Quantifying the impact of image degradation on Deep Learning models in face recognition systems

Leandro Dias Carneiro[1], **Flavio de Barros Vidal**[2]

[1]Criminalistics Institute of the Federal District Civil Police – Brasília – DF – Brazil
[2]Dep. of Computer Science – University of Brasília – Brasília – DF – Brazil

leandro.carneiro@pcdf.df.gov.br, fbvidal@unb.br

Abstract. *Significant advancements in computer vision, particularly in facial recognition systems, have been witnessed in recent years. However, it is imperative to comprehend how these systems perform under real-world conditions, specifically when confronted with degraded images. This paper presents a comprehensive analysis of the impact of image degradation on facial recognition systems that rely on deep neural networks. The study evaluates three facial detection algorithms and eight facial recognition algorithms, with experiments conducted on four diverse datasets. A total of 14 types of image degradations, encompassing pure and mixed variations, were employed at six different intensity levels. Three distinct types of image pairs were generated to encompass various scenarios. The primary objective of this research is to enhance the understanding and assessment of facial recognition system outcomes, thereby strengthening the overall analysis of these systems. On average, the models had a minimum impact of 17% and a maximum of 43% for the datasets used in the experiment.*

1. Introduction

Controlled facial recognition has been a research subject for several decades. However, in recent years, we have witnessed significant advancements in terms of accuracy and performance, as highlighted in important studies [Bansal et al. 2021, Grm et al. 2018]. With the popularization of Convolutional Neural Networks (CNNs), facial recognition has become increasingly successful, particularly for images captured in controlled environments, following the protocols and best practices outlined in ICAO-9303 (Machine Readable Travel Documents) [Doc 2008], and the Facial Identification Scientific Working Group (FISWG) [1]. A prominent example of the success of the facial recognition dataset is the Labeled Faces in the Wild (LFW) [Huang et al. 2007], widely used as a performance benchmark for such applications. Many algorithms now achieve accuracy scores close to 100%, indicating the relative ease of recognizing faces in this dataset [de Freitas Pereira et al. 2022]. However, despite the achievements in many application domains, facial recognition encounters challenges in uncontrolled environments where image capture conditions are adverse. In applications such as autonomous cars, public surveillance, low-light areas, or low-quality capture equipment, external images often fail to meet the ideal criteria for accurate processing [de Freitas Pereira et al. 2022, Grm et al. 2018]. Numerous studies have explored non-controlled environments, and the results consistently indicate lower scores than controlled environments [Schlett et al. 2022].

[1]Available at https://www.fiswg.org/.

Following the assumptions described above, this work proposes to understand the performance impact on facial recognition systems when provided with degraded images as input, representing a more closing as possible degradations found in uncontrolled natural environments. Applying a sequence of degradations into original image databases allows for simulating real-world scenarios where uncontrolled environmental factors are prevalent. The standard processing flow of multiple systems will be analyzed, varying the degradation's input images, type, and intensity.

This work is organized as follows: Section 2 describes an overview of essential studies focusing on the influence of image degradation influence in deep learning models. Section 3 is presented our proposed methodology to evaluate degradation effects in face recognition models. Sections 4 and 5 show the most important results and conclusions, including further works.

2. Related Works

According to [Karahan et al. 2016], three facial recognition algorithms were evaluated against motion blur, noise, compression, color distortions, and occlusions. The focus of this evaluation was to identify the influence of these degradations on the performance of each algorithm. Three algorithms were used during the experiments: AlexNet [Krizhevsky et al. 2012], VGG-Face [Parkhi et al. 2015], and GoogLeNet [Tang et al. 2017]. The results indicated that motion blur, noise, and occlusion caused a significant decrease in the algorithms' performance, while color distortions (balance and contrast) had less impactful results. The author used the LFW dataset [Huang et al. 2007] as the data source for the experiment, which is commonly used as a performance benchmark for models. Still, it is considered relatively easy for the most recent algorithms.

In [Grm et al. 2018] studied the effects of different image quality covariates on facial recognition. The study examined noise, blur, pixel absence, brightness, compression, and color at different incidence levels. The study also utilized the LFW dataset [Huang et al. 2007] and four facial recognition algorithms: AlexNet [Krizhevsky et al. 2012], VGG-Face [Parkhi et al. 2015], GoogLeNet [Tang et al. 2017], and SqueezeNet [Iandola et al. 2016]. As a result, the study identified the strengths and weaknesses of the models used, revealing that high noise, blur, pixel absence, and low brightness significantly hindered the algorithms' performance. However, contrast modifications and compressions had a milder influence.

Liu et al. [Liu et al. 2019] researched the performance of various CNN models in tasks such as facial recognition and detection, object recognition, and image classification using both image and video files. To evaluate the models, they applied multiple degradations, including downsampling, noise, blur, and occlusions. Different datasets were used for each task, such as CIFAR-10 for object recognition, MSRA-CFW for facial recognition, FDDB for facial detection, SVHN for digit recognition, and ImageNet for image classification. Transfer learning was utilized to develop a deep neural network model specifically for low-resolution images.

