UNIVERSIDADE D
COIMBRA

David Alexandre Mendes Carreira

DEEP FACE RECOGNITION FOR ONLINE STUDENT IDENTIFICATION

Thesis submitted to the University of Coimbra in fulfilment of the requirements of the Master's Degree in Engineering Physics under the scientific supervision of PhD David Portugal and MsC José Faria and presented to the Physics Department of the Faculty of Sciences and Technology of the University of Coimbra.

June 2023

Contents

Acknowledgments

Abstract

Resumo

Chapter 1

Introduction

1.1 Context

The outbreak of the COVID-19 pandemic tested the entire world on several levels and changed the concept of what is ”normal” thereafter. The devastating health, economic and social consequences that COVID caused, spanned a need to develop novel solutions, for almost every aspect of our lives, that facilitate the adaptation to the new world we're living in.

Educational systems were no exception. In the midst of the pandemic, governments around the world forced institutions to shut down and stop the customary in-person regimen of teaching. By April 2020, most universities transitioned to an adapted remote learning [90] that lacked proper support due to the unanticipated nature of the events, leading to new challenges, in particular, the legitimacy of moments of evaluation performed remotely. To counter this problem, different approaches can be taken, namely, changing the method of evaluation, suppressing it altogether [7] or, when possible, implement a continuous monitoring solution such as TrustID ref?. However, there are still unresolved issues that must be addressed in order to implement an end-to-end solution capable of assuring the success of such systems.

One core aspect of them is the face verification task, therefore, the data obtained directly influences the performance. Due to the purpose of the application and expected devices to be used, what is obtained can be classified as from an unconstrained nature. Even though the capture of image is consensual, there is no way of controlling the conditions of capturing the visual data and consequent results. This can be attributed to the fact that it is anticipated that the system will be executed in a laptop or a smartphone, thus the capture device might not be ideal. The more probable input method will be a webcam or the smartphone's front facing camera, so a high variation in pose, resolution, illumination, etc. is not unforeseeable.

Another detail that must be regarded, is the processing power available to execute the system. It is common for the equipments used to have a deficiency of it[1], which is not suitable for high-demanding applications, as its improved accuracy comes at the cost of increased computational overhead, which can make real-time continuous monitoring unfeasible.

In conclusion, the method of choice must take the aforesaid into consideration and be a trade-off between accuracy and computational strain, while also being invariant, to a certain degree, to the posed challenges of capturing the required data.

1.2 Dissertation structure

This dissertation will be divided into different chapters that partitions themselves into sections and subsections. Chapter one relates to the introduction of the dissertation, it will present the context and motivation behind the problem and structure of the document. The document continues to the second chapter, it starts with an overview of the History of AI, carries out a survey about the topic's

[1] According to the February 2023 Steam hardware survey, roughly 5% of its users do not have a dedicated GPU.

State-Of-The-Art, presented and summarized through the step-by-step analysis of the pipeline of a Face Recognition system, and ends with a comparison table of the discussed methods. In chapter number three, the implemented methods and experiments are described. The forth chapter will present and discuss the results. Finally, chapter five, will draw conclusions of the work achieved in the past several months and prospects for the future.

Chapter 2

State of The Art

2.1 History of AI

The following sections present a broad overview of the history of Artificial Intelligence (AI) by presenting important articles in order for the reader to be able to have a notion of the progress that has been made over the past decades, the hardships encountered and how important AI is in our lives.

Philosophy

On October 1950, in his article *Computing Machinery and Intelligence*, Alan Turing questioned: "Can machines think?" [80]. At the time, the question was too meaningless to answer since not only the theory but also the technology available weren't developed enough. Nonetheless, Turing still predicted that in the future there would be computers that could, effectively, display human-like intelligence and discernment under the conditions proposed on the aforementioned article.

Relevant events to the birth of AI

The breakthroughs of AI are predominant, and its importance in our everyday life is undeniable, but the theory behind it has several early roots. The interest in

the area grew immensely with, for example, all the Turing's theoretical research, the proposal of the first mathematical Artificial Neuron model in 1943 by Warren McCulloch and Walter Pitts (based of binary inputs and output) [58] and in 1949 Donald Hebb revolutionized the way the artificial neurons were treated by proposing what is known as the Hebb's rule[1]. Taking into consideration the latter two, but specially Hebb's proposals, Belmont Farley and Westley Clark implemented in 1954 one of the first successful Artificial Neural Networks (ANN), also called Perceptron, composed of two layers of 128 artificial neurons with weighted inputs [29]. Over the span of approximately ten years, multiple researches were performed attempting to computerize the human brain. However, only in 1956, during the *Dartmouth Summer Research Project on Artificial Intelligence* [57], was the term "Artificial Intelligence" firstly proposed by John McCarthy *et al.*, beginning what is now considered to be the birth of AI [101].

The fading of general interest

The succeeding two decades following the Dartmouth conference were filled with important developments, with special emphasis in the works published in 1958 by Frank Rosenblatt (generalized the Farley and Clark training to multilayer networks rather than only two) [70], the 1959 General Problem Solver implemented by Allen Newel *et al.* (a program intended to work as a universal problem solver that was capable of solving exercises such as the Towers of Hanoi[2]) [62] and the ELIZA a natural language processing tool program developed by Joseph Weizenbaum between 1964 and 1966 [88]. Unfortunately, part of the interest and development around AI met an unforeseen fade after criticisms about the exagger-

[1] "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." [35], meaning that when two neurons fire together their relation is strengthened.

[2] *The Towers of Hanoi* is a game with 3 stacks of increasingly smaller disks. The goal is to stack them one at a time, so that they are arranged in a decreasing radius manner.

ated public funding [33] and the Marvin Minksy and Seymour Papert 1969 book *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain* [60] that reported on the problems of the Perceptron network. The overall sentiments regarding this topic of research was of doubt and fear of no progress, mainly due to the spending and two issues raised by Minsky and Papert: the ANN couldn't solve linear inseparable problems[3] and there were limitations due to a lack of sufficient computing power to handle the processing of multi-layer large networks.

A better approach

Minksy and Papert raised important questions, but it shouldn't have discouraged other researchers from further trying, since they failed to acknowledge alternative approaches that had already solved those exact problems. As previously stated, the model proposed by McCulloch and Pitts, later improved by the Farley-Clark implementation and, finally, Rosenblatt, couldn't handle linearly inseparable classes. A possible solution for cases like this started being studied in the 1960s [42, 71] and, although it didn't produce relevant results, in 1965 Alexey Ivakhnenko and Valentin Lapa [40] were, indeed, successful in implementing what is nowadays considered to be the first deep learning network of its kind [73]. In 1971 Ivakhnenko also published an article describing a deep learning network with 8 layers that was already able to create hierarchical internal representations [41].

The years progressed, in 1979 Kunihiko Fukushima introduced the first Convolutional Neural Network (CNN) in a structural sense, due to its similarity to the architecture of modern ones of this category. Ten years later, Yann LeCun *et al.* applied for the first time a revolutionizing training algorithm called Backpropagation to a CNN [46], creating what is now a pillar for most of the modern competition winning networks in computer vision [73] and employing the term "convolution"

[3] That is, if two sets X and Y in $\mathbb{R}^d$ can't be divided by a hyperplane such that the elements of X and Y stay on opposing sides, then we're dealing with linear inseparable classes [28]

for the first known time [51]. He also introduced the MNIST (**M**odified **N**ational **I**nstitute of **S**tandards and **T**echnology) dataset, a collection of handwritten digits [48], that to this day is still one of the most famous benchmarks in Machine Learning. Backpropagation can be traced back many decades, but the modern version was first described by Seppo Linnainmaa (1970) [52], implemented for the first time by Stuart Dreyfus (1973) [26] and, finally in 1986, David Rummelhart *et al.* popularized it in the Neural Network's (NN) domain by demonstrating the growing usefulness of internal representations [72].

