UNIVERSIDADE DE SAO PAULO

ESCOLA POLITÉCNICA

PATRICIA ALEJANDRA PACHECO REINA

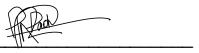
Convolutional Neural Network for Distortion Classification in Face Images

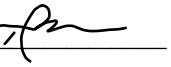
São Paulo

2021

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 13 de Setembro de 2021

Assinatura do autor: 

Assinatura do orientador: 

Catalogação-na-publicação

Pacheco Reina, Patricia Alejandra
Convolutional neural network for distortion classification in face images / P. A. Pacheco Reina -- versão corr. -- São Paulo, 2021.

96 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.CNN 2.Classificação de distorção 3.Qualidade da imagem 4.FIQA I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

PATRICIA ALEJANDRA PACHECO REINA

Convolutional Neural Network for Distortion Classification in Face Images

Revised Version

M.Sc. Thesis presented to the Electric Engineering Post Graduation Program, at the Escola Politécnica da Universidade de São Paulo (EPUSP), Brazil for the Masters in Science diploma.

Concentration Area: Computer Engineering

Advisor: Ph.D Graça Bressan

São Paulo

2021

Abstract

Face processing algorithms are becoming more popular in recent days due to the great domain of application in which they can be used. As a consequence, research about the quality of face images is also increasing. The current approach to Face Image Quality Assessment (FIQA) is focused on improving the performance of face recognition systems, as a result, current FIQA algorithms don't provide an indication of quality, but a performance estimation for face recognition algorithms. This approach makes the FIQA algorithms potentially unsuited for other scenarios regarding face images, and susceptible to inherit the limitations of face recognition. The present work tackles the main limitations of the current FIQA algorithms by proposing a new approach based on the distortions affecting the images. We developed two models based on Convolutional Neural Networks (CNN), to classify facial images according to the type and the degree of the distortion present in them. The models' output provides qualitative information about the quality of facial images, useful for face recognition systems, as well as other face processing algorithms. Additionally, the proposed method can be a starting point to image enhancement processes like denoising, and deblurring. Two other contributions can be outlined from this work: a comprehensive study about the impact of blur, noise, brightness, contrast, and JPEG compression in face processing algorithms; and a new dataset for image quality assessment and distortion classification in face images.

Keywords: CNN. Distortion classification. Image quality. FIQA.

Resumo

Os algoritmos de processamento facial estão se tornando mais populares nos últimos dias devido ao grande domínio de aplicação em que podem ser usados. Como consequência, as pesquisas sobre a qualidade das imagens faciais também estão aumentando. A abordagem atual para Avaliação da Qualidade da Imagem Facial (FIQA) é focada em melhorar o desempenho dos sistemas de reconhecimento facial, como resultado, os algoritmos FIQA atuais não fornecem uma indicação de qualidade e sim uma estimativa de desempenho para algoritmos de reconhecimento facial. Essa abordagem torna os algoritmos FIQA potencialmente inadequados para outros cenários relacionados a imagens faciais e suscetíveis a herdar as limitações do reconhecimento facial. O presente trabalho aborda as principais limitações dos algoritmos FIQA atuais ao propor uma nova abordagem baseada nas distorções que afetam as imagens. Desenvolvemos dois modelos baseados em Redes Neurais Convolucionais (CNN), para classificar as imagens faciais de acordo com o tipo e o grau de distorção nelas presente. A saída dos modelos fornece informação qualitativa sobre a qualidade das imagens faciais, útil para sistemas de reconhecimento facial, bem como outros algoritmos de processamento facial. Além disso, o método proposto pode ser um ponto de partida para processos de aprimoramento de imagem, como remoção de ruído e desfoque. Duas outras contribuições podem ser delineadas a partir deste trabalho: um estudo detalhado sobre o impacto de desfoque, ruído, brilho, contraste e compressão JPEG em algoritmos de processamento facial; e um novo conjunto de dados para avaliação de qualidade de imagem e classificação de distorção em imagens faciais.

Palavras-chave: CNN. Classificação de distorção. Qualidade da imagem. FIQA.

Figure Index

Figure 3.4-1. Examples of the images resulting from Gaussian noise degradation. From left to right: LFW, APPA-REAL, IBUG. 35
Figure 3.4-2 Examples of the images resulting from Gaussian blur degradation. From left to right: LFW, APPA-REAL, IBUG. 35
Figure 3.4-3. Examples of the images resulting from motion blur degradation. From left to right: LFW, APPA-REAL, IBUG. 36
Figure 3.4-4. Examples of the images resulting from contrast degradation. From left to right: LFW, APPA-REAL, IBUG. 37
Figure 3.4-5. Examples of the images resulting from brightness degradation. From left to right: LFW, APPA-REAL, IBUG. 37
Figure 3.4-6 Examples of the images resulting from JPEG quality degradation. From left to right: LFW, APPA-REAL, IBUG. 38
Figure 3.5-1. FaceNet behavior across noise levels. 39
Figure 3.5-2. DEX behavior across noise levels. 40
Figure 3.5-3. DAN behavior across noise levels. 40
Figure 3.5-4. FaceNet behavior across gaussian blur levels. 41
Figure 3.5-5 DEX behavior across gaussian blur levels. 41
Figure 3.5-6. DAN. behavior across gaussian blur levels. 41
Figure 3.5-7. FaceNet behavior across motion blur levels. 42
Figure 3.5-8. DEX behavior across motion blur levels. 42
Figure 3.5-9. DAN behavior across motion blur levels. 43
Figure 3.5-10. FaceNet behavior across contrast degradations. 43
Figure 3.5-11. DEX behavior across contrast degradations. 44
Figure 3.5-12. DAN behavior across contrast degradations. 44
Figure 3.5-13. FaceNet behavior across low brightness. 45
Figure 3.5-14. DEX behavior across low brightness 45
Figure 3.5-15. DAN behavior across low brightness 45
Figure 3.5-16. FaceNet behavior across high brightness. 46
Figure 3.5-17. DEX behavior across high brightness. 46
Figure 3.5-18. DAN behavior across high brightness. 46
Figure 3.5-19. FaceNet behavior across JPEG quality levels. 47
Figure 3.5-20. DEX behavior across JPEG quality levels. 47
Figure 3.5-21. DAN behavior across JPEG quality levels. 47
Figure 4.1-1. Example of a typical CNN architecture. (GOODFELLOW; BENGIO; COURVILLE, 2016) 50

Figure 4.2-1. Initial CNN architecture. 54
Figure 5.1-1. Final CNN model. 57
Figure 5.1-2. Loss and Accuracy during training. 58
*Figure 5.1-3. Final model for 128*128 images.* 58
Figure 5.1-4. Loss and Accuracy during training. 59
*Figure 5.2-1. Normalized Confission Matrix for the 256*256 model in the test set.* 61
*Figure 5.2-2. Normalized Confission Matrix for the 128*128 model in the test set.* 64
Figure 7.2-1. Experiment Results. MAE values obtained in every scenario for real and apparent age estimation. 73
Figure 10.2-1. Samples of the datatset created for distortion classification in face images. 91
Figure 10.2-2. Samples of the images obtained from the IDEAL_LIVE_DFD. 93

Table Index

Table 3.2-1. Face Processing Algorithms 33
Table 3.3-1. Datasets selected for the study 34
Table 4.2-1. Classes definition according to type and degree of distortion. 53
Table 4.2-2. Initial hyperparameter configuration for training 54
*Table 5.2-1.Classification Report for the 256*256 model* 60
*Table 5.2-2. Error analysis for the 256*256 model* 62
*Table 5.2-3. Classification Report generated with the 128*128 model in the test set.* 63
*Table 5.2-4. Error Analysis for the 128*128 model.* 63
Table 6.1-1. Parameter configuration of the functions used in (GUNASEKAR; GHOSH; BOVIK, 2014). 66
Table 6.1-2. Mapping from parameter configuration to distortion magnitude 67
*Table 6.2-1. Classification Report for the 256*256 model in the validation set.* 68
*Table 6.2-2. Error analysis for the 256*256 model in the validation set.* 68
*Table 6.2-3. Classification Report for the 128*128 model in the validation set.* 69
*Table 6.2-4. Error analysis of the 128*128 model in the validation set.* 69
Table 7.1-1. Scenarios description 72
Table 10.1-1. Metrics results from the noise experiment 85
Table 10.1-2. Metrics degradation from the noise experiment in percentage 85
Table 10.1-3. Metrics degradation from the Gaussian blur experiment 86
Table 10.1-4. Metrics degradation from the Gaussian blur experiment in percentage 86
Table 10.1-5. Metrics degradation from the motion blur experiment 86
Table 10.1-6. Metrics degradation from the motion blur experiment in percentage 87
Table 10.1-7. Metrics degradation from the contrast experiment 87
Table 10.1-8. Metrics degradation from the contrast experiment in percentage 87
Table 10.1-9. Metrics degradation from the low brightness experiment 88
Table 10.1-10. Metrics degradation from the low brightness experiment in percentage 88
Table 10.1-11. Metrics degradation from the high brightness experiment 88
Table 10.1-12. Metrics degradation from the high brightness experiment in percentage 89
Table 10.1-13. Metrics degradation from the JPEG experiment 89
Table 10.1-14. Metrics degradation from the JPEG experiment in percentage 89

Abreviations List

60PFCD	60 Person Face Comparison Database
AFLW	Annotated Facial Landmarks in the Wild
AI	Artificial Intelligence
ANN	Artificial Neural Network
APPA-REAL	Real and Apparent Age
AUC	Area Under de Curve
AVEC	Audio/Visual Emotion and Depression Recognition Challenge
BIQI	Blind Image Quality Index
BLIINDS-II	BLind Image Integrity Notator using DCT Statistics
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
CACD	Cross-Age Celebrity Dataset
CASIA	Chinese Academy of Sciences
CCTV	Closed-Circuit Television
CIFAR	Canadian Institute for Advanced Research
CNN	Convolutional Neural Network
CORNIA	Codebook Representation for No-Reference Image Assessment
CPU	Central Process Unit
CSIQ	Categorical Image Quality
CSR	Cascade Shape Regression
DAN	Deep Alignment Network
DCNN	Deep Convolutional Neural Network
DEX	Deep Expectation
DFD	Distorted Face Database
DIIVINE	Distortion Identification-based Image Verity and Integrity Evaluation
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
FABO	Bimodal Face and Body
FAR	False Alarm Rate
FERET	Facial Recognition Technology
FFHQ	Flickr Face High Quality
FG-NET	Face and Gesture Recognition Network

FIQA	Face Image Quality Assessment
FQA	Face Quality Assessment
FR	Full Reference
FRGC	Face Recognition Grand Challenge
FSIM	Feature Similarity Index
GAN	Generative Adversial Network
GIST	Efficient Data Encoding for Deep Neural Network Training
GLCM	Gray Level Co-occurrence Matrix
GPC	Gaussian Process Classifier
GPU	Graphic Processing Unit
GRU	Gated Recurrent Units
HELEN	Helen Facial Feature Dataset
HOG	Histogram of Oriented Gradients
HQ	High Quality
HQV	Human Quality Values
HVS	Human Visual System
IBUG	Intelligent Behaviour Understanding Group
ICA	Independent Component Analysis
IDEAL	Intelligent Data Exploration and Analysis Laboratory
IEEE	Institute of Electrical and Electronics Engineers
IFC	Information Fidelity Criterion
IJB-A	IARPA Janus Benchmark A
IMDB	Internet Movie Database
IMDB-WIKI	IMDB Wiki Faces Dataset
IQA	Image Quality Assessment
JPEG	Joint Photographic Experts Group
KNN	K-Nearest Neighbors
LAP	Looking At People
LBP	Local Binary Patterns
LBPH	Local Binary Pattern Histogram
LFPW	Labeled Face Parts in the Wild
LFW	Labeled Faces in the Wild
LIVE	Laboratory for Image & Video Engineering
LSTM	Long Short Term Memory
LTP	Local Ternary Pattern
MAE	Mean Absolute Error

ML	Machine Learning
MLBP	Modified Local Binary Patterns
MLP	Multilayer Perceptron
MORPH	A longitudinal image database of normal adult age-progression
MOS	Mean Opinion Score
MQV	Matcher Quality Values
MSM	Mutual Subspace Method
MS-SSIM	Multiscale Structural Similarity Index
Multi-PIE	Multi Pose, Illumination, and Expression
NLPR	National Laboratory of Pattern Recognition
NR	No Reference
NRQ LBP	No Reference Image Quality Assessment based on LBP
ORL	AT&T Face Dataset
PCA	Principal Component Analysis
POC	Proof of Concept
PSNR	Peak Signal to Noise Ratio
PSNRM	Modified PSNR
RBM	Restricted Boltzmann Machines
ReLU	Rectified Linear Unit
RMS	Root Mean Square
RMSE	Root Mean Square Error
RMSprop	Root Mean Square propagation
RNN	Recurrent Neural Network
RR	Reduced Reference
SGD	Stochastic Gradient Descent
SNR	Signal to Noise Ratio
SSEQ	Spatial-Spectral Entropy-based Quality
SSIM	Structural Similarity Index
SVHN	Street View House Numbers Dataset
SVM	Support Vector Machine
TID2008	Tampere Image Database
VCL@FER	Video Communication Laboratory at The Faculty of Electrical Engineering and Computing of the University of Zagreb
VGG-16	Very Deep Convolutional Networks for Large-Scale Image Recognition
VGGFace2	A large scale image dataset for face recognition

VIF	Visual Information Fidelity
WN	White Noise

Summary

1 Introduction 13
1.1 Motivation 14
1.2 Problem 17
1.3 Objectives 17
1.4 Methodology 19
1.5 Thesis Outline 19
2 State of the Art 21
2.1 Introduction 21
2.2 FIQA 21
2.2.1 Limitations of FIQA methods 26
2.3 Distortion classification methods 26
2.4 Summary 28
2.5 Conclusions 29
3 Impact of image distortion in face processing algorithms 30
3.1 Introduction 30
3.2 Face Processing Algorithms 30
3.2.1 FaceNet 30
3.2.2 Deep Expectation (DEX) 31
3.2.3 Deep Alignment Network (DAN) 32
3.3 Datasets 33
3.4 Distortions 34
3.5 Results 38
3.5.1 Noise 39
3.5.2 Blur 40
3.5.3 Contrast 43
3.5.4 Brightness 44
3.5.5 JPEG Compression 46
3.6 Summary 48
3.7 Conclusions 48
4 Model description 49
4.1 Deep Learning 49
4.1.1 CNN 50
4.1.2 Hyperparameters 51
4.2 Model Overview 53
4.2.1 Class definition 53

4.2.2 Model Architecture 53
4.3 Dataset creation 55
5 Results 57
5.1 Training Process 57
5.1.1 Second Model 58
5.2 Tests Results 59
5.2.1 256*256 model 59
5.2.2 128*128 model 62
5.3 Conclusions 64
6 Validation 66
6.1 Validation dataset 66
6.2 Validation Results 67
6.2.1 256*256 model 67
6.2.2 128*128 model 68
6.3 Conclusions 69
7 Distortion Classification as Previous Stage for Face Processing: Use Case. 71
7.1 Experiment Setup 71
7.2 Results 72
7.3 Conclusions 73
8 General Conclusions 74
8.1 Summary 74
8.2 Limitations 74
8.3 Conclusions 75
8.4 Future work 76
9 Bibliography 77
10 Appendix 85
10.1 Tables 85
10.2 Datasets samples 90

1 Introduction

Face processing algorithms are becoming more popular in recent days due to the great domain of applications in which they can be applied. Face identification, face verification, and face antispoofing are widely used for security and access control. Other methods like gender classification, age estimation, and emotion detection are also gaining attention thanks to their application in advertising and recommendation systems. As a consequence, research about the quality of face images is also increasing. Several studies have tackled this issue from different perspectives: either studying the quality requirements to achieve acceptable performance or focussing on improving said performance in the presence of low-quality images.

Face Image Quality Assessment (FIQA/FQA) is a subset of Image Quality Assessment (IQA), specifically dedicated to evaluating face image quality. Like IQA methods, FIQA algorithms can be classified as Full Reference (FR), Reduce Reference (RR), and No-Reference (NR), depending on the need and availability of original undistorted images (YOGITA; PATIL, 2015). FR methods require both the image to assess and an original image to compare to, to provide a quality measure. RR does not need the original image, but information concerning its characteristics. Finally, NR methods are capable of providing an image quality measure without any extra information apart from the image itself. Because of its nature, NR methods constitute the main focus of the recent publications in this domain, with Deep Learning (DL) architectures as the most promising approach (OKARMA, 2019).

One important characteristic of IQA methods is that regardless of the type, the image quality is described by a score. Said score is set to replicate the human notion of image quality. To achieve that, IQA datasets come with Mean Opinion Score (MOS) values or Differential MOS values, obtained as a result of processing the opinion of human observers. The goal of the IQA methods is to increase their correlation with the human scores (OKARMA, 2019).

FIQA methods differ from traditional IQA algorithms in that they are limited in terms of datasets. According to our research, there is only one dataset for face image quality currently available, the IDEAL-LIVE dataset (GUNASEKAR; GHOSH; BOVIK, 2014), where 215 reference images were used to create distorted ones, however, the authors aren't clear about the total amount of images generated, and only the original images are available for download. An important annotation is that most publications on FIQA are targeted to improve face recognition performance, and the datasets used to train and validate such methods are the same ones used to benchmark face recognition algorithms. As a consequence, they don't count with MOS or Differential MOS values, instead, a score representing the performance of a specific face recognition algorithm is generated as ground-truth. In other words, current FIQA algorithms don't provide an indication of quality, but a performance estimation for face recognition algorithms.

This is an important limitation, and although current FIQA methods can be useful for face recognition, their application in other scenarios is limited. Additionally, using face recognition datasets to train and test FIQA algorithms bounds their generalization to the conditions of these datasets. In that regard, a recent paper by (TERHÖRST et al., 2020) demonstrated that current FIQA algorithms have a demographic bias similar to the one observed

in face recognition systems. As stated in (TERHÖRST et al., 2020), one of the limitations of face recognition algorithms is their bias against specific demographics, which is mostly attributed to the ethnic distribution within the datasets used for training and testing. As a consequence, FIQA algorithms are inheriting said bias, making them susceptible to unfair results.

Besides that, the traditional approach to image quality is also limited. A score-based system can replicate human quality assessment, however, it doesn't provide much information about the image properties in terms of quality. At best, one can infer the general condition of the image, but with the score alone is not possible to know whether the image has noise, blur, compression artifacts, or any other distortion. That information could be beneficial not only to image enhancement algorithms but to other image processing systems that deal with images of different quality.

Furthermore, in a recent paper by (BLAU; MICHAELI, 2018), the authors mathematically demonstrated that distortion metrics and perceptual quality measures are anti-correlated, posing a problem to evaluate the results of enhancement algorithms like denoising, deblurring, and super-resolution. In their experiments, they proved that both traditional and modern distortion metrics such as Peak Signal to Noise Ratio (PSNR) (OKARMA, 2019), Root Mean Square Error (RMSE), Multiscale Structural Similarity Index (MS-SSIM) (OKARMA, 2019), Information Fidelity Criterion (IFC) (OKARMA, 2019), Visual Information Fidelity (VIF) (SHEIKH; BOVIK, 2006) are not reliable to correctly assess image quality in terms of distortion.

One alternative to solve the aforementioned limitations is to assess the image quality in terms of the distortions present in the image. A distortion classification method able to classify the type and degree of the distortion affecting the images, provides qualitative information useful for face recognition algorithms, as well as for other face processing systems. Additionally, can be a starting point to image enhancement processes like denoising, deblurring, and super-resolution. At the moment, such an approach hasn't been proposed to assess the quality of face images, so there is room for experimentation and innovation on that front.

1.1 Motivation

As stated in the introductory section, there is a great demand for research and deployment of face processing algorithms like face recognition, age estimation, emotion recognition, among others. Although their tasks are different, all these algorithms take face images as their input signal, and like any signal processing system, the input quality plays an important role in the final performance of the algorithms. A recent study carried on by (DODGE; KARAM, 2016) evaluates the impact of image quality in DL algorithms by analysing the performance of well-known image classification methods under different levels of blur, noise, contrast, JPEG, and JPEG 2000 compressions. The methods selected for the study were: the Caffe Reference Model (JIA et al., 2014), the VGG-CNN-S (CHATFIELD et al., 2014), the VGG-16 (SIMONYAN; ZISSERMAN, 2015), and the GoogleNet (SZEGEDY et al., 2015a). The images used correspond to a subset of the ImageNet 2012 database (KRIZHEVSKY; SUTSKEVER; HINTON, 2017), from which the authors generated the distorted ones.

The study concluded that blur and noise have a relevant influence on deep learning performance, while contrast and compressions only affect performance when the images reach very low quality. Another important conclusion is that even if the models only detect small changes in the first layers of the networks, they propagate throughout the architecture causing a bigger impact on the output.

Even though this study doesn't focus on face processing algorithms, the fact that the state-of-the-art algorithms dedicated to processing face images are based on DL architectures makes the aforementioned results very important to understand their behaviour under these distortions.

In a publication by (DUTTA; VELDHUIS; SPREEUWERS, 2012), the authors review the impact of image quality on face recognition performance. The authors focused on face recognition for forensic evaluation, where the images available usually come from CCTV cameras, with low quality. To assess the impact of image quality in this scenario, the authors used a commercial face recognition system[1] and simulated the open set recognition scenario in a way that not all individuals in the test set are present in the reference set. The database used for this purpose was the Multi Pose, Illumination, and Expression (MultiPIE) dataset (GROSS et al., 2010), and the distortions taken into consideration were pose variation, motion blur, illumination, resolution, and gaussian noise. In this study, each image was distorted varying just one parameter at a time. The conclusions of the paper placed the difference in pose between the test image and the reference image as the main factor affecting face recognition performance, with a 50% decrease in performance. Once the pose is the same, Gaussian noise and resolution account for a 35% decrease, whereas the influence of motion blur and illumination account for a 20% decrease.

A paper on the impact of image distortions in face recognition by (JATURAWAT; PHANKOKKRUAD, 2017), evaluated the face recognition accuracy of three well-known algorithms: Eigenfaces (TURK; PENTLAND, 1991), Fisherfaces (BELHUMEUR; HESPANHA; KRIEGMAN, 1997), and LBPH (KADIR et al., 2015), under unconstrained conditions. The experiments considered a variety of poses and expressions, as well as different light exposures, noise levels, and resolution. The results showed that all three algorithms were severely affected by the considered distortions, with an emphasis on resolution and light exposure as the distortions which caused the greater effects.

Focusing on image quality, the authors of (LI et al., 2019) surveyed the approaches to deal with low-quality images in face recognition. They stated that the main challenges lay in the first stages of the face recognition pipeline: face detection and face alignment. According to this survey, face detection is particularly impacted by low-resolution images, and for the case of face alignment, the best performing algorithms aren't trained to consider image distortions, so it could be concluded that in the presence of low-quality images, their performance will suffer.

Their survey showed two main approaches for dealing with low-quality images: pre-processing algorithms oriented to enhance the image's quality: restoration methods, deblurring, denoising, super-resolution methods; and DL algorithms that comprise several stages of the

[1] https://www.cognitec.com/facevacs-technology.html

face recognition pipeline. According to the authors, the most promising approach to improve face recognition with low-quality images is DL, however, the research is still not clear as to how to address specific issues like blur, noise, and low-resolution representation.

Another study by (MEHMOOD; SELWAL, 2020) made a review of face recognition methods and the factors affecting their accuracy. The study divided the algorithms into appearance-based methods, feature-based methods, and hybrid methods, and evaluated their strengths and limitations while listing the main factors affecting face recognition. According to this paper, there are intrinsic and extrinsic factors that impact face recognition accuracy. The main intrinsic factors are aging, facial expressions, and plastic surgery; while illumination, pose variation, occlusion, noise, and low-resolution account for the major extrinsic factors. The conclusions of this study show that there are multiple approaches to improve performance under the presence of specific intrinsic and/or extrinsic factors: namely pose correction, super-resolution, image enhancement, and DL. Although some methods have shown promising results, there is no clear path towards a face recognition method strongly enough to perform well in both constrained and unconstrained conditions.

Research has also been made about the influence of image quality in other face processing algorithms like age estimations and facial expression classification. In (TIAN; CHEN, 2012), the authors study the effect of image resolution in facial expression recognition. The authors used two databases: Cohn-Kanade (LUCEY et al., 2010) and FABO (GUNES; PICCARDI, 2006), and modified the images to obtain five different resolutions. Since facial expression recognition algorithms are composed of three main stages: face acquisition, face feature extraction and representation, and facial recognition, the authors decided to evaluate the performance of the available methods in each stage. Several methods were evaluated in each stage. The results obtained in this study were consistent across all stages, a decrease in performance in the selected methods was observed when dealing with low-resolution images.