Roy et al. [Roy et al. 2018] investigated the effect of image degradations on deep neural network architectures for image classification. They evaluated the performance of CNN models and proposed new configurations to improve accuracy. Their study included

benchmark models such as MobileNet, VGG16, VGG19, ResNet50, InceptionV3, and CapsuleNet. Image degradations like noise, blur, and JPEG compression were applied to assess the models' robustness. The experiments demonstrated that Gaussian noise decreased the network performance, with VGG architecture showing greater resilience. Salt and pepper noise affected all models, particularly MobileNet. Motion blur and Gaussian blur degraded the performance of all models, but VGG architecture exhibited more consistency in the digit dataset. Regarding JPEG compression, ResNet and MobileNet experienced significant degradation, while VGG architecture demonstrated more stable performance.

Pei et al. [Pei et al. 2018] conducted studies on the performance degradation of image classification models when using degraded images. They explored whether including degraded images in the training dataset could improve network performance. Degradations like fog noise, brightness alteration, motion blur, and fisheye lens distortion were employed. The study used the CNN models AlexNet and VGGNet-16 with datasets such as Caltech-256, PASCAL VOC, and ImageNet. The results indicated that network performance decreased when the training dataset did not include levels of degradation similar to the test images. Moreover, the study revealed that important features were not effectively captured in the hidden layers of CNNs, which could explain the low performance observed.

Aljarrah [Aljarrah 2021] investigated the impact of degradations on CNN performance for image classification tasks. The GoogLeNet network and the ImageNet dataset were used in the study. Degradations such as contrast reduction, noise addition, and occlusion were applied to the images. Motion blur had the most significant adverse effect on performance, followed by occlusion, while noise had a milder impact. Although degraded images were not used during the training phase, the author suggested that including such images in the training dataset could be considered in future work to enhance the GoogLeNet network's performance.

3. Proposed Methodology

To comprehensively assess the influence of degradations on facial recognition systems, we meticulously curated a collection of eight state-of-the-art facial recognition models alongside three leading face detection models. Subsequently, we carefully selected four prominent benchmark datasets commonly employed in this domain. To simulate real-world conditions, we systematically introduced 14 distinct types of degradations, and each applied across seven intensity levels (ranging from the original undegraded image to six gradually increasing degradation levels) to the datasets. Furthermore, we generated three different types of comparison pairs, as follows: **Pair 1 -** Standard image (non-degraded) vs. Copy of the standard image (degraded); **Pair 2 -** Standard image (non-degraded) vs. Questioned image (degraded); **Pair 3 -** Standard image (degraded) vs. Questioned image (degraded).

Consequently, 24 pipelines were formed by combining the three face detection models with the eight facial recognition models, resulting in an extensive ensemble for the facial recognition system. These pipelines were then evaluated using the degraded images, with accuracy, precision, and recall metrics recorded for each intensity level. The subsequent subsections offer a detailed account of the face recognition and detec-

tion models employed, the specific datasets utilized, the creation of image pairs, and a comprehensive overview of the types and intensity levels of degradations applied.

3.1. Face Detection and Recognition Models

In choosing the models built under the structure of methods using deep learning and that perform facial recognition, the models that, at least, were considered state-of-the-art (SOTA) models were listed. These models were incorporated into the implementation provided by the framework described in [Serengil], called DeepFace, which brings together models that perform recognition and face detection, constituting the workflow of a traditional facial recognition system. Using deep learning. The following facial recognition models are present in this framework: Facebook DeepFace [Taigman et al. 2014], DeepID [Sun et al. 2014], FaceNet [Schroff et al. 2015], VGG-Face [Parkhi et al. 2015], OpenFace [Amos et al. 2016], ArcFace [Deng et al. 2019], and SFace [Boutros et al. 2022]. Also included in this framework are SOTA models for face detection in images as Dlib [King 2009], MTCNN [Zhang et al. 2016], and RetinaFace [Deng et al. 2020].

Indeed, in most facial recognition systems, facial detection algorithms are typically the first step in the face recognition pipeline. Their primary task is to locate and identify faces within an image or video stream. These algorithms employ techniques such as deep learning-based methods or traditional computer vision approaches to analyze the input data and identify regions of interest that likely contain faces.

3.2. Datasets

To allow the evaluation of the influence of degradations on the input image, the main sets of images that were used in the validation of the chosen facial recognition models were picked, as described below:

LFW: The Labeled Faces in the Wild (LFW) dataset is a widely used benchmark dataset in face recognition. It was created by researchers at the University of Massachusetts, Massachusetts, USA. LFW was designed to evaluate and compare the performance of face recognition algorithms in unconstrained, real-world scenarios.

SCFace: The SCface (Surveillance Cameras Face Database) dataset is a widely used benchmark dataset for face recognition research. It was created by the Signal Processing Laboratory (LTS5) at École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland. The SCface dataset contains facial images of 130 subjects, with 52 images per subject, in different cameras, distance capture, and resolutions, resulting in 6.760 images. The most applicable images to this study were relative to the camera called mugshot_rotation_all, which has 1.170 images.

