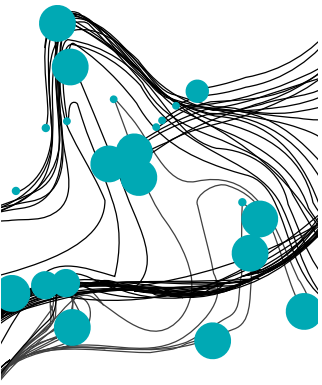
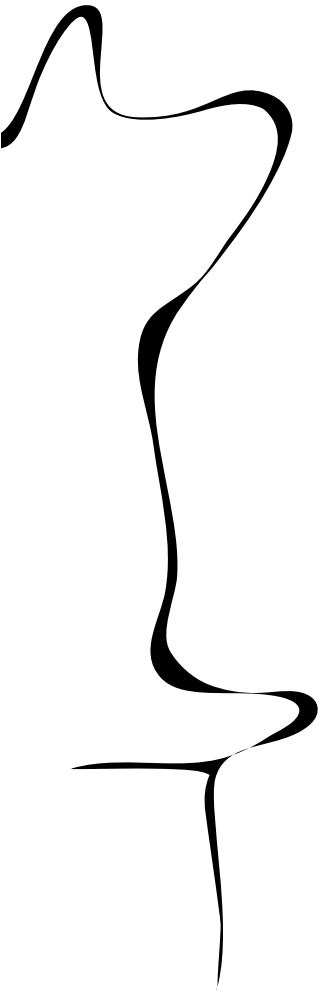


.55711

DMB

DATA MANAGEMENT AND BIOMETRICS

MORPHING ROBUST FACE RECOGNITION



A.D.Gavra

BACHELOR'S ASSIGNMENT

Committee:
L. Spreeuwers

July, 2025

Data Management and Biometrics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Morphing Robust Face Recognition

ANA DARIA GAVRA, University of Twente, The Netherlands

Applications for safe access control, such as those in border management, rely on systems based on face, fingerprint, or iris recognition. As face recognition systems become increasingly prevalent, their vulnerability to morphing attacks has become a relevant topic of discussion. An example of a morphing attack is in airports, with some countries permitting the release of electronic Machine Readable Travel Documents with photos submitted by the applicant which are subsequently approved or rejected. By examining ways to avert morphing attacks (i.e., techniques to reject or accept an image), this paper proposes a system that considers how various image quality properties may influence morph rejection. The threshold used in the decision-making process, as well as the final prediction score, is based on image quality properties combined with the prediction score given by a face recognition system. In doing so, the possible relationship between an image's quality properties, its similarity score, and BRISQUE is investigated. Further, to assess the validity of our performance enhancement claims, the chosen face recognition system will be tested both with and without taking image quality properties into account.

Additional Key Words and Phrases: Morphing Attacks, Biometrics, Face Morphing, Image Quality Assessment, BRISQUE

1 INTRODUCTION

Over the past decades, electronic documents storing biometric information have gradually replaced paper documentation [7]. Within this context, face, fingerprint, or iris recognition biometric systems are largely deployed in various access control applications. Face recognition systems (FRS) are primarily deployed in the context of authenticating users through ID verification services [30]. An example of such a service is the Automated Border Control (ABC) systems usually found in airports. The way ABC systems (or eGates) work is that the passenger presents their electronic Machine Readable Travel Document (eMRTD) to the ABC to verify their identity [23, 29]. Although identity can be confirmed by comparing it to the image on the document, many countries allow applicants to submit self-taken photos for their eMRTD, which are subsequently accepted or not. This aspect makes face recognition systems vulnerable to face morphing attacks [22].

Morphing is a visual effect commonly used in films and animations that smoothly transforms one image or shape into another [31]. A face morphing attack utilizes this technique to blend two or more images of different individuals–a criminal and an accomplice–to fool face recognition systems. An example of this can be seen in Figure 3.

In assessing whether an image is bona fide or morphed, the decision to accept or reject is made using a pre-defined threshold. The primary issue at hand is that choosing this threshold has its advantages and disadvantages, depending on the point of view, which is illustrated in Figure 1. On the one hand, setting a very high threshold can result in successfully rejecting a high number of morphs, as shown at the top; however, this comes with the downside of rejecting many innocent individuals as well. On the other hand, looking at the bottom figure, setting a low threshold could mean the acceptance of a higher number of criminals (i.e., morphs) [27]. Factors that influence the decision-making process that can be overlooked are image quality properties, such as sharpness, blurriness, and quality. For example, image compression is an aspect that could influence whether an image is rejected or accepted [23].

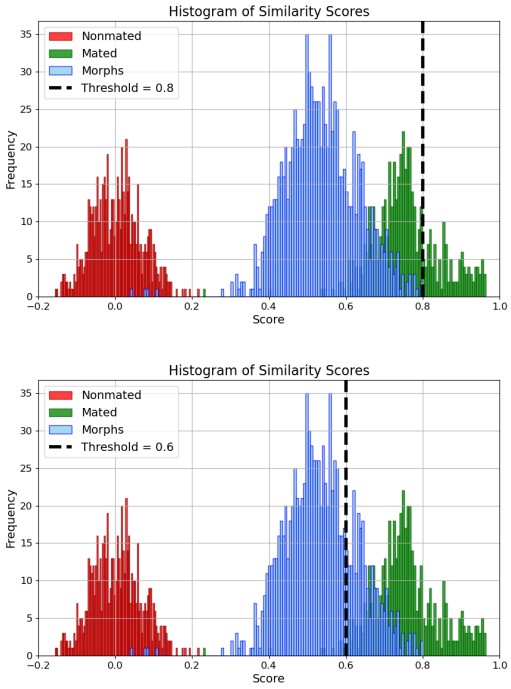


Fig. 1. Example illustrations of varying thresholds: the top figure shows a high threshold of 0.8 and the bottom one shows a lower threshold of 0.6

To address this issue, we propose a system that takes into account the quality properties of the input images in the decision-making process, aiming to achieve better performance in morph rejection. Figure 2 illustrates the proposed system's architecture.

Considering this, the project is centered around the following research question and its corresponding sub-questions:

RQ: How can image quality properties be used to enhance the effectiveness of a face recognition system in rejecting morphs?

(1) **SRQ1:** How is the effectiveness of morph rejection measured?
(2) **SRQ2:** What image quality properties should be taken into account to enhance morph rejection?
(3) **SRQ3:** How much do image property metrics influence the decision-making process of morph rejection?

Author's address: Ana Daria Gavra, a.d.gavra@student.utwente.nl, University of Twente, P.O. Box 217, Enschede, The Netherlands, 7500AE.

TScIT 43, July 4, 2025, Enschede, The Netherlands

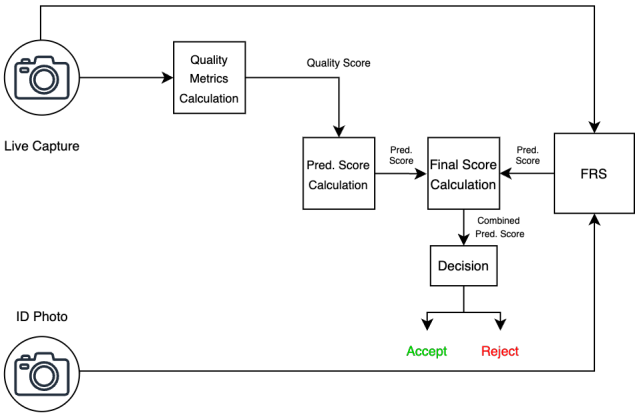


Fig. 2. Overview of proposed system architecture

2 BACKGROUND

Creating a morph comprises three main stages. The first stage involves establishing a *correspondence* between the samples being used. The following stage, known as *warping*, involves the distortion of the sample images such that their corresponding features align geometrically. Lastly, in the *blending* stage, the color values of the warped images are combined to produce the final morphed face image. Besides the three stages presented, some post-processing steps can be taken to improve the morph's quality. For example, blurring and sharpening could be used to eliminate unnatural color gradients and edges, and using histogram equalization can help achieve realistic color histogram shapes. The aforementioned post-processing steps aid in achieving a higher quality morph and introduce a higher probability of a criminal passing ABC gates [23].

Looking at the generic structure of an automated biometric system, these are usually comprised of the following [18]:

- A live capture device (e.g., a camera),
- A database with biometric information and other personal data,
- A feature extraction algorithm,
- Comparison and decision algorithms that establish whether the two samples (i.e., ID photo and live capture) belong to the same source.

