Development of handcrafted and deep based methods for face and facial expression recognition

Mohamed Kas

THÈSE DE DOCTORAT EN CO-TUTELLE

DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE TECHNOLOGIE DE BELFORT-MONTBÉLIARD

ET DE L'ÉTABLISSEMENT FACULTÉ DES SCIENCES Á L'UNIVERSITÉ IBN TOFAIL AU MAROC

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Informatique

par
MOHAMED KAS

Development of handcrafted and deep based methods for face and facial expression recognition

Thèse présentée et soutenue à UTBM Montbéliard, le 07 juillet 2021
Composition du Jury :

M. Hichem SNOUSSI	Professeur des universités Université de Technologie de Troyes (UTT)	Rapporteur
M. Abdellatif ELAFIA	Professeur des universités ENSIAS, Université Mohammed V de Rabat	Rapporteur
M. Ludovic MACAIRE	Professeur des universités Université de Lille	Examinateur
M. Habib BENLAHMAR	Professeur des universités Faculté des Sciences Ben M'sik, Université Hassan II, Casablanca	Examinateur
M. Yassine RUICHEK	Professeur des universités Université Bourgogne - Franche-Comté	Directeur de thèse
M. Messoussi ROCHDI	Professeur des universités Faculté des Sciences, Université Ibn Tofail, Kénitra	Co-Directeur de thèse
M. Youssef EL MERABET	Maître de conférences Faculté des Sciences, Université Ibn Tofail, Kénitra	Co-encadrant de these

N° | X | X | X |

ACKNOWLEDGEMENTS

This Ph.D. journey started in 2016 thanks to a collaboration between the University of Technologies Belfort-Montbéliard France and the Faculty of Science of Ibn Tofail University Morocco. All my researches have been carried out jointly between the LASTID laboratory in Kénitra and CIAD one in Montbéliard. This collaboration helped my thesis to be supervised by many professors leading to more interactions and scientific exchanges.

First and foremost, I would like to express my special appreciation and thanks to the defense committee for accepting to judge my work. I thank Professors Hichem Snoussi and Abdellatif Elafia for their considered time to read and evaluate this dissertation. Also, I thank Professors Ludovic Macaire and Habib Benlahmar for being examiners of my thesis defense.

I would like to thank my supervisor at UTBM, Prof Yassine Ruichek, for all the support, encouragement, and guidance he granted me. With his guidance and constant feedback, this Ph.D. achieved many contributions and publications. He gives interest and time to each scientific work with his prominent recommendations and numerous corrections. I thank him for his efforts to correct my writings and for providing the necessary computation machines to perform the experiments. He is endowed with perfection, dynamism, openness, and fascinating scientific standards.

I gratefully thank Prof Rochdi Messoussi, my supervisor at Ibn Tofail University and director of LASTID laboratory and with whom I started this Ph.D. I thank him for believing in me, taking me as one of his Ph.D. students, and helping me to establish this collaboration with UTBM. Also, he demonstrated excellent listening and communication capacities during my research stays in Morocco, providing all the requested materials and supports. Prof Messoussi always tries to build a research environment that promotes the interaction between the labmates and produces more scientific interactions.

In addition to my directors, I sincerely thank Prof Youssef El Merabet for his continuous guidance and support as my Co-advisor. This thesis has started thanks to him as he advised me to start a Ph.D. with the LASTID laboratory right after defending my engineer degree. He introduced to me the field of this thesis, the laboratories, and the possible collaborations that can be established. I thank him for all the scientific exchanges we had throughout this thesis and the corrections he made to my journal and conference papers.

I gratefully thank Prof Raja Touahni for all her efforts for all the administrative procedures

related to my thesis and its defense, scholarships, and for her high availability and mental support. She gave many valuable pieces of advice that will help me in my professional and personal life.

I would like to thank all the members of the two laboratories for the lively and friendly atmosphere in the office and our generous and pleasant exchanges, inside and outside the laboratory. Their support was invaluable throughout my thesis, making my time both memorable and enjoyable. I really enjoyed the daily coffee breaks I had with my labmates and professors.

Last but not least, I would like to extend my sincere gratitude to my parents for their continued support, which has helped me stay strong and focused on the thesis work. I also extend my thanks to my sisters, my brother, and all my relatives and friends for their affection and encouragement.

CONTENTS

1 Introduction **1**

1.1 Face perception 1

1.2 Facial Analysis motivations 4

1.3 Targeted Facial Tasks 8

1.4 Outline of the PhD thesis dissertation 11

2 Facial analysis Literature **13**

2.1 Introduction 13

2.2 Facial Image Classification Framework 14

2.3 Feature Extraction 15

2.3.1 Holistic description 16

2.3.2 Local description 16

2.3.2.1 Local Binary Patterns LBP 16

2.3.2.2 LBP-like descriptors 18

2.3.3 Learnable features 21

2.3.3.1 Principal Component Analysis Network (PCANet) 22

2.3.3.2 Compact Binary Face Descriptor (CBFD) 23

2.3.3.3 Deep Convolutional Neural Networks 24

2.4 Classification 25

2.4.1 Nearest Neighbor 25

2.4.2 Support Vector Machines 27

2.4.3 Neural Networks 28

2.5 Face recognition state-of-the-art 29

2.6 Facial expression recognition state-of-the-art 31

2.7 Conclusion 33

3 Local description-based face recognition 35

3.1 Introduction . . . 35

3.2 Mixed Neighborhood Topology Cross Decoded Patterns . . . 36

3.2.1 Neighborhood topology . . . 36

3.2.2 Pattern encoding . . . 38

3.2.3 MNTCDP feature vector . . . 40

3.3 Face recognition system using MNTCDP . . . 42

3.4 Experimental Analysis . . . 43

3.4.1 Datasets . . . 43

3.4.1.1 ORL . . . 44

3.4.1.2 YALE . . . 44

3.4.1.3 Extended YALE B . . . 44

3.4.1.4 FERET . . . 45

3.4.1.5 AR . . . 45

3.4.2 Experiments configuration . . . 46

3.4.3 Experimental evaluation against LBP-like descriptors . . . 48

3.4.3.1 Performance analysis on ORL: Experiment #1 . . . 48

3.4.3.2 Performance analysis on FERET: Experiment #2 . . . 49

3.4.3.3 Performance analysis on YALE: Experiment #3 . . . 50

3.4.3.4 Performance analysis on Extended YALE B: Experiments #4 and #5 . . . 50

3.4.3.5 Performance analysis on AR Face Database: Experiment #6 . . . 52

3.4.4 Experimental evaluation against PCANet2 and CBFD deep features 53

3.5 Comparison with state-of-the-art systems . . . 55

3.6 Implementation in the Human Support Robot (HSR) of UTBM . . . 66

3.7 Conclusion . . . 67

4 GAN-based Profile face recognition 69

4.1 Introduction . . . 69

4.2 Face pose translation literature review . . . 70

4.3 Proposed GAN-based PIFR framework . . . 71
4.3.1 Overall framework . . . 72
4.3.2 Generative Adversarial Networks . . . 73
4.3.2.1 Generator models . . . 74
4.3.2.2 Discriminator models . . . 76
4.3.3 CNN based face classification . . . 76
4.4 Experimental Analysis and Discussions . . . 77
4.4.1 Database . . . 78
4.4.2 GAN's architectures evaluation . . . 79
4.4.3 Evaluation of the proposed PIFR framework . . . 82
4.4.4 Implementation and execution time analysis . . . 83
4.5 Conclusion . . . 84

5 Facial Expression Recognition 87

5.1 Introduction . . . 87
5.2 Static Person-Independent FER . . . 88
5.2.1 OPD-GQMBP: New handcarfted descriptor for FER . . . 88
5.2.2 Overall FER framework . . . 91
5.2.3 Experimental Analysis and Discussions . . . 93
5.2.3.1 Experimental datasets . . . 94
5.2.3.2 Evaluation of OPD-GQMBP neighborhood size configuration 96
5.2.3.3 Comparative analysis against state-of-the-art handcrafted and deep feature methods . . . 97
5.2.3.4 Comparison against state-of-the-art FER systems . . . 99
5.2.3.5 Confusion matrix-based analysis for the FER . . . 100
5.2.3.6 Implementation and Execution time . . . 104
5.3 Dynamic Person-Independent FER . . . 106
5.3.0.1 Longer Short Term Memory . . . 106
5.3.0.2 Deep CNN-LSTM for Dynamic FER . . . 107
5.3.1 Experimental Analysis . . . 110

5.3.1.1 Dynamic FER datasets 110
5.3.1.2 Evaluation of the proposed CNN-LSTM 111
5.3.1.3 Comparison against state-of-the-art 113
5.4 Conclusion 114

6 Conclusion **117**

6.1 Thesis Summary 117
6.2 Future Works and Perspectives 119
6.3 Publications 121

1

INTRODUCTION

1.1/ FACE PERCEPTION

The face of a human being carries enough information to extract cognitive classification about his identity, facial expression, gender, and estimating his age. The face is considered as the primary communication channel rich in information observed by the environment and interpreted with high precision. Moreover, facial media is the most universal across all communications channels. It can be understood regardless the social, geographical, gender, and age belonging. The face particularities highlighted in the distinguishing of specific visual features related to the person's identity, emotional state, and other biometrics led to the computer vision-based facial analysis. The strategy consists of emulating the human perception skills on a computer through machine learning techniques, which proved that computers could perform perception tasks by learning and not programming similarly to humankind. The motivation also comes from the nature of the facial features, which are visual and not intrinsic or to be concluded as verbal language. The computer vision discipline cares about how computers can process the outputs of digital images and videos and understand them to make decisions. It covers all analysis performed by our biological vision pipeline, including seeing a visual stimulus, processing it, and then extracting semantic information that can be used to make a decision or exploited in other processes.

In our social life, we spend more time interpreting faces than any other single stimulus through face perception. It is classified as the human brain's most sophisticated perceptual skill [1]. Since their birth, infants demonstrate fundamental facial processing capacities and show keen interest in faces [2]. Babies (1–3 days) can detect faces even with rotations reaching 45 degrees [3]. However, the studies found that this interest in faces is not sustained in early childhood cycles as the child grows. The interest is reduced in children aged 1 to 4 months. It re-emerges and seems to reach the peak lately on the first year, but it decreases slightly over the next two years of growth [4]. The observed re-emergence may be motivated by the child's self motor abilities and experiences [4].

Infants of two days old can mimic adult facial expressions, demonstrate the ability to recognize details such as the shape of the mouth and eyes, and move their muscles to create similar patterns on their faces [5]. Despite this ability, newborns so far are not conscious of the emotional content represented through facial expressions. Five-month-old infants pay equal attention to a person's image-making an anxious facial expression and making a happy facial expression, and showing similar Event-Related Potentials (ERPs) for both expressions. Nevertheless, seven-month-olds pay more attention to the fearful face, and their ERPs-based response for the scarred face is much stronger than the one for the happy face. Hence, this statement highlights an increased attentional and cognitive focus toward fear that reflects the emotion's threat-salient nature [6]. Face perception is a highly complex function for the human visual system (Figure 1.1). Some non-faces categories share visual properties with faces and potentially resemble faces (such as fruits or animals).

Figure 1.1: Examples of face-like objects

Faces represent a very homogeneous visual category and are visually similar (surface, facial elements, general structure). Finally, the perception of faces must remain effective in a natural environment and, therefore, despite many physical conditions that modify faces' visual appearance (orientation of the head, relative size, level of lighting, partial vision, expression of emotions, etc.). Face perception allows at a glance to (1) detect faces in the scene and differentiate them from other categories (objects, animals) (face detection); (2) discriminate between several identities (individual face discrimination); (3) recognize familiar characters (recognition of faces). These processes are carried out efficiently by the human brain despite significant variations in relative size, the head's orientation, facial expression. Indeed, humans are performing well in perceiving faces. A face can be detected in a visual scene in 100ms [7], and a face can be perceived as

familiar in 200ms [8]. This efficiency makes it very difficult to understand the mechanisms involved. Understanding how the human brain perceives faces is, therefore, an essential issue for neuroscience. To this end, identifying the brain structures involved is a crucial step. The search for these brain structures began with neuropsychology and the description of patients with brain damage and then developed strongly 20-30 years ago to develop functional neuroimaging. Our ability to recognize a face seems so easy, fast, and automatic that it is difficult for us to imagine the mechanisms involved. As early as the 19th century, despite the absence of scientific evidence, some scientists wondered. For example, Francis Galton, who was interested in the physiognomy of faces, wrote in 1883: " The general expression of a face is the sum of a multitude of small details, which are viewed in such rapid succession that we seem to perceive them all at a single glance." Later, psychology made the same hypothesis, suggesting that faces are perceived as a whole, as a global unit [9]. This ability to perceive the multiple elements of a face simultaneously within a single, global representation is called a holistic process [10; 11]. Several shreds of evidence support a holistic mechanism of the perception of the faces. One compelling evidence for a holistic process of facial perception comes from the composite illusion of faces [12]. This illusion is based on the fact that one part of a face cannot be perceived without being combined with the rest parts. For example, two identical upper halves will be perceived as different if they are associated with two different lower halves (Figure 1.2). Another proof comes from the face inversion effect, which is a phenomenon where identifying inverted (upside-down) faces compared to upright faces is much more complicated than doing the same for non-facial objects. Behavioral studies have shown that a face's perception is disturbed when faces are inverted [13; 14; 15]. Thus, faces' perception is not based on the individual elements (nose, mouth, and eyes, are the same upright and inverted) but on a global representation only when the face is seen upright.