FEI: The FEI (Facial Expression Images) face database is a Brazilian face database that contains a set of face images taken between June 2005 and March 2006 at the Artificial Intelligence Laboratory of FEI in São Bernardo do Campo, São Paulo, Brazil. There are 14 images for each of the 200 individuals, a total of 2.800. All images are colorful and taken against a homogeneous white background in an upright frontal position with profile rotation of up to about 180 degrees. The scale might vary about 10%, and the original size of each image is 640x480 pixels. All

faces are mainly represented by students and staff at FEI, between 19 and 40 years old, with distinct appearances, hairstyles, and adornments.

GUFD: The GUFD (Glasgow Unfamiliar Face Dataset) contains 5 photos of 303 individuals, a total of 3.028 images. All images are colorful and taken in a controlled environment. The image's spatial dimensions are about 2.288 x 1.712 pixels.

Given that the SCFace, FEI, and GUFD datasets contain faces captured in a controlled environment, where various angles of a person's face are captured (ranging from 90 degrees to the left to 90 degrees to the right), and to keep only the frontal face images from the datasets, the mentioned datasets were subjected to the frontal face detector from the Dlib library.

The objective is to evaluate the effect of degradations in the facial recognition pipeline. It was decided to generate a "clean version" of all data sets, in which the worst face detector (Dlib presented the lowest performance in face detection) was used in all images to guarantee that the face recognition pipeline would not be interrupted by failures encountered during the face detection stage. Thus, Table 1 presents the real number of images used in the proposed approach.

Dataset	Total images on original dataset	Total images used	Total of classes
LFW	13.233	13.233	5.749
SCFace	1.170	833	130
FEI	2.800	2.450	200
GUFD	3.028	1.774	304

Table 1. Clean version of the chosen Data sets.

3.3. Pair Generation

Pair 1: Standard Image (non-degraded) vs. Copy of the standard image (degraded). Degradations (at various intensity levels) were applied to a copy of the standard image and then subjected to the facial recognition system. The goal of generating this pair is to determine the degradation intensity level the algorithm maintains to recognize the pair of identical images (one non-degraded and the other degraded) as the same person. So, for this reason, there are no negative pairs in this generation.

Pair 2: Standard image (non-degraded) vs. Questioned image (degraded). Degradations (at various intensity levels) were applied to all images in the dataset (except the standard image) and subjected to facial recognition systems. The purpose of generating this pair is to simulate the use of the system when one image is of relatively good quality, such as images from official documents, while the other image is obtained in an uncontrolled environment with degradations applied to it.

Pair 3: Standard image (degraded) vs. Questioned image (degraded). In this pair generation, degradations are applied to both images. The objective is to examine cases where both images were obtained in an uncontrolled environment (such as images from social media or public places) and assess the algorithm's performance in such scenarios.

3.4. Types of Degradations

The degradation of a digital image refers to the loss of information and subsequent decline in quality. Numerous factors can contribute to this deterioration, including defects in the

capture sensor, interferences during image transmission, compression techniques, and the introduction of noise, among other potential causes [Ashraf].

Therefore, to evaluate the effects of degradations on facial images, a simple protocol was elaborated, but capable of allowing the proposed analysis: Initially, it was decided to analyze the effect of a single degradation and later, incorporate a second degradation to the initial one, sequentially, and thus evaluate the effect caused. The chosen degradations were defined and listed below, following compiled from [Grm et al. 2018, Karahan et al. 2016, Roy et al. 2018, Liu et al. 2019, Liu et al. 2019, Pei et al. 2018, Aljarrah 2021] and detailed in Tables 2 and 3. Figure 1 describes an image sample collected from the dataset with single and mixed degradations as presented in Tables of the 2 and 3.

3.4.1. Single and Mixed Degradations

Degradation	Description
Gaussian Noise	Gaussian noise can be added to an image during its capture due to sensor issues or during transmission through a channel.
Salt and Pepper Noise	Another common noise is salt and pepper noise. This impulse noise is typically observed in images due to intense disturbances. It is characterized by randomly replacing original pixel values with black and white pixels.
Gaussian Blur	The generation of blurred images (Gaussian blur) was achieved using Gaussian filters with different filter window sizes (kernels).
Motion Blur	Motion blur usually occurs due to camera instability or the movement of the object/person being filmed. This type of blur is commonly observed in mobile device recordings due to the lack of stability of the person recording.
Brightness and Darkening	To evaluate the behavior of systems with various exposure levels, gradual changes in brightness and darkening were made to the images.
Downsampled	In captures taken in an uncontrolled environment, it is common for the suspect to be far from the camera's capture point, resulting in the face image being represented by only a few pixels. For this experiment, gradual downsizing of the images in the databases was performed.
JPEG Compression	To analyze the impact of compression on facial recognition models, the JPEG algorithm was used, a lossy compression algorithm. Thus, it is possible to indicate the desired compression level to be applied to the image, where a higher coefficient corresponds to higher compression and, consequently, greater degradation of the resulting image.

Table 2. List of Single degradations.

