Order number: . . . . .

Thesis submitted to the

UNIVERSITY OF MOHAMED BOUDIAF – MSILA

FACULTY OF MATHMATICS AND COMPUTER SCIENCE
DEPARTEMENT OF COMPUTER SCIENCE

In partial fulfillment of the requirements for the degree of

Master in **Computer science**

By

Amer Ouali Djamel Eddine

Achour Abd Eselm

Title of the thesis

Face feeling recognition algorithm using an artificial intelligent method

Under the supervision of

Dr. Mohamed Sahraoui

Composition of the jury

Dr. Abdelbaset Barket	University of Msila	President
Dr. Mohamed Sahraoui	University of Msila	Reporter
Dr. Imed-Eddine Debbi	University of Msila	Examiner

Join, 2022

ACKNOWLEDGMENTS

In the name of God, the most merciful,

The most merciful and his prophet Muhammad.

« صلى الله عليه وسلم »

We thank Almighty God.

We would like to express our deep appreciation to the final year project manager, "Dr. Mohamed Sahraoui", for his guidance andsupervision and for providing the necessary information regarding the AMF project.

We thank all members of the jury for agreeing to screen this work.

We offer all my thanks to our families, relatives and especially our parents, who have given to us all the support we need.

We thank both parts of our department: Professors and administration for their efforts to train us throughout the five years.

CONTENTS

LIST OF FIGURES

LIST OF ACRONYMS AND SYMBOLS

AI	Artificial Intelligence
IR	Image Recognition
SIFT	scale-invariant feature transform
HOG	histogram of oriented gradients
HMM	Hidden Markov Model
CNN	Convolution Neural Network
FER	Facial Emotion Recognition
LBP	Local Binary Pattern
MLP	Multi-Layer Perceptron
ReLU	Rectified Linear Unit
FACS	Facial Action Coding System
ASM	Active Shape Model
AAM	Active Appearance Model
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
ICA	Independent Component Analysis
CLM	Constrained Local Model
CLNF	Constrained Local Field
BN	Bayesian Network
SVM	Support Vector Machines
AU	Action Units
BOVM	Bag of Visual words
JAFFE	Japanese Female Facial Expression dataset
CK	Cohn-Kanade Dataset

LIST OF EQUATIONS

LIST OF TABLES

GENERAL INTRODUCTION

According to the remarkable development of smart and powerful devices, several related services are being developed. These services are designed to alternate human-centric services based on computer vision data extraction. One of the important fields is the automatic verification and recognition of individuals based on their physical appearance, behavioral traits and special traits, through biometric methods such as the face, fingerprint, signature and hand geometry. In particular, facial recognition technology can be used in many ways and in different manners. One of the latest use of this technology is emotion recognition which can be performed using various features, such as in [1] – [3]. Among these features, facial expressions are the most common due to reason that it contains many features useful for recognizing emotions.

Emotion recognition is the use of computers to detect human faces and analyze performance characteristics information [4,5]. The machine realizes the purpose of recognizing and understanding humans from emotional expression.

In order to reduce the false detection, improve the recognition accuracy and offer more robustness, machine learning techniques are introduced in this area. Recently, with the use of deep Learning in particular Convolutional Neural Networks (CNNs) [6], many features can be extracted, and learn how to have a decent facial expression recognition system. It should be noted that in the case of facial expressions, a lot of clues come from a few parts of the face, for example, the mouth and eyes, whereas other parts, such as the ears and hair, play a Small role in the output of emotional recognition such as happiness, sadness or anger [7].

In This work, we propose a new Framework based on deep learning to recognize facial expressions, which takes the above observation into account and uses the attention mechanism to focus on the prominent part of the face. By using an attentional convolutional network and unlike the used CNN algorithms that are performed using at Most four layers, our algorithm bases on the use of additional layers capable of achieving a very high resolution rate. More specifically, our work in this dissertation makes the following contributions:

• We propose an approach based on the attentional convolutional network, which can focus on the feature-rich parts, however, it outshines the face in modern brilliant works in precision.

• We train it on different data sets with an evaluation of the accuracy rate since it represents the

most important criteria used in the validation of such solutions.

The remainder of this dissertation is constructed through the following chapters:

Chapter1 is focused on image recognition to give an overview of image recognition technology and its fields as the global area of our work, in particular, we have focused on the field of facial recognition.

Chapter2 Focuses on the sub-field of facial emotions recognition as our special field of work, and presents the model based on convolutional neural networks CNN to recognize facial expression.

Chapter3 presents our proposed framework through the presenting of the environment of work, our conception and the different results with other previous works on different data sets.

CHAPTER 1

Image Recognition

1 Introduction

In this chapter, we give an overview of the image recognition domain in order to provide a global view of this domain focusing on the facial recognition field that presents the area to which facial emotions recognition belongs.

2 What is Image Recognition?

Image Recognition (IR), in the context of machine vision, is a term used for computer technologies that enable them to identify objects such as: places, actions, writing, or identify specific people, animals, or objects in images using special algorithms. It is an umbrella term for the process of training computers to see like humans and the mechanism of combining a camera and an artificial intelligence program to achieve image recognition [5].

The IR involves analyzing the pixels of the image to identify the image content as specific objects as shown in Figure 1.1.

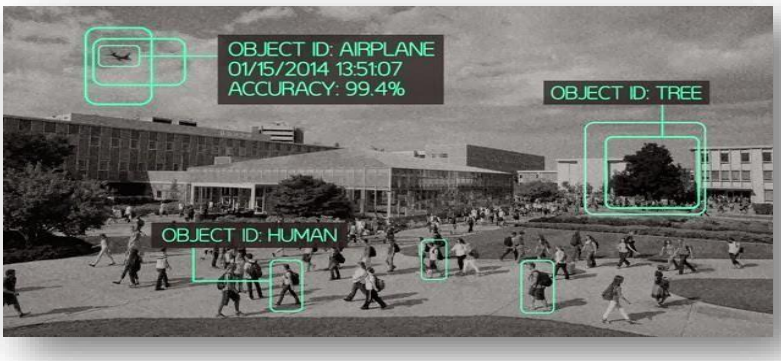

Figure 1.1 Recognizer Object and shapes.

By using of algorithms based on AI, the performance becomes better because the IR task requires high computing power [6].

3 Brief History

Image processing began to be studied around the 1920s, for the transmission of images by the submarine cable from New York to London. Indeed, the first image scan with data compression to send faxes from London to New York [1].

The 1960s saw the real start of image processing when computers began to be powerful enough to work with images [1].

The first attempts on automatic information extraction started with the beginning of the 70s , and in the late 1990s, it became possible to process a large amount of data at high speed with the evolution of general-purpose computers [4]. The mainstream method was to extract

a feature vector (called the image local features) from the image and apply a machine learning method to perform image recognition.

In the 2000 era, handcrafted features such as scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG) as image local features that are designed based on the knowledge of researchers, have been actively researched [4].

Next, in the late 2010s, deep learning to perform feature extraction process through learning has come under the spotlight. A handcrafted feature is not necessarily optimal because it extracts and expresses feature values using a designed algorithm based on the knowledge of researchers [4].

4 Image Recognition Applications

The image recognition system is still a challenging problem in the field of computer vision. It has received great attention during the past years due to its multiple applications in various fields. Although there is a strong research effort in this field, its systems are far from ideal for adequate performance in all real-world situations. In this section, we presents a brief survey of the most used applications in the field of image recognition.

4.1 Google Image Recognition

It is a completely free online image recognition tool that is very user-friendly in such a way that the search engine scans the web in two clicks in order to send you images that are identical to the reference photo (JPEG or PNG) initially downloaded.

Figure 1.1 Google Image Recognition.

4.2 IBM Image Detection

IBM's image recognition tool is one of the best of its kind. Its advanced technology allows not only to detect human faces, but also to determine the approximate age and gender before showing you similar pictures.

Figure 1.2 IBM Image Detection.

4.3 **Amazon Recognition**

Able to recognize objects, patterns and faces in the image. Thanks to its facial recognition functions, it is especially capable of searching for faces, comparing them and recognizing celebrities, however, it necessities creating an account which requires a login.

Figure 1.3 Amazon Recognition.

4.4 **Imagga**

It is a digital image recognition API with a library for sorting, organizing, and displaying images. This analysis and report tool offers an automated image tagging and category management solution capable of handling large amounts of images.

Figure 1.4 Imagga's API

4.5 **Gum Gum**

It is perfect for marketers and graphic designers looking for images that are relevant to their brand. Also, the software is considered one of the best AI photo editors [12].

Figure 1.5 GumGum API

4.6 **VISUA**

VISUA is a software capable of identifying logos, symbols and brands. It helps to monitor and control the usage of all the multimedia graphic content through monitoring of social networks, broadcast media and retail websites. It thus makes it possible to secure the brand and detect counterfeits.

Figure 1.6 VISUA API

4.7 **Clarifai**

It is typically used by federal and commercial organizations. This free advanced image recognition API powered by AI is able to tag, organize and interpret images and videos.

Figure 1.7 Clarifai API

5 Domains of Image Recognition

The concept of computer vision is based on teaching computers to extract, analyze, process and understand an image or series of images at. The pixel level, to gain a high-level understanding of digital images or videos [10,11], among its fields:

5.1 **Visual search**

It is the use of images from the real world to produce more reliable and accurate online searches. Thus, online sites may suggest items that have a theme, pattern, or otherwise relate to

consumer behaviors and interests [6].

Visual search means that people use an image to search for information instead of typing in a keyword or query. Where an Internet user, for example, uses an image of a red short-sleeved shirt instead of typing the words “red short-sleeved shirt” in an application or search engine. For its part, the search engine will suggest a set of images related to the image used by the user [8].

5.2 **Image organization**

As image and video content proliferates as technology advances, machine learning image recognition makes organizing them into categories easier and more efficient. The goal is to improve accessibility, improve search and discovery, and seamlessly share content. Many current image auto-organizing applications also use facial recognition, which is a specific task in the image recognition field [6].

5.3 **Content moderation**

It is a mandatory step for all brands or companies that want to implement an effective communication strategy. To control their brand image on the web, they no longer have a choice: they have to use this strategy to secure their social space [9].

This type of automated content moderation can be an essential tool to ensure that community spaces are targeted, safe and fulfill their intended purposes. Brands and companies must implement an effective censorship policy in order to maintain their online reputation [6].

6 Face recognition

As we have mentioned previously, the last years have witnessed extensive use of computer vision, and one of the most important area of its use is the technology of human face recognition that represents the domain to which our work belongs.

6.1 **Definition**

Face recognition is a method of identifying or confirming an individual's identity using their face. Facial recognition systems can be used to identify people in photos or videos. It is a class of biometric mostly used as technology for security and law enforcement. It bases on identifying and measuring facial features. Face recognition systems have seen wider uses in recent times for smartphones and other forms of technology such as robotics. Since it involves the measurement of the physiological characteristics of a human being, its systems are categorized as biometrics. Although the accuracy of facial recognition systems as a biometric technology is lower than iris recognition and fingerprint recognition, it is still widely used [13]

Hence, recognition systems are used all over the world today by governments and private companies as well as institutions and their effectiveness varies. On the other side, some systems were previously canceled due to ineffectiveness with allegations that they were violating the privacy of citizens. As example, Meta announced that it plans to shut down the facial recognition system on Facebook with Facebook deletes the facial scanning data of more than one billion users, this change will represent one of the largest shifts in the use of facial recognition in the history of technology [14].

6.2 Facial recognition use cases

Herein we put light on the areas in which it is used the facial recognition technology.

6.2.1 Find Missing Persons

Facial recognition can be used to find missing children, the mentally ill, and victims of human trafficking. When missing individuals and individuals are added to a database, law enforcement can be alerted as soon as they are identified, whether it's in a store, airport, or other public place [16].

The same is also used for lost pets which attempts to help owners reunite them with their lost pets such as the use of the Finding Rover application which uses face recognition (animal's face in this case) to match photos uploaded by pet owners to a database of pet photos in navigators and then immediately alert owners if their pet is found.

Figure 1.8 Find Missing Persons.

6.2.2 Facial Recognition for Security and Surveillance

Integrating facial recognition into security systems can greatly enhance security and surveillance effectiveness across sectors, and reduce overall costs. It can determine when individuals are in the camera's field of view, recognize anyone already in the system's database, and automatically send alerts for targeted human interventions. As examples: Preventing unknown or prohibited persons from entering buildings and smart residential security systems. So we consider facial recognition technology to facilitate or in other words seek to achieve security and surveillance today [15].

