostor scores, the two sets are easily separable, leading to an EER of 0 and a DET curve that is not significant at all.
Vice versa, the subimage (b) shows the distribution of scores obtained from the maximum distances: in this case there is overlap between genuine and impostor scores, the two sets are not easily separable and this leads to an EER other than 0 and a slightly more significant DET curve.

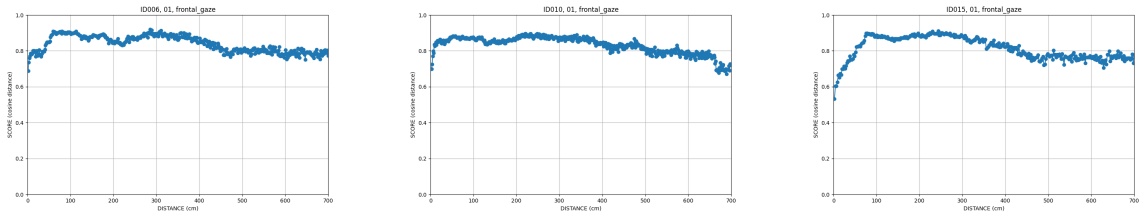
6.2.2 In-depth study about the variation of face recognition accuracy as a function of distance

Having also available the depth version of the frames that make up the dataset, it was possible to study and show through simple graphs how face recognition ability (i.e. score) varies as the

distance of the face from the camera varies.

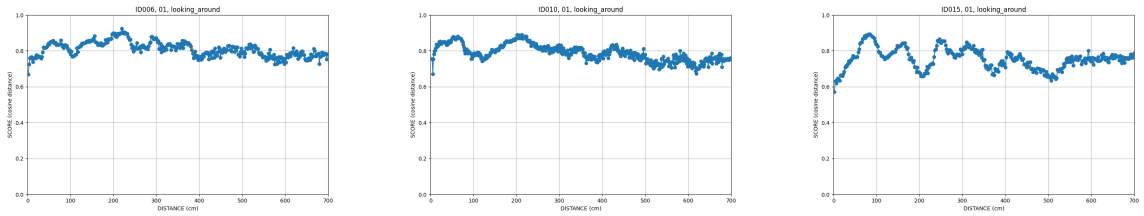
These graphs are of particular importance as they highlight interesting patterns and features common to all subjects, shown below:

- score curves referring to videos in which the subject assumes frontal gaze as a pose appear very stable, while score curves referring to videos in which the subject looks around by turning his head show obvious fluctuations, due precisely to the continuous rotations and variations in orientation of the face with respect to the camera →



(c) frontal gaze video of subject 006 (d) frontal gaze video of subject 0010 (e) frontal gaze video of subject 015

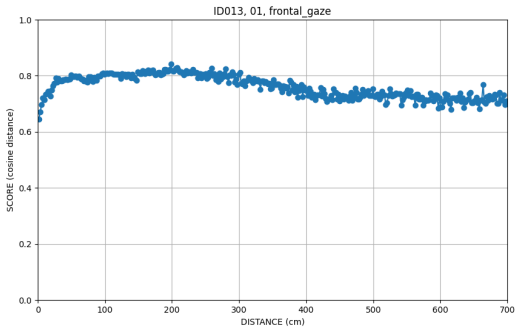
Figure 6.1: examples of score curves about frontal gaze videos.



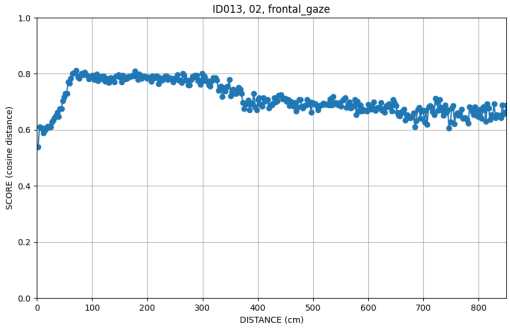
(a) looking around video of subject 006 (b) looking around video of subject 010 (c) looking around video of subject 015

Figure 6.2: examples of score curves about looking around videos

- curves referring to videos in which the subject takes frontal gaze as pose are very constant if the video belongs to the first session (environment illuminated by natural and artificial light), while curves referring to videos in which the subject took frontal gaze as pose in the second session show more fluctuations and a more irregular pattern, most likely due to poor lighting →

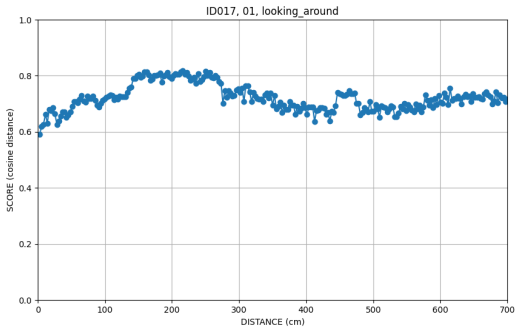


(a) curve of subject ID013 while assuming frontal gaze as pose in session 01

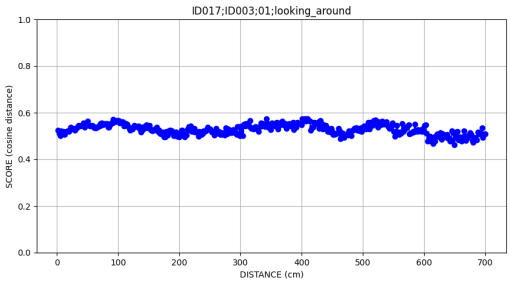


(b) curve of subject ID013 while assuming frontal gaze as pose in session 02

- curves referring to genuine comparisons are always better than those referring to impostor comparisons →