Item	Degradations	Description
1	Gaussian Blur →JPEG Compression	First, the image was degraded with Gaussian blur, and then the degraded image was subjected to JPEG compression. The purpose of this degradation sequence is to simulate images produced by low-quality cameras (with low sharpness) that undergo compression for transmission.
2	Gaussian Blur →Downsampled	Initially, the image was degraded with Gaussian blur, and then the degraded image was downsampled. The purpose of this degradation sequence is to simulate images produced by low-quality cameras (with low sharpness) where the target of interest is far from the capture point. This degradation sequence has been studied in [Liu et al. 2019]; however, downsampled was applied before Gaussian blur in the mentioned study.
3	Gaussian Blur →Brightness →JPEG Compression	The image was degraded with Gaussian blur, then the resulting image was exposed to a brightness adjustment and finally compressed using the JPEG algorithm. The purpose of this degradation sequence is to simulate images produced by low-quality cameras (with low sharpness) where the target of interest is exposed to high luminosity, such as sunlight, and then compressed for transmission.
4	Gaussian Blur →Darkening →JPEG Compression	In this sequence, the image was degraded with Gaussian blur, then the resulting image was subjected to a darkening process and finally compressed using the JPEG algorithm. The purpose of this degradation sequence is to simulate images produced by low-quality cameras (with low sharpness) where the target of interest is exposed to low luminosity, such as nighttime captures, and then compressed for transmission.
5	Gaussian Blur →Darkening →Downsampled	As the degradation #4 without applying the JPEG Compression, including downsizing. The purpose of this degradation sequence is to simulate images produced by low-quality cameras (with low sharpness) where the target of interest is exposed to low luminosity, such as nighttime captures, and is far from the capture point (camera).
6	Gaussian Blur →Darkening →Downsampled →JPEG Compression	As the degradation #5 applying the JPEG Compression. The purpose of this degradation sequence is to simulate images produced by low-quality cameras (with low sharpness) where the target of interest is exposed to low luminosity, is far from the capture point, such as distant suspects and nighttime captures, and is then compressed for transmission.

Table 3. List of Mixed Degradations.

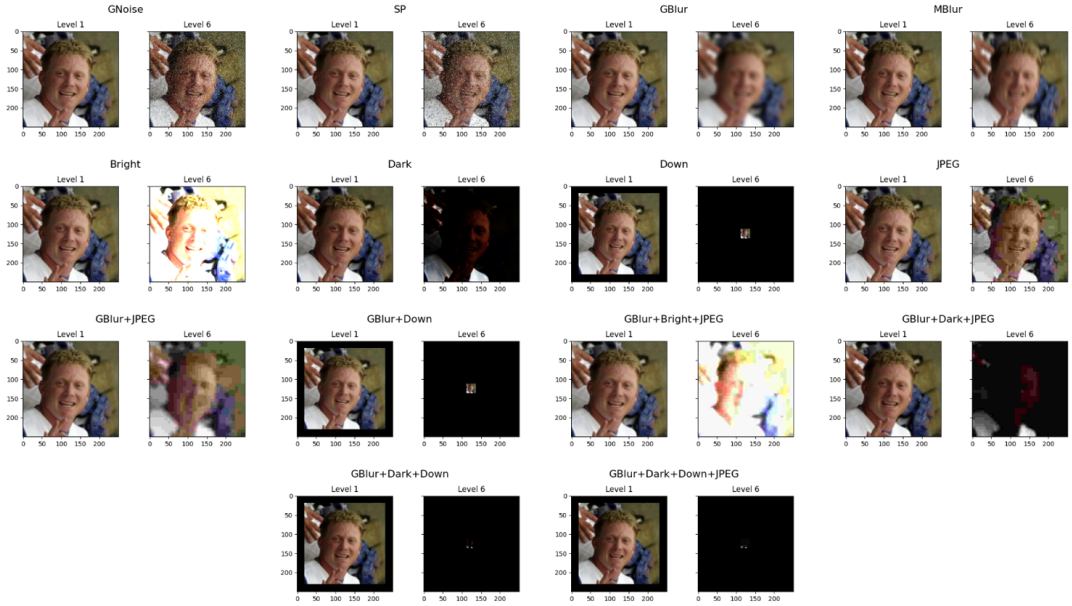

Figure 1. Single and Mixed degradations sample image.

All tested images were degraded in 6(six) levels of intensity. The following table shows the parameters used to generate each intensity level (Table 4).

Degradation	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Factor
Gaussian Blur	3	7	11	15	19	23	kernel size
Motion Blur	3	5	7	9	11	13	kernel size
Brightness	(1.0, 1.0)	(1.0, 1.5)	(1.0, 2)	(1.0, 2.5)	(1.0, 3)	(1.0, 3.5)	(alpha, beta)
Downsampled	85	70	55	40	25	10	scale percent
Darkening	0.9	0.7	0.5	0.4	0.3	0.2	gamma
JPEG Compression	85	70	55	30	15	5	quality factor
Gaussian Noise	0.001	0.01	0.03	0.07	0.1	0.15	percent of noise
Salt and Pepper	0.001	0.01	0.03	0.07	0.1	0.15	percent of noise

Table 4. Degradation Levels.

