

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO



Face Detection and Recognition in Unconstrained Scenarios

Anabela Machado Reigoto

Master's Degree in Electrical and Computers Engineering

Supervisor: Pedro Miguel Carvalho

Co-Supervisor: Paula Viana

September 23, 2019

Resumo

A detecção e o reconhecimento facial têm sido objetos de estudo ativo nos últimos anos, devido às suas diversas aplicações, como biométrica, controle de acesso, vigilância por vídeo e muitos outros sistemas interativos humano-máquina. Ambas são tarefas desafiadoras, especialmente em cenários sem constrangimentos, onde é mais propícia a ocorrência de oclusões e duma maior variabilidade das condições de iluminação, poses de cabeça e expressões faciais.

Até ao momento, as propostas para detecção e reconhecimento facial com melhor desempenho foram desenvolvidas recorrendo a técnicas de *deep learning* baseadas em redes neuronais convolucionais (CNN). No entanto, para treinar uma rede neuronal capaz detetar e reconhecer adequadamente uma face é necessário um grande número de imagens faciais, o que implica um alto custo computacional. Além disso, estas imagens requerem uma anotação correspondente, geralmente feita à mão, aumentando assim o esforço humano envolvido na preparação do conjunto de dados. Deste modo, há uma necessidade evidente de encontrar uma solução que permita a redução de dados e, ainda assim, realize de forma adequada a detecção e o reconhecimento facial.

Esta dissertação procura propor uma solução para o problema acima referido. Para tal, foram analisadas estratégias de reconhecimento facial. Estas são tipicamente divididas em dois tipos principais: verificação - “Estas são a mesma pessoa?” - e classificação - “Quem é esta pessoa?”. As duas abordagens foram implementadas, a primeira utilizando o detetor facial da biblioteca Dlib, seguido pela rede neuronal FaceNet, e a última aplicando *transfer learning* num modelo pré-treinado para detecção de objetos baseado numa CNN. Antes da implementação, foi necessário reunir e anotar manualmente um conjunto de imagens apropriado para os propósitos desta pesquisa.

A análise dos resultados dos dois métodos desenvolvidos revelou que um melhor desempenho poderia ser alcançado através do método baseado na classificação. No entanto, impulsionado pela necessidade de redução do conjunto de dados de treino, foi proposto um método de fusão, derivado da combinação entre os *outputs* da classificação e da verificação. Com este novo método, a quantidade de imagens anotadas diminuiu $\sim 93\%$, e o desempenho do reconhecimento facial sofreu apenas um pequeno impacto, diminuindo de $96,35\%$ para $93,96\%$.

Abstract

Face detection and recognition have been active objects of study for the past several years, due to their wide range of applications, such as biometrics, access control, video surveillance and many other interactive human-machine systems. These are challenging tasks, especially in unconstrained scenarios, which are more propitious to the occurrence of occlusions and of a wider variety of illumination conditions, head poses, and facial expressions.

So far, the outperforming proposals for either face detection and recognition were achieved through deep learning techniques based on convolutional neural networks (CNN). However, in order to train a neural network to detect and recognize a face properly, a vast number of face images is required, implying a high computational cost. Additionally, these data need their correspondent annotations, commonly hand-crafted, increasing the human effort in the dataset preparation. Therefore, there is an evident need to find a solution that allows the reduction of data samples and yet perform proper face detection and recognition.

This dissertation attempts to propose a solution to the problem described above. For this purpose, face recognition strategies were analyzed. These are typically divided into two main types: verification - “Are these the same person?” - and classification - “Who is this person?”. The two approaches were implemented, the first using the state-of-the-art Dlib’s CNN face detector followed by the FaceNet network, and the latter applying transfer learning in an object detection pre-trained model based on a CNN. Prior to the implementation, it was necessary to collect and manually annotate an adequate dataset for these research purposes.

The analysis of both developed methods’ results revealed that higher performance could be accomplished with the classification based method. Nevertheless, moved by the need for reduction of the training set, it was proposed a fusion method, derived from the combination of classification and verification outputs. With this novel method, the amount of annotated images decreased $\sim 93\%$, and the facial recognition performance suffered only a small negative impact, decreasing from 96.35% to 93.96%.

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor, Professor Pedro Miguel Carvalho, for all the guidance and support; to my co-supervisor, Professor Paula Viana; to Luís Vilaça; and to INESC TEC for giving the means to fulfill my goals.

I would also like to acknowledge the FotoInMotion project, funded by the Framework Programme H2020 of the European Commission and the CHIC project co-financed by the European Union through the European Regional Development Fund (Cohesion Fund, European Social Fund, European Structural, and Investment Funds) under COMPETE 2020 (Programme Competitiveness and Internationalization).

I also wish to sincerely thank all my friends and my family for all the encouragement they gave me.

A special thanks goes to my mother, for always being my backbone.

Anabela Machado Reigoto

*“Look wide, and even when you think you are looking wide,
look wider still.”*

Robert Baden-Powell

Contents

1	Introduction	1
1.1	Context and Motivation	1
1.2	Objectives	2
1.3	Contributions	3
1.4	Document Structure	3
2	Literature review	5
2.1	Face Detection	5
2.1.1	Datasets	5
2.1.2	Algorithms	8
2.2	Facial Recognition	9
2.2.1	Datasets	9
2.2.2	Algorithms	11
2.3	Towards Expression Recognition	12
2.3.1	Datasets	13
2.3.2	Algorithms	14
2.4	Discussion	16
3	Problem Characterization and Methodology	19
3.1	Problem Definition	19
3.2	Proposed Solution	20
3.3	Methodology	21
3.3.1	Verification Based Method	21
3.3.2	Classification Method	24
3.3.3	Fusion Method	24
4	Datasets Preparation	25
4.1	Base Dataset	25
4.2	Ground Truth Dataset	29
4.3	Noisy Dataset	30
4.4	Face Detection Dataset	31
4.5	<i>Gallery Set</i>	32
5	Assessment Strategy	35
5.1	Overall Assessment Strategy	35
5.2	Verification Based Method	36
5.3	Classification Method	38
5.4	Metrics	38

6	Results	41
6.1	Verification Based Method	41
6.1.1	Experimentation strategy	41
6.1.2	<i>Threshold Value</i>	42
6.1.3	Results Analysis	43
6.2	Classification Method	53
6.2.1	Model Parameters	53
6.2.2	Results Analysis	54
6.3	Comparative Analysis	55
6.4	Fusion Method	56
7	Conclusion	59
	References	61

List of Figures

2.1	Facial landmark annotations with different number of points, 17 points, 29 points, 51 points and 68 points. (Extracted from [1].)	5
2.2	Controlled conditions datasets examples. (Extracted, from left to right, from IMM, FERET, MUCT and AR datasets.)	6
2.3	Unconstrained conditions datasets examples. (Extracted, from left to right, from LFPW, 300-W, AFLW and Helen datasets.)	7
2.4	Face recognition datasets examples. (Extracted, from left to right, from CASIA WebFace, UMD, MegaFace and VGGFacen datasets.)	11
2.5	Basic Facial Expressions, Anger, Disgust, Fear, Happy, Contempt, Sadness and Surprise, respectively. (Extracted from the CK+ dataset.)	13
2.6	Usual Facial Expression Recognition System Diagram of Blocks.	14
3.1	Verification Based Method system model structure.	21
3.2	Dlib's CNN Face Detector performance example.	22
3.3	FaceNet face recognition process.	23
3.4	Triplet loss-based training process.	23
3.5	Classification Method system model structure.	24
3.6	Proposed Fusion Method model structure.	24
4.1	Donald Trump's dataset images examples.	26
4.2	Marcelo Rebelo de Sousa's dataset images examples.	26
4.3	Ricardo Rio's dataset images examples.	27
4.4	Rui Moreira's dataset images examples.	27
4.5	Distribution of the different head poses among the Base Dataset.	28
4.6	Distribution of the different image file sizes among the Base Dataset.	28
4.7	Distribution of the different images dimensions among the Base Dataset.	29
4.8	Distribution of amount of people per image among the Base Dataset.	29
4.9	Donald Trump's Base Dataset image with the ground truth annotated bounding box, represented in green and its cropped version.	30
4.10	Ground truth annotated bounding box diagonal, represented in blue.	30
4.11	Noisy Dataset modified bounding box examples, represented in red. Ground truth bounding boxes represented in green.	31
4.12	Bounding box resulting of the Dlib's CNN Face Detector represented in red. Ground truth annotated bounding box represented in green.	32
4.13	Bounding box resulting of the Dlib's CNN Face Detector expanded by 25% represented in red. Ground truth annotated bounding box represented in green.	32
4.14	<i>Gallery Set</i> examples.	33
5.1	Examples of personalities' images under different conditions.	35

5.2	Proposed face recognition verification strategy pipeline.	37
5.3	<i>Probe</i> and <i>Galley Sets</i>	38
5.4	Practical example of the used designations.	39
5.5	Intersection over Union computation.	40
6.1	ROC curve between Rank-1 $R(\tau)$ and $FAR(\tau)$	42
6.2	Frontal Gallery Set results with Noisy Dataset as <i>Probe Set</i>	44
6.3	Frontal Gallery Set results with Face Detection Dataset as <i>Probe Set</i>	45
6.4	Example of a Marcelo Rebelo de Sousa's incorrect <i>top match</i>	47
6.5	3 Images Gallery Set results with Noisy Dataset as <i>Probe Set</i>	47
6.6	3 Images Gallery Set results with Face Detection Dataset as <i>Probe Set</i>	48
6.7	5 Images Gallery Set results with Noisy Dataset as <i>Probe Set</i>	49
6.8	5 Images Gallery Set results with Face Detection Dataset as <i>Probe Set</i>	50
6.9	Average results of each Gallery Set.	51
6.10	Average results of each personality dataset.	52
6.11	F-measure of Fusion Method.	58

List of Tables

6.1	False Acceptance Rates for the threshold $\tau = 1$	43
6.2	Average Results.	53
6.3	5-Folds Cross Validation average mAP and AR for validation set.	55
6.4	5-Folds Cross Validation average mAP and AR for test set.	55
6.5	F-measure for the Classification and the Fusion Method.	55
6.6	F-measure for Classification and Fusion Methods.	57

Acronyms and Symbols

300-W	300 Faces in-the-Wild Challenge
AAM	Active Appearance Model
AFEW	Acted Facial Expressions in the Wild
AFLW	Annotated Facial Landmarks in the Wild
API	Application Programming Interface
AFW	Annotated Faces in the Wild
ASM	Active Shape Model
AU	Action Units
BDBN	Boosted DBN
CK+	Extended Cohn-Kanade
CNN	Convolutional Neural Network
COFW	Caltech Occluded Faces in the Wild
DBN	Deep Belief Network
DPM	Deformable Parts Model
EBGM	Elastic Bunch Graph Matching
FACS	Facial Action Coding System
FAUs	Facial Action Unit
Fddb	Face Detection Data Set and Benchmark
FER	Facial Expression Recognition
FER-2013	Facial Expression Recognition 2013
FERET	Facial Recognition Technology
FR	Face Recognition
GPU	Graphics Processing Unit
HOG	Histogram of Oriented Gradients
HGPP	Histogram of Gabor Phase Patterns
ICA	Independent Component Analysis
ICF	Integral Channel Features
JAFFE	Japanese Female Facial Expression
LBP	Local Binary Patterns
LDP	Local Directional Pattern
LBP-TOP	LBP on three orthogonal planes
LDA	Linear Discriminant Analysis
LFPW	Labeled Face Parts in the Wild
LFW	Labeled Faces in the Wild
LGBP	Local Gabor Binary Patterns
LLE	Locally Linear Embedding
LPP	Locality Preserving Projection

LRC	Linear Regression Classifier
MB	Megabyte
ML	Machine Learning
PCA	Principal Component Analysis
PHOG	Pyramid Histogram of Oriented Gradients
R-CNN	Region-based Convolutional Neural Network
ROI	Region Of Interest
RaFD	Radboud Faces Database
SFEW	Static Facial Expressions in the Wild
SIFT	Scale-invariant Feature Transform
SSD	Single Shot Detector
SFEW	Static Facial Expressions in the Wild
SIFT	Scale-invariant Feature Transform
SSD	Single Shot Detector
SURF	Speeded Up Robust Features
XM2VTS	Extended M2VTS Database

Chapter 1

Introduction

1.1 Context and Motivation

In recent years, we have been observing a notable increase in the available multimedia content and in the number of platforms through which it can be accessed. This has risen the urgency of finding improved ways to associate annotations to the available content, in order to increase its value, allow a better exploitation and, eventually, permit its reuse. However, this is a complex and lengthy process. Apart from the considerable amount of time it requires and the human cost associated with it, the annotations are susceptible to errors caused by human hand and, possibly, its fatigue. In this light, alternative annotation approaches were sought, leading to the development of various automatic methods and, ultimately, to the evolution of deep learning based techniques, often employed in cloud environments.

Within the field of content annotation and visual analysis, one should highlight the importance and multi-functionality of visual resources as face detection and recognition. Their practical application ranges from scenarios such as surveillance, biometrics and clinical monitoring to criminal identification and interactive games, among many other human-computer systems. Therefore, and since facial recognition relies on face detection, both topics have been the object of several research proposals and projects. Due to the technological advances of the past several years, namely in the scope of facial detection using intelligent networks, significant improvements have been achieved in the recognition of human faces.

However, despite all these advances in both face detection and recognition, flaws can still be found, especially in their application in unconstrained scenarios. Within these, the processes of face detection and recognition can be seriously compromised by occlusions, happening due to the presence of sunglasses, by the various head poses depicted, by the low quality of some images or by less favorable illumination conditions - common characteristics of real-life scenarios. For this reason, even using advanced techniques, like machine learning (ML), these require a large amount of training data for proper detection and recognition. The training set is commonly composed of images and their correspondent annotations, which contain the face bounding box and the respective label, indicative of the person's identity. The collection and annotation of several

data samples represent an additional human effort; besides, these hand-crafted annotations may also be compromised by its inherent human subjectivity, which here may affect the consideration of the head/face region for each image.

Facial recognition derives from object detection and classification approaches, and this can be performed through two different processes - face verification/ authentication and face classification/ identification. On the one hand, face verification is concerned with the validation of identity, based on an image of an individual's face; its function is to authenticate or to reject the claimed identity. On the other hand, face identification attempts to identify a particular subject, based on the image of its face. This classification relies on a set of pre-identified people pictures, usually a labeled dataset. Being two different strategies with the same purpose, these types of recognition methods are deserving of investigation and comparison, especially when addressing facial analysis.

In this context, this dissertation proposes to study existing verification and classification approaches. It is further analyzed in this work how the combination of the two distinct strategies can be valuable to the field, particularly by reducing the quantity of the annotated data required to achieve satisfactory results. By doing so, it is expected that this study will contribute to the improvement of the already existing algorithms, by training intelligent networks and bringing together different solutions for face detection and recognition to obtain more efficient results.

1.2 Objectives

In this investigation, it was intended to perform facial analysis on images, to then proceed to their automatic annotation, in terms of both face detection and recognition. This content annotation becomes even more interesting within the context in which this dissertation is proposed, that pays particular attention to photojournalism images, and, in specific, to local and national audiovisual broadcast content. To conduct the investigation, the following objectives were defined:

- To analyze strategies for face detection and recognition, namely the classification and the verification;
- To analyze the impact of the possible introduced noise by the face detection on the verification based methods results;
- To analyze the performance of a classification based method for face recognition with reduced training data;
- To compare the strategies of verification and classification and define a combination approach to overcome their individual limitations and/or enable smaller training datasets.

1.3 Contributions

The main contributions of the work are:

- Preparation of a dataset of images and corresponding annotations containing four distinct personalities, from which additional variations were generated;
- Comparative analysis between verification and classification face recognition strategies;
- Proposal for an approach based on the fusion of two strategies that enables a significant reduction in the training data.

1.4 Document Structure

The dissertation will be further divided into six sections. In Chapter Two, a literature review will be presented, to expose the existing algorithms and datasets. Chapter Three will be devoted to the problem characterization and to the used methodology explanation. Thereafter, Chapter Four will focus on the datasets preparation and Chapter Five on the assessment strategy for the proposed methodologies. At last, the results of the suggested methods are exposed in Chapter Six, followed by the conclusions of the study, displayed in Chapter Seven.

Chapter 2

Literature review

This chapter encompasses an analysis of existing relevant algorithms for both face detection and recognition, as well as the extension to facial expression. The algorithms are described and contextualized on relevant trends in recent years. Given their growing relevance, specifically in the context of deep learning techniques, existing relevant datasets are characterized.

2.1 Face Detection

2.1.1 Datasets

There is a vast number of state-of-the-art datasets for face detection available; these were collected under both controlled and uncontrolled environments and each one labeled according to a different number of facial landmark points, as shown in Figure 2.1.

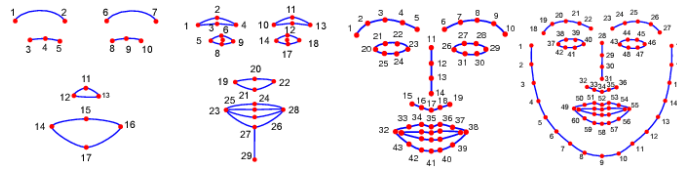


Figure 2.1: Facial landmark annotations with different number of points, 17 points, 29 points, 51 points and 68 points. (Extracted from [1].)

Controlled conditions datasets

Controlled datasets generally cover subject images taken under lab predefined conditions, such as a certain degree of illumination and occlusion.

Extended M2VTS Database (XM2VTS) [2] was built through video recording of 295 subjects, including four recordings of each subject. each one of these recordings included a speaking headshot and a rotating headshot. Also available, are synchronized video and speech data, a 3D face model of each subject and 2360 color face images labeled with 68 points. This database is only available at some cost.

AR database [3] is a publicly available database consisting of over 4000 frontal color images of 126 subjects, 70 male and 56 female, captured in different illumination conditions. Moreover, some of them display facial expressions – neutral, smile, anger or a scream – and others contain facial occlusions, provoked, for example, by sunglasses or a scarf. In [4], each image of this dataset was manually annotated with 22 facial features points.

The IMM database [5] contains 240 color images of 40 subjects, 33 male and 7 female, manually labeled with 58 points, discerning the eyebrows, eyes, nose, mouth, and jaw.

MUCT database [6] collected 3755 color images of 276 different subjects' faces, with each image labeled with 76 landmark points. This database has the particularity of presenting examples of diverse age and ethnicity groups, acquired under different lighting conditions.

PUT database [7] consists of 9971 high resolution color images of 100 subjects captured in partially controlled conditions. Each image was annotated with 30 landmark points and, additionally, a subset of 2193 near-frontal images was labeled with 194 points.

Facial Recognition Technology (FERET) [8] database contains 14126 images of 1199 different subjects, including about 20 head poses differing on yaw angles. The frontal images also differ in illumination and facial expression.

BioID database [9] collected 1521 gray level frontal face images of 23 subjects. The images were taken under a variety of illumination conditions and backgrounds, being each one manually labeled with 20 points. Despite being captured in a lab environment, this is a dataset that aims to simulate "real world" conditions.

Some examples of the aforementioned controlled conditions datasets are presented in Figure 2.2.



Figure 2.2: Controlled conditions datasets examples. (Extracted, from left to right, from IMM, FERET, MUCT and AR datasets.)

Unconstrained conditions datasets

Uncontrolled conditions datasets are intended to capture unconstrained real-world scenarios and are usually obtained through websites like Facebook or Flickr.

Face Detection Data Set and Benchmark (FDDB) [10] is a gray scale and color dataset of 28451 images with a total of 5171 faces. In these images were included occlusions, difficult head poses, low-resolution, and blurred faces. Face regions were annotated as elliptical regions.

Labeled Faces in the Wild (LFW) database [11] consists of 13233 face images of 5749 different subjects collected from the web. These faces were detected by Viola-Jones face detector and it can be noted that 1680 of these appear in two or more images, and the remaining only appear in a single image. Each image is labeled with the name of the person in the center of the image, not having any manually landmarks points annotated.

Annotated Facial Landmarks in the Wild (AFLW) database [12] is a large-scale database built using Flickr images that includes a wide variety of head poses, expressions, and individuals from different ethnic, age and gender groups. This database contains 21997 images with a total of 25993 faces, having each image an annotation of 21 landmarks points.

Labeled Face Parts in the Wild (LFPW) database [13] contains 1432 facial images downloaded from internet search sites, such as Google.com, Flickr.com, and Yahoo.com, using simple text queries. It provides a list of image URLs, some of which are no longer available. Each face was labeled with 35 points by workers on Amazon Mechanical Turk (MTurk), with only 29 of these being used in this database proposal.

Annotated Faces in the Wild (AFW) database [14] was organized through Flickr images and contains 205 images with a wide range of face pose and scale variations in cluttered backgrounds. Each face annotations include the correspondent rectangular bounding box and 6 landmark points.

Helen database [15] is a high-resolution database which contains 2330 face images obtained from Flickr, where each face is labeled with 194 points.

Caltech Occluded Faces in the Wild (COFW) database [16] consists of 1007 face images; these display a significantly large variety of occlusions due to different poses, the use of accessories (e.g. sunglasses and hats) and the interactions with objects, such as hands and microphones. All images were hand-annotated with 29 landmark points and each point is labeled whether it is occluded or not.

300 Faces in-the-Wild Challenge (300-W) [17] database is a combination of face images from different databases (LFPW, Helen, AFW, and XM2VTS). It comprises another 135 images in difficult poses and expressions. All these datasets' images were re-labeled with 68 points.

Figure 2.3 shows four examples of the referred unconstrained conditions datasets.



Figure 2.3: Unconstrained conditions datasets examples. (Extracted, from left to right, from LFPW, 300-W, AFLW and Helen datasets.)

2.1.2 Algorithms

Face detection algorithms can be divided into two main categories, Traditional Methods and Deep Learning Methods [18].

Traditional methods are understood as procedures in which feature extraction is hand-crafted. Therefore, the widely known Viola-Jones [19] face detector can be classified as a traditional face detection algorithm. Viola-Jones face detector represents a mark in terms of face detection algorithms on account of its outperforming results, having had an enormous impact. It is based on the construction of an integral image with an AdaBoost learning cascade structure using Harr-like features, considered in the literature as a referenced effective method and still used since it allows a real-time face detection with considerable accuracy. A great amount of work on the potential of Viola-Jones face detector was carried out, having been investigated features like SURF (Speeded Up Robust Features) [20], HOG (Histogram of Oriented Gradients) [21], SIFT (Scale-invariant Feature Transform) [22] and LBP (Local Binary Patterns) [23] on an identical structure of this face detector.