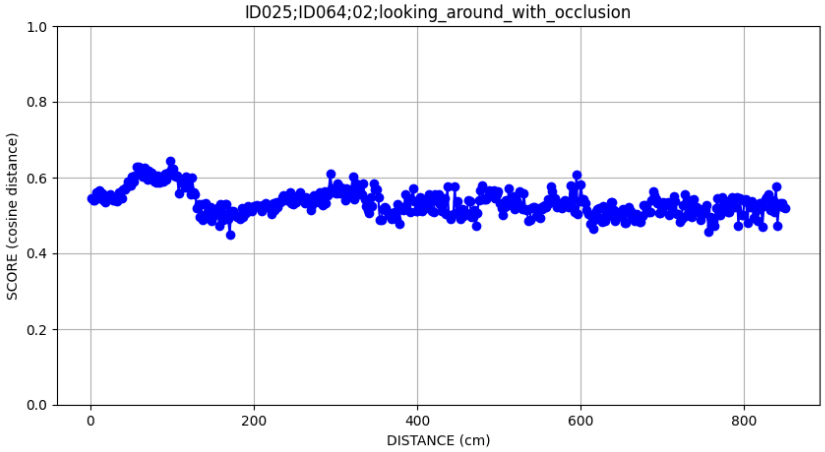


(c) curve of subject ID017 while assuming looking around as pose in a genuine comparison

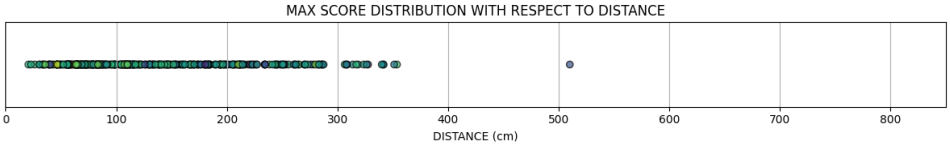


(d) curve of subject ID017 while assuming looking around as pose in an impostor comparison

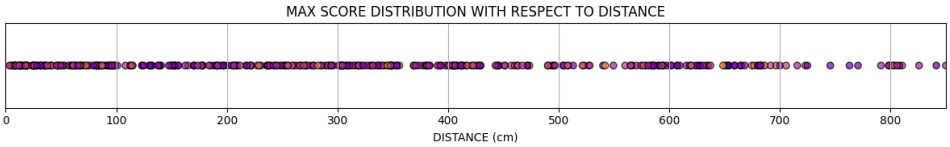
- it is true, as stated in the previous point, that curves referring to impostor comparisons show lower scores than genuine comparisons, but at least they are more constant, even though they are looking around with occlusion as pose → in other words, if the comparison is impostor the fluctuations due to sudden changes in the orientation of the face with respect to the camera are less noticeable →



- finally, the maximum score is almost always noticeable between 20 centimeters and 350 centimeters in genuine comparisons while it is more scattered in impostor comparisons →



(e) distribution of maximum scores as a function of distance for genuine comparisons



(f) distribution of maximum scores as a function of distance for impostor comparisons

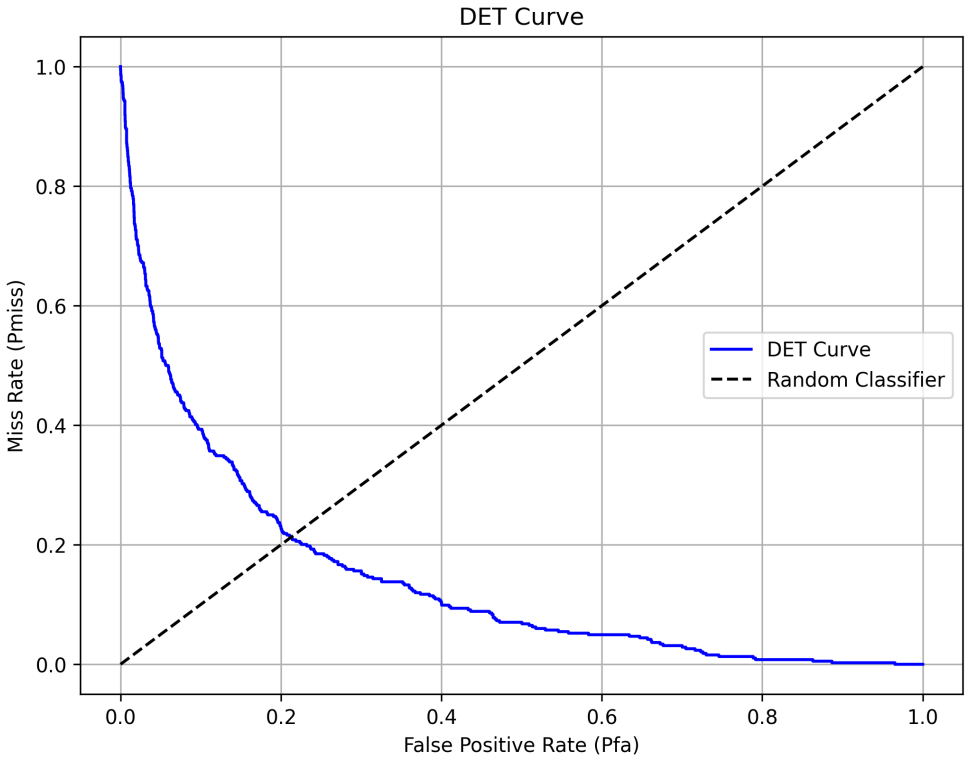
The trends that characterize these curves and the patterns present serve to emphasize, once again, how crucial the influence of certain factors (first and foremost distance, illumination and occlusions) is in the ability to recognize the face and associate it with the correct subject to which it belongs.