The importance of Convolutional Neural Networks

The study on Neural Networks continued and there were improvements on all types of architectures [36, 89] with special highlight to pioneering Neural Networks processed by GPUs[4] (standard NN in 2004 by [63] and CNN in 2006 by [14]). But there's a well deserved particular attention related to the developments of CNNs due to their great performance in image related tasks when compared to others networks, as proven by LeCun in his 1998 paper [48]. Some relevant examples: in 2003 the MNIST record was broken by Patrice Simard *et al.* [75], achieving an error rate of 0.4% (whereas a non-convolutional neural network by the same authors took the second place with 0.7%); three years later, the same benchmark had a new set low of 0.39% by Marc'Aurelio Ranzato *et al.* [68]; in 2009 a CNN by Yang *et al.* was the first network of this type to win an official international competition (TRECVid) [96]; a GPU implementation of a CNN [19] achieved superhuman vision performance in a competition (IJCNN 2011) in a *German Traffic Sign Recognition Benchmark* with a 0.56% error rate (0.78% for the best human performance, 1.69% for the second-best neural network contestant and 3.86% for the best non-neural method [76]). This last example conjoined with non-convolutional methods [66, 21] and the previously cited [14, 63], reinforces how fundamental GPUs were to further develop neural networks. To supplement

[4] Graphics Processing Unit

even more the importance of CNNs and GPUs, only a year later, Alex Krizhevsky *et al.* proposed a Deep CNN trained by GPUs that was the first one to win the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), achieving an error rate of 15.3% while the second place obtained 26.2% [44].

The year of 2012 was very important for Deep Learning, CNNs and Computer Vision, due to all the attention brought to many researches on this topic after several systems of this kind won image analysis competitions ([18, 20] and the very important previously mentioned [44]), beginning what's considered to be the start of the new wave, we're currently in, of interest in Artificial Intelligence, specially in the aforesaid topics [51].

2.2 Face Recognition

Face Recognition (FR) is a thoroughly debated and extensively researched task in the Computer Vision community for more than two decades [67], popularized in the early 1990s with the introduction of the Eigenfaces [81] or Fisherfaces [64] approaches. These methods projected faces in a low-dimensional subspace assuming certain distributions, but lacked the ability to handle uncontrolled facial changes that broke said assumptions, henceforth, bringing about face recognition approaches through local-features [17, 2] that, even though, presented considerable results, weren't distinctive or compact. Beginning in 2010, methods based on learnable filters arose [98, 49], but unfortunately revealed limitations when nonlinear variations were at stake.

Earlier methods for FR worked appropriately when the data was handpicked or generated on a constrained environment, however, they didn't scale adequately in the real world were there are large fluctuations in, particularly, pose, age, illumination, background scenario, the presence of facial occlusion [67] and many unimaginable more. These shortcomings can be dealt with by using Deep Learning, a framework of techniques that solves the nonlinear inseparable classes problem ref., more specifically a structure called Convolutional Neural Network (CNN) [85].

CNNs are an Artificial Neural Network (ANN) that exhibit a better performance on image or video-based tasks compared to other methods [48]. They were greatly hailed in 2012, after the AlexNet [44] victory, by a great margin, in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Just two years later, DeepFace [77] revolutionized the benchmarks scores by achieving state-of-the-art results that approached human performance, reinforcing even further the importance of Deep Learning and shifting the research path to be taken [85].

Given what has been stated so far and the proven robustness, performance, and overall results in computer vision ref. won competitions, the methods discussed in this dissertation will therefore deal exclusively with Deep Learning approaches.

For more information on other methods, please refer to [45].

2.3 A Face Recognition System

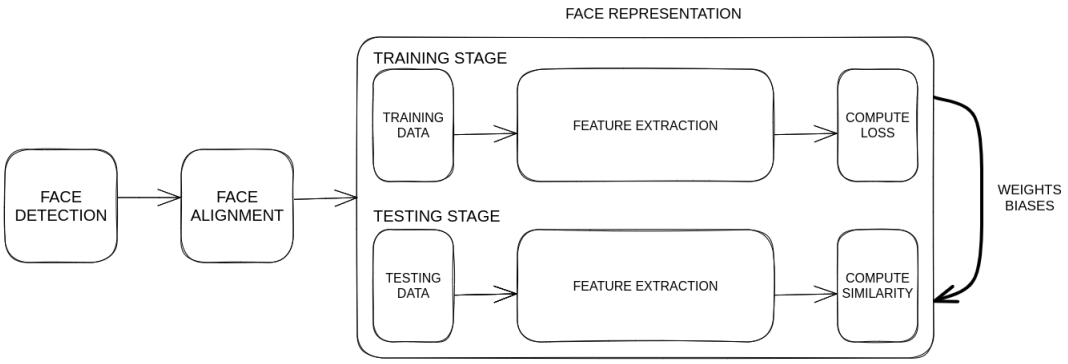


Figure 1: A typical face recognition pipeline, guided by the approach in [85].

According to Ranjan *et al.* [67], the goal of a FR system is to find, process and learn from a face, gathering as much information as possible, and as a result, it is one of the most widely implemented biometric system solutions in light of its versatility when facing real world application [27], such as military, public security and daily life.

By and large, all end-to-end automatic face recognition systems follow a sequential and modular[5] pipeline (Figure 1) composed of three pillar stages [85]: face detection, face alignment and face representation. First an image or video feed is used as an input then, as the name suggests, the **face detection** module is responsible for finding a face. Next, the **face alignment** phase applies spatial transformations to the data in order to normalize the faces' pictures (or frames, in the case where a video is used) to a standardized view. Finally, the **face representation** stage, makes use of deep learning techniques to learn discriminative

[5] Sequential because each stage relies on the output from the previous ones, and modular in the sense that each stage employs its own method and it can be modified to better adapt to specific tasks.

features that will allow the recognition.

All three stages have their individual importance and methods of implementation[6]. **Face detection** is achievable through classical approaches [83, 9] or deep methods, among them is [25] and the widely applied [104]. **Face alignment**, once again, can be accomplished through traditional measures [22, 56] or more modern ones, namely [39] or the aforementioned [104] which concurrently performs detection and alignment. To conclude, the **face representation** module is no exception, and can also be divided in two groups, regarding the methodology used. Some conventional systems were already mentioned, such as [64, 81], and the deep learning ones are the object of discussion of this dissertation and will be reviewed along the following sections, therefore, the focus will be on describing, with particular interest, the face representation stage.

2.3.1 Face Detection

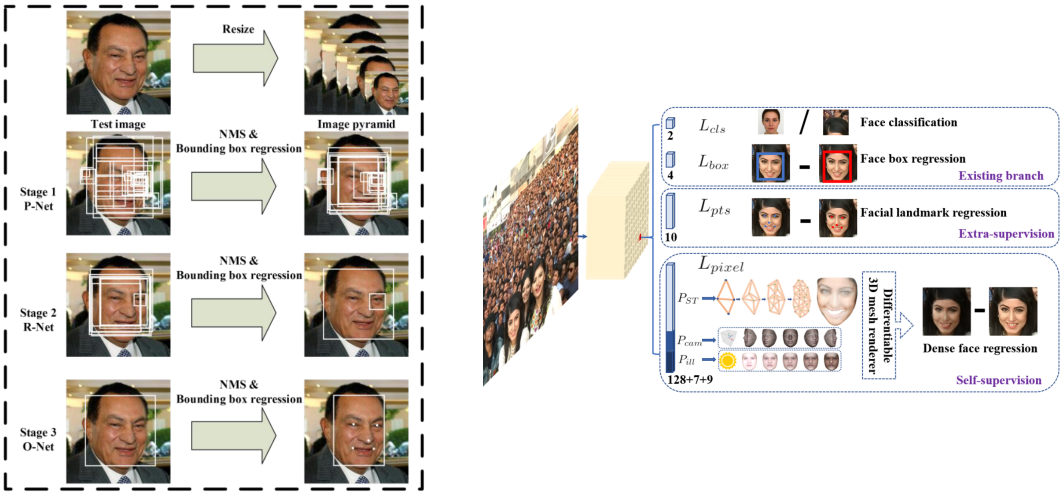
Face detection is the first step in any automatic facial recognition system. Given an input image to a face detector module, it is in charge of detecting every face in the picture and returning bounding-boxes coordinates, for each one, with a certain confidence score [27, 67].