Also included in the traditional methods are algorithms based on structured models, which consider faces as objects constituted by several parts. Algorithms proposed in [24], [25], [26] and [27] learn and apply deformable parts model (DPM) [28] to deal with potential variances and deformations between facial parts, generally requiring less training data.

More recent approaches combine various features in channels with the idea of integral images, achieving a higher accuracy [29]. The Integral Channel Features (ICF) proposed in [29] was used in representative works, as presented in [26] and [30].

Despite their good results, hand-crafted features are less meticulous in matters of detecting faces with significant changes in poses, expressions, occlusion, and illumination [18].

Recently, a greater amount of research focused on deep learning approaches was conducted, in order to cope with computer vision problems and to allow it to be used as a basis for face detection algorithms. In deep learning based methods features are not defined by the investigator, but rather learned by the network. Thus, deep learning approaches can be more accurate than the traditional methods once trained with a large amount of data and consequently, they represent significant progress for face detection and exhibit outperforming results compared to the traditional ones.

Face detection deep learning based algorithms can be subdivided into three distinct classes: cascade convolutional neural network (CNN), region-based convolutional neural network (RCNN) and Single Shot Detector (SSD) based algorithms.

Sun et al. [31] early proposed a three-level cascade CNN which predicted five facial landmark points achieving a favorable accuracy. Nevertheless, the necessity to model each point by a convolutional network enhance the complexity of the proposal. Departing from the theory of Sun et al. [31], several other models, such as [32] and [33] were advanced. Zhang et al. [32] took advantage of the multi-task learning idea with the intent of increasing the performance of the existing models. Multiple tasks, predictions of the head's position, the identification of facial landmark points and facial attributes are considered together; they share and learn common deep layers,

which results in a shared representation. Whereas Zhang et al. [33] proposed an auto-encoder network to perform the cascade facial landmark search.

In Cascade CNN [34] it is proposed a cascaded scheme to CNNs with a strong detection capability and high performance. More recently, several cascade CNN based algorithms were suggested, as [35], [36] and [37]. The cascade structure proposal is to reject simple negative samples at the first levels and refine the results at the last layers [18]. These methods achieve a high computational speed, though they are not able to improve the detection of crowded, tiny and blurry faces.

Almost in parallel with the development of CNN, R-CNN based algorithms were also being researched. R-CNN algorithms as [38] make use of a region proposal method to identify potential face regions of interest (ROIs) and to further refine these regions for face detection. These methods are slower in training since the feature extractions are repeated for each different ROI identified. To address this problem Fast R-CNN [39] and Faster R-CNN [40] were proposed, increasing significantly the computational speed of the model. Faster R-CNN based models as [41] and [42] use a detector capable of covering objects at different scales. The major drawback to the adoption of these methods is their limitations on detecting tiny faces. To overcome this difficulty, [43] and [44] models are trained including lower-level convolutional layers features, i.e. integrating the features from the face surrounding to train the model.

The development of SSD [45] was intended to perform detection at multiple scales using multi-scale feature maps to increase accuracy. Despite the improvements, SSD is not appropriated to detect small faces. To address this problem, more recent approaches like S3FD [46], FaceBoxes [47], Scaleface [48] and HR-ER [49] use similar techniques but either improving the matching strategy or assigning layers with specific scale ranges, improving the detection performance on tiny faces.

More technical and deepened description of the overall methods can be found in [18], [50] and [51].

2.2 Facial Recognition

Face recognition (FR) as a machine identification of a person by his facial image represents a challenging task in the field of image analysis and computer vision, which continues to draw researchers in diverse areas due to its various applications. The FR intends to verify/identify an individual's identity by comparing an input image data against stored images data. Therefore, there are two main categories in face recognition: **verification/authentication** - 1:1 match -, and **identification/recognition** - 1:N matching.

2.2.1 Datasets

With the rise of machine learning techniques, an important aspect when implementing a face recognition system is the training data required to learn different face representations. These datasets are commonly characterized by the presence of labels indicative of the individuals' identities. As suggested by Zhough et al. [52], improvements in performance in an FR system are

achieved when a vast amount of data is combining with deep learning. Early methods for deep FR were trained with private datasets, derived by companies' internally label face sets, such as Facebook [53], with 4M images of 4K people, and Google [54], composed by 200M of 3M individuals. Although the promising results obtained through this, the following researchers could not develop and compare their methods without public available training sets. Therefore, some other datasets have been provided over time. Of these, the most widely used ones are described below.

CASIA WebFace [55] provided the first widely used public training set. Composed of 500K samples collected by the web. This contains crop face images of 10K different celebrities. Some individuals, usually the more famous ones, comprise the majority of the dataset, while others are only represented by a few images.

VGGFace [56] is proposed as a dataset to train deep models and is composed of about 2.6 million images of 2622 subjects. Therefore, and contrarily to what is presented in the CASIA dataset, in this, each individual is represented in about 1000 samples. Most of these are high-quality frontal faces. Moreover, the VGGFace2 [57] dataset was proposed as an improved version of the latter, to overcome some limitations of its first version. VGGFace2 comprises 3M images of 9131 persons, in which are covered a wide range of head poses, age, and ethnicity, hence addressing a large range of intra-class variations.

Proposed by Basal et al. [58], UMDFaces is a face dataset which in the annotations were performed by both humans and deep-based face analysis tools. There are provided annotations for key points, face poses angles and gender. It contains both images and video frames to represent a total of 8227 subjects.

MS-Celeb-1M [59] and Megaface [60] are two large-scale training datasets, which have several individuals, but a limited number of samples per subject. The breadth of this, allows the model to train with a sufficiently variable representation of the different subjects. The MS-Celeb-1m consists of 10M images of 100k personalities. Megaface has about 1M images of 690K persons, which might appear in two samples as in 2469, having a marked uneven distribution of the set images.

More recently, the IMDb-Face [61] dataset was announced as a set derived by the MS-Celeb-1M and MegaFace. This novel dataset claims to be the largest noise-controlled face set, with 59K data samples for 1.7 million subjects.

Alongside these, another common dataset used for face recognition systems is the aforementioned LFW dataset. Some representative examples were extracted from some of the commonly used datasets of the FR process and present in Figure 2.4.



Figure 2.4: Face recognition datasets examples. (Extracted, from left to right, from CASIA Web-Face, UMD, MegaFace and VGGFace datasets.)

2.2.2 Algorithms

Traditional face recognition algorithms are commonly classified into **holistic** and **feature-based** methods.

Holistic methods use the whole face image as input to face recognition system, using a global representation of the entire face region, frequently by projecting it into a lower dimensional space, while retaining the most relevant information. Among this category, the most popular approach is EigenFaces [62], in which principal component analysis (PCA) is applied to a set of training face images, projecting the original data into a lower dimensional subspace, finding the eigenvectors of the set of the faces - eigenfaces - that account for the most variance in the data distribution, i.e. that maximize the total scatter across all classes. To identify faces, the set of training face images are saved as collections of weights, defining the contribution each eigenface has to that image. Also, the new face - the one intended to be identified - is projected into the subspace spanned by the eigenfaces. Face recognition can be performed comparing these weights with all weights of the training set, to find the closest match.

Based on this, more linear projection approaches were proposed in the literature, such as independent component analysis (ICA) [63] and linear regression classifier (LRC) [64]. Linear projection approaches project the face onto a linear subspace spanned by the eigenface images. These may fail to represent adequately faces, once face patterns, in the higher dimensional space, lie on a complex non-linear and non-convex manifold. So, one issue with PCA-based approaches is that these use a projection that maximizes the variance across all the images in the training set. This implies that the top eigenvectors might have a negative impact on the recognition accuracy since they might correspond to intra-personal variations that are not relevant for the recognition task (e.g. illumination conditions, pose variation or expression changes). Other methods, such as Linear Discriminant Analysis (LDA) [65][66], also known as FisherFaces, were proposed aiming to overcome these shortcomings.

With this in mind, the non-linear strategies were introduced, such as kernel PCA [67], kernel ICA [68], locality preserving projection (LPP) [69] or locally linear embedding (LLE) [70]. Non-linear strategies explore the non-linear structure of face patterns and typically use kernel techniques. These strategies consist of mapping the input face sample into a higher dimensional

space in which the variety of the faces is linear and simplified, to further application of the linear approaches techniques.

As for the feature-based methods, these process an image to identify and extract some individual features of the face, such as eyes, nose, mouth, etc., to a further calculation of their geometric relationships. The classification of a face is through the comparison and combination of the corresponding local features statistics. In this category, local appearance features relieved certain advantages over holistic feature, since these are more robust to local changes, as, for example, occlusions, facial expression, or even misalignment of the face.

Within these, Local Binary Pattern (LBP)[71] is commonly used as a representative method. This describes the changes in the neighborhood around the central pixel. Consequently, it is invariant to monotonic intensity of the pixels and slight illumination modifications. Some variants from the original LBP were developed, like Histogram of Gabor Phase Patterns (HGPP) [72] and Local Gabor Binary Patterns (LGBP) [73]. More feature-based algorithms were presented over the years, such as Elastic Bunch Graph Matching (EBGM) [74] and Local Directional Pattern (LDP) [75]. EBGM represents face features as a graph, where nodes store local information about the face landmarks and edges represent their relations, for example, the distance between the nodes. The face recognition is through a graph similarity function, that is used for matching. As for LDP, this encodes directional (eight directions) pattern features to extract high-order local information.

Moreover, due to technological advances, artificial intelligent strategies were proposed, seeking to resolve the shortcomings of the aforementioned algorithms and approaches. Deep learning approach Deepface [53] was successfully applied for the face recognition process. Deepface is a Facebook's developed algorithm, which trains a deep CNN, using their private aforementioned training set, to classify faces. Additionally, it is used a siamese network architecture, where pairs of face samples are applied to the same CNN, in order to obtain representations than are then compare using the Euclidean distance. Within this context, the main goal of the training is to minimize the distance between pairs of faces representing the same identity and maximize the distance between distinct individuals' faces pairs.

Through siamese networks, satisfactory outcomes were achieved, and thus more algorithms base on this architecture were proposed in the literature, such as Google's FaceNet [54] and OpenFace [76]. Although the three present similar results, in several researches, such as [56], the FaceNet network is pointed out as the outperforming one.

2.3 Towards Expression Recognition

A possible extension of face detection and recognition is expression recognition. To perform this task it is first necessary to detect faces in the image. While the detection can be for expression recognition, the identification of a specif person can add value. By preceding expression recognition in a processing pipeline, the noise of face detection and face recognition can propagate and affect the subsequent modules.

2.3.1 Datasets

There are some publicly available datasets for facial expression recognition, as presented in [77], being the following described ones noted as the most commonly adopted to test facial expression recognition (FER) algorithms.

The Multi-PIE [78] database is composed of more than 750000 images of 337 subjects, captured from 15 viewpoints and with 19 different illumination conditions. A subset was annotated with either 68 or 39 facial features points, depending on the head pose. Six facial expressions were also labeled: neutral, smile, surprise, squint, disgust and scream.

The Extended Cohn-Kanade (CK+) dataset [79] consists of 593 video sequences of frontal faces from 123 subjects. These videos show the change from a neutral expression to any other expression in a lab controlled scenario. Labeled with seven basic expressions, anger, contempt, disgust, fear, happy, sadness and surprise, the captured emotions are either posed or spontaneous expressions. Examples of these seven basic expressions are shown in Figure 2.5.



Figure 2.5: Basic Facial Expressions, Anger, Disgust, Fear, Happy, Contempt, Sadness and Surprise, respectively. (Extracted from the CK+ dataset.)

The Japanese Female Facial Expression (JAFPE) [80] is a lab controlled database containing ten Japanese female's expressions. There are represented seven basic expressions, anger, disgust, fear, happiness, sadness, surprise and neutral in a total of 213 gray scale images.

Inspired on the FLW dataset, the Acted Facial Expressions in the Wild (AFEW) dataset [81], first proposed in [82], was built collecting video clips from different scenes in movies available on the web. Composed by images of individuals coming from diverse races, gender and age groups, these videos contain various facial expressions, head movements, exhibit different occlusions; some of them even present more than one subject. This dataset is labeled with seven expressions classes: anger, disgust, fear, happiness, sadness, surprise and neutral. In [83] AFEW dataset is divided in three data sets mutually exclusive: Train (773 samples), Val (383 samples) and Test (653 samples).

The Static Facial Expressions in the Wild (SFEW) dataset [84] consists of static frames from AFEW dataset, being each image annotated with one of the seven expression classes. The most commonly used version is SFEW 2.0, also divided into three sets in [85]: Train (880 samples), Val (383 samples) and Test (372 samples).

The Facial Expression Recognition 2013 (FER-2013) dataset [86] was introduced in the contest Challenges in Representation Learning, as a large-scale database whose images were extracted from Google image search API. Incorrectly labeled images were then rejected and the remaining were resized to 48x48 pixels and converted to gray scale. It contains a total of 35887 images labeled with seven different expressions: anger, disgust, fear, happiness, sadness, surprise and neutral.

The lab controlled MMI database [87, 88] consists of static images and videos of frontal or profile faces. The image sequences were annotated with six basic emotions and facial muscle actions through Facial Action Coding System (FACS) [89] Action Units, starting each one with a neutral expression, reaching an expression nearly the middle, returning to the neutral expression at the end.

The Radboud Faces Database (RaFD) [90] annotated with eight expressions: anger, contempt, disgust, fear, happiness, sadness, surprise and neutral, is a controlled condition dataset. It is composed of images from 67 subjects in three distinct directions, front, left and right.

In the majority of the publicly available datasets, face expressions are labeled with happiness, sadness, anger, surprise, disgust, and fear, being these defined by Ekman and Friesen [91] in 1971 as the six basic cross-cultural emotions. Normally, alongside these six basic emotions, neutral and/or contempt emotions are also annotated [77]. Recent researches defend that these pre-established emotions cannot be considered universal [92], this being probably one of the strongest reasons why emotions recognition still faces significant challenges.

2.3.2 Algorithms

Facial expression recognition systems are generally composed of three main steps, Face Detection and Face Alignment, Feature Extraction, and Facial Expression Classification [93], [94]. A facial expression recognition systems model is represented in Figure 2.6.

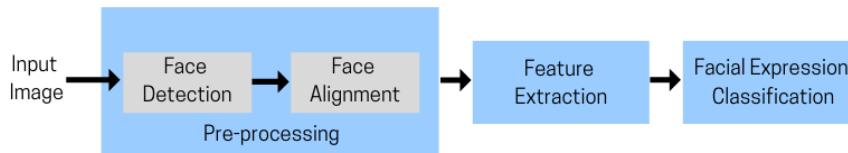


Figure 2.6: Usual Facial Expression Recognition System Diagram of Blocks.

Feature extraction aims to create and identify the most effective basic representations from each facial expression to be identified; hence, there is a large variance between distinct categories and a low intra-categories difference [94]. Subsequently, the facial expression classifier associates the extracted features into the respective emotion category through the corresponding classification mechanism.

Traditional facial expression features extraction techniques, which usually operated with hand-crafted features, can be classified into two distinct categories: geometric-based methods, appearance-based methods, and multi-feature fusion methods [94].

Geometric features aim to characterize the face shape by measuring geometric properties, like distances or deformations. As so, geometric-based methods are based on facial landmarks and/or components location followed by the extraction of geometric features through them. Early proposed Facial Action Coding System (FACS) [89] code Action Units (AU), which are used to define the contraction of one or more facial muscles, representing a specific facial movement, such as an eyebrow lift or a nose wrinkle. Facial Action Units (FAUs) methods are one of the most common geometric-based systems, either using the preliminary FACS approach, which only defines 26 AUs or a large set of AUs. Another typically geometric-based methods are the Active Shape Model (ASM) [95] approaches. ASM extracts geometric features through facial feature points. As an extension of ASM, Active Appearance Model (AAM) [96] combines both geometric and texture information to model a global geometrical face representation, to analyze and identify the emotion.

Appearance-based methods assume that a facial expression requires a local change in texture, using textural information by analyzing the intensity values of the pixels. Example of these approaches is the employment of filters, such as LBP [97], LBP on three orthogonal planes (LBP - TOP) [98], Gabor filters [99], [100], Local Phase Quantization [101], in the whole face or in particular face regions to encode the texture.

Multi-feature fusion methods combine various features to more effectively recognize facial expressions. For example, in [102] Gabor and LBP are combined, where the error rate was minimized. Also, in [103], SIFT and PHOG are used to determine local texture and shape information.

Aiming to avoid the hand-crafted features extraction needed by traditional methods, which can be considered an exhaustive task, and driven by the tremendous advances in GPU technology allied with the appearance of large enough training data, deep learning approaches began to appear, reaching outstanding performances.

Deep learning is based on a hierarchical architecture, through which the input data is subjected to various nonlinear transformations. As so, the feature representation in each layer becomes more abstract, attempting to obtain high-level representations to properly train the network with essential characteristics extracted from the data. Deep learning approaches for facial expression recognition can be characterized into two different based methods: Deep Belief Network (DBN) and CNN [94].

DBN is a graphical model composed of multiple layers, being the input data received in the lowest layer [104, 105, 106]. The top two layers have indirect connections and the remaining units in the higher layers, hidden units, are trained to learn the dependencies between the units in the adjacent lower layers. As so, DBN is pre-trained aiming to reconstruct the input data and further fine-tuned, improving the accuracy of the model. An example is Liu et al. [107] proposal, a boosted DBN (BDBN) consisting of an iterative process of feature representation, feature selection, and classifier training, in a unified loop framework.

From the deep learning methods, CNN is the most commonly used approach, having had significant improvements for facial expression recognition. Over the years, several CNN models were presented. Ng et al. [94] proposed a two-stage supervised fine-tuning process, being first performed on relevant FER datasets, followed by a fine-tuning based on the target dataset, to improve models adaptation to a specific dataset. A combination of multiple deep CNNs is proposed by Yu and Zhang [108], in which each CNN is pre-trained in a larger dataset and subsequently fine-tuned on the target dataset, SFEW 2.0.

A more recent approach by Arriaga et al. [109] presents a general building framework to design real-time CNNs to identify the gender and the emotion simultaneously in real-time. The proposed models aimed to reduce the number of parameters learned in the CNN, in order to make the implementation possible in a real-time system. Alizadeh and Fazel [110] proposed a CNN for FER, by training it from scratch with different depths using raw pixel data and Histogram of Oriented Gradients (HOG) features. The results showed that the combination of both features did not improve the model accuracy, concluding that CNNs are capable of intrinsically learn the key facial features using only raw pixel data. In [93], a loss function is defined in order to model informative local facial regions and expression recognition creating a model able to learn facial relevance maps and expression-specific features with the purpose of obtaining a more precise recognition.

More complete reviews in facial expression recognition can be in [77], [94] and [111].

The CNN became the most popular approach for facial expression recognition [109, 110, 93], due to their successes and to the fact that they represent a significant advantage: CNNs can intrinsically learn features, avoiding the hand-crafted extraction. The main weakness in their study is the reduced number of significantly large publicly available FER datasets, as an extensive amount of data is essential to properly train the CNNs to avoid overfitting. Additionally, it is worthy to note the extreme subjectivity implicit to this area, once both face detection and facial expression recognition rely on aspects such as age, gender, ethnic and level of expressiveness [77]. For the aforementioned reasons and regardless of all advances, the recognition of emotions remains a challenge.

2.4 Discussion

Over the years, several algorithms were proposed in the literature to performed both face detection and recognition, and additionally to the recognition of facial expressions. From this, it is concluded that innovative machine learning approaches achieved outperforming results in the resolution of the three tasks. These techniques, more explicitly when concerning deep network models, require a large amount of face data samples. On top of these datasets, there is a need to annotate some facial features, such as with key points or face bounding boxes. For both face recognition and facial expression recognition, an additional effort is needed, labeling the individuals' identity or the represented facial expression. Typically performed manually, both collection and annotation processes are lengthy and labor-intensive tasks. In terms of performance of face recognition systems,

and focusing on the scenario in which this dissertation emerges, the necessity of a vast number of face images increases, normally being above the thousands. Also, since it is intended for the recognition of specific personalities, none of the aforementioned public datasets could have been used, thus being demanded the construction of a proper dataset.

Chapter 3

Problem Characterization and Methodology

3.1 Problem Definition

The development of effective and successful face detection and recognition approaches has been a recurrent research topic in the scientific community. As stated in the Introduction, both the technological advances and contemporary concerns have raised the demand for face detection and recognition technologies to be used. Several proposals have been put forward, and being the face detection an inherent step in a face recognition system, from the latter it can be distinguished two major categories: i) one that uses a classifier, (e.g. through convolutional neural networks) to detect and classify the ROI in the image which contains a specific person's face (**classification**); and ii) other that requires that face detection is firstly performed and then confronts the resulting ROI with a labeled face image and decides if they belong to the same person (**verification**). From the research on these two types of strategies, it becomes evident that the first requires a large number of input images for each individual intended to be recognized and that the second presupposes an initial step of face detection that might influence the final identification result.

Despite the very significant advances conquered in the fields of face detection and recognition, questions about the most efficient method to perform these processes are not yet fully resolved, due to several factors which hinder the processes, such as the admissible distribution range of facial parts (eyes, nose, mouth, chin, etc.) and the wide diversity of human faces (related to age, gender or race). Face detection and recognition become increasingly difficult in uncontrolled scenarios, where other restrictions may arise with more frequency, namely the presence of occlusions (e.g., due to sunglasses or hands), the diversity of head poses and the low-quality of some images.

Machine learning methods and, more specifically, deep learning techniques, have been providing very promising outcomes, but at the expense of heavy requirements in terms of training data. This is exacerbated when the goal is to identify subjects in unconstrained scenarios, such as in broadcasting or journalism content, since several constraints (e.g., budget or limited human

resources) may make it unfeasible to allocate all the desired means for image collection and face annotation.

This project aims to address the immediately above-mentioned problem, by defining and implementing a method that enables the reduction of the required data samples, and consequently minimizes the associated effort, making only a slight trade-off with the performance.