As is described in [9], during enrollment at an ABC gate, a feature vector is extracted from the ID photo, which serves as a reference sample, and a live capture. The feature vector obtained from the live capture is compared against the reference (claimed identity), resulting in a final biometric comparison score. The aforementioned score is compared against a pre-defined threshold yielding acceptance or rejection.

There is an alternative scenario in which no live capture device is present; in such cases, the decision to reject or accept is strictly based on the ID photo and the manual check of its validity conducted by ABC gate personnel [21]. But the task of manual checking is becoming more and more difficult as the quality of morphs is increasing. Looking at the three images in Figure 3, to the human eye, it seems fair to assume that the photo in the middle can be the passport photo of the individual on the left or the one on the right, whereas, in fact, it is a morph of both. Therefore, this paper will only focus on the case in which a live capture is present for extra information.

Individual 1 Passport Photo Individual 2

Fig. 3. Example of how two different individuals can use a morph as a valid passport photo (passport photo is a morph of the two individuals)

As mentioned prior, the proposed system will take into account an ID photo and a live capture of the individual present. When someone applies for an eMRTD, the image they provide has to meet specific quality standards. Consequently, an assumption that can be made is that the ID photo that will be inputted into the system will respect these standards. An overview of ICAO standards [10], which are widely used and referenced, can be seen in Table 1. Analyzing these specifications provides a good starting point for identifying the key quality properties to be studied in this research.

Requirement	**ICAO Specification**
Photo Age	Must not be older than 6 months
Photo Size	35–40mm width
Framing	Face must take up 70–80% of photo
Quality	Photo must be clear and of sharp focus, have appropriate brightness and contrast, be taken with uniform lighting and not show shadows or light reflections
Coloring	Photo must show natural and uniform skin color
Resolution	Photo needs to be of high resolution
Face Expression	Subject must have neutral expression, with eyes open and clearly visible, mouth closed
Subject Positioning	Subject must be looking straight at camera, front facing and must be centered in image
Background	Photo must have a light uniform background
Glasses	Glasses can be worn if frames are thin and lenses are not reflective
Head Coverings	Can be worn for religious reasons, but facial features from bottom of chin to top of forehead and both face edges must be clearly visible

Table 1. Summary of ICAO passport photo requirements

To measure the degradation of an image's quality, image quality assessment (IQA) techniques are typically employed. Depending on the amount of information available about the original image, IQA metrics range from full-reference (i.e., the original is available for degradation comparison) to no-reference (i.e., no original is available). An example of a widely used no-reference IQA metric is BRISQUE [15], a natural scene statistic (NSS) based model, which quantifies the "naturalness" of an image by measuring its deviation from the expected patterns of undistorted images.

3 RELATED WORK

Although this paper looks into enhancing morph rejection, related work in the areas of morph detection and face recognition offers great insights into how image quality properties influence the acceptance/rejection decision-making process.

Ghost or blur artifacts are residuals of face morphing caused by misplaced landmarks. These artifacts are strong indicators of a morphed image. Szabó's [27] research into face recognition tackles the effect of warping in creating a more robust system. By warping all faces to a standard "average face", there is a possibility of more easily finding these artifacts. Further, other image quality properties, such as blurriness, sharpness, and brightness, can have an effect in better identifying morphing artifacts.

Another aspect that has been treated in trying to spot morphing artifacts is the color histogram of the provided image, as some morphs prove to have uneven color histograms. As shown in [28], by leveraging denoising techniques in the HSV color space, exposing morphing artifacts proves to be possible.

Additionally, image quality has proven to have significant influence on performance. Scherhag et al. [22] show that, from a digital perspective, morph attack detection systems are robust enough to detect morphs. However, after the same images are printed and scanned (such as the case of ABC systems checks in airports), they have a very low success rate. The compression that an image undergoes affects its BRISQUE score, with morphed images having very high BRISQUE values after compression [23]. Further, recent research shows that FRS are susceptible to images of lower resolution, one of the reasons being the lack of low resolution training data [4, 12]. Besides issues with training data, [11] points out that most FRS treat all data the same, without taking into account quality aspects, thus resulting in poorer performance for lower quality images.

4 PROPOSED METHODOLOGY

4.1 System Overview

In contrast to the state of the art, as previously discussed, the system proposed (Figure 2) incorporates quality aspects. From Figure 2, it can be seen that extra modules have been added, specifically the "Quality Metric Calculation" and "Prediction Score Calculation". The purpose of these two modules is the following: first, extract from the live capture its relevant quality aspects and generate a quality score, and second, output a prediction score for the system based on this. Moving forward, in the process to accept or reject, both the quality prediction score and the score generated by the FRS will be considered.

The methodology for performance metrics, face recognition systems, the dataset used, the decision-making process, image quality and prediction score calculation are elaborated upon in the following subsections.

4.2 Performance Metrics

In biometric systems and morphing attacks, various metrics evaluate effectiveness. Since this project focuses on measuring morph rejection effectiveness, rather than morph detection, metrics such as Attack Presentation Classification Error Rate (APCER) and Bonafide Attack Presentation Classification Error Rate (BPCER) are set aside, as these are meant for assessing Morphing Attack Detection (MAD) algorithm accuracy.

In assessing the effectiveness of a system, two aspects need to be considered: its behavior and overall vulnerability. To gain insights into the system's behavior, the following metrics will be utilized [18]:

- **False Match Rate (FMR)**: The proportion of the completed biometric non-mated comparison trials that result in a false match.
- **False Non-Match Rate (FNMR)**: The proportion of the completed biometric mated comparison trials that result in a false non-match.

Additionally, we will consider the True Non-Match Rate (TNMR) and True Match Rate (TMR), which, by definition, are the complements of the False Match Rate (FMR) and False Non-Match Rate (FNMR), respectively. All aforementioned metrics can be visualised using:

- **Receiver Operating Characteristic (ROC)**: plots TMR against FMR [33],
- **Detection Error Trade Off (DET)**: plots FMR against FNMR [32].

To evaluate the overall vulnerability of the system, we will use the Mated Morph Presentation Match Rate (MMPMR), which is defined as follows in [21]:

$$MMPMR(\tau) = \frac{1}{M} \cdot \sum_{m=1}^{M} \left\{ \left[\min_{n=1,\ldots,N_m} S_m^n \right] > \tau \right\},$$

where τ is the verification threshold, S_m^n is the mated morph comparison score of the n-th subject of morph m, M is the total number of morphed images and N_m the total number of subjects contributing to morph m.

4.3 Face Recognition System

Due to variations in behavior and implementation, this project will only focus on one FRS and its associated behavior. Consequently, a literature review of the state of the art has been performed in order to make a final decision within this research's scope.

There is a series of key properties that need to be looked into to properly assess what the best choice would be: the size and complexity of the system, its accuracy, the datasets it has been trained on, its performance with occluded faces, and how it reacts to various photo quality. All chosen systems respect Frontex guidelines (i.e. False Acceptance Rate ≤ 0.01, False Rejection Rate ≤ 0.5) [1], and had their performance tested on the same data, so we will use the same to compare them between each other. Table A.1 presents an overview of all systems reviewed. The table is constructed using the information given by each model's paper.

Based on all the presented information, the project will use AdaFace [11], since recent papers show that it outperforms its popular peers, such as ArcFace. Moreover, AdaFace was designed with considerations for face occlusion and varying image quality, two areas highly relevant to the goals of this project.

4.4 Dataset

To replicate the real life scenario of eGates, the dataset used needs to contain two types of images, one which adheres to the standards specified in [10], representing the passport photo, and another representing the live capture. For the latter, this could mean:

- Harsh face shadows,
- Busy background,
- Varying expression, (e.g., the subject is smiling)
- Occlusion of the face (i.e., the subject might not be looking straight at camera, may be wearing accessories, or may have hair covering their face).

The dataset that will be used is FRLL-Morphs [19, 20], which uses faces sourced from the Face Research London Lab [5]. One advantage of this dataset is that it offers morphs of people that resemble one another. For example, there are no morphs of a man and a woman, or of an elderly person and a young person. Additionally, the dataset provides images of people with various hairstyles, facial hair, makeup, and from different ethnic backgrounds and age groups. FRLL-Morphs only includes front-facing images with either a smiling or neutral expression. Therefore, to satisfy the need of varying live capture photos, we are also using quarter profiles provided by [5].