Focused on age estimation algorithms, (NGUYEN; CHO; PARK, 2015) and (NGUYEN et al., 2015) proposed methods to overcome the effects of optical and motion blur in age estimation. According to the authors, the main studies and proposals for age estimation haven't dealt with blurred images, which makes them susceptible to this kind of distortion. Both methods are based on the same approach, which consists in first identifying the type and degree of blur affecting the images, and then applying the adequate age estimator for the type of blur detected previously. The results obtained in both papers showed that age estimation accuracy is enhanced when using this approach, compared with the traditional methods that don't consider blurring effects.

Several conclusions can be drawn based on the previously cited literature. First, image quality does influence face processing algorithms, constituting an important factor to take into account when deploying such systems. Second, among the properties that describe image quality, image resolution seems to be the main factor affecting performance. Third, although there are several investigations about improving performance under certain image conditions, there is a lack of comprehensive studies focused in analysing the impact of specific distortions, like blur, noise, contrast, and compression, on the performance of face processing algorithms and to what extent that impact is relevant to the different tasks within the face processing domain. Fourth, there is no one approach to deal with the effect of image quality, several

methods have been proposed to tackle this issue, with the most common being image pre-processing to decrease the degree of a specific distortion, distortion-specific algorithms, and end-to-end systems based on DL.

Deep Learning has been successfully used to deal with distorted images in image classification tasks involving natural scene images. The best performing algorithms registered in the literature were proposed by (ZHOU; SONG; CHEUNG, 2017) (DIAMOND et al., 2017) (KIM et al., 2017) (DODGE; KARAM, 2018) (SANDLER et al., 2018) (BYUN et al., 2019), all consisting in very deep CNNs (DCNN). However, a recent paper by (HA et al., 2019) suggested the use of distortion-specific deep neural networks, also known as expert networks or dedicated networks, to correctly process images according to the distortions affecting them. According to (HA et al., 2019), the accuracy levels reached by these expert networks are comparable if not higher than the ones showed by the aforementioned DCNN architectures, with the advantage of requiring considerably less computational cost, making it easier to deploy in different platforms.

The adoption of either one of these approaches can be greatly benefited by the knowledge of the distortion affecting the images as well as the degree to which it is present. This constitutes the main motivation for this work. A distortion classification method for face images can be the first step for many of the face processing algorithms currently deployed in real-world environments, providing qualitative knowledge about the face image's quality in a way that isn't now available.

1.2 Problem

The domain of image processing applications, specifically face processing algorithms, has increased considerably, and with it, the interest in algorithms to assess and enhance face images. In recent years, different proposals have been made for dealing with images of varying qualities, with promising results. However, there still room for improvement and new approaches. Distortion classification in face images is a topic of little research, with a need for a method that can identify the distortions present in face images and provide qualitative information about them.

1.3 Objectives

The main objective of this work is to develop a CNN-based method that can correctly classify distortions in face images, as well as to estimate the degree to which a particular distortion is affecting the images.

To accomplish the said objective, the present work is divided into specific objectives and associated tasks:

1) Study the impact of blur, noise, brightness, contrast, and JPEG compression in the performance of face processing algorithms: three face processing algorithms will be evaluated under different levels of gaussian blur, motion blur, gaussian noise, contrast,

brightness, and JPEG compression. The obtained results will be the basis for the design of the final distortion classification method.

a) Select and implement three different face processing algorithms for evaluation.

b) Apply ten levels of the aforementioned distortions into the validation databases for each of the selected algorithms.

c) Evaluate the results obtained in each algorithm for each of the distortions applied.

2) Create the database to train and test the distortion classification method: using the results obtained in the previous tasks, a set of distortions and distortions levels will be selected to be the target of the distortion classification method, and a dataset with distorted and undistorted images will be created:

a) Choose the type and degree of the distortions to be classified.

b) Select a set of high-quality face images with variations in gender, ethnicity, and age.

c) Apply the established degrees of previously selected distortions to the original set.

3) Design and implement the distortion classification method:

a) Design a CNN architecture for distortions classification in face images.

b) Train the designed architecture with a subset of the dataset created before.

c) Test the obtained model in the remaining subset of images from the dataset created before.

4) Validate the method:

a) Evaluate the performance of the distortion classification method in another set of images.

b) Evaluate the potential of our distortion classification model as a previous stage for face processing algorithms.

At the end of this investigation, it is expected to obtain the following results:

- A comprehensive study analysing the impact of image quality on face processing algorithms.

- A dataset of distorted and undistorted images to be used for face quality assessment algorithms.

- A distortion classification method for face images that can accurately recognize both the distortion affecting the images, as well as its magnitude.

1.4 Methodology

To achieve the aforementioned objectives, the next methodology will be followed:

- Review the state of the art of face quality assessment methods as well as distortion classification methods
- Study the impact of several image distortions in the performance of face processing algorithms.
- Select the type and degree of the distortions for the distortion classification method.
- Create a dataset for training and testing the classification method.
- Design the architecture of the classification method based on the review of previous distortion classification algorithms.
- Implement the proposed method.
- Validate the method in a public IQA dataset design for face images.
- Create a Proof Of Concept (POC).

1.5 Thesis Outline

The remaining part of the text is structured as follows:

Chapter 2 contains a summary of the state of the art of face image quality assessment algorithms (FIQA), as well as their limitations. The state of the art of distortion classification methods is also presented.

Chapter 3 presents a study of the impact of image quality in three face processing algorithms. The study analyses the performance of the algorithms under Gaussian blur, Gaussian noise, motion blur, low brightness, high brightness, contrast degradation, and JPEG compression.

Chapter 4 describes the proposal for the distortion classification architecture, as well as the methodology for the creation of the dataset.

Chapter 5 describes the results of the training and testing processes.

Chapter 6 exposes the results obtained during our method's validation on the IDEAL_LIVE_DFD.

Chapter 7 presents a use case scenario where our model is used as a previous stage to a face processing algorithm. An experiment is performed to assess its suitability in the aforementioned scenario.

Chapter 8 is dedicated to the conclusions of our work.

2 State of the Art

2.1 Introduction

Section 2.2 of this chapter contains a review of the main approaches proposed in the last two decades to assess the quality of face images. Even though face quality assessment has become a subject of great interest, almost all its efforts have been towards improving face recognition performance in real-life environments, so the proposals are set in this domain, and as such, have its intrinsic biases. A discussion about the current solutions is presented in 2.2.1.

Section 2.3 of this chapter exposes a review of the state-of-the-art methods for distortion classification in images. Little research has been made for distortion classification in face images, and so, the methods reviewed in this section are not particularly dedicated to face images but to natural scene images, however, they represent the state of the art in the distortion classification task.

2.2 FIQA

The first FIQA methods focused on assessing and verifying the image's compliance with the requirements of the ISO/IEC 19794-5 for facial biometrics (HSU; SHAH; MARTIN, 2006). According to the ISO/IEC 19794-5, there are five categories in which face images should be evaluated to meet said requirements: Format, Digital, Photographic, Scene, and Others. In 2006, the authors of (HSU; SHAH; MARTIN, 2006), proposed a framework to assess face images according to the aforementioned requirements to improve face recognition performance. To develop the framework the authors looked into the specific requirements of each category and used different methods to measure each property, for example, measuring spatial sharpness and linear motion blur to quantify focus. They tested three models to combine all the metrics into a final quality score, obtaining the best results with neural networks. Given the goal of this work, the quality score is closely related to the performance estimation of a face recognition system with the assessed image. In this case, the authors used the FaceIt (VISIONICS, 2004) to validate the previous statement.

In (ABDEL-MOTTALEB; MAHOOR, 2007), the authors presented several algorithms for assessing the quality of facial images concerning the effects of blurring, lighting conditions, head pose, and facial expressions. The authors developed individual metrics to measure each property and then computed the correlation between the metrics and the expected performance on face recognition algorithms based on the Eigenface technique (TURK; PENTLAND, 1991). To validate the proposed algorithms for face quality assessment, the authors used several face databases such as the Facial Recognition Technology database FERET) (PHILLIPS et al., 1998), the face database from West Virginia University[2], and the Cohn-Kanade face database (LUCEY et al., 2010). The focus of this work was to provide quality scores for face images that ultimately indicate the expected performance of a face recognition algorithm. An important

[2] https://biic.wvu.edu/data-sets/multispectral-dataset

distinction is that the authors proposed individual algorithms to assess face quality depending on specific properties like focus, illumination, pose, and expression.

The authors of (GAO et al., 2007) recognized that the main cause for poor performance in biometric tasks is low-quality samples. To address this issue, they proposed a method to standardize face image quality on defects categories. From the basis that the acquisition process results in imperfect images, the authors defined four aspects to categorize image defects or distortions: defects caused by the environment, defects caused by camera conditions, defects caused by user face conditions, and defects caused by user-camera positioning. The proposed methodology consists of evaluating the images according to each category, computing a face quality score for each one. All the scores are later normalized and mapped to obtain an overall quality score that indicates how good a sample is for biometric recognition. The authors did not mention a validation methodology for the proposed approach.

A statistical-learning-based assessment scheme to evaluate face image quality is presented by (LIAO et al., 2012). The facial features are extracted using a Gabor filter (CLARK; BOVIK, 1989), and a hierarchical binary decision tree based on Support Vector Machine (SVM) (AWAD et al., 2015) is later employed as a classifier. The images are labeled into five categories: excellent, good, average, fair, and poor. For training, the authors constructed a database of 22720 images of mainly digital scanned face pictures, originally taken using film. Each image was assessed by 10 persons with five-level labels. According to the authors, the performance of the proposed metric is consistent with the Human Visual System (HVS), The main advantage of this approach to the previously presented, is that the focus is face image quality, independently of the future use of the image, however, they did not test their method in public databases. Additionally, there isn't much information about the dataset in terms of image size or image distortions to correctly assess the efficacy of the method. The image examples presented in the paper showed grey scale images of poor quality.

An approach for assessing face image quality from video is proposed by (RAGHAVENDRA et al., 2014). The author's major interest was to develop a software to act as a previous stage for face recognition systems. The proposal consists of using Gray Level Co-occurrence Matrix (GLCM) to measure the statistical features of face images. The justification for that is in the author's belief that accurately measuring variations that are present in the face skin texture should reflect the overall quality of the face image, and the GLCM provides that measure (RAGHAVENDRA et al., 2014). An important aspect of this work is that images with no frontal faces are automatically categorized as having "bad quality", whereas the rest are categorized as "fair quality" and/or "good quality". This is not the only algorithm that employs pose estimator to assess quality, however, in this case, only non-frontal images are considered to be bad.

In (BOURLAI et al., 2014), the authors proposed a face quality metric based on photometric measures for efficient face recognition. The properties measured were: brightness, contrast, sharpness, focus, and illuminations. The authors studied the available photometric measures and compared their performance to ultimately choose the better ones for their purpose. The authors integrated the photometric measures with a neural network to obtain the final overall classification: good (0) or bad (1). To test the efficiency of face recognition with face

quality, five face recognition algorithms were chosen: PittPatt[3], Local Ternary Pattern (LTP) (SHARMA; ARORA, 2013), Local Binary Pattern (LBP) (ZHAO, 2011), Independent Component Analysis (ICA) (BARTLETT; MOVELLAN; SEJNOWSKI, 2002), and Principal Component Analysis (PCA) (FUKUI, 2014). The results obtained showed an increase in face recognition performance when selecting only good images from the face quality metric results.

A paper by (CHEN et al., 2015) proposed a face quality assessment framework based on the learning to rank methodology, for face recognition purposes. The learning to rank methodology is based on the assumption that if a face recognition algorithm performs better with one image than with another, the first one must have better quality (CHEN et al., 2015). Applying this premise, a series of equations and constraints are formulated to estimate an image's quality. Several experiments were carried on to test the proposed framework. The authors formed three datasets DB1, DB2, and DB3. The images from the first set had better quality than the second and the third sets, and the images from the second set were better than the ones in the third set. To create the first set the authors used images from the FERET database, the Face Recognition Grand Challenge (FRGC) dataset (PHILLIPS et al., 2005), and a Chinese ID card photo database. The Labeled Faces in the Wilds (LFW) (HUANG et al., 2007) and the Annotated Facial Landmarks in the Wild (AFLW) (KÖSTINGER et al., 2011) datasets were used to create DB2. DB3 consisted of non-face natural images in which the face detector generates false positive detection results. The results obtained with the experiments showed good accuracy when selecting high-quality images for face recognition, however, an important limitation of this work is that it only considers non-face images as bad quality images.

In (VIGNESH; PRIYA; CHANNAPPAYYA, 2016), the authors proposed a face quality assessment method for face recognition in surveillance video. The authors propose to measure the image quality by mimicking the recognition capability of a given FR algorithm using a CNN. The goal of this research is to efficiently select the best images for face recognition among the pool of images available from surveillance videos, defining the image quality as the potential performance of the face recognition system with that image. This configuration ensures that the face quality assessment method can adapt to the needs of the specific face recognition system. A CNN is used to model the performance of a FR system. In the conducted experiments, the authors used LBP and Histogram of Oriented Gradients (HOG) (DALAL; TRIGGS, 2005), as features extractor, and Mutual Subspace Method (MSM) (SAKANO; MUKAWA, 2000) for face image set matching. The authors validated their proposal in the ChockePoint dataset (WONG et al., 2011), and achieved state-of-the-art results in improving face recognition performance in the surveillance scenario.

Another face image quality assessment method focused on the surveillance video scenario is proposed in (KHRYASHCHEV et al., 2017). The image quality score was computed considering eight measures: image resolution, sharpness, symmetry, a measure of the symmetry of landmarks points S, a quality measure K (based on learning to rank) (CHEN et al., 2015), and two no-reference image quality metrics NRQ LBP, and the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE). The learning to rank methodology was used to classify the images according to these measures. The database used for testing was the 60 Person Face Comparison Database (60PFCD) (KHRYASHCHEV et al., 2017), composed of 600 images of

[3] https://www.crunchbase.com/organization/pittpatt

60 subjects. To test the efficiency of their proposal in face recognition, the Openface library[4] was used as a facial recognition system. The results obtained by the authors showed that by applying their proposed method for face quality, they achieved a 15-18% increase in face recognition performance.

In a paper by (KHASTAVANEH; EBRAHIMPOUR-KOMLEH; JOUDAKI, 2018), a face image quality assessment method based on photometric features and classification techniques is proposed. This study aims to provide a face quality assessment method that effectively selects the images that have enough quality for face recognition systems. The method consists in extracting a series of photometric features, normalize the results, and then feed them into a classifier. The features extracted correspond to measures of brightness, contrast, focus, and illumination. For the classification process, five different techniques were studied: K-Nearest Neighbors (KNN) (KAUR; JINDAL, 2016), Multilayer Perceptron (MLP) (KAUR; JINDAL, 2016), SVM (KAUR; JINDAL, 2016), Decision Tree (DT) (TCHENDJOU et al., 2016) and Gaussian Process Classifier GPC (RASMUSSEN; WILLIAMS, 2006). The experiments carried on showed that the MLP outperforms the other classifiers in terms of f1-score and accuracy measures. The dataset used in the experiments was the National Laboratory of Pattern Recognition (NLPR) face dataset (KHASTAVANEH; EBRAHIMPOUR-KOMLEH; JOUDAKI, 2018), composed of 450 frontal face images of different lighting and background conditions.

A face quality assessment framework based on face features is proposed in (BHATTACHARYA; ROUTRAY, 2018). The main goal of the research is to provide an easy-to-use method to evaluate face image quality to improve face detection and recognition performance. The framework is composed of a deep neural network whose input are feature vectors previously obtained using four different techniques: LBP, HOG, Efficient Data Encoding for Deep Neural Network Training (GIST) (JAIN et al., 2018), and CNN with transfer learning (HUSSAIN; BIRD; FARIA, 2018). The output is a qualitative classification of the image quality: “good”, “bad”, and “average”. The authors used several datasets for the training and testing processes, images from the Yale dataset[5], the FERET, and the AT-T Face Dataset (ORL) (ABBAS; SAFI; RIJAB, 2017) conformed the good quality images subset, the average subset was obtained by combining images from the Celeb-Face dataset (CAO; LI; ZHANG, 2018) and LFW dataset, and finally, the bad quality images subset was conformed of false-positive results of face detection algorithms. The results obtained in the validation experiment are comparable with the ones in (CHEN et al., 2015), also having the limitation of only considering false-positive results of face detection algorithms as bad quality images.

The authors of (BEST-ROWDEN; JAIN, 2018) proposed and compared two models for the prediction of face image quality, one based on human quality ratings (HQV) and the other one based on quality values computed from similarity scores, named MQV. The models proposed were trained and tested over a variety of unconstrained face images from the LFW database, and the IARPA Janus Benchmark A (IJB-A) dataset (GROTHER; NGAN, 2017). The extraction of the image features was carried out using a CNN initially trained for face recognition purposes. This paper had several conclusions, the most important are (1) human

[4] https://awesomeopensource.com/project/cmusatyalab/openface

[5] http://vision.ucsd.edu/~iskwak/ExtYaleDatabase/Yale Face Database.htm

ratings are correlated with face recognition performance of unconstrained images, and (2) automatic prediction of human quality ratings (HQV) is more accurate than a prediction of score-based face quality values (MQV).

In (ZENG et al., 2018), a face image quality assessment method is proposed to improve face verification tasks in forensic science. The authors proposed a quantitative analysis based on the verification performance of the face verification system, as a measure of the face image quality. The method is Full Reference as it needs reference or neutral images to compute a verification performance score that will be established as a benchmark for the other images. The similarity between the verification performance score is computed, and from that, a quality score is calculated. The method takes into consideration different factors that affect the verification performance like imaging angles, facial expressions, face occlusion among others. Besides the important bias that face recognition oriented FIQA methods have, this works adds the limitation of being full reference, making it difficult to implement in a wider range of scenarios.

A paper by (HERNANDEZ-ORTEGA UAM et al., 2019), proposed a CNN to predict face images suitability for face recognition purposes. The system is non-reference, uses a performance-based ground-truth, and has a numerical output between 0 and 1. According to the authors, "(…), *the quality measure is directly proportional to the expected accuracy of the recognition process when using a specific face image".* The authors used the BioLab framework (MALTONI et al., 2009) to generate the quality scores (ground-truth) and fine-tuned a pre-existing CNN trained for face recognition to obtain their regression model. The VGGFace2 database (CAO et al., 2018) was used to fine-tune and validate the model. All images were collected in unconstrained conditions. The main advantage of this proposal is the use of automatically generated quality scores to form a database, which avoids bias introduction from human operators. However, it has a big limitation, which is that the image quality is closely related to a specific face recognition algorithm, making the model potentially unsuited for other face recognition systems.

In (ZHUANG et al., 2019), the authors proposed a face image quality assessment framework for face recognition purposes. To develop the framework the authors trained a DCNN to output a general facial quality metric that considers brightness, contrast, blurriness, occlusion, and pose. The FERET database and the KinectFaceDB database (MIN; KOSE; DUGELAY, 2014) were used to train and test de proposed framework. The developed framework is capable of providing measurements for brightness, symmetry, contrast, and sharpness, as well as an overall quality metric composed of the weighted sum of said measurements. The results obtained showed well within-database performance as well as cross-database. The experiments also showed that the overall quality score was closely correlated to face recognition performance. However, the paper doesn't provide a comparison with other FIQA approaches.

In (LIJUN et al., 2019), multi-task learning is employed to assess face image quality. The authors propose a CNN architecture consisting of two modules. The first module is responsible for the feature extraction step, using a lightweight CNN. The second one, called the Quality Score Fusion Module, takes the features extracted before and computes the alignment score, the corrosion score, the deflection score, and the clarity score. Said scores are later fused

together to output an overall quality score. The authors pointed out the lack of databases for this type of research, which led them to collect images and labeled them following a specific methodology described in (LIJUN et al., 2019). The experiments conducted led to satisfactory results, as they compared their method to two other face quality assessment algorithms and obtained the best accuracy and the least execution time. Another set of experiments showed the increase in face recognition performance when applied to the proposed method as a previous stage. These last experiments were carried on the CASIA-Webface (YI et al., 2014) and the Ms-Celeb-1M database (GUO et al., 2016).

2.2.1 Limitations of FIQA methods

The previous section reviews the state of the art of the face image quality assessment methods, and in such, their evolution throughout the years. Although the majority of the methods registered good results in their respective publications, neither one of them has been widely adopted for face quality assessment purposes. Additionally, important limitations of these works can be outlined.

First, given that their focus is to improve face recognition related tasks, their deployment is limited to that domain. For most of these methods, the quality score is actually a measure of the expected performance for a specific face recognition algorithm, so they act more like a performance estimator than as a quality assessment system. As a consequence, their usage is highly dependent on the selected face recognition method.

Secondly, given that their focus is face recognition, face pose is often considered as a distortion, and in some cases, it is the only condition for bad quality images. Similarly, false-positive results from face detection methods, meaning images with no faces, are often classified with low-quality scores or as bad images. This poses a problem in that images are been classified as having bad quality regardless of the actual properties that define image quality, like focus, brightness, contrast, etc.

Finally, the datasets used for training and testing most of the aforementioned methods weren’t created for face quality assessment, but for face recognition benchmarking. This is a great limitation because these methods are based on machine/deep learning architectures, which means that their generalization capabilities are highly dependent on the training data. If the datasets aren’t good enough to represent the different conditions of image quality, the methods won't be able to generalize to those conditions, and thus their application will be limited. Additionally, using these datasets for image quality assessment can lead to bias introduction when generating the ground-truth.

2.3 Distortion classification methods

In 2010, the authors of (CHETOUANI; BEGHDADI; DERICHE, 2010) published a paper that deals with a limitation present in most IQA methods, that is their inability to effectively predict the image quality for all degradations. According to the authors, this limitation exists due to the direct link between the distortion’s specificities and the efficiency of IQA algorithms. To solve said limitation, this paper proposes an IQA method that first

classifies the distortion and then applies the corresponding image quality evaluator according to the results of the first step.

The authors considered eight different degradations or distortions, such as gaussian blur, image denoising, JPEG compression, additive Gaussian noise, spatially correlated noise, impulse noise, quantization noise, and JPEG2000 compressions. The distortion classification method consists of two steps: feature extraction, and classification. The feature extraction step was done using existing IQA metrics like PSNRM (PONOMARENKO et al., 2007), SSIM (OKARMA, 2019), and SNR (OKARMA, 2019). An ANN is used to classify the image distortion according to the features previously extracted. The authors used an MLP as architecture for the ANN and the TID2008 database (PONOMARENKO et al., 2009) for training and testing. The results obtained in the testing stage showed high accuracy levels for the majority of the degradations.

A paper by (KANG et al., 2015) proposed a Multi-Task CNN to simultaneously assess the image quality as well as identify its distortions. The proposed architecture is loosely based on the work of (KANG et al., 2014), the authors modified it to obtain a multitasking CNN. The network is formed by two convolutional layers, each with a pooling operation, two fully connected layers, and one output layer with two output functions, linear regression for IQ and logistic regression as a multiclass output for the distortion identification. To evaluate the method`s performance, the authors used three databases, the LIVE dataset (SHEIKH et al., 2006), the TID2008 (PONOMARENKO et al., 2009), and the CSIQ database (LIU; PEDERSEN; HARDEBERG, 2014). The results obtained with the experiments were satisfactory as the IQA prediction levels were similar to the best-known methods: PSNR (OKARMA, 2019), SSIM (OKARMA, 2019), FSIM (OKARMA, 2019), DIIVINE (KRISHNA MOORTHY; CONRAD BOVIK, 2011), BLIINDS-II (SAAD; BOVIK; CHARRIER, 2012), and BRISQUE (MITTAL; MOORTHY; BOVIK, 211AD). As for the classification task, the proposed method outperformed the aforementioned algorithms.

Another approach for IQA was proposed by (ALAQL; GHAZINOUR; CHANG, 2016) consisting of an improvement on the BIQI framework (MOORTHY; BOVIK, 2010) to achieve better results in image distortion classification. The BIQI framework consisted of a two-step IQA method that first classified the distortion present in the images and then applied a distortion-specific IQA method to estimate its quality. This paper proposes a set of features to classify image distortion more efficiently. The authors focused on the same distortions in the LIVE database (SHEIKH et al., 2006): JPEG, JPEG2000 (JP2K), white noise (WN), Gaussian Blur, and Fast Fading. A total of 197 features are collected using different IQA algorithms such as BIQI (MOORTHY; BOVIK, 2010), SSEQ (LIU et al., 2014), DIVINE (KRISHNA MOORTHY; CONRAD BOVIK, 2011), BRISQUE (MITTAL; MOORTHY; BOVIK, 211AD), BLIINDS-II (SAAD; BOVIK; CHARRIER, 2012), among others. The classification task was carried on with several techniques to evaluate their efficiency, finally stating that the best performing classifier on the Laboratory for Image & Video Engineering (LIVE) dataset was the multiclass classifier with logistic regression as a base classifier (MultiClassClassifier-Logistic ECC), obtaining an accuracy of 96.41%.