3.2 Proposed Solution

This dissertation studies state-of-the-art methods for the two mentioned strategies of face recognition, to understand their advantages and shortcomings, but especially to evaluate how the reduction of the dataset and the previous face detection needed phase can affect the results. The ultimate goal of the investigation is to propose a solution for face detection and recognition with an adequate trade-off between accurate recognition rates and the required number of data. The resulting solution is intended to be used in a content annotation system to automatically associate metadata with multimedia content. To achieve the proposed goals, a phased approach was followed.

First, as an essential step, an appropriate dataset was prepared. It was composed of images captured in real scenarios and of the corresponding manual annotations. Different levels and types of noise were added to each annotated bounding box (ROI of the image containing only the face), creating distinct versions of the dataset. Concurrently, the Dlib's CNN Face Detector was applied to the originally collected data, in order to obtain another dataset version. All these versions were generated with the intent of analyzing the effect of different factors on a verification based strategy.

Secondly, the implications of a classification strategy, using a convolutional neural network, were analyzed varying the number of training samples. The use of this type of technique implied a large number of images for each known individual's identity (i.e., for each class), typically several hundred or thousands of images. Given that this type of requirement is incompatible with the target scenarios, transfer learning was used as a first step to reduce the number of required images. Following, the effect of reducing the training set was studied, even when increasing the number of training iterations. This was intended to help understand the tendency and impact of using limited available data.

Lastly, the final step consisted of evaluating the possible combination of both strategies aiming to overcome individual weaknesses while, if possible, reducing even further the data needed. It was studied the impact of the training dataset, in terms of its dimension, to identify a low amount of training data and yet a good face detection and recognition.

The envisioned application scenario was derived from the INESC TEC project CHIC, where this dissertation was inserted. It consists of the application of face recognition algorithms to audiovisual broadcasting content, whose outcomes are to be used in content annotation, promoting content reuse, management, and personalization. CHIC has a special focus on regional and local news, implying the need to recognize famous as well as less known individuals, the latter meaning that it may be more difficult to collect adequate training data. To further clarify this problem, a use case example follows.

Joaquim is a Portuguese reporter covering regional news. Preparing for the upcoming elections for regional bodies, Joaquim is required to prepare for the analysis of audiovisual content from multiple sources, past, present, and future. To achieve this goal, he needs to configure an automatic annotation system to recognize some of the candidates for Vila Real Mayor, including the front runner: Rui Santos. Joaquim starts by searching the web for photos of the candidates, collecting a number of them. In addition, he must also annotate them by marking the appropriate regions in the image and labeling them. A reporter's time is very short and costly, hence he needs to be as efficient as possible. Fortunately, Joaquim has at his disposal a tool powered by machine (deep) learning algorithms that can process images and videos to detect and recognize faces, and that minimizes the number of training samples, i.e., the number of images to be collected and annotated.

3.3 Methodology

In this section, a high-level description of the methodology used in the development of each proposed method is presented. Furthermore, a more detailed explanation of the state-of-the-art used architectures is provided. In addition to these architectures, the OpenCV libraries were also used for the dataset preparation. All the algorithms were implemented in the programming language Python and tested on desktops with GPU.

3.3.1 Verification Based Method

The face recognition process in the Verification Based Method was performed by the comparison of an input face image and a set of face samples of known individuals to the system. Consequently, the system required a pre-processing phase in which the face displayed in the input image was detected through the **Dlib's CNN Face Detector**. The posterior comparison was made by the **FaceNet** siamese network, resulting in a distance that directly represented the similarity between two faces. Thereby, the output of this system is an ordered vector of Euclidean distances, in which the first position represents the smaller distance obtained between the input face and the most similar one from the known subjects. The input face image is then classified with the label which corresponds to this first position, $label_V$, being the correspondent Euclidean distance the associated *distance score*. Verification Based Method system has the following Figure 3.1 model structure.

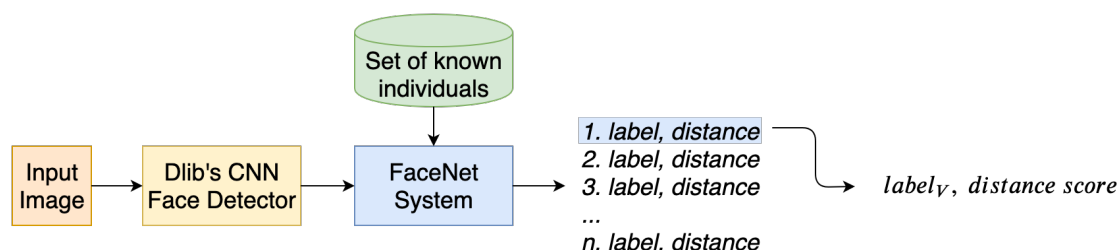


Figure 3.1: Verification Based Method system model structure.

Dlib's CNN Face Detector

Dlib is a C++ cross-platform software library that provides machine learning algorithms and tools to create real-world applicable complex software. Initial released in 2002 and originally written in C++, Dlib has valid and easy to implement Python bindings.

The used Dlib's Python example CNN Face Detector ¹ uses a Max-Margin Object Detection (MMOD) model with CNN based features. Having a simple training process, this implements a model that was pre-trained with a manually labeled dataset, composed of images from ImageNet, AFLW, Pascal VOC, VGGFace, WIDER, and Face Scrub datasets.

This face detector is capable of detecting faces in almost every angle and is robust in detecting faces with occlusions. Meant to be executed on GPU, its performance depends on the available memory, which often requiring the resizing of the images when these are too heavy. In terms of resolution, for example, an 8MPx image takes about 10GB of RAM. As the experiments were performed in a computer with significantly less RAM, it was required to resize the input images, concerning their aspect ratio and resolution (width $\times$ height). This might have had a negative impact, once the non-detection of faces in some images could have been due to their resizing.

Additionally, Dlib CNN Face Detector is not able to detect small faces, once the MMOD model was pre-trained for a minimum face size of 80×80 pixels. From a posterior analysis of its performance, it was observed that it produced small bounding boxes and regularly clip parts of the forehead and chin. An example of the Dlib's CNN Face Detector resulting bounding boxes is shown in Figure 3.2.

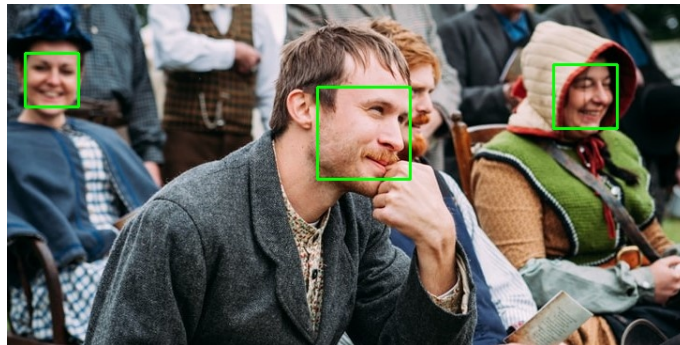


Figure 3.2: Dlib's CNN Face Detector performance example.

¹Available in http://dlib.net/cnn_face_detector.py.html

FaceNet System

FaceNet [54] system was proposed in 2015 by Google. This system transforms the input face images into a compact Euclidean space, where the distances between them directly correspond to a measure of their face similarity. To do so, the FaceNet makes use of a siamese network that is composed of two identical fully connected CNNs. For the face recognition process, it is given to each one of the networks a face image as input, which is subsequently encoding into a 128-dimensional space representation followed by an L2 normalization resulting in a face embedding. Therefore, FaceNet trains its output to be a compact 128-dimension embedding, i.e., each face is encoded in 128-bytes, using a triplet-based loss function based on LMNN [112]. The system output is the Euclidean distance between the two faces embeddings, as shown in Figure 3.3, having the network been pre-trained so that the more similar the faces, smaller is the distance.

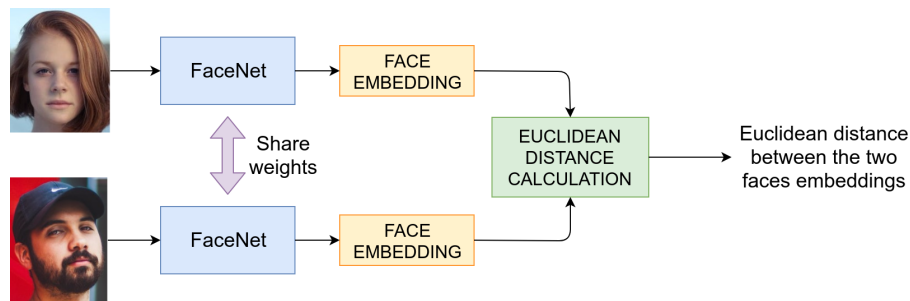


Figure 3.3: FaceNet face recognition process.

The train of the model was conducted giving to the network a three data samples - an *anchor* image, a *positive* and a *negative*. Representing the *anchor* a given individual's face, the *positive* sample was another image of that individual and the *negative* the face image of another subject. Triplet-based loss function during the training minimizes the distance between the anchor and the positive and maximizes the distance between the anchor and the negative, as illustrated in Figure 3.4. Consequently, the facial recognition verification through FaceNet is performed according to a certain threshold value: below that value, the two input face images are considered from the same person, and above as belonging to different people.

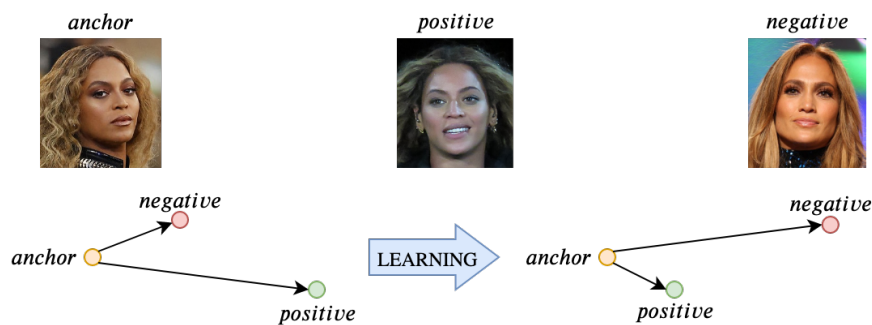


Figure 3.4: Triplet loss-based training process.

3.3.2 Classification Method

The Classification Method system was implemented using the **Tensorflow Object Detection API**. The TensorFlow Object Detection API ² is an open source framework that simplifies the construction and the training of object detection models. As the name suggests, this is implemented on top of Tensorflow, which is an open source library to develop and train machine learning models. Due to its ease of use, this API was chosen for the employment of the transfer learning technique in a pre-trained object detection model ³. This is a Faster-RCNN ResNet 101 model trained with the COCO dataset [113], able to detect and classify 90 classes. Transfer learning was used since the available amount of data was not enough to properly train the neural network from scratch. The pre-trained model was then trained with a training set of the individuals to be recognized, resulting in a fine-tuned model able to detect and classify them. Therefore, this model first detects the ROI, which probably contains a face. The outputs are the ROI, the classified identity label for that ROI, $label_F$, and a *confidence score* of that classification, as demonstrated in Figure 3.5.

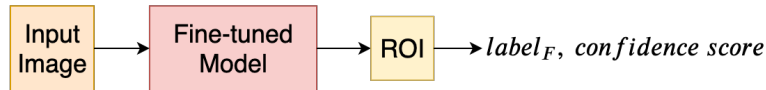


Figure 3.5: Classification Method system model structure.

3.3.3 Fusion Method

The Fusion Method combines the outputs of both the above methods. An input image is fed into the Classification Method, resulting in an ROI with the correspondent attributed $label_C$ and *confidence score*. The ROI is then introduced directly into the FaceNet network, skipping the face detection phase of the Verification Based Method, from which is obtained the identified $label_V$ and the correspondent *distance score*. Further, in the Fusion model, both methods outputs are combined according to different parameters, resulting in a final classification, $label_F$. This process is demonstrated in Figure 3.6. The Fusion Method implementation will be discussed in Section 6.4 of Chapter 6 (Results), since some of the fusion parameters emerged from the analysis of the verification and classification methods results.

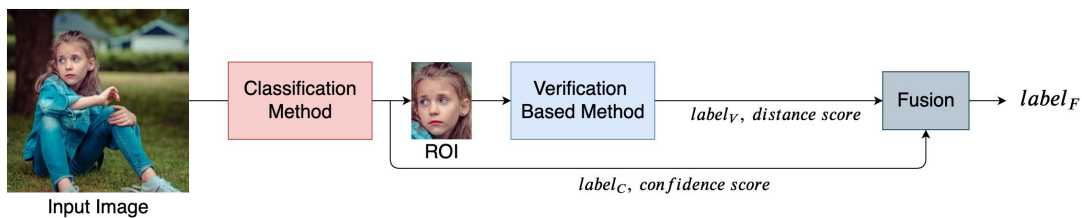


Figure 3.6: Proposed Fusion Method model structure.

²Available in https://github.com/tensorflow/models/tree/master/research/object_detection

³Available in https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md as faster_rcnn_resnet101_coco.

Chapter 4

Datasets Preparation

An essential component in the development of a face recognition system is the dataset. Commonly, this contains the face images of different people along with the respective annotations, which include the label - the person's identity - and the person's face bounding box. The datasets can be composed of images containing the individuals in a scene, possibly with other people, but may also just be cropped images with only the subject's face, excluding the need for bounding box annotations. In the particular case of face/identity verification, two datasets come into play:

- *Gallery Set* ($\mathcal{G}$) - Set of images whose identity is known to the system, having one or more images for the same subject.
- *Probe Set* ($\mathcal{P}$) - Set of images presented to the system for recognition, which may contain images of the individuals of the *Gallery Set* or not.

For the overall goal of this dissertation, it was necessary to prepare different adequate datasets for the assessment of the distinct strategies and of the proposed fusion approach.

4.1 Base Dataset

The Base Dataset is composed of diverse images of four distinct personalities, Donald Trump, Marcelo Rebelo de Sousa, Ricardo Rio, and Rui Moreira captured in real scenarios. The justification for the choice of these personalities derives from the integration of this dissertation on the CHIC project that is being developed by INESC TEC, which is, as introduced in Chapter 3, focused on analyzing strategies for the recognition of local, national and international individuals for automatic annotation of journalistic audiovisual content. As so, the dataset includes well known national and international personalities, but also more local ones, like Rui Moreira and, especially, Ricardo Rio. This is intended to align with the greater difficulty in collecting adequate images for less known people.

For each personality, the dataset has images with different resolutions and these can show the subject by himself or accompanied by one or more people. It also includes images where the individual was posing for the photograph and others where it was captured in unconstrained

circumstances. Therefore, the dataset exhibits several head poses and different illumination conditions, all aspects that might compromise facial recognition. The images were collected from either Google Images, the personality's Facebook or even from video frames.

More details on the Base Dataset of each personality will be given below, regarding aspects like the number of images, the number of persons per image, particular features, etc.

Donald Trump

Donald Trump's - the 45th President of the United States - dataset consists of 549 images, most of which have a resolution equal or superior to 1280×720 (~ 0.9 MPx) or above. It features a high diversity of facial expressions and in 36% of the images, the personality is accompanied by people, although in the majority, the remaining subjects appear as background and blurred. Figure 4.1 presents four examples of Donald Trump's dataset.



Figure 4.1: Donald Trump's dataset images examples.

Marcelo Rebelo de Sousa

The majority of the images in Marcelo Rebelo de Sousa's - President of Portugal at the time of this dissertation - dataset, which comprises a total of 520 images, show him accompanied. This is the set of images where the personality appears with the greatest age range and in which only about 22% of the images show the individual in a frontal pose. Some examples of this dataset are shown in Figure 4.2.



Figure 4.2: Marcelo Rebelo de Sousa's dataset images examples.

Ricardo Rio

Ricardo Rio's - at the time of this dissertation, Mayor of Braga - dataset contains 507 images, being some of these sampled from videos of interviews or election campaigns. Similar to what occurred in Marcelo Rebelo de Sousa's dataset, in about 70% of the images he is accompanied by one or more people. In Figure 4.3 are presented some examples for his dataset.



Figure 4.3: Ricardo Rio's dataset images examples.

Rui Moreira

Composed of 515 images, this dataset has the particularity of having some images where the personality is wearing glasses. As it happened with Donald Trump's dataset, in about 37% of the images, Rui Moreira - current Mayor of Porto - is accompanied by one or more people. A few examples of his dataset are displayed in Figure 4.4.

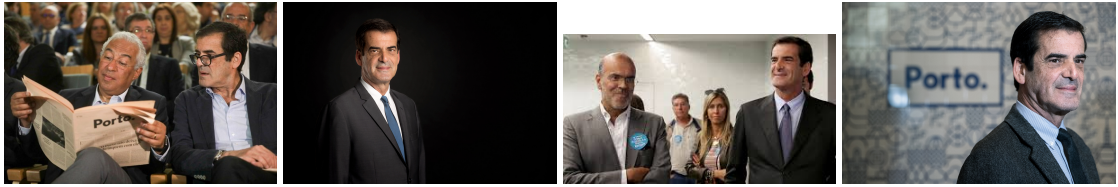


Figure 4.4: Rui Moreira's dataset images examples.

For a more detailed characterization of the Base Dataset, several important aspects were analyzed. Firstly, the number of existing images for each head pose (Frontal, Left, Right, Down, Up) was registered. This is relevant because the comparison can be between faces with different head positions, which will probably influence the final classification. Since this accounting was manual, thus one can argue that the classification for each position may contain a subjective influence. Furthermore, the file size of the images was analyzed, which may have more impact at the computational level and is related to the images' resolutions. This last characteristic was estimated by computing the number of megapixels (MPx), due to a wide range of dimensions (width $\times$ length) of the images. The image resolution is related to its quality, where smaller dimensions can imply a lower resolution of details, which, in turn, might detriment both face detection and recognition [114]. Additionally, the number of images in which the personality is alone was counted, as well as the number of images in which the personality is with one or more people. Figures 4.5, 4.6, 4.7 and 4.8 convey a summary of this analysis for each personality dataset.

It can be observed in Figure 4.5 that, in the majority of the images, the personalities exhibit a side head pose, to the left or the right. In about 22% of the pictures, the individual is in a frontal pose and only in a small percentage of images appear with an up or down head pose.

In Figure 4.6 it is shown that, for Marcelo and Ricardo datasets, most of the images file sizes, around 67%, are considered "Small", while in Trump's Dataset, 76% of images have a file size larger than 100KB. It is also shown that a substantial proportion of Rui's dataset has a file size between 10 KB and 1 MB, of which 41% are considered "Small" and 39% "Medium", being notable that Rui's is, from all the four, the only dataset that has images with a size lower than 10KB.

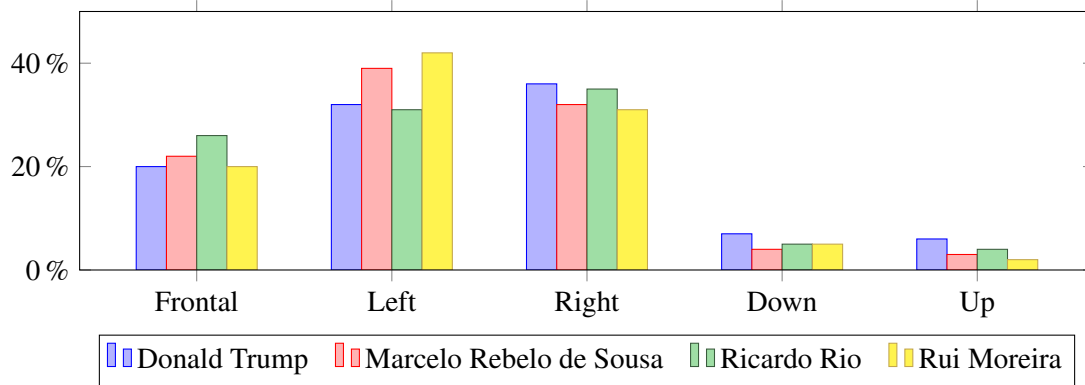


Figure 4.5: Distribution of the different head poses among the Base Dataset.

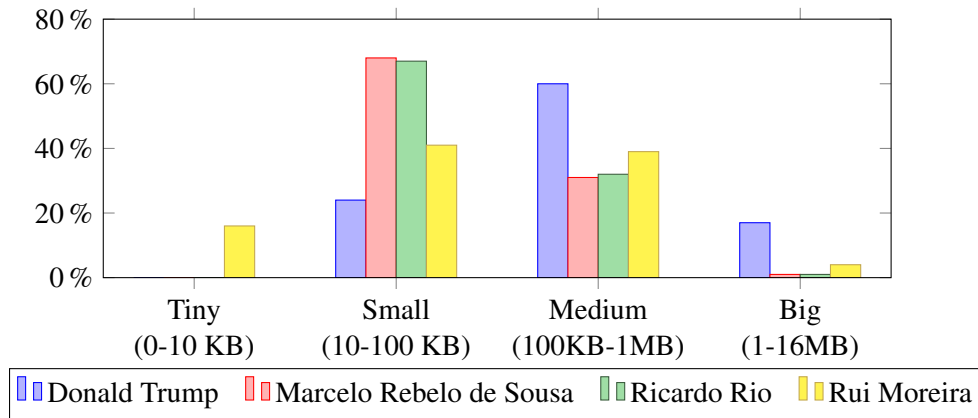


Figure 4.6: Distribution of the different image file sizes among the Base Dataset.

From Figure 4.7, it can be concluded that most of Marcelo's dataset has an image resolution between 0.3 and 0.5 MPx, typically corresponding to images with dimensions between 640×480 and 768×576 . Most of Ricardo's dataset has an image resolution between 0.5 and 0.8 MPx, the last referring to images with dimensions of about 1080×720 . Additionally, it can be estimated that 90% of the Rui's dataset has an image resolution bellow 1.3 MPx, usually with dimensions lower than 1280×1024 . It is also evident that Trump's dataset is by far the one with higher image resolutions, once more than half of his dataset has a resolution superior to 2 MPx, matching to images with dimensions above 1920×1080 .

At last, in Figure 4.8, it can be noticed that both Trump and Rui appear in the majority of its datasets alone, contrary to what happens with Ricardo, who in 50 % of his dataset appears accompanied by more than one person. It should also be noted that Marcelo appears alone almost with the same frequency as he appears accompanied by a person, or by 2 or more.