Previously employed traditional face detectors cite here are incapable of detecting facial information when faced with challenges such as variants in image resolution, age, pose, illumination, race, occlusions or accessories (masks, glasses, makeup) [27, 67]. The progress in deep learning and increasing GPU power led DCNNs to become a viable and reliable option that solves said problems in face detection.

These techniques can be included in different categories. A more analytical

[6] For a deeper and extensive study, please refer to: [99] in the case of classic face detection approaches and [59] for deep learning based methods; [87] addresses traditional face alignment methods and is complemented with [27] for more up-to-date techniques; and [45] tackles classic face representation (add the following if needed) while X supplements the deep learning ones.

perspective [27] distributes the methods, depending upon their architecture or purpose of application, over seven categories: multi-stage, single-stage, anchor-based, anchor-free, multi-task learning, CPU real-time and, finally, problem-oriented. Additionally, being as the face detection problem can be seen as a specific task in a general object detection situation, it is no surprise that several works inherit from them and, therefore, some bases are referenced throughout the next list.



a) b)

Figure 2: Comparison between **a)** MTCNN: multi-stage, CPU real-time and multi-task learning, and **b)** RetinaFace: single-stage, anchor-based, CPU real-time and multi-task learning. MTCNN [104] proposes a series of bounding boxes then, through a series of refinement stages, the best solution and landmarks are found. RetinaFace [25] accomplishes, in a single-stage, face classification and bounding box regression by evaluating anchors, landmark localization and dense 3D projection for facial correspondence.

→ **Multi-stage** methods [25] include all the coarse-to-fine facial detectors that work in similar manner to the following two phases. First, bounding box proposals are generated by sliding a window through the input. Then, over one or several subsequent stages, false positives are rejected and the approved bounding boxes are refined. To complement, one widely applied object detection protocol that inspired face detection methods and perfectly describes the steps mentioned above is Faster R-CNN [69]. However, these methods can be slower and have a more

complex way of training [94].

→ **Single-stage** approaches [25] are the ones that perform classification and bounding box regression without the necessity of a proposal stage, producing highly dense face locations and scales. This structure takes inspiration, once again, from general object detectors, for example, the Single Shot MultiBox detector, commonly referred to as SSD [54]. Finally, the methods included in this class are more efficient, but can incur in compromised accuracy, when compared to multi-stage.

→ **Anchor-based** techniques [55, 25, 102] detect faces by predefining anchors with different settings (scales, strides, number, etc.) on the feature maps, then performing classification and bounding box regression on them until an acceptable output is found. As proven by Liu and Tang *et al.* [55], the choice of anchors highly influences the results of prediction. Hence, it is necessary to fine-tune them on a situation-by-situation basis, otherwise, there is a limitation in generalization. Furthermore, higher densities of anchors directly generate an increase in computational overhead.

→ **Anchor-free** procedures, obviously, do not need predefined anchors in order to find faces. Alternatively, these methods address the face detection by using different techniques. For example, DenseBox [38] which attempts to predict faces by processing each pixel as a bounding box, or CenterFace [94] that treats face detection as a key-point estimation problem by predicting the center of the face and bounding boxes. Even so, relating to the accuracy of anchor-free approaches, there's still room for improvement for false positives and stability in the training stage [27].

→ **Multi-task learning** are all the methodologies that conjointly performs other

tasks, namely facial landmark[7] localization, during face classification and bounding box regression [27]. CenterFace [94] is one example, and so it is the widely implemented MTCNN [104], which correlated bounding boxes and face landmarks. RetinaFace [25] is another state-of-the-art approach, it mutually detects faces, respective landmarks and performs dense 3D face regression.

→ **CPU real-time** methods, as the name suggests, include the detectors that can run on a single CPU core, in real-time, for VGA-resolution input images. A face detector can achieve great results in terms of accuracy, but for real world applications, its use can be too computational heavy, therefore, can't be deployed in real time (specially in devices that do not have a GPU) [27]. MTCNN [104], Faceboxes [105], CenterFace [94] or RetinaFace [25] are examples of this category.

→ **Problem-oriented** is a category that includes the detectors that are projected to resolve a wide range of specific problems, for example, faces that are tiny, partially occluded, blurred or scale-invariant face detection [27]. PyramidBox [78] is an example that solves the partial occluded and blurry faces, and HR [37] tackles the tiny faces challenge.

Although this distribution can create some overlap among the categories, it is superior due to the simplicity of inferring what defines each category and being a more fine-grained way of classifying techniques when compared to others, namely the dual categorical division by [67] that groups the methods in region[8] or sliding-window[9] based.

[7] A facial landmark is a key-point in a face that contributes with important geometric information, namely the eyes, nose, mouth, etc. [30]

[8] Region-based approaches creates thousands of generic object-proposals for every image, and subsequently, a DCNN classifies if a face is present in any of them.

[9] Sliding-window approaches centers on using a DCNN to compute a face detection score and bounding box at every location in a feature map.

2.3.2 Face Alignment

Face Alignment, or facial landmark detection [13], is the second stage of the face recognition pipeline, and has the objective of calibrating the detected face to a canonical layout, through landmark-based or landmark-free approaches, in order to leverage the core final stage of face representation [27].

Despite the fact that traditional face alignment methods are very accurate, that only occurs in constrained circumstances. Therefore, once again, to address that issue, deep learning-based methods are the solution to perform an accurate facial landmark localization that realistically scales to real world scenarios [30].

Furthermore, face alignment, can be accomplished through two categories of methods: landmark-based and landmark-free.

→ **Landmark-based alignment** is a category of methods that exploits the facial landmarks with the aim of, through spatial transformations, calibrating the face to an established layout [27]. This can be accomplished through: coordinate regression, heatmap regression or 3D Model Fitting. **Coordinate regression-based** methodologies [30, 53, 104] consider the landmark localization as a numerical objective, i.e. a regression, thus an image is fed to a DCNN and it will output a vector of landmark coordinates. **Heatmap Regression** [24, 91, 15] is different from coordinate regression because, although it is a numerical objective task, the output is not a coordinate vector, but a map of likelihood of landmarks' locations. Finally, **3D Model Fitting** [8, 13, 93]is the category that integrates methods that consider the relation between 2D facial landmarks and the 3D shape of a generic face. The particularity of them is the reconstruction of the 3D face from a 2D face image that is then projected over a plane in order to obtain the landmarks.

→ **Landmark-free alignment**, on the other hand, integrates the approaches that do not rely on landmarks as a reference to align the face, in contrast, these type of methods incorporate the alignment into a DCNN that gives, as a result, an

aligned face [27]. An example of an end-to-end method that does not depend on facial landmarks is RDCFace [106], and it rectifies distortions, applies alignment transformations and executes face representation. Hayat et al. [34] proposes a method that deals with extreme head poses. The process to register faces in an image with high pose variance can be quite challenging and often demands complex pre-processing, namely landmark localization, therefore, to address that, a DCNN is employed that does not rely on landmark localization and concomitantly register and represent faces.

As can be seen from the previous section, this step in the face recognition process can be accomplished, very sporadically, through standalone methods that process the detected face from the previous stage, but generally joint detection and alignment methods (and sometimes even face representation), previously referenced in the multi-task learning definition, are the optimal choice [13].

2.3.3 Face Representation

Finally, Face Representation is the last stage of the Face Recognition process. It is responsible for processing the aligned face from the previous stage and mapping the produced feature representation to a feature space, in which features from the same person are closer together and those that are different stand further apart from each other [27].

According to the literature [27, 50, 67, 74, 85], there's a consensus about how Face Recognition can be performed in two settings of operation: face verification and face identification. This distinction is only made possible due to the approaches available in the Face Representation stage that can leverage one, the other or both.