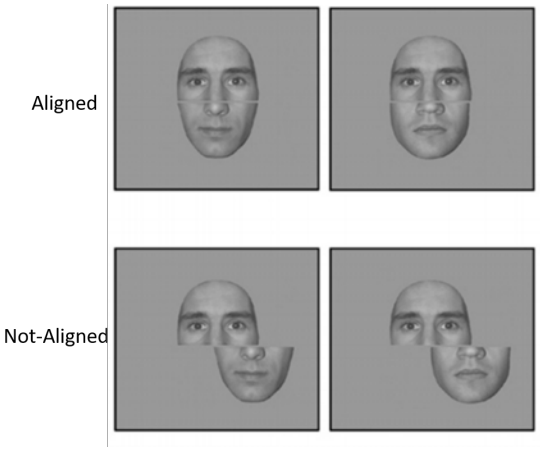

Figure 1.2: Lower and Upper parts missalignement effect on face percetion

From a theoretical point of view, the privileged theory concerning the organization of the visual system is a hierarchical organization with initially a perception "element by element"

then a perception of these elements in a global representation [16; 17]. This view is almost incompatible with a holistic perception of faces and the results presented above in which global representation plays a significant role (composite illusion, inversion effect). Another model for organizing facial perception called "coarse-to-fine" has been proposed [11], where faces would always be seen as a whole, first in a rough representation only to categorize a face as a face. Then, this global representation would be enriched in detail until reaching enough information to individualize the face and determine its identity.

In 1997, the authors in [18] published an functional Magnetic Resonance Imaging (fMRI) study introducing the so-called "functional localizer" approach. This approach consists of presenting participants with faces and control stimuli (familiar objects) and asking them to complete a simple task on one of the two categories. A brain area is defined as selective to faces if it responds more to faces than to objects. This area is defined as a region of interest, and the role of this area is then further investigated using other fMRI experiments. Using this approach, the authors identified the most particular face area and localized it in the right spindle-shaped gyrus, in its posterior and middle parts (Figure 1.3). This area was named FFA (Fusiform Face Area). The authors considered this area to be the only important modulus for facial perception, ignoring the other selective facial areas highlighted in their study. The faces' perception would no longer be based on a vast network, but mainly on a single region, the most selective to faces. Subsequent studies in the 2000s were greatly influenced by this view and often focused their analyzes on FFA.

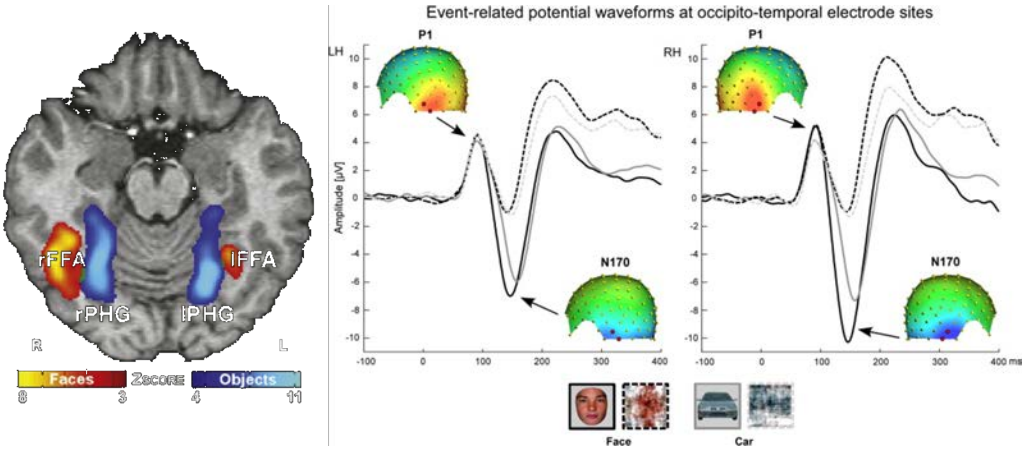

Figure 1.3: Left: fMRI image of the FFA, Right: ERPs responses of the FFA to Face/Object [19]

1.2/ FACIAL ANALYSIS MOTIVATIONS

Discovering the dedicated human FFA responsible for faces' perception within a human-to-human context gave birth to computer-based facial analysis to incorporate human-to-

machine interactions. The facial analysis takes advantage of the growing progress on three main research pillars: Computer vision, Pattern recognition, and Texture/Color feature extraction.

With the rise of the internet and modern technologies, we can now confidently confirm that we live in a society of images. Nowadays, anyone can use their smartphone's camera to take a photo or video and share it on the web and social media. Large amounts of video are uploaded to platforms like YouTube. All these images that flood the internet are full of data that could be valuable for businesses. However, to be able to collect and analyze this data, computers need to be able to "see" an image and understand its content. This is the main objective of Computer Vision. The field of Computer Vision brings together multiple techniques from various fields of engineering or computer science. In general, the different methods aim to reproduce human vision. In order to understand the content of images, machines must extract a description: an object, a description, a 3D model, etc. Specific computer vision systems may also require image processing, namely simplification or an increase in its content. Examples include normalizing the image's photometric properties, cropping its edges, or removing "noise" such as digital artifacts induced by low light. Allowing a computer to emulate biological vision is not straightforward. To this day, computer vision still fails to match the human vision. Part of the reason is that we still do not know how human vision works. We need to understand how perceptual organs such as the eyes work, but also how the brain interprets this perception. Although research in this area is advancing rapidly, we are still far from unraveling all the mysteries of vision. Another challenge is linked to the complexity of the visual world. An object can be perceived from multiple angles, partially hidden by other objects in various lighting conditions. However, a proper computer vision system must perceive the content in any situation and extract information from it. In fact, computer vision represents a real scientific challenge.

Despite the difficulties associated with computer vision development, significant advances have been made over the years that have already enabled Computer Vision to perform many tasks. It is effective for optical character recognition, also called "OCR." It is also used in automatic checkouts. The photogrammetry for generating 3D models is also based on computer vision. It is also used for machine inspection for medical imaging analysis. In the field of automobile safety, computer vision is used for the detection of dangers. With the emergence of self-driving cars, Computer Vision nowadays takes center stage in the automotive industry as it is what will allow vehicles to "see" on the road. The facial recognition technologies of the most recent smartphones, such as the famous Face ID of the latest Apple iPhones, are also based on Computer Vision. Automatic surveillance cameras also exploit this technology.

On the other hand, the pattern recognition field concerns the automatic discovery of patterns in data through computer algorithms and the use of these patterns to carry out

actions such as classifying data into different categories. Machine learning , which originates from artificial intelligence, is commonly used to refer to supervised learning methods. In contrast, data mining places more emphasis on unsupervised methods and a stronger connection to professional use. Pattern recognition has its functions in engineering. The term is popular in computer vision: one of the major conferences on computer vision is called the Computer Vision and Pattern Recognition Conference. In pattern recognition, it may be more interesting to formalize, explain, and visualize the pattern, while machine learning has traditionally focused on maximizing recognition rates. However, all of these fields have evolved considerably from their roots in artificial intelligence, engineering, and statistics. They have become more and more similar by integrating each other's developments and ideas.

In machine learning, pattern recognition works on labeling a given input, and it was referred to as discriminant analysis in statistics back to 1963. The most general application of pattern recognition is classification, where a classifier learns to predict a label to the processed input according to the training classes. However, pattern recognition is a more general issue that encompasses other types of output as well. Other examples are regression, which predicts a true-valued output to each input; sequence tagging gives a class to each pattern of a sequence of values (e.g., part of speech markup). Pattern recognition techniques generally aim to learn a logical response for all possible inputs and perform the "most probable" match of the inputs, considering their statistical variations. This is opposed to pattern matching, which looks to exact matches in the input with pre-existing patterns stored in a database. A typical example of a pattern-matching algorithm is regular expression matching. It searches for patterns of a given kind in text data, which is included in many text editors and word processors' search capabilities. Unlike pattern recognition, pattern matching is generally not a type of machine learning. However, pattern-matching algorithms (especially with reasonably general and carefully tailored shapes) sometimes provide output similar in quality to pattern recognition algorithms.

The flexibility of pattern recognition algorithms and their application alongside computer vision motivated human face analysis through machine learning. The facial analysis relies on processing the visual features by extracting them first and then finding their patterns. The visual components are the combination of three features: textural, color, and shape. The computation of relevant facial characteristics that help machine learning frameworks to predict the face identity or the dominant emotion according to the input face image relies on the discriminative power of the applied textural, color, and shape feature extraction methods.

The textural analysis is widely present in our natural world and plays a significant role in various critical applications. Any object or pattern's visual appearance can be represented in a texture form at a certain level by its size, shape, organization, and proportions

of its parts. The texture is detected on both artificial and natural objects such as on wood, plants, materials, and skin. Texture analysis becomes a fundamental branch of image processing and computer vision by exploring how objects can be textured. Hence, many applications can be redefined as texture classification tasks, including face recognition and content-based image retrieval [20]. Moreover, texture classification has been adopted in medical image analysis and helps to achieve good results in congestive heart failure [21], human skin analysis [22], brain degenerative diseases [23], etc. During past decades, texture classification gained too much attention due to its difficulties in terms of variability and inhomogeneity, such as scale changes, variable illumination, surface shape variability, and imaging conditions.

The description or representation of shapes is an essential topic in image analysis for object recognition and classification. Descriptions are given in terms of properties of the objects contained in the images and relationships between them. These properties correspond to the characteristics of the position, size, and shape of the objects. Each shape or image to be stored in the database is processed to obtain the shape features. The shape features are then used by the various shape representation techniques to organize and efficiently retrieve the useful shape information into index structures. Shape-based feature extraction has been widely used to develop face detectors that demonstrate outstanding performance even when the face is affected by affine transformations. Also, the recognition of emotions from facial expression relies on computing shape information on the regions that are believed to present emotion-related visual components (eyes, mouth, nouse).

On the other hand, color is also important and the most straightforward feature that humans perceive when looking at an image. The human visual system, by nature, is more sensitive to color information than to grayscale, so color is the first candidate used for feature extraction. However, the facial analysis relies on dealing with the color features as three separated plans RGB, where each one of them is fed to extract the textural and shape-based characteristics. Then, a strategy is to gather the three extracted features and form the final one describing the color input face image.

After all, the discussed theoretical advances in computer vision and pattern recognition depend mainly on computation power. Nowadays, many calculation devices are available for the research community with more capabilities and lower prices than in the early ’80s. Also, the appearance of ready machine learning environments based on cloud computing helps the researchers to focus on coding their design ideas rather than solving hardware and software issues that were used in face analysis earlier. Moreover, the machine learning dedicated environment like Python and MATLAB managed to offer standard programming platforms so the researchers can share their source codes, which accelerated machine learning progress based on reusing, enhancing, and sharing pipelines.

1.3/ TARGETED FACIAL TASKS

The research objectives of this thesis concern the development of new concepts for image segmentation and region classification. This involves implementing new descriptors, whether color, texture, or shape, to characterize regions and propose new deep learning architectures for the various applications linked to facial analysis. We restrict our focus on face recognition and person-independent facial expressions classification tasks, which are more challenging, especially in unconstrained environments. Face recognition is an increasingly popular technology, based on artificial intelligence, to identify a person in a photo or video by comparing their face with those stored in a database. This technology relies on capturing the visual features, that are converted into data (pixels) and processed to calculate a low dimensionality feature space. The obtained space is compared and matched with the reference database to find the most similar identity to the detected face. Typically, traditional facial recognition frameworks analyze around 80 facial feature regions referred to as nodal points. These features include geometric measurements as the distance between the eyes and the length of the nose. These visual and geometric characteristics differ from one person to another, helping facial recognition be accurate in identifying technology. The most recent new technologies are based on the texture, shape, and color analysis of the skin that is unique to each individual, leading to more precise results. In the early '90s, facial recognition began to gain popularity when the US Department of Defense searched for a technology that could detect people. In early 2001, facial recognition made an impression when it was first used in a public space during Super Bowl XXXV in Tampa. The authorities then used it to detect possible criminals and terrorists among the spectators. Systems were then deployed to other at-risk areas of the United States to monitor potential criminal activities. Currently, face recognition is regarded as one of the top three biometric technologies for identifying a person and the fastest-growing biometric technology. Its market could reach a value of 7.7 billion dollars by 2022. This technology is gradually used for surveillance and security purposes, especially by governments and authorities who incorporate it into video surveillance systems. Indeed, facial recognition has been used in France, the United Kingdom, and the United States. For example, Nice's city is experimenting with facial recognition for surveillance purposes as part of its Carnival, while the US government is using it at airports to identify individuals whose visas have expired. Companies in various industries are also more and more exploiting facial recognition, such as health, marketing, or tourism. It is also found on many services and products intended for the general public. For example, since the iPhone X exhibition in 2017, Apple smartphones feature Face ID technology allowing users to unlock them by showing their faces to the front camera. A 3D scanner compares more than 30,000 characteristics to verify user identity accurately. Face ID also allows validating purchases with Apple Pay. For its part, Facebook is developing Deep-

Face technology, which automatically identifies people's faces in photos uploaded to the social network with 97% accuracy. Every time a Facebook user is "tagged" on a photo, their facial features are mapped by the system.