Also, protect schools from Threats in such a way that monitoring systems can instantly recognize faces when expelled students, dangerous parents, bandits, drug dealers, or other people who may pose a threat or danger enter the school. Therefore, the face recognition feature can reduce the risk of violent acts [16].

Figure 1.9 Facial Recognition for Security and Surveillance.

6.2.3 Facial Recognition for Health and Safety and Diagnose Diseases

Since the COVID-19 pandemic struck in early 2020, the health and safety of individuals in public and private spaces has become a priority. This has led to a series of global measures that include the mandatory wearing of masks outside the home and temperature checks when entering public places such as train stations, offices, shopping malls, public institutions and even universities. This makes a great use case for facial recognition to help keep individuals healthy by ensuring that masks are worn properly, especially in public places [15].

It can also be used to diagnose diseases that cause detectable changes in appearance. For example, the National Human Genome Institute Research Institute uses facial recognition technology to detect a rare disease called DiGeorge syndrome. This helped diagnose 96% of cases. As algorithms become increasingly sophisticated, they will become an invaluable diagnostic tool for all types of people [16].

Figure 1.10 . FR for Security and Surveillance.

6.2.4 Biometric border checks

The Europe Large Information Technology Systems Agency (EU-LISA) has developed an entry/exit system to record biometric data for all non-EU citizens crossing the external borders of the European Union. This will include the face data. The new system should

eliminate the need for manual stamping of passports. This system has supposed to efficiently help the guards to discover travelers who are trying to use multiple identities. The system is scheduled to be operational in 2022 [17].

6.2.5 Facial Recognition for Access Control

Access control is selectively restricting access to specific places or resources using facial recognition technology. Many areas can be cited as examples: Access control systems for commercial and residential facilities where facial recognition is used to grant access to authorized individuals, family members or pre-registered guests. Airport access control systems to manage passenger boarding that represents one of the many obstacles faced by air travel. In recent years, airport self-service kiosks and access control doors have used facial recognition technology for passenger interests in passenger and flight attendant operations as well as immigration control. As well as systems for controlling access to limited resources and equipment...etc. [15]. Furthermore, a variety of phones use facial recognition to unlock phones. This technology is an effective way to protect personal data and ensure the presence of the phone [16].

Figure 1.11 FR for Access Control.

6.3 Facial recognition techniques

In this section, we introduce the different steps of the 6-step facial recognition technology. It is also resumed in the Figure1.12 as follow:

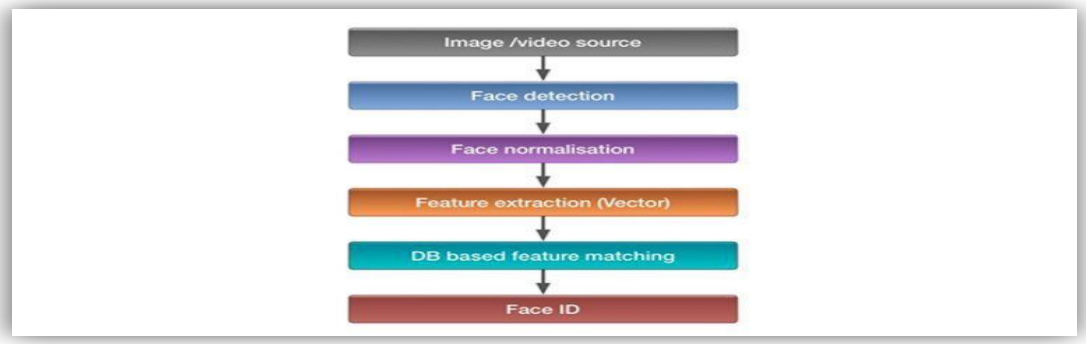

Figure 1.12 Crucial elements of the typical face recognition system.

The first step, is considered the basic step, i.e. specific to the inputs. As for the second step, face detection is the step we are most interested in in this chapter.

For the third step, face normalization is to complete the face localization effort, that is, the use of algorithms after the face alignment, so that feature-independent problems (rotation, brightness, background and occlusion) can be reduced.

The fourth step is to properly extract the useful features from the image data to maintain accuracy.

After extracting the features and specifications from step 4, step 5 in turn matches them to the database. Finally, step 6 is for outputs.

6.3.1 Face detection

It represents the first operation in the face recognition system. Its reliability has a major influence on the performance and usability of the entire face recognition system [7]. The goal of this step is to identify and locate all the faces in the image.

6.3.1.1 **Face detection challenges**

The challenges associated with face detection can be attributed to the following factors [18] :

- **Pose:** As is presented in Figure1.13, the images of a face vary due to the relative camera-face pose (frontal, 45 degree, profile, upside down), and some facial features such as an eye or the nose may become partially or wholly occluded.

Figure 1.13 Pose Variations

- **Presence or absence of structural components:** Facial features such as beards, mustaches, and glasses may or may not be present and there is a great deal of variability among these components including shape, color, and size.
- **Facial expression:** The appearance of faces are directly affected by a person's facial expression. Figure 1.14 shows an example of this influence.

Figure 1.14 Expression variation

- **Occlusion:** Faces may be partially occluded by other objects. As an example, in an image with a group of people, some faces may partially occlude other faces

Figure 1.15 Examples of occlusions.

- **Image orientation:** Face images directly vary for different rotations about the camera's optical axis. Imaging conditions: When the image is formed, factors such as lighting (spectra, source Distribution and intensity) and camera characteristics (sensor response, lenses) affect the appearance of the face, such as shown in the Figure 1.16.

Figure 1.16 Examples of Image orientation.

6.3.1.2 **Face detection methods**

A comprehensive study on face detection methods in [19] grouped the various methods into three categories: knowledge-based, Template Matching, and appearance-based methods.

a) knowledge-based methods

These methods depend on knowing the main characteristics of human faces and using them to know the features of the face and its relationship.

Knowledge based methods can divided into two different approaches which are top-down and bottom-up methods.

- **Top-Down Methods**

The methods based on this approach use rules to describe features of a face derived from knowledge of human face. For example, a face image is consist of two eyes that are symmetric to each other, a nose and a mouth. The relationships between features can be represented by their relative distances and positions. The problem of this approach is as follow, it is difficult to translate human knowledge into well-defined rules. If this method chooses to use strict rules, than it may fail while detecting faces because it may not pass all the rules defined. On the other hand if the rules are too general than there may be many false detections.

The authors of [22] perform one popular work of this approach. They have used a hierarchical knowledge-based method to detect faces. Their system consists of three levels of rules. At the highest level, all possible face candidates are found by scanning a window over the input image and applying a set of rules at each location. The rules at higher level are general descriptions of what a face looks like while the rules at lower levels rely on details of facial features. A multi-resolution hierarchy of images is created by averaging and subsampling (Figure 1.17).

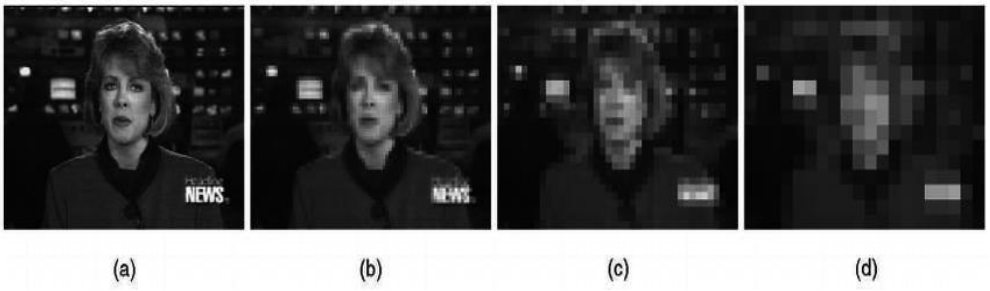

Figure 1.17 . Original and corresponding low resolution images: (a) n = 1, original image. (b) n = 4. (c) n = 8. (d) n = 16. Each square cell consists of n × n pixels in which the intensity of each pixel is replaced by the average intensity of the pixels in that cell.

Another example is presented in Figure 1.18. It bases on coded rules which are based on the characteristics of human face that are used to locate the face candidates in the lowest resolution presented by the center part of the face (Region-1) and the upper round part of the face (Region-2). The first region has a four cells with basically uniform intensity, and the second one has a basically uniform intensity. The difference between the average gray values of the center part and the upper round part is significant and plays the important role in face detection [20].

Figure 1.18 Typical face used in knowledge-based top-down methods

- Bottom-Up Methods

In this approach, the purpose of different algorithms is to find variant features of faces for the detection process. The underlying assumption is based on the observation that humans can effortlessly detect faces and objects in different poses and lighting conditions and so, there must be existing properties or features that are invariant over this variability. Numerous methods have been proposed that first detect facial features and then infer the presence of a

face. Facial features such as eyebrows, eyes, nose, mouth, and hair-line are commonly extracted using edge detectors. Based on the extracted features, a statistical model is built to describe their relationships and to verify the existence of the face. One problem with these feature-based algorithms is that the image features can be severely corrupted due to illumination, noise, and occlusion. Feature boundaries can be weakened for faces, while shadows can cause numerous strong edges which together render perceptual grouping algorithms useless [19].

b) Template matching methods

In template matching category, a standard face pattern (usually frontal) is manually predefined or parameterized by a function. Given an input image, the correlation values with the standard patterns are computed for the face contour, eyes, nose, and mouth independently. The existence of a face is determined based on the correlation values. This approach has the advantage of being simple to implement. However, it has proven to be inadequate for face detection since it cannot effectively deal with variation in scale, pose and shape.
Multiresolution, multiscale, sub-templates and deformable templates have subsequently been proposed to achieve scale and shape invariance [19].

Template matching algorithms can be separated into two sub-categories which are, algorithms using predefined templates and the others are algorithms using deformable templates.

- **Predefined templates**

An early attempt to detect frontal faces in photographs is reported by the authors of [23]. They have used several sub-templates for the eyes, nose, mouth, and face contour to model a face. Each sub-template is defined in terms of line segments. Lines in the input image are extracted based on greatest gradient change and then matched against the sub-templates. The correlations between sub-images and contour templates are computed first to detect candidate locations of faces, then, matching with the other sub-templates is performed at the candidate positions. As presented in Figure 1.19, the authors of [24 have used a small set of spatial image invariants to describe the space of face patterns. The key insight for designing the invariant is that, while variations in illumination change the individual brightness of different parts of faces (such as eyes, cheeks, and forehead), the relative brightness of these parts remain largely unchanged. The template is composed of 16 regions (the gray boxes) and 23 relations (shown by arrows). Determining pairwise ratios of the brightness of a few such regions and retaining just the “directions” of these ratios provides a robust invariant.

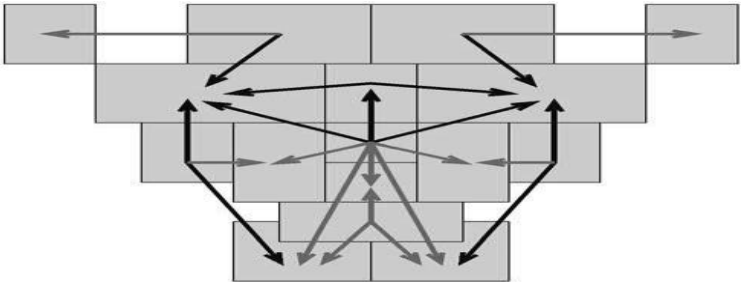

Figure 1.19 A 14x16-pixel ratio template for face localization.

- **Deformable Templates**

In this approach, facial features are described by parameterized templates. An energy function is defined to link edges, peaks, and valleys in the input image with their corresponding parameters in the template. The best fit of the elastic model is found by minimizing an energy function of the parameters. Although their experimental results demonstrate a good performance in tracking non-rigid features, one draw-back of this approach is that the deformable template must be initialized in the proximity of the object of interest [25].