→ **Face verification**, also referred to as **face authentication**, is a one-to-one match, and it's the action of verifying if the query face matches the identity that's

being claimed. These principles are used in biometric systems such as self-service immigration clearance using E-passport. [50]

→ **Face identification**, also called **face recognition**, is a one-to-many correlation process that compares a query face to a database of faces and associates it to the corresponding match (or matches). A typical use case is to identify someone in a watchlist or surveillance videos. [50]

The overall pipeline comes to a conclusion in this module, however, in reality, it goes further than that. As can be seen in (Figure 1), due to its importance for the face recognition problem, it's highlighted the inherent pipeline of the Face Representation stage, henceforth, it shall be discussed in depth in the next section.

2.4 Face Representation Pipeline

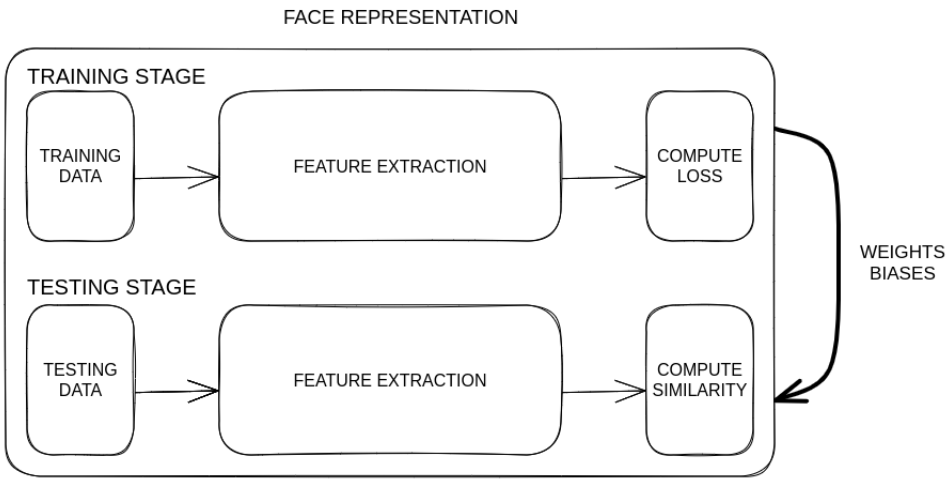


Figure 3: Face Representation pipeline, guided by the approach in [85].

As show in (Figure 3), Face Representation is a two-step module composed of a training and testing stage. So as to be capable of performing face recognition, in either a verification or identification manner, a face representation system needs to learn robust, invariant and discriminative features that can distinguish identities [67].

To meet these requirements, the feature extractor must first be trained properly by taking data from previous stages and outputting a feature representation that's compared to the desired value using a loss function[10] [47, 85]. After that, everything is ready for the testing stage, where the face recognition *per se* occurs by calculating a similarity score for the feature representation produced by the trained feature extractor, and dictating if the identity belongs to the same person (face verification) or if it matches any identity (face identification) [67].

[10] Also referred to as a cost or objective function. [47]

2.4.1 Datasets: Training and Testing Data

As been discussed throughout this dissertation, Deep Learning techniques can solve the problem of handling unconstrained scenarios, where there are variations in pose, illumination, occlusion, and so forth. To support this, in the past few years, datasets have been developed with this in mind so to be able to provide a large and diverse set of both training data, allowing for adequate regularization to unseen circumstances, and testing data that benchmarks the face recognition system in, as similar as possible, unconstrained real world scenarios [27].

Training Data

When developing a deep face recognition system it's essential to keep in mind its necessity to adapt, and that's where the dataset used for training comes at play. Large training datasets are essential for face recognition [65], but large-scale is not enough. There must be a balance between the depth (number of unique identities) and the breadth/width (number of images per identity) [6, 11], and it will lead to different effects.

On one hand, a training dataset that is deep will help the face recognition system to produce more discriminative feature representations, since it will have a great number of identities to learn from. On the other hand, a wider set will have more images per identity, therefore, variations in pose, expressions, illuminations, occlusions, background clutter, image quality, accessories, and so forth [5] can be introduced and ultimately lead to feature representations more robust to them.

In the following pages, a mix of training datasets, of overall relevance and more geared towards the purpose of this dissertation, will be reviewed.

→ **CASIA-WebFace** [97], composed of 494,414 face images and 10,575 identities, was proposed as a novel dataset to overcome the problem of data dependence in face recognition and improve comparability across different methods. By training

on the same dataset, methods can be better evaluated and compared.

→ **VGGFace** [65] was published alongside a homonymous face recognition method and, once again, with the objective of combating the lack of available large scale public datasets. It contains 2,6 million images and 2,622 different identities and a curated version, where incorrect image labels were hand-removed by humans, has 800,000 images for the same amount of identities.

→ **MS-Celeb-1M**'s [32] first intention was to provide a novel benchmark to identify celebrities that solves name ambiguities by linking a face with an entity key in a knowledge base. Second, it aimed at solving the gap in available large-scale datasets by providing a training set with, approximately, 10 million images and 100 thousand identities. Unfortunately, it is a dataset known for the presence of noisy labels.

→ **MegaFace** [61] introduced a benchmark for million-scale face recognition and provided a public large-scale training dataset that integrated 4,753,320 faces over 672,057 identities. The main difference compared to the previously mentioned datasets is that MegaFace does not use celebrities as subjects, in contrast it leverages the photographs released by Flickr under the Creative Commons license.

→ **VGGFace2** [11] is another large-scale dataset, and its main goals are: 1) covering numerous identities, 2) reduce labeling noise through automatic and manual filtering and, finally, 3) represent more realistic unconstrained scenarios due to a novel dataset generation pipeline that gathers images with a broad range of poses, age, illumination and ethnicity. All in all, this resulted in a dataset comprised of 3,31 million faces of 9131 subjects.

→ **UMDFaces-Videos** [6] is a video-based dataset composed of 22,075 videos of 3,107 subjects with 3,735,476 human annotated frames with great variation in image quality, pose, expressions and lightning. It was proposed during a study how

the performance of a face verification models is impacted by the effects of: 1) the type of media used for training (only videos or still images vs a mixture of both), 2) the width and depth of a dataset, 3) the label's noise and 4) the alignment of the faces.

→ **Celeb-500k** [10] is another large-scale proposed with two issues in mind: the disparity in the scale of public datasets when compared with private ones, and determining the impact in performance from intra- and inter-class variations. That being so, Celeb-500k, consisting of 50 million images from 500 thousand persons, and Celeb-500k-2R, a cleaned version of the previous, comprised of 25 million aligned faces of 245 thousand identities, are released.

→ **IMDb-Face** [84] proposes a new dataset with based on a manually cleaned revision of MS-Celeb-1M and MegaFace. The growing demand for large-scale datasets introduced a new variable to take into consideration: the time available to annotate the data. Datasets that are well-annotated and have an enormous amount of data are notably expensive and time-consuming to develop. Therefore, automatic measures to clean the data were used, so it's expected for a certain degree of noise to be introduced in a dataset. After selecting a subset from both the originals datasets, 2 million images were manually cleaned and resulted in 1,7 million images of 59 thousand celebrities.

→ **QMUL-SurvFace** [16] is a dataset introduced as benchmark in the *Surveillance Face recognition Challenge* for face recognition in a surveillance context. By data-mining 17 public person re-identification datasets, it achieves 463,507 facial images of 15,573 identities collected in uncooperative surveillance scenarios. Consequently, it presents a high variance in resolution, motion blur, pose, occlusion, illumination and background clutter.

→ **MS1MV2** [23] is another well know dataset. It was proposed in the ArcFace

face recognition method paper and consists of a semi-automatic refinement of the previously mentioned MS-Celeb-1M, resulting in 5,8 million images of 85 thousand identities.

→ **Glint360K** [4] is a training set presented in the Partial FC method paper. It was generated by merging and cleaning the aforementioned Celeb-500K and MS1MV2 datasets, which resulted in 17 million images of 360 thousand individuals.

→ **CAFR** (rev 2021) [107]

→ **MDMFR** [82]

→ **WebFace260M** [109] takes a giant leap in closing the gap between public available datasets and private ones. Partnered with a time-constrained face recognition protocol, the original paper presented an enormous 260 million faces and 4 million identities noisy dataset, an automatically cleaned, high quality training set with 42 million faces over 2 million identities (WebFace42M), and a smaller scale training dataset derived from the WebFace42M that has 10% of its data (WebFace4M).