6.3 Results for V-MAD

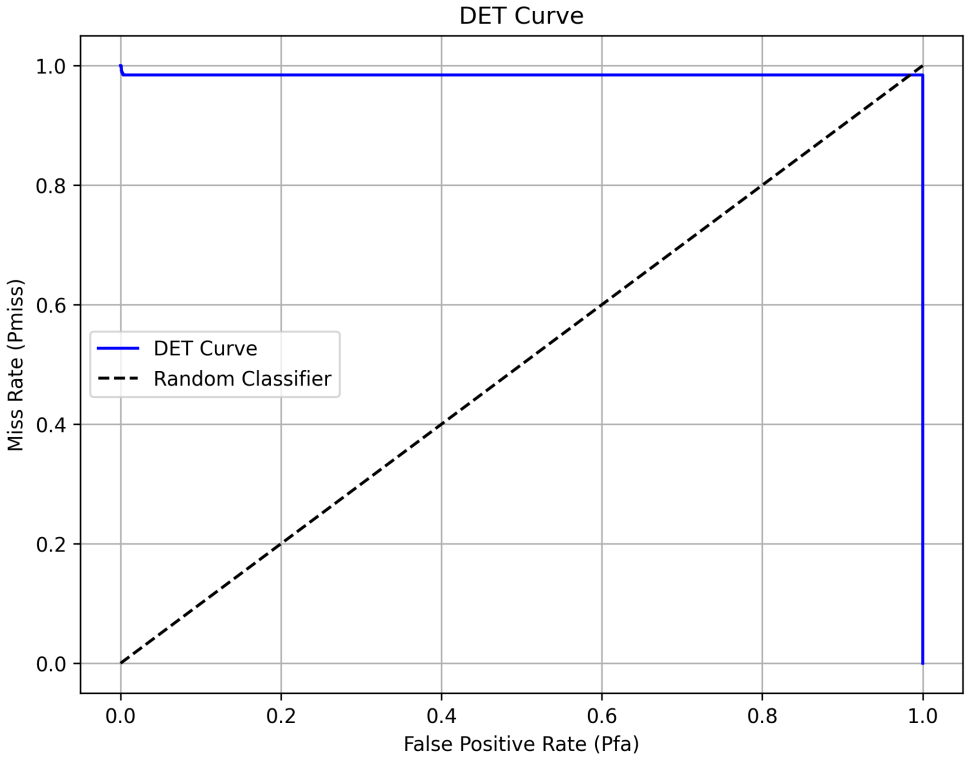
6.3.1 Drawing of DET curves

From the databases produced, it was possible to draw the DET curves in much the same way as described in the previous chapter, again both considering the mean, the minimum and maximum scores per video.

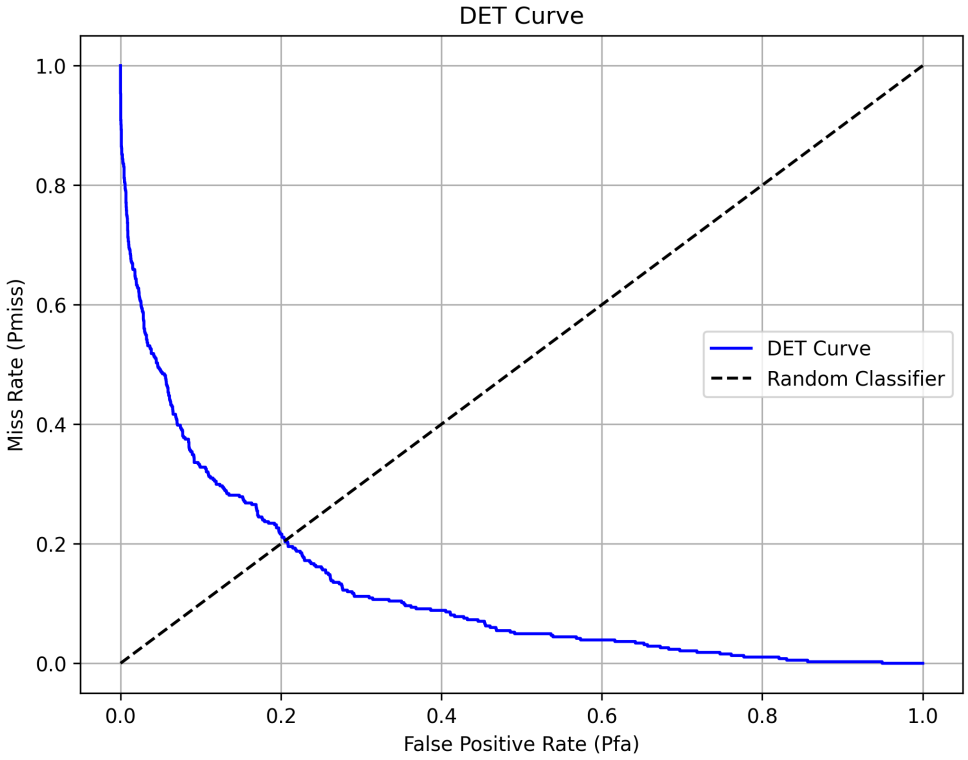
The DET curve produced considering the mean of the scores shows, for example, an EER of 0.212:



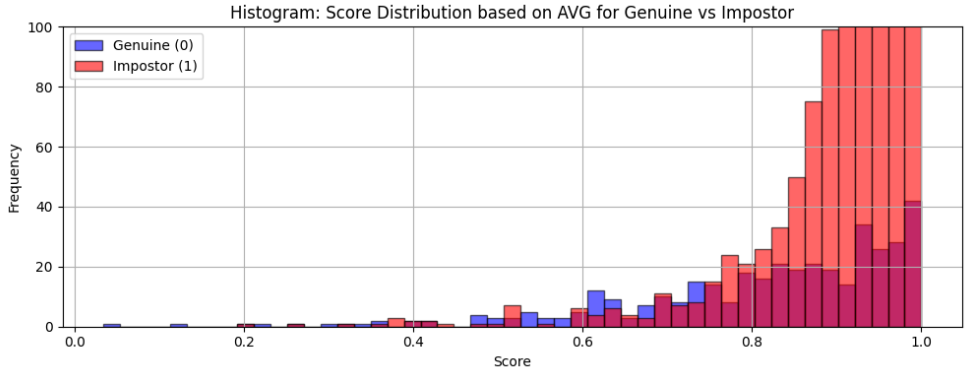
The aggregation strategy based on the maximums scores is instead ineffective, as the DET curve produced with this fusion strategy reports a very high EER, equivalent to 0.984:



Instead, the DET curve produced by considering the minimum scores for each video leads to a very low EER, even better than the average case, equivalent to 0.204:



As can be easily observed, the DET curves just shown and referring to the morphing task turn out to be much more reasonable than those produced for the face recognition task, and this is a consequence of the fact that it is the starting scores (which we recall were produced with an SVM classifier) that turn out to be more robust and realistic than the simple cosine distance. As proof of this, one only needs to look at the distribution of mean scores in both the genuine and impostor cases, which, while in the face recognition case showed often non-overlapping or barely overlapping scores (as reported and evidenced by the images of the distribution of scores in both the genuine and impostor cases in the previous chapter) leading to a trivial classification with error rates close to 0, shows here instead a wide overlap of scores and consequently a more realistic and reasonable margin of error:

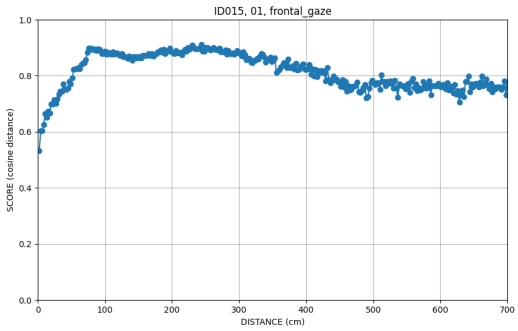


6.3.2 In-depth study about the variation of morphing detection accuracy as a function of distance

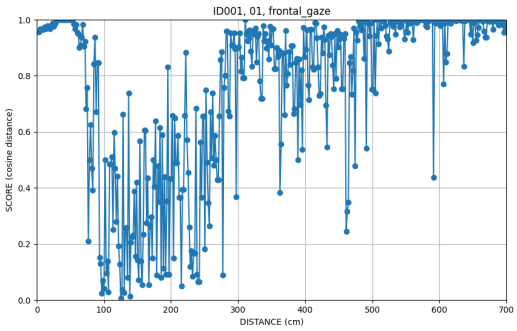
Proceeding then with the same pipeline used in the previous chapter in the case of the face recognition task, the next planned step required the study of the variation of morphing detection accuracy as a function of the distance of the face from the camera.

Well, comparing the graphs produced with those in the previous chapter reveals both similarities and differences:

- unlike the face recognition case, in the morphing detection task the scores for the same video show much more rapid and abrupt fluctuations, even with minimal variation in distance →

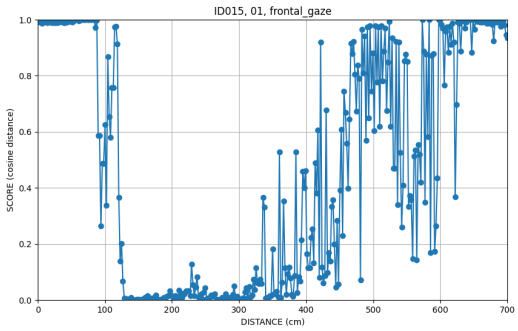


(g) scoring trend curve in the very first video of subject ID001 in the case of facial recognition

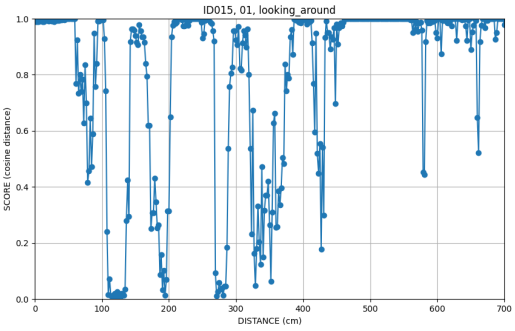


(h) scoring trend curve in the very first video of subject ID001 in the case of morphing detection

- in the case of morphing detection both morphing scores referring to videos in which the subject assumes frontal gaze and looking around poses fluctuate greatly, whereas in the case of facial recognition if the pose was frontal gaze the score was stable and only fluctuated if the pose was looking around → however it is interesting to note that in all videos in which the subject assumes frontal gaze pose the score stops fluctuating between 0.5 and 4 meters →

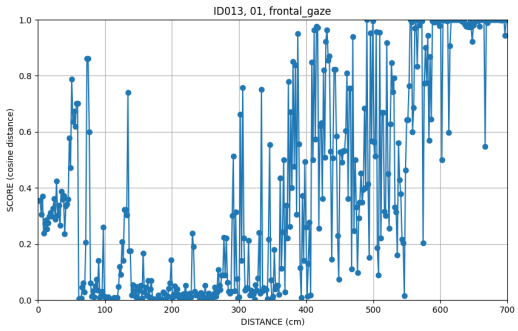


(i) curve of subject ID015 as it assumes frontal gaze pose

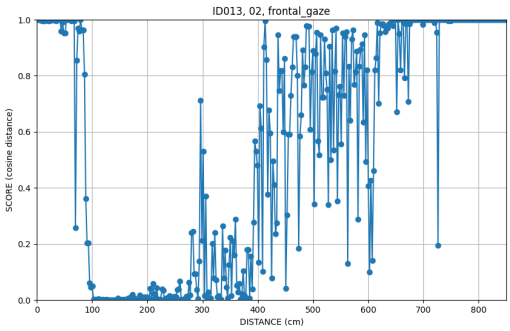


(j) curve of subject ID015 while assuming looking around pose

- unlike the case of facial recognition, in which the videos acquired in the second session showed much greater fluctuations than in the first session, this is not the case in the case of morphing detection, in which precisely the trend of the score curves remains almost the same regardless of the acquisition sequence → this shows that the scores computed for morphing turn out to be much more robust to changes in brightness than those calculated by simple cosine distance →

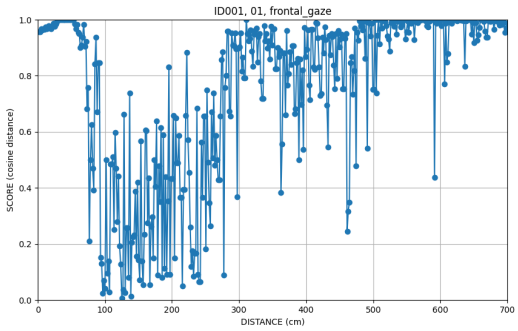


(k) curve of subject ID013 while assuming frontal gaze pose in session 01

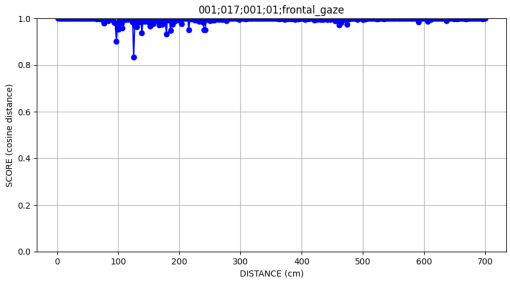


(l) curve of subject ID013 while assuming frontal gaze pose in session 02

- unlike the case of the facial recognition task, in which the score curves referring to genuine comparisons were much more stable than those referring to impostor comparisons, in the case of morphing it is the exact opposite, that is the curves referring to genuine comparisons are very erratic and have many fluctuations, while the score curves obtained from impostor comparisons are much more stable →