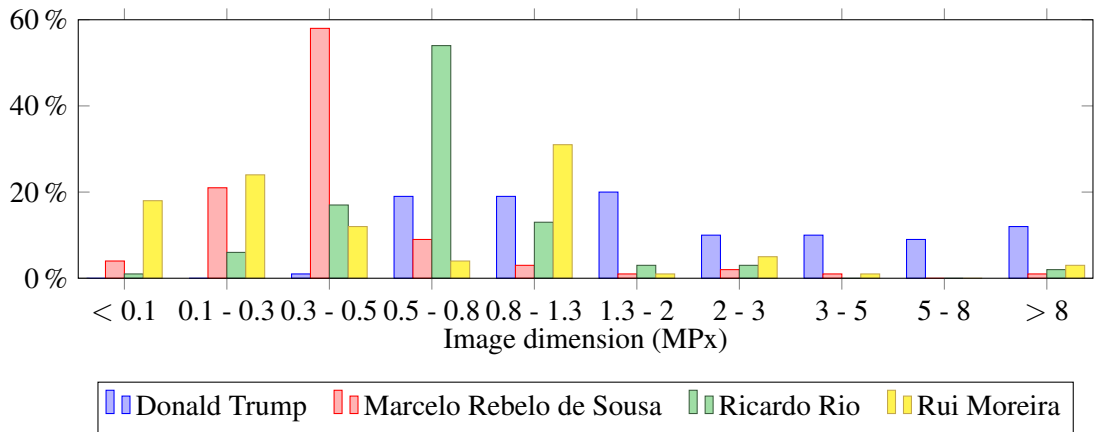


Figure 4.7: Distribution of the different images dimensions among the Base Dataset.

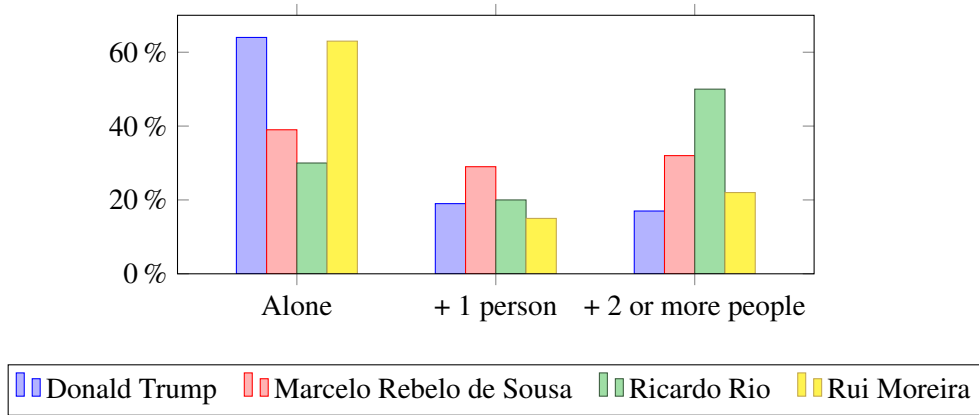


Figure 4.8: Distribution of amount of people per image among the Base Dataset.

4.2 Ground Truth Dataset

Alongside with aforementioned dataset, it was necessary to generate the corresponding annotation of the personality's face. For each image, it was performed a handcrafted annotation in an XML file, taking into consideration the bounding box that included the whole head of the individual, excluding any other person present in the image.

Based on these annotations, an additional Ground Truth Dataset (GT) for each personality was created, where the base images were cropped according to the corresponding annotated bounding box, similar to what it is illustrated in Figure 4.9. This is inline with commonly used datasets for facial recognition, such as LFW [11] or CASIA-WebFace [55], which contains images including larger or narrower head regions, with only the annotation of the person's identity label.

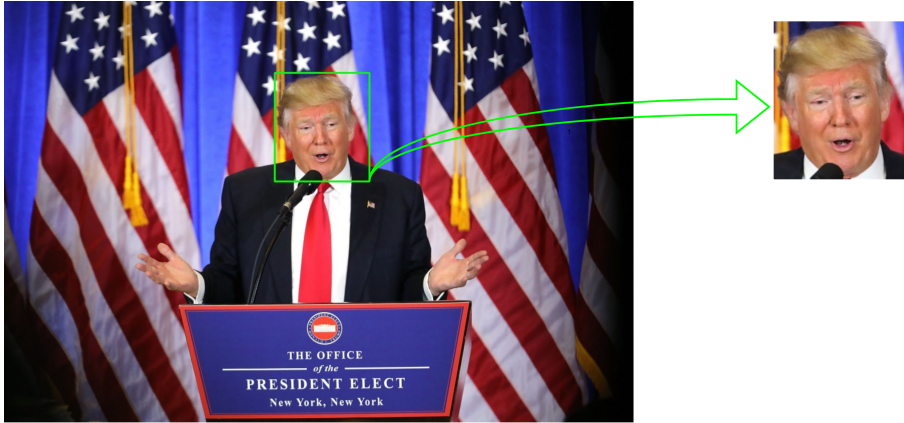


Figure 4.9: Donald Trump's Base Dataset image with the ground truth annotated bounding box, represented in green and its cropped version.

4.3 Noisy Dataset

In order to analyze the impact of possible noise emanating from facial detection into the face recognition pipeline, random modifications were artificially added to the annotated bounding boxes before the generation of the cropped versions. Those modifications concerned two main types of noise: expansion and shift of the bounding box. The expansion noise could be applied in all directions or just one (left, right, top or bottom) and shift noise could be introduced in any direction.

Having as base the Base Dataset annotations, the different noises were randomly added in diverse percentages (1%, 5%, 10%, 15%, and 20%) to every image bounding box of the personalities' subsets. The quantity, in pixels, of the introduced noise was computed as the applied percentage multiplied by the annotated face bounding box diagonal length, as represented in Figure 4.10. This procedure is presented in Algorithm 1.

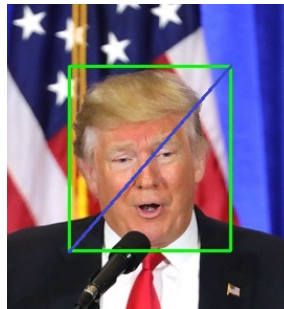


Figure 4.10: Ground truth annotated bounding box diagonal, represented in blue.

As a result of applying Algorithm 1, the Noisy Dataset is composed of the crops of the Base Dataset images by the distorted bounding boxes. Five Noisy Datasets were created for each one of the four personalities, one per percentage of noise introduced. Figure 4.11 shows a few examples of the resulting noisy bounding boxes compared with the ground truth ones.

Algorithm 1 Add noise to Ground Truth Dataset

```

1: for image in Ground Truth Dataset do
2:    $(x_{min}, y_{min}, x_{max}, y_{max}) \leftarrow$  annotated bounding box
3:    $bb\_height \leftarrow x_{max} - x_{min}$ 
4:    $bb\_width \leftarrow y_{max} - y_{min}$ 
5:    $bb\_diagonal \leftarrow \sqrt{bb\_height^2 + bb\_width^2}$ 
6:    $margin \leftarrow percentage \times bb\_diagonal$   $\triangleright percentage \in \{1\%, 5\%, 10\%, 15\%, 20\%\}$ 
7:    $noise \leftarrow$  random type of noise
8:    $noisy\_image \leftarrow \text{ApplyNoise}(\text{image}, x_{min}, y_{min}, x_{max}, y_{max}, noise)$ 
9: end for

```

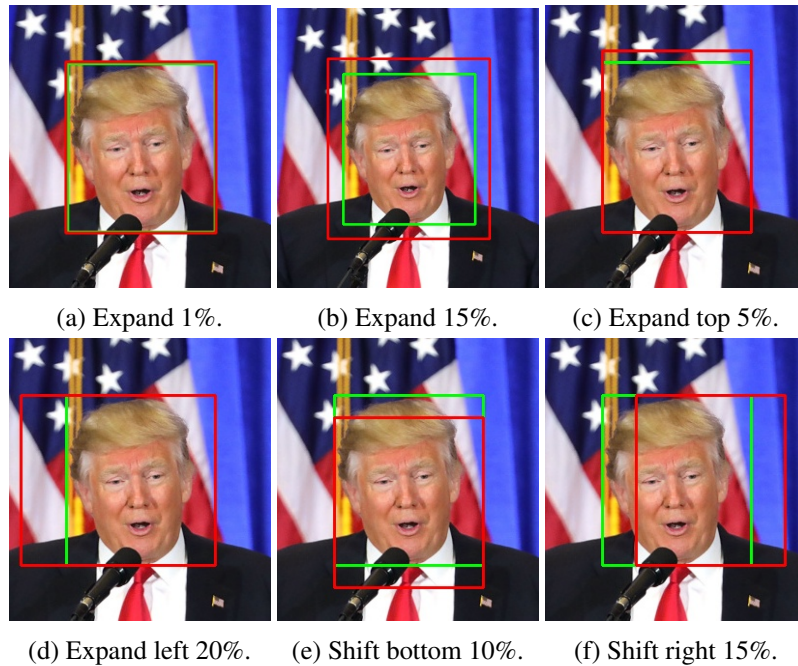


Figure 4.11: Noisy Dataset modified bounding box examples, represented in red. Ground truth bounding boxes represented in green.

4.4 Face Detection Dataset

To evaluate the entire Verification Based System, the face detection phase was implemented using the Dlib's CNN Face Detector ¹ on the Base Dataset, in order to analyze how the system would perform with a state-of-the-art face detector and the impact that this could bring. As a result, the Face Detection Dataset was created, compiling the crops of every detected face in each image. An example of the bounding box resulting from face detection is shown in Figure 4.12. Consequently, this set of data includes not only the faces of the personalities but also those of the other people in the images.

¹ Available in http://dlib.net/cnn_face_detector.py.html

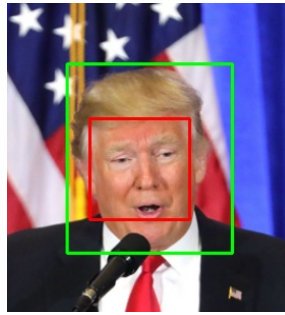


Figure 4.12: Bounding box resulting of the Dlib's CNN Face Detector represented in red. Ground truth annotated bounding box represented in green.

At this point, a quick inspection of Dlib's CNN Face Detector results showed that it produces a significant small crop of the detected faces when compared to the head annotated bounding boxes, as demonstrated in Figure 4.12. Hence, an alternative version of this dataset was generated through the expansion of the detected faces bounding boxes. This expansion was performed in all directions, with a margin of 25% of the diagonal length of the bounding box, thus generating a second dataset based on Dlib's face detection. Designated as Extended Face Detection Dataset, this second dataset presents results more similar to the ground truth annotations (Figure 4.13) and was defined to analyze the impact of Dlib's more reduced detected face regions.

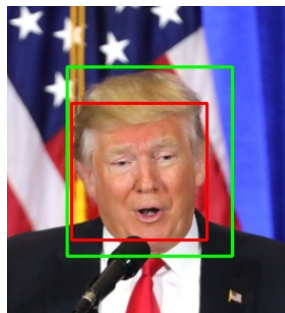


Figure 4.13: Bounding box resulting of the Dlib's CNN Face Detector expanded by 25% represented in red. Ground truth annotated bounding box represented in green.

4.5 Gallery Set

With the evaluation process came the need to collect a dataset with the individuals known to the system, with which the identity verification would be processed. This dataset is composed of five images of twelve distinct personalities, including the four belonging to the Base Dataset. These five images include a front view photo and two images per side view of the person's head. All images have been cropped in such a way that only the personality head appears. Some of the images belonging to this dataset are presented in the following Figure 4.14, depicting the represented personalities.



Figure 4.14: *Gallery Set* examples.

For evaluation purposes, the results were generated for the following three variants of the *Gallery Set*:

- **Frontal Gallery Set (FGS)** - containing only the front view images of the twelve personalities;
- **3 Images Gallery Set (3GS)** - containing three images for each personality, the one of the front view and one per side view (left and right);
- **5 Images Gallery Set (5GS)** - containing all five images per personality.

Chapter 5

Assessment Strategy

Two essential elements must be considered to achieve an adequate evaluation of the results: the datasets - images with which the results were produced, described in Chapter 4 - and the metrics, which evaluate the performance of the proposed models.

As stated in Chapter 3, the goal of this dissertation is to study face recognition techniques and research a strategy aiming at a proper trade-off between performance and required training data. From this main goal, more explicit evaluation objectives have emerged and, hence, the need to collect the various *Gallery* and *Probe Sets* described in Chapter 4. The performance of the proposed face recognition strategies was evaluated through different metrics.

This chapter begins by explaining the assessment strategy, basing the purpose of constructed datasets and relating them to metrics, that are further described.

5.1 Overall Assessment Strategy

The first required step was the collection of a [Base Dataset](#) with subsets for each one of the four personalities, having each subset a reduced number of images (about 500) that attempt to represent the individuals in real-life scenarios, as described in Chapter 4. This dataset has images with different aspects that might hinder face detection and/or recognition, such as illumination conditions, head pose variation, image resolution and even the presence of more than one person, as shown in Figure 5.1. This last one is especially important for the proposed face recognition methods, once it turns the identification task harder and computationally more costly than it would be if only one face/person was exhibited in the image.



Figure 5.1: Examples of personalities' images under different conditions.

To complement these images, the personality identity (label) and the bounding box annotations of the personality's face were hand-crafted in XML files. These annotations allow the knowledge of the true personality identity and also its face location in the image, and are therefore used as a basis for establishing a comparison with the results of the other datasets. It is important to emphasize that the annotation and labeling of the bounding boxes was a hard and time-consuming manual task. Although these were outlined to consider the whole head region, the inherent subjectivity of a manual task and the fatigue during its execution may have led to few less consistent annotations, where the bounding box considers less or more of the head region.

As referred in the Chapter 3, two strategies for facial recognition were assessed: **verification/authentication** and **identification/recognition**. These distinct recognition approaches have an impact on the type of assessment, there being different evaluation metrics for each one. These will be discussed in Section 5.2 and 5.3.

5.2 Verification Based Method

In a typical verification method, a *probe* - face image whose identity is to be verified - is presented to the system claiming to be a certain person. The system then computes the similarity between the *probe* and that person's image stored in the *Gallery Set*, to determine whether the two images correspond to the same person or not. Therefore, verification systems are generally characterized as a 1:1 matching system. Conversely, the proposed method computes the similarity between the *probe* and all images of the *Gallery Set*, i.e., the *probe* verification is performed with the entire *Gallery Set*, thus being a 1:N matching. This is consistent with the proposed application scenario, where a given stream is to be processed with knowledge if one or more of the personalities are present.

The proposed method is based on a siamese network, more specifically, on FaceNet [54] siamese network. In this method, the similarity between a *probe* and an image from the gallery is computed by the Euclidean distance between the two encoding representations (embedding) of the two faces. For that, and assuming that the *Gallery Set* already contains cropped images of the known individuals' faces, it is essential to detect the faces in the input image before feeding it to the siamese network. This is because it is considered that the input image does not contain the face cropped and can contain more than one face. As expected, the smaller this Euclidean distance, the more similar are the faces, and so the verification decision is made by taking into account a certain threshold value, above which the system classifies the faces as belonging to different persons. Since the comparison is executed between the *probe* and the whole *Gallery Set*, the output of the proposed system is a vector of Euclidean distances. The resulting vector is subsequently sorted in ascending order, meaning that the first position of the vector corresponds to the distance between the face of the input image and the most similar face of the gallery, according to the network.

The proposed verification based approach follows the pipeline depicted in Figure 5.2: face detection is performed on an input image; each detected face is provided, one at a time, as an input to the siamese network, being the other input each image from the *Gallery Set*; the result is, then,

the vector of the Euclidean distances. For this strategy's assessment, the **Rank-1** recognition rate [115] is used.

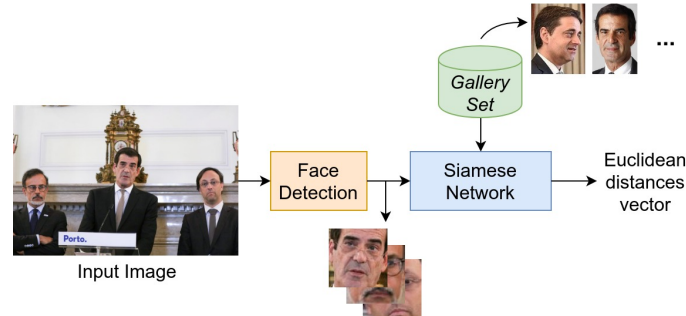


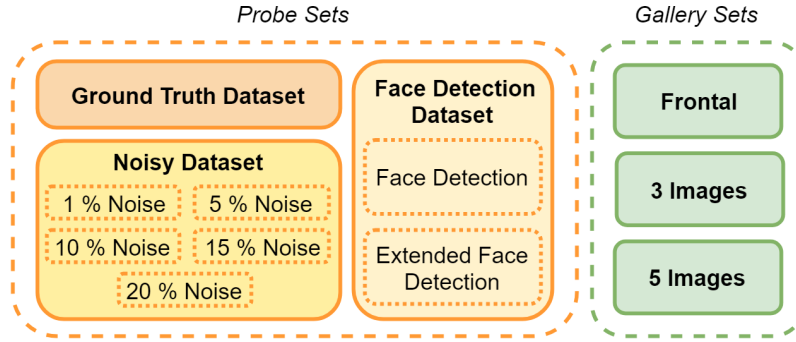
Figure 5.2: Proposed face recognition verification strategy pipeline.

Having this system assembled, it was necessary to collect the *Gallery Set*. This is composed of several images for each known individual and contains also other personalities beyond the four intended to be recognized. See Figure 4.14.

It was also required to collect the *Probe Set*. For a more accurate analysis of the results, three *Probe Sets* were collected, aiming to cover three different evaluation purposes:

- **Ground Truth Dataset** - to evaluate the verification/authentication of the system itself, and for such purpose, the face detection phase is not considered. It derives from the Base Dataset and its bounding box annotations, resulting in a set of cropped images of the personalities' face. These resulting images represent what would be a nearly perfect face detection result and for that reason will be used as the basis of comparison for the other *Probe Sets* results;
- **Noisy Dataset** - To assess the impact of the face detection phase, which represents a crucial step in the pipeline, since the comparison is made between the *Gallery Set* and the detected face bounding box. An invalid or misplaced detection or even a non-detection of a face can imply an incorrect and unfairly verification. Therefore, it is also important to analyze how these possible levels of noises associated with the face detection influence the result of the verification;
- **Face Detection Dataset** - To evaluate the overall system performance, the face detection phase was implemented using the Dlib's Face Detector on the Base Dataset.

To sum up, for the assessment of the verification strategy, each image of the three generated datasets - Ground Truth Dataset, Noisy Dataset, and Face Detection Dataset - were compared with the *Gallery Set*, which resulted in vectors of Euclidean distances for each one of those. Figure 5.3 illustrates the organization of the different datasets for both *Probe* and *Gallery Sets*.

Figure 5.3: *Probe and Galley Sets.*

5.3 Classification Method

For the identification/recognition strategy, a pre-trained object detection model was used as a starting model for the recognition of the intended four individuals. This machine learning method is known as transfer learning [116]. For this process, 500 images of each individual were selected from the Base Dataset, to compose a more balanced train set, to achieve a fair classification model for the recognition of the four personalities. To this strategy assessment, Cross Validation was applied, which is a model validation technique to evaluate the generalization capacity of a model in an independent dataset. More specifically, it was used the **K-Fold Cross Validation**, commonly used to evaluate machine learning methods with limited data.

5.4 Metrics

To evaluate the face recognition verification method of the Ground Truth, Noisy and Face Detection Datasets on the three *Gallery Sets*, the resulting Euclidean distances vectors were sorted in ascending order, in other words, ordered from the personality of gallery more similar to the *probe* to the most distinct.

Adopting similar designations used by Phillips et. al [115], this comparison between a *probe* p_j and each gallery image g_i produces the Euclidean distance d_{ij} , which will be referred to as *distance score* and whose negation could be used as similarity score. Assuming that the identity of a *probe*, $id(p_j)$, is present in the *Gallery Set* with one or more images, a correct match is the one where $id(p_j) = id(g_*)$, being g_* one of the gallery images correspondent to that identity. If d_{*j} is the n th smallest one, the probe p_j rank is n , given by $rank(p_j)$. A practical example of these designations is demonstrated in Figure 5.4.

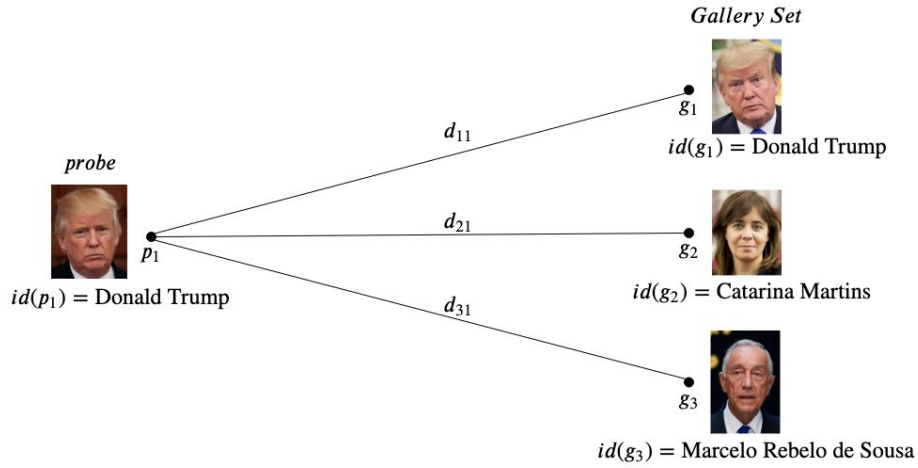


Figure 5.4: Pratical example of the used designations.