→ **DigiFace-1M** [5] is a novel approach that revolutionizes the way of training face recognition models. It is a fully synthetic dataset that proposes mitigating three very relevant problems present in the majority of the conventional datasets: 1) ethical issues, 2) label noise and 3) data bias. The dataset is divided in two parts: part one contains 720 thousand images from 10 thousand identities and part two has 500 thousand images with 100 thousand identities, for a total of 1,22 million images and 110 thousand unique identities.

Although some of the previously referenced datasets do have a benchmarking component, they're generally employed to train algorithms, opposed to the following datasets that are described as "benchmarks". A common denominator throughout the machine learning domain, and more specifically deep learning, is

a general difficulty in separating concepts with a single line. This is a non-linear science in its nature.

Dataset	Year	Availability	Images/vids	Depth	Breadth	Distribution	Description
CASIA-WebFace	2014						
VGGFace	2015						
MS-Celeb-1M	2016						
MegaFace	2016						
VGGFace2	2017						
UMDFaces-Videos	2017						
Celeb-500k	2018						
IMDb-Face	2018						
QMUL-SurvFace	2018						
MS1MV2	2019						
Glint360k	2021						
CAFR	2021*						
WebFace260M	2021						
MDMFR	2022						
DigiFace-1M	2023						

Table 1

Testing Data

After the training is completed the performance of the system needs to be evaluated on different challenges to properly access if it scales (or generalizes or applies or performs in) to real-world scenarios. A test dataset can be chosen for specific hurdles, for instance, cross-pose, cross-age, racial variations, quality assessment, and so forth [27].

→ **LFW**

→ **YTF**

→ **IJB-C**

→ **MegaFace**

→ **Trillion Pairs**

→ **DiF**

→ **Fair Face**

→ **RFW**

→ **XQLFW** (2022)

→ **FaVCI2D** (2022)

2.4.2 Feature Extraction

There are several types of Neural Networks architectures, but Convolutional Neural Networks (CNNs or Convnets) are probably the most widely implemented model overall [95, 51] with successful applications in the domains of Computer

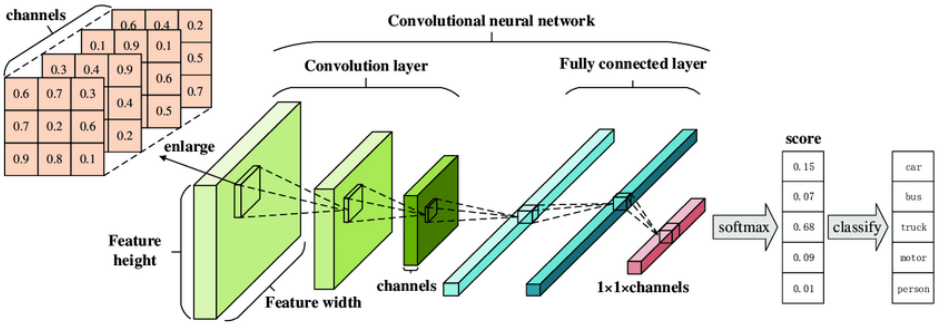


Figure 4: Architecture of a Convolutional Neural Network [43].

Vision [44, 77, 79, 103] or Natural Language Processing[1, 86, 92]. In the CNN category itself there are different variants, but they all abide the fundamental structure of a feedforward hierarchical multi-layer network (Figure 4). Feedforward because the information only flows in a singular direction without cycling [100], hierarchical because the higher complexity internal representations are learned from lower ones [47, 108] and multi-layer because it is composed of a series of stages, blocks or layers: the raw data is fed to an input layer, forwarded to a sequence of intercalating convolutional and pooling layers, transmitted to a stage of one or more fully-connected layers [47, 95, 31, 3]. The convolutional layer is designed to extract feature representations by being composed of kernels (or filter banks [47]) that compute feature maps through element-wise product, to which is applied a nonlinear activation function [31, 95]. Next is the pooling layer, that's responsible for reducing the spatial size of the input data [31] and joining identical features [47]. Finally, the fully connected layers, and their core function is to perform high logic and generate semantic information [31].

Using CNNs for Computer Vision tasks is not an arbitrary choice, but due to the fact that the network design can benefit from the intrinsic characteristics of the input data, consequently performing really well in image related applications [47, 12]. In the first place, images have an array-like structure with numerous elements, namely, each pixel has an assigned value organized in a grid-like manner [95]. In

the second place, there's an inherent correlation between local groups of values, which creates distinguishable motifs [47]. Finally, the local values of images are invariant to location, that is, a certain composition should have the same value independently of the spatial location in the picture [47]. Therefore, the following key, unique features potentiate the previously stated efficient performance [12]:

1. Designed to process multidimensional arrays [47];
2. Shared weights between the same features in different locations;
3. Automatically identifies the relevant features without any human supervision, hence, small amounts of preprocessing [3, 51];
4. Local connections (or receptive fields/sparse connectivity) [3];
5. Pooling layers that reduces the spatial size of the input data.

The ensemble of features 2, 4 and 5 enable an invariance of the network to small shifts, distortions and rotations [31, 47], while 2, 3, 4 and 5 helps to reduce the complexity of the model, and as a result training it is easier[31, 51].

Transfer Learning

2.4.3 Loss

Chapter 3

Methodology

3.1 Joint Face Detection and Alignment

3.2 Face Representation

Chapter 4

Results

Chapter 5

Conclusion

Bibliography

[1] Ossama Abdel-Hamid et al. "Convolutional Neural Networks for Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.10 (Oct. 2014), pp. 1533–1545. ISSN: 2329-9304. DOI: 10.1109/TASLP.2014.2339736 (cit. on p. 25).

[2] T. Ahonen, A. Hadid, and M. Pietikäinen. "Face Description with Local Binary Patterns: Application to Face Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.12 (2006). Cited By :4611, pp. 2037–2041. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2006.244 (cit. on p. 9).

[3] Laith Alzubaidi et al. "Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions". In: *Journal of Big Data* 8.1 (Mar. 2021), p. 53. ISSN: 2196-1115. DOI: 10.1186/s40537-021-00444-8. (Visited on 02/09/2023) (cit. on pp. 25, 26).

[4] Xiang An et al. *Partial FC: Training 10 Million Identities on a Single Machine*. Comment: 8 pages, 9 figures. Jan. 2021. arXiv: 2010.05222 [cs]. (Visited on 04/28/2023) (cit. on p. 22).

[5] Gwangbin Bae et al. "DigiFace-1M: 1 Million Digital Face Images for Face Recognition". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Datasets. Waikoloa, HI, USA: IEEE, Jan. 2023, pp. 3515–3524. ISBN: 978-1-66549-346-8. DOI: 10.1109/WACV56688.2023.00352. (Visited on 02/28/2023) (cit. on pp. 19, 22).

[6] Ankan Bansal et al. *The Do's and Don'ts for CNN-based Face Verification*. Comment: 10 pages including references, added more experiments on deeper vs wider dataset (section 3.2). Sept. 2017. arXiv: 1705.07426 [cs]. (Visited on 04/28/2023) (cit. on pp. 19, 20).

[7] Maria Barron Rodriguez et al. *Remote Learning During the Global School Lockdown: Multi-Country Lessons*. World Bank, Aug. 2021. DOI: 10.1596/36141. (Visited on 03/13/2023) (cit. on p. 1).

[8] Chandrasekhar Bhagavatula et al. *Faster Than Real-time Facial Alignment: A 3D Spatial Transformer Network Approach in Unconstrained Poses*. Comment: International Conference on Computer Vision (ICCV) 2017. Sept. 2017. arXiv: 1707.05653 [cs]. (Visited on 04/15/2023) (cit. on p. 15).

[9] S. Charles Brubaker et al. "On the Design of Cascades of Boosted Ensembles for Face Detection". In: *International Journal of Computer Vision* 77.1 (May 2008), pp. 65–86. ISSN: 1573-1405. DOI: 10.1007/s11263-007-0060-1 (cit. on p. 11).