In [26], the authors have described a face representation method with both shape and intensity information .They start with sets of training images in which sampled contours such as the eye boundary, nose, chin/cheek are manually labeled, and a vector of sample points is used to represent shape. They used a Point Distribution Model (PDM) to characterize the shape vectors over an ensemble of individuals and an approach to represent shape-normalized intensity appearance. A face-shape PDM can be used to locate faces in new images by using Active Shape Model (ASM) search to estimate the face location and shape parameters. The face patch is then deformed to the average shape and intensity parameters are extracted. The shape and intensity parameters can be used together for classification.

c) Appearance Based Methods

In this approach the templates are learned from examples in images. In general, appearance-based methods rely on techniques from statistical analysis and machine learning to find the relevant characteristics of face and non-face images. The learned characteristics are in the form of distribution models or discriminant functions that are consequently used for face detection. Meanwhile, to overcome performance issues, dimensionality reduction is usually carried out for the sake of computation efficiency and detection efficacy.

Many algorithms use the appearance-based methods. Mostly used are: Eigen faces, distribution-based methods, neural networks, support vector machines, sparse network of winnows, Naïve Bayes classifiers, hidden Markov model and information the orifical and

inductive learning approaches [19]. Figures 1.20 and 1.22 shows two example of twp different algorithms. The first system consists of two components, distribution-based models for face/non-face patterns and a multilayer perceptron classifier. Each face and non-face example is first normalized and processed to a 19 X 19-pixel image and treated as a 361-dimensional vector or pattern. Next, the patterns are grouped into six faces and six non-faces clusters using a modified k-means algorithm. Figure 1.21 shows the distance measures.

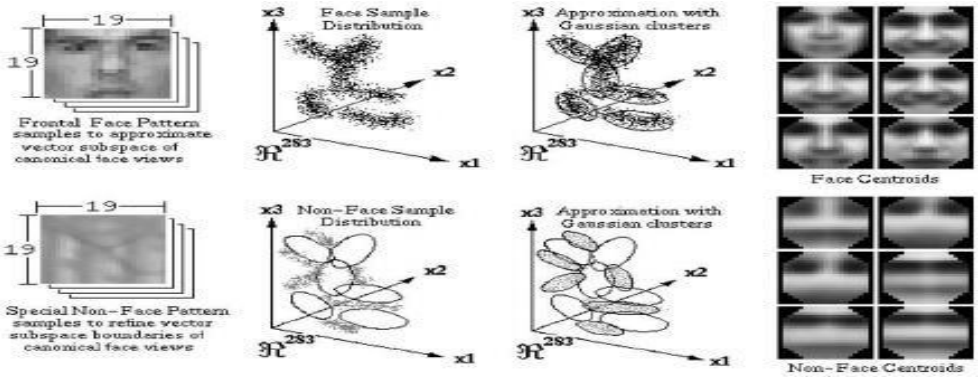

Figure 1.20 Face and non-face clusters method

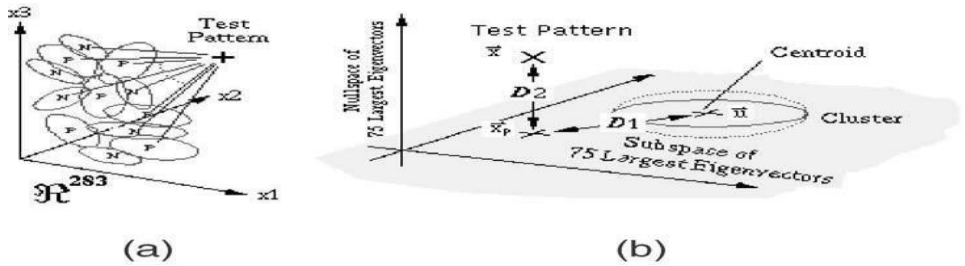

Figure 1.21 The distance measures used by Sung and Poggio.

In the second system (Figure1.22), a Hidden Markov Model (HMM) for a pattern recognition problem is used in such a way that a number of hidden states need to be decided first to form a model. Then, a training process of the HMM is needed to learn the transitional probability between states from the examples where each example is represented as a sequence of observations. After the HMM has been trained, the output probability of an observation determines the class to which it belongs. HMM-based methods usually treat a face pattern as a sequence of observation vectors where each vector is a strip of pixels, as shown in Figure 1.22 (a). During training and testing, an image is scanned in some order (usually from top to bottom) and an observation is taken as a block of pixels, as shown in Figure 1.22 (a). For face patterns, the boundaries between strips of pixels are represented by probabilistic transitions between states, as shown in Figure 1.22 (b), and the image data within a region is modeled by a multivariate Gaussian distribution.

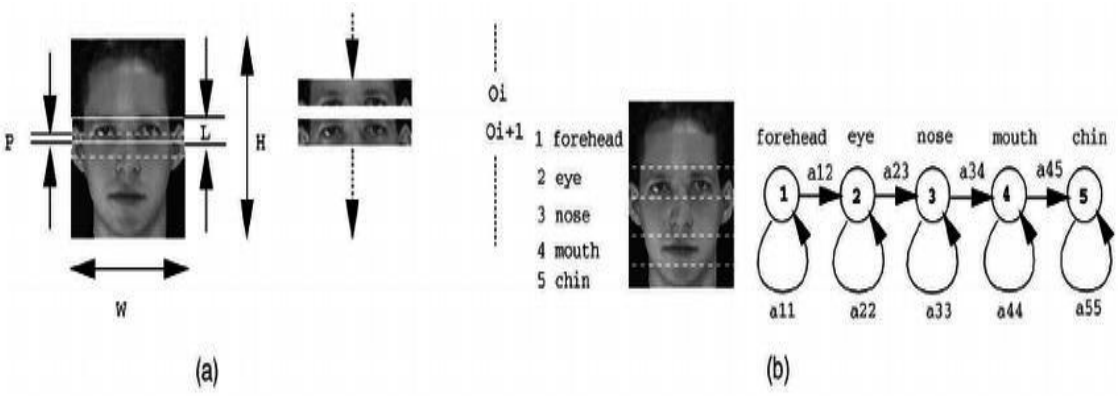

Figure 1.22 Hidden Markov model for face localization.

6.3.1.3 **Feature invariant approaches**

These algorithms aim to find structural features that exist even when the pose, viewpoint or lighting conditions vary.

- Facial Features

It basses on the segmentation of the face from a cluttered background for face identification. It uses an edge map (Canny detector) and heuristics to remove and group edges so that only the ones on the face contour are preserved. An ellipse is then fit to the boundary between the head region and the background. This algorithm achieves 80 percent accuracy on a database of 48 images with cluttered backgrounds. Recently, Amit et al. presented a method for shape detection and applied it to detect frontal-view faces in still intensity images. Detection follows two stages: focusing and intensive classification [28].

- Texture

Human faces have a distinct texture that can be used to separate them from different objects. The idea of this method is to infers the presence of a face through the identification of face-like textures. The textures are computed using second-order statistical features (SGLD) on sub-images of 16 X 16 pixels. Three types of features are considered: skin, hair, and others. A combination between a cascade correlation neural network for supervised classification of textures and a Kohonen self-organizing feature map to form clusters for different texture classes is needed [21].

7 Conclusion

In this chapter, we have focused on image recognition generally and specifically face recognition techniques and fields in order to give to the reader a global view of the original domain focusing on face detection techniques since it represent the first step in the face emotional recognition as a sub field of face recognition field. In the next chapter, we have focused on the sub-domain of facial emotions recognition as it represents the sub-domain that we are interested in.

CHAPTER 2

Facial emotions recognition

1 Introduction

Emotions are expressed when interacting and communicating with others. The main body element that present these emotions is the face. Although, emotions recognition is used by many companies due to its great importance nowadays, studying how to recognize and read them remains a challenging task for smart technology which is used to do the job. In this chapter, we have introduced some concepts about emotions and facial expressions, as well as Fields of application, then defined CNNs and its architectures, and finally the architecture of facial recognition systems.

2 Facial expressions and emotions

2.1 Emotion

The term Emotion is derived from Latin word "Emover" which means to move or to agitate [29]. It represents the reaction to person, object or event and can be expressed in a variety of ways such as anger, fear, joy, happiness, sadness, surprise etc.

Don Hockenbury and Sandra E. Hockenbury suggest in their book entitled "Discovering Psychology" [30], that an emotion is a complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response.

In addition to trying to define what emotions are, researchers have also tried to identify and classify the different types of emotions. The descriptions and insights have changed over time [31]:

In 1972, psychologist Paul Eckman suggested that there are six basic emotions that are universal throughout human cultures: fear, disgust, anger, surprise, happiness, and sadness [32].

In the 1980s, Robert Plutchik introduced another emotion classification system known as the "wheel of emotions." This model demonstrated how different emotions can be combined or mixed together, much the way an artist mixes primary colors to create other colors [33].

In 1999, Eckman expanded his list to include a number of other basic emotions, including embarrassment, excitement, contempt, shame, pride, satisfaction, and amusement [32].

Figure 2.1 presents the emotions that are so-called “primary” emotions, which are respectively: joy, fear, anger, sadness, disgust and surprise [34].

Figure 2.1 Facial expressions of primary emotions.

2.2 Facial emotions recognition

Facial Emotions Recognition (FER) is a technology used for analyzing sentiments by different sources, such as pictures and videos. It belongs to the family of technologies often referred as 'affective computing', Hence, FER bases on computer's capabilities to recognize and interpret human emotions and affective states and it often builds on Artificial Intelligence technologies [35].

Facial expressions are forms of non-verbal communication, providing hints for human emotions. For decades, decoding such emotion expressions has been a research interest in the field of psychology but also to the Human Computer Interaction field. Recently, the high diffusion of cameras and the technological advances in biometrics analysis, machine learning and pattern recognition have played a prominent role in the development of the FER technology [36].

3 Fields of application

Potential uses of FER cover a wide range of applications. In order to distinct between them, an attempt of classification is made as follow [36]:

3.1 Provision of personalized services

Analyzing facial emotions process is performed in order to show personal messages in an environment capable of identifying people. As examples: Based on the analysis of instantaneous facial expressions, music that matches the specific feeling is presented. Also, analyzing facial expressions to determine people's reaction to movies.

3.2 Customer behavior analysis and advertising

In this category, facial recognition system is used for marketing purposes, such as analyzing the feelings of customers while shopping in stores, focusing on goods or arranging them inside the store.

3.3 Healthcare

The main purpose of this category of applications is the monitoring of human's conditions

such as: Monitoring patients' conditions during treatment, screening for autism or neurodegenerative diseases and anticipate psychotic disorders or depression to identify users who need help.

3.4 Employment

The main cases of using FER in this category are: Helping recruits make decisions, identifying uninterested candidates in a job interview and monitor the mode and interest of employees.

3.5 Other

Some other areas can use the FER for at least on purpose such as: Education, public safety and crime detection

4 Facial emotions recognition methods

Facial expression recognition has been an active area of research over the past few decades, and it is still challenging due to the high intra-class variation and the big gap between the visual features that can be captured through a camera (in the format of pixels) and the required processing features used in emotion recognition. To overcome this problem, in the past the algorithms used visual hand-crafted features, such as Dense SIFT, Histogram of Oriented Gradients (HOG) or LBP. Followed by a classifier trained on a database of images or videos. Most of these works perform reasonably well on datasets of images captured in a controlled condition but fail to perform as well on more challenging datasets with more image variation and partial faces. More recently, Deep learning enables to automatically infer a hierarchy representation of the visual information and many features can be extracted and learned for a decent facial expression recognition system. One of the main methods for doing so is what is known as Convolutional Neural Networks (CNNs) [37].

4.1 Convolutional Neural Networks

CNNs are a type of deep learning algorithm that is used to process data that has a spatial or temporal relationship. CNNs are similar to other neural networks, but they have an added layer of complexity due to the fact that they use a series of convolutional layers. Convolutional layers are an essential component of Convolutional Neural Networks (CNNs). The name “convolutional neural network” indicates that the network employs a mathematical operation called convolution. Convolution is a special linear operation. Convolutional networks are simply neural networks that use convolution at the place matrix multiplication in at least one of their layers. The picture below represents a typical CNN architecture.