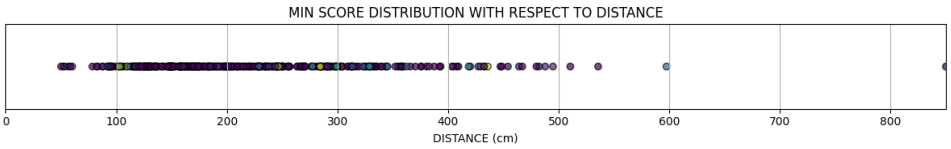


(m) scoring trend curve in the very first genuine comparison

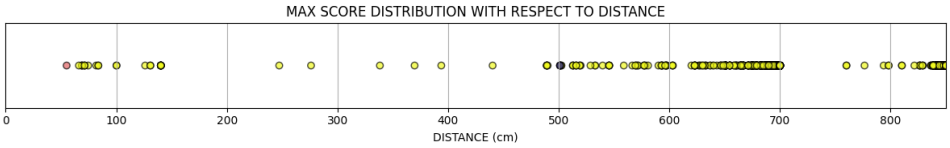


(n) scoring trend curve in the very first comparison impostor

- finally, as the two graphs below clearly show, an interesting peculiarity emerges, namely that a low distance between the subject and the camera (generally 0.5 to 4 meters) is necessary to establish with a high degree of confidence that the photo is not morphed, but instead it is at a medium to long distance (5 to 7 meters) that a photo can be classified as morphed with high confidence →



(o) distribution of minimum morphing scores as a function of distance for genuine comparisons



(p) distribution of maximum morphing scores as a function of distance for impostor comparisons

From the above comparisons, therefore, it appears that the morphing scores produced by the classifier described above are not only more accurate and reasonable, but also much more stable and robust to changes in distance, pose and illumination than the scores derived from the cosine distance.

Chapter 7

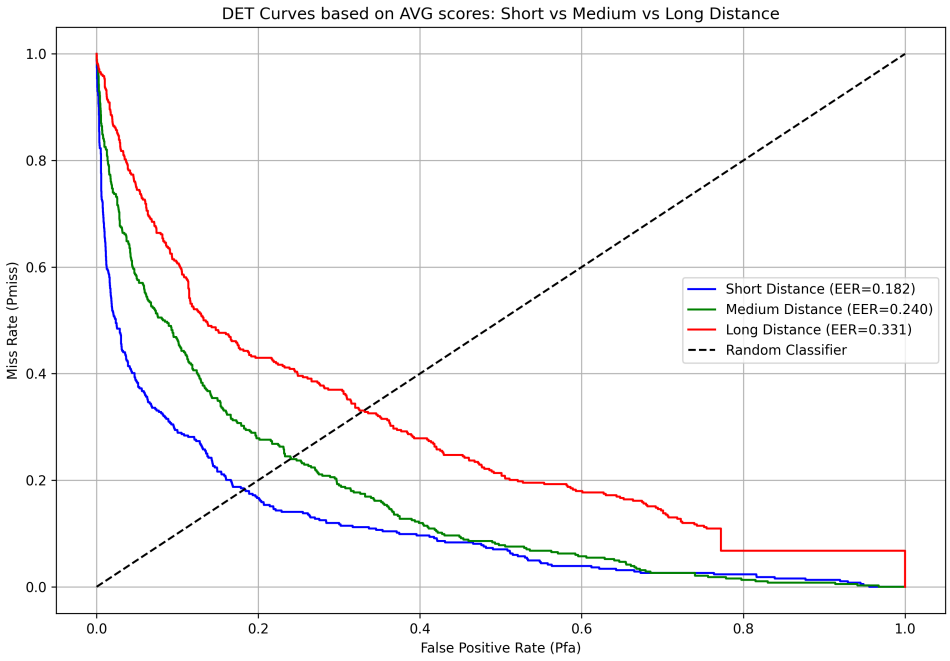
FINAL CONSIDERATIONS AND SPECIFIC METRICS

In this chapter we exploit task-specific metrics of Morphing Attack Detection and we will study how they change in relation to distance to extract final considerations about the usefulness of using videos and the aggregation techniques chosen.

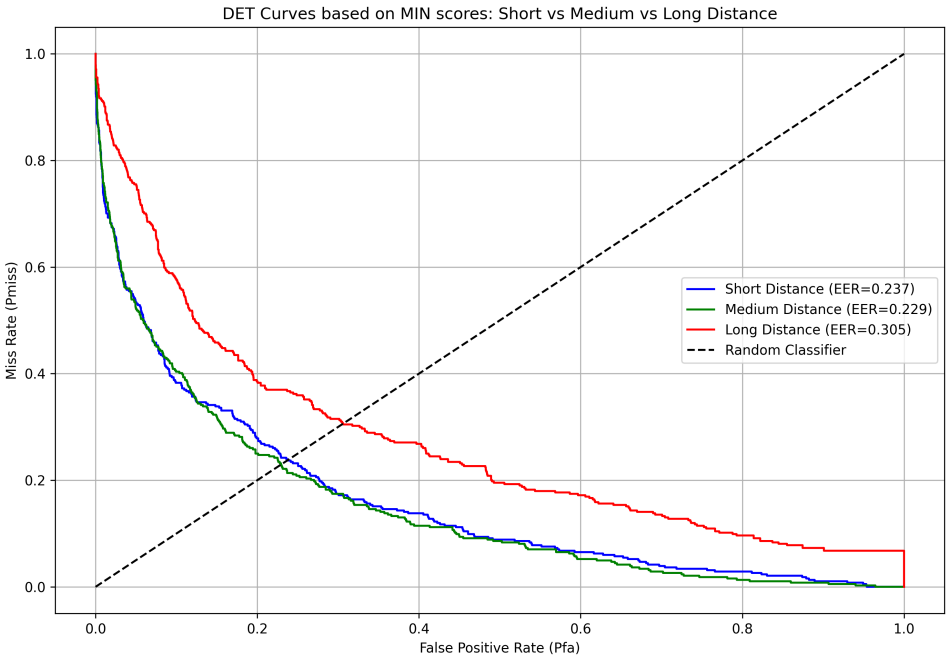
7.1 DET curves in relation to distance

The Detection Error Tradeoff curves, a tool already presented several times throughout this thesis, are particularly relevant when produced by considering the distance of the scores returned by the classifier.