[10] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. "Celeb-500K: A Large Training Dataset for Face Recognition". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. Oct. 2018, pp. 2406–2410. DOI: 10.1109/ICIP.2018.8451704 (cit. on p. 21).

[11] Qiong Cao et al. *VGGFace2: A Dataset for Recognising Faces across Pose and Age*. Comment: This paper has been accepted by IEEE Conference on Automatic Face and Gesture Recognition (F&G), 2018. (Oral). May 2018. arXiv: 1710.08092 [cs]. (Visited on 04/27/2023) (cit. on pp. 19, 20).

[12] Weipeng Cao et al. "A Review on Neural Networks with Random Weights". In: *Neurocomputing* 275 (Jan. 2018), pp. 278–287. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2017.08.040. (Visited on 02/09/2023) (cit. on pp. 25, 26).

[13] Fengju Chang et al. *FacePoseNet: Making a Case for Landmark-Free Face Alignment*. Aug. 2017. arXiv: 1708.07517 [cs]. (Visited on 04/15/2023) (cit. on pp. 15, 16).

[14] Kumar Chellapilla, Sidd Puri, and Patrice Simard. "High Performance Convolutional Neural Networks for Document Processing". In: () (cit. on p. 7).

[15] Lisha Chen, Hui Su, and Qiang Ji. "Face Alignment With Kernel Density Deep Neural Network". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 6991–7001. ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00709. (Visited on 04/15/2023) (cit. on p. 15).

[16] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. *Surveillance Face Recognition Challenge*. Comment: The QMUL-SurvFace challenge is publicly available at https://qmul-survface.github.io/. Aug. 2018. arXiv: 1804.09691 [cs]. (Visited on 04/28/2023) (cit. on p. 21).

[17] Chengjun Liu and H. Wechsler. "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition". In: *IEEE Transactions on Image Processing* 11.4 (Apr. 2002), pp. 467–476. ISSN: 1941-0042. DOI: 10.1109/TIP.2002.999679 (cit. on p. 9).

[18] D.C. Ciresan et al. "Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images". In: *NIPS* 25 (2012). Export Date: 26 January 2023; Cited By: 92, pp. 2852–2860 (cit. on p. 8).

[19] Dan Cireşan et al. "A Committee of Neural Networks for Traffic Sign Classification". In: *The 2011 International Joint Conference on Neural Networks*. July 2011, pp. 1918–1921. DOI: 10.1109/IJCNN.2011.6033458 (cit. on p. 7).

[20] Dan C. Cireşan et al. "Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*. Ed. by Kensaku Mori et al. Lecture

Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 411–418. ISBN: 978-3-642-40763-5. DOI: 10.1007/978-3-642-40763-5_51 (cit. on p. 8).

[21] Dan Claudiu Cireşan et al. "Deep, Big, Simple Neural Nets for Handwritten Digit Recognition". In: *Neural Computation* 22.12 (Dec. 2010), pp. 3207–3220. ISSN: 0899-7667. DOI: 10.1162/NECO_a_00052. (Visited on 01/25/2023) (cit. on p. 7).

[22] T.F Cootes et al. "View-Based Active Appearance Models". In: *Image and Vision Computing* 20.9 (Aug. 2002), pp. 657–664. ISSN: 0262-8856. DOI: 10.1016/S0262-8856(02)00055-0 (cit. on p. 11).

[23] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: () (cit. on p. 21).

[24] Jiankang Deng et al. *Joint Multi-view Face Alignment in the Wild*. Comment: submit to IEEE Transactions on Image Processing. Aug. 2017. arXiv: 1708.06023 [cs]. (Visited on 04/15/2023) (cit. on p. 15).

[25] Jiankang Deng et al. *RetinaFace: Single-stage Dense Face Localisation in the Wild*. May 2019. arXiv: 1905.00641 [cs]. (Visited on 04/13/2023) (cit. on pp. 11–14).

[26] Stuart E. Dreyfus. "The Computational Solution of Optimal Control Problems with Time Lag". In: *IEEE Transactions on Automatic Control* 18.4 (1973). Cited by: 32, pp. 383–385. DOI: 10.1109/TAC.1973.1100330 (cit. on p. 7).

[27] Hang Du et al. "The Elements of End-to-end Deep Face Recognition: A Survey of Recent Advances". In: *ACM Computing Surveys* 54.10s (Jan. 2022), pp. 1–42. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3507902. (Visited on 03/07/2023) (cit. on pp. 10–16, 19, 24).

[28] D. Elizondo. “The Linear Separability Problem: Some Testing Methods”. In: *IEEE Transactions on Neural Networks* 17.2 (Mar. 2006), pp. 330–344. ISSN: 1045-9227. DOI: 10.1109/TNN.2005.860871. (Visited on 01/24/2023) (cit. on p. 6).

[29] B. Farley and W. Clark. “Simulation of Self-Organizing Systems by Digital Computer”. In: *Transactions of the IRE Professional Group on Information Theory* 4.4 (1954), pp. 76–84. DOI: 10.1109/TIT.1954.1057468 (cit. on p. 5).

[30] Zhen-Hua Feng et al. *Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks*. Comment: 11 pages, 6 figures, 6 tables. Oct. 2018. arXiv: 1711.06753 [cs]. (Visited on 04/14/2023) (cit. on pp. 14, 15).

[31] Jiuxiang Gu et al. “Recent Advances in Convolutional Neural Networks”. In: *Pattern Recognition* 77 (May 2018), pp. 354–377. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2017.10.013. (Visited on 02/09/2023) (cit. on pp. 25, 26).

[32] Yandong Guo et al. *MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition*. July 2016. arXiv: 1607.08221 [cs]. (Visited on 04/25/2023) (cit. on p. 20).

[33] Michael Haenlein and Andreas Kaplan. “A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence”. In: *California Management Review* 61 (July 2019), p. 000812561986492. DOI: 10.1177/0008125619864925 (cit. on p. 6).

[34] Munawar Hayat et al. “Joint Registration and Representation Learning for Unconstrained Face Identification”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, July 2017, pp. 1551–1560. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.169. (Visited on 04/15/2023) (cit. on p. 16).

[35] Donald Olding Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, 1949. ISBN: 978-0-471-36727-7 (cit. on p. 5).

[36] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. (Visited on 01/25/2023) (cit. on p. 7).

[37] Peiyun Hu and Deva Ramanan. "Finding Tiny Faces". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, July 2017, pp. 1522–1530. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.166. (Visited on 04/14/2023) (cit. on p. 14).

[38] Lichao Huang et al. *DenseBox: Unifying Landmark Localization with End to End Object Detection*. Sept. 2015. arXiv: 1509.04874 [cs]. (Visited on 04/13/2023) (cit. on p. 13).

[39] Xiehe Huang et al. *PropagationNet: Propagate Points to Curve to Learn Structure Information*. Comment: 10 pages, 8 figures, 8 tables, CVPR2020. June 2020. arXiv: 2006.14308 [cs]. (Visited on 04/08/2023) (cit. on p. 11).

[40] A G Ivakhnenko and V G Lapa. "Cybernetic Predicting Devices". In: () (cit. on p. 6).

[41] A. G. Ivakhnenko. "Polynomial Theory of Complex Systems". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-1.4 (1971), pp. 364–378. DOI: 10.1109/TSMC.1971.4308320 (cit. on p. 6).

[42] Roger David Joseph. *Contributions to Perceptron Theory*. Cornell Aeronautical Laboratory, 1960 (cit. on p. 6).

[43] Xu Kang, Bin Song, and Fengyao Sun. "A Deep Similarity Metric Method Based on Incomplete Data for Traffic Anomaly Detection in IoT". In: *Applied Sciences* 9 (Jan. 2019), p. 135. DOI: 10.3390/app9010135 (cit. on p. 25).

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. (Visited on 01/26/2023) (cit. on pp. 8, 9, 25).

[45] Erik Learned-Miller et al. "Labeled Faces in the Wild: A Survey". In: *Advances in Face Detection and Facial Image Analysis*. Ed. by Michal Kawulok, M. Emre Celebi, and Bogdan Smolka. Cham: Springer International Publishing, 2016, pp. 189–248. ISBN: 978-3-319-25956-7 978-3-319-25958-1. DOI: 10.1007/978-3-319-25958-1_8. (Visited on 03/09/2023) (cit. on pp. 10, 11).