4. Results

All experiments are addressed to observe and quantify the effects of degraded images on facial recognition systems. The resultant findings are presented graphically, structured in the following manner: each page corresponds to a particular dataset, facial recognition algorithm, and pair generation method. Each page is divided into three horizontal blocks, each representing a specific pair generation method and subdivided into three columns, each corresponding to a distinct detection model. Within each column are two rows denoting positive and negative samples. Reinforcing, we show only all meaningful results. The full results list is available in the project's repository on the website GitHub Repository [2]. To better display, the name of the assessed degradations were abbreviated as follows: Gaussian Blur (GBlur), Motion Blur (MBlur), Brightness (Bright), Downsampled (Down), Darkening (Dark), JPEG Compression (JPEG), Gaussian Noise (GNoise), and, Salt and Pepper (SP).

[2]Available in https://github.com/fbvidal/Paper233907-ENIAC2023.

The analysis was organized from the observation of two perspectives: Pair generation and Model perspective, divided into datasets. Following the proposed organizing of the analysis, we are focused on gaining insights into the facial pair generation process and the performance of different models among various datasets: **Pair Generation Analysis -** We examine the process of generating pairs of faces for the recognition and detection task. We assess the effectiveness of the pair generation technique in capturing relevant facial features and variations; **Model Perspective -** We focus on evaluating the performance of different face recognition and detection models. We analyze how these models perform on various datasets, highlighting their strengths and weaknesses, subdivided by datasets.

4.1. Pair Generation Analysis

When analyzing the results, it was observed that the recall metric better represented the study. Considering that the number of negative pairs is much higher than the number of positive pairs, the accuracy metric ended up following the trend of the negative pairs curve. Furthermore, it was possible to observe different behaviors for the pairs of images generated, as follows:

4.1.1. Pair 1: Standard image (non-degraded) vs. Copy of the standard image (degraded)

Upon examining the generation of Pair 1, consisting of a standard image (non-degraded) versus a copy of the standard image (degraded), the study proved relevant as it quantified the decline in system performance for each type and intensity of degradation. As two identical images form the pair, one non-degraded and the other degraded, the pair is formed only by the same person. So, because of this, there is no negative information about this pair generation. Regarding the impact of degradations on the pipelines, it was observed that, as expected, mixed degradations had a greater impact compared to single degradations. The combinations of degradations exhibited the highest level (Figure 2) of impact on the pipelines, respectively, during the experiments: GBlur →Dark →Down →JPEG; GBlur →Dark →JPEG; GBlur →Bright →JPEG; GBlur →Dark →Down.

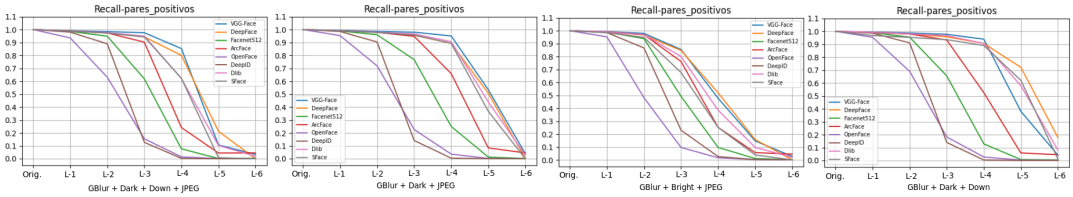

Figure 2. Most impact degradation on pair 1 (mtcnn).

On the other side, the degradations that have less impact in the experiments are described in Figures 3, were JPEG, Dark, and Bright.

4.1.2. Pair 2: Standard image (non-degraded) vs. Questioned image (degraded)

The study thoroughly analyzed Pair 2, comprising a standard image (non-degraded) compared to a questioned image (degraded). This examination proved highly pertinent as

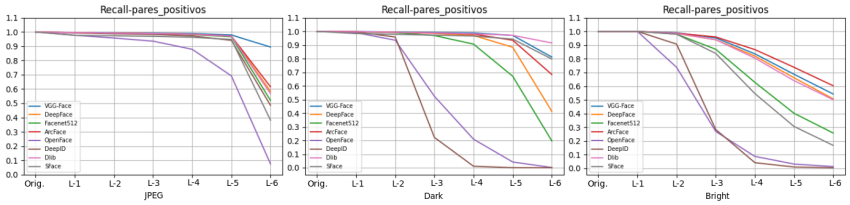

Figure 3. Less impact degradation on pair 1 (mtcnn).

it precisely measured the decrease in system performance associated with each specific type and intensity of degradation. In terms of impact degradation, the research findings indicated which ones produced more impact in the pipelines, respectively (Figure 4): GBlur →Dark →Down →JPEG; GBlur →Dark →JPEG; GBlur →Bright →JPEG; GBlur →Dark →Down.