That is, by dividing the scores returned according to the distance of the face into short distance (less than 2.5 meters), medium distance (between 2.5 and 5 meters) and long distance (more than 5 meters) the DET curves show, both aggregating by mean and minimum, how morphing recognition is more accurate and less prone to error if the distance is short or medium rather than long, as logic would suggest and as in fact the images below show, where the worst EER is in fact always found to be that at the long distance-based DET curve:



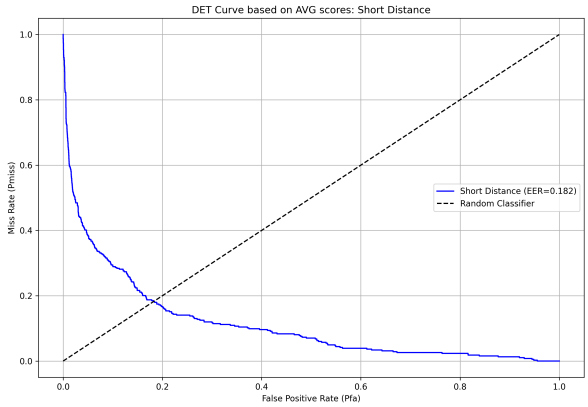
(a) DET curve as a function of distance considering the average scores per video



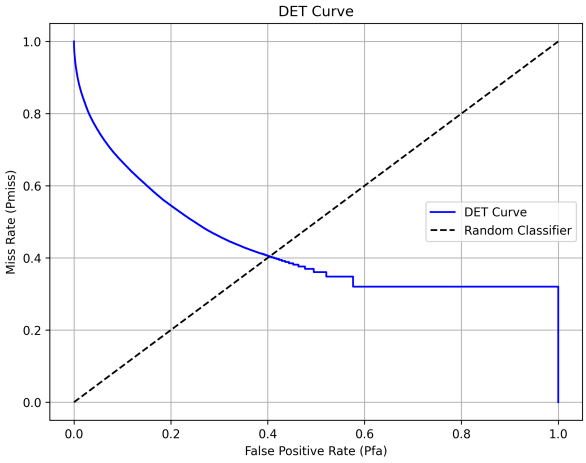
(b) DET curve as a function of distance considering the minimum scores per video

It should be noted, as was the case in previous chapters, that in the DET curves shown above, the one produced from the maximum scores per video was not included because it was totally non-significant, as it would report a very high error.

Finally, the DET curve below shows significantly and indisputably the great usefulness of aggregation operations.
In fact, if we were to consider the individual scores frame by frame, i.e. without any kind of aggregation, we would obtain decidedly degraded performance, characterized by an EER greater than 0.4, as opposed to an EER that, including simple aggregation operations based on average or minimum, we have shown to be in the best case equal to 0.18 and in the worst case 0.33:



(c) DET curve on short distance with min as aggregation



(d) DET curve without any kind of aggregation

Figure 7.1: comparison between DET curve produced with and without aggregation.

Chapter 8

CONCLUSIONS

It is appropriate to distinguish the final considerations between those concerning the present thesis and those concerning the termination of the academic course addressed.

With regard to the present thesis, it addressed the problem of morphing attack detection in the biometric domain by proposing and evaluating an innovative solution based on video analysis, called V-MAD (Video-based Morphing Attack Detection).
The entire process was developed along several stages, from data collection to in-depth analysis of classification metrics.

An initial significant contribution of this thesis was the construction from scratch of the Gaze-Way dataset, which was acquired, cleaned and labeled entirely independently, with attention to compliance with ICAO specifications and variability in subject, pose and environmental conditions.
This represented not only a technical challenge but also an essential step in the experimental validation of the proposed models.

The analysis of the results indicates that the embeddings generated with ArcFace are particularly effective in the face verification task, suggesting that the dataset used is relatively simple from this perspective.
In contrast, the morphing detection task appears to be more challenging, with embeddings showing greater robustness, realism and stability, especially when combined with supervised classifiers.

In particular, greater resilience and robustness of the system was observed with respect to variations in distance, illumination and subject pose.

A general observation is that the systems tend to perform better when the subject is at a limited distance.
As the DET curves and MAP matrices indicate, recognition accuracy decreases as the distance increases, with performance degrading significantly at higher ranges.
This distinction represents novel and valuable information that could guide future implementations in airport and security settings.

Finally, the thesis has shown that strategies for aggregating MAD scores from video sequences, even through simple aggregation methods such as mean and minimum, can outperform traditional D-MAD approaches in accuracy: in fact, they have resulted in an EER that, while far from optimality, is certainly very promising and worthy of further investigation and refinement. In fact, the inclusion of image quality indices has further improved the results, paving the way for future developments in which data quality can guide the interpretation of the biometric signal.

With regard to the conclusion of the academic course, however, it should be emphasized that this thesis was not only a research course, but also and above all an experience of personal, technical and human growth.
If as a result of my bachelor's degree I had discovered the importance of collaboration and discussion, today I feel I have taken a further step in the maturation of my way of thinking and working.

The difficulty of starting from raw data, the initial uncertainty for what concerns the scarce literature on V-MAD, the practical problems encountered and solved are all elements that, today, I recognize as decisive moments of learning: in fact, one must always remember that it is not only in success that one grows, but also and above all in error, obstacle and restart.

Working on a real, innovative problem without pre-packaged solutions made me realize how much the work of the modern engineer/researcher can no longer be reduced to the sterile application of technical notions.