A CNN for distortion classification is presented in (BUCZKOWSKI; STASINSKI, 2019). The authors proposed two similar CNN architectures differing only on the number of

trainable parameters. The network is formed by three convolutional layers each followed by a max-pooling layer and a dropout layer, and all followed by two fully connected layers, where the last one has 4 neurons indicating the four classes considered: blur, additive noise, JPEG compression, and JPEG2000 compression. The dataset employed for the training and testing was composed of images from three IQA databases: the Categorical Image Quality (CSIQ) (LIU; PEDERSEN; HARDEBERG, 2014), LIVE2006 (SHEIKH et al., 2006), and VCL@FER (ZARIC et al., 2012). The results obtained in the testing stage showed good levels of accuracy for both models, with 92.12% and 94.14% respectively. The CNNs proposed were benchmarked against the Blind Image Quality Index (BIQI) framework (MOORTHY; BOVIK, 2010), where both CNNs outperformed said method by at least 10%.

In recent years, image classification under low-quality images has been the object of much research. Different approaches have been proposed, with two main directions: very large neural networks that can deal with images with different qualities; and two steps systems that firstly classify the distortions and then compute image classifications with dedicated networks for the particular distortion. Both approaches have demonstrated good performance and high accuracy levels, however, the latter approach comes with the additional advantage of requiring less computational power. In (HA et al., 2019), the authors proposed a low-cost classification method for distorted images using the two-steps approach. The CNN architecture used for the first stage is formed by four blocks of convolution layers and max-pooling layer of sizes 3*3 and 2*2 respectively, and two fully connected layers. The network was trained to distinguish between clear and distorted images with blur, noise, and low light, each distortion with three degrees, making a total of 10 classes. The authors call this network a Tiny CNN because of its simple architecture and relatively low hardware requirements. The databases used for training and testing were the Canadian Institute for Advanced Research (CIFAR-100) dataset (KRIZHEVSKY, 2009), the Caltech-256 (GRIFFIN; HOLUB; PERONA, 2007), and the Street View House Numbers (SVHN) dataset (NETZER et al., 2011).

2.4 Summary

As reviewed in previous sections, traditional FIQA algorithms are mainly focused on improving face recognition performance, and as a consequence, are limited in terms of their applicability outside of the face recognition domain. Nevertheless, reviewing their evolution is important to evaluate the best techniques available at the moment. From section 2.2 it is possible to extract the most common approaches: SVM and similar ML techniques, the learning to rank methodology, and deep neural network architectures with an emphasis on CNNs. From those, the latter seems to be the one with better performance and with more promising results.

On the other hand, the state-of-the-art methods for distortion classification in natural scene images are also based on CNN architectures. The last two reviewed methods are very interesting as they achieve high levels of accuracy and have relatively simple architectures Another important remark about the method proposed by (HA et al., 2019) is that their network is optimized for low-cost deployment, which makes it very attractive for real-time solutions.

2.5 Conclusions

This chapter presented a review of the state-of-the-art algorithms for FIQA and distortion classification, where the main proposals of both fields were described and analysed. Important annotations were made in subsection 2.2.1 about the limitations of the current FIQA methods available. Section 2.4 summarized the main aspects of both reviews, from what was decided to use the proposals of (BUCZKOWSKI; STASINSKI, 2019) and (HA et al., 2019) as the bases for our distortion classification method for face images.

3 Impact of image distortion in face processing algorithms

3.1 Introduction

This chapter is dedicated to analysing to what extent a selected group of distortions affects face processing algorithms. To achieve that goal, three different face processing algorithms were tested with images under different conditions of blur, noise, contrast, brightness, and compression.

The methodology adopted for the study is based on the work of (DODGE; KARAM, 2016). However, a few changes were made to adapt it to our goal. The main differences in our approach are that the selected algorithms are focused on different tasks as opposed to one, and that each algorithm was tested with a dataset and a set of metrics corresponding to the task in question. Also, three additional distortions were considered as a part of our study: motion blur, low brightness, and high brightness.

Sections 3.2, 3.3, and 3.4 contain details about the algorithms, the datasets, the metrics, and the distortions analysed in our study. In Section 3.5, a discussion of the results obtained with the experiments is presented. Finally, section 3.6 contains the conclusions of this chapter.

3.2 Face Processing Algorithms

3.2.1 FaceNet

FaceNet is a deep learning system that generates face embeddings for face recognition tasks, such as face identification and face verification, proposed by (SCHROFF; PHILBIN, 2015) in 2015. Since then, the proposed methodology has been widely used thanks to its good results in benchmark datasets like LFW and YouTube Faces DB (WOLF; HASSNER; MAOZ, 2011).

The main contribution of FaceNet is the introduction of a new loss for deep learning architectures, specifically made for face recognition purposes: the triplet loss. The authors described their motivation for this loss as follows "(…) *we strive for an embedding* $f(x)$*, from an image* x *into a feature space* R_d*, such that the squared distance between all faces, (...), of the same identity is small, whereas the squared distance between a pair of face images from different identities is large*.".

FaceNet uses two DCNN as base architectures: the Zeiler&Fergus (ZEILER; FERGUS, 2014) style networks and the Inception (SZEGEDY et al., 2015b) type networks. The authors proposed different variations of the aforementioned architectures in order to evaluate their performance under different image conditions and with different configurations. The results showed that the Inception model trained with 224*224 images achieved the best results with a fraction of the parameter of the Zeiler&Fergus based architectures.

The authors used accuracy (1), and the validation rate at a fixed False Acceptance Rate (FAR) of 0.001 as metrics to measure performance. The FAR is the probability that two images of different identities are classified as the same (2) (SCHROFF; PHILBIN, 2015). The

validation rate @FAR = 0.001 indicates the proportion of face image pairs that FaceNet can correctly identify as the same identity (3) (SCHROFF; PHILBIN, 2015) while keeping the FAR to 0.001.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$FAR = \frac{FP}{FP+TN} \quad (2)$$

$$Validation\ Rate = \frac{TP}{TP+FP} \quad (3)$$

$$TP = True\ Positives, TN = True\ Negatives, FP = False\ Positives, FN = False\ Negatives$$

For this study, an implementation of the FaceNet system based on the Inception architecture was chosen, and the performance of the algorithms was evaluated using accuracy, and validation rate under the same FAR used by the authors. Table 3.2-1 shows the details of the model used for the study.

3.2.2 Deep Expectation (DEX)

The DEX algorithm consists of a deep learning architecture for age estimation from a single face image and without the use of facial landmarks (ROTHE; TIMOFTE; VAN GOOL, 2018). The pipeline of the entire system consists of four main stages: face detection, face alignment and resize, feature extraction, and age estimation.

The face detection stage is done by applying a face detector both in the original image and the image rotated in 5 degrees steps from -60 to 60, plus an additional detection with -90, 90, and 180 degrees. The face with the best detection score across all the operations gets selected as the face image. The alignment stage is made simply by rotating the image according to the rotation angle of the image with the highest face score.

The feature extraction stage is achieved with a VGG-16 based CNN previously trained with the ImageNet dataset for image classification. The CNN network has a total of 16 layers, 13 convolutional, and 3 fully connected. The authors finetuned the VGG-16 architecture using a new dataset for age estimation called IMDB Wiki Faces Dataset (IMDB-WIKI) (ROTHE; TIMOFTE; VAN GOOL, 2018).

For the age prediction stage, the authors reformulated the problem as a classification task instead of a regression one. Being a classification problem, the authors defined age classes that covered a range of ages. According to the authors, this formulation "*increases robustness during training and accuracy during testing*".

The authors carried on several experiments to validate their proposal. The datasets used were the IMDB-WIKI, the Face and Gesture Recognition Network (FG-NET) (PANIS et al., 2016), the MORPH (K. RICANEK JR. AND T. TESAFAYE, 2006), the Cross-Age Celebrity Dataset (CACD) (CHEN; CHEN; HSU, 2014), and the Looking At People (LAP) dataset (ESCALERA et al., 2015). To measure the model's performance the authors used the mean absolute error (MAE) (4) in years and the e-error (5) (ESCALERA et al., 2015) for the datasets where there is no ground-truth. The results showed state-of-the-art results in the FG-NET and the MORPH datasets for real age, and in the LAP dataset for apparent age. The method was also validated in the APPA-REAL dataset by (CLAPES et al., 2018). Table 3.2-1 shows the details of the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} abs(y_i - \hat{y}_i) \qquad (4)$$

$n = number\ of\ samples$

$y_i = ground - truth\ for\ the\ i^{st}\ example$

$\hat{y}_i = predicted\ age\ for\ the\ i^{st}\ example$

$$\varepsilon - error = 1 - e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (5)$$

$\mu = mean\ of\ the\ normal\ distribution$

$\sigma = standard\ deviation$

3.2.3 Deep Alignment Network (DAN)

The DAN method consists of a CNN for image alignment proposed by (KOWALSKI; NARUNIEC; TRZCINSKI, 2017). The proposal is inspired in the Cascade Shape Regression (CSR) (XIONG; DE LA TORRE, 2013) framework, which consists of a combination of a sequence of regressors to approximate nonlinear mapping between the initial shape of the face and the desired frontal face (XIONG; DE LA TORRE, 2013). Like the CSR framework, the DAN method starts with an estimation of the face shape. From that, the model substitutes each CSR iteration for a deep neural network for both feature extraction and regressions. The authors stated that the main difference between the two approaches is that the DAN method extracts features from the entire face rather than the patches around the landmark positions. This is achieved by adding an extra input to each DAN stage, consisting of a landmark heatmap that indicates the current landmark locations within the face image. Each DAN stage takes three inputs: the face image, a landmark heatmap, and the feature image from a dense layer connected to the penultimate layer of the previous stage.

The authors performed several validation experiments using the 300W private test set (SAGONAS et al., 2013) and the 300W public test subsets (SAGONAS et al., 2013): the Intelligent Behaviour Understanding Group (IBUG), the Labeled Face Parts in the Wild (LFPW), and the Helen Facial Feature Dataset (HELEN). They used "*the mean distance between the localized landmarks and the ground truth landmarks divided by the inter- ocular distance*" as a measure of error, and evaluated their model's performance by computing the average of such error (mean error). According to the authors, they considered each image with an inter-ocular normalized error of 0.08 or greater as failure. Having that definition, they also

calculated the failure rate (5) and used it as another measure of performance. The results showed that the proposed model decreased the state-of-the-art failure rate by a margin of over 70%.

$$Failure\ Rate = \frac{Total\ failures}{Total\ Samples} \quad (6)$$

Table 3.2-1. Face Processing Algorithms

Architectures	Task	Metrics	Framework
DEX[6]	Age Estimation	MAE (years)	Keras
DAN[7]	Face Alignment	Mean Error, Failure Rate	Theano
FaceNet[8]	Face Recognition	Accuracy, Validation Rate @FAR = 0.001	Tensorflow

3.3 Datasets

The LFW (HUANG et al., 2007) will be employed to evaluate the performance of the FaceNet algorithm. The LFW dataset is composed of 13233 face images corresponding to 5749 individuals. All images were extracted from the internet, available as 250x250 pixel JPEG images, most of them in colour. The images are the result of the Viola-Jones (VIOLA; JONES, 2001) face detection algorithm and have been rescaled and cropped to the aforementioned size. The dataset comprehends a variety of scenarios with different head poses, resolutions, facial expressions, ages, genders, ethnicities, accessories, make-up, occlusions, and background. Due to the diversity of its composition, the LFW dataset has been widely used as a benchmark dataset for face recognition algorithms such as FaceNet.

To evaluate the performance of the DEX algorithm, the Real and Apparent Age (APPA-REAL) dataset (CLAPES et al., 2018) was used. The dataset contains 7591 images of 7000 individuals with ages ranging from 0 to 91 years, in unconstrained environments, and with varying resolutions. The APPA-REAL allows testing age estimation algorithms in both real and apparent age. For the study, only the validation set, was used, containing 1500 images.

Lastly, the challenging subset of the 300W dataset was used to assess the performance of the DAN method. This subset is called IBUG (SAGONAS et al., 2013) and consists of 135 images obtained from the Internet, with large variations of poses, expressions, and resolutions. The dataset provides landmark annotations for face alignment, obtained employing the Multi-PIE annotation scheme (GROSS et al., 2010).

[6] https://github.com/yu4u/age-gender-estimation
[7] https://github.com/MarekKowalski/DeepAlignmentNetwork
[8] https://github.com/davidsandberg/facenet

Table 3.3-1. Datasets selected for the study

Datasets	Purpose	No. of Images	Resolution
LFW	Face Recognition	13233	250*250
IBUG	Face Alignment	135	variable
APPA-REAL	Age Estimation	1500	variable

3.4 Distortions

Digital images may be subjected to several distortions from the acquisition stage of any image processing system. To illustrate the effects of image quality in face processing algorithms, five different distortions were contemplated: noise, blur, contrast, brightness, and JPEG.

Noise can be caused by low-quality camera sensors, or by the environmental conditions at the moment of the acquisition (MEHMOOD; SELWAL, 2020). For this study, we modelled the noise as a Gaussian distribution using the Gaussian function of the Scikit-image[9] library for Python, with mean equal to 0 and variance ranging from 0.01 to 0.1 in steps of 0.01. Figure 3.4-1 shows examples of the obtained images for the different datasets used.

Blur can result from unfocused camera lenses or moving targets (DODGE; KARAM, 2016). For this study, we simulated both motion blur and Gaussian blur. The first one was achieved by filtering the images with different sized kernels with value 1/(kernel size). Figure 3.4-2 shows examples of the output images. For the latter, the Scikit-image library was also used to model the Gaussian blur, where we varied the standard deviation of the kernels from 1 to 9 in order to obtained different levels of blur (Figure 3.4-3), with the most degradation corresponding to the largest standard deviations.

[9] https://scikit-image.org/

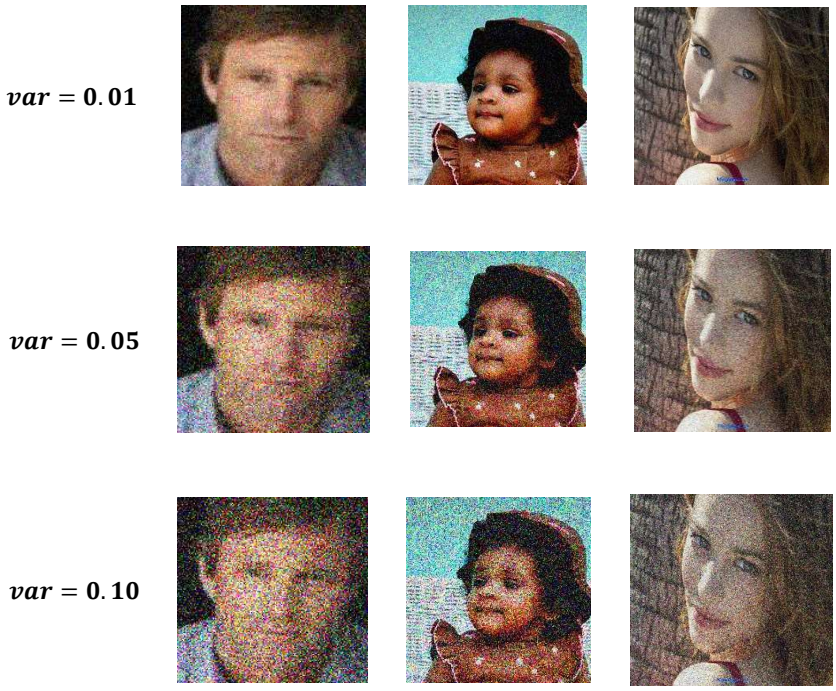


Figure 3.4-1. Examples of the images resulting from Gaussian noise degradation. From left to right: LFW, APPA-REAL, IBUG.

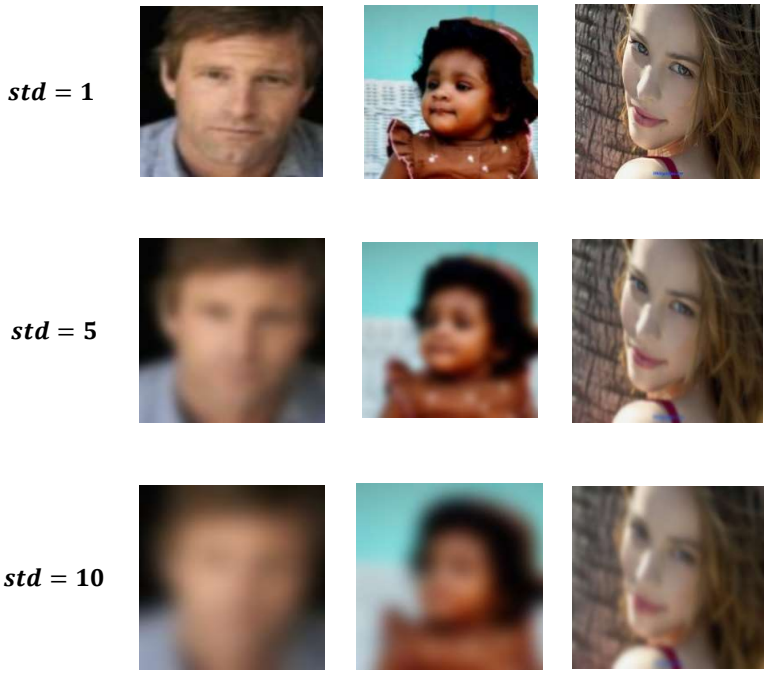


Figure 3.4-2 Examples of the images resulting from Gaussian blur degradation. From left to right: LFW, APPA-REAL, IBUG.

Figure 3.4-3. Examples of the images resulting from motion blur degradation. From left to right: LFW, APPA-REAL, IBUG.

To evaluate the effect of reduced contrast in face processing, we used the Pillow[10] library for Python and simulated different levels of contrast by changing the contrast factor in the images from 1 to 0, where 0 means no contrast. Figure 3.4-4 shows examples of the obtained images.

Illumination is an important aspect of image quality, affecting both the human and the machine capacity of recognizing details. One way to simulate low and high illumination conditions is through image brightness. In that sense, we simulated 10 stages of high and low brightness by altering the brightness factor of the images using the Pillow library in Python. For low brightness, we altered the brightness factor from 1 to 0, in steps of 0.1. For high brightness, the established range was 1.2-3.0 with steps of 0.2. Figure 3.4-5 shows the obtained images.

JPEG compression is often cited as a distortion to study due to its intrinsic characteristics, meaning, it is a type of compression that provokes loss in the final result. As was stated in the study carried on by (DODGE; KARAM, 2016), it is interesting to analyse if the algorithms are affected by the quality of the compression and in what measure it is relevant. To evaluate the influence of JPEG compression in the performance of the algorithms, the Pillow library was used to obtain 10 levels of quality ranging from 5 to 95 in steps of 10. Examples of the obtained are shown in Figure 3.4-6.

[10] https://pillow.readthedocs.io/en/stable/

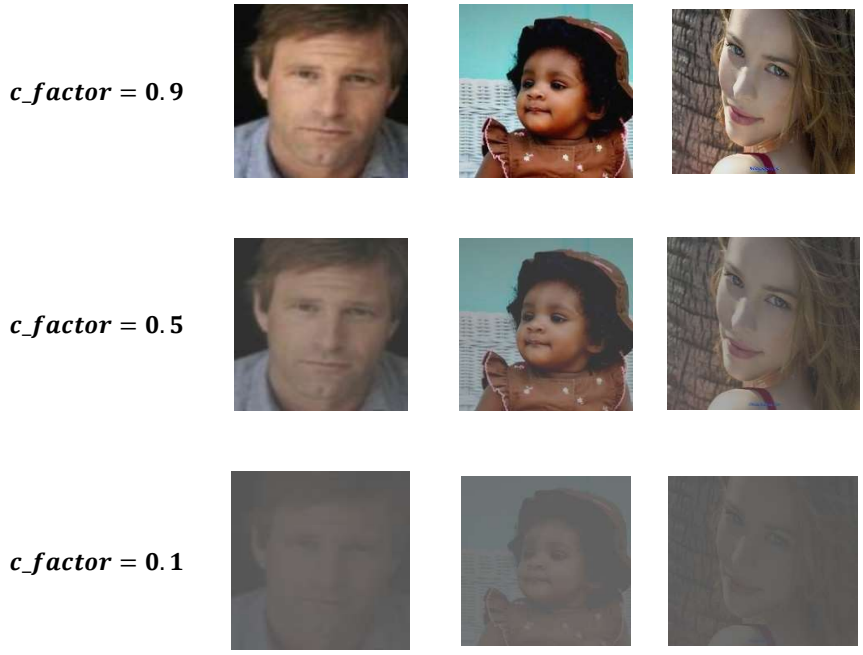


Figure 3.4-4. Examples of the images resulting from contrast degradation. From left to right: LFW, APPA-REAL, IBUG.

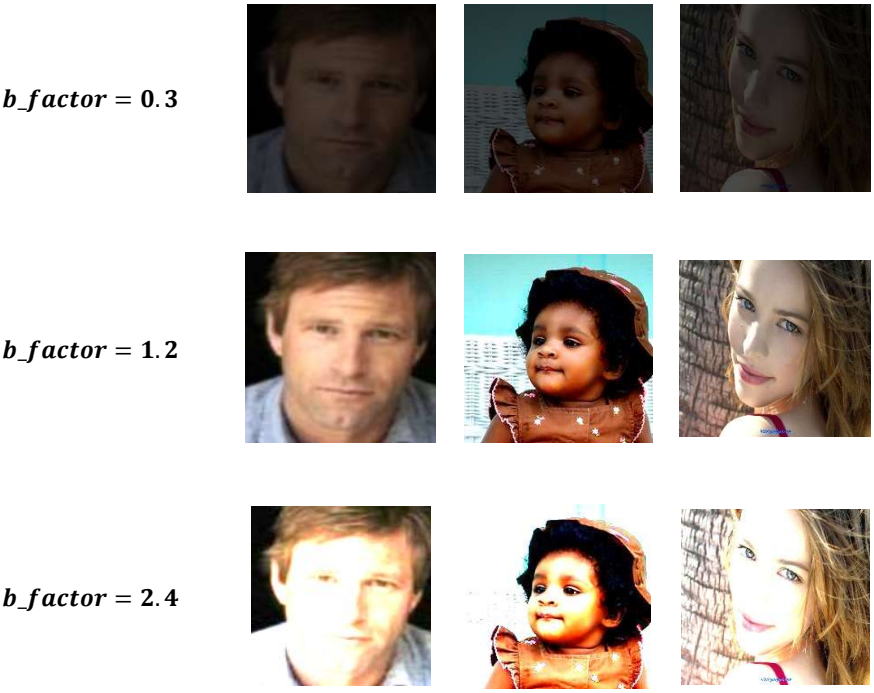


Figure 3.4-5. Examples of the images resulting from brightness degradation. From left to right: LFW, APPA-REAL, IBUG.

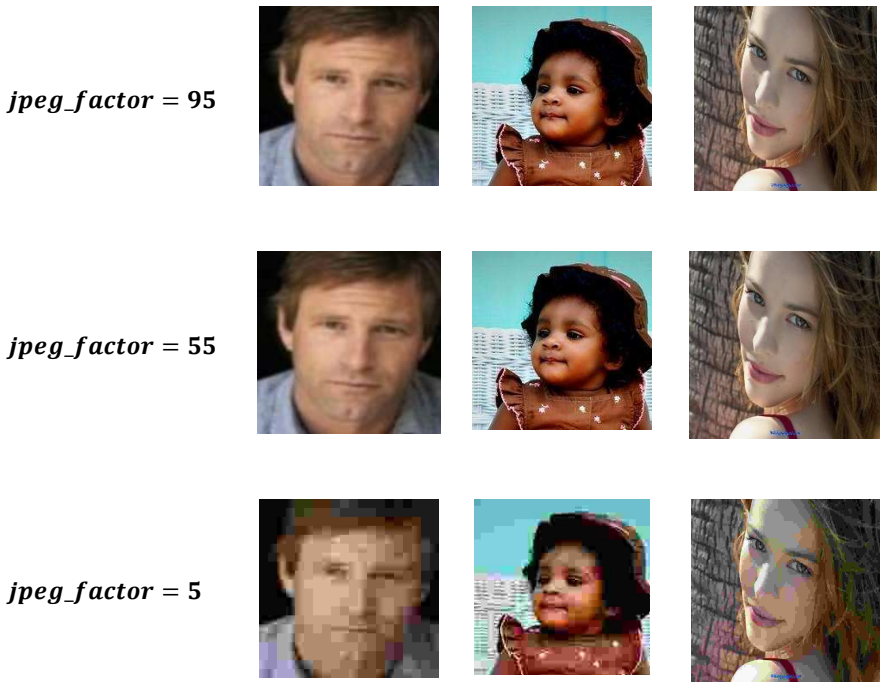


Figure 3.4-6 Examples of the images resulting from JPEG quality degradation. From left to right: LFW, APPA-REAL, IBUG.