[46] Y. LeCun et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4 (1989), pp. 541–551. DOI: 10.1162/neco.1989.1.4.541 (cit. on p. 6).

[47] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* 521.7553 (May 2015), pp. 436–444. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14539 (cit. on pp. 18, 25, 26).

[48] Yann LeCun et al. "Gradient-Based Learning Applied to Document Recognition". In: (1998) (cit. on pp. 7, 9).

[49] Z. Lei, M. Pietikainen, and S.Z. Li. "Learning Discriminant Face Descriptor". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.2 (2014). Cited By :287, pp. 289–302. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2013.112 (cit. on p. 9).

[50] Stan Z. Li and Anil K. Jain, eds. *Handbook of Face Recognition*. London: Springer, 2011. ISBN: 978-0-85729-931-4 978-0-85729-932-1. DOI: 10.1007/978-0-85729-932-1. (Visited on 02/14/2023) (cit. on pp. 16, 17).

[51] Zewen Li et al. "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects". In: *IEEE Transactions on Neural Networks*

and Learning Systems 33.12 (Dec. 2022), pp. 6999–7019. ISSN: 2162-2388. DOI: 10.1109/TNNLS.2021.3084827 (cit. on pp. 7, 8, 24, 26).

[52] Seppo Linnainmaa. "The Representation of the Cumulative Rounding Error of an Algorithm as a Taylor Expansion of the Local Rounding Errors". PhD thesis. Master's Thesis (in Finnish), Univ. Helsinki, 1970 (cit. on p. 7).

[53] Hao Liu et al. "Two-Stream Transformer Networks for Video-Based Face Alignment". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.11 (Nov. 2018), pp. 2546–2554. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2017.2734779 (cit. on p. 15).

[54] Wei Liu et al. "SSD: Single Shot MultiBox Detector". In: vol. 9905. Comment: ECCV 2016. 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0_2. arXiv: 1512.02325 [cs]. (Visited on 04/13/2023) (cit. on p. 13).

[55] Yang Liu et al. *HAMBox: Delving into Online High-quality Anchors Mining for Detecting Outer Faces*. Comment: 9 pages, 6 figures. arXiv admin note: text overlap with 1802.09058 by other authors. Dec. 2019. arXiv: 1912.09231 [cs]. (Visited on 04/13/2023) (cit. on p. 13).

[56] Brais Martinez et al. "Local Evidence Aggregation for Regression-Based Facial Point Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.5 (May 2013), pp. 1149–1163. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2012.205 (cit. on p. 11).

[57] J McCarthy et al. "A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE". In: () (cit. on p. 5).

[58] Warren S Mcculloch and Walter Pitts. "A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY". In: () (cit. on p. 5).

[59] Shervin Minaee et al. *Going Deeper Into Face Detection: A Survey*. Mar. 2021 (cit. on p. 11).

[60] Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press, 1969 (cit. on p. 6).

[61] Aaron Nech and Ira Kemelmacher-Shlizerman. "Level Playing Field for Million Scale Face Recognition". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Datasets. Honolulu, HI: IEEE, July 2017, pp. 3406–3415. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.363. (Visited on 02/28/2023) (cit. on p. 20).

[62] Allen Newell, John C Shaw, and Herbert A Simon. "Report on a General Problem Solving Program". In: *IFIP Congress*. Vol. 256. Pittsburgh, PA. 1959, p. 64 (cit. on p. 5).

[63] Kyoung-Su Oh and Keechul Jung. "GPU Implementation of Neural Networks". In: *Pattern Recognition* 37.6 (June 2004), pp. 1311–1314. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2004.01.013 (cit. on p. 7).

[64] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.7 (July 1997), pp. 711–720. ISSN: 1939-3539. DOI: 10.1109/34.598228 (cit. on pp. 9, 11).

[65] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition". In: *Procedings of the British Machine Vision Conference 2015*. VGG-Face. Swansea: British Machine Vision Association, 2015, pp. 41.1–41.12. ISBN: 978-1-901725-53-7. DOI: 10.5244/C.29.41. (Visited on 02/27/2023) (cit. on pp. 19, 20).

[66] Rajat Raina, Anand Madhavan, and Andrew Y. Ng. "Large-Scale Deep Unsupervised Learning Using Graphics Processors". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. New York, NY, USA: Association for Computing Machinery, June 2009, pp. 873–880. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553486. (Visited on 01/25/2023) (cit. on p. 7).

[67] Rajeev Ranjan et al. “Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans”. In: *IEEE Signal Processing Magazine* 35.1 (Jan. 2018), pp. 66–83. ISSN: 1558-0792. DOI: 10.1109/MSP.2017.2764116 (cit. on pp. 9–11, 14, 16, 18).

[68] Marc’ aurelio Ranzato et al. “Efficient Learning of Sparse Representations with an Energy-Based Model”. In: *Advances in Neural Information Processing Systems*. Vol. 19. MIT Press, 2006. (Visited on 01/25/2023) (cit. on p. 7).

[69] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. Comment: Extended tech report. Jan. 2016. arXiv: 1506.01497 [cs]. (Visited on 04/13/2023) (cit. on p. 12).

[70] F. Rosenblatt. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.” In: *Psychological Review* 65 (1958), pp. 386–408. ISSN: 1939-1471(Electronic),0033-295X(Print). DOI: 10.1037/h0042519 (cit. on p. 5).

[71] Frank Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, 1962 (cit. on p. 6).

[72] DE Rumelhart, GE Hinton, and RJ Williams. *Learning Internal Representations by Error Propagation”, in\Parallel Distributed Processing”, DE Rumelhart, JL McClelland Eds*. 1986 (cit. on p. 7).

[73] Jürgen Schmidhuber. “Deep Learning in Neural Networks: An Overview”. In: *Neural Networks* 61 (Jan. 2015), pp. 85–117. ISSN: 08936080. DOI: 10.1016/j.neunet.2014.09.003. (Visited on 01/24/2023) (cit. on p. 6).

[74] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682. arXiv: 1503.03832 [cs]. (Visited on 02/16/2023) (cit. on p. 16).

[75] P.Y. Simard, D. Steinkraus, and J.C. Platt. "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis". In: *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* Vol. 1. Edinburgh, UK: IEEE Comput. Soc, 2003, pp. 958–963. ISBN: 978-0-7695-1960-9. DOI: 10.1109/ICDAR.2003.1227801. (Visited on 01/25/2023) (cit. on p. 7).

[76] J. Stallkamp et al. "Man vs. Computer: Benchmarking Machine Learning Algorithms for Traffic Sign Recognition". In: *Neural Networks*. Selected Papers from IJCNN 2011 32 (Aug. 2012), pp. 323–332. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2012.02.016. (Visited on 01/25/2023) (cit. on p. 7).

[77] Yaniv Taigman et al. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, June 2014, pp. 1701–1708. ISBN: 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.220. (Visited on 02/13/2023) (cit. on pp. 9, 25).

[78] Xu Tang et al. *PyramidBox: A Context-assisted Single Shot Face Detector*. Comment: 21 pages, 12 figures. Aug. 2018. arXiv: 1803.07737 [cs]. (Visited on 04/14/2023) (cit. on p. 14).

[79] Jonathan Tompson et al. *Efficient Object Localization Using Convolutional Networks*. Comment: 8 pages with 1 page of citations. June 2015. arXiv: 1411.4280 [cs]. (Visited on 02/13/2023) (cit. on p. 25).

[80] A. M. Turing. "I.—COMPUTING MACHINERY AND INTELLIGENCE". In: *Mind* LIX.236 (Oct. 1950), pp. 433–460. ISSN: 1460-2113, 0026-4423. DOI: 10.1093/mind/LIX.236.433. (Visited on 01/13/2023) (cit. on p. 4).

[81] Matthew Turk and Alex Pentland. "Eigenfaces for Recognition". In: *Journal of Cognitive Neuroscience* 3.1 (Jan. 1991), pp. 71–86. ISSN: 0898-929X. DOI: 10.1162/jocn.1991.3.1.71. (Visited on 03/07/2023) (cit. on pp. 9, 11).