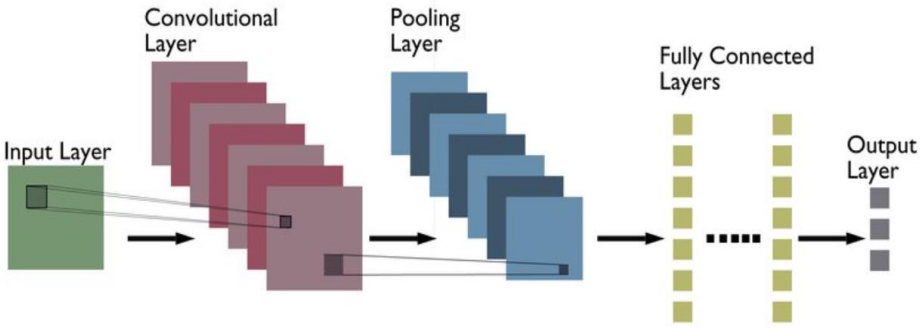

Figure 2.2 Standard architecture of a convolutional neural network.

The first part of a CNN is the convolutional part. It functions as an image feature extractor. The image passes through a succession of filters, or convolution kernels, to transform it into new images called convolution maps "feature maps". Some intermediate filters reduce the resolution of the image by a local maximum operation. In the end, the cards of convolutions are flattened and concatenated into a feature vector, called a CNN code. The output result of the convolutional part is connected to the input of a second part, consisting of fully connected layers which consists of combining the characteristics of all the network to classify the image in the output which is a layer comprising one neuron per class. Finally a numerical values generally normalized between 0 and 1 is obtained to present the probability distribution on the classes [38].

4.1.1 The convolution layer (CONV)

Three hyper parameters are used to size the volume of the convolution layer: the depth, stride and margin.

1. Layer depth: number of convolution nuclei (or number of neurons associated with the same receptive field).
2. Stride: Controls the overlap ofreceptive fields. The smaller the pitch, the more the receptive fields overlap and the greater the output volume will be.
3. Margin (at 0 or 'zero padding'): sometimes it is convenient to put zeros at the boundary of the input volume. The size of this 'zero-padding' is the third hyper parameter. This margin controls the spatial dimension of the output volume. In particular, he sometimes desirable to keep the same area as the input volume [39].

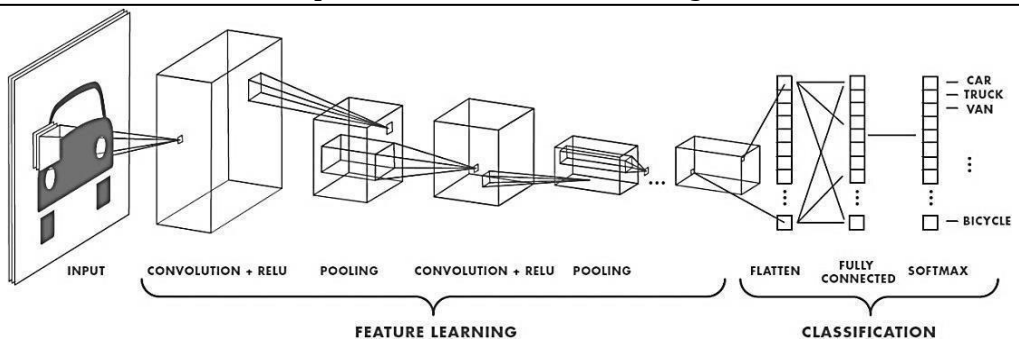

Figure 2.3 Example of a network composed of many convolutional layers. Filters are applied to each image used for training at different resolutions, and the output of each convolved image is used as input to the next layer.

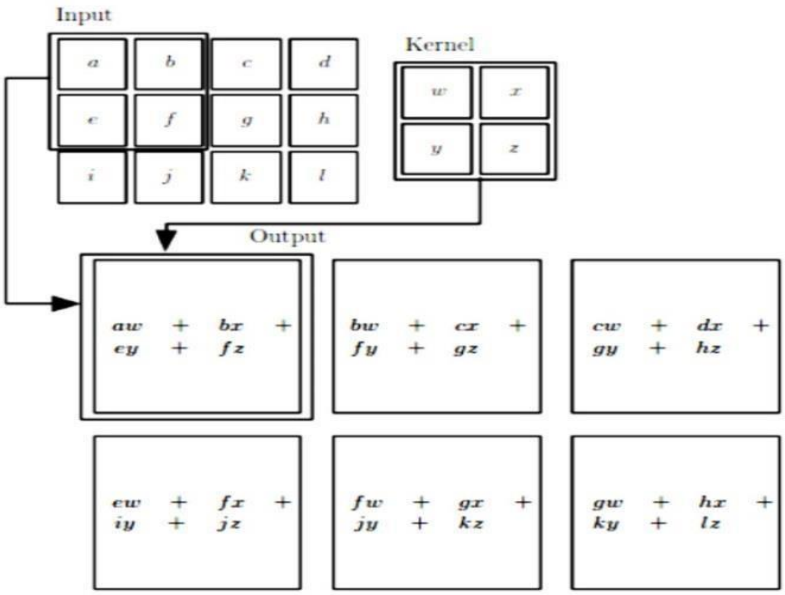

Figure 2.4 Example of a 2D convolution.

4.1.2 Pooling layer

Pooling is a form of image sub-sampling, it allows to gradually reduce the size of the representations in order to reduce the amount of parameters and computation in the network as well as the invariance to small translations, it is therefore common to periodically insert a pooling layer between two successive convolutional layers of a CNN architecture to control over-learning. The pooling operation also created a form of translational invariance. The pooling layer works independently on each depth slice of the input and scales it only at the surface level.

The most common form is a pooling layer with filters of size 2x2 (width/height) and as output value the maximum input value. In this case, we speak of “Max-Pool 2x2” as is presented in Figure2.5. It is possible to use other pooling functions than the maximum. It can

be used an “average pooling”, the output is the average of the values of the input patch. Pooling allows big gains in computing power. However, due to the aggressive reduction in the size of the representation and therefore the associated loss of information, the current trend is to use small filters (2x2 type). It is also possible to avoid the pooling layer but this involves a greater risk of over- learning [41].

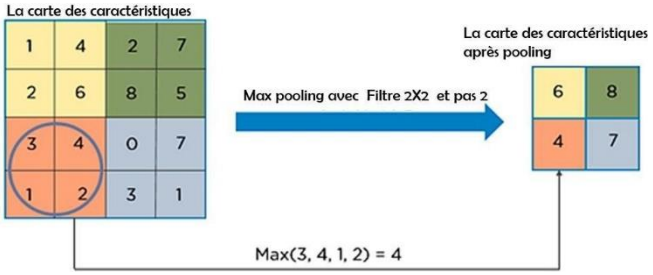

Figure 2.5 Pooling with a 2x2 filter.

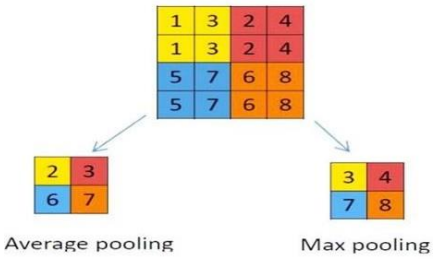

Figure 2.6 Average and Max pooling

4.1.3 Fully Connected Layers

After extracting the characteristics of the inputs, a perceptron or an MLP (Multi-Layer Perceptron) is attached to the end of the network. The perceptron takes as input the extracted features and produces a vector of N dimensions where N is the class number or each element is the probability of belonging to a class. Each probability is calculated using the softmax function ref in the case where the classes are exclusively mutual [42].

4.1.4 Activation functions

The activation function is a mathematical function applied to a signal at the output of an artificial neuron. The term "activation function" comes from the biological equivalent "activation potential", the stimulation threshold, which once reached, triggers a neuron response. The activation function is often a non-linear function. Their purpose is to allow neural networks to learn more complex functions than simple linear regression because multiplying the weights of a hidden layer is just a linear transformation. Example of activation function is the ReLu (Rectified Linear Units) which is a function that eliminates all negative values [41].

4.2 Facial emotions recognition and CNN:

The typical procedure for the emotion recognition method with CNN is depicted in Figure 2.7. The first stage is the face detection and tracking. It involves the process of locating face regions from the input data, and tracking the face region in every frames of video. The second stage is facial feature extraction and representation, which is responsible for extracting and representing the facial variations caused by facial expressions. Finally, the facial expression classification [43].

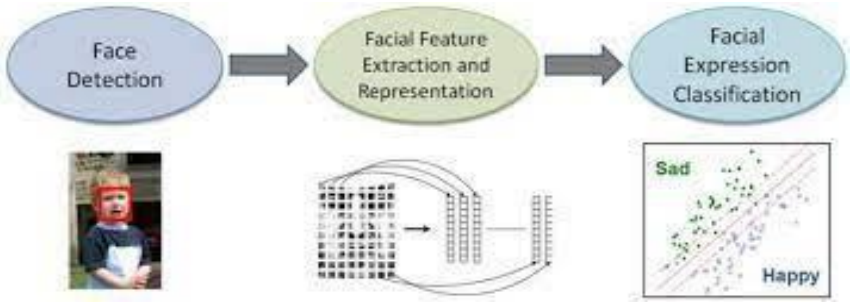

Figure 2.7 Steps of FER.

4.2.1 Face detection

It represents the first operation in the face recognition system. Its reliability has a major influence on the performance and usability of the entire face recognition system [44]. The goal of this step is to identify and locate all the faces in the image.

The techniques used at this step are explained in the previous chapter (Chapter1) generally known as a part of the field of pattern recognition.

4.2.2 Feature extraction

After face detection, the next step in FER is feature extraction. The main aim of facial feature extraction is to extract an effective and efficient representation of facial components without any loss of face information. Geometric-based and appearance-based features are the two feature extraction categories of techniques used for facial motion and deformation of facial features. The input image may be either a static image or image sequence. Based on the input image, a suitable facial feature extraction algorithm is applied to extract either the local, global or hybrid features. The extracted features are considerably reduced in size, which is given as input to the classifier and that significantly helps the classifier to make the decision easier in identifying and recognising the facial expression [45].

We present a generalized view of facial feature extraction methods and an extensive review on recent feature extraction techniques in FER.

4.2.2.1 Gometric features [45]

It represents the shape and location of the facial components or points of pre-defined facial features. They include the mouth, eyes, eyebrows and nose, which are drawn to form distinct vectors to represent the geometry of the face as shown in figure 2.8.

Figure 2.8 Detection of static geometric facial features.

However, expressions affect the relative shapes and positions of different facial features. Thus, basic facial expressions can be recognized by measuring the displacement of important facial components.

In the case of image sequences as input, the Facial Action Coding System (FACS) is the most famous of several different techniques for recognizing facial signals and is generally used as part of reference mental exploration. This system helps distinguish facial movements based on the analysis of facial movements. The FACS contains different working units (AUs) related to specific muscle contractions.

On the other hand, static image input uses a model-based approach, such as the Active Shape Model (ASM) , the Active Appearance Model (AAM) and the Static Attribute Transformation algorithm for scale (SIFT) to extract facial features.

The geometry-based method is more suitable for real-time facial images as features that can be easily identified and traced, but it requires precise face detection technology.

4.2.2.2 Appearance Features:

The Appearance-based approaches focus on the transient features of the face such as wrinkles, bulges and indentations which describe changes in facial texture, density, histograms and pixel values. Among the algorithms that are applied to extract feature descriptors: PCA, Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Gabor wavelet, LBP.

In recent years, Gabor wavelet and LBP are extensively used to extract feature descriptors. Gabor wavelets are a well-known representative feature for extracting texture information effectively. Zhang et al. have [46]. Investigated and compared geometry-based and Gabor-based method and the result shows that Gabor wavelet outperforms in performance and considered as a more powerful tool for feature extraction [45].