3.5 Results

To comprehend the results obtained with the experiments, it is important to understand their methodology. The DEX and DAN algorithms have only one task each, so the experiments consisted of evaluating their performance on the specific task, under images with different distortions at different magnitudes. However, FaceNet is a more complex system designed to generate embedding for face recognition tasks such as face identification and face verification. Face identification consists in assigning an identity to a face through a one-to-many operation, where the embeddings of the unknown face are compared with the ones in the dataset in order to output the corresponding identity. Face verification, on the other hand, is a one-to-one operation, where the task is to check if the person's embeddings are close enough to the embeddings of the identity he or she claims to be.

To evaluate the FaceNet performance under different quality conditions, the experiments followed the same methodology proposed by (HUANG et al., 2007), where the system has to classify a pair of images as belonging to the same person or different ones, according to previously established pairs of matched and mismatched persons from the dataset. In other words, the experiments will be evaluating the algorithm's performance in a verification-like operation.

The website for the LFW dataset states that it is "*very difficult to extrapolate from performance in verification to performance in 1:N recognition*", although, given the nature of

these two tasks, it is safe to assume that any changes in the algorithm performance during verification operations, will be more noticeable during identification.

3.5.1 Noise

Figure 3.5-1 shows the behaviour of both accuracy and validation rate at a FAR = 0.001 for the FaceNet algorithm, across the different levels of Gaussian noise. The behaviour represented in the graph demonstrates that both metrics are affected by the noise, however, there is a significant difference between the overall accuracy of the model and the validation rate when the FAR is set to 0.001. Even at the lowest variance levels, the validation rate suffers considerably more compared to the accuracy. The algorithm appears to be robust in terms of accuracy, however, as was stated before, a bigger impact could be seen in the identification task. Both metrics get worst as the variance increase, with the accuracy reaching a little less than 80% in the highest level of noise, and the validation rate getting to 5%. The exact values can be seen in Table 10.1-1 in the Appendix section.

The behaviour of the DEX algorithm under the different levels of noise can be seen in Figure 3.5-2, where both classifications, apparent and real age, seem to be equally affected. The graph shows how the error in both estimations significantly increases even at the lowest noise levels. It is also noticeable how both curves plateau with variance = 0.06 and higher, reaching a MAE of approximately 16 years for the case of the real age, and 15 years for the apparent age.

Finally, Figure 3.5-3 shows how the DAN algorithm behaves across different levels of noise. Similar to the previous algorithms, its metrics worsen under the presence of noise, with the failure rate being the most affected metric. The degradation observed in the mean error is still noticeable, as can be seen in Table 9.1-2, the mean error reported for a variance of 0.01 represents a 43.7% increase compared to the value achieved without distortions. For the case of the failure rate, the results are very alarming, as it reached a maximum of 0.985 (Table 10.1-1).

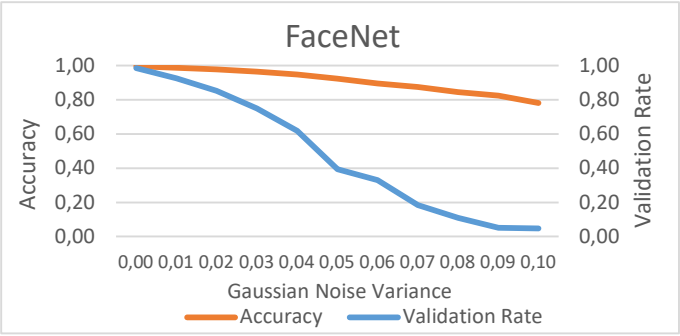


Figure 3.5-1. FaceNet behavior across noise levels.

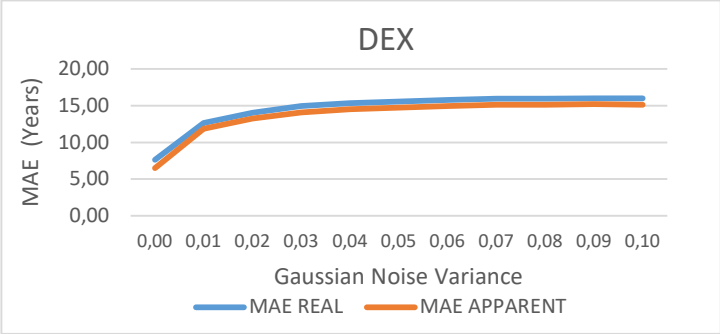


Figure 3.5-2. DEX behavior across noise levels.

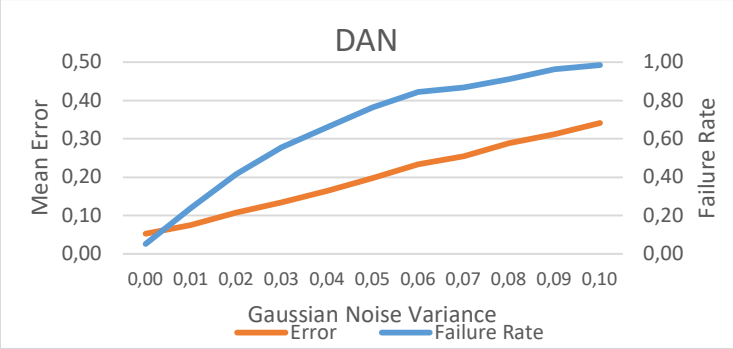


Figure 3.5-3. DAN behavior across noise levels.

3.5.2 Blur

3.5.2.1 Gaussian Blur

Figure 3.5-4 shows the behavior of the FaceNet algorithm as the images get increasingly blurred with a gaussian distribution. Similar to the graph in 3.5-1, the results of the validation rate at a fixed FAR of 0.001 are worse than the overall accuracy. However, in this case, the accuracy reached lower levels than in the previous experiment, with the final value below the 70% mark. A significant decline in both accuracy and validation rate is observed after a standard deviation of 3.0, where up to that point the accuracy stayed above the 95% mark, and the validation rate was approximately 85%, however, from that on, both metrics started decreasing at a higher rate.

Figure 3.5-5 shows the curves corresponding to the MAE values for the apparent and the real age classification with the DEX algorithm. As can be seen in the graph, both curves behave in the same way. It is interesting to observe a slight improvement in both metrics under a gaussian blur with a standard deviation of 1.0. Since blurring techniques are used for denoising, might be the case that some of the images in the dataset were noisy, and the smoothness caused by that level of blur helped achieve better results. From that point on, both metrics worsen, reaching MAE values of approximately 14 for the real age, and 13 for the apparent.

The DAN behaviour across the ten levels of Gaussian blur is shown in Figure 3.5-6. Both the error and the failure rate curves present similar behaviours, however, under a standard deviation of 1.0, the error stayed practically the same, where the failure rate almost doubled (Table 10.1-3). Similar to the noise experiments, the DAN performance worsens under the presence of blur, however, the mean error reached higher values with blurred images than with noisy images, in turn, the failure rate was less affected compared with the previous experiment.

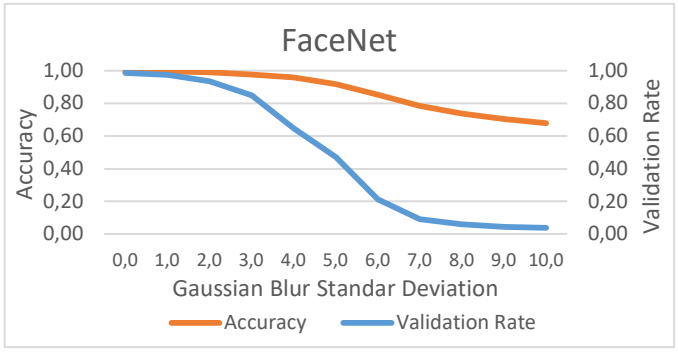


Figure 3.5-4. FaceNet behavior across gaussian blur levels.

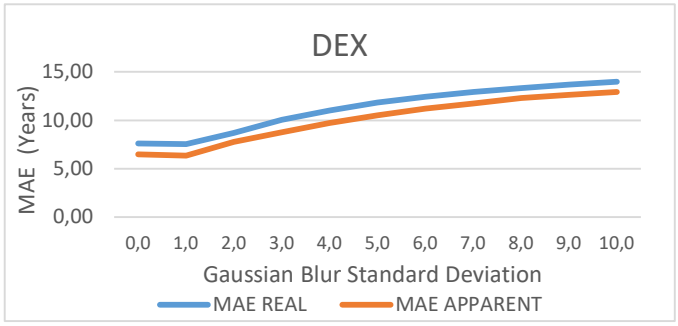


Figure 3.5-5 DEX behavior across gaussian blur levels.

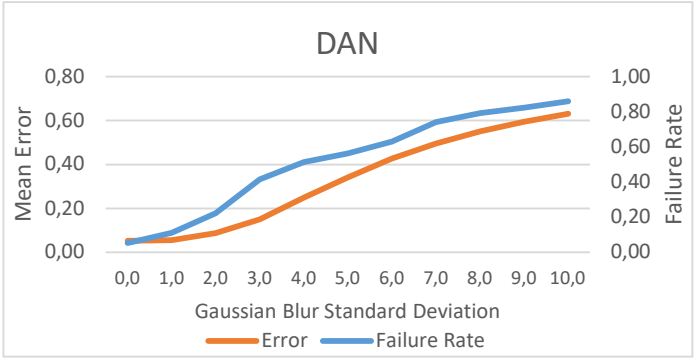


Figure 3.5-6. DAN. behavior across gaussian blur levels.

3.5.2.2 Motion blur

Figure 3.5-7 shows the curves of the accuracy and the validation rate obtained with this experiment for the FaceNet algorithm. Contrary to the results observed with noise and gaussian blur, the motion blur impacted significantly less than the previous distortions. The overall accuracy stayed almost constant across all kernel sizes, slightly decreasing towards the bigger ones. The validation rate at FAR = 0.001 shows a bigger decrease rate than the overall accuracy, reaching a minimum value of approximately 68%, which is significantly higher than the values obtained in the previous experiments.

Motion blur also had a lesser impact on the DEX algorithm than the previous distortions. Figure 3.5-8 shows how both metrics increase as the kernels get bigger, reaching MAE values of approximately 9 years for the apparent age, and 10 for the real. The graph shows a slight improvement in both metrics under the smaller kernel, as was the case with gaussian blur distortions.

The performance of the DAN algorithms is shown in Figure 3.5-9. Both metrics increase as the kernels get bigger. For the mean error, little variation is observed for kernel sizes of 3, 5, and 7, where the performance shows similar results to the undistorted images (Table 10.1-5). The same effect is observed towards the end of the curve, where the value of the error with kernel sizes of 19 and 21 remains practically the same. On the other hand, the values reached by the failure rate are significantly smaller than the ones obtained in the previous experiments.

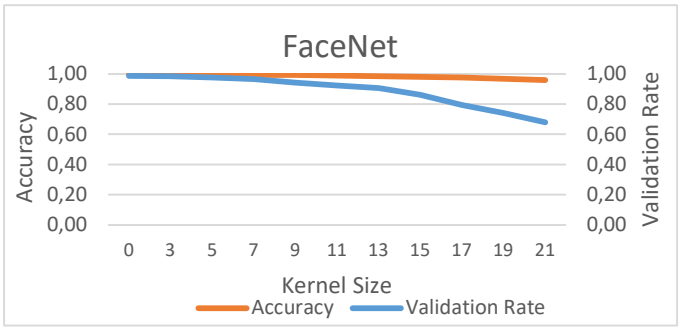


Figure 3.5-7. FaceNet behavior across motion blur levels.

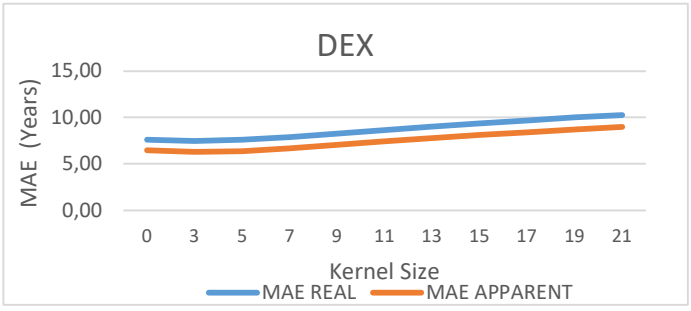


Figure 3.5-8. DEX behavior across motion blur levels.

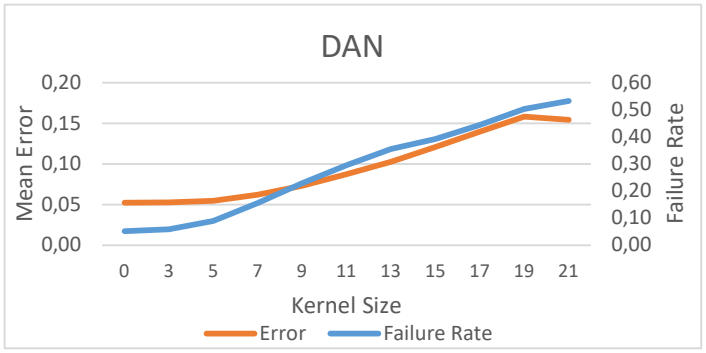


Figure 3.5-9. DAN behavior across motion blur levels.

3.5.3 Contrast

The impact of contrast in the FaceNet and the DAN algorithms is very similar, as is shown in Figures 3.5-10 and 3.5-12, respectively. Both algorithms show good and stable performance across most of the contrast factors, worsen only with contrast factors of 0.2 and 0.1. As was shown in section 3.4, such low contrast factors produce highly degraded images. For the case of the DEX algorithm, the contrast's impact is more noticeable, as can be seen in Figure 3.5-11. For the real age, the MAE reached 15 years, and for the apparent age estimation, the maximum MAE was 14 years.

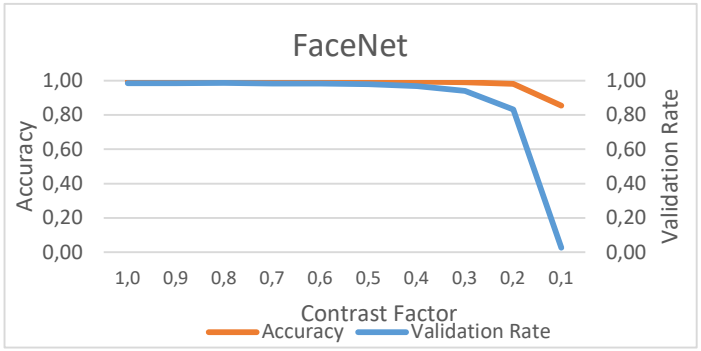


Figure 3.5-10. FaceNet behavior across contrast degradations.

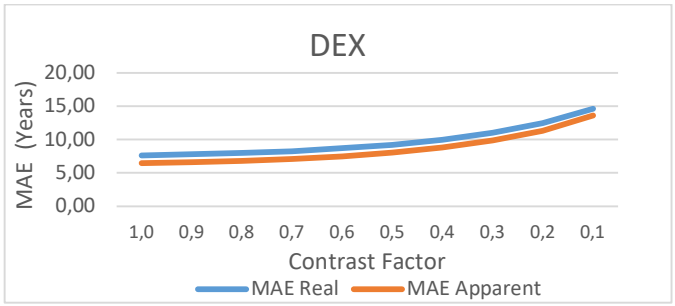


Figure 3.5-11. DEX behavior across contrast degradations.

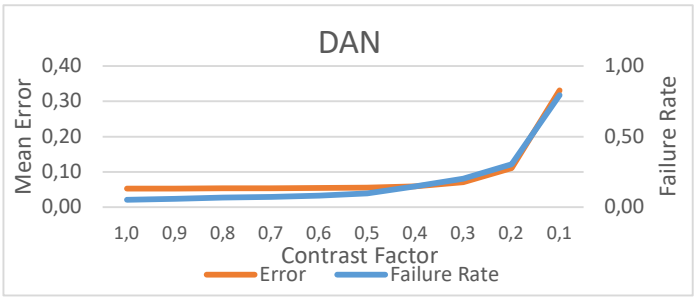


Figure 3.5-12. DAN behavior across contrast degradations.

3.5.4 Brightness

3.5.4.1 Low brightness

Similar to the effect observed in Figures 3.5-10 and 3.5-12, low brightness seems to have a small impact on the FaceNet and the DAN algorithms. Figures 3.5-13 and 3-15 show the behaviour of the corresponding metrics for both algorithms, and like the previous experiment, the curves remained stable for most of the brightness factors, changing only with severely degraded images. Figure 3.514 shows the behaviour of the MAE values for the real and the apparent age estimation with the DEX algorithm. The curves describe similar trajectories to the ones in Figure 3.11, however, the values obtained in this experiment indicate a smaller impact of low brightness than of contrast.

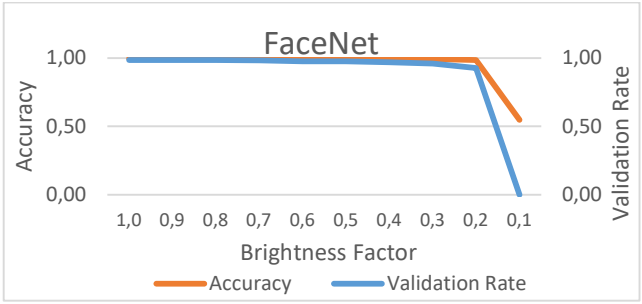


Figure 3.5-13. FaceNet behavior across low brightness.

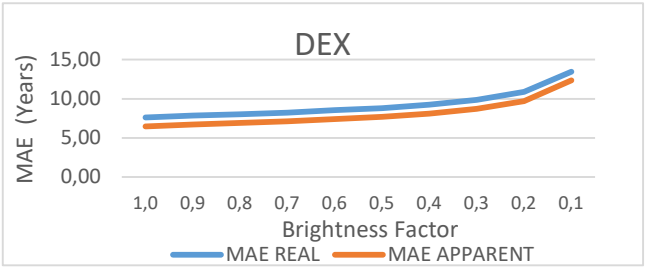


Figure 3.5-14. DEX behavior across low brightness

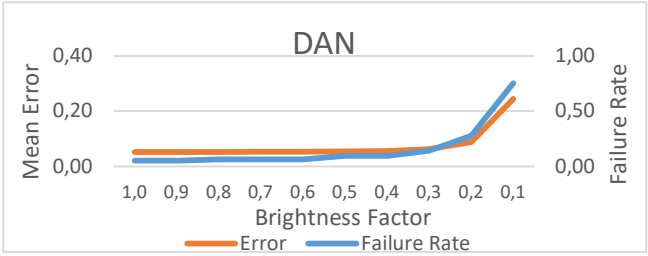


Figure 3.5-15. DAN behavior across low brightness

3.5.4.2 High brightness

Figure 3.5-16 shows the accuracy and validation rate values of the FaceNet algorithm across the different brightness factors used in this experiment. The results indicate that excess brightness has a bigger impact than the opposite situation in similar degradation levels, since both metrics were more affected during this experiment. It can also be observed that with brightness factors of 1.2, 1.4, and 1.6, the algorithm can still deliver good results.

The MAE values for the DEX algorithm are shown in Figure 3.5-17. As has been observed in previous experiments, both values increased as the images become more degraded. In this case, we can also see that both metrics remained stable with images at a brightness factor of 1.2.

The failure rate and the mean error of the DAN algorithm are graphed in Figure 3.5-18. The mean error curve shows good behaviour, increasing only up to 0.082 in the highest brightness value (Table 10.1-11). The failure rate curve denotes a bigger impact from high brightness; however, its failure rate under the most degraded images is the lowest the algorithm had reached at similar points in the previous experiments.

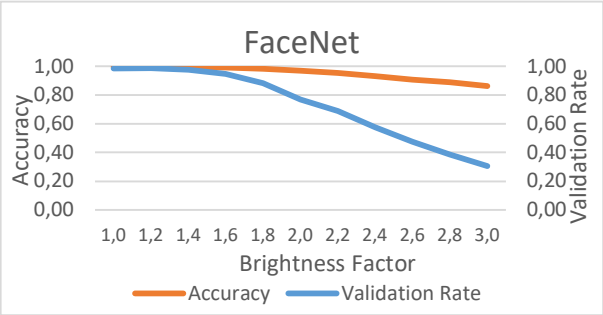


Figure 3.5-16. FaceNet behavior across high brightness.

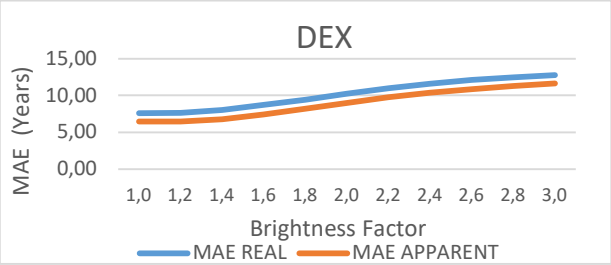


Figure 3.5-17. DEX behavior across high brightness.

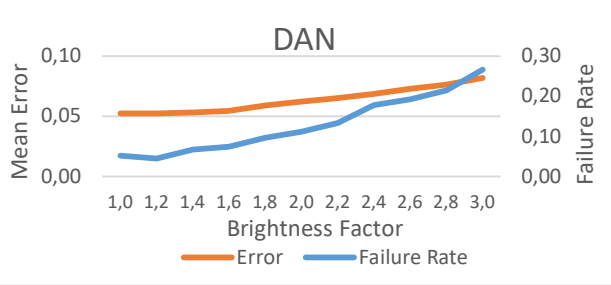


Figure 3.5-18. DAN behavior across high brightness.

3.5.5 JPEG Compression

The last analysed distortion was the JPEG compression. In this case, the goal was to observe the effect of different qualities of compression in the performance of the algorithms.

Figures 3.5-19, 3.5-20, and 3.5-21 show that the three algorithms are robust under different compression qualities. The only noticeable impact occurred, in all three of them, at the lowest quality factors.

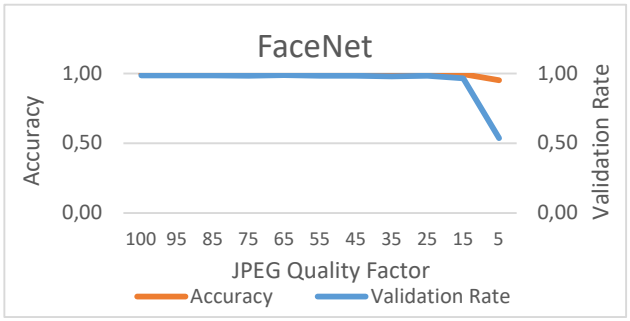


Figure 3.5-19. FaceNet behavior across JPEG quality levels.

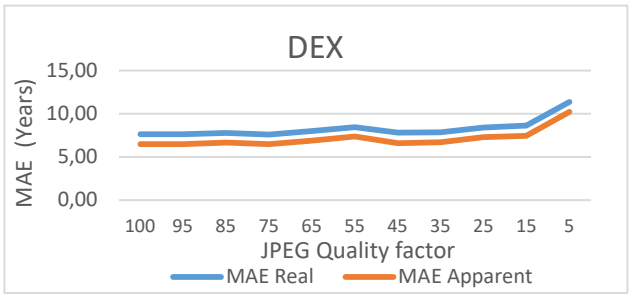


Figure 3.5-20. DEX behavior across JPEG quality levels.

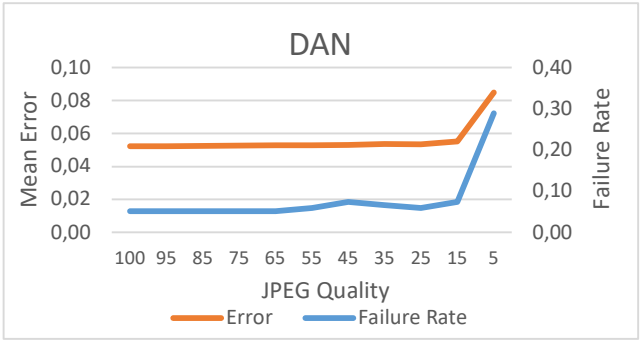


Figure 3.5-21. DAN behavior across JPEG quality levels.

3.6 Summary

The results obtained with the experiments show that even though the algorithms don't behave exactly the same, patterns can be observed. In that sense, a series of remarks can be outlined regarding the impact of each distortion in these algorithms.

First, noise and blur, in their Gaussian distribution, constitute the bigger threats to face processing performance in terms of image quality. Both distortions noticeably impacted the algorithms metrics even at the lowest levels of degradation.

Second, even though Gaussian blur severely impacted the performance of the algorithms, motion blur didn't have the same effect. The results show significantly less influence throughout the majority of kernel sizes. This is an interesting result because it indicates that not all blur constitutes a threat to performance, unfocused images and lack of detail have a bigger impact on performance than motion.

Third, contrast, low brightness, and JPEG compression all seem to have a small impact on performance. According to the graphs, metrics worsen only with the highest levels of degradation in each distortion.