[82] Naeem Ullah et al. “A Novel DeepMaskNet Model for Face Mask Detection and Masked Facial Recognition”. In: *Journal of King Saud University - Computer and Information Sciences* 34.10, Part B (Nov. 2022), pp. 9905–9914. ISSN: 1319-1578. DOI: 10.1016/j.jksuci.2021.12.017. (Visited on 02/27/2023) (cit. on p. 22).

[83] P. Viola and M. Jones. “Rapid Object Detection Using a Boosted Cascade of Simple Features”. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.* Vol. 1. Dec. 2001, pp. I–I. DOI: 10.1109/CVPR.2001.990517 (cit. on p. 11).

[84] Fei Wang et al. *The Devil of Face Recognition Is in the Noise.* Comment: accepted to ECCV’18. July 2018. arXiv: 1807.11649 [cs]. (Visited on 04/28/2023) (cit. on p. 21).

[85] Mei Wang and Weihong Deng. “Deep Face Recognition: A Survey”. In: *Neurocomputing* 429 (Mar. 2021), pp. 215–244. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.10.081. (Visited on 02/27/2023) (cit. on pp. 9, 10, 16, 18).

[86] Mingxuan Wang et al. “genCNN: A Convolutional Architecture for Word Sequence Prediction”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Beijing, China: Association for Computational Linguistics, July 2015, pp. 1567–1576. DOI: 10.3115/v1/P15-1151. (Visited on 02/13/2023) (cit. on p. 25).

[87] Nannan Wang et al. “Facial Feature Point Detection: A Comprehensive Survey”. In: *Neurocomputing* 275 (Jan. 2018), pp. 50–65. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2017.05.013 (cit. on p. 11).

[88] Joseph Weizenbaum. “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine”. In: *Commu-*

nications of the ACM 9.1 (Jan. 1966), pp. 36–45. ISSN: 0001-0782. DOI: 10.1145/365153.365168. (Visited on 01/18/2023) (cit. on p. 5).

[89] J. Weng, N. Ahuja, and T.S. Huang. "Cresceptron: A Self-Organizing Neural Network Which Grows Adaptively". In: *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*. Vol. 1. June 1992, 576–581 vol.1. DOI: 10.1109/IJCNN.1992.287150 (cit. on p. 7).

[90] Chris J Winstead. "Remote Microelectronics Laboratory Education in the COVID-19 Pandemic". In: *2022 Intermountain Engineering, Technology and Computing (IETC)*. May 2022, pp. 1–6. DOI: 10.1109/IETC54973.2022.9796805 (cit. on p. 1).

[91] Wayne Wu et al. *Look at Boundary: A Boundary-Aware Face Alignment Algorithm*. Comment: Accepted to CVPR 2018. Project page: https://wywu.github.io/projects/ May 2018. arXiv: 1805.10483 [cs]. (Visited on 04/15/2023) (cit. on p. 15).

[92] Lingyun Xiang et al. "A Convolutional Neural Network-Based Linguistic Steganalysis for Synonym Substitution Steganography". In: *Mathematical Biosciences and Engineering* 17.2 (2020), pp. 1041–1058. ISSN: 1551-0018. DOI: 10.3934/mbe.2020055. (Visited on 02/13/2023) (cit. on p. 25).

[93] Shengtao Xiao et al. "Recurrent 3D-2D Dual Learning for Large-Pose Facial Landmark Detection". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017, pp. 1642–1651. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.181. (Visited on 04/15/2023) (cit. on p. 15).

[94] Yuanyuan Xu et al. *CenterFace: Joint Face Detection and Alignment Using Face as Point*. Comment: 11 pages, 3 figures. A demo of CenterFace can be available at https://github.com/Star-Clouds/CenterFace. Nov. 2019. arXiv: 1911.03599 [cs]. (Visited on 04/13/2023) (cit. on pp. 13, 14).

[95] Rikiya Yamashita et al. “Convolutional Neural Networks: An Overview and Application in Radiology”. In: *Insights into Imaging* 9.4 (Aug. 2018), pp. 611–629. ISSN: 1869-4101. DOI: 10.1007/s13244-018-0639-9. (Visited on 02/09/2023) (cit. on pp. 24, 25).

[96] Ming Yang et al. “Detecting Human Actions in Surveillance Videos”. In: 2009 TREC Video Retrieval Evaluation Notebook Papers. Cited by: 26. 2009 (cit. on p. 7).

[97] Dong Yi et al. *Learning Face Representation from Scratch*. Nov. 2014. arXiv: 1411.7923 [cs]. (Visited on 04/25/2023) (cit. on p. 19).

[98] Z. Cao et al. “Face Recognition with Learning-Based Descriptor”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. June 2010, pp. 2707–2714. ISBN: 1063-6919. DOI: 10.1109/CVPR.2010.5539992 (cit. on p. 9).

[99] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. “A Survey on Face Detection in the Wild: Past, Present and Future”. In: *Computer Vision and Image Understanding* 138 (Sept. 2015), pp. 1–24. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2015.03.015 (cit. on p. 11).

[100] Andreas Zell. “Simulation Neuronaler Netze”. In: 1994 (cit. on p. 25).

[101] Caiming Zhang and Yang Lu. “Study on Artificial Intelligence: The State of the Art and Future Prospects”. In: *Journal of Industrial Information Integration* 23 (Sept. 2021), p. 100224. ISSN: 2452414X. DOI: 10.1016/j.jii.2021.100224. (Visited on 01/11/2023) (cit. on p. 5).

[102] Changzheng Zhang, Xiang Xu, and Dandan Tu. *Face Detection Using Improved Faster RCNN*. Feb. 2018. arXiv: 1802.02142 [cs]. (Visited on 04/13/2023) (cit. on p. 13).

[103] Yu-Dong Zhang et al. “Improved Breast Cancer Classification Through Combining Graph Convolutional Network and Convolutional Neural Network”. In: *Information Processing & Management* 58.2 (Mar. 2021), p. 102439. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2020.102439 (cit. on p. 25).

[104] Kaipeng Zhang et al. “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks”. In: *IEEE Signal Processing Letters* 23.10 (Oct. 2016). Comment: Submitted to IEEE Signal Processing Letters, pp. 1499–1503. ISSN: 1070-9908, 1558-2361. DOI: 10.1109/LSP.2016.2603342. arXiv: 1604.02878 [cs]. (Visited on 04/13/2023) (cit. on pp. 11, 12, 14, 15).

[105] Shifeng Zhang et al. *FaceBoxes: A CPU Real-time Face Detector with High Accuracy*. Comment: Accepted by IJCB 2017; Added references; Released codes. Dec. 2018. arXiv: 1708.05234 [cs]. (Visited on 04/14/2023) (cit. on p. 14).

[106] He Zhao et al. “RDCFace: Radial Distortion Correction for Face Recognition”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2020, pp. 7718–7727. ISBN: 978-1-72817-168-5. DOI: 10.1109/CVPR42600.2020.00774. (Visited on 02/28/2023) (cit. on p. 16).

[107] Jian Zhao, Shuicheng Yan, and Jiashi Feng. “Towards Age-Invariant Face Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.1 (Jan. 2022), pp. 474–487. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2020.3011426 (cit. on p. 22).

[108] Xinqi Zhu and Michael Bain. *B-CNN: Branch Convolutional Neural Network for Hierarchical Classification*. Comment: 9 pages, 8 figures. Oct. 2017. DOI: 10.48550/arXiv.1709.09890. arXiv: 1709.09890 [cs]. (Visited on 02/13/2023) (cit. on p. 25).

[109] Zheng Zhu et al. “WebFace260M: A Benchmark Unveiling the Power of Million-Scale Deep Face Recognition”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Datasets. Nashville, TN, USA: IEEE, June 2021, pp. 10487–10497. ISBN: 978-1-66544-509-2. DOI: 10.1109/CVPR46437.2021.01035. (Visited on 02/28/2023) (cit. on p. 22).