On the other hand, facial expression recognition works on detecting the dominant emotion expressed through the face. Over the past two decades, computer vision and pattern recognition communities have shown a great interest in analyzing and recognizing facial expressions. Initially inspired by cognitive science researchers' discoveries, the computer vision and scientific research community envisioned developing systems capable of recognizing facial expressions in videos or static images. Most of these facial expression analysis systems attempt to classify expressions into a few broad emotional categories, such as joy, sadness, anger, surprise, fear, and disgust.

Facial expression is the most expressive way for humans to communicate emotions and signal intentions, which conveys non-verbal communication signals in face-to-face interactions. A facial expression is a visible representation of a person's activity, intention, personality, and psychopathology. Facial expression, along with other gestures, transmits non-verbal communication signals in face-to-face human interaction. These clues can also complement speech by helping the listener get the desired meaning from spoken words. They play an essential role in our relationships. They can reveal a person's attention, personality, intention, and psychological state. These are interactive signals that can regulate our interactions with the environment and other people in our neighborhood.
This task serves diverse applications that interest many markets. It can be helpful to children diagnosed with autism to understand their social environment better. Moreover, such technology would evaluate E-learning contents and public services more efficiently, accurately, and in real-time. Furthermore, the industry of Human Support Robot would develop robots qualified to adapt their interactions according to the emotional atmosphere. In real-world scenarios, the desired system is expected to recognize unseen individuals' emotions in real-time, which makes this task among the toughest ones in computer vision. In the literature, the recognition of facial expression presents four different levels, as can be seen in Figure 1.4, which shows the difficulty levels regarding the way the emotion is expressed (Spontaneous vs. Posed) and the person expressing it (the same person as the train or a different one). Spontaneous emotions are hard to be classified since each individual expresses a given emotion differently compared to another person.

Furthermore, this fact often leads to interclass samples interferences meaning that two emotion classes are represented over two images with the same overall appearance. Moreover, the recording of spontaneous emotions must be performed. Simultaneously, the subjects are not aware of it, which is very hard to establish since the subjects should deliver authorization to record and use their images/videos. Therefore, few works were interested in spontaneous emotion recognition and focused only on verifying the

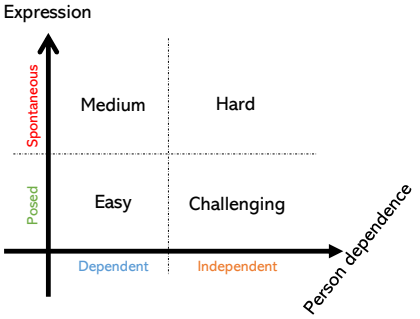

Figure 1.4: Facial Expression Recognition difficulty levels

assigned labels whether they match the corresponding observation or not as reported in [24]. The majority of available databases for spontaneous facial expression are collected from the web, based on saving the search engines' images (Google and Flicker mainly) by specifying the emotions-related keywords. The well-known and widely used databases of this kind are FER, AFEW. On the other side, posed expressions are obtained by requesting the subjects uniformly perform the facial expressions to avoid intra-classes similarities that would confuse the classification task. The subjects are usually skilled persons (actors), so their expressions could be computationally classified. The second challenge of facial expression recognition relies on correctly decoding individuals' observations not taking part in the train session. In this thesis, we deal with person-independent posed FER.

The developed framework for face recognition and facial expression classification, within the context of this thesis, should fulfill the following criteria:

- The developed systems must be reliable in terms of recognition/identification results by achieving very high recognition rates.
- Provide a stable performance when dealing with an unconstrained environment such as the lighting conditions, background inhomogeneity, the face's position, facial expressions, and, most importantly, the recognition of obscured faces.
- As with recognition rate, execution time is also an essential metric for evaluating facial recognition architecture, knowing that most future applications will be real-time.
- Compatibility with low-end devices such as robot systems, prototyping platforms (Raspberry Pi, Arduino), and mobile devices.
- Re-usability of the developed source codes by sharing them with the research community, which will offer opportunities for more enhancements and further implementations.

1.4/ OUTLINE OF THE PHD THESIS DISSERTATION

In order to give the readers and community a comprehensive presentation of the contributions introduced in this thesis and also preparing their background to a fluent reading experience, the rest of the thesis dissertation is organized as follows:

- Chapter 2 reviews the typical facial analysis configuration. It discusses in-depth the key steps composing such a framework, namely feature extraction and classification. This chapter also highlights some existing works from the literature devoted to face recognition and facial expression analysis.
- Chapter 3 presents our proposed framework for face recognition based on a new local features descriptor referred to as Mixed Neighborhood Topology Cross Decoded Patterns (MNTCDP). It explains the overall architecture and the considered benchmark for a comprehensive evaluation. It also discusses a real implementation of the proposed framework within a Human Support Robot of Toyota.
- Chapter 4 is devoted to introducing a brand new solution to deal with the recognition of profile images. It is based on Generative Adversarial Network-based image translation (GANs). We trained a GAN to generate a profile input's frontal face and then processed it with existing frontal recognition systems. Also, Chapter 4 includes a comprehensive explanation of the concept of GAN for image translation purposes.
- Chapter 5 exhibits our contributions to the facial expression recognition task that cover both static and dynamic-based scenarios. The static-based FER framework relies on extracting textural and shape features from specific face landmarks that carry enough information to detect the dominant emotion through the SVM classifier. On the other hand, dynamic FER contribution incorporates Long Term Short Memory (LSTM) deep network to encode the temporal information efficiently with a guiding attention map to focus on the emotion-related landmarks and guarantee the person-independent constraint. This chapter includes comprehensive comparisons with the state-of-the-art works.

2

FACIAL ANALYSIS LITERATURE

2.1/ INTRODUCTION

Facial analysis frameworks are based on a standard image classification pipeline, aiming to develop a system capable of automatically assigning a label to an image. Such frameworks rely on machine learning techniques to develop a computational capacity emulating human skills, and their life cycle includes two phases. The first phase is to train the computer to construct a model representing the relationship between the input space, which is the image pixels, and the decision to be made represented as labels or classes. The second evaluates the framework by calculating performance indicators such as precision, accuracy, recall, and computation complexity on unseen inputs that simulate a real use case of the developed framework. If the reported metrics satisfy and fulfill the requirements, we proceed to implement the considered image classification system for general use.

Machine learning algorithms are trained following one of two strategies, namely supervised and non-supervised learning. The difference between them is the way the machine reaches the knowledge. The unsupervised algorithm relies on the algorithm to autonomously discover a pattern linking the inputs between them based on their similarities. It works mainly on clustering the inputs as subcategories representing a classification space. On the other hand, the supervised manner requires labeling all the observations for the training. Therefore, the learning is much easier than the unsupervised case, but it depends on the labeling quantity and quality. Facial analysis frameworks are achieved through supervised learning with full annotations regarding the challenging nature of such applications.

Despite the progress made in face recognition in terms of computation speed and performance accuracy, the identification process is not guaranteed all the time, and it is vulnerable when the exterior environment is wild. Thus, the computational abilities to correctly classify a facial image depends on many parameters related to the environment such as lighting and contrast, face position, facial expressions, image quality, and also

those related to the framework itself in terms of the overall recognition framework, including pre-processing steps, extraction of the image feature, and the decision/classification algorithm. These two last issues, i.e., feature extraction and classifier designation, constitute the two critical sub-problems in facial analysis. Many literature approaches have achieved good face recognition performance thanks to the extraction of discriminant features, which can be considered as the critical stage in the overall recognition framework. Indeed, extracting non-discriminating features that may ignore important details and descriptions severely affects recognition performance even if we adopt a sophisticated classifier. Consequently, most facial and image classification research gives more interest to extracting the characteristics, giving birth to various face descriptors. Face recognition systems typically rely on three categories of feature extractors: global (or holistic methods), local descriptors, and deep features.

2.2/ Facial Image Classification Framework

Image-based classification systems share a typical architecture regardless of the targeted application. Despite more than 20 years of extensive research and the large number of papers published in journals and conferences dedicated to this area, we still can not claim that artificial systems can measure to human performance regarding the discussed face-related challenges: difficult light imaging conditions, head pose, aging, facial expressions, and occlusions. However, the literature works share a standard structure and lead to a generic facial analysis system, which involves three basic steps as illustrated in Figure 2.1. This system's input can be an image or a video stream of the face to be recognized, while its output is the person identification label and/or its emotional state.

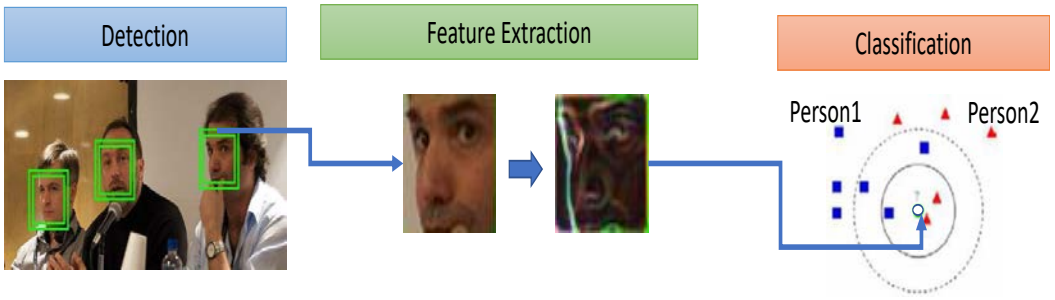

Figure 2.1: Generic architecture of an image-based facial analysis framework

- The first step is **face detection**, which is defined as the process of extracting the bounding box containing only the face. In state-of-the-art databases, the provided images are mostly cropped so that the face detection step can be skipped.
- After detecting the face, the following step works to obtain relevant and discriminant

descriptions from the facial appearance performed during the **feature extraction** phase. This phase also performs dimensionality reduction, which is an essential task in all pattern recognition applications. The feature extraction remains the distinguishing one within the whole framework since it represents the first processing on the raw data and the left stages depend on the quality of this extraction. Moreover, feature extraction is shared with other classification applications of pattern recognition. Therefore, the literature experienced the proposal of many works devoted to enhancing the feature extraction process resulting in an important number of techniques. For face recognition, the feature extraction can be performed globally or locally on the image. The global or holistic methods include filter and wavelet transformation-based ones, which try to compute one feature vector from the whole facial image. On the other hand, the local methods treat pixel by pixel exploring small neighborhoods to detect all the variations.

- The obtained feature vector will be fed to the **face classification** step, which needs a training phase to make a classifier or more capable of recognizing probe images. There are three classifier categories: 1) Similarity-based ones known as the nearest neighbor rule, where the approach relies on grouping the similar patterns into the same class by establishing a distance metric; 2) Probabilistic approach using Bayes decision rule, which is a conditional probability model minimizing the misclassification probability. Naive Bayes classifiers have been especially popular for text classification; 3) Decision boundaries methods such as support vector machine SVM, used for binary classification, underly a given feature space into two zones representing two classes.

A new image classification architecture started emerging after 2010 with Deep Convolutional Neural Networks' outstanding machine learning performance. The deep facial analysis relies on learning convolution weights and the decision ones (fully connected neurons) via backpropagation and loss-based optimization. Unlike handcrafted systems where the feature extraction and the classification are disconnected, where only the classifier is trained to get the maximum from the static extracted features. The deep CNNs learning considers mainly the earlier layers as they have a large learnable parameter set compared to the fully connected layers. Figure 2.2 illustrates the configuration of deep CNN-based image classification.

2.3/ FEATURE EXTRACTION

Feature extraction aims to encode the image space into another that motivates the person-related visual features than the common ones and performs dimensionality re-

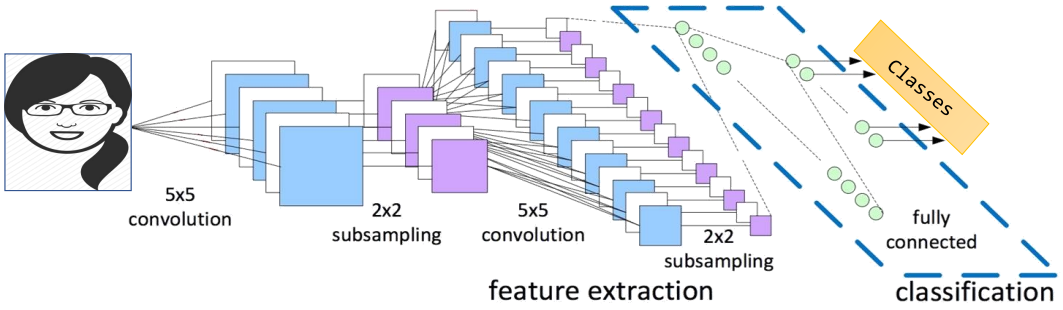

Figure 2.2: CNN-based architecture for image facial analysis

duction on the input space delivering a feature vector. There are three categories of feature computation: holistic, local, and learnable features.