It requires creativity, critical thinking, but also the ability to communicate, collaborate and adapt. In an increasingly digital and interconnected world, where data travel but people often remain distant, the figure of the engineer must know how to be open to dialogue, context and the social responsibility of his research.
Living in today's nontrivial and multifaceted world requires accepting and attempting to understand its complexity, identifying and translating abstract patterns that emerge into concrete solutions to real problems, and above all sharing one's knowledge so that progress is collective and not just individual.

So I end this hard but fascinating journey with gratitude and awareness: gratitude to those who have accompanied and guided me and awareness that I am only at the beginning of a journey in which what really makes a difference is not only what one knows, but what one chooses to do with what one has learned.

Bibliography

[1] Matteo Ferrara, Annalisa Franco, Davide Maltoni, *"The Magic Passport"*

[2] Matteo Ferrara, Annalisa Franco, *"Morph Creation and Vulnerability of Face Recognition Systems to Morphing"*, 2022

[3] Gomez M., *"Is your biometric system robust to morphing attacks?"*, 2017

[4] ICAO, *"Biometric Deployment of Machine Readable Travel Documents"*, 2004

[5] M. Ferrara, A. Franco, D. Maio, D. Maltoni, *"Face Image Conformance to ISO/ICAO standards in Machine Readable Travel Documents"*, IEEE, 2012

[6] M. Ferrara, A. Franco, D. Maltoni, Y. Sun, *"On the Impact of Alterations on Face Photo Recognition Accuracy"*, 2013

[7] Ferrara M., Franco A., Maltoni D., *"On the effects of image alterations on face recognition accuracy"*, 2016

[8] Scherhag U., Rathgeb C., Merkle J., Breithaupt R., Busch C., *"Face recognition systems under morphing attacks: a survey"*, IEEE, 2019

[9] IATA, *"Airport with Automated Border Control Systems"*, 2014

[10] Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, Christoph Busch, *"Face Morphing Attack Generation & Detection: A Comprehensive Survey"*, 2020

[11] Gomez M., *"Predicting the vulnerability of biometric systems to attacks based on morphed biometric information"*, 2018

[12] FRONTEX-Research and Development Unit, *"Best Practice Technical Guidelines for Automated Border Control (ABC) Systems"*, 2012

[13] N. Damer, A. M. Saladie, S. Zienert, Y. Wainakh, P. Terh, F. Kirchbuchner, A. Kuijper, *"To detect or not to detect: The right faces to morph"*, 2019

[14] A. Rottcher, U. Scherhag, C. Busch, *"Finding the suitable doppelganger for a face morphing attack"*, 2020

[15] Valetine J., *"Optimizing blending operations"*, Hydrocarbon Engineering, 2006

[16] O. Celiktutan, S. Ulukaya, B. Sankur, *"A comparative study of face landmarking techniques"*, 2013

[17] Seibold C., *"Detection of face morphing attacks by deep learning"*, 2017

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, *"Generative adversarial nets"*, 2014

[19] Yu H., Liao X., *"Free form deformation image registration using edge information measure"*, 2008

[20] Yu C., Chen X., Xie Q., Li G., Yin L., Han H., *"Image deformation using modified moving least squares with outlines"*, IEEE, 2017

[21] Rogers D.F., Adams J.A., *"Mathematical elements for computer graphics"*, 1989

[22] Wolberg G., *"Digital image warping", IEEE, 1994*

[23] Zinelabidine Boulkenafet, Jukka Komulainen, Abdenour Hadid, *"Face spoofing detection using colour texture analysis"*, IEEE 2016

[24] Y. Weng, L. Wang, X. Li, M. Chai, K. Zhou, *"Hair interpolation for portrait morphing"*

[25] Venkatesh S., *"Can gan generated morphs threaten face recognition systems equally as landmark based morphs?"*, 2020

[26] Makrushin A., Neubert T., Dittmann J., *"Automatic Generation and Detection of Visually Faultless Facial Morphs"*, 2017

[27] Young A.W., Burton A.M., *"Recognizing faces"*, 2017

[28] K. Raja, M. Ferrara, A. Franco, L. Spreeuwers, I. Batskos, F. De Wit, M. Gomez-Barrero, U. Scherhag, D. Fischer, S. Venkatesh, *"Morphing attack detection-database, evaluation platform and benchmarking"*, IEEE, 2020

[29] Matteo Ferrara, Annalisa Franco, Davide Maltoni, *"Face morphing detection in the presence of printing/scanning and heterogeneous image sources"*, 2021

[30] Raghavendra R., Raja K. B., Busch C., *"Detecting morphed face images"*, 2016

[31] Spreeuwers L., Schils M., Veldhuis R., *"Towards robust evaluation of face morphing detection"*, 2018

[32] Ministry of the Interior National Office for Identity Data and Kingdom Relations, *"State of the art of morphing detection"*, 2020

[33] Robertson D.J., *"Morphed passport photo detection by human observers"*, 2020

[34] Robertson D.J., Kramer R.S.S., Burton A.M., *"Fraudulent ID using face morphs: experiments on human and automatic recognition"*, 2017

[35] Scherhag U., Rathgeb C., Busch C., *"Morph detection from single face image: a multi-algorithm fusion approach"*, 2018

[36] Du X., Zhang R, *"Fusing color and texture features for blurred face recognition"*, 2014

[37] M. Rabbani, *"Jpeg2000: Image compression fundamentals, standards and practice"*, Journal of Electronic Imaging, 2002

[38] Raghavendra R., Li G., *"Multimodality for reliable single image based face morphing attack detection"*, IEEE, 2022

[39] S. Venkatesh, R. Ramachandra, K. Raja, L. Spreeuwers, R. Veldhuis, C. Busch, *"Morphed face detection based on deep color residual noise"*, IEEE, 2019

[40] K. Raja, S. Venkatesh, R. Christoph Busch, *"Transferable deep-cnn features for detecting digital and print-scanned morphed face images"*, IEEE, 2017

[41] Venktatesh S., *"Multilevel fusion of deep features for face morphing attack detection"*, 2022

[42] Neubert T., *"Face morphing detection: an approach based on image degradation analysis"*, 2017

[43] Singh P., Chandler D. M., *"F-MAD: A feature-based extension of the most apparent distortion algorithm for image quality assessment"*, 2013