An aspect that has to be noted is that in the photos provided by [5], all subjects are wearing a white T-shirt against a white background. This is an issue that can be overlooked, as the pipeline that is being used crops the images with face alignment before being fed into the FRS. Therefore, we can ensure that we are focusing only on image qualities in assessing the performance of our model. Examples of the dataset can be found in Appendix B. Pictures with a neutral expression with $[0°, 0°, 0°]$ position are a proxy for a passport photo, whereas the rest serve as proxies for live captures.

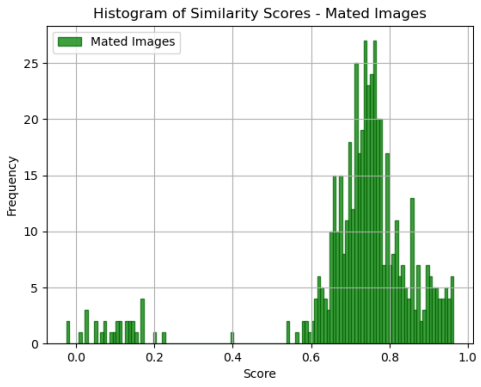


Fig. 4. Similarity scores for mated images

A preliminary data cleanup was conducted to ensure that results are not affected moving forward. As shown in Figure 4, some images perform poorly even without manipulations. Therefore, these have been removed from the dataset, as they will likely skew experimental results. Additionally, photos that do not adhere to ICAO regulations, such as bangs covering relevant marks of the face, have also been removed.

4.5 Image Quality Properties

Based on the ICAO photo requirements that have been presented in section 2, but also the related work presented in section 3, the image quality properties that will be investigated are:

- Brightness,
- Sharpness,
- Contrast,
- Noise.

Of the two images that are going to be inputted into the system, the passport photo is assumed to attain to quality standards specified for eMRTD. Therefore, only live capture photos will be manipulated. By manipulation we mean that a varying quality factor was applied to the original photo to generate a new one. The new photo is then inputted into the system to get a new a similarity score. Examples of manipulation can be seen in Appendix C. In this phase of the project, only one type of manipulation is applied to a photo. This way, we can ensure that we can clearly see each manipulation type's effect. All results will be separated by pose, expression and manipulation type, so it can be clearly seen how each one affects model performance.

After all similarity scores are computed and the effect of each type of manipulation is visualised, one should be able to conclude which image quality properties should be taken into account moving forward, but also to categorize live capture images into High Quality images and Low Quality images, which are defined as follows:

- **High Quality (HQ) Image**: an image that we are sure will always perform well (i.e., always results in a similarity score higher than threshold τ),
- **Low Quality (LQ) Image**: an image that has varying behavior, performance depends on the case.

Moving further, since all manipulations are done manually, a mapping between HQ images, LQ images and an IQA metric needs to be made. If a relationship between the aforementioned cannot be found, we will resort to using a manual quality metric based on the manipulations we have performed that will act as a proxy.

4.6 Image Quality Assessment

This project will only focus on no-reference IQA metrics. Most available metrics integrated with PyTorch are NSS based models and have a low complexity, therefore not increasing the size of the project by much. While there are also training-based models, such as CORNIA [35], they are not as easy to integrate. There is also the option of using deep features for IQA, recent papers showing that deep features are highly correlated with NSS based metrics and are highly accurate [11, 36]. However, incorporating deep features may complicate our existing architecture for similarity scores, and may limit our control over image manipulation processes, resulting in reduced interpretability of the results.

One last aspect that has to be treated is using individual metrics for the image quality properties that will be tested. Since there are already metrics that take into account multiple image quality properties in score calculation, adding an individual metric for each

image quality property would make the project more complex than necessary.

Taking everything into account, for this project, we will be using blind/referenceless image spatial quality evaluator (BRISQUE). BRISQUE has been chosen over other IQA metrics, such as NIQE [16] and BIQI [17], because it is better suited for the kind of image quality properties we will be looking at during this project. Along with its easy integration, high use in academia and its architecture, BRISQUE proves to be the best fit for this project, as it makes the mapping between image quality and BRISQUE scores a facile process.

4.7 Setting the Threshold

The process of setting a threshold for the system to be tested is pretty straightforward. The threshold will be set using impostor pairs and a FMR of 0.01%, as stated in [1]. For our specific dataset, impostor pairs of morphs (passport photo) and genuine people (live capture) are fed into the FRS to get similarity scores. After all pairs have gone through the FRS, the threshold is the 99.99th percentile value, which is equivalent to FMR of 0.01%.

4.8 Testing the Proposed System

In order to properly assess the effectiveness of the proposed system, we first need the performance metrics of the FRS for morph rejection. Moving forward, there are two scenarios under which the proposed system will be tested.

- Scenario 1: Static Threshold
  After calculating the threshold for the system, we will simply compare the difference between adding the quality prediction score to the decision making process and without adding it. This way, we can see how much image quality can affect decision making.
- Scenario 2: Dynamic Threshold
  Since we are able to also distinguish between high quality images and low quality images, it is fair to assume that for higher quality images the threshold should be higher. Therefore, based on the quality metric, we will also set the threshold of the system for the decision-making process.

In both cases, the prediction score that will be compared against the threshold is an average of the quality metric prediction score and the FRS prediction score:

$$Final\ Score = \frac{Quality - Based\ Pred.\ Score + FRS\ Pred.\ Score}{2}$$

5 FINDINGS

5.1 Relationship between Quality Manipulation and Similarity Scores

Each photo in the dataset has been manipulated at various levels and put in the FRS to obtain a new similarity score, as stated in section 4. To gain an overall view of the effect of each manipulation type (i.e., brightness, contrast, sharpness, noise), the mean scores for the whole dataset have been plotted at each manipulation level. All results, divided by pose, expression, and type, can be seen in Appendix E.1.

Looking at the results, pose appears to be a significant factor in the performance of a FRS. Figure 5 illustrates how, in the case where only pose differs, similarity scores degrade quicker for 45° angles. From top to bottom, comparing the first and third graphs, even the highest point differs significantly. In each graph, the highest point is represented by the original, meaning level 1. For front-facing images, originals can achieve similarity scores close to 1, whereas quarter-facing images start around 0.7.

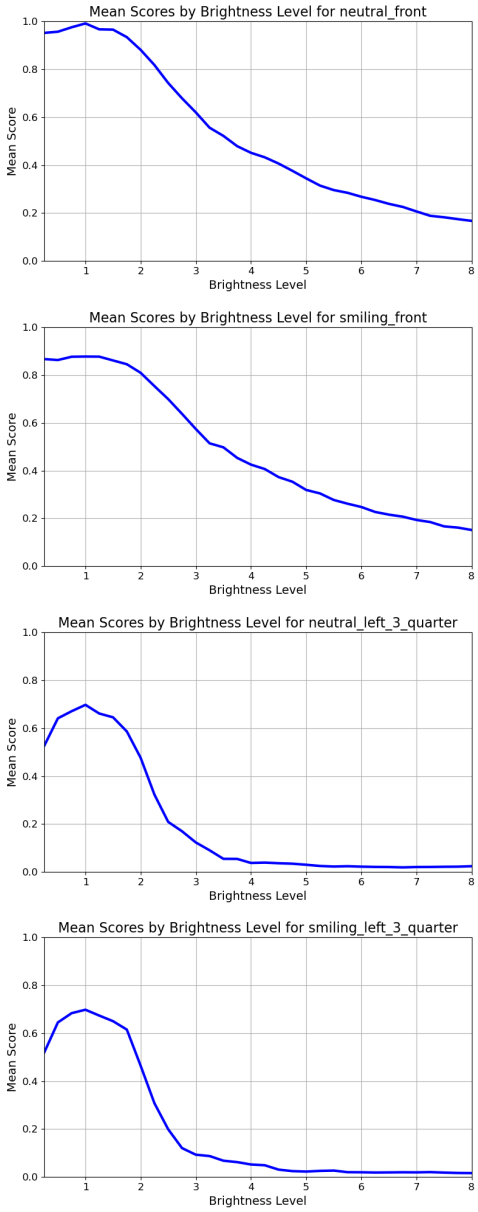


Fig. 5. Results of brightness manipulation impact over the dataset similarity scores, split by pose and expression

Furthermore, expression has a minimal effect on performance. In the case of quarter-facing images, the bottom two graphs in Figure 5 are almost identical. For front-facing images, there is a difference between the top two graphs for low manipulation levels; however, performance remains largely the same. It can be concluded that front-facing images are more robust against manipulations, with brightness having the biggest effect on them out of the four.

Considering all manipulation types, it is evident that all should be taken into account moving forward. Although front-facing photos are mainly affected by brightness manipulations, quarter-facing images prove to be susceptible to all four types. Consequently, moving forward, all results will be split by pose.