2.3.1/ HOLISTIC DESCRIPTION

Holistic face recognition uses global information of faces to perform face recognition. The global information of faces is essentially represented by a small number of features derived directly from face images' pixel information. This small number of features captures the variance between different individual faces and identifies individuals uniquely. Holistic methods use the entire face as input. In these methods, each face image is represented as a high dimensional signal vector by listing all the pixels as one vector and not a matrix. The face recognition literature was enriched by the proposal of several global methods, including approaches based on different transforms such as wavelet sub-bands [25], Gabor filters [26], optimal matrix factorization [27] and steerable pyramid transform [28], independent component analysis (ICA) method [29], Zernike moments method [30], global Gabor-Zernike feature descriptor [31], principal components analysis (PCA) method [32], linear discriminant analysis (LDA) based Fisherface method [33], etc. It is stated by the literature researches that the global techniques are fast in extracting and computing the similarities between the features. However, they demonstrated many weaknesses, which can be found in [34]. Figure 2.3 shows an example of the extracted global features from a face image by 2-levels wavelets

2.3.2/ LOCAL DESCRIPTION

2.3.2.1/ LOCAL BINARY PATTERNS LBP

The first local descriptor was introduced in 2002 by [35] and referred to as Local Binary Patterns. After proving a remarkable feature extraction quality and efficiency in texture classification, the LBP was adopted also in other classification problems and mainly in face recognition thanks to its ability to effectively compute the feature patch of facial im-

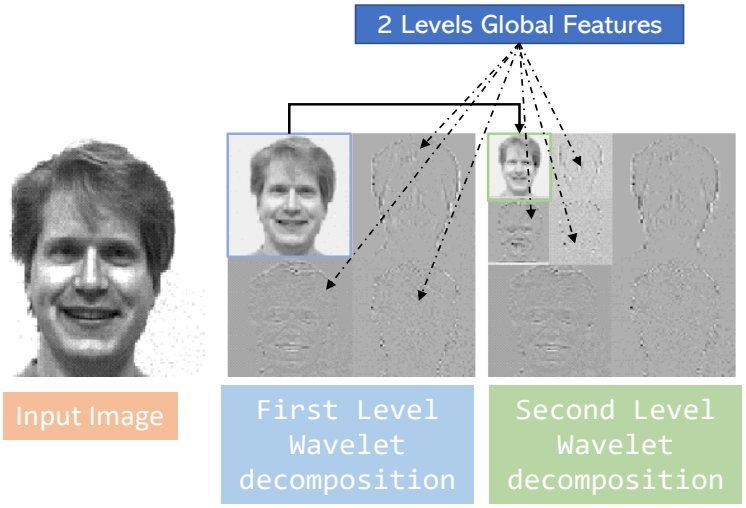

Figure 2.3: Example of extracting global features using wavelet transformations

ages. As described in Figure 2.4, the LBP descriptor works on encoding the structure around each pixel of the image and generates its label, which is performed by thresholding the value of a concerned pixel with those of its 3×3 sub-block neighbors, where the strictly negative are encoded with 0 and the positive ones with 1. The eight binary values are concatenated starting from the top-left position in clockwise direction to form the 8-bits code and its corresponding decimal representation is pixel label. This process can be formulated as given in Eq 2.1 :

$$LBP\,(I_c) = \sum_{p=0}^{P-1} \Xi(I_p, I_c) \times 2^p \tag{2.1}$$

where P refers to the number of neighbors which is 8 adopting a 3×3 block size while I_p refers to values of the neighbor pixels with $p = [0, 1, ...P - 1]$. Thus, the basic LBP allows to reach a discriminative power of 256 (2^8) possible different patterns. $\Xi(x, y)$ is Heaviside step function (cf Eq. 5.8).

$$\Xi(x, y) = \begin{cases} 1 & , x \geq y \\ 0 & , x < y \end{cases} \tag{2.2}$$

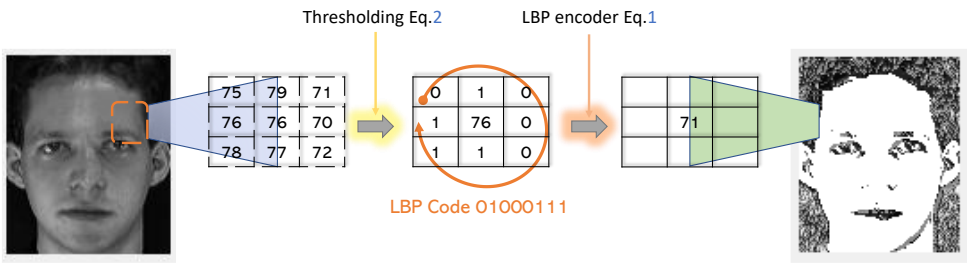

Figure 2.4: The LBP pixel transformation process.

If the coordinates of the center pixel are (x_c, y_c) then the coordinates of his P neighbors (x_p, y_p) on the edge of the circle with radius R can be calculated with the sinus and cosines:

$$x_p = x_c + R.cos(\frac{2\pi p}{P}) \tag{2.3}$$

$$y_p = y_c + R.sin(\frac{2\pi p}{P}) \tag{2.4}$$

In Figure 2.5 the neighborhood was expanded to capture dominant feature with large-scale structures. The neighborhood can be denoted by a pair (P, R) where P is the sampling points on a circle of radius of R. Therefore, there are 2^P different output values.

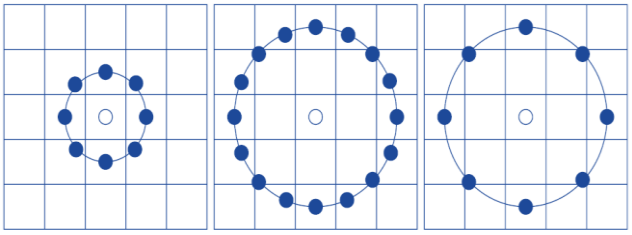

Figure 2.5: The circular (8,1),(16,2) and (8,2) neighborhoods.

2.3.2.2/ LBP-LIKE DESCRIPTORS

The LBP descriptor proved an outstanding performance in many applications, leading to an important expansion of LBP-like methods and inspiring the researchers to develop new ones.. Indeed, after Ojala's proposal [35] and thanks to the demonstrated flexibility and efficiency, the overall LBP-like concept has proven very prominent, and a great variety of LBP variants have been proposed in the literature to overcome the weaknesses and enhance the strengths of the original LBP method covering discriminative power, robustness, and applicability of LBP. Tan and Triggs introduced a three-level operator referred

to as local ternary patterns (LTP), which is a generalization of LBP proving more effective than the original operator in face recognition application [36]. The concept behind is based on three values (1, 0, or -1) encoding the difference between the center pixel and its neighbor pixels considering a Δ threshold. Motivated by the basic LBP operator, Zhang et al. [37] proposed a higher-order local derivative pattern (LDP) for face recognition, which provides more detailed descriptions of facial images. However, the LDP presents more noise sensitivity than LBP. CS-LBP [38] was developed for image matching based on SIFT framework. CS-LBP could extract the feature of a given region with less dimensionality than SIFT and LBP. Based on the CS-LBP descriptor, Fu and Wei [39] proposed centralized binary pattern (CBP) for facial expression recognition. CBP considers the information in the center pixel by comparing its value to the average of all eight pixels in the neighborhood and giving it the largest weight. However, it is not easy to manually set a suitable CS-LBP threshold in all methods using thresholding like LTP. Also for face recognition, Rivera et al. [40] proposed local directional number pattern (LDN), which relies on computing edge responses and taking the top directional numbers. It encodes intensity variations and structural information of texture using major direction indices (directional numbers) to distinguish different gray level intensity transitions among the similar structural patterns. Recently, Wankou et al. [34], inspired by Weber's Law, proposed two adaptive local feature descriptors, referred to as Adaptive Local Ternary Pattern (ALTP) and center-symmetric Adaptive Local Ternary Pattern (CS-ALTP) for face recognition. QBP proposed by Zeng et al. [41] is another recent variant of LBP, called quad binary pattern. QBP, where the interest is to improve the mean-shift tracking accuracy with more robustness and lower computational complexity, is computed over 2×2 pixel blocks using the mean of the four pixel values of each block as its threshold. Pattern recognition field experiences ongoing and extensive researches on local descriptors based applications, as can be found on new published works [42; 43; 44]. Indeed, there is always a need to propose a robust local descriptor with high discriminative power and numerous powerful LBP variants continue to be developed in the literature. Notable recent methods include statistical binary patterns (SBP2, SBP3 and SBP4) [45], local quadruple pattern (LQPAT) [46], local neighborhood difference pattern (LNDP) [47], local directional ternary pattern (LDTP) [44], etc. More recently, Issam et al. [44] proposed local directional ternary pattern (LDTP) for texture classification. The LDTP operator consists in encoding both contrast information and directional pattern features in a compact way based on local derivative variations. LDTP conveys valuable information about the nature of textures by capturing local structures using both LTP's [36] and LDP's [37] concepts simultaneously.

The outstanding performance of success of LBP method in many applications related to computer vision, has motivated much new researches to propose enhanced LBP variants. Regarding its flexibility and ease of implementation, the LBP operator can be worked out

to meet the requirements of various applications, which include non traditional texture problems such as dynamic texture and scene classification, medical image analysis, face recognition, etc. The new extensions of LBP try to enhance the following aspects of the basic LBP descriptor:

1. Neighborhood topology and sampling: This extension focuses on defining more prominent neighborhood area and sampling the pixels in an effective way to cover most of the textural transitions.
2. Thresholding and quantization: Thresholding process is one of the key elements of the LBP philosophy. The researchers develop sophisticated thresholding kernel functions to perform the binarization and quantization of gray level transitions to multiple levels based on thresholding settings.
3. Encoding and regrouping: This category relys on spliting the pixels of the neghborhood area into groups of patterns, to be combined, to improve distinctiveness.
4. Combining complementary features: Current trend in local image and video descriptors is to combine multiple complementary LBP-like descriptors, or LBP-like with non-LBP descriptors in the objective of exploring the advantages of different concepts.

More details about the taxonomy of existing LBP variants can be found in [48] where their merits and demerits and their underlying connections were analyzed. A wide variety of LBP-like methods have been proposed in pattern recognition literature, usually developed to extract features to achieve specific applications (see Table 2.1). Unlike global approaches, local descriptors, which overcome the limitations mentioned above, divide the whole face into smaller image patches, from which local features are extracted and combined to build one face descriptor. The authors in [31] reported that local description is non-sensitive to nonessential and irrelevant patterns of facial and textural images . Heisele et al. [49] evaluated global and local descriptors to disclose that the local descriptors outperform global ones. Note that most of LBP variants have been proposed usually for one specific application, which are intended by the researchers and then each variant is tuned to reach it best performance according to the specific application. Indeed, as we will see later on this paper, majority of LBP variants have been proposed for texture analysis and, to a lesser degree, to fulfill the specifications and deal with the challenges of different applications including dynamic texture and scene classification, medical image analysis, etc . However, the performance of most of these LBP extensions have not been yet evaluated on face recognition problem. Hence, the state-of-the-art lacks of an extensive review of handcrafted descriptors for face recognition, which should be done by performing a large scale empirical study on challenging face datasets to investigate the weakness and strengths of LBP-like methods.

Table 2.1: summary of texture descriptors tested.