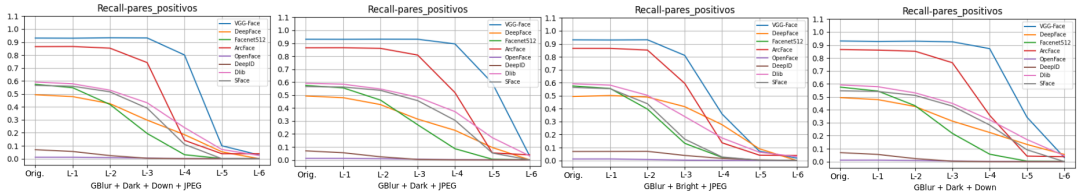

Figure 4. Most impact degradation on pair 2 (mtcnn).

On the other hand, the degradation that produced less impact was (Figure 5): JPEG, Dark, and Bright.

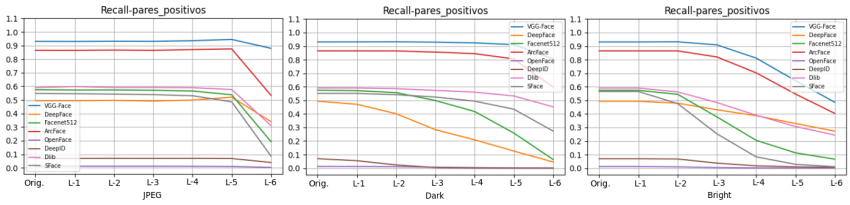

Figure 5. Less impact degradation on pair 2 (mtcnn).

4.1.3. Pair 3: Standard image (degraded) vs. Questioned image (degraded)

Concerning the generation of the third pair of images, comprising a degraded standard image versus a degraded questioned image, the evaluation not only quantified the impact of degradations on the facial recognition systems but also uncovered an exceedingly alarming behavior of the algorithms. It was observed that, at a certain point, the algorithms identified all images as indistinguishable, irrespective of the individuals depicted. This occurred when the degradation intensity reached a severe level that the algorithms lost their capacity to differentiate between different individuals and instead treated them as the ”same person.” This behavior raises serious concerns, particularly in applications operating in uncontrolled environments, as both images may have undergone degradation before being processed by the facial recognition algorithm. All observations can be observed in the Figure 6.

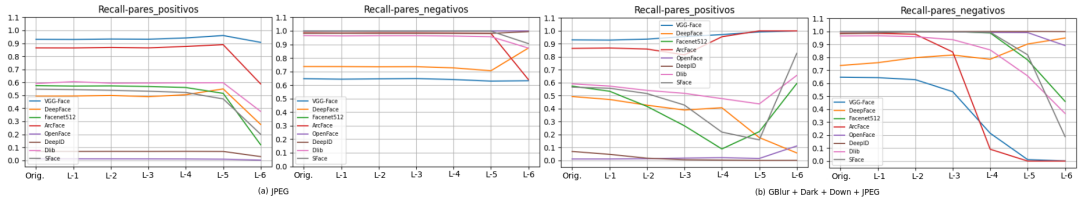

Figure 6. Less (a) and most (b) impact degradation on pair 3 (mtcnn).

A meticulous analysis of the performance graphs of the algorithms reveals a consistent and worrisome pattern: For algorithms that initially demonstrate high performance on non-degraded images, a gradual decline in the metric score is observed as the degradation intensity increases, which aligns with expectations. However, at a specific threshold, as evident in the graph representing positive pair samples, the declining score unexpectedly rises to 100%. This indicates that all positive pairs (representing the same person) are classified correctly as identical. This behavior raises suspicions as it contradicts the anticipated trend of degraded images resulting in diminished algorithm performance. Concurrently, an examination of the graph representing negative pair samples demonstrates that, at the same threshold where the positive pair curve starts to rise, the negative pair curve drops to 0, indicating that no negative pairs were accurately identified. When the positive pair curve reaches 1, and the negative pair curve reaches 0, the algorithms lose their ability to correctly identify pairs altogether, effectively categorizing all pairs as the ”same person.” A similar trend is observed for algorithms with initially low performance. However, since these algorithms already exhibit poor performance on non-degraded images, the expected initial decline in performance is less discernible. Only after surpassing a particular threshold (intensity level) does the rise in the positive pair curve and the decline in the negative pair curve become evident. These observations underscore significant limitations and raise concerns regarding the algorithms' reliability and ability to accurately identify and distinguish between pairs, particularly under highly degraded image conditions.

In a hypothetical scenario, an expert analyzing the results might erroneously conclude that two images belong to the same person. However, as demonstrated in the experiments, the system could have already lost the capability to differentiate between individuals. In law enforcement, this behavior can lead the expert to support suspicions or, in more severe cases, result in unjust arrests. By analyzing degraded images, they may consider the image of a suspect and an innocent citizen as belonging to the same person, initiating investigations, legal proceedings, and even convictions based on misinterpreted information.

4.2. Models analysis

Analyzing the results from the model's perspective, it is possible to observe different behaviors. The analysis was separated by datasets as follows.