Fourth, high brightness had a bigger impact on the low to middle levels of degradation, than the behaviour observed with low brightness at similar levels. However, at their highest levels of degradation, low brightness had a bigger effect on performance than its counterpart.

3.7 Conclusions

The focus of this chapter was to study the behaviour of three different face processing algorithms under the presence of noise, blur, contrast, brightness, and JPEG compression, at different levels. The goal was to draw conclusions about the impact of these distortions on face processing algorithms and obtain a more insightful understanding of the influence of quality in these types of algorithms.

Based on the results, a series of remarks were summarized in the previous section. From those remarks, we can conclude that the analysed algorithms, and potentially others, are unsuited for unconstrained environments where noise and blur, resembling Gaussian distributions, might be present. On the positive side, their deployment in scenarios with different conditions of contrast, JPEG compression, and brightness, would not be compromised unless the images are severely distorted.

Additionally, the information presented in this chapter might be useful to develop adequate solutions for face image quality assessment methods, oriented to improve face processing performance with images of different qualities. In that sense, we believe that identifying the type and degree of the distortion affecting face images, in conjunction with the information presented in this chapter, could lead to the development of more robust face processing systems.

4 Model description

This chapter is dedicated to describing the methodology adopted in the implementation of the distortion classification method.

Section 4.1 contains an explanation of the Deep Learning technology and its intricacies, as well as a detailed characterization of the Convolutional Neural Networks. Based on that, section 4.2 introduces the architecture and specification of the proposed CNN model. An analysis of the available datasets is presented in section 4.3, as well as the methodology for the creation of a new one. Section 4.4 describes the methodology that will be adopted to validate the model. Finally, section 4.5 summarizes the chapter and illustrates the progress made to achieve the final goal of this work.

4.1 Deep Learning

Machine Learning (ML) is a subsection of Artificial Intelligence (AI) intended to replicate human behaviour to perform specific tasks. It depends on concepts and knowledge to design features from which it can make inferences (LECUN; BENGIO; HINTON, 2015). This very characteristic constitutes a limitation when we as humans, are incapable of translating our knowledge into features. Tasks like face recognition, image classification, and language translation fall under that scenario. A solution for that is Deep Learning, a subsection of ML that learns high-level, abstract features from simpler representations, establishing a hierarchy. DL uses neurons as the fundamental logistic units, which take different signals as inputs and process them with nonlinear operations as they transfer each output to the next layer of neurons (SCHMIDHUBER, 2015). A key component in the learning process is the backpropagation operation, which updates the network's weights in order to obtain the model that best describes the particular scenario. To perform that, a loss function is defined to measure how close are the results obtained in each forward pass, to the ground-truth. From that, the cost function of the entire network is computed, and the network's weights are updated using the gradient of the cost function with respect to each weight. This way of updating the weights is called Gradient Descent (GOODFELLOW; BENGIO; COURVILLE, 2016).

DL algorithms can be categorized according to their approach to the learning process as supervised, semi-supervised, and unsupervised (LECUN; BENGIO; HINTON, 2015). There is also a subset of algorithms following the technique of Reinforcement Learning (RL), usually within the group of semi-supervised or unsupervised learning (ALOM et al., 2019).

The supervised learning approach consists on training the DL algorithm using labeled data. This type of architecture needs a set of inputs, and their correspondent labels or desire outputs. The goal of the training is to optimize the DL network to achieve maximum approximation with the ground-truth labels (LECUN; BENGIO; HINTON, 2015). Architectures like Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) such as Long Short Term Memory (LSTM), and Gated Recurrent Units (GRU), are examples of architectures that use the supervised learning approach (ALOM et al., 2019).

Semi-supervised learning is a technique in which the training process occurs with a set of labeled data, and a set of unlabelled data. When using semi-supervised learning, it is common to assume that points that are close to each other in a region are more likely to share the same labels, and that data tends to form clusters that separate inputs with similar labels (CHAPELLE; SCHÖLKOPF; ZIEN, 2006). Some architectures used for this approach are LSTM, GRU, and Generative Adversarial Networks (GAN).

The unsupervised learning approach is used when there is no labeled data. In this case, the network needs to learn the unknown relations within the input data (LECUN; BENGIO; HINTON, 2015). Examples of unsupervised tasks are clustering, dimensionality reductions, and generative techniques. Auto-Encoders (AE), Restricted Boltzmann Machines (RBM), and GAN architecture are some of the usually employed networks for these types of tasks. RNNs like LSTM and GRU can also be used with this approach (ALOM et al., 2019).

As mentioned above, many types of architecture can be used with DL, depending on the desired task. Our case falls into the category of image classification, and as described in section 2.3, the state-of-the-art results in distortion classification, as well as other computer vision tasks, are achieved with CNN. For that reason, we chose that type of architecture as the base for our model. The following section details the characteristics of the CNN, as well as the required techniques and parameter configurations.

4.1.1 CNN

The CNN architecture consists in a combination of three types of layers: convolution, max-pooling, and fully connected (also called dense). Figure 4.1-1 shows an example of a typical CNN architecture. The convolution layers extract features from the input images, and as they propagate through deeper layers, the network learns higher-level representations. The pooling layer reduces the size of the feature maps, and a set of fully connected layers are often used as classifier. (GOODFELLOW; BENGIO; COURVILLE, 2016)

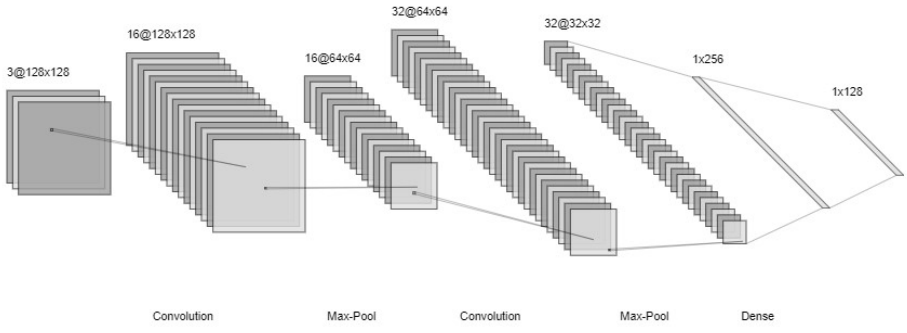


Figure 4.1-1. Example of a typical CNN architecture. (GOODFELLOW; BENGIO; COURVILLE, 2016)

The behaviour of a convolutional layer is defined in Equation 6 (GOODFELLOW; BENGIO; COURVILLE, 2016), where x_j^l represents the output of the layer, x_i^{l-1} represents a feature map from the previous layer, which is convolved with the kernel for the l^{th} layer k_{ij}^l, b_j^l represents the l^{th} layer bias, and M_j stands for the number of kernels to convolve the previous feature maps with.

$$x_j^l = f(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l) \quad (7)$$

As stated before, the pooling layer performs a dimensionality reduction. The amount of feature maps stays the same, but their size is reduced following the layer configuration. Usually, the pooling kernels are of size 2*2 with a stride of 2, which results in output matrices of half the size as they were before. The dimensionality reduction is achieved by taking either the minimum, maximum or the average value of the numbers in the feature maps within the area formed by the pooling kernel (GOODFELLOW; BENGIO; COURVILLE, 2016).

The classification portion of the network is performed by a combination of fully connected layers, also called dense layers, which take the flattened representation of the feature maps obtained with the convolution operations as their input. These layers apply a non-linear function to their inputs, usually sigmoid, tanh, or ReLU (GOODFELLOW; BENGIO; COURVILLE, 2016). There is no rule for the amount of fully connected layers needed to perform classification, it mostly depends on the architecture itself, however, is common to see between two and four layers.

4.1.2 Hyperparameters

When designing a DL network, it is important to configure a set of hyperparameters that define the network architecture. For the case of CNNs, it is important to set the number of convolutional kernels in each layer, as well as the size, the stride, and the padding. The stride is the step that the convolution kernels take, and the padding refers to the desired output size after the convolution (GOODFELLOW; BENGIO; COURVILLE, 2016). If configured as '*same',* padding will be added so the output feature map is the same size as the input, when '*valid'* is set, no padding will be added. It is also important to set up the hidden units in the fully connected layers and the kernel size of the pooling layer.

After configuring the network's architecture, it is also important to configure the hyperparameters that will determine the behavior of the training process. For this, the learning rate is the most important hyperparameter to configure since it determines how fast the network learns. However, a large learning rate doesn't guaranty fast and successful training, instead, is common for the network to diverge when the learning rate is set to a large value (SCHMIDHUBER, 2015). With smaller learning rates the network takes longer to train, but it tends to be more stable, unless the value is too small, in which case the network could get stuck at a local minimum (SCHMIDHUBER, 2015). A common configuration is to gradually decrease the learning rate during training.

Other important hyperparameters are β_1 and β_2, which are optimization parameters, namely momentum, and the magnitude of the gradient. Optimization methods are used to improve the network's weights update in the backpropagation step. The main optimizers available are Stochastic Gradient Descend (SGD), Adagrad, AdaDelta, RMSprop, and Adam (GOODFELLOW; BENGIO; COURVILLE, 2016). Of them, the Adam method has achieved the best results over the years, improving accuracy and helping the training process converge better. It uses both β_1 and β_2, and it considers the direction of the gradient instead of its absolute value when updating the network's weights, and also calculates the adaptative learning rate (ALOM et al., 2019). These two hyperparameters are usually configured to 0.9, and 0.999 respectively, since good results are obtained with that configuration (GOODFELLOW; BENGIO; COURVILLE, 2016).

Another important hyperparameter to configure before training is the batch size. Several studies indicate that performing weights update after each training example doesn't lead to successful results. In the same way, performing said update after all the training examples have been processed is also detrimental (LECUN; BENGIO; HINTON, 2015). The authors suggest using batches instead, where the backpropagation is done after a subset of examples is fed forward. Choosing the right size for the batches is important to avoid overfitting and achieve rapid convergence. The batches need to be supported by the CPU/GPU memory of the hardware used for training, and they are usually set to a power of 2: 32, 64, 128, etc (GOODFELLOW; BENGIO; COURVILLE, 2016).

The two main challenges of ML problems, and therefore of DL, are underfitting and overfitting. Underfitting occurs when the model is unable to learn the patterns and relations within the training data, and therefore cannot perform adequate predictions in new data. It is characterized by a large training error (or loss), and it usually means that the model is not complex enough (GOODFELLOW; BENGIO; COURVILLE, 2016).

Overfitting, on the other hand, occurs when the model learns the patterns and relations within the training data too well, to the point of not being able to generalize to unseen data. Overfitting can be recognized when the gap between the training error (or loss) and the test error is too big. A model that is too complex, or a small training set, are some of the main causes of overfitting (GOODFELLOW; BENGIO; COURVILLE, 2016).

Given that the quality and the quantity of the available data is usually a constraint when developing DL architectures, and that this is also a main cause of overfitting, several techniques have been proposed to avoid overfitting from happening. L2 regularization, Dropout, and Early Stopping are amongst the most popular ones (GOODFELLOW; BENGIO; COURVILLE, 2016). L2 regularization, also known as weight decay, uses a regularization parameter λ that modifies the cost function to obtain smaller network weights after backpropagation. Dropout is a simpler regularization technique that consists of setting to zero a random set of activations, the number of activations to be modified in this manner is calculated with the $keep\_prob$ hyperparameter, which establishes the activation percentage that will be kept unmodified. The early stopping technique consists in stopping the training process whenever a condition is reached, usually lack of improvement in the metrics (GOODFELLOW; BENGIO; COURVILLE, 2016).

4.2 Model Overview

4.2.1 Class definition

Our model was designed to classify six distortions: Gaussian noise, Gaussian blur, motion blur, low brightness, high brightness, and JPEG compression. All distortions were divided into levels, to provide additional information about the degradation suffered by the images. The distortions were selected according to the results obtained in Chapter 3. As the Gaussian noise and Gaussian blur showed the biggest impact on performance, three levels were defined for both distortions. For the others, two levels were defined. Including undistorted images, a total of 15 classes were defined to be the target of our distortion classification system. In the classes representing distortions, the higher the level the bigger the degradation. Table 4.2-1 summarizes the classes and parameter definitions. The parameters observed in the table are the same as the ones used in Chapter 3 to generate distorted images: variance (Gaussian Noise), standard deviation (Gaussian Blur), kernel size (Motion Blur), brightness factor, and JPEG quality factor.

Table 4.2-1. Classes definition according to type and degree of distortion.

Classes	Definition
Clean	Undistorted images
Gaussian Noise 1	var ϵ (0.005-0.02)
Gaussian Noise 2	var ϵ (0.05-0.065)
Gaussian Noise 3	var ϵ (0.1-0.25)
Gaussian Blur 1	std ϵ (0.5-2.5)
Gaussian Blur 2	std ϵ (4.5-6.0)
Gaussian Blur 3	std ϵ (8.5-10.0)
Motion Blur 1	k_size ϵ (7-13)
Motion Blur 2	k_size ϵ (17-23)
Low Brightness 1	f ϵ (0.8-0.5)
Low Brightness 2	f ϵ (0.3-0.05)
High Brightness 1	f ϵ (1.6-1.9)
High Brightness 2	f ϵ (2.7-3.0)
JPEG 1	q ϵ (80-35)
JPEG 2	f ϵ (20-5)

4.2.2 Model Architecture

The initial architecture of the model was based on the one proposed by (HA et al., 2019) for their Tiny CNN. As was described in Chapter 2, that Tiny CNN architecture was trained to classify natural scene images according to the presence or absence of three main distortions, at three levels. Since our goal is to identify more distortions, some changes were made. Our initial architecture had five blocks of convolutional and max-pooling layers, instead of the four proposed by (HA et al., 2019).

The kernel sizes were kept to 3 for each convolutional layer, with a stride of 1 and no padding. The number of kernels was initially set to 128, 256, 512, 128, and 64, in that order. The size of the max-pooling layer was 2*2, with a stride of 2. After each block, a dropout layer was added to perform regularization, with a $keep\_prob$ of 0.9. The last max-pooling layer was a global max-pooling layer, which in addition to reducing the dimensionality of the feature maps, also flattens the data and prepare it for the fully connected layers. The first fully connected layer had 256 hidden units with ReLU activation, and the last one had 15 units and Softmax activation. Softmax is a type of activation function commonly used in multi-class classification problems to compute the probability distribution of the numbers generated by the FC layers. The output of this function is the probability of each class (number between 0 and 1), with the target class having the highest value. The sum of all probabilities is equal to 1. Its mathematical formulation is shown in Equation 8 (GOODFELLOW; BENGIO; COURVILLE, 2016).

$$f(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (8)$$

The model's initial architecture is shown in Table 4.2-1, and its training configuration is described in Table 4.2-2.

Figure 4.2-1. Initial CNN architecture. (Source: Author)

Table 4.2-2. Initial configuration for training

Hyperparameter	**Configuration**
Metrics	Loss, Accuracy
Learning rate	0.0001
Learning rate decay	0.1 at loss plateau
Batch size	64
Early stopping	After no change in validation loss for 20 epochs
Optimizer	Adam
Loss function	Categorical cross-entropoy

The implementation and the training of the proposed architecture were done using the Keras[11] framework with Tensorflow[12] as backend. Image pre-processing consisted in normalizing the inputs to [0, 1]. The same operation is needed in deployment.

4.3 Dataset creation

As stated in subsection 2.2.1 of chapter 2, one of the limitations of traditional FIQA algorithms is that they are trained and validated with datasets designed for face recognition tasks. Some of the most popular datasets are the FERET dataset, the LFW dataset, and the Celeb-Face dataset. These datasets are used as a benchmark for face recognition tasks because of their composition: face images usually gathered from the internet, with individuals of different genders, ages, and ethnicity. They are created to resemble real-world environments, and as such, have images with different quality levels. The problem with that is that those datasets weren't created for image quality assessment, so even if they have images of different quality levels, they are not properly distributed and don't have the annotations to relate the images with a particular distortion at any given magnitude. In conclusion, those datasets are not suited for training and testing a distortion classification method.

In (GUNASEKAR; GHOSH; BOVIK, 2014), the authors created the IDEAL-LIVE Distorted Face Database (DFD) for face image quality assessment. The dataset is composed of 215 reference images gathered from the internet, each with one or more frontal faces. According to the authors, all the images are of good quality and without any visible distortions. The images were resized so the faces occupy a space of approximately 80*64 pixels, and then distorted images were obtained considering Gaussian noise, Gaussian blur, and JPEG compression. The authors aren't clear about the total amount of images generated, and only the original images are available to download.

Given the lack of datasets for quality assessment in face images, we decided to create our own dataset of face images for distortion classification. The first step was to select a group of undistorted high-quality images containing one frontal face to be the reference images. To that end, a subset of images from the Flickr Faces HQ dataset (FFHQ) (KARRAS; LAINE; AILA, 2019) was selected. The FFHQ dataset is composed of 70 000 high-quality images of 1024*1024 pixels, each containing one aligned frontal face. The images were collected from the Flickr[13] site, and have a variety of ages, genders, and ethnicity.

A general rule of thumb when working with DL and images, is to train with a minimum of 1000 examples per class for image classification. This rule is based on the first ImageNet[14] contest, where 1000 different classes were represented by 1 million images (RUSSAKOVSKY et al., 2015). For the purposes of our dataset, a subset of 6000 images was carefully selected to guaranty a balanced representation of ages, genders, and ethnicities. Having that set as a reference, 84 000 images were obtained through the application of the functions described in

[11] https://keras.io/
[12] https://www.tensorflow.org/
[13] https://www.flickr.com/
[14] https://image-net.org/challenges/LSVRC/

Chapter 3 for Gaussian noise, Gaussian blur, motion blur, brightness, and JPEG compression. The final dataset has a total of 90 000 images. All images were resized to 256*256 pixels to reduce the computational cost. The dataset was divided in the following manner: 85% for training, 10% for validation, and 5% for testing. Figure 10.2-1 in the Appendix section shows examples of the obtained images.

5 Results

5.1 Training Process

The training process for the distortion classification model consisted in experimenting with different configurations of the initial model described in Chapter 4, where one or more hyperparameters were changed in each iteration until the desired performance was achieved. Said performance was measured through the loss and the accuracy in both training and validation sets.

Another aspect of the training process was the reduction of the model's parameters once the metrics reached good values. This was mainly achieved by decreasing the number of filters in each convolution layer. Figure 5.1-1 shows the final model, its architecture is similar to the one described in chapter 4, but with fewer convolution filters and only two dropout layers. As a result, the final model has **531 887** trainable parameters, while the initial one had **2,183,631**.

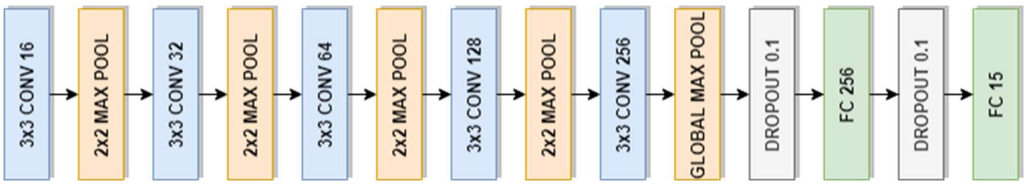


Figure 5.1-1. Final CNN model.

Both the training set and the validation set were used during training. Figure 5.1-2 shows the behaviour of the loss and the accuracy in both sets. The loss was approximately **0.05** in the training set and around **0.07** in the validation set. As for the accuracy, the model reached approximately **0.98** in both sets.

The training process was implemented in the Google Colaboratory[15] environment with a setup consisting of a dual-core Intel Xeon 79 CPU, an NVIDIA Tesla T4 GPU, 12.72 GB of RAM, and 76GB of available storage. Each epoch took approximately 5 minutes to complete.

[15] https://colab.research.google.com/notebooks/

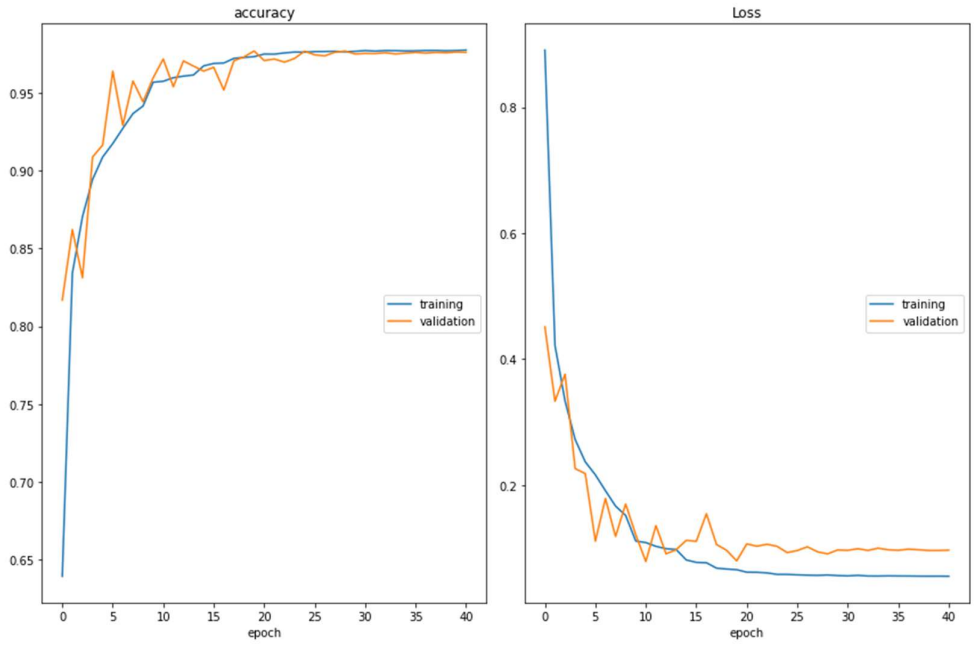


Figure 5.1-2. Loss and Accuracy during training.

5.1.1 Second Model

After obtaining satisfactory results with the model from Figure 5.1-1, a second model targeting images of 128*128 pixels was also trained. The intuition behind this is that image distortions are very susceptible to resizing (HA et al., 2019) (AHN; KANG; SOHN, 2018), and information might be lost and/or artificially created when transforming an image to a specific size. Given that CNN models need a fixed input size, resizing is a must, however, its effects can be mitigated if the scale of the transformation is small enough (AHN; KANG; SOHN, 2018). In that sense, a distortion classification model trained with 128*128 images was obtained. This second model will be used in images smaller than 256*256 but bigger than 64*64.

Figure 5.1-3 and Figure 5.1-4 show its architecture and the behaviour of the loss and the accuracy during training. The number of parameters is the same as the model trained with 256*256 images since the only alteration was the elimination of the max pool layer after the first convolution.

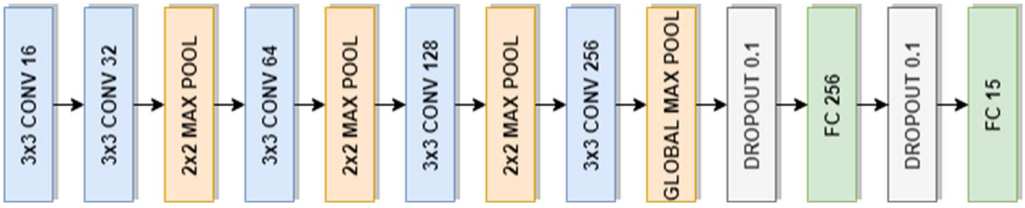


*Figure 5.1-3. Final model for 128*128 images.*

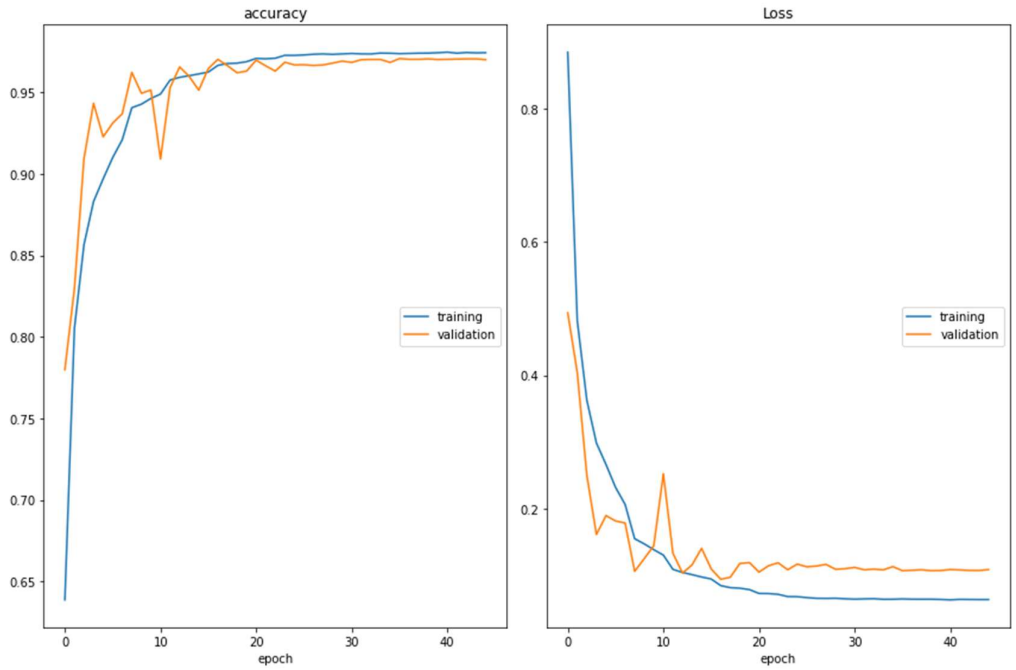


Figure 5.1-4. Loss and Accuracy during training.