The **Rank-1** recognition rate is based on the percentage of probe searches returning d_{*j} in the first rank-ordered result, i.e the $\text{rank}(p_j) = 1$. This first rank-ordered result is also referred as the *top match*. A *probe* is correctly verified if $\text{rank}(p_j) = 1$ and its *distance score* is below a operational threshold value τ . For evaluation purposes, two Rank-1 recognition rates were computed, considering or not the operational threshold. These are given by the Equations 5.1 and 5.2, respectively.

$$\text{Rank-1}R(\tau) = \frac{|\{p_j : \text{rank}(p_j) = 1 \text{ and } d_{*j} < \tau\}|}{|\mathcal{P}|} \quad (5.1)$$

$$\text{Rank-1}R = \frac{|\{p_j : \text{rank}(p_j) = 1\}|}{|\mathcal{P}|} \quad (5.2)$$

Additionally, the False Acceptance Rate (FAR) was computed using Equation 5.3 as the fraction of wrong matches, whose distance d_{ij} is inferior to the threshold value with regards of the total number of wrong attempts:

$$\text{FAR}(\tau) = \frac{|\{d_{ij} : \text{id}(p_j) \neq \text{id}(g_j) \text{ and } d_{ij} < \tau\}|}{|\mathcal{P}| |\mathcal{G}| - 1} \quad (5.3)$$

For the classification strategy, the assessment was performed using **5-Folds Cross-Validation**. For this task, from the entire Classification Dataset - with 500 Base Dataset images per individual - 50 images of each were randomly selected for the test set, totaling 200 images. The remaining 450 images per personality were randomly split into five folds. Each of the folds containing 90 images. Then, the transfer learning process was conducted using four folds - 360 images per personality - to fine-tune the pre-trained model, and the fifth was used to validate the model. This iterative process was repeated until each folder had been used as validation set. The performance metrics **mean average precision (mAP)** and **mean average recall (mAR)** were computed for the five iterations with the validation set during the training and subsequently with the test set, and used to evaluate the fine-tuned model. The average precision (AP) is the average value of the precision

across eleven recall values, from 0 to 1. Consequently, the mAP was calculated by the mean AP over the four classes and overall IoU threshold values, from 0.5 to 0.95 with a step of 0.05. As for the average recall (AR), this is maximum recall given a fixed number of detections per image averaged over all IoU threshold values.

For a more correct comparison of the performance of the Verification Based and Classification Methods, and for further analysis of the proposed Fusion Method - which combines the two previous approaches - it was used the metric **F-measure**, also known as *F-Score* or *F1-Score*. This metric, given by the Equation 5.6, considers both Precision and Recall of the classification process to compute a single performance score, representing the harmonic mean between both. The precision metric, defined by the equation 5.4, represents the percentage of positive detections/identifications that were actually correct, according to an IoU (Intersection over Union) overlap threshold operational threshold, i.e., the percentage of total relevant results correctly detected and classified. The recall metric, defined by the equation 5.5, represents the percentage of face images in the *Gallery Set* that were correctly detected and classified, also according to an IoU overlap threshold; in other words, the percentage of positive detections/identifications that were considered as relevant results.

$$Precision = \frac{\sum TruePositives}{\sum TruePositives + \sum FalsePositives} \quad (5.4)$$

$$Recall = \frac{\sum TruePositives}{\sum TruePositives + \sum FalseNegatives} \quad (5.5)$$

The Intersection over Union between the ground truth bounding box and the bounding box obtained from the face detection or from the Classification Method is computed as expressed in the following Figure 5.5.

$$IoU = \frac{Overlap\ Area}{Union\ Area}$$


Figure 5.5: Intersection over Union computation.

$$F - measure = 2 \frac{Precision \times Recall}{Precision + Recall} \quad (5.6)$$

Chapter 6

Results

This chapter presents and analyses all the results obtained by the Assessment Strategy, aforementioned in Chapter 5. Once the evaluation was carried out for the Verification Based Method and the Classification Method, the results of these are listed in two separate sections, being the third section dedicated to their comparative analysis. These two methods experiments attempted to detect and classify faces of the four intended personalities. Therefore, the results are displayed individually, and their average was computed to assess both methods as an all. The performance of the Fusion Method is analyzed in the last section of this chapter.

6.1 Verification Based Method

6.1.1 Experimentation strategy

To evaluate the Verification Based Method, whose assessment strategy is reported in Section 5.2, several experiments were performed. The aim was to obtain an adequate set of results for the evaluation of different aspects that might influence the efficiency of this face recognition method. The only difference between the executed tests was the dataset chosen as *Probe Set*. For each *Probe Set*, the performance metrics were computed using the three variations of the *Gallery Set*: FGS, 3GS, and 5GS.

In the initial stage of the evaluation process, the definition of a baseline for these experimental results was required. This was obtained from the application of this method to the recognition of the four specific individuals and the most accurate approximation to the ideal results was obtained using the **Ground Truth Dataset** as *Probe Set*. As a result, the obtained performance refers only to this face recognition system, not taking into account the impact of a previous facial detection phase. Therefore, the obtained results with the GT are used as a baseline throughout the analysis of this verification based method.

Two experiments were conducted using the following datasets as *Probe Set*:

- **Noisy Dataset** - composed of five subsets - 1%, 5%, 10%, 15% and 20% - and used to analyze how different levels of noise can influence the face recognition;
- **Face Detection Dataset** - composed of two subsets, one resulting from the application of a face detector on the Base Dataset and the other from a 25% expansion of the latter subset bounding boxes, and used to evaluate the entire system implementing the face detection phase.

The next section examines the question of the operation threshold value, through which the identity of an individual is verified, and asserts its most adequate value. The results of the experiments conducted are further analyzed for the two *Probe Sets*.

6.1.2 Threshold Value

The operation threshold (τ) corresponds to the value upon which a decision about an identity is undertaken. As stated in Chapter 5, for each given *probe* to the verification system, the output is a vector of Euclidean distances, which corresponds to *distance scores* between the *probe* and each image of the *Gallery Set*. The *probe* p_j is recognized as a gallery individual g_i if the *distance score* d_{ij} is below the operation threshold.

Hence, before proceeding to the assessment of the results, it was necessary to study and define the operational threshold value. This analysis was based on the False Acceptance Rate (FAR), with the aim to maximize Rank-1 $R(\tau)$ and minimize $\text{FAR}(\tau)$ and but yet considering that the maximization of the first implies the maximization of latter. Different tests were performed using several threshold values, ranging from 0 to 4, with a step of 0.1, having the *a priori* knowledge that the *distance score* $d \in [0.0, 4.0]$ [54]. Through the obtained results, the receiver operating characteristic (ROC) curve, presented in Figure 6.1, was constructed, by plotting the Rank-1 $R(\tau)$ against the $\text{FAR}(\tau)$ at the various threshold values.

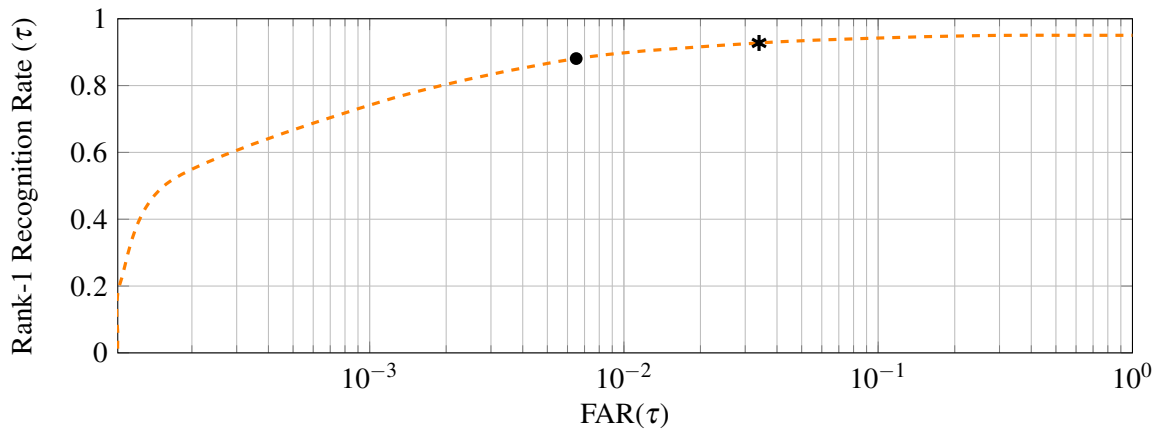


Figure 6.1: ROC curve between Rank-1 $R(\tau)$ and $\text{FAR}(\tau)$.

The optimal maximization of Rank-1 $R(\tau) - \text{FAR}(\tau)$ was obtained for $\tau = 1.1$ (*). However, aspiring for a more reliable recognition and an even lower FAR, the chosen threshold value was 1 (●). In Table 6.1 are displayed the False Acceptance Rates for the Ground Truth Dataset. The presented results for $\tau = 1$ are consistent with reports in the state-of-the-art [117, 118], where FAR values equal to 1% are considered low and reasonable. Hence, in the following sections the value of τ will be set to 1.

Ground Truth Dataset	FAR ($\tau = 1$)		
	FGS	3GS	5GS
Donald Trump	0.75%	0.53%	0.46%
Marcelo Rebelo de Sousa	0.65%	2.61%	2.36%
Ricardo Rio	0.00%	0.03%	0.19%
Rui Moreira	0.02%	0.02%	0.18%
Average	0.35%	0.80%	0.80%

Table 6.1: False Acceptance Rates for the threshold $\tau = 1$.

6.1.3 Results Analysis

6.1.3.1 Frontal Gallery Set

This section presents the results using the Frontal Gallery Set. Hence, each *probe* will be compared with twelve images, from the twelve known personalities. Figures 6.2 and 6.3 show the results for the two *Probe Sets*, Noisy and Face Detection Datasets, respectively, using as baseline the results for the Ground Truth Dataset.

By analyzing Figure 6.2, it can be asserted that, when going through this face recognition process, the four personalities exhibit similar behavior. As expected, all Rank-1 R outperform the Rank-1 $R(\tau = 1)$. This happens due to the fact that the Rank-1 R metric only takes into account the correct *top matches*, not verifying if the corresponding *distance score* is below the threshold value. Thus, a *probe* p_j is always identified as the person in the gallery g_i , with which it has a lower *distance score* d_{ij} . Therefore, this classification is not considered highly reliable, once it does not take into account a threshold value, implying a FAR of 100%.

It can also be noted that, as the artificially added noise increases, the recognition rate decreases. This performance decrease was expected, since the comparison is made with incrementally modified bounding boxes that may leave out face components, such as the chin, the forehead or even the nose in non-frontal pictures. The decline in performance is even more noticeable when applying the threshold value. Once even when the *top match* is correct, the corresponding *distance score* could be higher than the threshold value, and consequently the *probe* not classified with any identity.

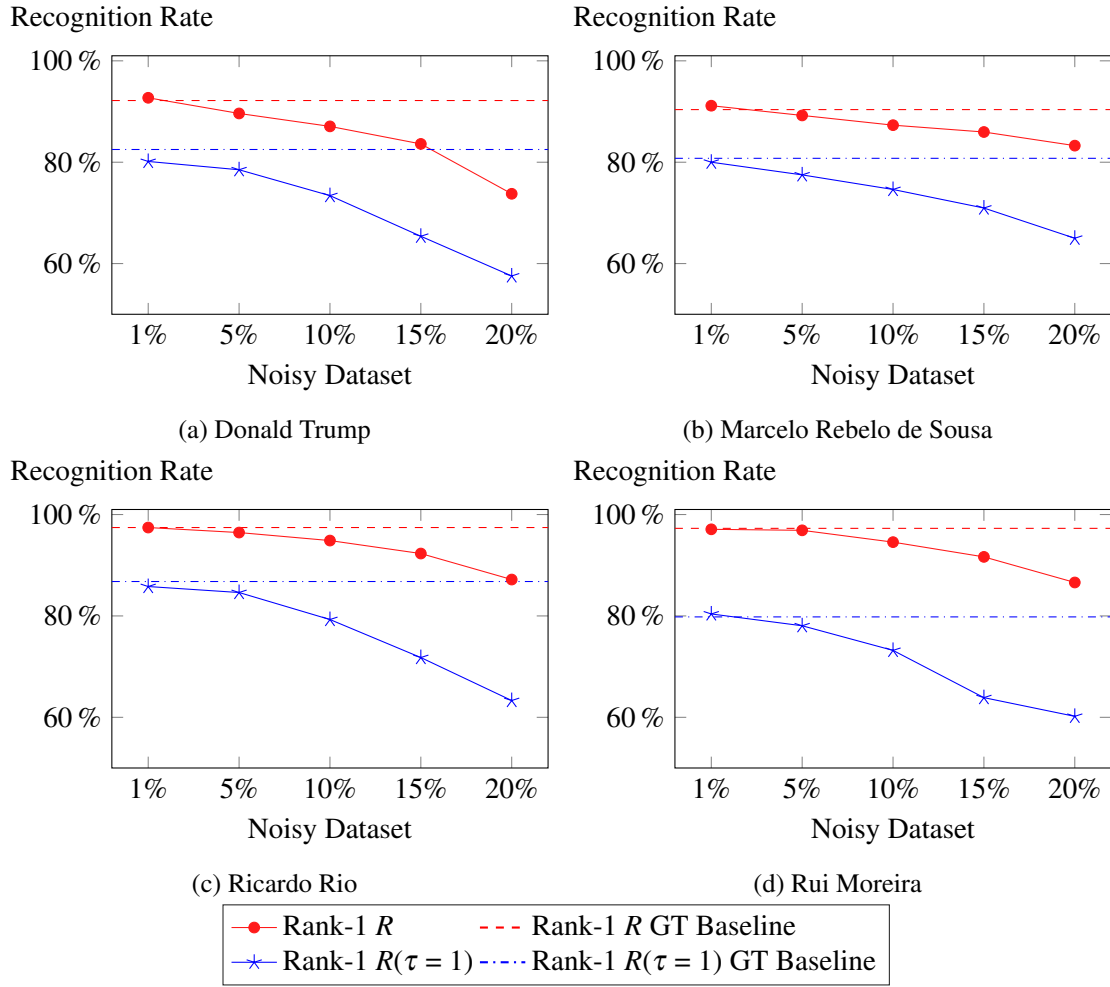


Figure 6.2: Frontal Gallery Set results with Noisy Dataset as *Probe Set*.

By making a comparative analysis between the four personalities' results, it can be concluded that the recognition rates for Donald Trump, both Rank-1 R and Rank-1 $R(\tau = 1)$, are the ones that decrease the most as the percentage of applied noise increases from 1% to 20%. This behavior can be explained by the fact that his dataset has a lower percentage of frontal pictures and it is characterized by a wide range of facial expressions. The comparison with a frontal neutral - with no facial expression - image, characteristic of this *Gallery Set*, might be unfair relative to the comparison executed with the other three personalities, where the diversity of facial expressions is significantly lower.

Marcelo Rebelo de Sousa's dataset displays the lowest Rank-1 R performance, with 90.38%; however, it also exhibits the lowest discrepancy of both Rank-1 R and Rank-1 $R(\tau = 1)$ between the results for 1% and 20% Noisy Datasets.

Both Ricardo Rio and Rui Moreira present the highest performances for Rank-1 R , having a similar decreasing behavior of about 10%. Nonetheless, Rui exhibits the highest gap between the Rank-1 R and the Rank-1 $R(\tau = 1)$. After a quick inspection, it was concluded that this was because the Rui Moreira's frontal gallery photo was taken under more contrasting illumination

conditions when compared with his entire dataset, leading to higher *distance scores* and, consequently, to a lower Rank-1 $R(\tau = 1)$.

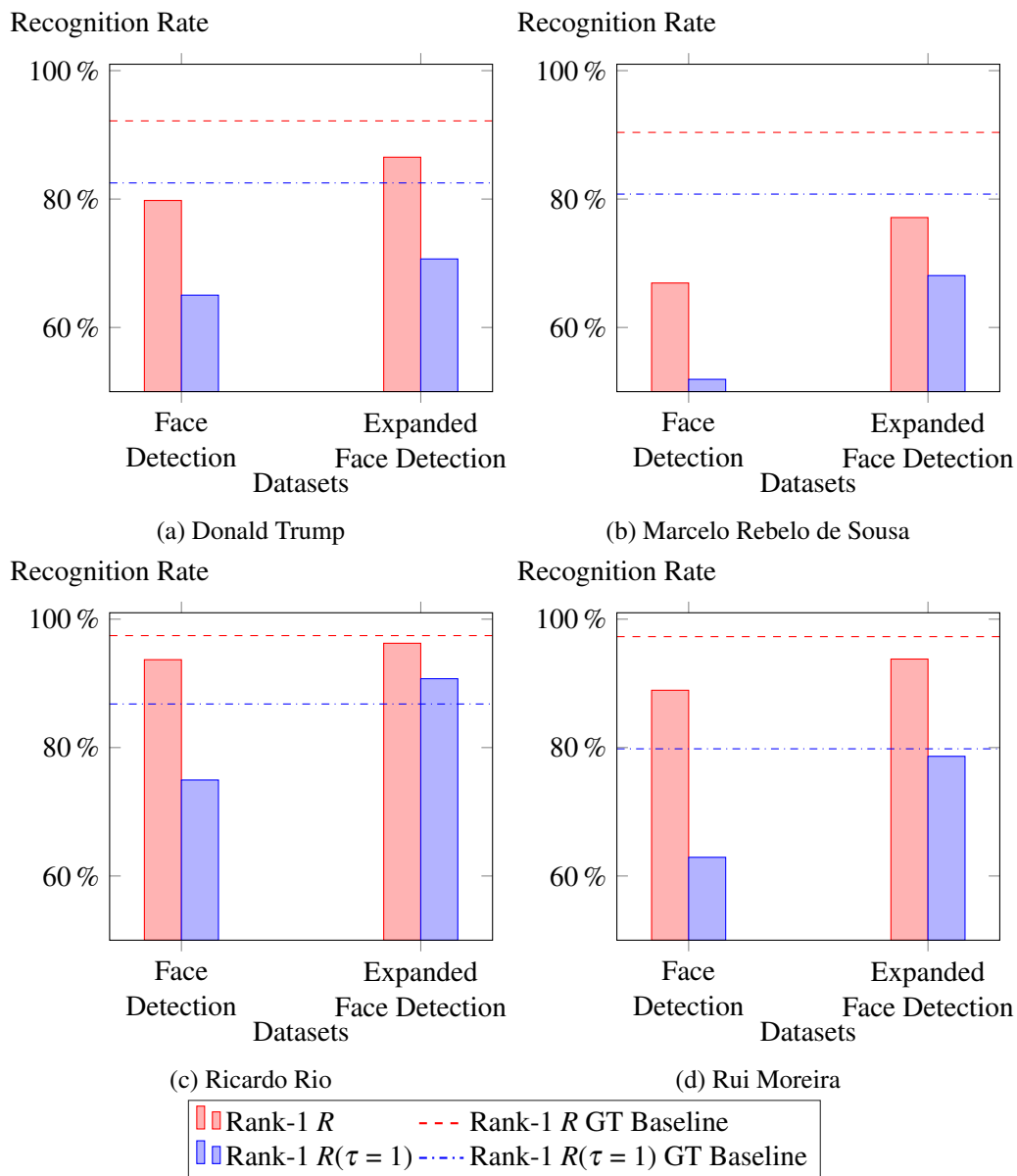


Figure 6.3: Frontal Gallery Set results with Face Detection Dataset as *Probe Set*.

As presented in Figure 6.3, the recognition rates for the Face Detection Datasets are considerable distinct within the four classes. However, it is evident the overall improvement of performance when using the Expanded Face Detection Dataset. Also, closer results of R and $R(\tau = 1)$ are obtain with this *Probe Set*, since the expansion resulted in bounding boxes more identical to the ground truth annotations, as proven in Figure 4.13 in Chapter 4. This leads, in turn, to lower *distances scores*, reflected in the better results of this dataset, relatively to the ones with the Face Detection Dataset.

Once again, the processes of face recognition only taking into account the correct *top matches*, Rank-1 R , for both Ricardo Rio and Rui Moreira have higher performances. Figure 6.3 b) shows that Marcelo Rebelo de Sousa had significantly low recognition rates. In the search for the reason behind these results, it was observed that the detection of Marcelo Rebelo de Sousa's face is very defective. This could be justified by the reduced dimension of his images, as demonstrated in Figure 4.7 in Chapter 4, and by the small proportion of his head in relation to the respective images; both factors hinder the detection. Consequently, and regarding all the four personalities, the given recognition rates also reflect the non-detection of faces in some images. Considering that fewer faces from Marcelo were detected, it is also reasonable that his results show a lower overall performance.

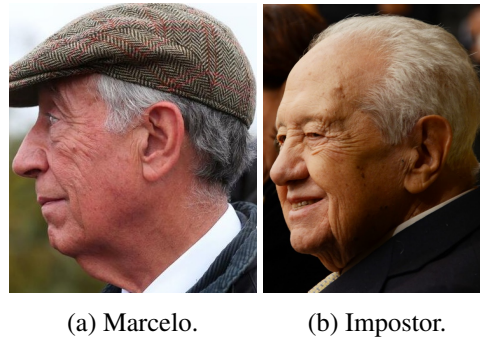
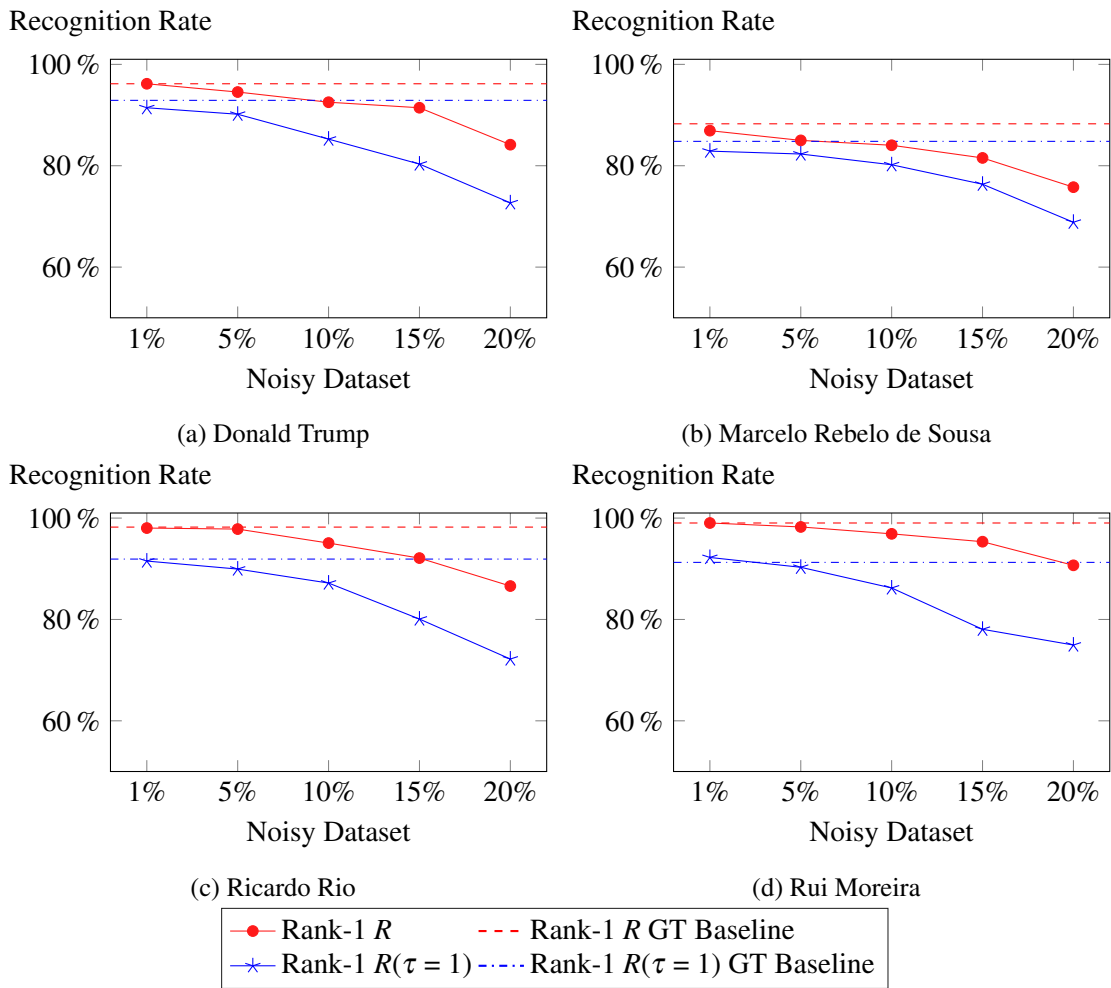
The presented results indicate that the use of a real detector in this verification system had a significant impact since it could add noise to the bounding box's results (shift or changes in dimension), but it could also fail to detect the face, which was never, in this context, compared. This is important information considering that face detection is a necessary step in a real pipeline.