The results shown in Appendix E.1 offer a great starting point for certifying what HQ and LQ images are in the case of our dataset. By following the manipulation results in Appendix E.1, we can establish manipulation boundaries that enable us to certify whether an image is of high or low quality. In our case, high quality means that an image will always output a similarity score above 0.6, the threshold of our current system. Therefore, four new datasets have been created: one for HQ images and one for LQ images, with each dataset divided into front-facing images and quarter-facing images. One thing to note is that, in the case of LQ images, we have only gone outside of bounds slightly, as taking all possible values into account would result in mainly poorly performing images. For example, we will not consider images with a brightness level higher than 5, as this will always result in images with similarity scores below 0.1.

With some finetuning, Table 2 has been created. By following the intervals indicated, one should be able to create images that exhibit predictable behavior. The similarity score plots for all images created using Table 2 can be found in Appendix D. The similarity score distributions indicate that, in the case of front-facing images, behavior remains predictable, with the data distribution gradually shifting to the left as quality decreases (i.e., higher manipulation levels). In contrast, for quarter-facing images, the model exhibits bimodal behavior. As a result, in the case of quarter-facing images, we can only identify manipulation boundaries that will yield positive results, but we cannot reliably predict outcomes in other scenarios. This is unexpected, as the model we chose, AdaFace, is built keeping in mind face occlusion and image quality.

Quality Property	**High Quality (HQ)**		**Low Quality (LQ)**	
	$[0°, 0°, 0°]$	$[\pm45°, 0°, 0°]$	$[0°, 0°, 0°]$	$[\pm45°, 0°, 0°]$
Brightness	$(0.25,\ 3)$	$(0.75,\ 1.5)$	$(3,\ 4.5)$	$(1.5,\ 2)$
Sharpness	–	$(0.25,\ 2.5)$	–	$(2.5,\ 3)$
Contrast	–	$(0.75,\ 1.25)$	–	$(1.25,\ 2)$
Noise	–	$(0,\ 10)$	–	$(10,\ 15)$

Table 2. Finetuned intervals for generating HQ and LQ live captures. If an interval is not specified, any value can be used.

5.2 Relationship between Similarity Scores, Quality Properties and BRISQUE

Having established an understanding of the FRS and the impact various manipulations have on its performance, the potential relationship between similarity scores and IQA metrics, specifically BRISQUE, will be investigated. The expected outcome is for the distribution of BRISQUE scores to exhibit a behavior pattern similar to similarity score distributions. Specifically, it is anticipated that as image quality diminishes, its similarity score lowers, and its BRISQUE becomes higher.

Plotting the BRISQUE scores of the live captures used to create the visualizations in Appendix E.1 makes it fair to assume that this will be the case, as can be seen in the results of Appendix E.2. Looking at Figure 6, as quality manipulation levels increase, so does the BRISQUE score. By comparing the behavior displayed in Figure 5 to the one in Figure 6, it's clear that there is a relation between an image's BRISQUE and its similarity score. One particularly interesting aspect to note is that, in some cases, higher sharpness actually lowers BRISQUE.

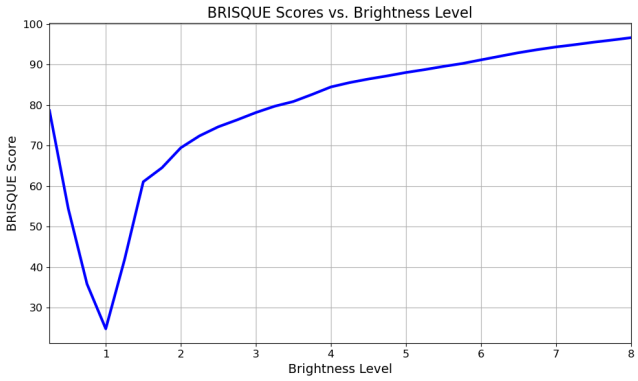


Fig. 6. Figure showing the relationship between the brightness level of an image and its BRISQUE score

Moving further, the BRISQUE scores of all live captures created using the boundaries specified in Table 2 were calculated and plotted, which yielded Figure 7. One aspect to note is that BRISQUE scores are much higher for front-facing images than quarter-facing ones. This is expected, as the manipulations applied to front-facing images are much harsher than the ones applied to quarter-facing ones.

In the case of front-facing images (top), there is a clear overlap between HQ and LQ images, despite their similarity score distributions not exhibiting the same behavior.

When examining the case of quarter-facing images (bottom), there is an overlap around a BRISQUE score of 25, which was expected due to the overlap between HQ and LQ image similarity scores, as visualized in Appendix D. Although the long tail in the LQ image distribution indicates that lower-quality images tend to have higher BRISQUE scores, this information is insufficient to draw a solid conclusion.

While it is clear that a relationship exists between the quality of our images, similarity scores and BRISQUE, this relationship cannot be used to predict the similarity scores an image will have. From this, it can be concluded that there isn't any relation between BRISQUE and similarity scores.

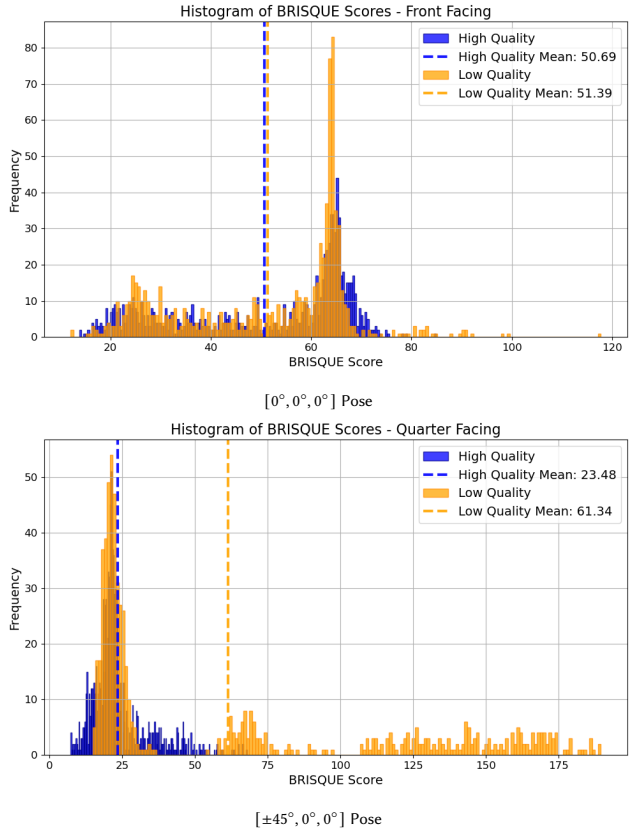


Fig. 7. BRISQUE score mappings of HQ and LQ images, separated by pose

5.3 Testing the System: Revisited

As it has been established that there is no relationship between similarity scores and BRISQUE, we will resort to calculating a manual quality metric using the information provided in Table 2. By using the manually manipulated data, we can easily assess what manipulations have been used and at what level. Consequently, we are able to calculate a quality metric that can act as a proxy for our proposed system.

In order to calculate this metric, weights have to be set for the four possible manipulations, as each has a different impact on the performance of a FRS. Therefore, we created five new live capture image datasets. In the first one, referred to as "default", all manipulations have been performed at various levels, while in the other four all but one manipulation has been performed. This way, we can establish the weight of each manipulation type based on how much the mean similarity score deviates from the default. The final formula that will be used to establish a quality metric for each photo is:

$$0.73B + 0.15C + 0.07S + 0.05N,$$

where B = Brightness Level Applied, C = Contrast Level Applied, S = Sharpness Level Applied, N = Noise Level Applied

Moving forward, a correlation between the quality metric and the prediction score of our FRS has to be found. Specifically, we need to find a function f, such that $s = f(q)$, where s is the prediction score and q is the quality metric. By training a polynomial model with the data we have available, meaning the quality score and associated prediction score of each photo in our dataset, we found $f(q) = -0.009793x^3 + 0.0995x^2 - 0.3488x + 0.9604$.

5.4 System Results

The distributions of the data that has been used for testing the proposed systems can be seen in Figure 8. All performance results for $FMR = 0.1$ and $FMR = 0.01$ have been summarized in Tables 3 and 4, respectively. One aspect to note is that, for calculating the MMPMR, a higher number of morphs than the ones showcased in Figure 8 has been used, but the threshold remained the same as specified in the tables. Looking at the relationship between the True Match Rate (TMR) and False Match Rate (FMR), some conclusions can be drawn.