As can be seen in Figure 5.1-4, both loss and accuracy have similar behaviours as the previous model. Loss in the training set was also around **0.05**, however, its value in the validation set was higher, approximately **0.09**. Accuracy in both sets reached values above **0.97**

The training process was carried in the same environment as the previous model, this time requiring around 3 minutes per epoch.

5.2 Tests Results

Once the models were fitted in the training and validation set, they were tested on the test set, which consists of 4500 images from the dataset created with the images from the FFHQ database.

5.2.1 256*256 model

The results of the model trained with 256*256 images are consistent with the ones obtained during training. The overall accuracy was approximately **0.9766**, which was expected considering the three subsets have the same distribution. The average prediction confidence was **0.9642**.

Table 5.2-1 shows the classification report generated with the scikit-learn[16] library. The metrics shown are Precision (9), Recall (10), and F1-Score (11). The results show good and

[16] https://scikit-learn.org/stable/

consistent behaviour across all classes, with a slight decrease in performance with images classified as HB1 and HB2, which correspond to images with different levels of excess brightness.

*Table 5.2-1.Classification Report for the 256*256 model*

Classes	Precision	Recall	F1-Score	Images
CL	1.00	0.97	0.98	300
GB1	1.00	1.00	1.00	300
GB2	1.00	1.00	1.00	300
GB3	1.00	1.00	1.00	300
GN1	1.00	1.00	1.00	300
GN2	1.00	1.00	1.00	300
GN3	1.00	1.00	1.00	300
HB1	0.88	0.91	0.89	300
HB2	0.91	0.88	0.89	300
JP1	1.00	0.93	0.97	300
JP2	0.94	1.00	0.97	300
LB1	0.97	0.99	0.98	300
LB2	0.99	1.00	1.00	300
MB1	1.00	0.97	0.99	300
MB2	0.97	1.00	0.99	300
Macro average	0.98	0.98	0.98	4500
Weighted Average	0.98	0.98	0.98	4500

$$Precision = \frac{true\ positives}{true\ positive + false\ positive} \quad (9)$$

$$Recall = \frac{true\ positives}{true\ positive + false\ negative} \quad (10)$$

$$F1 - Score = \frac{Precisio \times Recall}{Precisio + Recall} \quad (11)$$

Figure 5.2-1 shows the model's normalized confusion matrix obtained in the test set. Similar to the results obtained above, the model has almost perfect performance on all classes, with the exception of HB1 and HB2.

Another metric usually used when assessing classification models, is the ***top k accuracy***, which consists in measuring the accuracy of the model when the correct class is within its **k**

first predictions (SAWADA; KANEKO; SAGI, 2020). The top 1 accuracy is the same as overall accuracy, **0.9766**, the top 2 accuracy was **0.999**, and the top 3 accuracy was **1.0**.

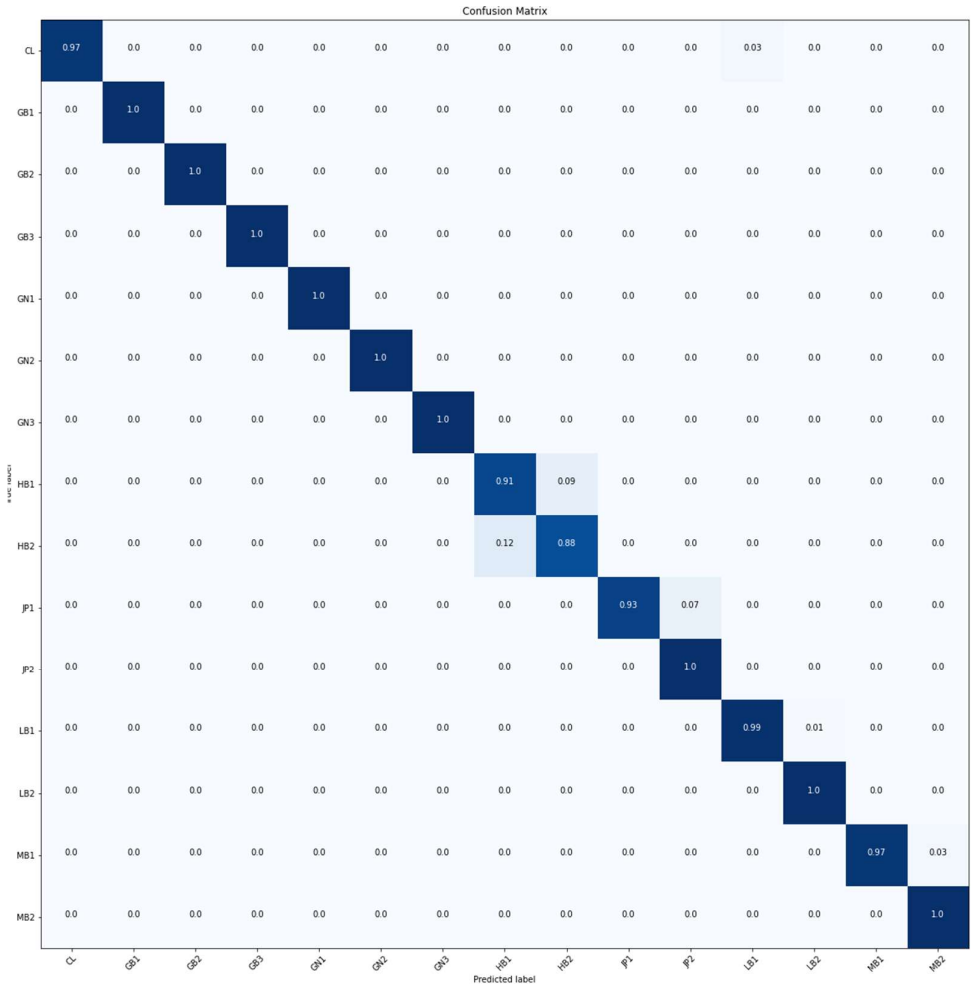


*Figure 5.2-1. Normalized Confusion Matrix for the 256*256 model in the test set.*

An error analysis was also conducted, Table 5.2-2 shows the details. There were 105 images misclassified from the test set, with images with excess brightness representing **61.9%** of all errors. On the other hand, almost half of the classes had zero errors, which constitutes excellent results.

*Table 5.2-2. Error analysis for the 256*256 model*

Classes	Errors	Percentage (Total Errors)	Percentage (Within Class)
CL	8	7.61%	2.67%
GB1	1	0.95%	0.33%
GB2	1	0.95%	0.33%
GB3	0	0.00%	0.00%
GN1	0	0.00%	0.00%
GN2	0	0.00%	0.00%
GN3	0	0.00%	0.00%
HB1	28	26.7%	9.33%
HB2	37	35.2%	13.3%
JP1	20	19.0%	6.67%
JP2	0	0.00%	0.00%
LB1	0	0.00%	0.00%
LB2	11	10.4%	3.67%
MB1	8	7.61%	2.67%
MB2	0	0.00%	0.00%

5.2.2 128*128 model

The results obtained with the second model were also good, although a slight decrease in performance was noticed compared with the model trained with 256*256 images. The overall accuracy was **0.9737**, and the average prediction confidence was **0.9685**.

Table 5.2-3 shows the results from the classification report generated using the scikit-learn library. The results are similar to the ones in Table 5.2-1 in that the model shows good performance in most classes, with the exception of HB1 and HB2. However, the decrease in performance in these classes is more noticeable with this model. A similar phenomenon is observed in the normalized confusion matrix in Figure 5.2-2.

The top 1 accuracy is the same as the overall accuracy, **0.9737**, the top 2 accuracy and top 3 accuracy had the same values as the ones obtained with the model trained with 256*256 images.

Table 5.2-4 shows the details of the error's distribution, this time a total of 154 images were misclassified, with **55.19%** corresponding to images with high brightness. Given that this model is similar to the one targeting 256*256 images, these results were expected.

*Table 5.2-3. Classification Report generated with the 128*128 model in the test set.*

Classes	Precision	Recall	F1-Score	Images
CL	1.00	0.98	0.99	300
GB1	1.00	1.00	1.00	300
GB2	1.00	1.00	1.00	300
GB3	1.00	1.00	1.00	300
GN1	1.00	1.00	1.00	300
GN2	1.00	1.00	1.00	300
GN3	1.00	1.00	1.00	300
HB1	0.95	0.76	0.84	300
HB2	0.80	0.96	0.87	300
JP1	1.00	0.96	0.98	300
JP2	0.96	1.00	0.98	300
LB1	0.98	0.99	0.99	300
LB2	0.99	1.00	0.99	300
MB1	1.00	0.97	0.99	300
MB2	0.97	1.00	0.99	300
Macro average	0.98	0.97	0.97	4500
Weighted Average	0.98	0.97	0.97	4500

*Table 5.2-4. Error Analysis for the 128*128 model.*

Classes	Errors	Percentage (Total Errors)	Percentage (Class)
CL	7	4.55%	2.33%
GB1	0	0.00%	0.00%
GB2	1	0.65%	0.33%
GB3	0	0.00%	0.00%
GN1	0	0.00%	0.00%
GN2	0	0.00%	0.00%
GN3	0	0.00%	0.00%
HB1	73	47.4%	24.33%
HB2	12	7.79%	4.00%
JP1	12	7.79%	4.00%
JP2	1	4.55%	2.33%
LB1	4	2.60%	1.33%
LB2	0	0.00%	0.00%
MB1	8	5.19%	2.67%
MB2	0	0.00%	0.00%

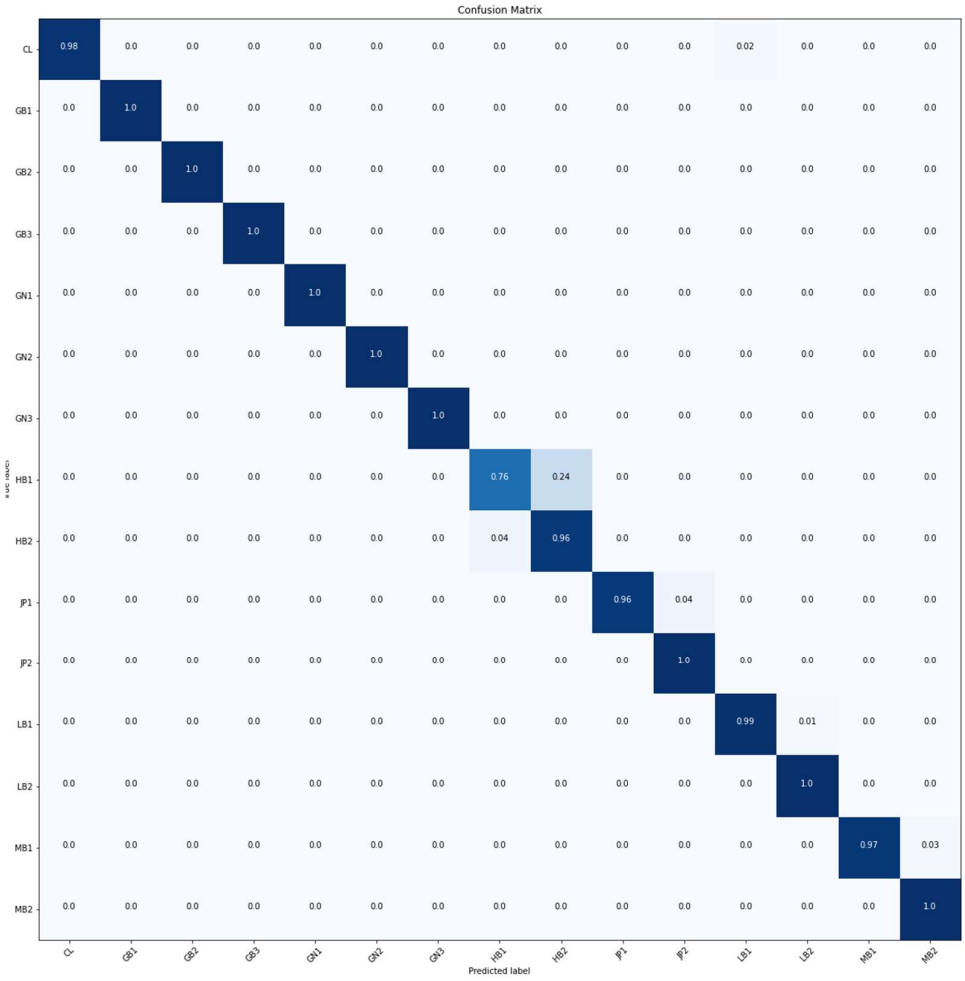


*Figure 5.2-2. Normalized Confusion Matrix for the 128*128 model in the test set.*

It is clear that the models' main weakness is recognizing excess brightness, which might be due to the way brightness alterations were simulated. To obtain low and high brightness images, the brightness factor is modified to increase or decrease the intensity of each pixel. The drawback of this approach is evident when dealing with face images with different skin tones, where it becomes hard to differentiate if a person has light or dark skin colour, or if the image is lacking or exceeding brightness.

5.3 Conclusions

This chapter exposes the results obtained during the training and testing of our distortion classification model. To ensure good results in images with different sizes, a second model targeting 128*128 images was also trained and tested.

The obtained results are very satisfactory, the overall accuracy reached by both models shows great performance in all classes, with the exception of high brightness. It is important to say that even though the decrease in performance in the HB1 and HB2 classes is noticeable when compared with the other classes, the results are still good. As mentioned above, the cause of this phenomenon might be associated with the way the images are generated. A different approach, one where the images are acquired in real-life scenarios with different lighting conditions, might result in better performance.

6 Validation

6.1 Validation dataset

The main limitation for the development of this work is the lack of publicly available datasets for quality assessment in face images. According to our research, the IDEAL-LIVE DFD is the only one, however, only the original images are available to download when using the provided link. The absence of previous works dedicated to distortion classification in face images, and the lack of pre-trained models for distortion classification in natural scene images, makes it impossible for us to conduct a comparison to validate our proposal. Given these circumstances, it was considered best to use the original images from the IDEAL-LIVE DFD and follow the details provided in (GUNASEKAR; GHOSH; BOVIK, 2014) to obtain the distorted versions. According to (GUNASEKAR; GHOSH; BOVIK, 2014), three MATLAB functions were used to generate the distorted images. For the Gaussian noise, the authors used the *imnoise()* function and configured the variance with the values shown in Table 6.1-1. For the Gaussian blur, the *imfilter()* function was used, and the standard deviation of the kernels was set up as it is shown in Table 6.1-1. Finally, the JPEG compression distortion was achieved using the *imwrite()* function, configuring the quality factor as illustrated in Table 6.1-1.

Table 6.1-1. Parameter configuration of the functions used in (GUNASEKAR; GHOSH; BOVIK, 2014).

Function	Parameter Configuration
imnoise()	$var = \{4.5 * 10^{-5}, 0.0001, 0.0003, 0.0009, 0.0025, 0.0065, 0.02, 0.05, 0.15, 0.36\}$
imfilter()	$std = \{0.4, 1.0, 2.3, 3.6, 4.5, 6.0, 7.4, 12.0, 20.0, 32.0\}$
imwrite()	$q = \{90, 60, 40, 25, 15, 10, 7.5, 5.0, 3.0, 2.0\}$

The validation process was focused on identifying the type and degree of the distortion present in the images. In that sense, the accuracy, precision, recall, and F1-Score were measured. In (GUNASEKAR; GHOSH; BOVIK, 2014), the authors did not assign levels to their degraded images, however, using Table 6.1-2, it is possible to label the distorted images according to our class definition and the parameters configuration provided by the authors. Table 6.1-2 shows said mapping.

Table 6.1-2. Mapping from parameter configuration to distortion magnitude

Labels	Parameters
Clean	Original images
Gaussian Noise 1	$var = \{4.5 * 10^{-5}, 0.0001, 0.0003, 0.0009, 0.0025, 0.0065, 0.02\}$
Gaussian Noise 2	$var = \{0.05\}$
Gaussian Noise 3	$var = \{\ 0.15, 0.36\}$
Gaussian Blur 1	$std = \{0.4, 1.0, 2.3\}$
Gaussian Blur 2	$std = \{3.6, 4.5, 6.0\}$
Gaussian Blur 3	$std = \{7.4, 12.0, 20.0, 32.0\}$
JPEG 1	$q = \{90, 60, 40\}$
JPEG 2	$q = \{25, 15, 10, 7.5, 5.0, 3.0, 2.0\}$

6.2 Validation Results

To create our validation set, only the images with one frontal face were selected as references. The final dataset contains 431 images belonging to 9 classes: CL, GB1, GB2, GB3, GN1, GN2, GN3, JP1, and JP2.

Figure 10.2-2 in the Appendix section, shows examples of the dataset. The difference in quality between the reference images from the FFHQ dataset (Figure 10.2-1), and the ones in the IDEAL-LIVE DFD is quite noticeable, so a drop in performance is expected.

6.2.1 256*256 model

The overall accuracy obtained with this model was **0.8515**, the top 2 and top 3 accuracies were also calculated, obtaining **0.8955** and **0.9396**, respectively. The mean prediction confidence was **0.9580**.

Table 6.2-1 shows the classification report for this model on the validation dataset. The results show GB2, GB3, GN1, GN2, GN3, and JP1 as the classes where the model exhibits the best performance, with perfect scores in most metrics. On the other hand, poor performance was achieved when classifying clean images, as well as images with gaussian blur and JPEG compression in their lower levels. A closer look at Table 6.2-1 indicates two main problems. The first, recognizing clean images (recall), with only 8% of them correctly classified. The second problem is related to the precision, as the model classified more images with GB1 and JP1 than the actual number of images having these distortions.

*Table 6.2-1. Classification Report for the 256*256 model in the validation set.*

Classes	Precision	Recall	F1-Score	Images
CL	0.80	0.08	0.15	49
GB1	0.59	1.00	0.74	48
GB2	1.00	1.00	1.00	48
GB3	1.00	1.00	1.00	48
GN1	1.00	0,93	0.96	48
GN2	1.00	1.00	1.00	49
GN3	1.00	1.00	1.00	48
JP1	0.57	0.69	0.62	48
JP2	0.96	0.98	0.97	48
Macro average	0.88	0.85	0.83	431
Weighted average	0.88	0.85	0.83	431

Table 6.2-2 shows the error analysis. A total of 64 images were misclassified, with 45 belonging to the clean class (reference images). A more in-depth analysis of these errors revealed that, of those 45 clean images, 20 were classified as having Gaussian blur level 1 (GB1), 24 with JPEG compression level 1 (JP1), and 1 with low brightness level 1 (LB1). From this, is easy to understand that the results of the classification report from Table 6.2-1. Most errors come from the misclassification of the reference images, mainly as images with GB1 and JP1 distortions, which are the classes with lower precision scores. These results are expected since the reference images have significantly lower quality than the ones the model was trained with.

*Table 6.2-2. Error analysis for the 256*256 model in the validation set.*

Classes	Errors	Percentage (from total errors)	Percentage (within class)
CL	45	70.0%	91.8%
GB1	0	0.00%	0.00%
GB2	0	0.00%	0.00%
GB3	0	0.00%	0.00%
GN1	3	4.69%	8.33%
GN2	0	0.00%	0.00%
GN3	0	0.00%	0.00%
JP1	15	23.4%	31.2%
JP2	1	1.56%	2.22%

6.2.2 128*128 model

The results obtained with the model targeted to 128*128 images were slightly better than the ones shown above. The overall accuracy was **0.8608**, the top 2 and top 3 accuracies were **0.9304** and **0.9629**, respectively, and the mean prediction confidence was **0.9606**.

Table 6.2-3 shows the classification report for this model, its behavior is similar to the one in Table 6.2-1. The best performing classes were GB2, GB3, GN1, GN2, GN3, and JP2, and the worst were CL, GB1, and JP1.

*Table 6.2-3. Classification Report for the 128*128 model in the validation set.*

Classes	Precision	Recall	F1-Score	Images
CL	0.80	0.24	0.37	49
GB1	0.59	1.00	0.74	48
GB2	1.00	1.00	1.00	48
GB3	1.00	1.00	1.00	48
GN1	1.00	0.92	0.96	48
GN2	1.00	1.00	1.00	49
GN3	1.00	1.00	1.00	48
JP1	0.65	0.69	0.67	48
JP2	0.95	0.91	0.93	48
Macro average	0.89	0.86	0.85	431
Weighted average	0.89	0.86	0.87	431

Table 6.2-4 shows the results obtained from the error analysis. A total of 60 images were misclassified, from which **37** belonged to the clean class (reference images). Of those 37, 18 were classified as GB1, 17 as JP1, 1 as JP2, and 1 as LB1.

*Table 6.2-4. Error analysis of the 128*128 model in the validation set.*

Classes	Errors	Percentage (from total errors)	Percentage (within class)
CL	37	61.7%	75.5%
GB1	0	0.00%	0.00%
GB2	0	0.00%	0.00%
GB3	0	0.00%	0.00%
GN1	4	6.67%	8.33%
GN2	0	0.00%	0.00%
GN3	0	0.00%	0.00%
JP1	15	25.0%	31.2%
JP2	4	6.67%	8.89%

6.3 Conclusions

The focus of this chapter was to illustrate the validation process of our distortion classification models. To that end, a set of face images with Gaussian blur, Gaussian noise, and

JPEG compression was generated from the images provided by the IDEAL_LIVE_DFD. The set obtained allowed us to assess our models' performance in 9 out of the 15 classes our models were trained to output.

Both models exhibited the same behaviour in that they struggled to correctly classify the reference images. Between the two, the model targeting 128*128 images showed better performance, according to its overall accuracy, precision, recall, and F1-score results. As was stated before, given the differences in quality between the images used for training and the ones provided by the IDEAL_LIVE DFD as reference images, a drop in performance was expected. As we look closer into Table 6.2-1 and Table 6.2-3, we can see that the models can correctly recognize most distorted images, showing high recall and precision scores in most classes. Similarly, the error analysis in Table 6.2-2 and Table 6.2-4, show zero errors in most classes.

7 Distortion classification as previous stage for face processing: use case.

As was stated in the introductory section, the information of the distortion affecting a given face image, as well as the level of the degradation, might be useful in several scenarios. Face processing algorithms, with known quality restrictions, can use this information to filter the images that comply with the established requirements from the ones that don't.

In that sense, we have decided to conduct an experiment using our distortion classification model in combination with a face processing algorithm. The goal behind this experiment was to test its suitability in this type of scenario. The following sections describe the experiment setup, as well as our results.

7.1 Experiment Setup

The DEX algorithm, focused in age estimation, was used as our face processing algorithm. The experiment was carried on the test set of the APPA-REAL dataset, which contains 1978 images of different sizes and qualities. The images were not artificially distorted for this experiment, they were just resized to comply with the input restrictions of both models.

From the results illustrated in Chapter 3, we know that the DEX algorithm is heavily affected by noise and blur in their gaussian distributions, even at their lowest levels. We also know that the impact of brightness and JPEG compression is less noticeable unless the images are heavily degraded. Additionally, we ran our distortion classification models throughout the test set to see which distortions were present in the test set, the results are as follows: GB1, GB2, MB1, LB1, LB2, HB, JP1, and JP2. Considering that, we defined 8 scenarios in which we filtered out face images according to the output of our distortion classification models and the requirements of each scenario. Table 7.1-1 shows a description of the defined scenarios.

Our experiment consisted in two main stages: a distortion classification stage, and an age estimation stage. The first part of the experiment consisted in classifying the images according to the 15 classes our models were trained to output. Having that information, we filtered the images according to the constraints of the previously defined scenarios and then performed the age estimation with the DEX algorithm.

To measure the performance of the DEX algorithm we used the Mean Absolute Error (MAE) in years, which was also the metric used in Chapter 3 to assess performance (See equation 4). Given that the APPA-REAL dataset allows us to evaluate the performance of age estimation models in both the real age and the apparent age, we assess the DEX algorithm by calculating the MAE value for the real and the apparent age estimations. To compare the DEX performance in the different scenarios, we also ran the DEX algorithm throughout the entire test set and took the obtained MAE values as our references for comparison.