6.1.3.2 3 Images Gallery Set

This section presents face verification results using three images in the *Gallery Set* - one of a frontal pose, one left and one right side view, bringing a total of 36 images with which the *Probe Sets* were compared. The results are depicted in Figures 6.5 and 6.6.

Analyzing 3GS results for the Noisy Dataset (Figure 6.5) it becomes evident the overall approximation between the two computed recognition rates, Rank-1 R and Rank-1 $R(\tau = 1)$. This occurs once, in contrast to what happened in the previous experiment (Section 6.1.3.1), the *Probe Sets* comparison was carried out with both frontal and side faces. Taking into account that the majority of the datasets show the personalities in a left or right view position (Figure 4.5), it was reasonable to expect that the *distance scores* would be, in general, lower. This was corroborated by the fact that most of the correct *top matches* were obtained with one of the two images added per individual, which, in turn, led to a notable better performance, relative to the FGS.

Despite the closer recognition rates for the Marcelo Rebelo de Sousa's dataset, this was the only one in which the performance decreased from the Frontal and to 3 Images Gallery Set results. A posterior analysis demonstrated that incorrect *top matches* predominantly recognized Marcelo Rebelo de Sousa as another personality from the gallery, recurring to the side view added images of the latter. Two images of an example of incorrect recognition are shown in Figure 6.4. Therefore, some of the previous correct *top matches* obtained with frontal images were replaced by incorrect ones with a non-frontal face of the impostor.

Figure 6.4: Example of a Marcelo Rebelo de Sousa's incorrect *top match*.Figure 6.5: 3 Images Gallery Set results with Noisy Dataset as *Probe Set*.

As for the Face Detection results shown in Figure 6.6, it is relevant to highlight the better performance of the system in recognizing Donald Trump, when compared with the previous test. From this can be implied that face detection in his dataset is adequate, unlike in Marcelo's, but the recognition in the FGS experiment was due to the aforementioned reasons, mainly because of the diversity of the facial expressions. These justified its performance for the Noisy Dataset. Ricardo

and Rui also presented an improvement for the Rank-1 $R(\tau = 1)$. For all these three personalities, the outcomes with the Expanded Face Detection Dataset conducted to a recognition rate that is nearer to the baseline scenario. Once more, Marcelo's results for both Face Detection Datasets remained substantially below to the GT baseline.

One can conclude that the overall performance of the proposed face recognition model based on the verification was improved by the addition of one left and one right side view pictures to the *Gallery Set*. This addition of images allowed the system to know more than one head pose for each individual and, therefore, made the face recognition more reliable for unconstrained scenarios, under which photographs or videos more hardly capture a face in a perfect frontal position.

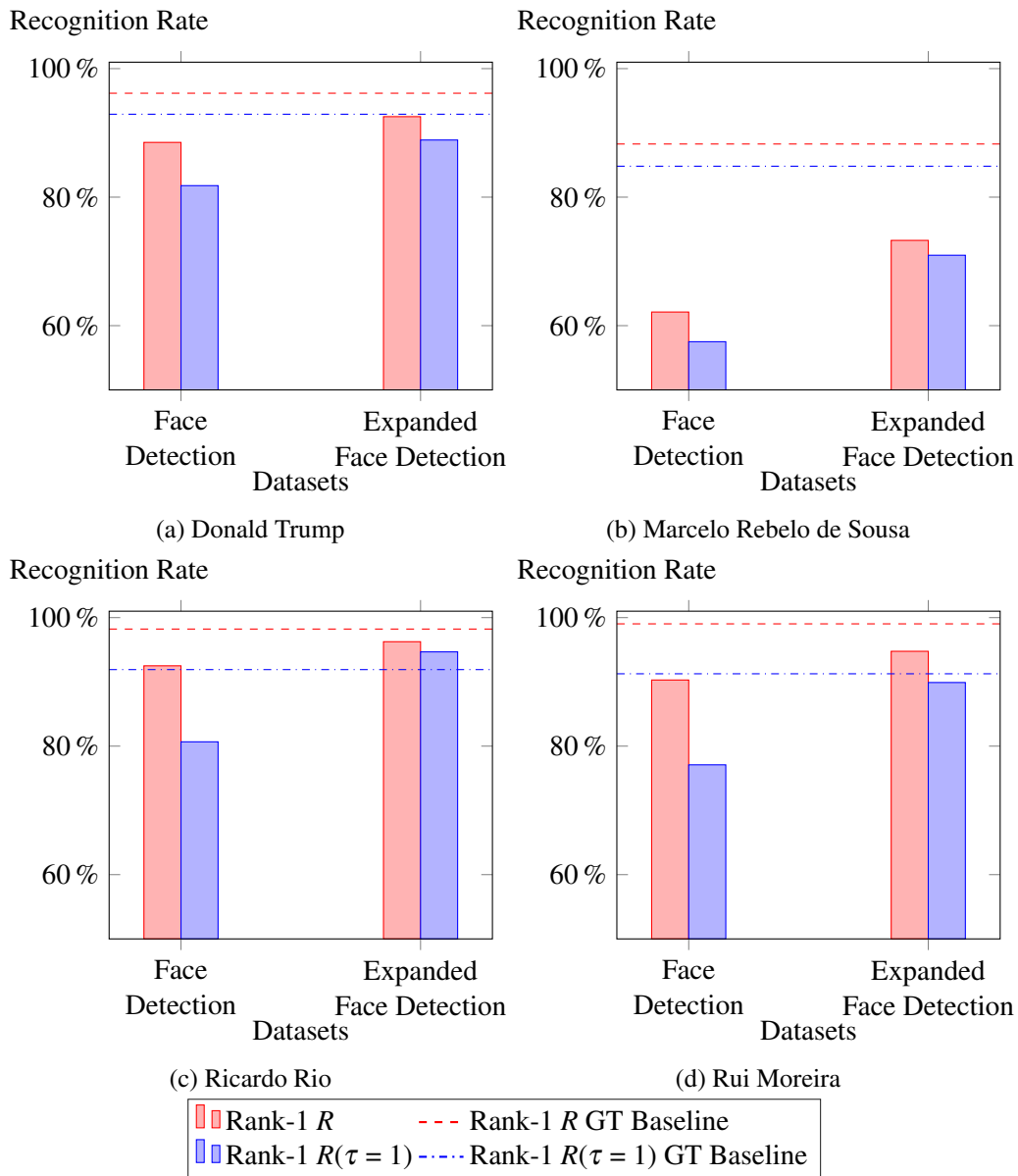


Figure 6.6: 3 Images Gallery Set results with Face Detection Dataset as *Probe Set*.

6.1.3.3 5 Images Gallery Set

In this section, the behavior of the face recognition verification based method is analyzed, using the entire *Gallery Set*, with 5 images - one frontal, two left and two right - for each one of the twelve known subjects. As a consequence, every *probe* will be compared with 60 face images.

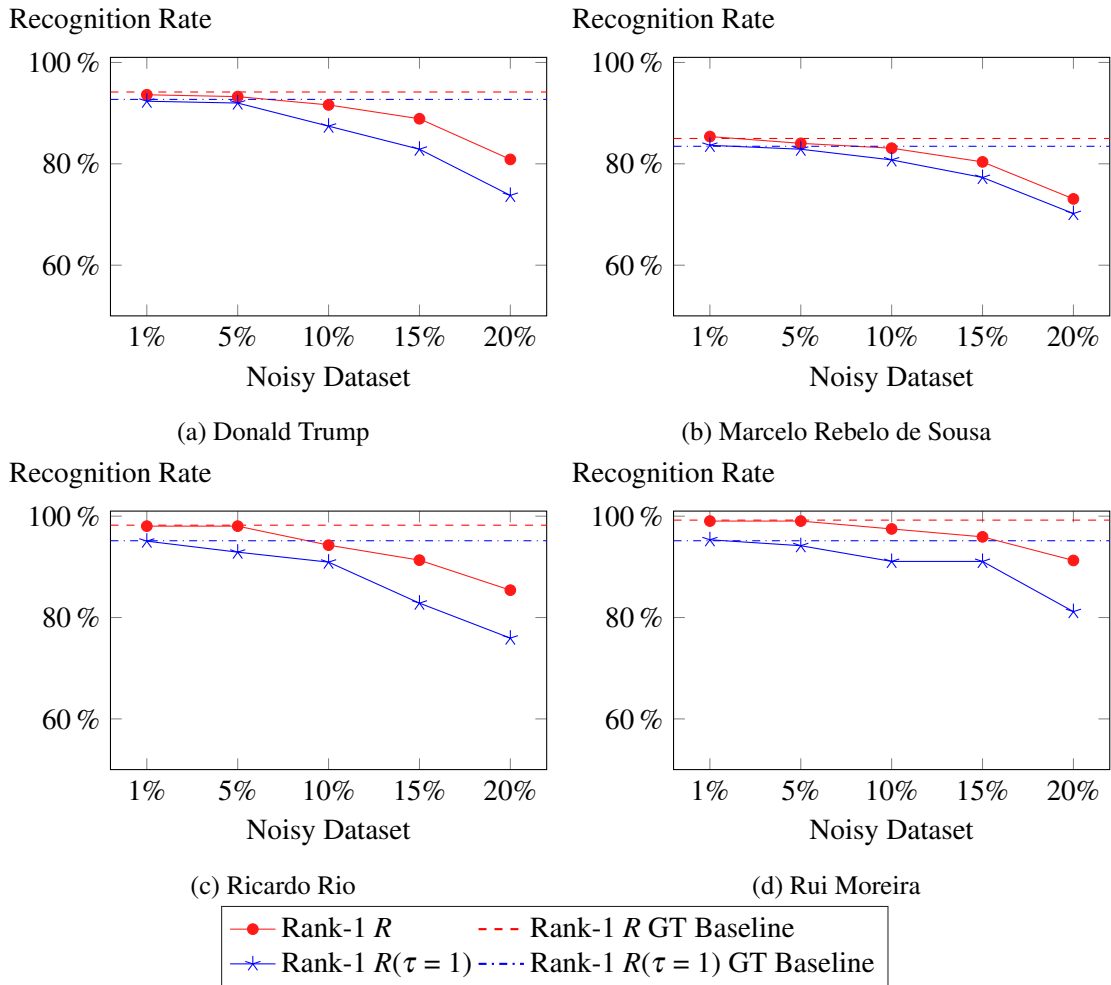


Figure 6.7: 5 Images Gallery Set results with Noisy Dataset as *Probe Set*.

For the 5 Images Gallery Set, Figure 6.7 shows a slight improvement of the face recognition performance. When the threshold was considered, higher reliable results were reached, with a Rank-1 $R(\tau)$ almost identical to the one of Rank-1 R , which in turn is practically the same as the GT performance for the 1%, 5% and 10% Noisy Datasets. This means that having more than one image per known individual, especially with a variety of non-frontal head poses, leads to a system capable of adequate face recognition, even if the face detection phase introduces a noise up to about 10% of the size of the face exhibited in the image.

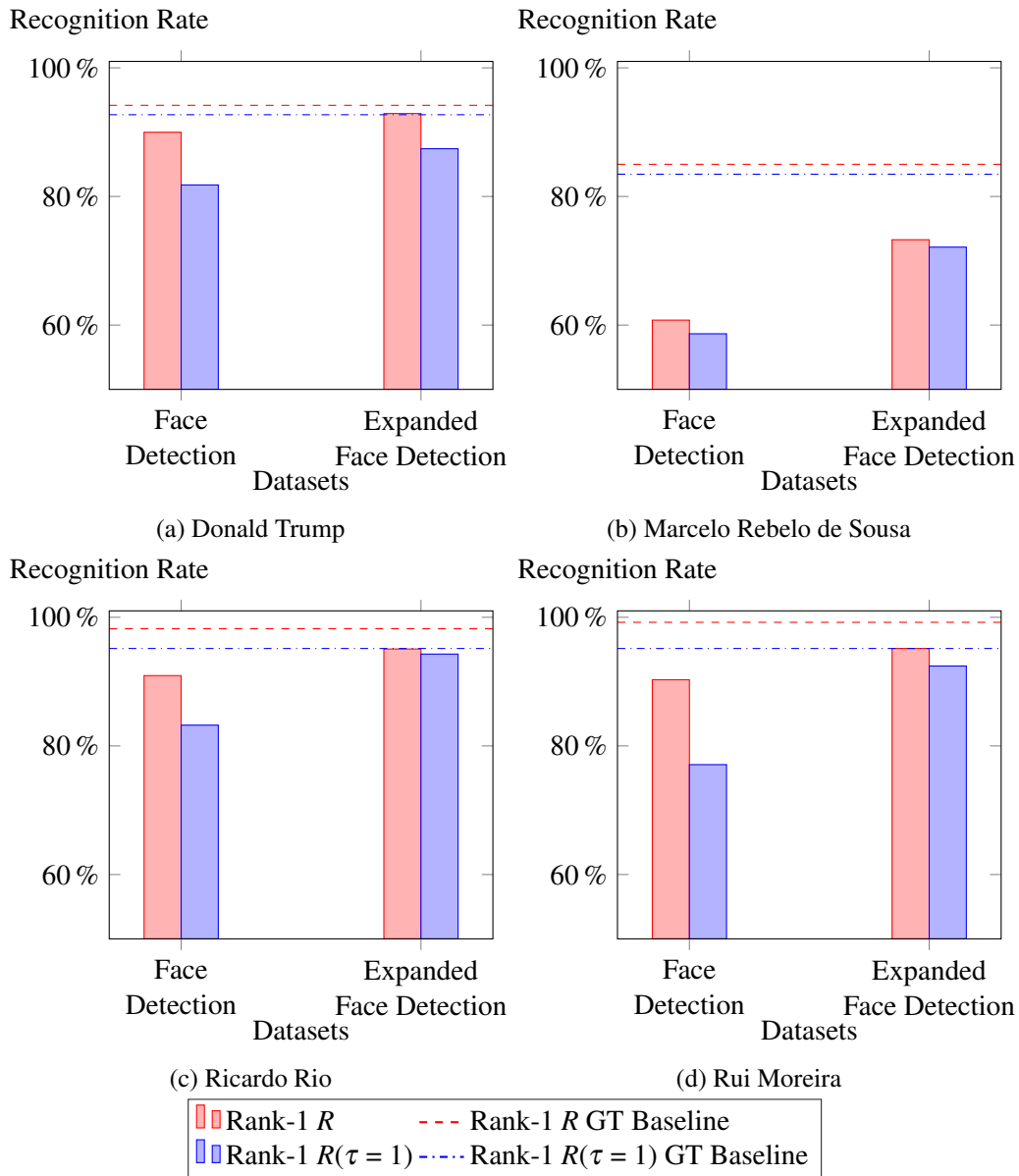
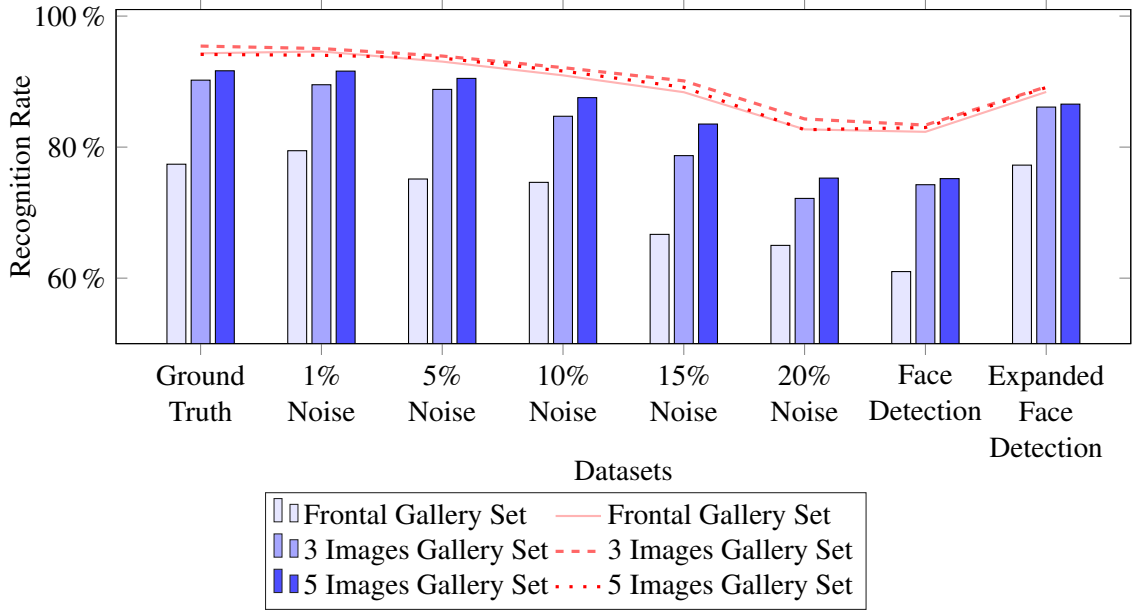


Figure 6.8: 5 Images Gallery Set results with Face Detection Dataset as *Probe Set*.

The results depicted in Figure 6.8 are very similar to Figure 6.6, unlike what happened with the Noisy Dataset. These are indicative of the impact of the face detector. Dlib's Face Detector usually detects a face with a tight bounding box, not considering the whole head region as the images in the *Gallery Set*, which often crop important facial features. When compared with images that take into account a different face region, lower *distance scores* were obtained, although the percentage of correct *top matches* was high. Hence, the comparison with the expanded bounding boxes resulted in better and more similar performances to the Ground Truth outcomes, once the applied expansion discarded part of the noise introduced by the face detector.

6.1.3.4 Conclusions

To understand the overall performance of the proposed face recognition method, the recognition rates for the four individuals were aggregated for each *Probe Set*. These results are presented in Table 6.2, and visually shown in Figure 6.9. It is also shown in Figure 6.10 the results of each *Probe Set* per personality dataset.

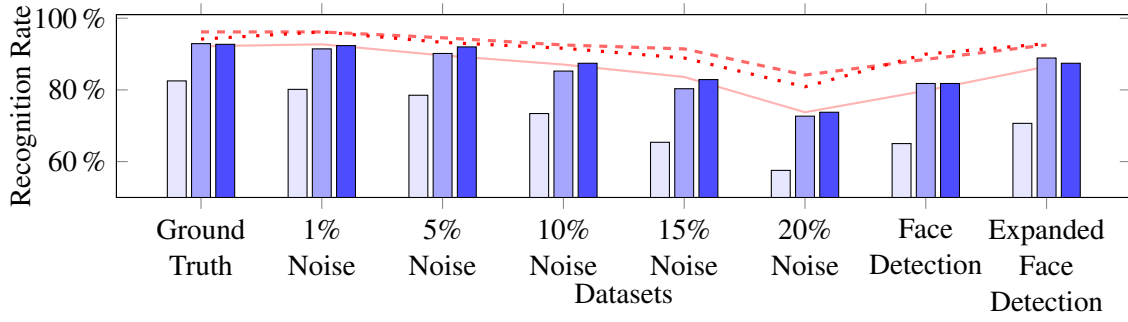


In red are the Rank-1 R results and the blue are the Rank-1 $R(\tau = 1)$ results.

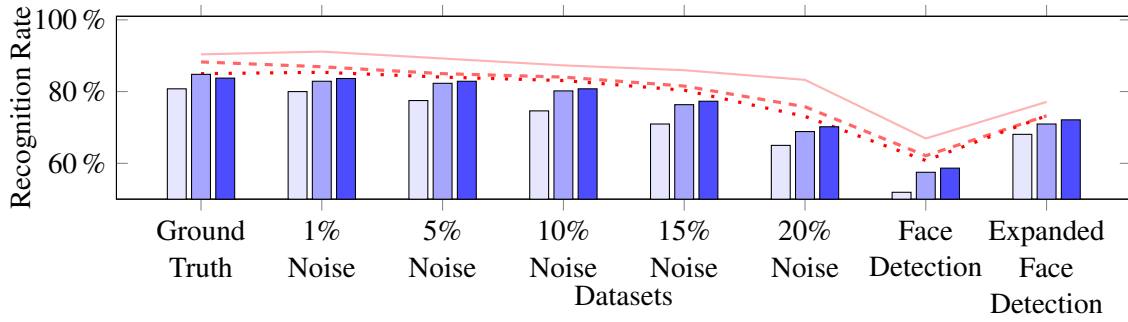
Figure 6.9: Average results of each Gallery Set.

In Figure 6.9, it is easily noted that the recognition performance of the system without considering the threshold value, given by the metric Rank-1 R , did not suffer a significant effect throughout the three *Gallery Sets*. This means that there was not a notable improvement in the performance when this facial recognition method was applied relying exclusively on the top match and the identification of a person resorting to one or two pictures per side view of its face, rather than to just one frontal image. Since this metric did not use a confidence value (threshold) to accept the *top match* as correct or not, it can be stated that its outcomes are not highly reliable and that it would be difficult to apply this method in a real system.