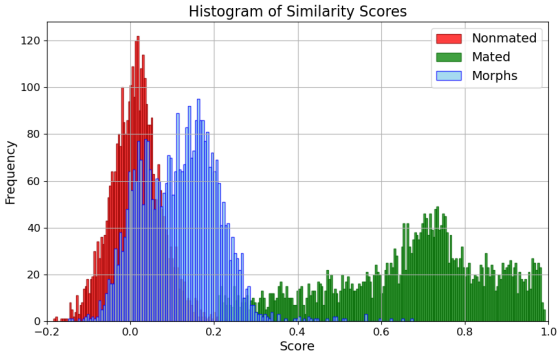


Fig. 8. Similarity score distributions for mated, nonmated, and morphs

Firstly, strictly comparing the systems with a static threshold, it is clear that taking quality metrics into account makes the FRS less robust, as the TMR decreased and the FMR increased. Additionally, this approach increased the vulnerability of our system, as it can be seen in the MMPMR column. Second, it can be seen that a dynamic threshold impacts TMR positively, but this comes at the cost of a higher FMR.

Method	TMR	FNMR	TNMR	FMR	MMPMR
Without QM	0.7434	0.2566	0.9990	0.0010	0.0025
With QM - Static	0.7182	0.2818	0.9958	0.0042	0.1022
With QM - Dynamic	0.8070	0.1930	0.9892	0.0108	–

Table 3. Performance metrics at FMR = 0.1, $\tau = 0.54$

Method	TMR	FNMR	TNMR	FMR	MMPMR
Without QM	0.6301	0.3699	0.9997	0.0003	0.0008
With QM - Static	0.5270	0.4730	0.9997	0.0003	0.0016
With QM - Dynamic	0.7247	0.2753	0.9963	0.0037	–

Table 4. Performance metrics at FMR = 0.01, $\tau = 0.63$

To properly assess the effect of taking quality into account, along with Tables 3 and 4, an indication of the system's trade-offs are visualized through ROC in Figure 9. From this visualization, we can confirm that taking quality metrics into account negatively impacts the system trade-off and not taking quality metrics remains the more robust approach.

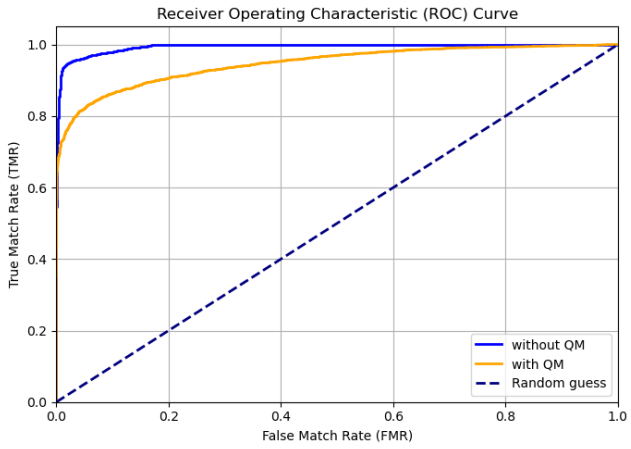


Fig. 9. ROC curve comparison for systems with and without QM

6 CONCLUSIONS

6.1 Answering SRQ1

As presented in section 4, but also from our findings, measuring the effectiveness of morph rejection is done by calculating the MMPMR of the FRS we are using. An FRS proves to be effective by being able to reject morphs even in the cases where the face presents severe occlusions which can be caused by pose, illumination, shadows and other factors that were treated in this paper.

6.2 Answering SRQ2

After analyzing all the results, it is evident that all proposed image quality properties influence similarity scores, with varying degrees of impact.

As shown in the results of different manipulation levels in Appendix E.1 and the weights used to calculate the quality metric, brightness and contrast have the most significant influences, while noise and sharpness have a more minor impact. These results align with expectations, as brightness and contrast cause some areas of the face to be illegible, due to a photo being too bright or too saturated, especially in the cases where a photo is not front-facing. Of course, adding an extremely high level of noise, for example, will also cause illegibility, but we are only looking at the manipulation levels that have been applied in section 4.5.

However, what was not expected was sharpness having the lowest impact, since blur also causes illegibility of face features. This may be attributed to the minimal sharpness level that has been applied during our manipulation phase. Nevertheless, all four image quality properties should be taken into account, as they all have an impact on performance and on each other, in some cases stronger than in others.

6.3 Answering SRQ3

It is evident that incorporating quality metrics does not enhance the effectiveness of morph rejection. To reiterate on section 5.4, although taking quality metrics into account and setting a dynamic threshold positively impact the True Match Rate, this comes at the cost of the False Match Rate also becoming higher. Additionally, only including quality metrics (i.e., no dynamic threshold) actually proves to have a negative impact on performance. The ROC curve further shows that the trade-off of the system proposed by us is actually worse than the original one, and that taking quality metrics into account makes our FRS less robust.

6.4 Answering RQ

While it is true that image quality properties have various effects on similarity scores, it does not seem like it is a good source of extra information in trying to make a FRS more robust against morphing attacks. Additionally, while it does look like there is some relation between the performance of an image, its quality and BRISQUE, this chosen IQA metric does not make it possible to find a mapping between the three, as one cannot conclude how good an image will perform based on its BRISQUE. This could be attributed to the fact that image quality properties serve as indicators of "perceptual quality", whereas FRS use various other sources of information to make a final decision on what score should be outputted.

7 FUTURE WORK

Some aspects that could be treated in the future are: trying to find another FRS, since AdaFace presented bimodal behavior from the start. Having a FRS that is simpler or uses another type of loss function could make it easier to model similarity score distributions against image quality properties. Since it proves to be a great influence for FRS performance, quality metrics should also take into account pose. Additionally, experimenting with deep features could be a starting point in trying to assess how quality influences a FRS decision on a deeper level. If further research would like to be conducted in the direction of this paper, another solution to map image quality properties to an IQA metric would be finding individual metrics for each quality property treated, or testing other IQA metrics similar to BRISQUE. Lastly, the bad performance of the proposed system could also be caused by the way the final score was calculated (the average between the quality-based prediction score and the FRS prediction score).

REFERENCES

[1] 2023. *Best Practice Technical Guidelines for Automated Border Control (ABC) Systems.* Research and Development Report 10.2819/86138. Frontex, European Border and Coast Guard Agency. https://www.frontex.europa.eu/assets/Publications/Research/Best_Practice_Technical_Guidelines_ABC.pdf Accessed: 2025-05-13.

[2] Mohamad Alansari, Oussama Abdul Hay, Sajid Javed, Abdulhadi Shoufan, Yahya Zweiri, and Naoufel Werghi. 2023. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access* 11 (2023), 35429–35446.

[3] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. 2018. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese conference on biometric recognition*. Springer, 428–438.

[4] Pranali Dandekar, Shailendra S. Aote, and Abhijeet Raipurkar. 2024. Low-resolution face recognition: Review, challenges and research directions. *Computers and Electrical Engineering* 120 (2024), 109846. https://doi.org/10.1016/j.compeleceng.2024.109846

[5] Lisa DeBruine and Benedict Jones. 2017. Face Research Lab London Set. https://doi.org/10.6084/m9.figshare.5047666.v5

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

[7] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. 2014. The magic passport. In *IEEE international joint conference on biometrics*. IEEE, 1–7.

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

[9] Mathias Ibsen, Christian Rathgeb, Daniel Fischer, Pawel Drozdowski, and Christoph Busch. 2022. Digital Face Manipulation in Biometric Systems. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, and Christoph Busch (Eds.). Springer Nature, Chapter 2, 27–45.

[10] International Civil Aviation Organization. 2003. Annex A: Photograph Guidelines. https://www.icao.int/Security/mrtd/Downloads/technical%20reports/annex_A-photograph_guidelines.pdf Accessed: 2025-05-16.

[11] Minchul Kim, Anil K Jain, and Xiaoming Liu. 2022. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18750–18759.

[12] Martin Knoche, Stefan Hörmann, and Gerhard Rigoll. 2021. Susceptibility to image resolution in face recognition and trainings strategies. *arXiv preprint arXiv:2107.03769* (2021).

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

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

[15] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* 21, 12 (2012), 4695–4708.

[16] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters* 20, 3 (2012), 209–212.

[17] Anush Krishna Moorthy and Alan Conrad Bovik. 2010. A two-step framework for constructing blind image quality indices. *IEEE Signal processing letters* 17, 5 (2010), 513–516.

[18] Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, and Christoph Busch. 2022. *Handbook of digital face manipulation and detection: from DeepFakes to morphing attacks*. Springer Nature.