Table 7.1-1. Scenarios description

Scenario	Description
No GB	Test set without images classified as GB1 or GB2
No MB1	Test set without images classified as MB1
No LB2	Test set without images classified as LB2
No JP2	Test set without images classified as JP2
No GB & MB	Test set without images classified as GB1, GB2, MB1
No GB &MB & LB2	Test set without images classified as GB1, GB2, MB1, LB2
No GB & MB & LB2 & JP2	Test set without images classified as GB1, GB2, MB1, LB2, JP2
Only CL	Only images classified as clean (CL)

7.2 Results

As stated in the previous section, the metric used to measure performance was the MAE in years, therefore a performance improvement means a decrease in that value.

The results obtained with our experiment are shown in Figure 7.2-1. The first two columns correspond to the MAE values obtained when using all the images from the test set. To their right, the graph shows the MAE values obtained in every scenario. As can be seen, the MAE values decreased whenever distorted face images were filtered out, when compared with the results obtained with the entire test set.

For the cases of images classified as MB1, LB2, and JP2, their exclusion from the set did not significantly improve performance. However, when excluding images classified as GB1 or GB2 (No GB scenario), a notable decrease in the MAE values is observed for both the real, and the apparent age estimation errors. The results obtained for that scenario represent a 14.76% improvement for the apparent age estimation and an 11.18% improvement for the case of real age estimation. Similarly, when filtering out images according to more than one distortion, as is shown in the last four scenarios, the results improved considerably. The best performance was achieved in the last scenario, where choosing only clean images caused a 22% decrease in the MAE value corresponding to the apparent age estimation and a 20% decrease for the case of the real age estimation MAE.

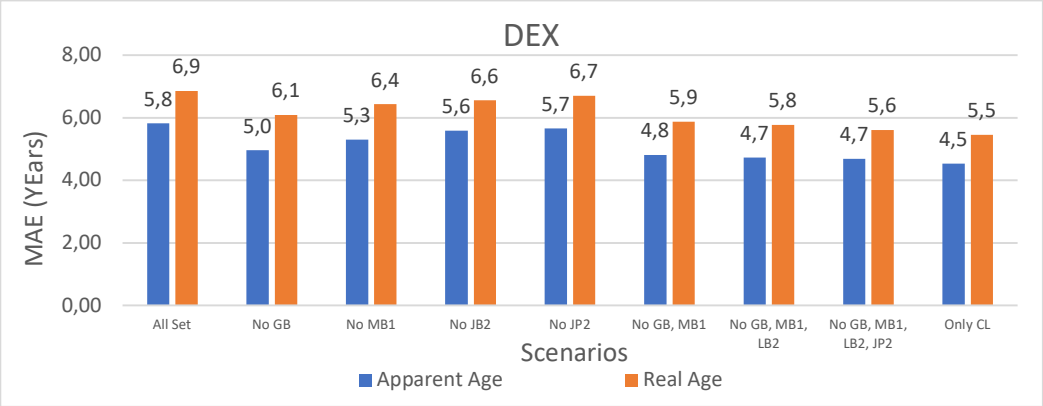


Figure 7.2-1. Experiment Results.

7.3 Conclusions

This chapter focuses on evaluating the suitability of our models to function as a previous stage for face processing algorithms. To that end, we used the DEX algorithm in combination with our models and defined 8 different scenarios to filter out face images according to specific distortions. Using the MAE values as a measure of performance, we evaluated the DEX algorithm across all the scenarios and compared the results with the MAE values obtained when running the DEX algorithm throughout the entire test set.

The results of the experiment were satisfactory, as it shows how the performance of the DEX algorithm improved whenever distorted images were filtered out. As was outlined in the previous section, not all distortions have the same impact, therefore their exclusion did not improve the DEX's performance in a similar way. However, for the case of Gaussian blur, as well as for the scenarios that filtered out face images according to more than one distortion, the performance improvement was noticeable. This constitutes a good endorsement of the benefits of using our distortion classification models as part of the face processing pipeline.

Although these results are promising, more tests are needed to ensure the suitability of this approach across the different tasks within the face processing domain.

8 General conclusions

8.1 Summary

The purpose of this work was to provide a different approach to face image quality assessment through distortion identification. Our main objective was to develop a model capable of classifying face images according to the distortion affecting them, in order to output qualitative information about their quality.

To achieve that goal, we first conducted a study to comprehend the impact of specific distortions on the performance of face processing algorithms. Based on the study's findings, described in Chapter 3, we defined the output of our distortion classification model to identify undistorted images, as well as images degraded by Gaussian blur, Gaussian noise, motion blur, low brightness, high brightness, and JPEG compression, at different degrees. Having that, we created a dataset for training and testing, using a subset of high-quality face images from the FFHQ dataset as references.

As stated in Chapter 5, resizing is a necessary part of image pre-processing for deep learning models. However, its effect could potentially diminish our model's ability to correctly identify distortions. In that sense, two models of similar architecture but different input configurations were trained and tested, in order to choose the best one for a given image size.

The results obtained during testing were very satisfactory, as our models achieved over 97% of accuracy in the test set. The classification report generated as part of our assessment also showed good performance in metrics such as Precision, Recall, and F1-Score.

We also tested our models' performance in the IDEAL_LIVE_DFD, where the results were also satisfactory. A drop in performance was observed when compared with the results obtained in our test set, however, it is attributable to the notable difference in quality between the reference (clean) images of the IDEAL-LIVE_DFD, and the ones in the FFHQ dataset.

Given that these distortion classification models constitute, to the best of our knowledge, a new approach for face image quality assessment, we were unable to compare their performance against other models. Instead, we conducted an experiment to show the advantages of using our models as a previous stage for face processing algorithms. The obtained results validate our initial premise about the models' usability since a notable increase in performance was achieved when processing only the images classified as clean, or having small degradations. Although these results are very promising, further experiments should be carried out to assess its suitability with other face processing algorithms.

8.2 Limitations

With this work, we have tackled some of the limitations outlined in Chapter 2 about the state of the art in face image quality assessment. However, our models also have some limitations that need to be handled in future works.

The main limitations of our work are mostly related to the images used to train our models. Firstly, we only considered images affected by one distortion. In real-world environments, images can have several distortions, and in those cases, our models might not be suited to process them. We also assumed that the distortion is present in the entire image, which is not always the case in real-world scenarios. Lastly, the fact that our training dataset is composed of artificially distorted images is also a limitation, since it is not the best representation of real images.

The other main limitation is given by the novelty of our approach in the face processing domain. Comparisons with similar models, as well as the evaluation of our models in publicly available datasets, could greatly contribute to improving our results. Since at the time of our research, we did not find other models focused on distortion classification in face images, or publicly available datasets for face image quality, this task will be performed as part of our future work.

8.3 Conclusions

The main goal of this work was to develop a CNN-based method to classify face images according to the type and degree of the distortion present in them. We achieved that by training and testing two CNN models able to identify different degrees of the following distortions: Gaussian noise, Gaussian blur, motion blur, low brightness, high brightness, and JPEG compression; as well as undistorted images.

We conducted several experiments to evaluate our models' performance, as well as their suitability as a previous stage for face processing algorithms. The results were satisfactory, as the models showed good performance both in the test set and in the IDEAL-LIVE-DFD. As for the results obtained in Chapter 7, the performance improvement showed by the DEX algorithm constitutes a good endorsement of the potential of our approach to improve face processing.

Our work has several contributions. First, our approach to FIQA tackled several of the current limitations outlined in Chapter 2. The models developed in this work provide us with information regarding the quality of face images without carrying in the biases of face recognition algorithms, this makes them suitable for other tasks inside the face processing domain.

Secondly, the findings presented in Chapter 2 about the impact of the aforementioned distortions in face processing performance constitute valuable information for the development of more robust face processing algorithms.

Lastly, as a result of our work, we also created a dataset for FIQA purposes with images containing faces of different ethnicities, genders, and ages, and a variety of types and degrees of distortions.

8.4 Future work

Our future work will be focused on tackling the limitations outlined in Section 8.2. The main goal will be to evaluate our models' performance in different sets of face images, where images can be degraded by more than one distortion. According to the results obtained during this assessment, modifications and improvements will be made to our models.

Another future work will be to develop a distortion detection method that besides identifying the type and degree of the distortion(s), can also find its (their) location in the image.

9 Bibliography

ABBAS, E. I.; SAFI, M. E.; RIJAB, K. S. **Face recognition rate using different classifier methods based on PCA**. International Conference on Current Research in Computer Science and Information Technology, ICCIT 2017. **Anais**...Institute of Electrical and Electronics Engineers Inc., 30 Jun. 2017

ABDEL-MOTTALEB, M.; MAHOOR, M. H. Algorithms for assessing the quality of facial images. **IEEE Computational Intelligence Magazine**, v. 2, n. 2, p. 10–17, 2007.

AHN, N.; KANG, B.; SOHN, K.-A. Image Distortion Detection using Convolutional Neural Network. **Proceedings - 4th Asian Conference on Pattern Recognition, ACPR 2017**, p. 226–231, 28 May 2018.

ALAQL, O.; GHAZINOUR, K.; CHANG, C. **Classification of Image Distortions Based on Features Evaluation**. 2016 IEEE International Symposium on Multimedia Classification. **Anais**...Institute of Electrical and Electronics Engineers (IEEE), 20 Jan. 2016

ALOM, M. Z. et al. A state-of-the-art survey on deep learning theory and architectures. **Electronics (Switzerland)**, v. 8, n. 3, p. 1–67, 2019.

AWAD, M. et al. Support Vector Machines for Classification. In: **Efficient Learning Machines**. [s.l.] Apress, 2015. p. 39–66.

BARTLETT, M. S.; MOVELLAN, J. R.; SEJNOWSKI, T. J. Face recognition by independent component analysis. **IEEE Transactions on Neural Networks**, v. 13, n. 6, p. 1450–1464, Nov. 2002.

BELHUMEUR, P. N.; HESPANHA, J. P.; KRIEGMAN, D. J. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 19, n. 7, p. 711–720, 1997.

BEST-ROWDEN, L.; JAIN, A. K. Learning face image quality from human assessments. **IEEE Transactions on Information Forensics and Security**, v. 13, n. 12, p. 3064–3077, 1 Dec. 2018.

BHATTACHARYA, S.; ROUTRAY, A. **Score based Face Quality Assessment (FQA)**. 2017 14th IEEE India Council International Conference, INDICON 2017. **Anais**...2018

BLAU, Y.; MICHAELI, T. **The Perception-Distortion Tradeoff**. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. **Anais**...IEEE Computer Society, 14 Dec. 2018

BOURLAI, T. et al. Design and evaluation of photometric image quality measures for effective face recognition. **IET Biometrics**, v. 3, n. 4, p. 314–324, 2014.

BUCZKOWSKI, M.; STASINSKI, R. **Convolutional Neural Network-Based Image Distortion Classification**. International Conference on Systems, Signals, and Image Processing (IWSSIP). **Anais**...Osijek, Croatia: IEEE Computer Society, 1 Jun. 2019

BYUN, Y. et al. Low-Complexity Dynamic Channel Scaling of Noise-Resilient CNN for Intelligent Edge Devices. **Proceedings of the 2019 Design, Automation and Test in Europe Conference and Exhibition, DATE 2019**, n. 2, p. 114–119, 2019.

CAO, J.; LI, Y.; ZHANG, Z. **Celeb-500K: A large training dataset for face**

recognition. Proceedings - International Conference on Image Processing, ICIP. **Anais**...IEEE Computer Society, 29 Aug. 2018

CAO, Q. et al. **VGGFace2: A dataset for recognising faces across pose and age**. 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). **Anais**...2018

CHAPELLE, O.; SCHÖLKOPF, B.; ZIEN, A. **Introduction to Semi-Supervised LearningSemi-Supervised Learning**, 2006.

CHATFIELD, K. et al. **Return of the Devil in the Details: Delving Deep into Convolutional Nets**. MVC 2014 - Proceedings the British Machine Vision Conference 2014. **Anais**...2014

CHEN, B.-C.; CHEN, C.-S.; HSU, W. H. **Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval**. European Conference on Computer Vision-ECCV 2014. **Anais**...2014

CHEN, J. et al. Face image quality assessment based on learning to rank. **IEEE Signal Processing Letters**, v. 22, n. 1, p. 90–94, 2015.

CHENG, B. et al. Robust emotion recognition from low quality and low bit rate video: A deep learning approach. **2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017**, v. 2017, p. 65–70, 2017.

CHETOUANI, A.; BEGHDADI, A.; DERICHE, M. **Image distortion analysis and classification scheme using a neural approach**. 2010 2nd European Workshop on Visual Information Processing, EUVIP2010. **Anais**...2010

CLAPES, A. et al. From apparent to real age: Gender, age, ethnic, makeup, and expression bias analysis in real age estimation. **IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops**, v. 2018–June, p. 2436–2445, 2018.

CLARK, M.; BOVIK, A. C. Experiments in segmenting texton patterns using localized spatial filters. **Pattern Recognition**, v. 22, n. 6, p. 707–717, 1 Jan. 1989.

DALAL, N.; TRIGGS, B. **Histograms of oriented gradients for human detection**. Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005. **Anais**...2005

DIAMOND, S. et al. **Dirty Pixels: Optimizing Image Classification Architectures for Raw Sensor DataarXiv: Computer Vision and Pattern Recognition**, 23 Jan. 2017.

DODGE, S. F.; KARAM, L. J. Quality Robust Mixtures of Deep Neural Networks. **IEEE Transactions on Image Processing**, v. 27, n. 11, p. 5553–5562, 1 Nov. 2018.

DODGE, S.; KARAM, L. **Understanding how image quality affects deep neural networks**. 2016 8th International Conference on Quality of Multimedia Experience, QoMEX 2016. **Anais**...Institute of Electrical and Electronics Engineers Inc., 23 Jun. 2016

DUTTA, A.; VELDHUIS, R.; SPREEUWERS, L. **The Impact of Image Quality on the Performance of Face Recognition**. 33rd Symposium on Information Theory in the Benelux and the 2nd Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux. **Anais**...2012

ESCALERA, S. et al. **ChaLearn Looking at People 2015: Apparent Age and**

Cultural Event Recognition datasets and results. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). **Anais**...2015

FUKUI, K. **Subspace Methods**, 2014.

GAO, X. et al. **Standardization of face image sample quality**. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **Anais**...2007

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [s.l.] MIT Press, 2016.

GRIFFIN, G.; HOLUB, A.; PERONA, P. **Caltech-256 Object Category DatasetCalTech Report**, 2007.

GROSS, R. et al. **Multi-PIE**. Proc Int Conf Autom Face Gesture Recognit. **Anais**...2010

GROTHER, P.; NGAN, M. **The IJB-A Face Identification Challenge Performance Report**, 2017.

GUNASEKAR, S.; GHOSH, J.; BOVIK, A. C. Face Detection on Distorted Images Augmented by Perceptual Quality-Aware Features. **IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY**, v. 9, n. 12, 2014.

GUNES, H.; PICCARDI, M. Bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. **Proceedings - International Conference on Pattern Recognition**, v. 1, p. 1148–1153, 2006.

GUO, Y. et al. **MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition**. 14th European Conference on Computer Vision – ECCV 2016. **Anais**...2016

HA, M. et al. Selective Deep Convolutional Neural Network for Low Cost Distorted Image Classification. **IEEE Access**, v. 7, p. 133030–133042, 2019.

HERNANDEZ-ORTEGA UAM, J. et al. **FaceQnet: Quality Assessment for Face Recognition based on Deep Learning**. ICB 2019. **Anais**...2019

HSU, R. L. V.; SHAH, J.; MARTIN, B. **Quality assessment of facial images**. Biometrics Symposium, BCC 2006. **Anais**...2006

HUANG, G. B. et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. **Tech Report**, 2007.

HUSSAIN, M.; BIRD, J. J.; FARIA, D. R. **A Study on CNN Transfer Learning for Image Classification**. Advances in Computational Intelligence Systems. UKCI 2018. Advances in Intelligent Systems and Computing. **Anais**...2018

JAIN, A. et al. **Gist: Efficient Data Encoding for Deep Neural Network Training**. 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA). **Anais**...2018

JATURAWAT, P.; PHANKOKKRUAD, M. An evaluation of face recognition algorithms and accuracy based on video in unconstrained factors. **Proceedings - 6th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2016**, n. November, p. 240–244, 2017.

JIA, Y. et al. **Caffe: Convolutional Architecture for Fast Feature Embedding**.

Proceedings of the 22nd ACM International Conference on Multimedia. **Anais**...2014

K. RICANEK JR. AND T. TESAFAYE. MORPH: A longitudinal image Age-progression, of normal adult. **Proc. 7th Int. Conf. Autom. Face Gesture Recognit**, p. 0–4, 2006.

KADIR, K. et al. **A comparative study between LBP and Haar-like features for Face Detection using OpenCV**. 2014 4th International Conference on Engineering Technology and Technopreneuship, ICE2T 2014. **Anais**...Institute of Electrical and Electronics Engineers Inc., 9 Jan. 2015

KANG, L. et al. **Convolutional Neural Networks for No-Reference Image Quality Assessment**. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). **Anais**...2014

KANG, L. et al. **Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks**. Proceedings - International Conference on Image Processing, ICIP. **Anais**...IEEE Computer Society, 9 Dec. 2015

KARRAS, T. (NVIDIA); LAINE, S. (NVIDIA); AILA, T. (NVIDIA). **A Style-Based Generator Architecture for Generative Adversarial Networks**. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). **Anais**...2019

KAUR, S.; JINDAL, S. A Survey on Machine Learning Algorithms. **International Journal of Innovative Research in Advanced Engineering (IJIRAE) Issue**, v. 3, n. 11, p. 6–14, 2016.

KHASTAVANEH, H.; EBRAHIMPOUR-KOMLEH, H.; JOUDAKI, M. **Face image quality assessment based on photometric features and classification techniques**. 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation, KBEI 2017. **Anais**...Institute of Electrical and Electronics Engineers Inc., 23 Mar. 2018

KHORRAMI, P. et al. **HOW DEEP NEURAL NETWORKS CAN IMPROVE EMOTION RECOGNITION ON VIDEO DATA**. 2016 IEEE International Conference on Image Processing, ICIP 2016 - Proceedings. **Anais**...2016

KHRYASHCHEV, V. et al. **Development of Face Image Quality Assessment Algorithms for Biometric Identification Tasks**. CSIT2017. **Anais**...2017

KIM, J. et al. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. **IEEE Signal Processing Magazine**, v. 34, n. 6, p. 130–141, 1 Nov. 2017.

KÖSTINGER, M. et al. **Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization**. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). **Anais**...2011

KOWALSKI, M.; NARUNIEC, J.; TRZCINSKI, T. **Deep Alignment Network: A Convolutional Neural Network for Robust Face Alignment**. 2017 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. **Anais**...2017

KRISHNA MOORTHY, A.; CONRAD BOVIK, A. Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. **IEEE Transactions on Image Processing**, v. 20, n. 12, p. 3350–3364, 2011.

KRIZHEVSKY, A. **Learning Multiple Layers of Features from Tiny**

ImagesTechnical Report, 2009.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. **Commun. ACM**, v. 60, n. 6, p. 84–90, 2017.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015.

LI, P. et al. Face Recognition in Low Quality Images: A Survey. **ACM Comput. Surv**, v. 1, n. April-, 2019.

LIAO, P. et al. **Facial image quality assessment based on support vector machines**. Proceedings - 2012 International Conference on Biomedical Engineering and Biotechnology, iCBEB 2012. **Anais**...2012

LIJUN, Z. et al. **Multi-branch Face Quality Assessment for Face Recognition**. International Conference on Communication Technology Proceedings, ICCT. **Anais**...Institute of Electrical and Electronics Engineers Inc., 1 Oct. 2019

LIU, L. et al. No-reference image quality assessment based on spatial and spectral entropies. **Signal Processing: Image Communication**, v. 29, n. 8, p. 856–863, 1 Sep. 2014.

LIU, X.; PEDERSEN, M.; HARDEBERG, J. Y. **CID:IQ – A New Image Quality Database**. International Conference on Image and Signal Processing. **Anais**...Springer, Cham, 2014

LUCEY, P. et al. **The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression**. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. **Anais**...2010

MALTONI, D. et al. **BioLab-ICAO: A new benchmark to evaluate applications assessing face image compliance to ISO/IEC 19794-5 standard**. Proceedings - International Conference on Image Processing, ICIP. **Anais**...IEEE Computer Society, 2009

MEHMOOD, R.; SELWAL, A. A Comprehensive Review on Face Recognition Methods and Factors Affecting Facial Recognition Accuracy. **Lecture Notes in Electrical Engineering**, v. 597, n. January, p. 455–467, 2020.

MIN, R.; KOSE, N.; DUGELAY, J.-L. KinectFaceDB: A Kinect Database for Face Recognition. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, v. 44, n. 11, p. 1534–1548, 2014.

MITTAL, A.; MOORTHY, A. K.; BOVIK, A. C. **Blind/Referenceless Image Spatial Quality Evaluator**. 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR). **Anais**...211AD

MOORTHY, A. K.; BOVIK, A. C. **A two-stage framework for blind image quality assessment**. Proceedings - International Conference on Image Processing, ICIP. **Anais**...2010

NETZER, Y. et al. **Reading Digits in Natural Images with Unsupervised Feature Learning**. NIPS Workshop on Deep Learning and Unsupervised Feature Learning. **Anais**...2011

NGUYEN, D. et al. Human Age Estimation Method Robust to Camera Sensor and/or Face Movement. **Sensors**, v. 15, n. 9, p. 21898–21930, 31 Aug. 2015.

NGUYEN, D.; CHO, S.; PARK, K. Age Estimation-Based Soft Biometrics Considering

Optical Blurring Based on Symmetrical Sub-Blocks for MLBP. **Symmetry**, v. 7, n. 4, p. 1882–1913, 19 Oct. 2015.

OKARMA, K. Current trends and advances in image quality assessment. **Elektronika ir Elektrotechnika**, v. 25, n. 3, p. 77–84, 2019.

PANIS, G. et al. Overview of research on facial ageing using the FG-NET ageing database. **IET Biometrics**, v. 5, n. 2, p. 37–46, 2016.

PHILLIPS, P. J. et al. **The FERET Evaluation Methodology for Face-Recognition Algorithms**, 1998.

PHILLIPS, P. J. et al. **Overview of the Face Recognition Grand Challenge**. Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005. **Anais**...2005

PONOMARENKO, N. et al. **ON BETWEEN-COEFFICIENT CONTRAST MASKING OF DCT BASIS FUNCTIONS**. Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, VPQM 2007. **Anais**...2007

PONOMARENKO, N. et al. TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics. **Advances of Modern Radioelectronics**, p. 30–45, 2009.

RAGHAVENDRA, R. et al. **Automatic face quality assessment from video using gray level co-occurrence matrix: An empirical study on automatic border control system**. Proceedings - International Conference on Pattern Recognition. **Anais**...Institute of Electrical and Electronics Engineers Inc., 4 Dec. 2014

RASMUSSEN, C. E.; WILLIAMS, C. K. I. **Gaussian Processes for Machine Learning**. [s.l.] MIT Press, 2006.

RINGEVAL, F. et al. **AVEC 2015-The 5th International Audio/Visual Emotion Challenge and Workshop**. Proceedings of the 23rd ACM International Conference on Multimedia. **Anais**...2015

ROTHE, R.; TIMOFTE, R.; VAN GOOL, L. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. **International Journal of Computer Vision**, v. 126, n. 2–4, p. 144–157, 2018.

RUSSAKOVSKY, O. et al. ImageNet Large Scale Visual Recognition Challenge. **International Journal of Computer Vision**, v. 115, n. 3, p. 211–252, 1 Dec. 2015.

SAAD, M. A.; BOVIK, A. C.; CHARRIER, C. Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain. **IEEE TRANSACTIONS ON IMAGE PROCESSING**, v. 21, n. 8, p. 3339, 2012.

SAGONAS, C. et al. **300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge**. 2013 IEEE International Conference on Computer Vision Workshops. **Anais**...2013

SAKANO, H.; MUKAWA, N. Kernel mutual subspace method for robust facial image recognition. **International Conference on Knowledge-Based Intelligent Electronic Systems, Proceedings, KES**, v. 1, p. 245–248, 2000.

SANDLER, M. et al. **MobileNetV2: Inverted Residuals and Linear Bottlenecks**. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. **Anais**...2018

SAWADA, A.; KANEKO, E.; SAGI, K. **Trade-offs in Top-k Classification Accuracies on Losses for Deep Learning**. [s.l: s.n.].

SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural Networks**, v. 61, p. 85–117, 2015.

SCHROFF, F.; PHILBIN, J. **FaceNet: A Unified Embedding for Face Recognition and Clustering**. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). **Anais**...Boston, MA: 2015

SHARMA, R.; ARORA, R. Face Recognition Using LTP Algorithm. **International Journal of Science and Research (IJSR) ISSN**, v. 4, 2013.

SHEIKH, H. R. et al. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. **IEEE TRANS. IMAGE PROCESSING**, v. 15, n. 11, p. 3440–3451, 2006.