Category	Complete name	Abbreviation	Application	Year	Ref
Combining with complementary features	Local Extreme Complete Trio Pattern	LECTP	Image retrieval	2014	[50]
Encoding and regrouping	Rotation-invariant features based on directional coding	DC	Texture classification	2018	[51]
	Statistical binary patterns (2)	SBP2	Texture classification	2017	[45]
	Statistical binary patterns (3)	SBP3	Texture classification	2017	[45]
	Statistical binary patterns (4)	SBP4	Texture classification	2017	[45]
	Quad Binary Pattern	QBP	Target tracking	2016	[41]
	Dominant Rotated Local Binary Patterns	DRLBP	Texture classification	2016	[52]
	Adaptive Local Ternary Pattern	ALTP	Face recognition	2016	[34]
	Center-Symmetric Cdaptive LTP	CSALTP	Face recognition	2016	[34]
	Magnitude Maximum Edge Position Octal Pattern	MMEPOP	Image retrieval	2015	[53]
	Sign Maximum Edge Position Octal Pattern	SMEPOP	Image retrieval	2015	[53]
	Adjacent Evaluation Completed LBP (S)	AECLBP-S	Texture classification	2015	[54]
	Adjacent Evaluation Completed LBP (M)	AECLBP-M	Texture classification	2015	[54]
	Adjacent Evaluation Completed LBP (S-MxC)	AECLBP-S-MxC	Texture classification	2015	[54]
	Adjacent Evaluation LTP	AELTP	Texture classification	2015	[54]
	Orthogonal Combination Of Local Ternary Patterns	OC-LTP	Infrared imagery recognition	2014	[55]
	Complete Robust Local Binary Pattern (M)	CRLBP-M	Texture classification	2013	[56]
	Complete Robust Local Binary Pattern (S)	CRLBP-S	Texture classification	2013	[56]
	Complete Robust Local Binary Pattern (S-MxC)	CRLBP-S-MxC	Texture classification	2013	[56]
	Local Gray Code Pattern	LGCP	Face expression analysis	2013	[57]
	Local Maximum Edge Binary Patterns	LMEBP	Image retrieval	2012	[58]
	Center-Symmetric Local Ternary Pattern	CS-LTP	Feature description	2010	[59]
	extended Center-Symmetric Local Ternary Patterns	eCS-LTP	Image retrieval	2011	[60]
	Center-symmetric Local Binary Patterns	CS-LBP	Texture classification	2006	[38]
	Gradient Texture Unit Coding	GTUC	Texture classification	2004	[61]
Neighborhood topology and sampling	Attractive-and-Repulsive Center-Symmetric Local Binary Patterns	ARCS-LBP	Texture classification	2019	[62]
	Local Optimal Oriented Pattern	LOOP	Species recognition	2018	[63]
	Repulsive-and-attractive local binary gradient contours	RALBGC	Texture classification	2018	[43]
	Local concave-and-convex micro structure patterns	LCCMSP	Texture classification	2018	[42]
	Local directional ternary pattern	LDTP	Texture classification	2018	[44]
	Local neighborhood difference pattern	LNDP	Texture image retrieval	2017	[47]
	Local quadruple pattern	LQPAT	Recognition and retrieval	2017	[46]
	Extended Local Graph Structure	ELGS	Texture classification	2016	[64]
	Diagonal Direction Binary Pattern	DDBP	Face recognition	2016	[65]
	Linear Directional Binary Pattern	LDBP	Face recognition	2016	[65]
	Directional Local Binary Patterns	dLBP$_\alpha$	Texture classification	2015	[66]
	Difference Symmetric Local Graph Structure	DSLGS	Finger vein recognition	2015	[67]
	eXtended Center-Symmetric Local Binary Pattern	XCS-LBP	Texture classification	2015	[68]
	Multi-Orientation Weighted Symmetric Local Graph Structure	MOW-SLGS	Finger vein recognition	2015	[69]
	Local Binary Patterns by neighborhoods	nLBPd	Texture classification	2015	[66]
	Symmetric Local Graph Structure	SLGS	Finger vein recognition	2015	[70]
	Local Extreme Sign Trio Pattern	LESTP	Image retrieval	2014	[50]
	Local Directional Number Pattern	LDN	Face expression analysis	2013	[40]
	Improved Binary Gradient Contours (1)	IBGC1	Texture classification	2012	[71]
	Local Graph Structure	LGS	Face recognition	2012	[72]
	Binary Gradient Contours (1)	BGC-1	Texture classification	2011	[73]
	Binary Gradient Contours (2)	BGC-2	Texture classification	2011	[73]
	Binary Gradient Contours (3)	BGC-3	Texture classification	2011	[73]
	Improved Center-Symmetric Texture Spectrum	ICSTS	Texture classification	2011	[74]
	Directional Binary Code	DBC	Face recognition	2010	[75]
	Local Derivative Pattern	LDP	Face recognition	2010	[37]
	Center-Symmetric Texture Spectrum	CSTS	Texture classification	2003	[74]
	Simplified Texture Spectrum	STS	Texture classification	2003	[76]
	Simplified Texture unit *	STU*	Texture classification	2003	[77]
	Simplified Texture Unit +	STU+	Texture classification	2003	[77]
Thresholding and quantization	Local Quantization Code Histogram	LQCH	Texture classification	2016	[78]
	Complete Robust Local Binary Pattern (C)	CRLBP-C	Texture classification	2013	[56]
	Robust Local Binary Pattern	RLBP	Texture classification	2013	[79]
	Local Binary Count	LBC	Texture classification	2012	[80]
	Completed Local Binary Count	CLBC	Texture classification	2012	[80]
	Improved Local Ternary Pattern	ILTP	Medical image analysis	2010	[81]
	Centralized Local Binary Pattern	CLBP	Face expression analysis	2008	[39]
	Local Ternary Pattern	LTP	Face recognition	2007	[36]
	Improved Local Binary Patterns	ILBP	Face detection	2004	[82]
Other Methods	Local Phase Quantization	LPQ	Texture classification	2008	[83]
	Texture spectrum (Δ)	TS(Δ)	Texture classification	1992	[84]
	Texture spectrum (0)	TS(0)	Texture classification	1990	[85]

2.3.3/ LEARNABLE FEATURES

During the last years, computer vision field is experiencing the birth of deep learning and deep features based methods, which improve the performance of recognition and classification process in many applications including, among others, face recognition, medical image analysis, content-based image retrieval.

2.3.3.1/ PRINCIPAL COMPONENT ANALYSIS NETWORK (PCANET)

The authors in [86] proposed a new and simple 2 stages deep learning architecture for various image classification tasks. The concept is based on the well known statistical procedure PCA, which is used to learn the data-adapting convolution filter banks from the training images at the first stage. The same learning operation is repeated exploiting the output of the first stage. The obtained output image patches are binarized using Heaviside step function, then are divided into blocks. Local histogram is computed in each block and final feature is the concatenation of all computed local histograms. The feature extraction of training samples utilizes a pre-trained model using MultiPIE database. The classification phase is performed using nearest neighbor (NN) classifier with chi-squared or cosine distance. The computation time of this deep learning method and the performance dependence to the pre-training database remain the major drawbacks of this architecture. The architecture of this deep feature method is illustrated in Figure 2.6.

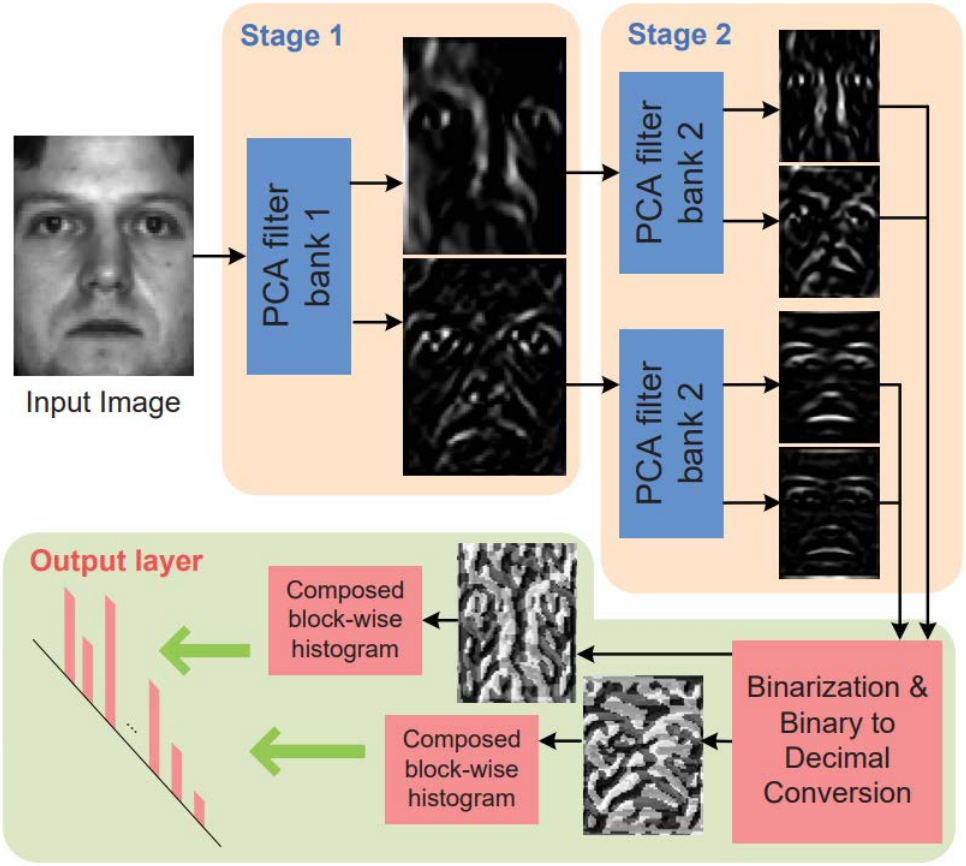

Figure 2.6: The PCANet2 learnable feature extraction method proposed in [86]

2.3.3.2/ COMPACT BINARY FACE DESCRIPTOR (CBFD)

Most of the state-of-the-art existing local binary descriptors are hand crafted, which required and continue requiring an important research effort to develop new variants by specifying sophisticated neighborhood topologies and thresholding functions. The authors in [87] developed a non hand-crafted feature learning method referred to as Compact Binary Face Descriptor (CBFD). As can be seen in Figure 2.7, the idea behind is to extract Pixel Difference Vectors (PDVs) from local blocks by calculating the threshold between each pixel and its neighbors. The obtained PDVs are then used and projected to form a feature mapping and build dictionaries of the training set. The compact binary codes are retrieved after removing the information redundancy in projected PDVs. Lastly, the compact binary codes are clustered into one histogram feature vector as the final representation for each face image. The feature extraction computing time of this learning descriptor is too much higher than the one required for extracting hand crafted descriptors. Moreover, the stability of the performance depends on the process of building training dictionaries, which may be affected by the random permutations in the train and test sets.

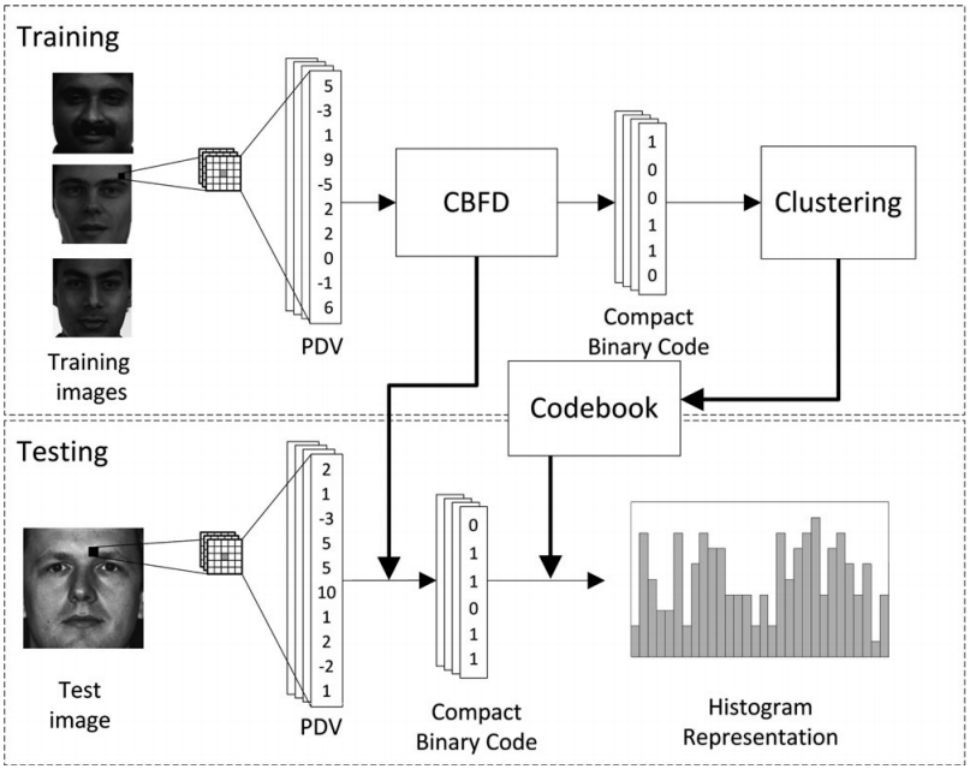

Figure 2.7: CBFD workflow for face feature extraction as proposed in [87]

2.3.3.3/ Deep Convolutional Neural Networks

Deep Convolutional Neural Networks also are adopted to extract deep features from an input image. The CNN are generally used end-to-end to perform a classification or regression task thanks to their fully connected layers, however we can extract deep features if remove the fully connected layers and keep only the convolutional ones. The output is a set of feature maps with low resolution (16×16 dimension). A CNN typically consists of convolutional layers, pooling layers, and fully connected layers. Convolutional layers are the core building blocks of a CNN.

We considered 10 deep networks, which are briefly introduced in the following, based on a recent survey [88]:

- AlexNet: Referring to its author Alex Krizhevsky, AlexNet was proposed in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2012). AlexNet is a deeper configuration of a LeNet5 network. Therefore, the high performance at this competition comes at the cost of a high computation that was possible using only graphic card units. It consists of five convolutional layers, two fully connected hidden layers, and one fully connected output layer.

- VGG: VGG network architectures were introduced by Simonyan and Zisserman in 2014. VGG stands for the Visual Geometry Group of Oxford University. Compared to LeNet and AlexNet, VGG networks are conceptually simple employing only stacked 3×3 convolutional layers combined with a max pooling layer to reduce the volume size, leading to two fully connected layers of 4096 nodes each, followed by a softmax classifier. VGG19 has three more convolutional layers than VGG16.

- ResNet: Residual learning networks were also proposed for the ILSVRC competition in 2015, introducing the Skip Connection concept to CNNs, which are known as recurrent networks. Typical ResNet models are implemented with double- or triple-layer skips that contain nonlinearities (ReLU) and batch normalization between them. The skip connection technique allows the training of 152 layers or more with fewer computations then AlexNet and VGG networks. In this study, we considered ResNet18, ResNet50, and ResNet101.

- DenseNet: Densely connected convolutional networks were inspired by the ResNet topology. They incorporate dense residual blocks composed of batch normalization, ReLU activation, and a 3×3 convolution. The ResNet models use the sum function as a skip connection, whereas DenseNet integrates the concatenation. Therefore, each input layer receives all outputs of the earlier versions. The concatenation process generates an output with a large number of channels, which makes DenseNet models computationally heavy.