[19] Eklavya Sarkar, Pavel Korshunov, Laurent Colbois, and Sébastien Marcel. 2020. Vulnerability Analysis of Face Morphing Attacks from Landmarks and Generative Adversarial Networks. *arXiv preprint* (Oct. 2020). https://arxiv.org/abs/2012.05344

[20] Eklavya Sarkar, Pavel Korshunov, Laurent Colbois, and Sébastien Marcel. 2022. Are GAN-based morphs threatening face recognition?. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2959–2963. https://doi.org/10.1109/ICASSP43922.2022.9746477

[21] Ulrich Scherhag, Andreas Nautsch, Christian Rathgeb, Marta Gomez-Barrero, Raymond NJ Veldhuis, Luuk Spreeuwers, Maikel Schils, Davide Maltoni, Patrick Grother, Sebastien Marcel, et al. 2017. Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 1–7.

[22] Ulrich Scherhag, Ramachandra Raghavendra, Kiran B Raja, Marta Gomez-Barrero, Christian Rathgeb, and Christoph Busch. 2017. On the vulnerability of face recognition systems towards morphed face attacks. In *2017 5th international workshop on biometrics and forensics (IWBF)*. IEEE, 1–6.

[23] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. 2019. Face Recognition Systems under Morphing Attacks: A Survey. *IEEE Access* 7 (2019), 23012–23026. https://doi.org/10.1109/ACCESS.2019.2899367

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

[25] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6398–6407.

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

[27] Krisztián Szabó. 2022. *Morphing robust face recognition*. B.S. thesis. University of Twente.

[28] Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, Luuk Spreeuwers, Raymond Veldhuis, and Christoph Busch. 2019. Morphed face detection based on deep color residual noise. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 1–6.

[29] Wikipedia contributors. 2024. Automated Border Control System – Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Automated_border_control_system Accessed: 2025-04-28.

[30] Wikipedia contributors. 2024. Facial Recognition System – Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Facial_recognition_system Accessed: 2025-04-28.

[31] Wikipedia contributors. 2024. Morphing – Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Morphing Accessed: 2025-04-28.

[32] Wikipedia contributors. 2025. Detection Error Tradeoff — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Detection_error_tradeoff. Accessed: 2025-05-14.

[33] Wikipedia contributors. 2025. Receiver Operating Characteristic — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Receiver_operating_characteristic. Accessed: 2025-05-14.

[34] Ivan William, Eko Hari Rachmawanto, Heru Agus Santoso, Christy Atika Sari, et al. 2019. Face recognition using facenet (survey, performance test, and comparison). In *2019 fourth international conference on informatics and computing (ICIC)*. IEEE, 1–6.

[35] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 1098–1105.

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

AI STATEMENT

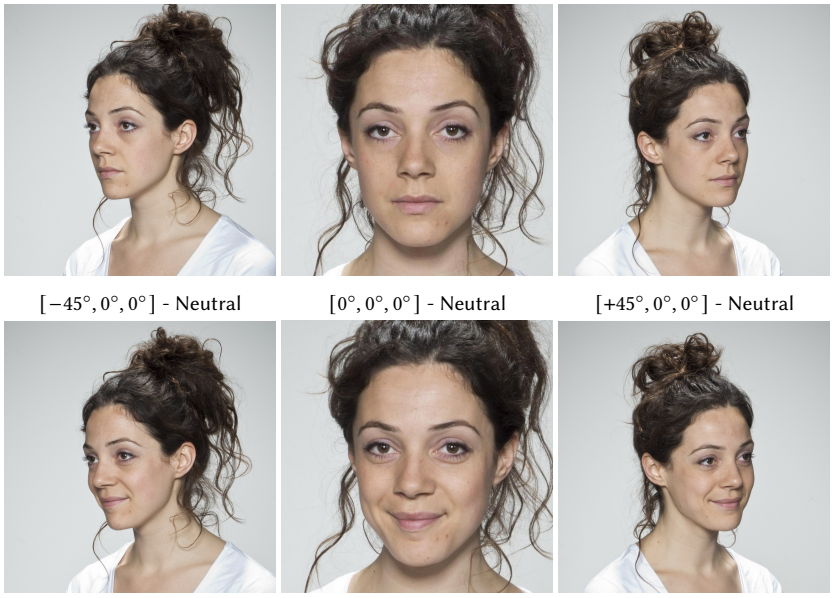
In writing this paper, Grammarly Pro was used for rephrasing purposes, for finding synonyms and for spelling and clarity checks. For the coding part of this research, ChatGPT was used for debugging, but also for shortening the coding process for simple tasks. For example, if I needed code that modified a data frame in a certain way or plotted data, I would use ChatGPT to generate the code. All code was modified to my needs and checked before being used.

A SUMMARY TABLE OF FACE RECOGNITION SYSTEM LITERATURE REVIEW

FRS	Loss	Backbone	Accuracy LFW	Accuracy IJB-C	Additional Comments
MobileFaceNets [3]	ArcFace	MobileFaceNet	99.28%	-	Lightweight, 4MB size
FaceNet [24, 34]	Triplet Loss	NN1-4	99.63%	-	Requires minimal alignment (i.e. tight crop around the face). Difference in performance between more complex architectures and smaller ones is statistically insignificant.
ArcFace [6]	ArcFace	ResNet50	99.83%	97.27%	Not stable for varying photo quality.
SphereFace [13]	A-Softmax	64-layer CNN	99.42%	-	Outperforms all other models on 64-layer CNN architecture.
CurricularFace [8]	Adaptive Curriculum Learning	ResNet100	99.8%	96.1%	-
Circle Loss [25]	Circle Loss	ResNet34 ResNet100	97.81% 98.5%	93.44% 93.59%	High flexibility in optimization.
MagFace [14]	MagFace	ResNet100	99.83%	95.97%	-
AdaFace [11]	Margin Based	ResNet100	99.83%	97.39%	Treats the case of face occlusion and varying image quality. Outperforms other models for mixed quality data. Offers smaller backbone options.
GhostFaceNets [2]	ArcFace	GhostFaceNet	99.76%	94.943%	Lightweight
DeepID3 [26]	-	DeepID	99.53%	-	-

Table A.1. Performance comparison table for all identified Face Recognition Models. Accuracy reported for IJB-C is for TAR@FAR=1e-4.

B DATASET EXAMPLES



[−45°, 0°, 0°] - Neutral [0°, 0°, 0°] - Neutral [+45°, 0°, 0°] - Neutral

[−45°, 0°, 0°] - Smiling [0°, 0°, 0°] - Smiling [+45°, 0°, 0°] - Smiling

C MANIPULATED IMAGES (NEUTRAL FRONT EXAMPLES ONLY)

Fig. C.1. Manipulated Images - Brightness Level
a positive value <1 makes the image darker, >1 makes the image brighter

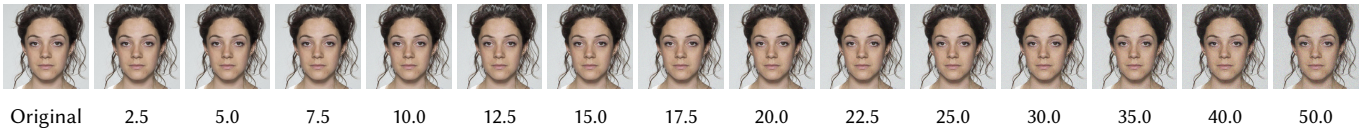


Fig. C.2. Manipulated Images - Noise Level
Gaussian noise is added with factor x, where x = standard deviation

Fig. C.3. Manipulated Images - Contrast Level
a positive value <1 makes the image less saturated, >1 makes the image more saturated