4.2.1. LFW Dataset

The LFW dataset consists of images captured in an uncontrolled environment, making it more challenging for the models to provide accurate matches. For the initial stage of

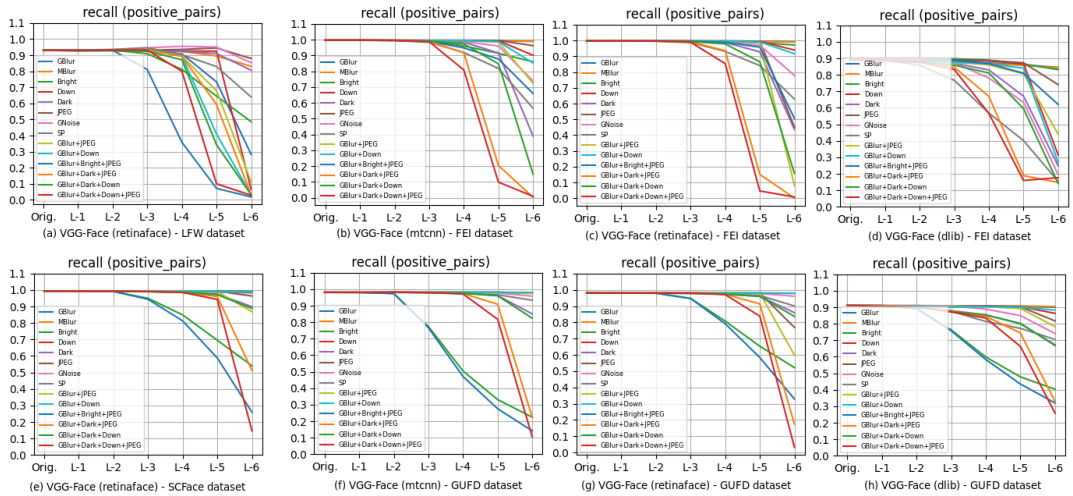

Figure 7. Impact of the degradations for specifics pipelines of pair 2.

non-degraded images, the best-performing in-order pipelines (with a close to 100% success rate) were: VGGFace+(all 3 detectors)(mtcnn: 97.89%, dlib: 97.61% and retinaface: 93.07%), ArcFace+dlib (96.05%), ArcFace+mtcnn (92.62%), SFace+dlib (91.25%), and Dlib+dlib (90.14%). For the last stage, with the highest degraded level, the best pipelines (with fewer numbers degradation curves at 0%) were: VGGFace+(all 3 detectors) (retinaface: 0 curves, mtcnn: 5 curves and dlib: 5 curves) and SFace+dlib (7 curves). At this point, it is noteworthy that VGGFace achieved the best results among all face recognition models (see Figure 7-(a)).

4.2.2. FEI Dataset

The FEI dataset consists of images captured in a controlled environment. However, the images were taken in different poses ranging from 90° left to 90° right, introducing difficulties for the pipelines. For the initial stage of degraded images, the best-performing pipelines were, respectively: VGGFace+retinaface (99.68%), VGGFace+mtcnn(99.63%), ArcFace+retinaface(95.06%), ArcFace+mtcnn (94.09%), Dlib+dlib(92.36%), Facenet512+retinaface (88.96%) and Facenet512+mtcnn (87.06%). It is remarkable that VGGFace+mtcnn and VGGFace+retinaface achieved the best results among all face recognition models (see Figure 7-(b) and (c)). In the last stage of degraded images, the best pipelines were VGGFace+(all 3 detectors)(dlib: 0 curves, mtcnn: 2 curves, and retinaface: 2 curves)). Here, VGGFace+dlib achieved the best results among all face recognition models (see Figure 7-(d)).

4.2.3. SCFace Dataset

Similar to the FEI dataset, the SCFace dataset consists of images captured in a controlled environment with different poses, ranging from 90° left to 90° right, posing challenges for the pipelines. For the initial stage of degraded images, the best-performing pipelines were: VGGFace+retinaface (99.43%) and VGGFace+mtcnn (98.70%). For this case, VGGFace+retinaface achieved the best results (see Figure 7-(e)). For the last stage of

degraded images, the best pipelines were: Arcface+(all 3 detectors)(0 curves) and VGGFace+(all 3 detectors)(0 curves). Again, VGGFace+retinaface was the best model.

4.2.4. GUFD Dataset

The GUFD dataset was captured in a controlled environment, with images taken in different poses ranging from 90° left to 90° right. For the initial stage of degraded images, the best-performing pipelines were: VGGFace+(all 3 detectors)(mtcnn: 98.10%, retinaface: 98.08%, dlib: 91.17%) and Dlib+dlib (94.06%). Here, VGGFace+mtcnn and VGGFace+retinaface achieved the best results (see Figure 7-(f) and (g)). In the last stage of degraded images, the best pipelines were: Arcface+(all 3 detectors)(0 curves), VGGFace+dlib (0 curves), and VGGFace+mtcnn (0 curves). And VGGFace+dlib achieved the best results among all face recognition models (see Figure 7-(h)).

As exposed, based on the experimental results of all tested datasets, it was consistently observed that VGGFace exhibited the best performance in non-degraded images or lower levels of degradation. Additionally, VGGFace demonstrated superior robustness when faced with the most intensive levels of image degradation.