The Gabor wavelet is one of the most popular feature description methods, and it is also one of the mainstream methods of facial descriptions. Gabor wavelets can well simulate mammalian visual neurons and capture salient visual features. The Gabor wavelet can extract spatial and frequency-domain information from multiple scales and multiple directions, which can enlarge the difference between classes. The function of the two-dimensional Gabor wavelet filter can be expressed in the following form [47]:

$$\psi_{\mu,v}(z) = \frac{\|k_{u,v}\|^2}{\sigma^2} e^{\left(-\|k_{u,v}\|^2 \|z\|^2/2\sigma^2\right)} \left(e^{jk_{u,v}z} - e^{-\sigma^2/2}\right) \qquad (1)$$

Where z = (x, y). $\|.\|$ Represents vector norm operation.

$$k_{u,v} = k_v e^{\varphi_{\varphi\mu}} \text{ and } \varphi_u = \pi\mu/8 \cdot k_v = k_{max}/f^v \qquad (2)$$

Where K_{max} is the maximum sampling frequency and $K_{max} = \pi/2$ is the sampling step in the frequency domain that usually has a value of $\sqrt{2}$. Parameter σ determines the size and the wavelength ratio of the Gauss filter window. Let ν and μ represent the scale and direction of the Gabor filter respectively. It usually has 5 scales $\nu \in \{0, \ldots, 4\}$ and 8 directions $\mu \in \{0,\ldots, 7\}$, as shown in Figure 2.9

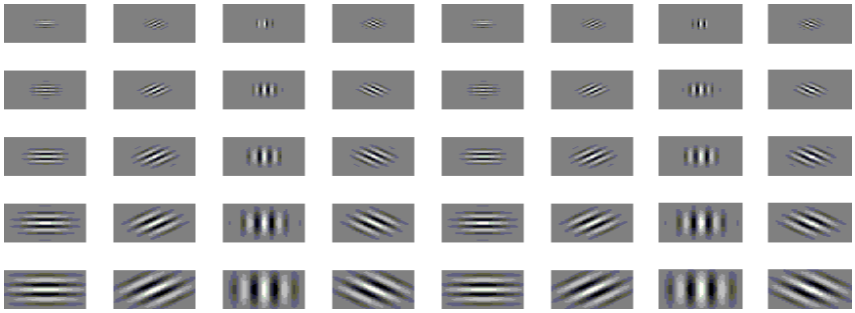

Figure 2.9 Five scales from top to bottom and 8 directions from left to right of the Gabor filter.

Let *I(x, y)* represents the pixel distribution of face images. Then, the convolution between *I(x, y)* and Gabor filter $\Psi_{u,v}(z)$ is the Gabor feature representation of the human face,

$$O_{\mu,v}(z) = I(x,y) * \psi_{\mu,v}(z) \qquad (3)$$

Where *represents convolution operation, $O_{u,v}(z)$ is the convolution *I(x, y)* and the Gabor kernel function is $\Psi_{u,v}(z)$ a multiresolution and multidirectional Gabor filter decomposition representation is obtained through the convolution of 40 Gabor kernels consisting of 5 scales and 8 directions. The result of the convolution consists of two parts, the real part and the imaginary part. However, usually use the corresponding amplitude of the convolution as the facial representation, which is shown in Figure 2.10.

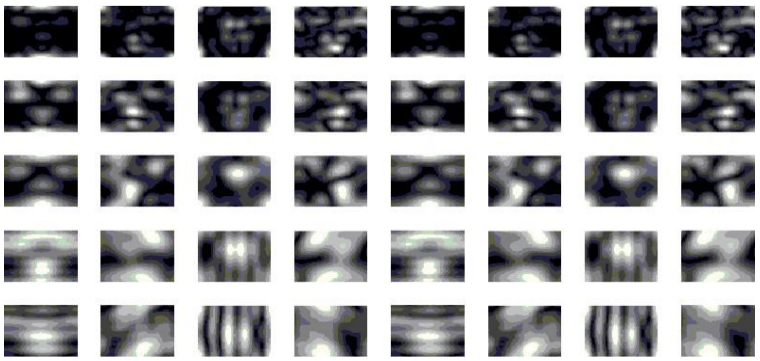

Figure 2.10 Multiscale and multidirectional Gabor amplitude representation of face images.

The LBP of the local texture feature extraction method based on the local binary model is an effective non-parametric method for local image texture descriptions. It uses the structure method to analyze the features of fixed windows and then uses the statistical method to extract the features. It is simple in calculation and can capture small detailed features in the image. It can also extract the local neighborhood relationship model that is more favorable for classification. The LBP operator was originally designed for image texture feature extraction. Each pixel of the image is calculated by binary values according to the gray value of the central pixel of its 3 × 3 neighborhood pixel [47].

$$S(f_p - f_c) = \begin{cases} 1, & f_p \geq f_c \\ 0, & f_p < f_c \end{cases} \qquad (4)$$

Where f_c is the gray value of the central pixel and f_p is the sampling point of the neighborhood pixels of the center pixel. Then, each sampling point in the neighborhood is assigned 8 neighborhood sampling points according to different weight coefficients 2^p . Finally, the values are added together to obtain the LBP value of the center point f_c. Figure 2.11 shows the feature extraction process of the LBP operator.

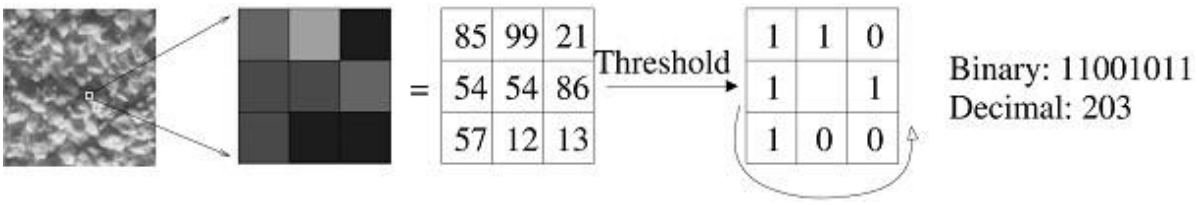

Figure 2.11 The LBP operator extraction process

A drawback is that the original LBP operator cannot capture larger scale texture structures. To solve this problem, the LBP operator is used to extract the features at different scales using (P, R) as the nearest neighbor region representing the central point. That is, there are P sampling points around the circle with a center of distance of R. By changing the number of P

and the distance of R, we can change the scale of feature extraction. In object recognition applications such as face recognition, LBP usually uses consistent patterns of texture descriptions. This pattern allows for only 0 or 1 jumps in binary encoding up to two times. For example, 00000000 (0 jumps), 01110000 (two jumps), and 11001111 (two jumps) are the uniform patterns, while 11001001 (4 jumps) and 01010011 (6 jumps) are not the uniform patterns [47].

CLNF uses what is known as Constrained Local Model (CLM) framework. CLM was coined by the authors of [48], and it is a class of methods for modelling deformable objects that posses a distinct set of features such is presented in Figure 2.12. This can be applied to a setting in which there is a face (deformable object) and one wants to detect the facial landmarks (features) [37].

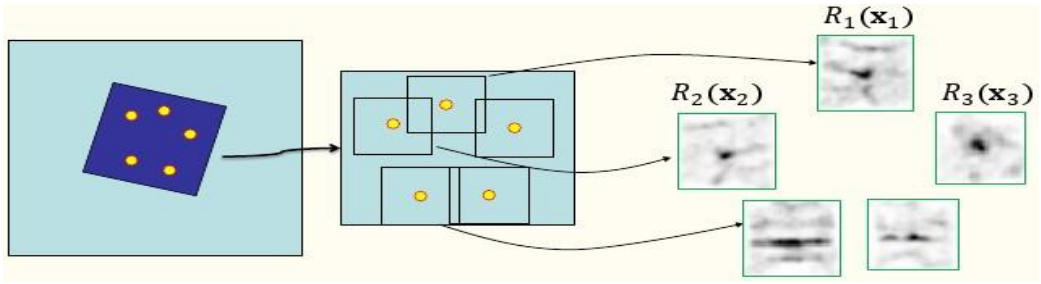

Figure 2.12 : Given a detected object on the image (left), a set of features locations are predicted (middle) and a "response image" R(x) is generated for each location (right).

It all starts by providing an estimate on where the location of the features are within the image. In the case of the face, a template of the landmarks seen from a frontal view over the area from a face detector is the first estimate. This is adjusted through multiple iterations until convergence presented in Figure 2.13. The overall workflow can be subdivided in three main components: a point distribution model, patch experts and a fitting approach [37].

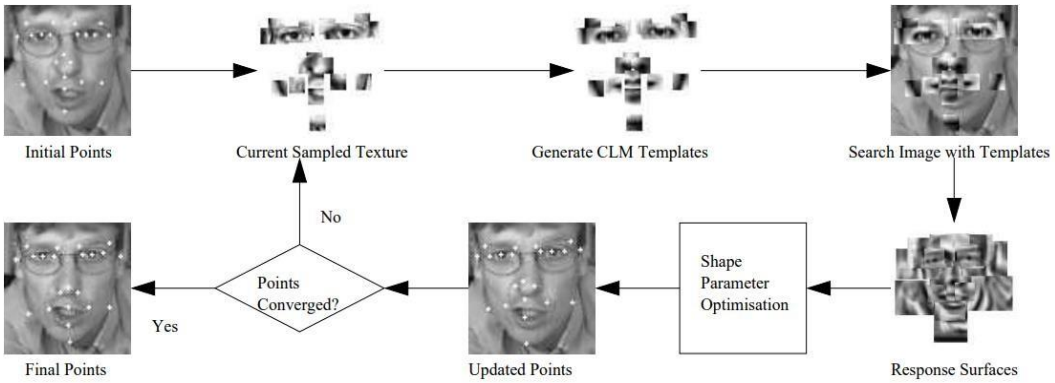

Figure 2.13 CLM Search Algorithm.

The Constrained Local Neural Field (CLNF) model is an instance of the CLM framework that includes a novel Local Neural Field (LNF) patch expert and a novel Non-uniform RLMS fitting technique. CLNF outperforms other state-of-the-art techniques when estimating landmarks in unseen lighting conditions and in the wild settings [37]. The content of this section significantly relies in the theory that can be found in [50].

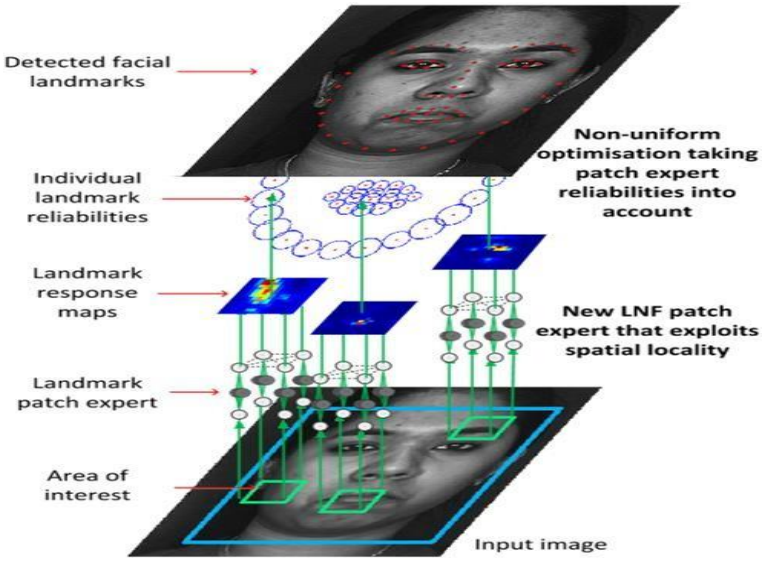

Figure 2.14 Overview of the CLNF model (showing only three patch experts).

4.2.3 Classification

The final stage of the FER method is based on machine learning theory precisely it is the classification module. The input to the classifier is a set of features which has been retrieved from face region in feature extraction stage. The feature vector is formed to describe the facial expression. The first part of Classification module is training. The training set of classifier consists of labeled data. Once the classifier is trained, it can recognize input images by assigning them a particular expression class label [43].

All approaches for classification of facial expression can be divided into two groups: frame based recognition which only relies on a single frame, image sequence based approaches exploited the temporal behaviors of facial. In different researches, various classifiers have been applied such as Neural Network, Bayesian Network (BN), Support Vector Machine (SVM), rule-based classifiers, and Hidden Markov Models (HMM) .Some studies in facial expression analysis, made their effort to classify action units (AU). Other studies classified each emotional state based on the extracted features. In Gabor filters with different frequencies and orientations were applied on face region, and SVM classification method used for recognizing four basic emotion [43].

In [52], an educational approach is proposed based on the intentional convolutional network,

which is able to focus on important parts of the face, and the use of visualization technology that is able to find important facial areas to detect various emotions, is based on the output of the classifier, that is, the parts of the image that have the strongest effect on the result of the classifier.

In [53], the authors have proposed a local learning framework that combines automatic features learned by CNNs and handcrafted features to predict the class naming of each test image. It is based on three steps:

First, an OR k-nearest neighbor model is applied to determine which training samples are closest to the input test image. Second, an individual SVM classifier is trained for everyone on the selected training samples. Finally, use the SVM classifier to predict the class feature only for the test image that has been trained.