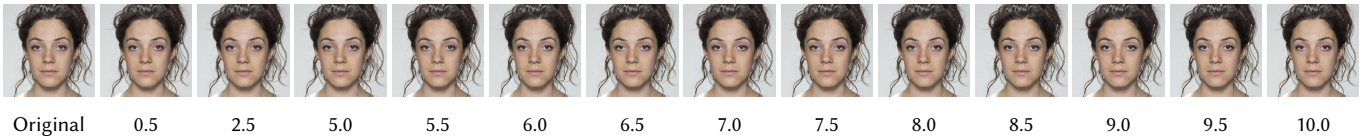
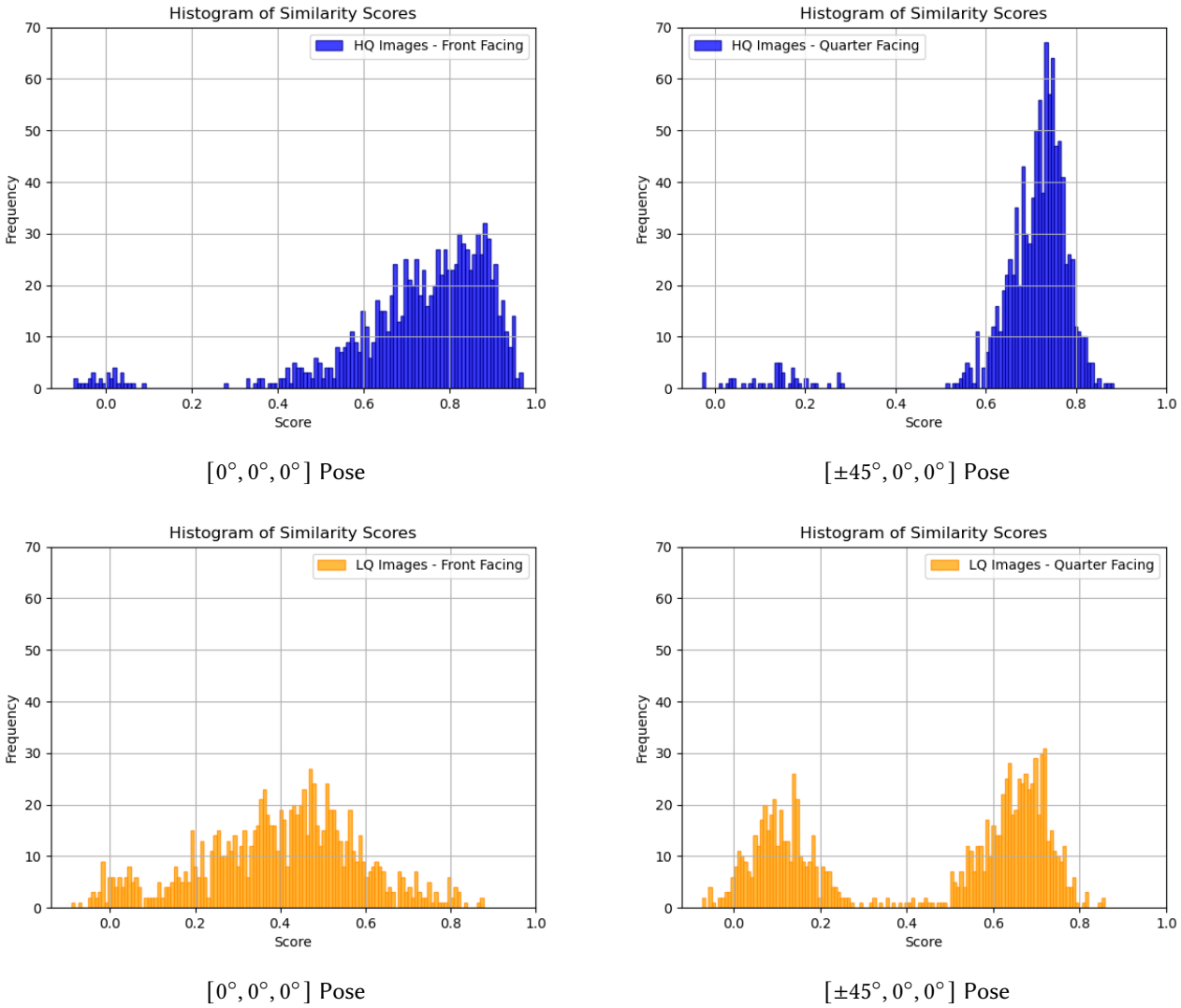


Fig. C.4. Manipulated Images - Sharpness Level
a positive value <1 makes the image blurrier, >1 makes the image sharper

D MATED SIMILARITY SCORE DISTRIBUTIONS FOR MANIPULATED IMAGES



$[0°, 0°, 0°]$ Pose $[\pm 45°, 0°, 0°]$ Pose

$[0°, 0°, 0°]$ Pose $[\pm 45°, 0°, 0°]$ Pose

Fig. D.1. Mated similarity scores for high quality (HQ) images (top) and low quality (LQ) images (bottom) separated by pose

E RESULTS

E.1 Influence of Manipulating Image Quality Properties on Similarity Scores

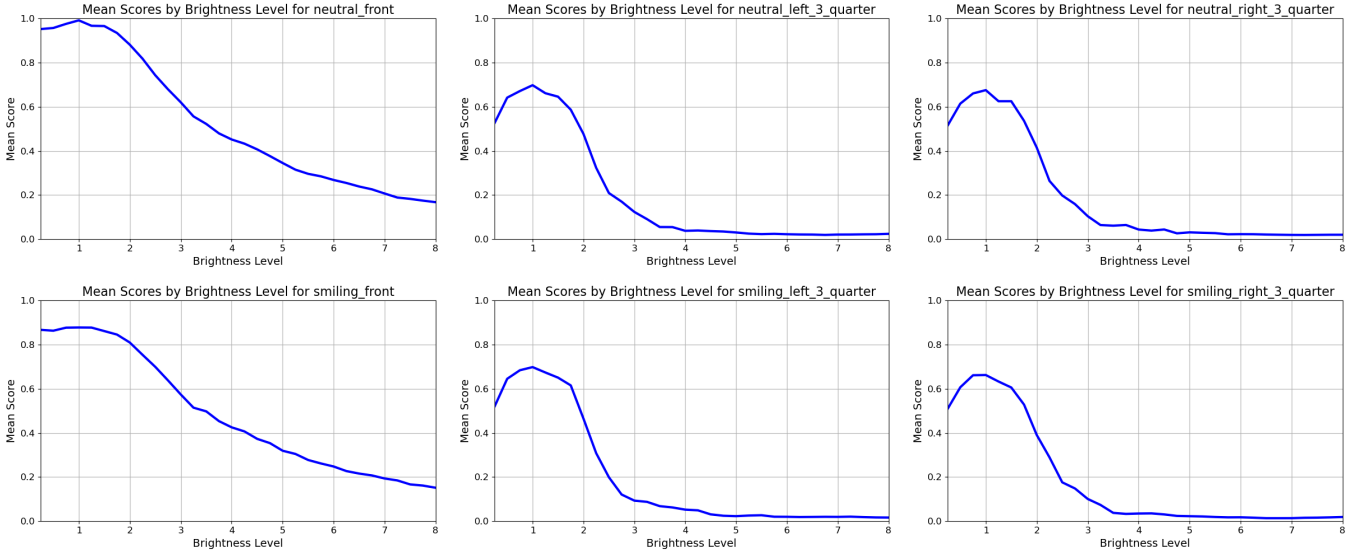


Fig. E.1.1. Effect of Brightness Level on Similarity Scores
Comparison between Passport (neutral front) and Live Capture (varying poses and expressions)

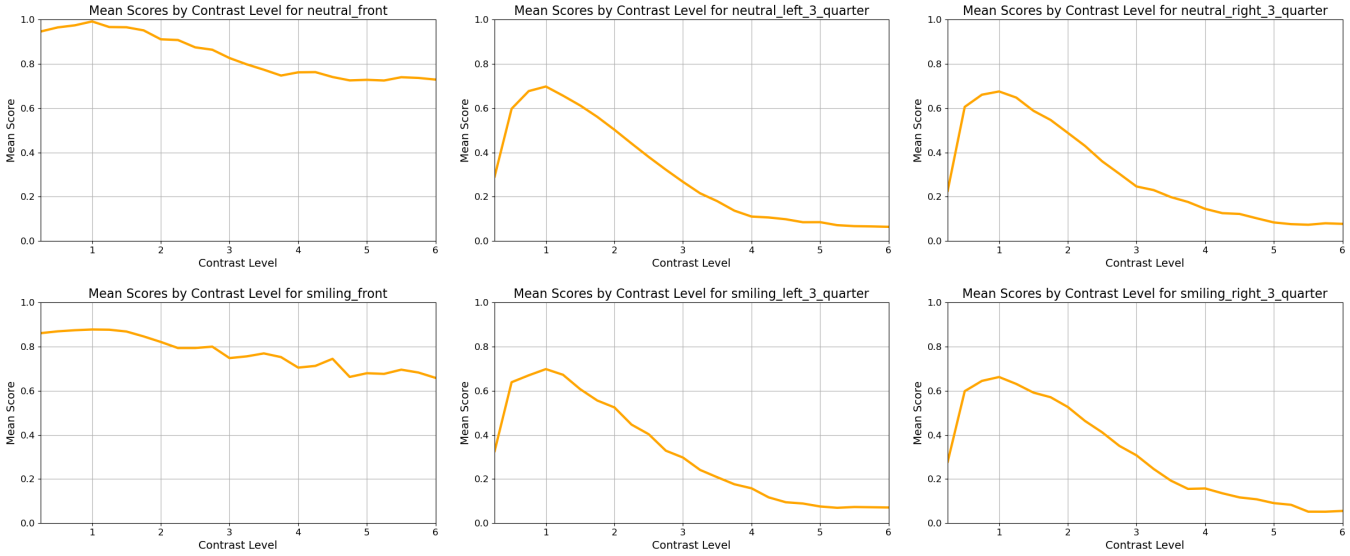


Fig. E.1.2. Effect of Contrast Level on Similarity Scores
Comparison between Passport (neutral front) and Live Capture (varying poses and expressions)