- Inception: Google proposed its own deep learning inspired by LeNet, referred to as Inception, stacking more convolutional layers deeper to achieve a better performance, which comes at the cost of heavy computations and a complex design. The philosophy of inception relies on concatenating the responses of different convolution filters at the same layer, forming the input of the next layer. Moreover, they used a 1×1 convolution filter as a feature reduction technique before jumping to the next layer. Google introduced four versions of the Inception architecture, Inception.v1 known as GoogLeNet with 27 layers, Inception.v2, Inception.v3, and Inception.v4 tackling batch normalization, factorization, and grid size control problems, respectively. Google proposed two versions of a residual network inspired by the performance of ResNet, known as InceptionResNet.v1 and InceptionResNet.v2 based on creating the skip connections on the previous Inception models. Inception-ResNet.v1 and Inception-ResNet.v2 networks have the same computational cost of Inception.v3 and Inception.v4, respectively.

To employ deep learning architectures for deep feature extraction in solving the facial analysis problems, we follow the basic procedure shown in Figure 2.11. Initially, the model was trained end-to-end on a big dataset, mainly the LFW database. Afterwards, the model is expected to achieve a good training performance using the validation set. We then proceed to the transfer learning technique to extract the features of the subject database that belongs to the same application as the database used for the initial training (same classes). Once the features are obtained, we train the classifier and evaluate the performance of each deep feature.

2.4/ CLASSIFICATION

The supervised classification objective is mainly to define rules making it possible to classify objects in classes from qualitative or quantitative variables characterizing these objects. The methods often extend to quantitative Y variables (regression). In the following, we introduce the widely used algorithms for face analysis.

2.4.1/ NEAREST NEIGHBOR

The Nearest Neighbor algorithm is the most simple classifier in machine learning literature. It relies on finding the most similar observations from the training data to the probe input. To do so, it computes the distance between the probe input and all the training samples. After, it assigns to the probe image the label of the training sample that has the lowest distance. The Nearest Neighbor classifier has a more generic algorithm known as

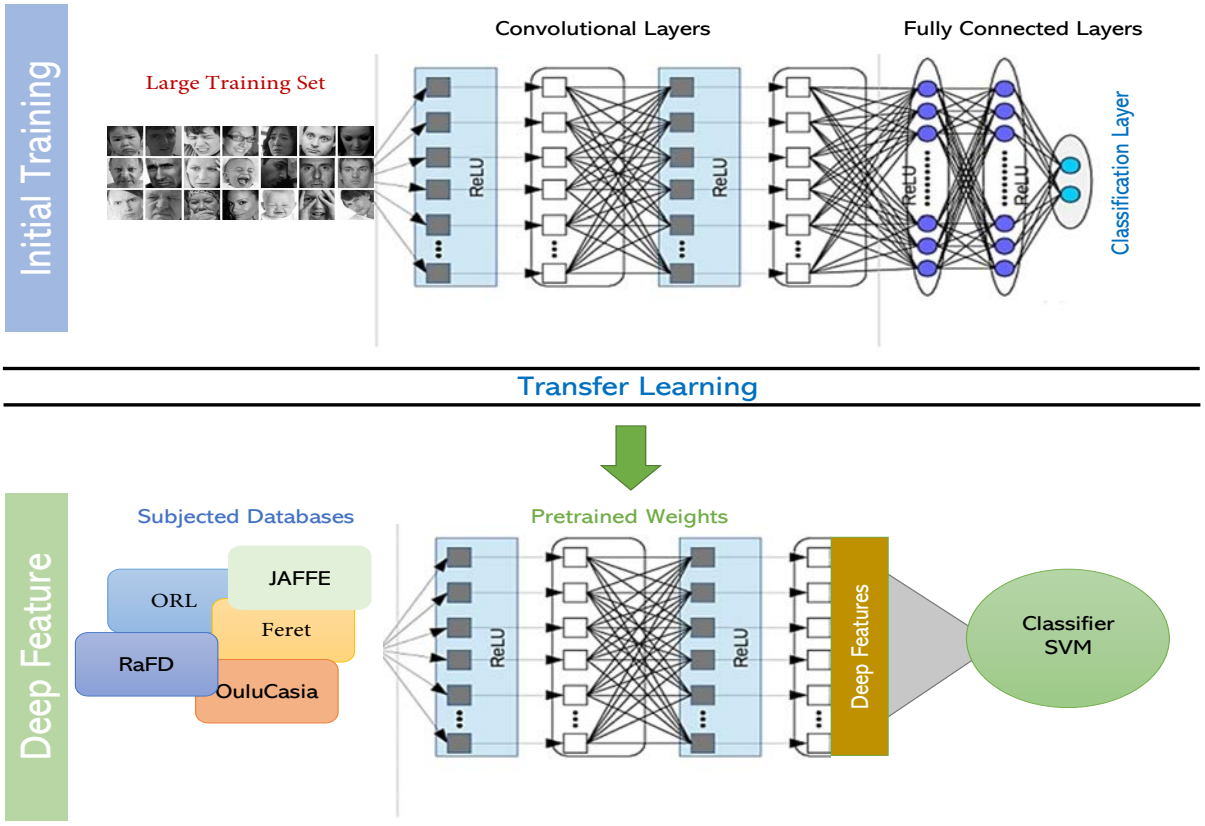

Figure 2.8: Deep-feature based transfer learning approach for face and facial expression recognition.

"k" Nearest Neighbor, where the algorithm this time looks for "k" similar training samples and then selects their majority class among. The Nearest Neighbor is a parameter-free classifier and has no kernels that would require more computation. Moreover, it is a lazy classifier since it does not perform any training and repeats the matching procedure for each probe image. However, this fact makes the Nearest Neighbor flexible to dynamic training sets as the new samples are taken into account when computing the distances. Table 2.2 lists the literature metrics used to compute the pairwise distance.

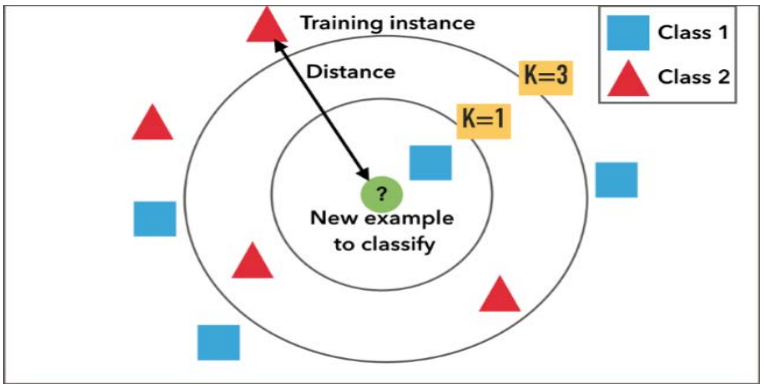

Figure 2.9: Example of Nearest Neighbor matching with $K = 1$ and $K = 3$

Table 2.2: Examples of Distance Metrics for Classification

Metric	Formula
Euclidean distance	$d(r,s) = (x_r - x_s)(x_r - x_s)'$
Standardized Euclidean distance	$d(r,s) = (x_r - x_s)trace(\Sigma)^{-1}(x_r - x_s)'$
Mahalanobis distance	$d(r,s) = (x_r - x_s)\Sigma^{-1}(x_r - x_s)'$
City Block metric	$d(r,s) = \sum_{j=1}^{n} \left\|x_{rj} - x_{sj}\right\|$
Minkowski metric	$d(r,s) = \sqrt[p]{\left(\sum_{j=1}^{n} \left\|x_{rj} - x_{sj}\right\|^{p}\right)}$
Cosine distance	$d(r,s) = \left(1 - \frac{x_r x_s'}{\sqrt{x_r' x_r}\sqrt{x_s' x_s)}}\right)$
Correlation distance	$d(r,s) = 1 - \frac{(x_r - \bar{x}_r)(x_s - \bar{x}_s)'}{\sqrt{(x_r - \bar{x}_r)(x_r - \bar{x}_r)'}\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}}$
Hamming distance	$d(r,s) = \frac{\#(x_{rj} \neq x_{sj})}{n}$
Jaccard distance	$d(r,s) = \frac{\#[(x_{rj} \neq x_{sj}) \wedge ((x_{rj} \neq 0) \vee (x_{sj} \neq 0))]}{\#[(x_{rj} \neq 0) \vee (x_{sj} \neq 0)]}$

x and x' denote a column vector and its transpose respectively.
x_r and x_s indicate the r^{th} and s^{th} samples in the data set, respectively.
x_{rj} indicates the j^{th} feature of the r^{th} sample in the data set.
$\bar{x}_r$ indicates the mean of all features in the r^{th} sample in the data set.
Σ is the sample covariance matrix.
The symbol # denotes counts; the number of instances satisfying the associated property.

2.4.2/ Support Vector Machines

The Support Vector Machine (SVM) is a discriminative classifier formally based on defining separating hyperplanes. In other words, given labeled training data (supervised learning), the algorithm outputs optimal hyperplanes dividing the two-dimensional space generating an SVM model based on the representation of the training data as points in space, mapped so that the examples of the separate categories or classes are divided by a dividing plane that maximizes the margin between different classes which allows to classify the query points and predict their classes.

Figure 3.3 illustrates the concept behind Support Vector Machines classification. On the left side, we have the input feature vectors calculated earlier using the handcrafted descriptors mapped by a complex curve. The classifier will rearrange these input points to reach a linear separation using a set of mathematical functions called kernels.

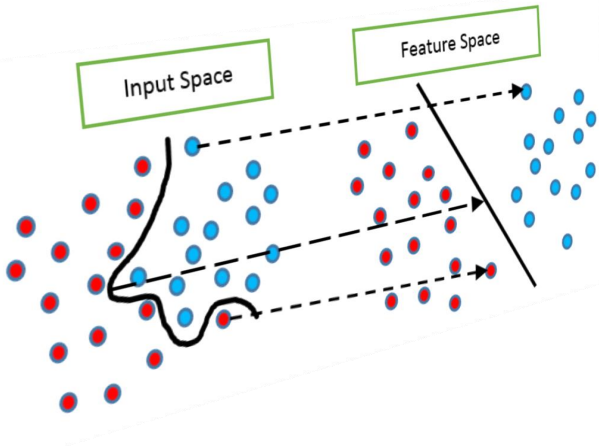

Figure 2.10: Feature vectors representation both in the input feature spaces.

We adopted lib-svm toolbox [89][1], which became a very popular toolbox for SVM classification regarding the complexity of classification in terms of the number of classes. This toolbox supports the four kernels of the binary SVM classifier: linear, polynomial, radial basis function (RBF) and sigmoid kernels as defined in Eq. 2.5.

$$\begin{cases} Linear & : u'v \\ Polynomial & : (\gamma u'v + C)^d \\ RBF & : e^{(-\gamma * |u-v|^2)} \\ Sigmoid & : \tanh(\gamma u'v + C) \end{cases} \tag{2.5}$$

2.4.3/ Neural Networks

Neural Networks are extensively used for classification purposes as well as regression-based tasks. Like SVM, a neural network-based classifier requires to be trained to generate a model that predicts the labels of probe images. The Neural classifier has the shape of an autoencoder organized in successive layers: an input layer, an output layer, and between the two one or more intermediate layers, also called hidden layers. There is no connection between neurons in the same layer, but every neuron in one layer is connected to all neurons in the next layer. The input layer has the same dimension as the feature vector extracted from the image, while the last layer dedicates one neuron to each class label. A hidden layer helps the autoencoder classifier model nonlinear relationships between the inputs and their labels. A single hidden layer is sufficient in theory, but having more hidden layers makes it easier to model a non-continuous discrimination function. The autoencoder classifier training relies on finding the configuration of the connection weights between neurons to associate with each input feature vector its corresponding

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvm/

label. Therefore, increasing the hidden layers extends the weight configurations to be checked.

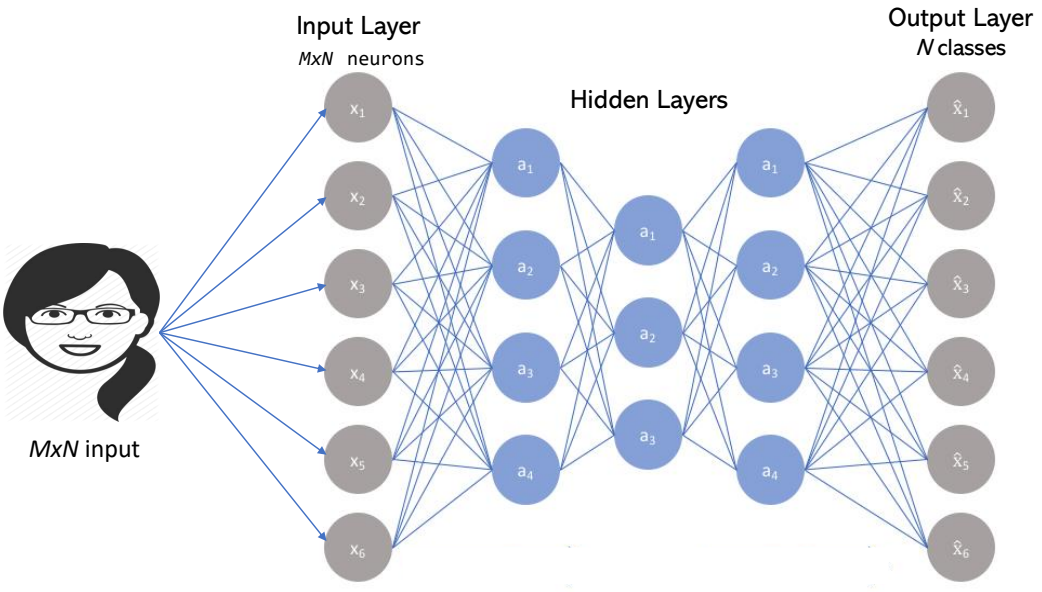

Figure 2.11: Neural Networks-based classification architecture.