In this work they demonstrate that they can bypass the current state of the art systems by combining automatic features learned by Convolutional Neural Networks (CNNs) with handcrafted features computed by the Bag of Visual Words (BOVW) model, They used three CNN models in this work namely VGG-face, VGG-f and VGG-13 which are based on horizontally inverted images using DSD (DSD) training to train their CNN models. Since the images in FER 2013 are of the same size, the VGG-13 was considered an excellent choice for FER 2013.

Then they only trained the latter from scratch using DSD. They used pre-trained and optimized versions, used random gradient descents using small batches of 512 images and the momentum rate set to 0.9, with input images scaling 64×64 pixels for the VGG-13. They divided the images into bins to create a spatial hierarchical representation of the BOVM model, using an application SVM classifier.

5 Conclusion

In this chapter, we have focused on facial emotions recognition, especially the use of CNNs architecture then we have explained the most popular theories and methods for facial expression recognition through the three steps: detecting faces in images, feature extraction and finally classification.

The next chapter explains our proposed solution using CNNs architecture for a facial emotions recognition.

CHAPTER 3

Proposed Framework

1 Introduction

In this chapter, we talk about the architecture of our work based on the Convolutional Neural Network, focusing specifically on all issues related to the CNN used in our work as well as the implementation and different results obtained compared to other solutions

2 Network Architecture

In our study, we address the reliance on convolutional neural networks, as they have recently been widely used in image processing and classification, specifically in recognizing the face and its seven expressions. in this section a study of the network structure of our model is presented.

Our work is inspired by the techniques presented by [52] and the approach proposed in [53], which combines automatic features learned by neural networks (CNNs) with handcrafted features. By exploiting both approaches and integrating further applied techniques to increase average accuracy and optimization, a comprehensive framework for deep learning that relies on a convolutional neural network is proposed to classify human face emotions as shown in the Figure 3.1.

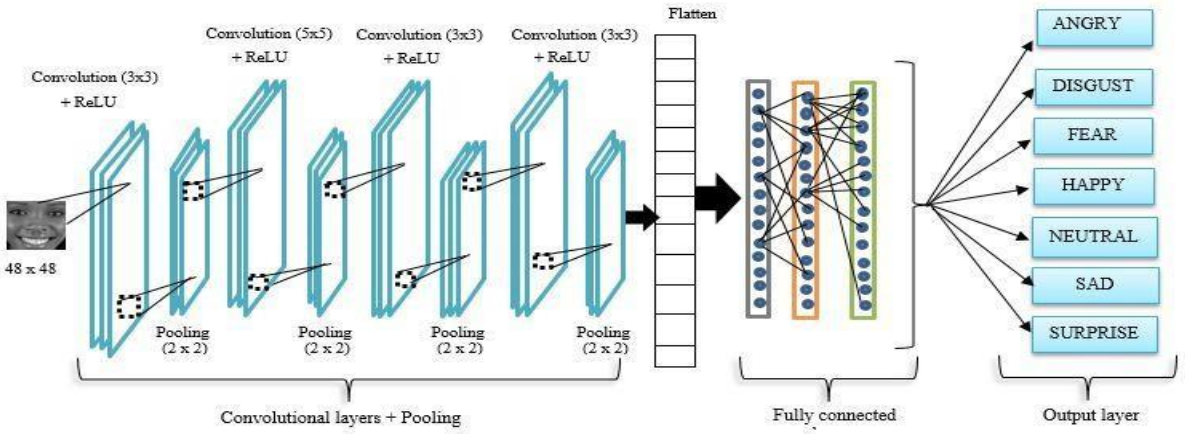

Figure 3.1 The proposed Method architecture.

1. Input layer: It represents the starting point of our system. Displays processing data as (48 x 48) gray level images.
2. Convolution layer: The layer takes as input a 48 x 48 image that passes through different filters that help to extract relevant features. For this reason, we have used this layer 4 times. The convolution layer is provided by the Conv2D function of a sequential model of the convolutional network:

   Conv2D (filters, kernel size, activation, padding, stride).

Filters: Present the number of convolution filters.

- The first convolutional layer contains 64 filters.
- The second has 128 filters.
- The third layer has 512 filters
- The last has 512 filters

Kernel size: an integer or tuple / List of 2 integers, specifying the size (height and width) of the 2D convolution filter.

- The first convolutional layer contains 3x3 kernel size.
- The second has 5x5-kernel size.
- The third layer has 3x3-kernel size.
- The last has 3x3-kernel size.

Activation: The activation function to use various choices (reread, elected, etc.) are available in our model. Hence, we have used activation function (Relu) in each layer.

Spatial aggregation layer: It represents the step that reduces the spatial size of the local characteristic maps. We have used pooling size (2 x 2) and type max after each convolution layer which is defined by:

MaxPooling (pool size).

Pool size = the max pooling window size.

Fully Connected (FC) layer: It represents the last part that contains the high-level reasoning, where the data is in the form of a vector of dimension (n x 1). In this part, we have used three fully connected layers with a Relu activation function. where this layer defines by:

Dense (Units, activation=ReLu).

Units : positive integer which represents the layer dimension.

The hidden layer in the first FC layer contains 256 neurons, the second FC layer contains 512 neurons, and the third FC layer contains 128 Nerve cell.

Finally the classification output layer which uses the softmax function to calculate the probability distribution of the seven classes. The formula used is defined by:

Dense (Num_classe, activation= softmax)

Num_classe: the classification classes (7 in our case).

3 Implementation

In this section, we present the environment of our implementation as well as the obtained results in comparison to other algorithms.

3.1 **Environment**

The Characteristics of the computer that we have used for training our algorithm:

Marque de PC : Dell Inc.

Operating system: Windows 10 pro.

Processeur: Intel(R) Core(TM) i5-4310U CPU @ 2.00GHz 2.60 GHz

RAM: 6 Go

System Type: 64-bit operating system, x64-based processor Pen and touch Touch support with 10 touch points

For the development, we have chosen Python for its simplicity and popularity in this field.

Programming language: Python 3.8 .8

Python development environment : spyder-2022, version : 4.2.5

Libraries: to do the database training and testing we have used the following libraries:

1. TensorFlow : It is a library for high-performance numerical calculations which makes it easy to build and deploy machine-learning applications.
2. Keras: It is a high-level Library for deep Learning, written in Python. It has been developed to allow rapid experimentation and quality research.
3. Numpy: Thee NumPy Library allows us to perform numerical calculations with Python. It introduces easier management of vectors and matrices.
4. Matplotlib: It is a feature rich Library for plotting high quality 2D graphs using python.
5. Open CV: This Library contains more than 2,500 optimized algorithms, which includes a comprehensive suite of both classic and modern computer vision and machine learning algorithms. These algorithms can be used to detect and recognize faces, identify objects, categorize human actions in videos, track camera movements, track moving objects, and extract 3D models of objects.

3.2 **Experimental results**

Herin, we present a detailed empirical analysis of our model in several facial expression recognition databases. We first provide a brief overview of the three databases used in this work, we then present the performance and compare the obtained results of our model with some promising recent work, basing on the task of categorizing each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

A. Datasets:

- The Facial Expression Recognition 2013 (FER2013) database: It was first introduced in the ICML 2013 workshope on challenges in Representation Learning [54]. The data consists of 48x48 pixel grayscale images offaces with 35887 images. The training set consists of 28709 examples and the public test set used for the leaderboard consists of 3589 examples. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount ofspace in each image. Four sample images from FER dataset are shown in Figure 3. 2.

Figure 3.2 Four sample images from FER database.

• The Extended Cohn-Kanade Dataset (CK+) [55]: It is a complete dataset for action unit and emotion-specified expression. CK+ dataset contains 593 video sequences from a total of 123 different subjects, ranging from18 to 50 years of age with a variety of genders and heritage. Each video shows a facial shift from the neutral expression to a targeted peak expression, recorded at 30 frames per second (FPS) with a resolution of either 640x490 or 640x480 pixels. Out of these videos, 327 are labeled with one of seven expression classes: anger, con- tempt, disgust, fear, happiness, sadness, and surprise. The CK+ database is widely regarded as the most extensively used laboratory-controlled facial expression classification database available, and is used in the majority of facial expression classification methods. Six sample images from this dataset are shown in Figure 3.3.

Figure 3.3 Six sample images from CK+ database.

• The Japanese Female Facial Expression dataset [56]: It consists of 213 images of different facial expressions from 10 different Japanese female subjects. Each subject was asked to do 7

facial expressions (6 basic facial expressions and neutral) and the images were annotated with average semantic ratings on each facial expression by 60 annotators. Sample images from this dataset are shown in Figure 3.4.

Figure 3.4 Six sample images from JAFFE database.

Images for fer2013, CK+ and JAFEE are categorized into 7 categories presented in Table I.

Emotions	Number of images FOR FER 2013	Number of images FOR CK+	Number of images FOR JAFFE
Angry = 0	4953	270	30
Disgust = 1	547	354	29
Fear = 2	5121	150	32
Happy = 3	8989	414	31
Neutral = 4	6198	108	30
Sad = 5	6077	168	31
Surprise = 6	4002	498	30

Table I. Numbers of images corresponding to each emotion.

B. Results Comparison and Analysis

For check the performance of our proposed model on the previous explained datasets, we have trained the different CNN models then reported the accuracy. Therefore, we have trained each model over the three datasets. Furthermore, we have used various hyper-parameters to fine- tune the model and train it on kaggle platform. For FER-2013 dataset, the different algorithms have trained for 50 epochs. For JAFFE and CK+ datasets, they have trained for 140 epochs from scratch. For the optimization of our algorithm, we have used Adam optimizer with

a learning rate of 0,001 with weight decay. It takes around 4 hours to train our models on FER dataset. For JAFFE and CK+, it takes a few minutes (from 5 to 10 minutes) to train a model since they have much fewer images.

For FER-2013 dataset, we have used around 28,709 images for train the model, 35,888 for validation, and 3,589 for testing. At this end, our algorithm is succeed to achieve the best accuracy through an accuracy rate of around 96.04 % , as shown in Figure 3.5.

```
Epoch 46/50
560/560 [==============================] - 1143s 2s/step - loss: 0.1256 -
accuracy: 0.9556 - val_loss: 0.0229 - val_accuracy: 0.9933
Epoch 47/50
560/560 [==============================] - 1144s 2s/step - loss: 0.1168 -
accuracy: 0.9608 - val_loss: 0.0297 - val_accuracy: 0.9905
Epoch 48/50
560/560 [==============================] - 1144s 2s/step - loss: 0.1217 -
accuracy: 0.9575 - val_loss: 0.0495 - val_accuracy: 0.9827
Epoch 49/50
560/560 [==============================] - 1141s 2s/step - loss: 0.1180 -
accuracy: 0.9585 - val_loss: 0.0165 - val_accuracy: 0.9936
Epoch 50/50
560/560 [==============================] - 1143s 2s/step - loss: 0.1168 -
accuracy: 0.9604 - val_loss: 0.0524 - val_accuracy: 0.9823
```

Figure 3.5 . The Result of train our model on FER 2013 dataset.

The comparison of the result of our model with some of previous works on FER 2013 are provided in Table II and Figure 3.6.

N°	Algorithm	Accuracy rate
1	Bag of Words [57]	67.4 %
2	VGG+SVM [53]	66.31%
3	GoogleNet [58]	65.2 %
4	Mollahosseini et al [59]	66.4 %
5	Gabor Wavelets [60]	70.02 %
6	Our algorithm	96.04 %

Table II. Obtained accuracies on FER 2013 dataset.

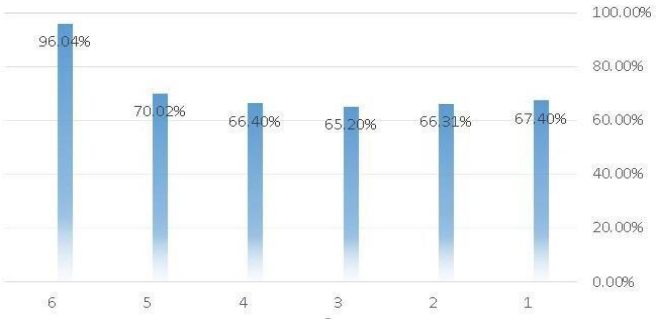

Figure 3.6 Obtained Accuracies on FER 2013 dataset.