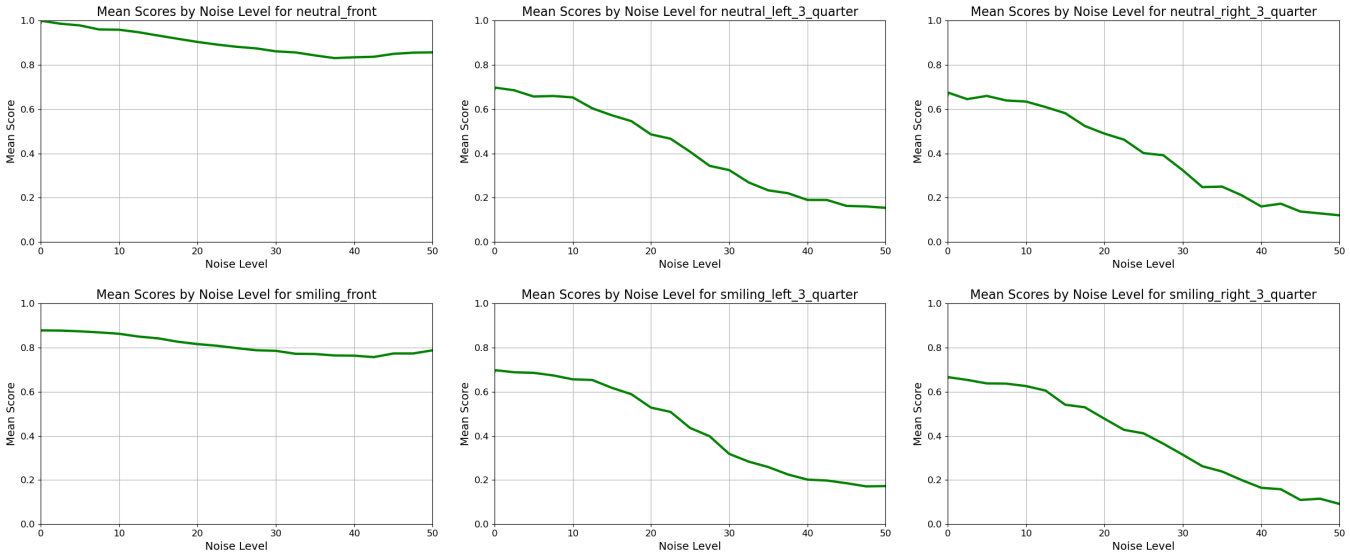


Fig. E.1.3. Effect of Noise Level on Similarity Scores
Comparison between Passport (neutral front) and Live Capture (varying poses and expressions)

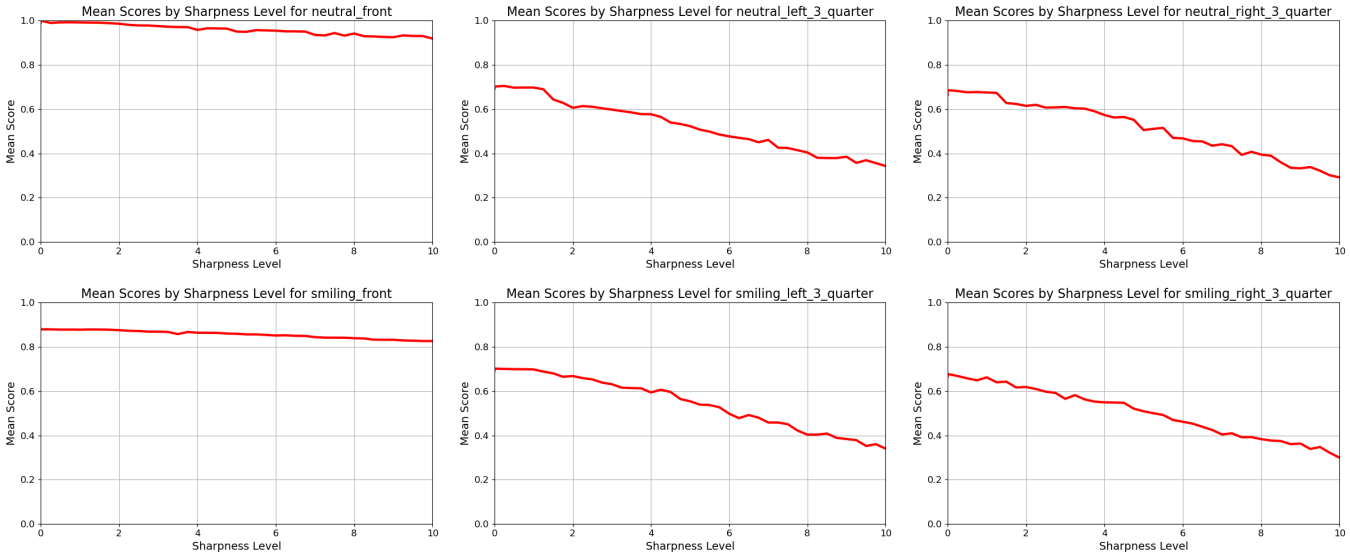


Fig. E.1.4. Effect of Sharpness Level on Similarity Scores
Comparison between Passport (neutral front) and Live Capture (varying poses and expressions)

E.2 Relationship between BRISQUE Scores and Various Quality Manipulation Levels

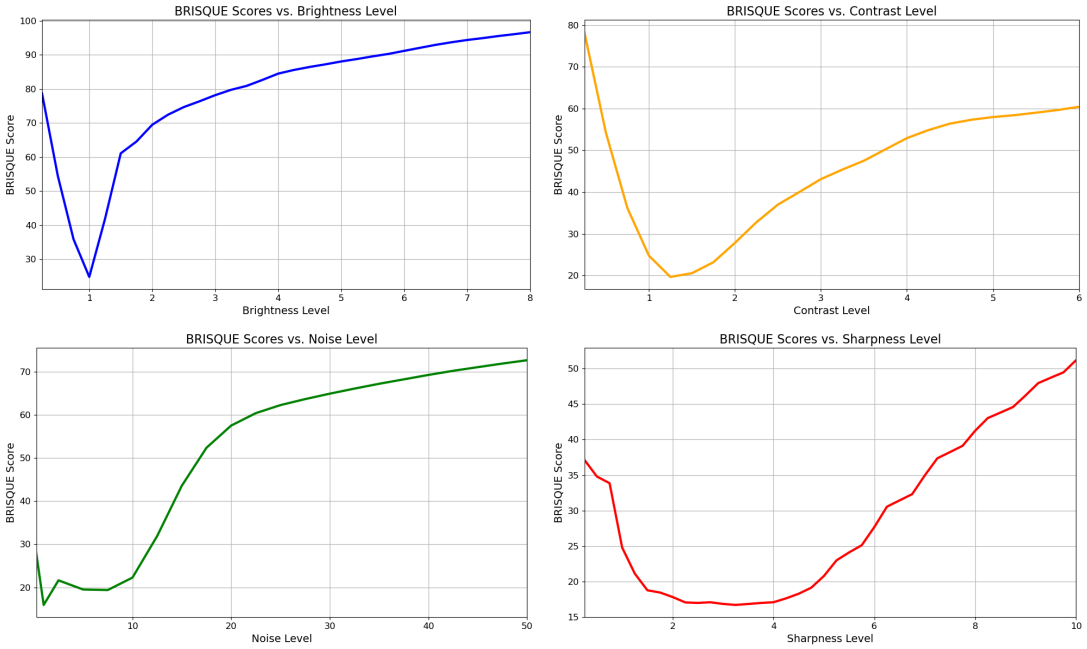


Fig. E.2.1. Relationship between BRISQUE Scores and Various Levels of Quality Manipulation