2.5/ FACE RECOGNITION STATE-OF-THE-ART

Fathima et al. [90] proposed a combination of Gabor wavelet and LDA for face recognition (HGWLDA). The HGWLDA convolves the grayscale face image with a set of multi orientations and scales Gabor filters. After that, the authors applied 2D-LDA to reduce the filter maps space and keep only the discriminant features. The classification is performed using the nearest neighbor with Euclidean distance. This work lacks a proper comparison with the state-of-the-art to judge the performance of the proposed HGWLDA technique. Advanced correlation filters and Walsh LBP (WLBP) face-based descriptor was introduced by Juefei et al. [91]. This technique works on generating 3D rotations from 2D samples using the 3D generic elastic model. Then, computing and concatenating the WLBP features from the 3D views and classify them through Class-Dependence Feature Analysis (KCFA). The recorded results on the LFW benchmark showed that this method managed to be competitive to deep-learning models. [92] proposed an effective feature extraction technique referred to as a Multi-sub-region-based correlation filter bank (MS-CFB) for robust face recognition. MS-CFB extracts independent features from non-overlapping blocks. After that, these features are concatenated to form the final face descriptor. The correlation of a probe image descriptor with the training ones is expected to give the correct class's correlation peak, which the authors used as the classification rule. The evaluation was performed on three state-of-the-art benchmarks proving the su-

periority of the MS-CFB in front of classical methods but no comparison against recent methods. [93] have developed a face recognition model based on combining SIFT features and Fisher vectors. They adopted PCA discriminative dimensionality reduction to overcome the computation resulted from Fisher vectors. The reduced space is further projected into a liner space to conclude the identity of input facial images. [94] proposed a multi-space face recognition system referred to as the Multi-Modal Deep Face Recognition (MM-DFR) framework. The modalities include the original holistic face image, uniformly sampled image patches, and extracted frontal face views from a computed 3D face model. A CNN extracts the deep features from each modality and concatenates them to form the feature vector fed to an auto-encoder classifier to predict the image label. The evaluation was performed on LFW. However, the results were not outperforming the state-of-the-art regarding the modalities considered and the computational complexity. [95] introduced a pose-invariant face recognition framework based on PCA for feature extraction and adaptive neuro-fuzzy inference system (ANFIS) classifier. However, this work is not complete in terms of evaluation that was limited to the ORL database. [96] develop a fast face recognition system based on DCT and PCA techniques for feature extraction and a Genetic Algorithm (GA) to compute discriminant features from DCT-PCA ones and remove the irrelevant bins. The minimum Euclidian distance (ED) is used to calculate the similarities and then classify the images. [97] proposed 2D and 3D-based modalities for face recognition through a hybrid transform to correct the pose of a 3D face using its texture and achieving efficiency and robustness to facial expressions. The SIFT descriptor is used to compute the features from corrected 3D views and fed to the iterative closest point (ICP) algorithm for the decision. This system is less sensitive and robust to facial expressions, which achieved a 98.6% verification rate and 96.1% identification rate on the complete FRGC v2 database. [98] proposed a new LBP-like face descriptor referred to as Orthogonal difference-local binary pattern (OD-LBP). Their scheme relies on computing three orthogonal transformations and then computing the differences between them, leading to 3 feature maps. Each feature map is divided into nine non-overlapping sub-regions to perform local histogram count. Then the local histogram vectors are concatenated and fed to the SVM algorithm to perform training and evaluation. The authors adopted five benchmarks but with a low number of individuals (less than 50), which is not challenging compared to other large datasets. [99] proposed a computationally efficient hybrid face recognition system that relies on both local and holistic description. The Local Gabor Binary Pattern Histogram Sequence (LGBPHS) method is employed to realize the feature extraction on the whole facial image. After that, the PCA technique is used for dimensionality reduction. The experimental evaluations on Extended Yale Face Database B demonstrated an improved recognition rate as compared to the basic PCA and Gabor wavelet techniques under illumination variations. [100] proposed a novel hybrid technique for face representation and recognition, which exploits both local and subspace features.

To extract the local features, the whole image is divided into sub-regions, while the global features are extracted directly from the whole image. After that, PCA and Fisher linear discriminant (FLD) techniques are introduced on the fused feature vector to reduce the dimensionality. The CMU-PIE, FERET, and AR face databases are used for the evaluation. [101] developed a new face recognition method based on SIFT features and PCA for space reduction and Nearest Neighbor as the classifier. The framework is referred to as SPCA-KNN and was evaluated on a dataset of 100 subjects with 1000 images for training and 500 for testing. However, this work lacks a comprehensive comparison with the state-of-the-art, and the evaluation should be on more benchmarks.

2.6/ FACIAL EXPRESSION RECOGNITION STATE-OF-THE-ART

The computer vision community has conducted many studies devoted to facial expression recognition (FER) by applying machine learning techniques. In this section, we briefly present some state-of-the-art FER frameworks to highlight some of the proposed architectures that rely on either handcrafted descriptors or deep-learning methods. Shan et al. [102] proposed an approach, referred to as Boosted-LBP, based on combining a basic LBP descriptor with the Adaboost algorithm to enhance the classification performance. They conducted experiments on CK+, MMI, and JAFFE databases, and found that the Boosted-LBP outperforms the basic LBP combined with a multi-class SVM classifier. Moreover, they reported that local methods (LBP) perform better than global methods (Gabor filters). Zhang et al. [103] proposed a novel facial expression recognition method using a local binary pattern (LBP) and local phase quantization (LPQ) based on a Gabor face image. First, Gabor wavelets are applied to capture the prominent visual attributes, which are separable and robust to illumination changes, by extracting multi-scale and multi-direction spatial frequency features from the face image. Then, the LBP and LPQ features based on the Gabor wavelet transform are fused for face representation. Considering that the dimensions of a fused feature are too large, the PCA-LDA algorithm is used to extract compressed features. Finally, the method is tested and verified using multi-class SVM classifiers. Lekdioui et al. [104] proposed an automatic FER framework based on a local appearance approach, extracting the features from seven regions of interest (ROIs) covering the left eyebrow, right eyebrow, left eye, right eye, eyebrows, nose, and mouth. They evaluated the LBP, LTP, and CLBP texture descriptors and their combination with the HOG operator cascading with a linear SVM classifier. They found that the concatenation of LTP and HOG leads to the best FER performance on three datasets (CK, FEED, and KDEF). Their framework strengths rely on extracting the appearance features from seven sub-images defined from landmarks carrying information about the facial expression class, in addition to combining the LBP-Like descriptor with the HOG operator.

However, this architecture presents certain drawbacks that we can point out. The seven extracted sub-images have different sizes and orientations, but their computed features have the same length. We found that the nose region of interest is vertically oriented compared to the eyebrow regions, which are horizontal, and the eye regions, which are almost square. Therefore, different amounts of information on different locations are represented over feature vectors of the same length. Furthermore, this study did not cover an important number of handcrafted methods, and no deep-learning method was evaluated. The method proposed by Makhmudkhujaev et al. [105] uses a new handcrafted LBP descriptor referred to as local prominent directional pattern (LPDP) for FER application. It is also an appearance-based approach exploring the benefits of extracting features from three patches: edge, curved edge, and corner-like texture maps. Their study focuses only on the handcrafted descriptor LPDP and its scheme to extract textural features. The authors also used a thresholding parameter to discriminate significant features from insignificant patterns in featureless/smooth regions of a face. Afterwards, a feature selection method is applied to reduce the dimensionality of the final feature vector because they use the spatial division on the input image. However, this system takes as input the entire face image, which makes it inconvenient for person-independent FER applications. In addition, no shape descriptor has been adopted in the overall framework, relying only on LPDP extracted features. Minchul et al. [106] used a convolutional neural network model to achieve facial expression recognition. They adopted and aligned cropped faces from FER-2013, SFEW2.0, CK+, KDEF, and Jaffe with respect to the landmark position of the eyes. The training data were augmented 10 times by flipping them. Five types of data input (raw, histogram equalization, isotropic smoothing, diffusion-based normalization, and difference of Gaussian) were tested. They then selected the one that showed the highest accuracy as a target structure for fine-parameter tuning. Finally, the CNN network with histogram equalization images was chosen as the baseline CNN model for further research. Yu et al. [107] proposed a method that contains a face detection module based on an ensemble of three state-of-the-art face detectors, JDA, DCNN, and MoT. Subsequently, a classification module composed of an ensemble of deep convolutional neural networks (CNNs) was adopted based on averaging the output responses. Each CNN model is initialized randomly and pretrained on the Facial Expression Recognition (FER) Challenge 2013 database. The pretrained models were then fine-tuned on the training set of SFEW 2.0. To combine multiple CNN models, they presented two schemes for learning the ensemble weights of the network responses: minimizing the log-likelihood loss and minimizing the hinge losses. According to the results reported in their study, the hinge loss performs slightly better than the log-like and single CNN models on the validation and test sets of the FER2013 and SEFW databases. Therefore, their framework is computationally heavy, and the outcomes are not very promising. Jung et al. [108] proposed a new CNN framework based on combining the temporal appearance and temporal ge-

ometry extracted from two CNN models. The faces in the input image sequences are detected, cropped, and rescaled to a pixel resolution of 64×64, and 49 landmark points are then extracted using the IntraFace algorithm. Finally, these two models are combined using an element-wise sum of the outputs of the last fully connected layers from the two temporal CNN models. Through several experiments conducted on the CK+, MMI, and Oulu-CASIA databases as well as numerous data from various data augmentation techniques, the framework built showed that the two models cooperate with each other. However, the joint model did not improve the recognition of all of the facial expressions and achieved the same performance as the temporal appearance and temporal geometry models of the Disgusted, Fear, Happy, and Surprised classes. In addition, the temporal appearance CNN model outperformed the geometry model on all tested databases. Most of the previous methods have considered the entire facial region as the input information, and have paid less attention to the sub-regions of human faces, which may lead to a large difference between the extracted and expected representations. Indeed, when the extracted information obtained from the entire face image is irrelevant, the final recognition result will be affected.

2.7/ CONCLUSION

This chapter was dedicated to present a comprehensive background on image-based facial analysis tasks. We presented the generic configuration of the classification framework, then presenting an overview of the different feature extraction and classification techniques. We focused on local-based face description since it is the topic of this thesis. We highlighted in-depth the concept of the Local Binary Patterns method and its flexibility to develop a wide variety of descriptors intended for different applications. Moreover, we discussed the state-of-the-art of face facial expression recognition underlining the limitations of existing methods.

3

LOCAL DESCRIPTION-BASED FACE RECOGNITION

3.1/ INTRODUCTION

This chapter is dedicated to introducing our contribution related to developing a new local descriptor while keeping the effectiveness and simplicity of the traditional LBP and addressing its weakness. We propose a conceptually simple, high-quality, and yet robust framework of LBP, referred to as Mixed Neighborhood Topology Cross Decoded Patterns (MNTCDP), for face recognition. The proposed MNTCDP descriptor is a ten bits code assigned to each sub-region of size 5×5, which aims at achieving both simplicity and efficiency at the same time. It computes and describes the relationship between the referenced pixel and its neighbors on a 5×5 pixel block by encoding gray-level difference based on two-level radius $(R = 1 \ \& \ R = 2)$ and multi-direction angles: $0°$, $45°$, $90°$ and $135°$. The idea behind is to combine the radius with the angle to extract a more detailed and discriminating description. To make the thresholding process more accurate, each pixel is compared to the average gray level of its 3×3 neighbor pixels within its 5×5 neighborhood encompassing. Unlike most of the existing hand-crafted descriptors, which encode the pixels considering simple neighborhood topology and pattern encoding, the MNTCDP descriptor proposes an advanced encoding way exploiting multi-radial and multi-orientation information simultaneously. This particularity gives the ability to the proposed descriptor to extract more relevant information than the existing descriptors, as shown later. The extracted feature vector is then fed to the simplest K-Nearest Neighbor classifier configured with City Block distance measure for classification. MNTCDP has the following outstanding advantages: As shown further, it allows considerably enhancing the discriminative power of LBP variants and their robustness to small variations (due to image noise) and has low computational complexity. At the feature extraction stage, there is no pre-learning process and no additional parameters to be learned. The main contributions can be summarized and briefed in the following points:

- We propose a novel LBP-like descriptor referred to as Mixed Neighborhood Topology Cross Decoded Patterns (MNTCDP), which combines two topological dimensions to describe a given pixel keeping a low computational complexity and simple conception. MNTCDP extracts the local features from 5×5 neighboring pixels that fulfill the lack of information required for a local descriptor.
- The performance of the proposed MNTCDP operator and its stability are evaluated on four benchmark databases. To make this investigation more meaningful, we record the average classification rate over ten random splits and different training images.
- We conducted a fair and systematic comparison between the proposed MNTCDP descriptor and, on the one hand, a large number of state-of-the-art LBP variants, which have been rarely evaluated in the face recognition field, and on the other hand, several recent state-of-the-art face recognition systems.
- The performance of the proposed MNTCDP descriptor is compared to the one achieved by two deep learning methods adopting the same experimental protocols.