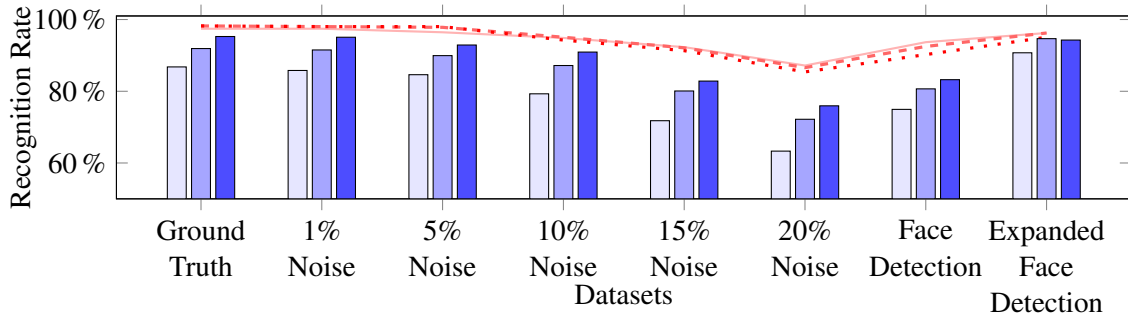
On the other hand, it is evident the $\sim 11\%$ increase of the recognition rate when the introducing the threshold, Rank-1 $R(\tau)$, from the Frontal to the 3 Images Gallery Set. Thus, such results support that by adding images, which show both left and right sides of the individual's face, to the *Gallery Set*, more reliable identifications can be achieved. The improvement accomplished with 5 images in the *Gallery Set* is not so pronounced when compared to the latter outcomes. Once the two additional images continue to present the subject in a side view, and also by the fact that they only show a little yaw variation in the head pose.



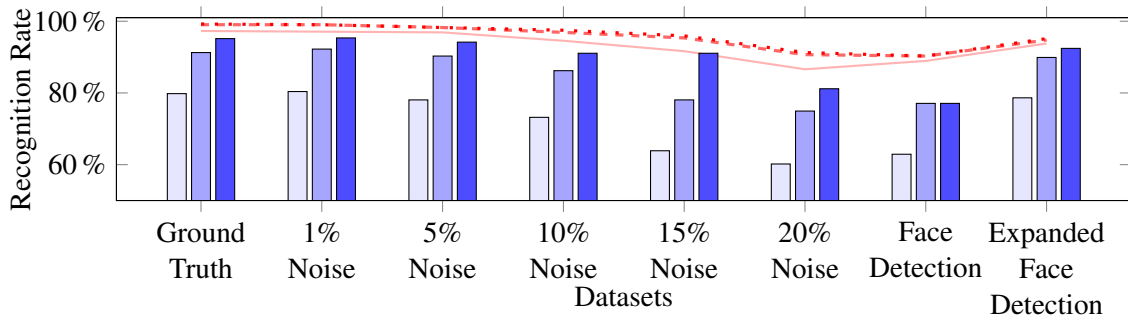
(a) Donald Trump



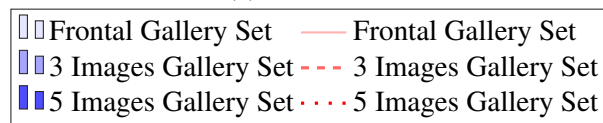
(b) Marcelo Rebelo de Sousa



(c) Ricardo Rio



(d) Rui Moreira



In red are the Rank-1 results and the blue are the Rank-1 ($\tau = 1$) results.

Figure 6.10: Average results of each personality dataset.

It is, then, possible to state that for the implementation of this method in a real-life scenario, the addition of these last two images would not be reasonable, since it would not significantly improve the recognition task and would impose a higher computational cost to the system. So, it is argued that a consistent facial recognition can be achieved recurring to three images per personality known to the system - one frontal, one left and one right. Therefore, in the following sections, when referring to the verification based method, the performance results presented are the ones obtained with the 3 Images Gallery Set.

Another interesting outcome was the recognition performance considering a confidence value τ , for the Face Detection Dataset, which resulted from the application of the Dlib's CNN Face Detector on the Base Dataset, as described in Section 4.4. These results proved that the use of a real face detector in a recognition system introduces noise that will harm the posterior identification. In this particular case, considering the similar results between the Face Detection Dataset and the 20% Noisy Dataset, it can be assumed that Dlib's CNN Face Detector might have added modifications of about 20% in his detection. The Expanded Face Detection Dataset, even though reduces part of the Dlib introduced noise, presented similar results to those of 10% Noise Dataset.

<i>Gallery Sets</i> <i>Probe Sets</i>	FGS		3GS		5GS	
	Rank-1 <i>R</i>	Rank-1 <i>R</i> (τ)	Rank-1 <i>R</i>	Rank-1 <i>R</i> (τ)	Rank-1 <i>R</i>	Rank-1 <i>R</i> (τ)
Ground Truth	94,32%	77,39%	95,42%	90,22%	94,15%	91,65%
1% Noise	94,60%	79,45%	95,04%	89,52%	94,02%	91,60%
5% Noise	93,05%	75,13%	93,90%	88,18%	93,59%	90,49%
10% Noise	90,95%	74,62%	92,13%	84,71%	91,61%	87,55%
15% Noise	88,38%	66,68%	90,11%	78,70%	89,13%	83,52%
20% Noise	82,71%	65,00%	84,30%	72,17%	82,65%	75,27%
Face Detection	82,33%	64,28%	83,36%	74,26%	82,99%	75,19%
Expanded Face Detection	88,42%	77,90%	89,20%	86,11%	89,10%	86,56%

Table 6.2: Average Results.

6.2 Classification Method

6.2.1 Model Parameters

As stated in the previous chapter, Tensorflow Object Detection API was used for the classification method, alongside with a pre-trained model with the state-of-the-art COCO dataset. This model is a Faster RCNN, trained to detect and classify 90 classes of objects, such as people, dogs, cars, etc. This classification method aimed to perform transfer learning, which a Machine Learning technique, on the pre-trained model, to obtain a fine-tuned model able to detect and classify the four selected personalities. The goal was, then, to start from a pre-trained model and adapt it without the need for several thousand of images for each intended label.

Several tests were conducted, changing several parameters on the model configuration, in order to achieve the best model for the recognition of the four intended classes. Most of the experiments were performed by modifying the number of steps, the batch size, the optimizer, the initial learning rate, and the image resizer, varying from one that maintained the aspect ratio of the image to another that applied a fixed resize.

Although it was initially defined, through the 5-Fold Cross Validation technique explained in Section 5.4, that the transfer learning would be performed with 360 images per class, other tests were conducted, decreasing the amount of images in the training set and increasing the number of steps. These experiments resulted in less robust outcomes, even with a higher number of steps. It was not considered reasonable to present the different results for the various configuration settings experienced, and so the elected fine-tuned model and one used in further presented results was accomplished with the following parameters:

- *model*: Faster RCNN
- *image resizer*: keep_aspect_ratio_resizer
 - min_dimension: 600
 - max_dimension: 1024
- *score converter*: SOFTMAX
- *optimizer*: Adam
- *initial learning rate*: 0.00001
- *batch size*: 1
- *steps*: 14400

6.2.2 Results Analysis

The obtained results for the validation and test sets are presented in the tables 6.3 and 6.4, respectively. These show the average of the computed performance metrics, mAP and AR, regarding the five combinations for the train and validation sets. With the validation set, the evaluation was performed during the training, while the metrics for the test set were conducted when the train was already concluded and with 50 images per class that were never presented to the system before. For both tests, the performance had similar behavior.

Class	mAP	AR@100
Donald Trump	78.19%	82.42%
Marcelo Rebelo de Sousa	69.38%	77.62%
Ricardo Rio	70.04%	78.18%
Rui Moreira	74.47%	79.36%
<i>Average</i>	72.98%	79.36%

Table 6.3: 5-Folds Cross Validation average mAP and AR for validation set.

Class	mAP	AR@100
Donald Trump	76.91%	81.64%
Marcelo Rebelo de Sousa	73.03%	78.88%
Ricardo Rio	65.05%	75.76%
Rui Moreira	76.54%	80.72%
<i>Average</i>	73.20%	79.22%

Table 6.4: 5-Folds Cross Validation average mAP and AR for test set.

6.3 Comparative Analysis

To perform a proper comparison between the two proposed face recognition methods, the F-measure metric was computed. This metric was calculated for the Verification Based Method having as *Probe Set* the Ground Truth Dataset and the Face Detection Dataset, both confronted with the 3 Images *Gallery Set*. Additionally, the F-measure metric was determined for the Classification Method, using the aforementioned configuration settings and a training set with 360 images for each class. These results are expressed in Table 6.5.

	<i>F-measure</i>		
	Verification Based Method		Classification Method
	Ground Truth (3GS)	Face Detection (3GS)	360 images Training Set
Donald Trump	96,32%	93,60%	97,96%
Marcelo Rebelo de Sousa	91,78%	77,72%	94,74%
Ricardo Rio	95,79%	86,28%	93,48%
Rui Moreira	95,43%	87,05%	98,99%
<i>Overall System</i>	94,87%	86,73%	96,35%

Table 6.5: F-measure for the Classification and the Fusion Method.

The results point to a better overall performance of the Classification Method. Although similar results were achieved for the Verification Based Method with the Ground Truth Dataset, these

do not reflect the influence of the face detection phase. So, even though the same metric was used to evaluate the performance, it would be more accurate to compare the classification results to the verification ones for the Face Detection Dataset. With this in mind, from Table 6.5 is inferred that the classification method outperforms the verification based one by $\sim 11\%$.

6.4 Fusion Method

Driven by the results achieved for the Verification Based and the Classification Methods, a Fusion Method was proposed, as stated in Chapter 3. This aimed to combine the outputs of both methods to achieve an even better performance attending this dissertation's task of face recognition. Taking this under consideration, the higher goal was to analyze how, with this Fusion Method, it would be possible to achieve similar or even outperforming results to the Classification Method, which presented the best individual performance, as concluded in the previous section.

Within the framework of these criteria, the Fusion Method was implemented through the following steps:

1. For each trained model, which differ in the number of images for the training set, its output was extracted. This output was composed by the detected bounding box by the neural network, the attributed label - which can be one of the established classes or a "N/A" one - and the confidence score for this classification;
2. The obtained ROIs, agnostic of label and confidence score, were given as input to the Verification Based Method system;
3. These inputs were compared with the 3 Images Gallery Set, and the identification was through the correct top matches whose the *distance scores* were below the threshold value. From this, the outputs were the classified label and the correspondent *distance score*;
4. Since the confidence score of the Classification Method was also within a range of $[0, 1]$, the *distance score* from the Verification Based Method was normalized. Traducing this *distance score* in a confidence score, and with the knowledge that lower scores probably represent a more accurate classification, it was used $1 - \text{the normalized distance score}$.
5. Finally, the combination of both outputs was performed using both labels and confidence scores. For the Classification Method, the label was referred to as $label_C$ and the confidence score as $score_C$. As for the posterior application of the Verification Based Method, the output label was referred to as $label_V$ and the confidence score, derived from the *distance score*, as $score_V$. The labels of the correspondent scores were only accepted through a threshold value, which for $score_C$ was 0.5 and for $score_V$ was 0.75. The first represents that were only considered the classifications obtained with a confidence above 50% and the latter, referent to the already set threshold value, being $0.75 = 1 - \text{the normalization of the } \tau \text{ value } 1$;

6. The fusion was conducted giving both classifications, $label_C$ and $label_V$, different weights, α and β , respectively. These weights were applied to the correspondents confidence scores, $score_C$ and $score_V$, and considering that $\alpha + \beta = 1$.

Consequently, the recognition of a person's identity by the Fusion Method was performed following the Algorithm 2, represented by $label_F$.

Algorithm 2 Fusion Method

```

1:  $label_C, score_C \leftarrow$  classification outputs
2:  $label_V, score_V \leftarrow$  verification outputs
3:  $\alpha \leftarrow$  weight of the classification
4:  $\beta \leftarrow$  weight of the verification
5: if  $score_C > 0.5$  and  $score_V < 0.75$  then
6:    $label_F \leftarrow label_C$ 
7: else
8:   if  $score_C < 0.5$  and  $score_V > 0.75$  then
9:      $label_F \leftarrow label_V$ 
10:  else
11:    if  $score_C < 0.5$  and  $score_V < 0.75$  then
12:       $label_F \leftarrow \text{"N/A"}$ 
13:    else
14:      if  $label_C = label_V$  then
15:         $label_F \leftarrow label_C$ 
16:      else
17:        if  $label_C \neq label_V$  and  $\alpha score_C > \beta score_V$  then
18:           $label_F \leftarrow label_C$ 
19:        else
20:          if  $label_C \neq label_V$  and  $\alpha score_C < \beta score_V$  then
21:             $label_F \leftarrow label_V$ 
22:          end if
23:        end if

```

Several tests were executed differing the α and β weights. The best performance for the recognition task was obtained having $\alpha = 0.3$ and $\beta = 0.7$. In Table 6.6, the results for the Fusion Method are compared with the ones of the Classification Method, using the same number of images for the training set. It is important to clarify that the presented numbers of images for the training set represent the amount of pictures per personality presented to the neural network, having the training set a total of images four times these numbers.

# Training Set	<i>F-measure</i>					
	360	200	100	48	32	24
Classification Method	96,35%	96,30%	95,47%	91,26%	86,40%	81,62%
Fusion Method	98,21%	97,92%	97,65%	94,43%	94,44%	93,96%

Table 6.6: F-measure for Classification and Fusion Methods.

As expected, from Table 6.6, it is evident that with the combination of the verification and the classification methods outperforming results were achieved. Through Figure 6.11, where these outcomes are visually shown, the improvement obtained with the latter proposed method is even more highlighted. It can be asserted that even when reducing the training set from 360 to 48 data samples, the Fusion Method always outperforms the Classification Method. This enhancement in performance is more pronounced when the training set is reduced to only 24 images, with a discrepancy of performance of $\sim 12\%$. It can also be observed that the Fusion Method had better outcomes for the six training sets, that the ones obtained with the verification method for the Face Detection Dataset, implying that these reduced the negative impact imposed by the required face detection phase in a face recognition system. Even when smaller training sets were used, the performance of the proposed approach was similar to the one reached with the original Classification Method.

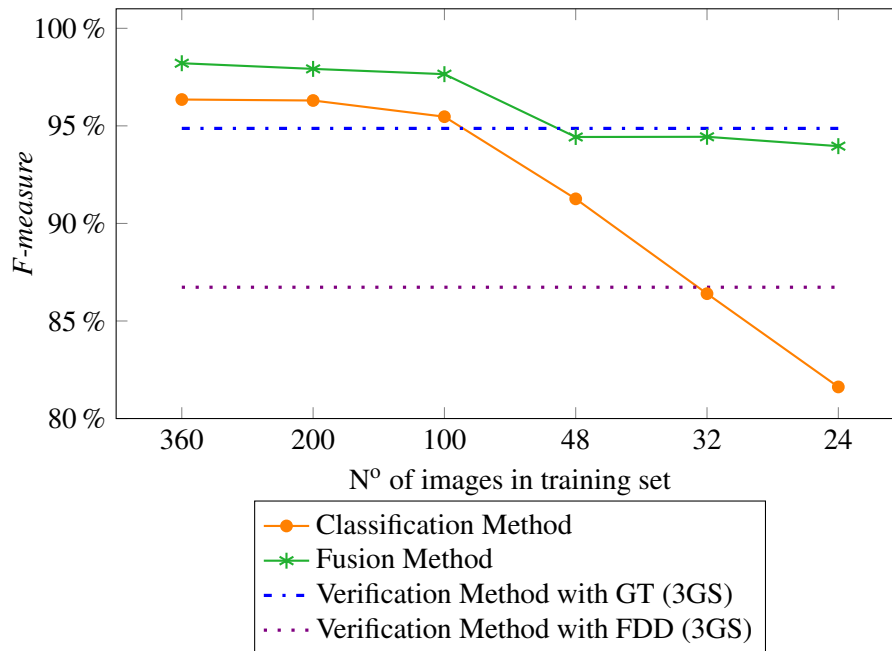


Figure 6.11: F-measure of Fusion Method.

With this in mind, it is concluded that by reducing the required amount of data for this latter face recognition method in $\sim 93\%$ and combining its results with a verification approach, it was possible to achieve a proper face recognition, with a small disadvantageous impact of 2% , relatively to the original classification system.

Chapter 7

Conclusion

The main goal of this work was to analyze different approaches for the face recognition problem in unconstrained scenarios. With that aim in view, a research of the algorithms proposed in the literature was conducted. It was also attended the fact that in facial recognition systems, a face detection phase is essential when addressing images from uncontrolled environments. Face recognition revealed several applications in the computer vision field, having been an active object of study from the past years, and despite all the positive outcomes achieved, this remains a challenging and unsolved problem. Additionally, the evidence from this study suggests two principal types of techniques for face recognition: verification and classification. In this context, two methodologies, one for each type of recognition, were implemented based on analyzed algorithms from the state-of-the-art.

The use case of this dissertation is the face recognition and its automatic annotation of local and national audiovisual contents. These contents represent unrestrained scenarios, which impose an increased difficulty on the verification/classification of a person's identity. To this end, it was defined as the primary task the recognition of four specific personalities, representative entities of the groups of people more probable to appear these audiovisual environments. To such purpose, it was required a dataset preparation. The collection of a significant number of images per individual has proven to be a difficult task, particularly when it came to local personalities. This provided further evidence that there was a need to minimize the data preparation effort.

In order to accomplish the principal objective, the face recognition developed methods, Verification Based Method and Classification Method, were applied making use of the collected data. Both approaches rely on machine learning state-of-the-art models. In the Verification Based Method, it was first used the Dlib's CNN Face Detector, being then the extracted ROIs fed into a FaceNet network, which compared them with a set of labeled faces. The face in the input ROI would be identified as belonging to the person whom the system pointed out as the most similar one within its known identities, and according to a threshold value. As for the Classification Method, an object detection pre-trained model was adopted, on which was performed transfer learning for the attainment of a neural network able to detect and classify the four intended personalities.

The experiments for the Verification Method corroborate the previous suspicions that the

needed preliminary face detection might detriment the final identity classification since the latter was performed on top of the obtained bounding boxes. Once the face detector can end up not detecting or leading to a resultant ROI that does not represent a face or that has a poor representation of one, the recognition system has a negative influence, resulting thus in a less robust method. With this method was obtained a performance of $\sim 87\%$, computed by the F-measure metric. However, it is aware that this methodology may have a limitation that could have influenced the results obtained, once this evaluation is regarding the exclusive use of one face detector. Further experiments with other face detectors would be needed to determine exactly how face detection affects face recognition.

With the experiments for the Classification Method, outperforming results, of $\sim 96,35\%$, were achieved. Even so, to perform accurate recognition, this method requires a vast quantity of annotated data, which in turn increases, even more, the computational cost of the training task. Nonetheless, a significant decrease in training data had a notable negative impact on performance. Taking into account the considerable effort in that data gathering, an alternative methodology was planned. Therefore, a Fusion Method was proposed. This combines the aforementioned methods, by introducing the Classification Method outputs - classification label and confidence score - in the Verification Based Method system. The findings of this novel approach indicates that it possible to reduce, by $\sim 93\%$, the amount of annotated images necessary, from 360 to 24 pictures per individual to recognize, to fine-tuning an object detection model, and still be able achieve a system for face recognition with a high performance of $\sim 93,96\%$. This latter presented strategy revealed interesting results, indicating that both human and machine efforts, commonly implied in a face recognition system, can be significantly diminished and yet secure an adequate classification, combining distinct techniques for the same aim.

Revisiting the point at which this dissertation was left, the following steps are suggested as future lines of work: expand the experiments to other representative entities, including the female gender and other human races; introduction of others face detectors from the literature, and analyze their different impacts; analyze other approaches for the implementation of the proposed Fusion Method; use a face detector on the Fusion Method instead of the ROIs from the Classification Method; analyze the impact of increasing the number of faces to recognize; test transfer learning in a pre-trained model to detect faces.

References

- [1] Junliang Xing, Zhiheng Niu, Junshi Huang, Weiming Hu, Xi Zhou, and Shuicheng Yan. Towards robust and accurate multi-view and partially-occluded face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:987–1001, 2018.
- [2] K. Messer, Jiri Matas, J. Kittler, K. Jonsson, Juergen Luetttin, and Gilbert Maître. XM2VTSDB: The extended M2VTS database. *Proc. of Audio- and Video-Based Person Authentication*, 04 2000.
- [3] A M. Martinez and Robert Benavente. The AR face database. *Tech. Rep. 24 CVC Technical Report*, 01 1998.
- [4] AR face database 22pt markup. URL: https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/tarfd_markup/tarfd_markup.html.
- [5] M.M. Nordstrom, M Larsen, J. Sierakowski, and M.B. Stegmann. The IMM face database. *Environment*, 22:1–4, 01 2004.
- [6] S. Milborrow, J. Morkel, and F. Nicolls. The MUCT Landmarked Face Database. *Pattern Recognition Association of South Africa*, 2010. <http://www.milbo.org/muct>.
- [7] Andrzej Kasiński, A. Florek, and Adam Schmidt. The put face database. *Image Processing and Communications*, 13:59–64, 01 2008.
- [8] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The FERET evaluation methodology for face-recognition algorithms. 22(10):15, 2000.
- [9] BioID face database | dataset for face detection | facedb. URL: <https://www.bioid.com/facedb/>.
- [10] Vidit Jain and Erik Learned-Miller. FDDB: A Benchmark for Face Detection in Unconstrained Settings. page 11, 2010.
- [11] Gary B. Huang, Marwan Mattar, Tamara L. Berg, and Erik G. Learned-Miller. Labeled faces in the wild : A database for studying face recognition in unconstrained environments. 2007.
- [12] Martin Kostinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2144–2151. IEEE, 2010. URL: <http://ieeexplore.ieee.org/document/6130513/>, doi:10.1109/ICCVW.2011.6130513.