As shown in the table II, our proposed algorithm realize an improvement of the accuracy rate by an average of 20.02 % compared to the other algorithms. This fact is coming due to the use of multiple enhanced strategies with more number of levels responsible to ensure more accuracy rate.

For JAFFE dataset, we have used 120 images for training, 23 images for validation and 70 images for testing. As result, our algorithm is succeed to achieve an accuracy rate of around 96.20 %, as shown in Figure 3.7.

```
- val_loss: 4.4994 - val_accuracy: 0.1875
Epoch 135/140
2/2 [==============================] - 3s 1s/step - loss: 0.1736 - accuracy: 0.9494
- val_loss: 4.8305 - val_accuracy: 0.1406
Epoch 136/140
2/2 [==============================] - 4s 1s/step - loss: 0.1564 - accuracy: 0.9747
- val_loss: 4.7120 - val_accuracy: 0.1875
Epoch 137/140
2/2 [==============================] - 4s 3s/step - loss: 0.1711 - accuracy: 0.9620
- val_loss: 4.7655 - val_accuracy: 0.1719
Epoch 138/140
2/2 [==============================] - 4s 1s/step - loss: 0.1595 - accuracy: 0.9620
- val_loss: 4.8132 - val_accuracy: 0.1562
Epoch 139/140
2/2 [==============================] - 5s 3s/step - loss: 0.1351 - accuracy: 0.9766
- val_loss: 4.7313 - val_accuracy: 0.1406
Epoch 140/140
2/2 [==============================] - 3s 1s/step - loss: 0.2232 - accuracy: 0.9620
- val_loss: 4.9268 - val_accuracy: 0.1094
```

Figure 3.7 The Result of train our model on JAFFE dataset.

The comparison with previous works on JAFFE dataset are shown in Table III as well as in Figure 3. 8.

N°	Algorithm	Accuracy rate
1	VGG+SVM [53]	89.02%
2	GoogleNet [58]	91.8 %
3	Mollahosseini et al[59]	95.31 %
4	Gabor Wavelets [60]	92.8 %
5	Our algorithm	96.20 %

Table III. Obtained accuracies on JAFFE dataset.

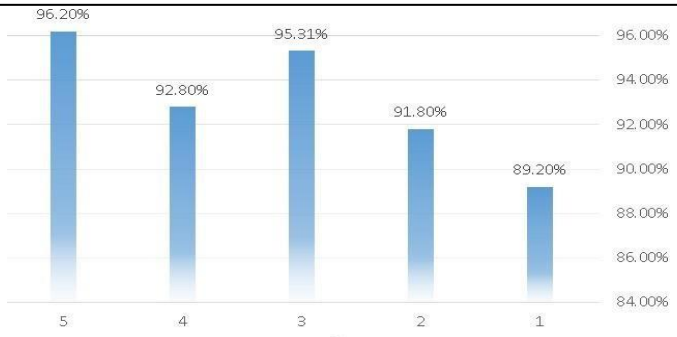

Figure 3.8 Obtained Accuracies on JAFFE dataset.

As shown in the table III, our proposed algorithm realize an improvement of the accuracy rate by an average of 4.3 % compared to the other algorithms.

For CK+ dataset, we use 70% of images for train the model, 10% for validation, and 20% for testing. We were able to achieve an accuracy rate of around 98.16%, as shown in Figure 3.9.

```
Epoch 137/140
15/15 [==============================] - 35s 2s/step - loss: 0.0134 - accuracy:
0.9956 - val_loss: 0.0016 - val_accuracy: 1.0000
Epoch 138/140
15/15 [==============================] - 35s 2s/step - loss: 0.0109 - accuracy:
0.9956 - val_loss: 0.0762 - val_accuracy: 0.9615
Epoch 139/140
15/15 [==============================] - 35s 2s/step - loss: 0.0067 - accuracy:
0.9967 - val_loss: 0.2272 - val_accuracy: 0.9125
Epoch 140/140
15/15 [==============================] - 35s 2s/step - loss: 0.0081 - accuracy:
0.9816 - val_loss: 0.0136 - val_accuracy: 0.9958
```

Figure 3.9 The Result of train our model on CK+ dataset.

The comparison of our model with previous works on the extended CK dataset are shown in Table I and Figure 3.10.

N°	Algorithm	Accuracy rate
1	VGG+SVM [53]	95.37%
2	GoogleNet [58]	97.2 %
3	Mollahosseini et al[59]	97.03 %
4	Gabor Wavelets [60]	98.03 %
5	Our algorithm	98.16 %

Table IV. Obtained accuracies on CK+ dataset.

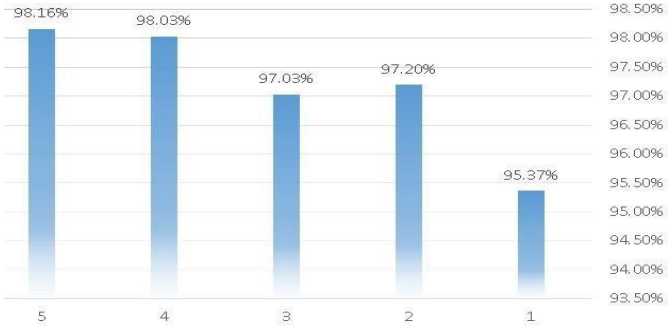

Figure 3.10 Obtained Accuracies on CK+ dataset.

As shown in the table IV, our proposed algorithm realize an improvement of the accuracy rate by an average of 1.25 % compared to the other algorithms.

C. Application practical images

Some practical images of our model after training and verifying it on human images, are shown.

Figure 3.11 .Example Surprise Emotions.

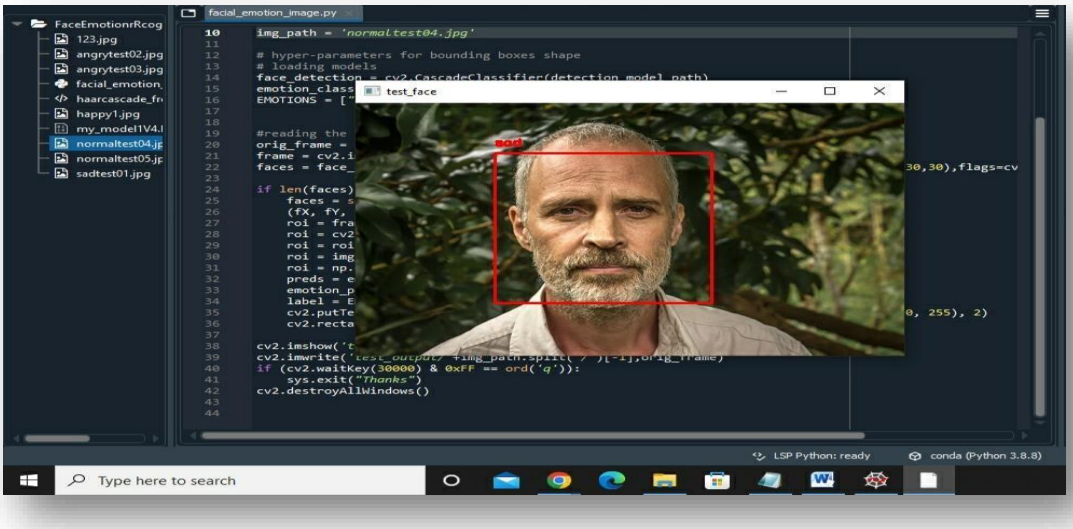

Figure 3.12 .Example Sad Emotions.

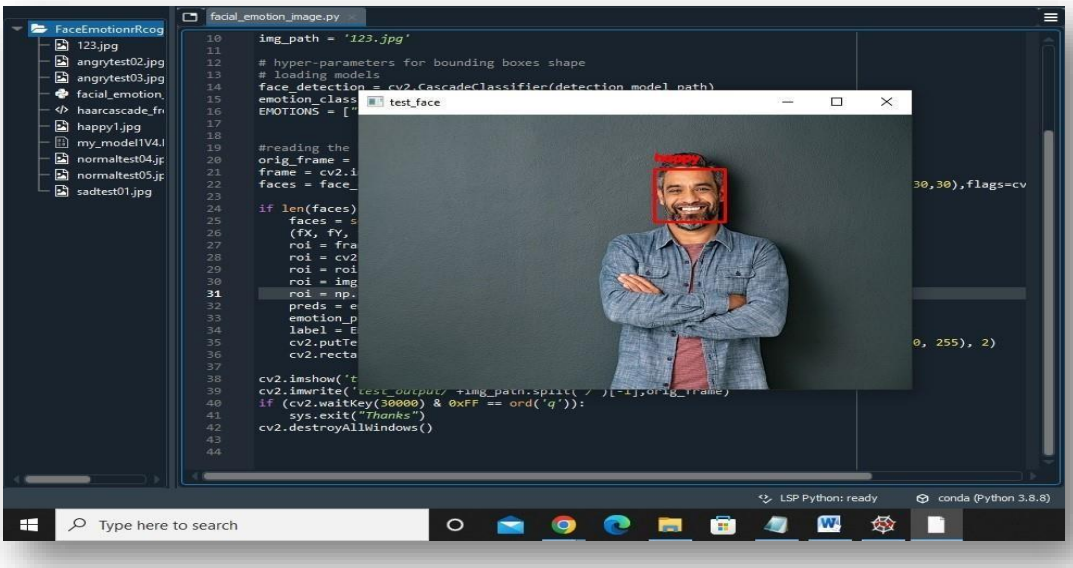

Figure 3.13 .Example Emotions.

4 Conclusion

In this chapter, we have presented our proposed convolutional neural network architectures, with the various results obtained from the database set FER2013, JAFFE and CK+.

The obtained results do not seem to be close to any previous work mentioned. In addition, these obtained results do show how well our model achieved the highest accuracy rate.

GENERAL CONCLUSION

In this dissertation, we have covered the basic concepts of image discovery, classification, and recognition. Second, we have discussed the different steps that an emotion recognition system relies on, namely: face detection, feature extraction, and classification. The goal is to design and implement an emotion recognition application.

Finally, we have presented our proposed algorithm that focuses on convolutional neural networks (CNN), our main objectives understanding human emotional behavior and improving the accuracy of emotion recognition on the database set FER2013, JAFFE and CK+, In addition, we have come up with the best ways to configure deep learning networks to capture key features individually or jointly. With the experimental results, our system has a great potential to detect emotions with high accuracy on various data.

As future work, we propose a more powerful facial emotions recognition algorithm that can be performed in 3D images.

BIBLIOGRAPHY

[1] *CH.Mohamed-Tahar,"Amélioration des images par un modèle de réseau de neurons", Memoir de fin d'études pour l'obtention du diplôme de Master en Informatique, univ.tlemcen, Université Abu-Bakr Belkaid –Tlemcen- Faculté des Sciences Department d'Informatique, 28 Septembre 2011.*
http://dspace.univ-tlemcen.dz/bitstream/112/1031/1/CHIKH-Mohammed-Tahar.pdf

[2] Kumari, Riya and Nikki, Shikha and Beg, Robin and Ranjan, Shashi and Gope, Sawan Kumar and Mallick, Ritesh Ranjan and Dutta, Arijit, "A Review of Image Detection, Recognition and Classification with the Help of Machine Learning and Artificial Intelligence (May 26, 2020)". International conference on Recent Trends in Artificial Intelligence, IOT, Smart Cities & Applications (ICAISC-2020). https://ssrn.com/abstract=3611339 or http://dx.doi.org/10.2139/ssrn.3611339.

[3] *Th.Hoang Le, "Applying Artificial Neural Networks for Face Recognition", Advances in Artificial Neural Systems,vol. 2011, Article ID 673016, 16 pages, 2011.*

[4] *"Tsubasa Hirakawa Takayoshi Yamashita" ,Journals & Books (Deep learning-based image recognition for autonomous driving) Chubu University, 1200 Matsumoto-cho, Kasugai, Aichi 487-8501, Japan , Accepted 26 November 2019.*

[5] What Does Image Recognition Mean?-Techopedia. https://www.techopedia.com/definition/33499/image-recognition#what-does-image-recognition-mean

[6] *"Denny Alquinta, Principal Cloud Architect-Oracle», livre .Qu'est-ce que l'image recognition AI ?,*
May 12, 2022 at. 9am PT.
https://www.oracle.com/fr/artificial-intelligence/image-recognition-ai-definition.