5. Conclusions

In this work, we proposed an approach to quantify the impact of 14 types of degradations on 24 different pipelines of facial recognition systems. Furthermore, 3 types of image pairs were generated to support different application scenarios, and all were submitted to 4 face image datasets.

All results demonstrated, as expected, a gradual decrease in the recall curve for positive pairs for the generation of Pairs 1 and 2 - Standard image (non-degraded) vs. Copy of standard image (degraded) and Standard image (non-degraded) vs. Questioned image (degraded). For Pair 3 - standard image (degraded) vs. questioned image (degraded), an extremely dangerous trend was identified, possibly causing misinterpretation of the results. This behavior was observed for most of the tested facial recognition pipelines. As described in the presented work, when two degraded images are submitted to the facial recognition system beyond a certain point of degradation intensity, the algorithm tends to infer that the two images belong to the same person, even when they are different individuals. In other words, regardless of whether the image pair is positive or negative, the algorithm produces a positive result. In a criminal investigation scenario, this behavior can lead to a suspect and an innocent citizen being considered the same person, leading the expert to error and, consequently, the other entities within the criminal justice system. This specific misunderstanding can result in irreparable material and moral violence against an individual. In this remark, it is important to study mechanisms in the future, such as face quality assessment algorithms, to identify the exact moment for each algorithm and type of degradation when the curve changes from a downward trend to an upward trend to avoid misinterpretation errors by experts.

Finally, the most severe degradations to the pipeline's performance were Gaussian Blur →Darkening →Downsampled →JPEG Compression, Gaussian Blur →Brightness →JPEG Compression, and Gaussian Blur →Darkening →Downsampled. On the other hand, the degradations that had a minimal impact on the face recognition pipelines were

JPEG Compression, Darkening, and Brightness. From the model's perspective, the most resilient face recognition model that consistently exhibited the best initial performance and minimal performance degradation under these degradations was VGGFace.

A further potential analysis of image degradation and face recognition deep models could involve investigating the impact of different degradations and conducting an in-depth examination of various degradation factors that affect the models' backbone. It can provide insights into the vulnerabilities and limitations of these models and provides improvements. With this information, researchers can assess the algorithms' ability to handle real scenarios where images are often subject to multiple degradation simultaneously. Additionally, exploring the transferability of the trained models across different datasets with varying degradation characteristics can shed light on the algorithms' capabilities and potential biases. These further analyses would contribute to developing more reliable and robust face recognition systems in real-world applications.

References

Aljarrah, I. A. (2021). Effect of image degradation on performance of convolutional neural networks. *International Journal of Communication Networks and Information Security*, 13(2):215–219.

Amos, B., Ludwiczuk, B., and Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

Ashraf, M. Image degradation and noise.

Bansal, A., Ranjan, R., Castillo, C. D., and Chellappa, R. (2021). Deep cnn face recognition: Looking at the past and the future. *Deep Learning-Based Face Analytics*, pages 1–20.

Boutros, F., Huber, M., Siebke, P., Rieber, T., and Damer, N. (2022). Sface: Privacy-friendly and accurate face recognition using synthetic data. *arXiv preprint arXiv:2206.10520*.

de Freitas Pereira, T., Schmidli, D., Linghu, Y., Zhang, X., Marcel, S., and Günther, M. (2022). Eight years of face recognition research: Reproducibility, achievements and open issues. *arXiv*, (2208.04040).

Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Doc, I. (2008). 9303, parts i, ii, iii, machine readable travel documents specifications.

Grm, K., Štruc, V., Artiges, A., Caron, M., and Ekenel, H. K. (2018). Strengths and weaknesses of deep learning models for face recognition against image degradations. *Iet Biometrics*, 7(1):81–89.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Karahan, S., Yildirum, M. K., Kirtac, K., Rende, F. S., Butun, G., and Ekenel, H. K. (2016). How image degradations affect deep cnn-based face recognition? In *2016 international conference of the biometrics special interest group (BIOSIG)*, pages 1–5. IEEE.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Liu, D., Cheng, B., Wang, Z., Zhang, H., and Huang, T. S. (2019). Enhance visual recognition under adverse conditions via deep networks. *IEEE Transactions on Image Processing*, 28(9):4401–4412.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press.

Pei, Y., Huang, Y., Zou, Q., Zang, H., Zhang, X., and Wang, S. (2018). Effects of image degradations to cnn-based image classification. *arXiv preprint arXiv:1810.05552*.

Roy, P., Ghosh, S., Bhattacharya, S., and Pal, U. (2018). Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*.

Schlett, T., Rathgeb, C., Henniger, O., Galbally, J., Fierrez, J., and Busch, C. (2022). Face image quality assessment: A literature survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–49.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Serengil, S. I. Serengil - deepface.

Sun, Y., Chen, Y., Wang, X., and Tang, X. (2014). Deep learning face representation by joint identification-verification. *Advances in neural information processing systems*, 27.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708.

Tang, P., Wang, H., and Kwong, S. (2017). G-ms2f: Googlenet based multi-stage feature fusion of deep cnn for scene recognition. *Neurocomputing*, 225:188–197.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.