- [13] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR 2011*, pages 545–552, June 2011. doi:10.1109/CVPR.2011.5995602.
- [14] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. pages 2879–2886, 06 2012. doi:10.1109/CVPR.2012.6248014.
- [15] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang. Interactive facial feature localization. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, volume 7574, pages 679–692. Springer Berlin Heidelberg, 2012. URL: http://link.springer.com/10.1007/978-3-642-33712-3_49, doi:10.1007/978-3-642-33712-3_49.
- [16] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. *2013 IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.
- [17] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403, Dec 2013. doi:10.1109/ICCVW.2013.59.
- [18] Yuqian Zhou, Ding Liu, and Thomas Huang. Survey of face detection on low-quality images. 2018. URL: <http://arxiv.org/abs/1804.07362>, arXiv:1804.07362.
- [19] Paul Viola and Michael Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57:137–154, 05 2004. doi:10.1023/B:VISI.0000013087.49260.fb.
- [20] Jianguo Li, Tao Wang, and Yimin Zhang. Face detection using SURF cascade. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2183–2190. IEEE, 2011. URL: <http://ieeexplore.ieee.org/document/6130518/>, doi:10.1109/ICCVW.2011.6130518.
- [21] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 1, pages 886–893. IEEE, 2005. URL: <http://ieeexplore.ieee.org/document/1467360/>, doi:10.1109/CVPR.2005.177.
- [22] David G. Lowe. Distinctive image features from scale-invariant keypoints. 60(2):91–110, 2004. URL: <http://link.springer.com/10.1023/B:VISI.0000029664.99615.94>, doi:10.1023/B:VISI.0000029664.99615.94.
- [23] Lun Zhang, Rufeng Chu, Shiming Xiang, Shengcai Liao, and Stan Li. Face detection based on multi-block LBP representation. pages 11–18, 01 2007. doi:10.1007/978-3-540-74549-5_2.
- [24] Xiangxin Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886. IEEE, 2012. URL: <http://ieeexplore.ieee.org/document/6248014/>, doi:10.1109/CVPR.2012.6248014.

- [25] J. Yan, Z. Lei, L. Wen, and S. Z. Li. The Fastest Deformable Part Model for Object Detection. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2497–2504, June 2014. doi:[10.1109/CVPR.2014.320](https://doi.org/10.1109/CVPR.2014.320).
- [26] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8692, pages 720–735. Springer International Publishing, 2014. URL: http://link.springer.com/10.1007/978-3-319-10593-2_47, doi:[10.1007/978-3-319-10593-2_47](https://doi.org/10.1007/978-3-319-10593-2_47).
- [27] Junjie Yan, Xucong Zhang, Zhen Lei, and Stan Z. Li. Real-time high performance deformable model for face detection in the wild. In *2013 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2013. URL: <http://ieeexplore.ieee.org/document/6612972/>, doi:[10.1109/ICB.2013.6612972](https://doi.org/10.1109/ICB.2013.6612972).
- [28] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sep. 2010. doi:[10.1109/TPAMI.2009.167](https://doi.org/10.1109/TPAMI.2009.167).
- [29] Piotr Dollar, Zhuowen Tu, Pietro Perona, and Serge Belongie. Integral channel features. In *Proceedings of the British Machine Vision Conference 2009*, pages 91.1–91.11. British Machine Vision Association, 2009. URL: <http://www.bmva.org/bmvc/2009/Papers/Paper244/Paper244.html>, doi:[10.5244/C.23.91](https://doi.org/10.5244/C.23.91).
- [30] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li. Aggregate channel features for multi-view face detection. *CoRR*, abs/1407.4023, 2014. URL: <http://arxiv.org/abs/1407.4023>, arXiv:[1407.4023](https://arxiv.org/abs/1407.4023).
- [31] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483. IEEE, 2013. URL: <http://ieeexplore.ieee.org/document/6619290/>, doi:[10.1109/CVPR.2013.446](https://doi.org/10.1109/CVPR.2013.446).
- [32] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8694, pages 94–108. Springer International Publishing, 2014. URL: http://link.springer.com/10.1007/978-3-319-10599-4_7, doi:[10.1007/978-3-319-10599-4_7](https://doi.org/10.1007/978-3-319-10599-4_7).
- [33] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8690, pages 1–16. Springer International Publishing, 2014. URL: http://link.springer.com/10.1007/978-3-319-10605-2_1, doi:[10.1007/978-3-319-10605-2_1](https://doi.org/10.1007/978-3-319-10605-2_1).
- [34] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5325–5334. IEEE, 2015. URL: <http://ieeexplore.ieee.org/document/7299170/>, doi:[10.1109/CVPR.2015.7299170](https://doi.org/10.1109/CVPR.2015.7299170).

- [35] H. Qin, J. Yan, X. Li, and X. Hu. Joint Training of Cascaded CNN for Face Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3456–3465, June 2016. doi:[10.1109/CVPR.2016.376](https://doi.org/10.1109/CVPR.2016.376).
- [36] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. 23(10):1499–1503, 2016. URL: <http://ieeexplore.ieee.org/document/7553523/>, doi:[10.1109/LSP.2016.2603342](https://doi.org/10.1109/LSP.2016.2603342).
- [37] Kaipeng Zhang, Zhanpeng Zhang, Hao Wang, Zhifeng Li, Yu Qiao, and Wei Liu. Detecting faces using inside cascaded contextual CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3190–3198. IEEE, 2017. URL: <http://ieeexplore.ieee.org/document/8237606/>, doi:[10.1109/ICCV.2017.344](https://doi.org/10.1109/ICCV.2017.344).
- [38] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. 2013. URL: <http://arxiv.org/abs/1311.2524>, arXiv:[1311.2524](https://arxiv.org/abs/1311.2524).
- [39] Ross Girshick. Fast r-CNN. 2015. URL: <http://arxiv.org/abs/1504.08083>, arXiv:[1504.08083](https://arxiv.org/abs/1504.08083).
- [40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-CNN: Towards real-time object detection with region proposal networks. 2015. URL: <http://arxiv.org/abs/1506.01497>, arXiv:[1506.01497](https://arxiv.org/abs/1506.01497).
- [41] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-CNN. 2016. URL: <http://arxiv.org/abs/1606.03473>, arXiv:[1606.03473](https://arxiv.org/abs/1606.03473).
- [42] Hao Wang, Zhifeng Li, Xing Ji, and Yitong Wang. Face r-CNN. 2017. URL: <http://arxiv.org/abs/1706.01061>, arXiv:[1706.01061](https://arxiv.org/abs/1706.01061).
- [43] Chenchen Zhu, Yutong Zheng, Khoa Luu, and Marios Savvides. CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection. 2016. URL: <http://arxiv.org/abs/1606.05413>, arXiv:[1606.05413](https://arxiv.org/abs/1606.05413).
- [44] Xudong Sun, Pengcheng Wu, and Steven C. H. Hoi. Face detection using deep learning: An improved faster RCNN approach. 2017. URL: <http://arxiv.org/abs/1701.08289>, arXiv:[1701.08289](https://arxiv.org/abs/1701.08289).
- [45] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. 9905:21–37, 2016. URL: <http://arxiv.org/abs/1512.02325>, arXiv:[1512.02325](https://arxiv.org/abs/1512.02325), doi:[10.1007/978-3-319-46448-0_2](https://doi.org/10.1007/978-3-319-46448-0_2).
- [46] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S³fd: Single shot scale-invariant face detector. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 192–201. IEEE, 2017. URL: <http://ieeexplore.ieee.org/document/8237292/>, doi:[10.1109/ICCV.2017.30](https://doi.org/10.1109/ICCV.2017.30).
- [47] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. FaceBoxes: A CPU real-time face detector with high accuracy. 2017. URL: <http://arxiv.org/abs/1708.05234>, arXiv:[1708.05234](https://arxiv.org/abs/1708.05234).

- [48] Shuo Yang, Yuanjun Xiong, Chen Change Loy, and Xiaoou Tang. Face detection through scale-friendly deep convolutional networks. 2017. URL: <http://arxiv.org/abs/1706.02863>, [arXiv:1706.02863](#).
- [49] Peiyun Hu and Deva Ramanan. Finding tiny faces. 2016. URL: <http://arxiv.org/abs/1612.04402>, [arXiv:1612.04402](#).
- [50] Nannan Wang, Xinbo Gao, Dacheng Tao, Heng Yang, and Xuelong Li. Facial feature point detection: A comprehensive survey. 275:50–65, 2018. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0925231217308202>, [doi:10.1016/j.neucom.2017.05.013](#).
- [51] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. A survey on face detection in the wild: Past, present and future. 138:1–24, 2015. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1077314215000727>, [doi:10.1016/j.cviu.2015.03.015](#).
- [52] Erjin Zhou, Zhimin Cao, and Qi Yin. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *CoRR*, abs/1501.04690, 2015. URL: <http://arxiv.org/abs/1501.04690>, [arXiv:1501.04690](#).
- [53] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, June 2014. [doi:10.1109/CVPR.2014.220](#).
- [54] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. pages 815–823. URL: <http://arxiv.org/abs/1503.03832>, [arXiv:1503.03832](#), [doi:10.1109/CVPR.2015.7298682](#).
- [55] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. URL: <http://arxiv.org/abs/1411.7923>, [arXiv:1411.7923](#).
- [56] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015.
- [57] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *CoRR*, abs/1710.08092, 2017. URL: <http://arxiv.org/abs/1710.08092>, [arXiv:1710.08092](#).
- [58] Ankan Bansal, Anirudh Nanduri, Carlos D. Castillo, Rajeev Ranjan, and Rama Chellappa. Umdfaces: An annotated face dataset for training deep networks. *CoRR*, abs/1611.01484, 2016. URL: <http://arxiv.org/abs/1611.01484>, [arXiv:1611.01484](#).
- [59] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *CoRR*, abs/1607.08221, 2016. URL: <http://arxiv.org/abs/1607.08221>, [arXiv:1607.08221](#).
- [60] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. *CoRR*, abs/1512.00596, 2015. URL: <http://arxiv.org/abs/1512.00596>, [arXiv:1512.00596](#).
- [61] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. *CoRR*, abs/1807.11649, 2018. URL: <http://arxiv.org/abs/1807.11649>, [arXiv:1807.11649](#).

- [62] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, June 1991. doi:[10.1109/CVPR.1991.139758](https://doi.org/10.1109/CVPR.1991.139758).
- [63] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, 13(6):1450–1464, Nov 2002. doi:[10.1109/TNN.2002.804287](https://doi.org/10.1109/TNN.2002.804287).
- [64] I. Naseem, R. Togneri, and M. Bennamoun. Linear regression for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):2106–2112, Nov 2010. doi:[10.1109/TPAMI.2010.128](https://doi.org/10.1109/TPAMI.2010.128).
- [65] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, July 1997. doi:[10.1109/34.598228](https://doi.org/10.1109/34.598228).
- [66] H. Hu, P. Zhang, and F. De La Torre. Face recognition using enhanced linear discriminant analysis. *IET Computer Vision*, 4(3):195–208, Sep. 2010. doi:[10.1049/iet-cvi.2009.0024](https://doi.org/10.1049/iet-cvi.2009.0024).
- [67] Kwang In Kim, Keechul Jung, and Hang Joon Kim. Face recognition using kernel principal component analysis. *IEEE Signal Processing Letters*, 9(2):40–42, 2002. doi:[10.1109/97.991133](https://doi.org/10.1109/97.991133).
- [68] Francis Bach and Michael Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 03 2003. doi:[10.1162/153244303768966085](https://doi.org/10.1162/153244303768966085).
- [69] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using laplacian faces. *IEEE transactions on pattern analysis and machine intelligence*, 27:328–40, 04 2005. doi:[10.1109/TPAMI.2005.55](https://doi.org/10.1109/TPAMI.2005.55).
- [70] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. URL: <https://science.sciencemag.org/content/290/5500/2323>, arXiv: <https://science.sciencemag.org/content/290/5500/2323.full.pdf>, doi:[10.1126/science.290.5500.2323](https://doi.org/10.1126/science.290.5500.2323).
- [71] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, Dec 2006. doi:[10.1109/TPAMI.2006.244](https://doi.org/10.1109/TPAMI.2006.244).
- [72] Baochang Zhang, Shiguang Shan, Xilin Chen, and Wen Gao. Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 16:57–68, 02 2007. doi:[10.1109/TIP.2006.884956](https://doi.org/10.1109/TIP.2006.884956).
- [73] Wenchao Zhang, Shiguang Shan, Xilin Chen, and Wen Gao. Local gabor binary patterns based on mutual information for face recognition. *Int. J. Image Graphics*, 7:777–793, 10 2007. doi:[10.1142/S021946780700291X](https://doi.org/10.1142/S021946780700291X).
- [74] L. Wiskott, J. . Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. In *Proceedings of International Conference on Image Processing*, volume 1, pages 129–132 vol.1, Oct 1997. doi:[10.1109/ICIP.1997.647401](https://doi.org/10.1109/ICIP.1997.647401).

- [75] Baochang Zhang, Yongsheng Gao, Sanqiang Zhao, and Jianzhuang Liu. Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor. *Image Processing, IEEE Transactions on*, 19:533 – 544, 03 2010. doi: [10.1109/TIP.2009.2035882](https://doi.org/10.1109/TIP.2009.2035882).
- [76] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. 2016.
- [77] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. 2018. URL: <http://arxiv.org/abs/1804.08348>, arXiv:1804.08348.
- [78] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. 28(5):807–813, 2010. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0262885609001711>, doi:10.1016/j.imavis.2009.08.002.
- [79] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101. IEEE, 2010. URL: <http://ieeexplore.ieee.org/document/5543262/>, doi:10.1109/CVPRW.2010.5543262.
- [80] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205. IEEE Comput. Soc, 1998. URL: <http://ieeexplore.ieee.org/document/670949/>, doi:10.1109/AFGR.1998.670949.
- [81] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. 19(3):34–41, 2012. URL: <http://ieeexplore.ieee.org/document/6200254/>, doi:10.1109/MMUL.2012.26.
- [82] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Acted Facial Expressions In The Wild Database. 2011.
- [83] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. From individual to group-level emotion recognition: EmotiW 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017*, pages 524–528. ACM Press, 2017. URL: <http://dl.acm.org/citation.cfm?doid=3136755.3143004>, doi:10.1145/3136755.3143004.
- [84] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112. IEEE, 2011. URL: <http://ieeexplore.ieee.org/document/6130508/>, doi:10.1109/ICCVW.2011.6130508.
- [85] Abhinav Dhall, O.V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: EmotiW 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, pages 423–426. ACM Press, 2015. URL: <http://dl.acm.org/citation.cfm?doid=2818346.2829994>, doi:10.1145/2818346.2829994.

- [86] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. 2013. URL: <http://arxiv.org/abs/1307.0414>, [arXiv:1307.0414](https://arxiv.org/abs/1307.0414).
- [87] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 317–321. IEEE, 2005. URL: <http://ieeexplore.ieee.org/document/1521424/>, [doi:10.1109/ICME.2005.1521424](https://doi.org/10.1109/ICME.2005.1521424).
- [88] Michel F. Valstar and Maja Pantic. Induced Disgust , Happiness and Surprise : an Addition to the MMI Facial Expression Database. 2010.
- [89] P.E. Ekman and W V. Friesen. Facial action coding system (facs). 01 2002.
- [90] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. Presentation and validation of the radboud faces database. 24(8):1377–1388, 2010. URL: <http://www.tandfonline.com/doi/abs/10.1080/02699930903485076>, [doi:10.1080/02699930903485076](https://doi.org/10.1080/02699930903485076).
- [91] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.
- [92] Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. Facial expressions of emotion are not culturally universal. 109(19):7241–7244, 2012. URL: <https://www.pnas.org/content/109/19/7241>, [doi:10.1073/pnas.1200155109](https://doi.org/10.1073/pnas.1200155109).
- [93] Pedro M. Ferreira, Filipe Marques, Jaime S. Cardoso, and Ana Rebelo. Physiological inspired deep neural networks for emotion recognition. 6:53930–53943, 2018. URL: <https://ieeexplore.ieee.org/document/8472816/>, [doi:10.1109/ACCESS.2018.2870063](https://doi.org/10.1109/ACCESS.2018.2870063).
- [94] Ting Zhang. Facial expression recognition based on deep learning: A survey. In Fatos Xhafa, Srikanta Patnaik, and Albert Y. Zomaya, editors, *Advances in Intelligent Systems and Interactive Applications*, volume 686, pages 345–352. Springer International Publishing, 2018. URL: http://link.springer.com/10.1007/978-3-319-69096-4_48, [doi:10.1007/978-3-319-69096-4_48](https://doi.org/10.1007/978-3-319-69096-4_48).
- [95] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models-their training and application. 61(1):38–59, 1995. URL: <http://linkinghub.elsevier.com/retrieve/pii/S1077314285710041>, [doi:10.1006/cviu.1995.1004](https://doi.org/10.1006/cviu.1995.1004).
- [96] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. 23(6):681–685, 2001. URL: <http://ieeexplore.ieee.org/document/927467/>, [doi:10.1109/34.927467](https://doi.org/10.1109/34.927467).
- [97] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. 27(6):803–816, 2009. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0262885608001844>, [doi:10.1016/j.imavis.2008.08.005](https://doi.org/10.1016/j.imavis.2008.08.005).

- [98] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. 29(6):915–928, 2007. URL: <http://ieeexplore.ieee.org/document/4160945/>, doi:10.1109/TPAMI.2007.1110.
- [99] Chengjun Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. 11(4):467–476, 2002. URL: <http://ieeexplore.ieee.org/document/999679/>, doi:10.1109/TIP.2002.999679.
- [100] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. The computer expression recognition toolbox (CERT). In *Face and Gesture 2011*, pages 298–305. IEEE, 2011. URL: <http://ieeexplore.ieee.org/document/5771414/>, doi:10.1109/FG.2011.5771414.
- [101] Z. Wang and Z. Ying. Facial expression recognition based on local phase quantization and sparse representation. In *2012 8th International Conference on Natural Computation*, pages 222–225, May 2012. doi:10.1109/ICNC.2012.6234551.
- [102] Thiago H. H. Zavaschi, Alceu S. Britto, Luiz E. S. Oliveira, and Alessandro L. Koerich. Fusion of feature sets and classifiers for facial expression recognition. 40(2):646 – 655, 2013. URL: <http://www.sciencedirect.com/science/article/pii/S095741741200930X>, doi:<https://doi.org/10.1016/j.eswa.2012.07.074>.
- [103] Z. Li, J. Imai, and M. Kaneko. Facial-component-based bag of words and phog descriptor for facial expression recognition. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 1353–1358, Oct 2009. doi:10.1109/ICSMC.2009.5346254.
- [104] Y. Lv, Z. Feng, and C. Xu. Facial expression recognition via deep learning. In *2014 International Conference on Smart Computing*, pages 303–308, Nov 2014. doi:10.1109/SMARTCOMP.2014.7043872.
- [105] Xiaoming Zhao, Xugan Shi, and Shiqing Zhang. Facial expression recognition via deep learning. *IETE Technical Review*, 32(5):347–355, 2015. URL: <https://doi.org/10.1080/02564602.2015.1017542>, arXiv:<https://doi.org/10.1080/02564602.2015.1017542>, doi:10.1080/02564602.2015.1017542.
- [106] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. 18(7):1527–1554, 2006. URL: <http://www.mitpressjournals.org/doi/10.1162/neco.2006.18.7.1527>, doi:10.1162/neco.2006.18.7.1527.
- [107] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, June 2014. doi:10.1109/CVPR.2014.233.
- [108] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, pages 435–442. ACM Press, 2015. URL: <http://dl.acm.org/citation.cfm?doid=2818346.2830595>, doi:10.1145/2818346.2830595.

- [109] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. Real-time Convolutional Neural Networks for Emotion and Gender Classification. *arXiv e-prints*, page arXiv:1710.07557, October 2017. [arXiv:1710.07557](https://arxiv.org/abs/1710.07557).
- [110] Shima Alizadeh and Azar Fazel. Convolutional neural networks for facial expression recognition. 2017. URL: <http://arxiv.org/abs/1704.06756>, [arXiv:1704.06756](https://arxiv.org/abs/1704.06756).
- [111] Ciprian Adrian Corneanu, Marc Oliu Simon, Jeffrey F. Cohn, and Sergio Escalera Guerrero. Survey on RGB, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. 38(8):1548–1568, 2016. URL: <https://ieeexplore.ieee.org/document/7374704/>, [doi:10.1109/TPAMI.2016.2515606](https://doi.org/10.1109/TPAMI.2016.2515606).
- [112] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, June 2009. URL: <http://dl.acm.org/citation.cfm?id=1577069.1577078>.
- [113] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL: <http://arxiv.org/abs/1405.0312>, [arXiv:1405.0312](https://arxiv.org/abs/1405.0312).
- [114] Tomasz Marciniak, Agata Chmielewska, Radoslaw Weychan, Marianna Parzych, and Adam Dabrowski. Influence of low resolution of images on reliability of face detection and recognition. 74(12):4329–4349. URL: <http://link.springer.com/10.1007/s11042-013-1568-8>, [doi:10.1007/s11042-013-1568-8](https://doi.org/10.1007/s11042-013-1568-8).
- [115] P. Jonathon Phillips, Patrick Grother, and Ross Micheals. Evaluation methods in face recognition. In Stan Z. Li and Anil K. Jain, editors, *Handbook of Face Recognition*, pages 551–574. Springer London. URL: http://link.springer.com/10.1007/978-0-85729-932-1_21, [doi:10.1007/978-0-85729-932-1_21](https://doi.org/10.1007/978-0-85729-932-1_21).
- [116] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. 22(10):1345–1359. URL: <https://ieeexplore.ieee.org/document/5288526/>, [doi:10.1109/TKDE.2009.191](https://doi.org/10.1109/TKDE.2009.191).
- [117] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 471–478. IEEE. URL: <https://ieeexplore.ieee.org/document/8614364/>, [doi:10.1109/SIBGRAPI.2018.00067](https://doi.org/10.1109/SIBGRAPI.2018.00067).
- [118] Yoanna Martinez-Diaz, Heydi Mendez-Vazquez, Leyanis Lopez-Avila, Leonardo Chang, L. Enrique Sucar, and Massimo Tistarelli. Toward more realistic face recognition evaluation protocols for the YouTube faces database. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 526–5268. IEEE. URL: <https://ieeexplore.ieee.org/document/8575544/>, [doi:10.1109/CVPRW.2018.00082](https://doi.org/10.1109/CVPRW.2018.00082).