0.1 Face Recognition

Face Recognition (FR) is a thoroughly debated and extensively researched task in the Computer Vision community for more than two decades [24], popularized in the early 1990s with the introduction of the Eigenfaces [28] or Fisherfaces [23] 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 [6, 2] that, even though, presented considerable results, weren't distinctive or compact. Beginning in 2010, methods based on learnable filters arose [35, 17], 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 [24] 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) [30].

CNNs are an Artificial Neural Network (ANN) that exhibit a better performance on image or video-based tasks compared to other methods [16]. They were greatly hailed in 2012, after the AlexNet [13] victory, by a great margin, in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Just two years later, DeepFace [26] 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 [30].

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 [14].

0.2 A Face Recognition System

According to Ranjan *et al.* [24], 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 [9], such as military, public security and daily life.

By and large, all end-to-end automatic face recognition systems follow a sequential and modular[1] pipeline (Figure 1) composed of three pillar stages [30]: 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 transformations to the data, such as crop and/or rotation, in order to normalize the faces' pictures (or frames, in the case where a video is used) to standardized coordinates. Finally, the **face representation** stage, makes use of deep learning techniques to learn discriminative features that will allow the recognition.

All three stages have their individual importance and methods of implementation[2]. Face detection is achievable through classical approaches [29, 4] or deep methods, among them is [8] and the widely applied [39]. Face alignment, once again, can be accomplished through traditional measures [7, 21] or more modern ones, namely [11] or the aforementioned [39] 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 [23, 28], and the deep learning ones are the objective 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.

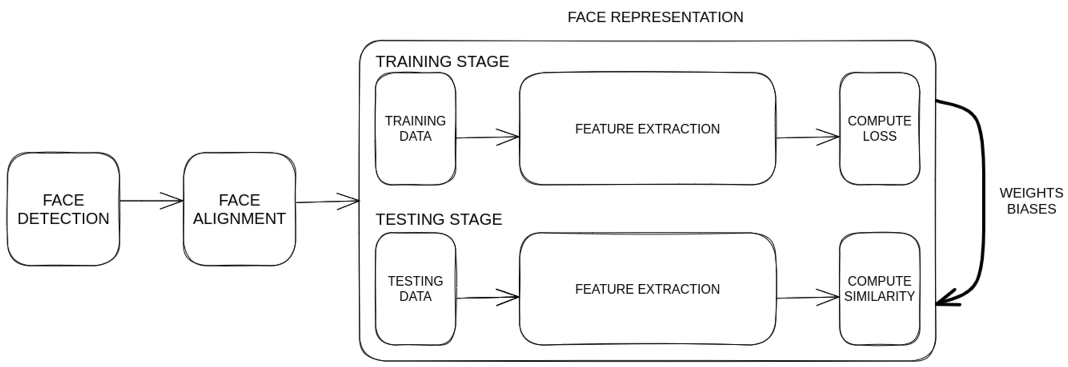

Figure 1: A typical face recognition pipeline, guided by the approach in [30].

[1] 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.

[2] For a deeper and extensive study, please refer to: [36] in the case of classic face detection approaches and [22] for deep learning based methods; [32] addresses traditional face alignment methods and is complemented with [9] for more up-to-date techniques; and [14] tackles classic face representation (add the following if needed) while X supplements the deep learning ones

0.2.1 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 [9, 24].

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) [9, 24]. 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 perspective [9] 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 basis are referenced throughout the next list.

Multi-stage methods [8] 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 [25].

Single-stage approaches [8] 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 [19]. Finally, the methods included in this class are more efficient, but can incur in compromised accuracy, when compared to multi-stage.

Anchor-based techniques [20, 8] detect faces by taking predefined anchors with different settings (scales, strides, number) set on the feature maps, then performing classification and bounding box regression on them until an acceptable output is found.

Anchor-free

Multi-task learning

CPU real-time

Problem-oriented

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 [24] that groups the methods in region[3] or sliding-window[4] based.

0.2.2 Face Alignment

0.2.3 Face Representation

0.3 Face Representation Pipeline

0.3.1 Convolutional Neural Networks

There are several types of Neural Networks architectures, but Convolutional Neural Networks (CNNs or Convnets) are probably the most widely implemented model overall [34, 18] with successful applications in the domains of Computer Vision [13, 26, 27, 38] or Natural Language Processing[1, 31, 33]. In the CNN category itself there are different variants, but they all abide the fundamental structure of a feedforward hierarchical multi-layer network (Figure 1). Feedforward because the information only flows in a singular direction without cycling [37], hierarchical because the higher complexity internal representations are learned from lower ones [15, 40] 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 [15, 34, 10, 3]. The convolutional layer is designed to extract feature representations by being composed of kernels (or filter banks [15]) that compute feature maps through element-wise product, to which is applied a nonlinear activation function [10, 34]. Next is the pooling layer, that's responsible for reducing the spatial size of the input data [10] and joining identical features

[3] 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.

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

[15]. Finally, the fully connected layers, and their core function is to perform high logic and generate semantic information [10].

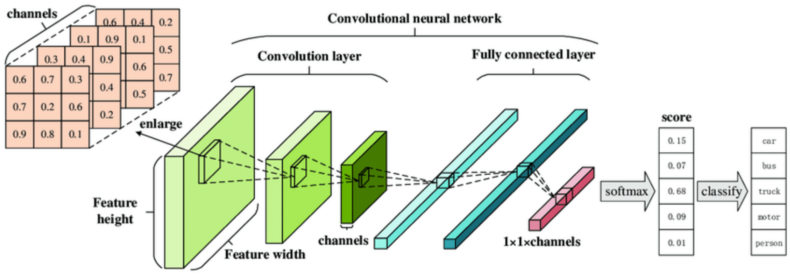

Figure 2: Architecture of a Convolutional Neural Network [12]

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 [15, 5]. 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 [34]. In the second place, there's an inherent correlation between local groups of values, which creates distinguishable motifs [15]. 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 [15]. Therefore, the following key, unique features potentiate the previously stated efficient performance [5]:

1. Designed to process multidimensional arrays [15];

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, 18];

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 [10, 15], while 2, 3, 4 and 5 helps to reduce the complexity of the model, and as a result training it is easier[10, 18].