[44] Ferrara M., Franco A., Maltoni D, *"Face demorphing"*, IEEE, 2018

[45] Scherhag U., *"On the vulnerability of face recognition systems towards morphed face attacks"*, 2017

[46] Guido Borghi, Gabriele Graffieti, Annalisa Franco, Davide Maltoni, *"Incremental Training of Face Morphing Detectors"*

[47] G. Graffieti, G. Borghi, D. Maltoni, *"Continual learning in real-life applications"* IEEE, 2022

[48] Tian Li, Anit Kumar Sahu, Ameet Talwalkar Virginia Smith, *"Federated learning: Challenges, methods, and future directions"*, 2019

[49] Z. Li, D. Hoiem, *"Learning without forgetting"*, IEEE, 2017

[50] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *"Overcoming catastrophic forgetting in neural networks"*, 2017

[51] Guido Borghi, Annalisa Franco, Nicolò di Domenico, Matteo Ferrara, Davide Maltoni, *"V-MAD: Video-based Morphing Attack Detection in Operational Scenarios"*, 2024

[52] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, *"Magface: A universal representation for face recognition and quality assessment"*, 2021, IEEE

[53] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer, *"Cr-fiqa: Face image quality assessment by learning sample relative classifiability"*, 2023, IEEE

[54] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, A. Kuijper, *"Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness"*, 2020, IEEE

[55] *"ISO/IEC 29794-5, Information technology, Biometric sample quality, Part 5: Face image data. "*

[56] Zander W. Blasingame, Chen Liu, *"Leveraging Diffusion For Strong and High Quality Face Morphing Attacks"*, IEEE, 2023

[57] Kaiwen Zhang, Yifan Zhou, Xudong Xu, Xingang Pan, Bo Dai, *"DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing"*, IEEE, 2023

[58] Marcel Grimmer, Christoph Busch, *"LADIMO: Face Morph Generation through Biometric Template Inversion with Latent Diffusion"*, IEEE, 2024

[59] Y. Wong, S. Chen, S. Mau, C. Sanderson, B.C. Lovell, *"Patch-based Probabilistic Image Quality Assessment for Face Selection and Improved Video-based Face Recognition"*, 2011, IEEE

[60] P J. Phillips, J. R. Beveridge, David Bolme, Bruce A. Draper, Geof H. Givens, Yui M. Lui, Hao Zhang, W T. Scruggs, Kevin W. Bowyer, Patrick J. Flynn, Su L. Cheng, *"The Challenge of Face Recognition From Digital Point-and-Shoot Cameras"*, 2013, IEEE

[61] Multiple Biometric Grand Challenge (MBGC) Version 2 Data collection, *"https://www.nist.gov/programs-projects/multiple-biometric-grand-challenge-mbgc"*, 2006

[62] Julien Vitay, *"YouTubeFacesDB"*

[63] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Anil K. Jain, *"Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A"*, 2015

[64] Ankan Bansal, Anirudh Nanduri, Carlos Castillo, Rajeev Ranjan, Rama Chellappa, *"UMDFaces: An Annotated Face Dataset for Training Deep Networks"*, 2016

[65] ISO/IEC 19794-5, Information technology, *"Biometric data interchange formats - Part 5: Face image data"*, 2011

[66] M. Ferrara, A. Franco, D. Maio, D. Maltoni, *"Face Image Conformance to ISO/ICAO standards in Machine Readable Travel Documents"*, IEEE, 2012

[67] Arnaldo Gualberto de Andrade e Silva, Herman Martins Gomes and Leonardo Vidal Batista, *"A collaborative deep multitask learning network for face image compliance to ISO/IEC 19794-5 standard"*, 2022

[68] D. Maio, D. Maltoni, *"Real-time face location on gray-scale static images"*, 2000

[69] Viola, M. Jones, *"Rapid object detection using a boosted cascade of simple features"*, IEEE, 2001

[70] N. Eveno, A. Caplier, P.Y. Coulon, *"A new color transformation for lips segmentation"*, IEEE, 2001

[71] Deng Jiankang, Guo Jia, Xue Niannan, Zafeiriou Stefanos, *"Arcface: Additive angular margin loss for deep face recognition"*, IEEE, 2019

[72] Laurens van der Maaten, Geoffrey Hinton, *"Visualizing Data using t-SNE"*, JMLR, 2008

[73] Matteo Ferrara, Annalisa Franco, Davide Maltoni, *"Decoupling texture blending and shape warping in face morphing"*, BIOSIG, 2019

[74] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Christoph Busch, *"Deep Face Representations for Differential Morphing Attack Detection"*, IEEE, 2020

[75] Matteo Ferrara, Annalisa Franco, Davide Maltoni, Christoph Busch, *"Morphing Attack Potential"*, IEEE, 2022

[76] Qiaoyun He, Zongyong Deng, Zuyuan He, Qijun Zhao, *"Optimal-Landmark-Guided Image Blending for Face Morphing Attacks"*, IJCB2023, 2024

[77] Guilherme Schardong, Tiago Novello, Hallison Paz, Iurii Medvedev, Vinícius da Silva, Luiz Velho, Nuno Gonçalves, *"Neural Implicit Morphing of Face Images"*, CVPR2024, 2024

[78] Una M. Kelly, Meike Nauta, Lu Liu, Luuk J. Spreeuwers, Raymond N. J. Veldhuis, *"Worst-Case Morphs using Wasserstein ALI and Improved MIPGAN"*, IEEE, 2023

[79] Raghavendra Ramachandra, Sushma Venkatesh, Naser Damer, Narayan Vetrekar, Rajendra Gad, *"Multispectral Imaging for Differential Face Morphing Attack Detection: A Preliminary Study"*, IEEE, 2023

[80] Nicolò Di Domenico, Guido Borghi, Annalisa Franco and Davide Maltoni, *"Combining identity features and artifact analysis for Differential Morphing Attack Detection"*, ICIAP, 2023

[81] Ria Shekhawat, Hailin Li, Raghavendra Ramachandra, Sushma Venkatesh, *"Towards Zero-Shot Differential Morphing Attack Detection with Multimodal Large Language Models"*, IEEE, 2025.