SHEIKH, H. R.; BOVIK, A. C. Image Information and Visual Quality. **IEEE TRANSACTIONS ON IMAGE PROCESSING**, v. 15, n. 2, p. 430–444, 2006.

SIMONYAN, K.; ZISSERMAN, A. **VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION**. International Conference on Learning Representations. **Anais**...2015

SZEGEDY, C. et al. **Going Deeper with Convolutions**. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). **Anais**...2015a

SZEGEDY, C. et al. **Going deeper with convolutions**. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). **Anais**...Boston, MA: 2015b

TCHENDJOU, G. T. et al. Evaluation of machine learning algorithms for image quality assessment. **2016 IEEE 22nd International Symposium on On-Line Testing and Robust System Design, IOLTS 2016**, p. 193–194, 2016.

TERHÖRST, P. et al. Face Quality Estimation and Its Correlation to Demographic and Non-Demographic Bias in Face Recognition. **arXiv**, 2020.

TIAN, Y.; CHEN, S. Understanding Effects of Image Resolution for Facial Expression Analysis. **Journal of Computer Vision and Image Processing**, n. NWPJ-201201-20, 2012.

TURK, M. A.; PENTLAND, A. P. **Face Recognition Using Eigenfaces**. Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. **Anais**...1991

VIGNESH, S.; PRIYA, K. V. S. N. L. M.; CHANNAPPAYYA, S. S. Face image quality assessment for face selection in surveillance video using convolutional neural networks. **2015 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2015**, n. November, p. 577–581, 2016.

VIOLA, P.; JONES, M. **Rapid Object Detection using a Boosted Cascade of Simple Features**. IEEE Conference on Computer Vision and Pattern Recognition. **Anais**...2001

VISIONICS, C. **FaceIt ® An award-winning facial recognition software engine**, 2004.

WOLF, L.; HASSNER, T.; MAOZ, I. **Face Recognition in Unconstrained Videos with Matched Background Similarity**. Computer Vision and Pattern Recognition (CVPR).

Anais...2011

WONG, Y. et al. **Patch-based Probabilistic Image Quality Assessment for Face Selection and Improved Video-based Face Recognition**. CVPR 2011 WORKSHOPS. **Anais**...2011

XIONG, X.; DE LA TORRE, F. **Supervised Descent Method and its Applications to Face Alignment**. 2013 IEEE Conference on Computer Vision and Pattern Recognition. **Anais**...2013

YI, D. et al. Learning Face Representation from Scratch. **ArXiv**, 2014.

YOGITA, H.; PATIL, H. Y. **A Survey on Image Quality Assessment Techniques, Challenges and Databases**. IJCA Proceedings on National Conference on Advances in Computing NCAC 2015. **Anais**...2015

ZARIC, A. et al. VCL@FER Image Quality Assessment Database. **Automatika Journal for Control, Measurement, Electronics, Computing and Communications**, p. 344–354, 2012.

ZEILER, M. D.; FERGUS, R. **Visualizing and Understanding Convolutional Networks**. 13th European Conference on Computer Vision – ECCV 2014. **Anais**...2014

ZENG, J. et al. **Face Image Quality Quantitative Assessment for Forensic Identification of Human Images**. Proceedings of the 2018 IEEE International Conference on Progress in Informatics and Computing, PIC 2018. **Anais**...Institute of Electrical and Electronics Engineers Inc., 2 Jul. 2018

ZHAO, Y. **Theories and applications of LBP: A survey**. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **Anais**...Springer, Berlin, Heidelberg, 2011

ZHOU, Y.; SONG, S.; CHEUNG, N.-M. **On classification of distorted images with deep convolutional neural networks**. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. **Anais**...2017

ZHUANG, N. et al. Recognition oriented facial image quality assessment via deep convolutional neural network. **Neurocomputing**, v. 358, p. 109–118, 17 Sep. 2019.

10 Appendix

10.1 Tables

Table 10.1-1. Metrics results from the noise experiment

NOISE VARIANCE	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
0,00	**0,9965**	**0,98567**	**6,46788**	**7,6086**	**0,052**	**0,0518**
0,01	0,989	0,926	11,835	12,659	0,075	0,237
0,02	0,977	0,853	13,219	14,062	0,108	0,415
0,03	0,965	0,751	14,087	14,940	0,134	0,556
0,04	0,948	0,618	14,511	15,361	0,164	0,659
0,05	0,924	0,393	14,757	15,551	0,197	0,763
0,06	0,897	0,330	14,958	15,778	0,233	0,844
0,07	0,876	0,183	15,096	15,913	0,255	0,867
0,08	0,845	0,109	15,112	15,948	0,289	0,911
0,09	0,825	0,052	15,196	16,012	0,312	0,963
0,10	0,782	0,048	15,129	15,956	0,342	0,985

Table 10.1-2. Metrics degradation from the noise experiment in percentage

NOISE VARIANCE	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
0,00	**0,997**	**0,986**	**6,468**	**7,609**	**0,052**	**0,052**
0,01	0,8	6,0	83,0	66,4	43,7	357,1
0,02	1,2	13,5	104,4	84,8	106,7	700,0
0,03	3,1	23,8	117,8	96,4	155,2	971,4
0,04	4,9	37,3	124,4	101,9	213,6	1171,4
0,05	7,2	60,2	128,2	104,4	276,9	1371,4
0,06	10,0	66,6	131,3	107,4	345,8	1528,6
0,07	12,1	81,4	133,4	109,1	387,0	1571,4
0,08	15,2	88,9	133,6	109,6	451,4	1657,1
0,09	17,2	94,8	135,0	110,4	496,5	1757,1
0,10	21,5	95,1	133,9	109,7	552,5	1800,0

Table 10.1-3. Metrics degradation from the Gaussian blur experiment

STANDARD DEVIATION	ACCURACY	VALIDATION RATEE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
0	**0,9965**	**0,9857**	**6,4679**	**7,6086**	**0,0524**	**0,0519**
1	0,9957	0,9747	6,3348	7,5249	0,0553	0,1111
2	0,9899	0,9348	7,7535	8,6961	0,0866	0,2222
3	0,9758	0,8490	8,7475	10,0413	0,1508	0,4148
4	0,9587	0,6483	9,7207	11,0002	0,2484	0,5138
5	0,9192	0,4717	10,5318	11,8231	0,3410	0,5630
6	0,8515	0,2130	11,1935	12,4259	0,4268	0,6296
7	0,7840	0,0920	11,7378	12,9049	0,4960	0,7407
8	0,7388	0,0610	12,2777	13,3167	0,5495	0,7926
9	0,7055	0,0437	12,6175	13,6724	0,5942	0,8222
10	0,6792	0,0390	12,9041	13,9599	0,6312	0,8596

Table 10.1-4. Metrics degradation from the Gaussian blur experiment in percentage

STANDARD DEVIATION	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
0	**0,997**	**0,986**	**6,468**	**7,609**	**0,052**	**0,052**
1	0,083	1,116	-2,058	-1,100	5,609	114,286
2	0,661	5,165	19,877	14,293	65,423	328,571
3	2,074	13,866	35,246	31,973	188,058	700,000
4	3,790	34,226	50,291	44,575	374,527	890,826
5	7,760	52,147	62,832	55,390	551,390	985,714
6	14,551	78,390	73,063	63,313	715,236	1114,286
7	21,325	90,666	81,478	69,608	847,411	1328,571
8	25,858	93,811	89,826	75,020	949,667	1428,571
9	29,202	95,570	95,079	79,695	1035,06	1485,714
10	31,844	96,043	99,511	83,474	1105,61	1557,786

Table 10.1-5. Metrics degradation from the motion blur experiment

KERNEL SIZE	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
0	**0,9965**	**0,9857**	**6,4679**	**7,6086**	**0,0524**	**0,0519**
3	0,9953	0,9853	6,2926	7,4780	0,0527	0,0593
5	0,9942	0,9767	6,3769	7,6107	0,0548	0,0889
7	0,9927	0,9650	6,6799	7,9063	0,0624	0,1556
9	0,9925	0,9417	7,0470	8,2531	0,0730	0,2296
11	0,9890	0,9240	7,4316	8,6433	0,0875	0,2963
13	0,9852	0,9043	7,7827	9,0208	0,1028	0,3556
15	0,9810	0,8623	8,1101	9,3635	0,1207	0,3926
17	0,9748	0,7930	8,4178	9,6787	0,1399	0,4444
19	0,9663	0,7397	8,7187	9,9900	0,1584	0,5037
21	0,9583	0,6803	8,9849	10,2712	0,1546	0,5333

Table 10.1-6. Metrics degradation from the motion blur experiment in percentage

KERNEL SIZE	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
0	**0,997**	**0,986**	**6,468**	**7,609**	**0,052**	**0,052**
3	0,117	0,034	-2,710	-1,717	0,758	14,286
5	0,234	0,913	-1,406	0,027	4,583	71,429
7	0,384	2,097	3,277	3,913	19,279	200,000
9	0,401	4,464	8,954	8,470	39,386	342,857
11	0,753	6,257	14,900	13,599	67,044	471,429
13	1,137	8,252	20,328	18,560	96,452	585,714
15	1,555	12,513	25,390	23,064	130,499	657,143
17	2,175	19,547	30,147	27,207	167,142	757,143
19	3,028	24,958	34,800	31,298	202,519	871,429
21	3,830	30,978	38,915	34,994	195,263	928,571

Table 10.1-7. Metrics degradation from the contrast experiment

CONTRAST FACTOR	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
1,0	**0,9965**	**0,9857**	**6,4679**	**7,6086**	**0,0524**	**0,0519**
0,9	0,9963	0,9857	6,6334	7,8000	0,0524	0,0593
0,8	0,9958	0,9863	6,8245	7,9865	0,0528	0,0667
0,7	0,9957	0,9847	7,0662	8,2576	0,0530	0,0741
0,6	0,9952	0,9837	7,5076	8,6945	0,0538	0,0815
0,5	0,9953	0,9790	8,0324	9,2205	0,0552	0,0963
0,4	0,9945	0,9683	8,7966	9,9751	0,0588	0,1481
0,3	0,9917	0,9407	9,8673	11,0236	0,0709	0,2000
0,2	0,9798	0,8323	11,3151	12,4393	0,1102	0,3037
0,1	0,8543	0,0273	13,5935	14,5980	0,3308	0,7926

Table 10.1-8. Metrics degradation from the contrast experiment in percentage

CONTRAST FACTOR	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
1,0	0,997	0,986	6,468	7,609	0,052	0,052
0,9	0,017	0,000	2,559	2,515	0,185	14,286
0,8	0,067	-0,067	5,513	4,966	0,910	28,571
0,7	0,083	0,101	9,251	8,530	1,304	42,857
0,6	0,133	0,203	16,075	14,272	2,707	57,143
0,5	0,117	0,677	24,188	21,185	5,490	85,714
0,4	0,201	1,759	36,005	31,102	12,336	185,714
0,3	0,485	4,565	52,558	44,882	35,385	285,714
0,2	1,673	15,557	74,943	63,489	110,463	485,714
0,1	14,267	97,227	110,170	91,861	531,864	1428,571

Table 10.1-9. Metrics degradation from the low brightness experiment

BRIGHTNESS FACTOR	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
1,0	**0,9965**	**0,9857**	**6,4679**	**7,6086**	**0,0524**	**0,0519**
0,9	0,9963	0,9850	6,7060	7,8234	0,0525	0,0519
0,8	0,9965	0,9850	6,8995	8,0156	0,0528	0,0667
0,7	0,9963	0,9830	7,1050	8,2215	0,0529	0,0667
0,6	0,9962	0,9760	7,4271	8,5286	0,0534	0,0667
0,5	0,9960	0,9753	7,6763	8,8173	0,0543	0,0963
0,4	0,9953	0,9683	8,1112	9,2502	0,0558	0,0963
0,3	0,9935	0,9610	8,7266	9,8644	0,0627	0,1407
0,2	0,9857	0,9267	9,6779	10,8681	0,0878	0,2815
0,1	0,5475	0,0003	12,3576	13,4337	0,2446	0,7556

Table 10.1-10. Metrics degradation from the low brightness experiment in percentage

BRIGHTNESS FACTOR	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
1,0	0,997	0,986	6,468	7,609	0,052	0,052
0,9	0,017	0,068	3,681	2,823	0,277	0,000
0,8	0,000	0,068	6,673	5,348	0,809	28,571
0,7	0,017	0,271	9,851	8,055	1,104	28,571
0,6	0,033	0,981	14,831	12,091	2,055	28,571
0,5	0,050	1,049	18,684	15,886	3,763	85,714
0,4	0,117	1,759	25,408	21,574	6,572	85,714
0,3	0,301	2,503	34,923	29,647	19,794	171,429
0,2	1,087	5,986	49,630	42,839	67,719	442,857
0,1	45,058	99,967	91,061	76,559	367,188	1357,143

Table 10.1-11. Metrics degradation from the high brightness experiment

BRIGHTNESS FACTOR	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
1,0	**0,9965**	**0,9857**	**6,4679**	**7,6086**	**0,0524**	**0,0519**
1,2	0,9952	0,9863	6,4763	7,6331	0,0524	0,0444
1,4	0,9942	0,9767	6,8012	8,0657	0,0532	0,0667
1,6	0,9920	0,9480	7,4323	8,6951	0,0547	0,0741
1,8	0,9822	0,8830	8,1965	9,4396	0,0591	0,0963
2,0	0,9695	0,7690	9,0058	10,2465	0,0623	0,1111
2,2	0,9540	0,6890	9,7599	11,0076	0,0651	0,1333
2,4	0,9318	0,5743	10,3806	11,6264	0,0687	0,1778
2,6	0,9085	0,4737	10,8840	12,0786	0,0729	0,1926
2,8	0,8882	0,3850	11,2700	12,4501	0,0764	0,2148
3,0	0,8618	0,3043	11,6299	12,7698	0,0819	0,2667

Table 10.1-12. Metrics degradation from the high brightness experiment in percentage

BRIGHTNESS FACTOR	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
1,0	**0,997**	**0,986**	**6,468**	**7,609**	**0,052**	**0,052**
1,2	0,133	-0,067	0,131	0,322	0,051	-14,286
1,4	0,234	0,913	5,153	6,007	1,552	28,571
1,6	0,452	3,822	14,910	14,279	4,419	42,857
1,8	1,438	10,416	26,727	24,064	12,921	85,714
2,0	2,709	21,982	39,238	34,669	18,924	114,286
2,2	4,265	30,098	50,898	44,672	24,375	157,143
2,4	6,490	41,732	60,495	52,805	31,276	242,857
2,6	8,831	51,944	68,278	58,749	39,284	271,429
2,8	10,871	60,940	74,245	63,631	45,876	314,286
3,0	13,514	69,125	79,810	67,833	56,474	414,286

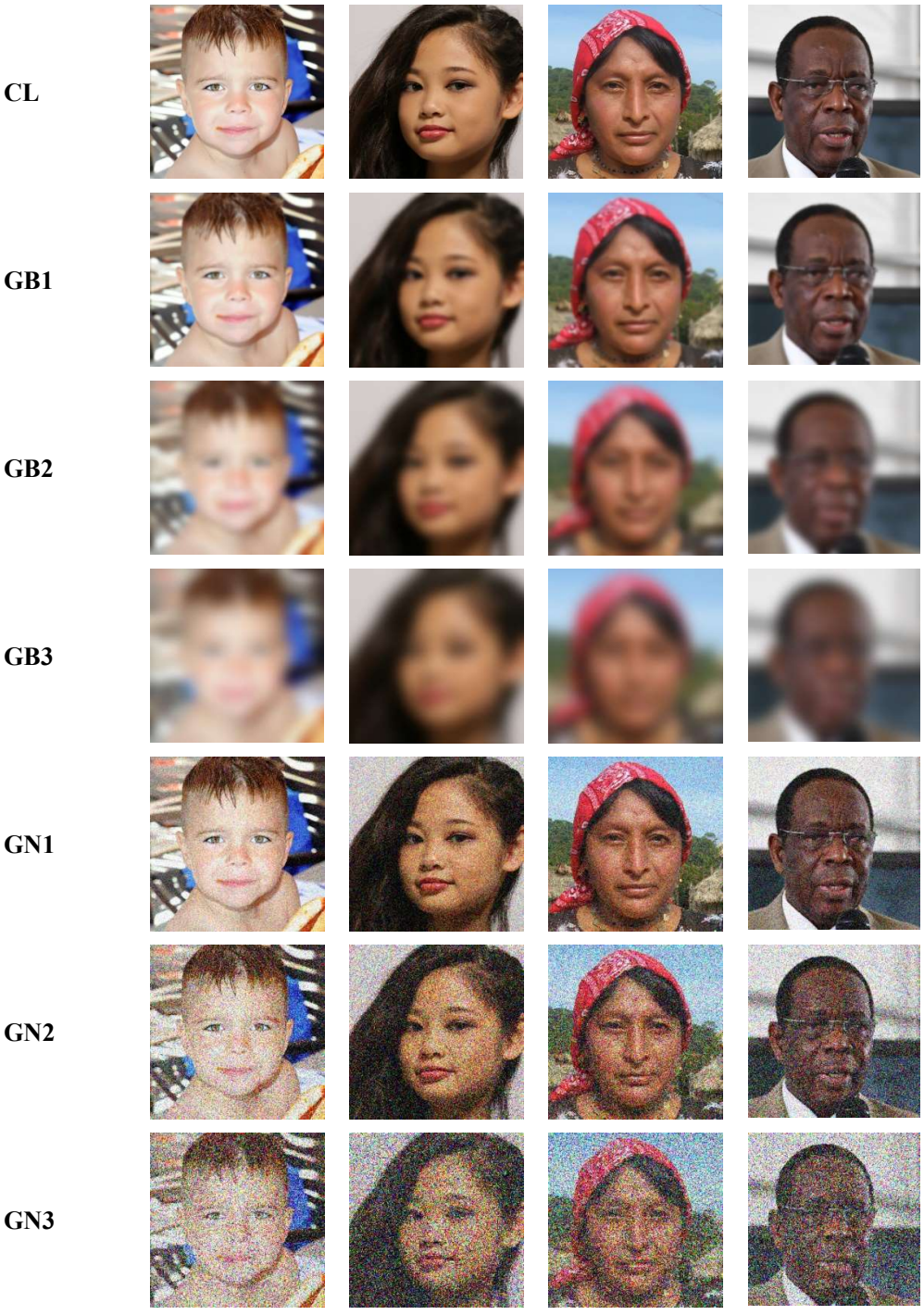
Table 10.1-13. Metrics degradation from the JPEG experiment

JPEG QUALITY	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
100	**0,99650**	**0,98567**	**6,46788**	**7,60864**	**0,05235**	**0,05185**
95	0,99583	0,98567	6,48624	7,62542	0,05239	0,05185
85	0,99617	0,98500	6,64621	7,76626	0,05248	0,05185
75	0,99567	0,98433	6,47174	7,61167	0,05267	0,05185
65	0,99600	0,98733	6,88184	7,98198	0,05282	0,05185
55	0,99567	0,98367	7,35502	8,42648	0,05295	0,05926
45	0,99550	0,98367	6,57975	7,79789	0,05303	0,07407
35	0,99583	0,97933	6,69005	7,85361	0,05361	0,06667
25	0,99550	0,98300	7,30135	8,40854	0,05352	0,05926
15	0,99317	0,96700	7,42562	8,62706	0,05519	0,07407
5	0,95067	0,53700	10,22565	11,37077	0,08484	0,28889

Table 10.1-14. Metrics degradation from the JPEG experiment in percentage

JPEG QUALITY	ACCURACY	VALIDATION RATE	MAE APPARENT	MAE REAL	ERROR	FAILURE RATE
100	**0,997**	**0,986**	**6,468**	**7,609**	**0,052**	**0,052**
95	0,067	0,000	0,284	0,221	0,067	0,000
85	0,033	0,068	2,757	2,072	0,239	0,000
75	0,083	0,136	0,060	0,040	0,610	0,000
65	0,050	-0,168	6,400	4,907	0,886	0,000
55	0,083	0,203	13,716	10,749	1,149	14,286
45	0,100	0,203	1,730	2,487	1,299	42,857
35	0,067	0,643	3,435	3,220	2,399	28,571
25	0,100	0,271	12,886	10,513	2,227	14,286
15	0,334	1,894	14,808	13,385	5,427	42,857
5	4,599	45,519	58,099	49,445	62,053	457,143

10.2 Datasets samples



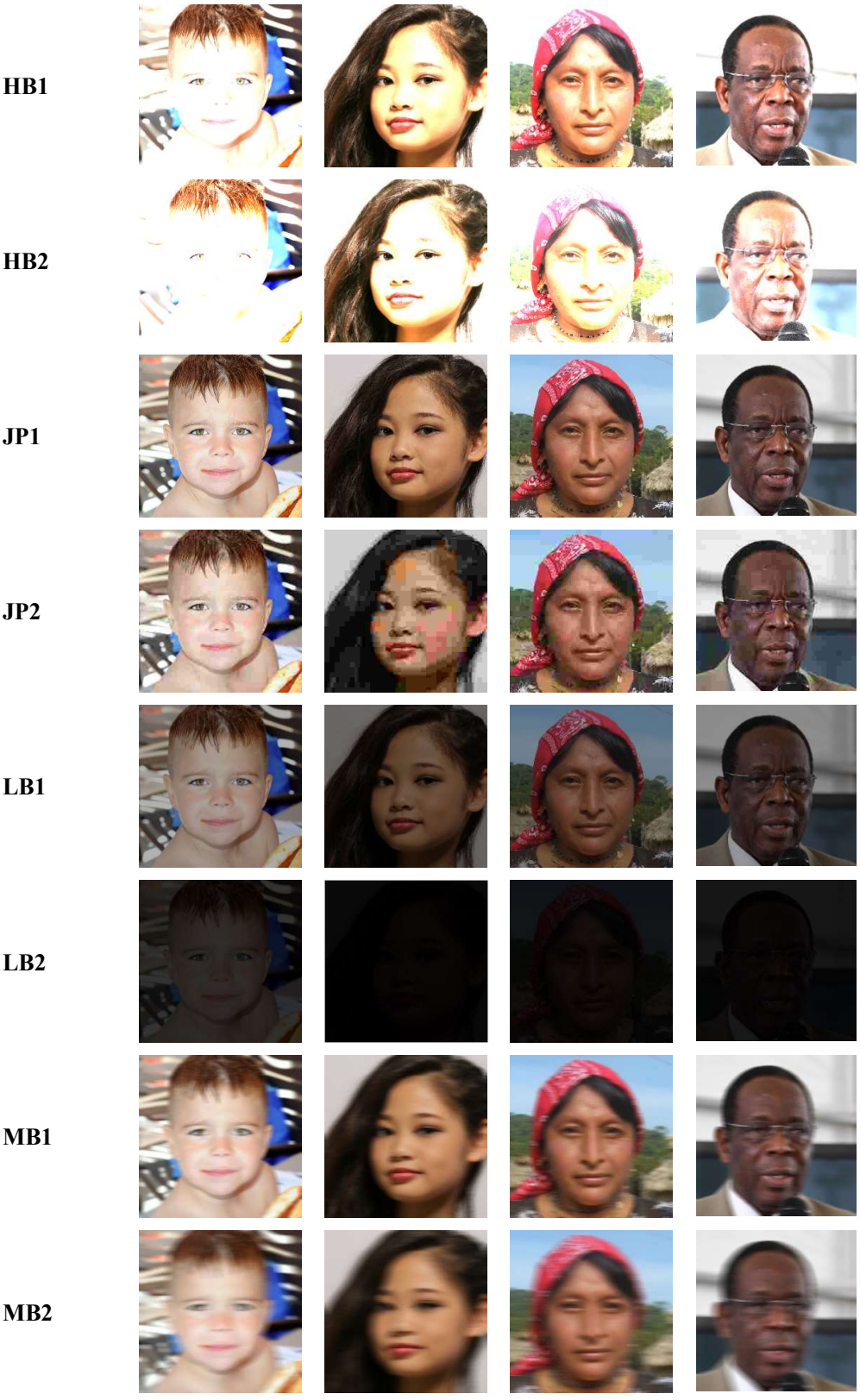


Figure 10.2-1. Samples of the datatset created for distortion classification in face images.

CL 

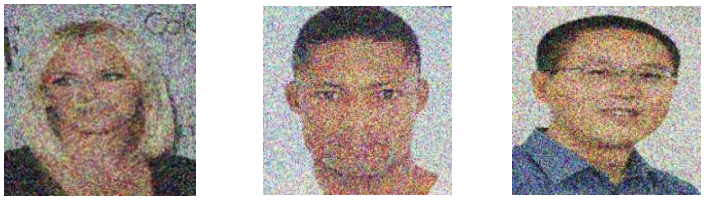
GB1 

GB2 

GB3 

GN1 

GN2 

GN3 

JP1 

JP2 

Figure 10.2-2. Samples of the images obtained from the IDEAL_LIVE_DFD.