[7] S. Z. Li and A. K. Jain, Handbook of Face Recognition, Springer Science & Business Media, 2005.

[8] *Qu'est-ce que la recherche visuelle ?-Agence 90*
Agence 90 Paris : 4 rue de Jarente, 75004 Paris, FRANCE 2019.

https://www.agence90.fr/quest-ce-que-la-recherche-visuelle

[9] *M. Antoine, "La modération automatique des contenus en ligne"-Netino, 2022 Netino by webhelp, 2022.*
https://netino.fr/la-moderation-automatique-des-contenus-en-ligne

[10] *Computer vision-wikipidia, In Article is an interdisciplinary scientific field that deals with how computers, Wikipedia 2022.*
https://en.wikipedia.org/wiki/Computer_vision#Definition

[11] *N. Babich, "What Is Computer Vision & How Does it Work? An Introduction", 2020 Xd Ideas, Adobe, 2020.*
https://xd.adobe.com/ideas/principles/emerging-technology/what-is-computer-vision-how-does-it-work

[12] *outils de reconnaissance d'image à connaître-Graphiste.com.*
https://graphiste.com/blog/outils-reconnaissance-image

[13] *What is Facial Recognition – Definition and Explanation-Kaspersky, 2022 AO Kaspersky Lab.?*
All Rights Reserved.
https://www.kaspersky.com/resource-center/definitions/what-is-facial-recognition

[14] *Facial recognition system-Wikipedia, In Article On 29 April, 2022.*
https://en.wikipedia.org/wiki/Facial_recognition_system

[15] *Top 7 Use Cases for Facial Recognition In 2022-FaceMe.*
https://www.cyberlink.com/faceme/insights/articles/228/How_is_Facial_Recognition_Used_in_2021

[16] *21 Amazing Uses for Face Recognition – Facial ... - Face First, In 2022.*
https://www.facefirst.com/blog/amazing-uses-for-face-recognition-facial-recognition-use-cases.

[17] *Where is facial recognition used?-THALES, 2022 Thales.* *https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/inspired/where-facial-recognition-used*

[18] K. Mishra, Challenges Faced by Facial Recognition System, A Journal, August 18. 2020

[19] M. -H. Yang, D. Kriegman and N. Ahuja, Detecting Faces in Images : A Survey, IEEE Transactions on pattern analysis and machine intelligence, February 2002, pp. 34-58.

[20] Z. Orman, A. Battal and E. Kemer, A Study On Face, Eye Detection And Gaze Estimation, International Journal of Computer Science & Engineering Survey 2(3), August 2011.

[21] Q. M. Rizvi, B. G. Agarwal and R. Beg, A Review on Face Detection Methods, February 2011.

[22] G. Yang and T. S. Huang, "Human Face Detection in Complex Background," Pattern Recognition, vol. 27, no. 1, pp. 53-63, 1994.

[23] T. Sakai, M. Nagao, and S. Fujibayashi, "Line Extraction and Pattern Detection in a Photograph," Pattern Recognition, vol. 1, pp. 233-248, 1969.

[24] P. Sinha, "Object Recognition via Image Invariants : A Case Study," Investigative Ophthalmology and Visual Science, vol. 35, no. 4, pp. 1735-1740, 1994.

[25] A. Yuille, P. Hallinan, and D. Cohen, "Feature Extraction from Faces Using Deformable Templates," Int'l J. Computer Vision, vol. 8, no. 2, pp. 99-111, 1992.

[26] A. Lanitis, C.J. Taylor, and T.F. Cootes, “An Automatic Face Identification System Usi
Appearance Models,” Image and Vision Computing, vol. 13, no. 5, pp. 393-401, 1995.

[27] K.-K. Sung and T. Poggio, “Example-Based Learning for View- Based Human Face De
Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 39-51, Jan. 1998.
.

[28] S.A. Sirohey, “Human Face Segmentation and Identification,” Technical Report CS-TR
Maryland, 1993.

[29] N. Burton, “What Is an Emotion?,A Journal of Reviews”, January 3. 2016 .

[30] D. Hockenbury and S.E. Hockenbury, “ Discovering Psychology”, Worth Publishers, 2012 (6th edition).

[31] K. Cherry, “emotions and Types of Emotional Responses”, A Journal of Reviews, February 25,2022 .

[32] P. Ekman, “Basic Emotions Handbook of Cognition and Emotion”, 2005, pp. 45-60.

[33] R. Plutchik, “In Search of the Basic Emotions”, A Journal of Reviews, 1984, pp. 511-513.

[34] D. DOZOLME, “La détection d’une expression faciale incongrue par rapport à un modèle de situation émotionnel:un défi neurocognitif? ”,doctorat, unversite Paris-sud , 16/10/2014 .

[35] B. Parkinson, “ Heart to Heart How Your Emotions Affect Other People”, Cambridge University Press, 2019.

[36] K. VEMOU and A. HORVATH, “ Facial Emotion Recognition, A Journal of Reviews”, 2021, pp. 1-2.

[37] D. SHAHROKHIAN, “Syna: Emotion Recognition based on Spatio-Temporal Machine Learning”, Master, Eindhoven University of Technology, 2017 .

[38] D. Kriese, A Brief Introduction to Neural Networks, German, 2007.

[39] M. M. Zakaria, “ Classification des images avec les réseaux de neurones convolutionnels”, Master, Université Abou Bakr Belkaid Tlemcen, 2017.

[40] What is a Convolutional Neural Network?, https://ch.mathworks.com/fr/solutions/deep-learning/convolutional-neuralnetwork.html .

[41] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, http://www. deeplearningbook.org, MIT Press, 2016.

[42] N. Foued, “Reconnaissance d’expression faciale à partir d’un visage réel”, Master, Université de 8 Mai 1945 – Guelma - Faculté des Mathématiques, d’Informatique et des Sciences de la matière, Juillet 2019 .

[43] M. D. Youcef, “ Deep Learning pour la classification des images, Master”, Université Abou Bakr Belkaid Tlemcen, 2017.

[44] M. Navran, N. M. Charkari and M. Mansoorizadeh, “Automatic Facial Emotion Recognition Method Based on Eye Region Changes,Journal of Information Systems and Telecommunication”, Vol 4, October-December 2016, pp. 223-224.

[45] S. Z. Li and A. K. Jain, “Handbook of Face Recognition, Springer Science & Business Media”, 2005.

[46] S. Rajan, P. Chenniappan, S. Devaraj and N. Madian, “Facial expression recognition techniques: acomprehensive survey, the Institution of Engineering and Technology”, vol13, May 2019, pp. 1031-1040.

[47] Zhang Z., Lyons M., and Schuster M. et al: “Comparison between geometry-based and Gabor-based expression recognition using multi-layer perceptron”. Proc. Int. Conf. Automatic Face and Gesture Recognition, Japan, April 1998, pp. 26–39 .

[48] Z. XIANG , H. TAN AND W. YE, “The Excellent Properties of a Dense Grid-Based HOG Feature on Face Recognition Compared to Gabor and LBP”, ieee access, vol 13, 2018.

[49] D. Cristinacce and T. F. Cootes, “Feature detection and tracking with constrained local models.” in BMVC, vol. 1, no. 2, 2006, p. 3.

[50] Constrained local models, https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/Models/clm.html, (Accessed on 07/17/2017).

[51] T. Baltrušaitis, “Automatic facial expression analysis,” Ph.D. dissertation,University of Cambridge, 2014.

[52] S. Minaee1 and A. Abdolrashidi , “Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network,” Expedia Group, University of California, 2019.

[53] G. Mariana-Iuliana, R.T. Ionescu and M. Popescu, Local Learning with Deep and Handcrafted Features for Facial Expression Recognition, In arXiv preprint arXiv:1804.10892, 2018.

[54] P. L. Carrier, A. Courville, I. J. Goodfellow, M. Mirza, and Y. Bengio, “FER-2013 face database,” University of Montreal, 2013.

[55] J.R. Lee, L. Wang, and A. Wong, “EmotionNet Nano: An Efficient Deep Convolutional Neural Network Design for Real-Time Facial Expression Recognition,” In Artificial Intellenge, 3:609673, 2021.

[56] X. Chang, F. Nie, Z. Ma, and Y. Yang, “Balanced k-Means and Min-Cut Clustering,” In arXiv:1411.6235, 2014.

[57] I.R. Tudor, M. Popescu, and C. Grozea, “Local learning to improve bag of visual words model for facial expression recognition,” In Workshop on challenges in representation learning, ICML, 2013.

[58] G. Panagiotis, I. Perikos, and I. Hatzi- lygeroudis, “Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER2013,” In Advances in Hybridization of Intelligent Methods. Springer, Cham, pp. 1-16, 2018.

[59] M. Ali, D. Chan, and M.H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016.

[60] J. L. Michael , M. Kamachi, and J. Gyoba, “Coding Facial Expressions with Gabor Wavelets,” In arXiv:2009.05938, 2020.

Abstract—Automatic recognition of human emotions has received increasing interest from researchers in the field of computer vision, which has led to the proposal of several methods. Many of them relied on handcrafted features and traditional fusion and classification techniques. The use of deep Learning techniques to automatically extract powerful features from multimedia information as well as their use for merging and classification are new trends that researchers are currently pursuing. In This work, we define a new accurate facial expression detection algorithm based on a deep Learning method, specifically on an intentional convolutional neural network capable of focusing on important parts of the face in an image or video database using more number of needed layers. As a result, our proposed algorithm significantly improves the accuracy rate compared to previously proposed models in database groups FER2013, JAFFE and CK+.

Index Terms—Deep Learning, CNN, Computer Vision, Image recognition.

Résumé—La reconnaissance automatique des émotions humaines suscite un intérêt croissant de la part des chercheurs dans le domaine de la vision par ordinateur, ce qui a conduit à la proposition de plusieurs méthodes. Beaucoup d'entre eux s'appuyaient sur des caractéristiques artisanales et des techniques traditionnelles de fusion et de classification. L'utilisation de techniques d'apprentissage en profondeur pour extraire automatiquement des fonctionnalités puissantes à partir d'informations multimédias ainsi que leur utilisation pour la fusion et la classification sont de nouvelles tendances que les chercheurs poursuivent actuellement. Dans ce travail, nous définissons un nouvel algorithme de détection d'expression faciale précis basé sur une méthode d'apprentissage en profondeur, en particulier sur un réseau neuronal convolutif intentionnel capable de se concentrer sur des parties importantes du visage dans une base de données d'images ou de vidéos en utilisant un plus grand nombre de couches nécessaires. En conséquence, notre algorithme proposé améliore considérablement le taux de précision par rapport aux modèles précédemment proposés dans les groupes de bases de données FER2013, JAFFE ET CK+.

Termes de l'index— apprentissage en profondeur, CNN, vision par ordinateur, reconnaissance d'images.

الملخص -حظي التعرف التلقائي على المشاعر الانسانية باهتمام متزايد من الباحثين في مجال الرؤية الحاسوبية، مما أدى إلى اقتراح عدة طرق. اعتمد الكثير منهم على الميزات المصنوعة يدويًا وتقنيات الاندماج والتصنيف التقليدية. يعد استخدام تقنيات التعلم العميق لاستخراج الميزات القوية تلقائيًا من معلومات الوسائط المتعددة بالإضافة إلى استخدامها للدمج والتصنيف من الاتجاهات الجديدة التي يتابعها الباحثون حاليًا. في هذا العمل، قمنا بتعريف خوارزمية جديدة دقيقة للكشف ً ء على طريقة التعلم العميق، وتحديدًا على شبكة عصبية تلافيفيه مقصودة قادرة على التركيز على عن تعبيرات الوجه بنا أجزاء مهمة من الوجه في قاعدة بيانات الصور أو الفيديو باستخدام عدد أكبر من الطبقات المطلوبة. نتيجة لذلك، تعمل مسبقً في. مجموعات قواعد الخوارزمية المقترحة لدينا على تحسين معدل الدقة بشكل كبير مقارنة بالنماذج المقترحة ا CK +.وJAFFE و البياناتFER2013

مصطلحات الفهرس -التعلم العميق، سي إن إن، رؤية الكمبيوتر، التعرف على الصور