3.2/ Mixed Neighborhood Topology Cross Decoded Patterns

The proposed MNTCDP face descriptor relies on two essential aspects: neighborhood topology and pattern encoding, which are the core components of a face image descriptor.

3.2.1/ Neighborhood topology

The essence of MNTCDP descriptor is to perform neighborhood topology and pattern encoding in the most informative directions contained within face image. For face recognition, useful face image information consists of two parts [109]: the configuration of facial components and the shape of each facial component. The shape of facial components is, in fact, rather regular. After geometric normalization of the face image, the central parts of several facial components, i.e., the eyebrows, eyes, nose, and mouth, extend either horizontally or vertically, while their ends converge in approximately diagonal directions ($\pi/4$ and $3\pi/4$). In addition, wrinkles in the forehead lie flat, while those in the cheeks are either raised or inclined.

Based on the above observations, neighborhood topology of MNTCDP is conducted as shown in Figure 3.1. In order to capture more discriminant information and without in-

creasing computational complexity, we adopted 5×5 block in MNTCDP descriptor, which allows combining radii (2) and angles (4), compared to 3×3 block supporting only angle variation. The pixels of such block cover at the same time, the four orientations which are [$0°$, $45°$, $90°$ and $135°$], and two radiuses [$R = 1$, $R = 2$] (see Figure 3.1). These sampled pixels allow describing the variance intra-person according to the changes that may occur in the pixels of the first neighborhood ($R = 1$) and the ones of the second neighborhood ($R = 2$) labeled with $B_{i;i=0,..,7}$.

As can be seen from Figure 3.1, we define two levels of pixels; level A containing eight pixels of the whole 3×3 neighborhood where the distance between them and the central pixel $\mathbf{I}_c$ is $R = 1$ and level B containing eight pixels which are evenly distributed on the periphery of the 5×5 neighborhood around the central pixel $\mathbf{I}_c$ ($R = 2$). Each pixel in level A is sampled with a pixel in level B in the same direction (the vertical and horizontal directions, and the two diagonal directions). This arrangement is based on combining the pixels of the level A along with horizontal and vertical directions or diagonal directions and the pixels of the level B with the other directions from as illustrated in Figure 3.2.

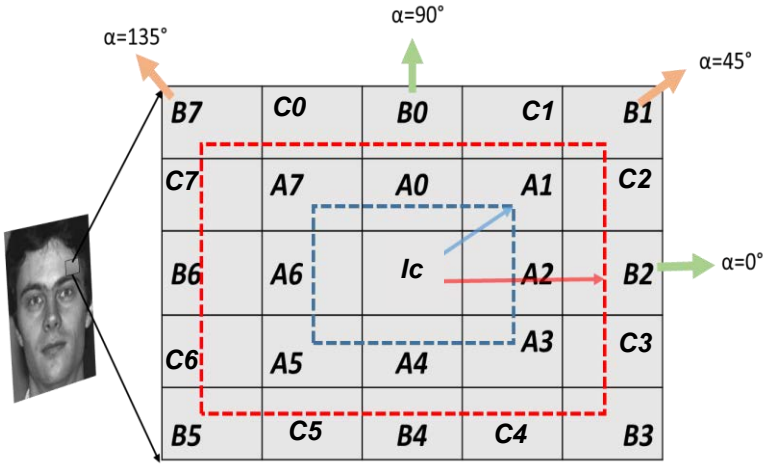

Figure 3.1: Local sampling topology of MNTCDP.

We define, as shown in Figure 3.2, two sampling groups SG_1 and SG_2. SG_1 contains the pixels in blue color, i.e., the even pixels of level A located in the vertical and horizontal directions $\{A_0,A_2,A_4,A_6\}$ and the odd ones of level B located in the two diagonal directions $\{B_1,B_3,B_5,B_7\}$, while SG_2 contains the pixels in green color, i.e., the odd pixels of level A located in the two diagonal directions $\{A_1,A_3,A_5,A_7\}$ and the even ones of level A located in the vertical and horizontal directions $\{B_0,B_2,B_4,B_6\}$. In a more manageable way, the sampling groups SG_1 and SG_2 are given as follows (cf. Eq. 3.1 and Eq. 3.2):

$$SG_1 = \{A_{2i}, B_{2i+1}\}; i = 0, 1..., 3 \tag{3.1}$$

$$SG_2 = \{A_{2i+1}, B_{2i}\}; i = 0, 1..., 3 \tag{3.2}$$

It is known that the average gray level is a widely accepted statistical parameter for texture analysis. The rest of the pixels within the 5×5 neighborhood, labeled as $C_{i;i=0,...,7}$ in Figures 3.1 and 3.2, are considered in order to calculate the average local gray level of the whole 3×3 neighborhood around each pixel in level A, which will be used in the pattern encoding phase. Based on the above neighborhood topology, eight average local gray levels labeled as $\{M_{A0},.....M_{A7}\}$, are computed as follows (cf. Eqs 3.3 and 3.4).

$$M_{A_{2i}} = \frac{\sum_{p\in\omega_{2i}} A_p + \sum_{q\in\upsilon_{2i}} C_q + B_{2i} + \mathbf{I}_c}{\mathbf{P}+1}; \tag{3.3}$$

$$M_{A_{2i+1}} = \frac{\sum_{p\in\omega_{2i+1}} A_p + \sum_{q\in\upsilon_{2i+1}} C_q + B_{2i+1} + \mathbf{I}_c}{\mathbf{P}+1}; \tag{3.4}$$

where i ∈ [0-3], ω_r is the quintuplet centered at element r and υ_s is the doublet with starting element s, extracted from the $\mathbf{P}$-cycle $\mathfrak{G}_{\mathbf{P}}$={0,1,2,....,7} (circular permutation of order $\mathbf{P}$), respectively.

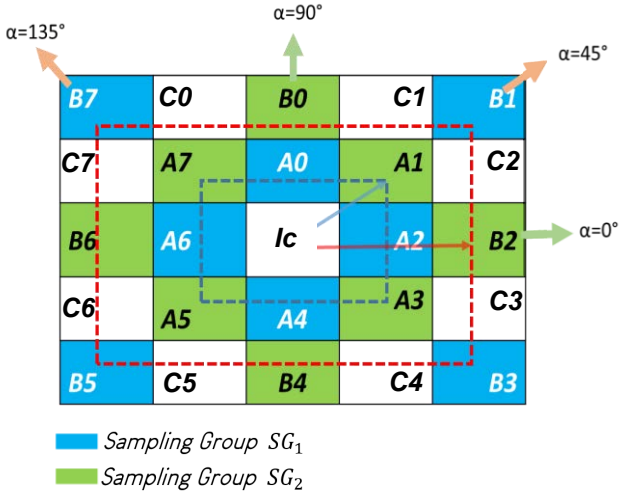

Figure 3.2: The two groups of sampling adopted in MNTCDP descriptor.

3.2.2/ PATTERN ENCODING

After defining the neighborhood topology, which permits to define pixels that will be used in the modeling, we present, as a second step, the pattern encoding scheme of the proposed MNTCDP descriptor. The local information is encoded using two encoders named Ec_1 and Ec_2 based, in addition to the central pixel, on the two groups of samples SG_1 and

SG_2 and the eight average local gray levels $\{M_{A0},\ldots\ldots M_{A7}\}$ defined previously. The two encoders Ec_1 and Ec_2 adopt the basic thresholding function where each pixel, except the central pixel, is compared to the average local gray level of the 3×3 image patch to which the pixel belongs to. The central pixel is compared locally to the average local gray level M_{Ec_1} (cf. Eq. 3.7) of the even pixels $A_{2i;i\in[0-3]}$ and the average local gray level M_{Ec_2} (cf. Eq. 3.8) of the odd pixels $A_{2i+1;i\in[0-3]}$ of level A, and incorporated in the modeling of the two encoders Ec_1 and Ec_2, respectively. The codes produced by the two encoders Ec_1 and Ec_2 associated to the two sampling groups SG_1 and SG_2 are computed as:

$$Ec_1(\chi) = \sum_{i=0}^{3} \Lambda(B_{2i+1}, M_{A_{2i+1}}) \times 2^i + \sum_{i=4}^{\mathbf{P}-1} \Lambda(A_{2(i-4)}, M_{A_{2(i-4)}}) \times 2^i + \Lambda(\mathbf{I}_c, M_{Ec_1}) \times 2^{\mathbf{P}}$$

(3.5)

$$Ec_2(\chi) = \sum_{i=0}^{3} \Lambda(B_{2i}, M_{A_{2i}}) \times 2^i + \sum_{i=4}^{\mathbf{P}-1} \Lambda(A_{2(i-4)+1}, M_{A_{2(i-4)+1}}) \times 2^i + \Lambda(\mathbf{I}_c, M_{Ec_2}) \times 2^{\mathbf{P}}$$

(3.6)

where

$$M_{Ec_1} = \frac{\sum_{i=0}^{3} A_{2i}}{\mathbf{P}/2} \tag{3.7}$$

$$M_{Ec_2} = \frac{\sum_{i=0}^{3} A_{2i+1}}{\mathbf{P}/2} \tag{3.8}$$

In equations 3.2.2 and 3.2.2, χ is the set of gray-scale values of a 5×5 square neighborhood.

To make the representation more robust, the coarse and fine information can be captured by multi-scale which can be made through a linear combination of different features obtained by several operators. In this paper, the final MNTCDP code obtained for each pixel of the image, is the concatenation of the two features generated by the two cross encoders Ec_1 and Ec_2:

$$MNTCDP = \langle Ec_1, Ec_2 \rangle \tag{3.9}$$

3.2.3/ MNTCDP FEATURE VECTOR

MNTCDP operator produces 1024 (2×2^9) possible patterns in a 5×5 neighborhood. As shown later in experimental results, feature images computed using MNTCDP encode useful relationships amongst neighborhood pixels, which help discriminate interclass facial images. The MNTCDP feature extraction process is more explained in Figure 3.3, where the final MNTCDP code is represented as the concatenation of two histograms corresponding to the cross encoder Ec_1 and Ec_2, the histograms are computed separately and combined at the end of the computation to be stored as the feature vector.

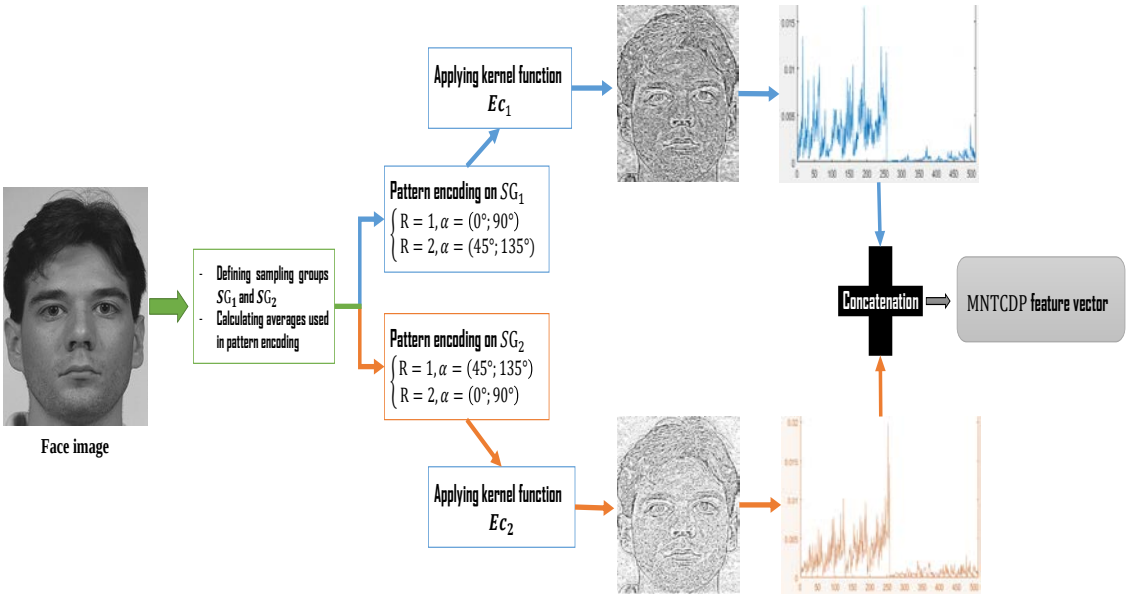

Figure 3.3: The overall framework of the MNTCDP feature vector calculation.

In order to visually show the effectiveness of the coding strategy of MNTCDP, Figure 3.4 illustrates examples of histogram-based matching of two sample images of the same class using LBP, LTP, $nLBPd$, $dLBP\alpha$ and MNTCDP. As we can see, the feature vector obtained by our proposed method carries more information than those of the other descriptors. Therefore, the coding image of MNTCDP can well reflect the structure of the face image.

After encoding each pixel in the face image using the two cross encoders Ec_1 and Ec_2, two code maps are produced. To incorporate more spatial information into the final MNTCDP descriptor, the obtained code maps are spatially divided into small spatial $w \times w$ non-overlapping portions referred to as blocks histograms of MNTCDP codes are extracted from each block. All these regional sub-histograms of dimensionality m are concatenated through Eq. 5.12 to form the holistic face representation of dimensionality $m \times w^2$. The overall framework of the MNTCDP based face representation approach is illustrated in Figure 3.3. This face representation can be directly used to measure the similarity be-

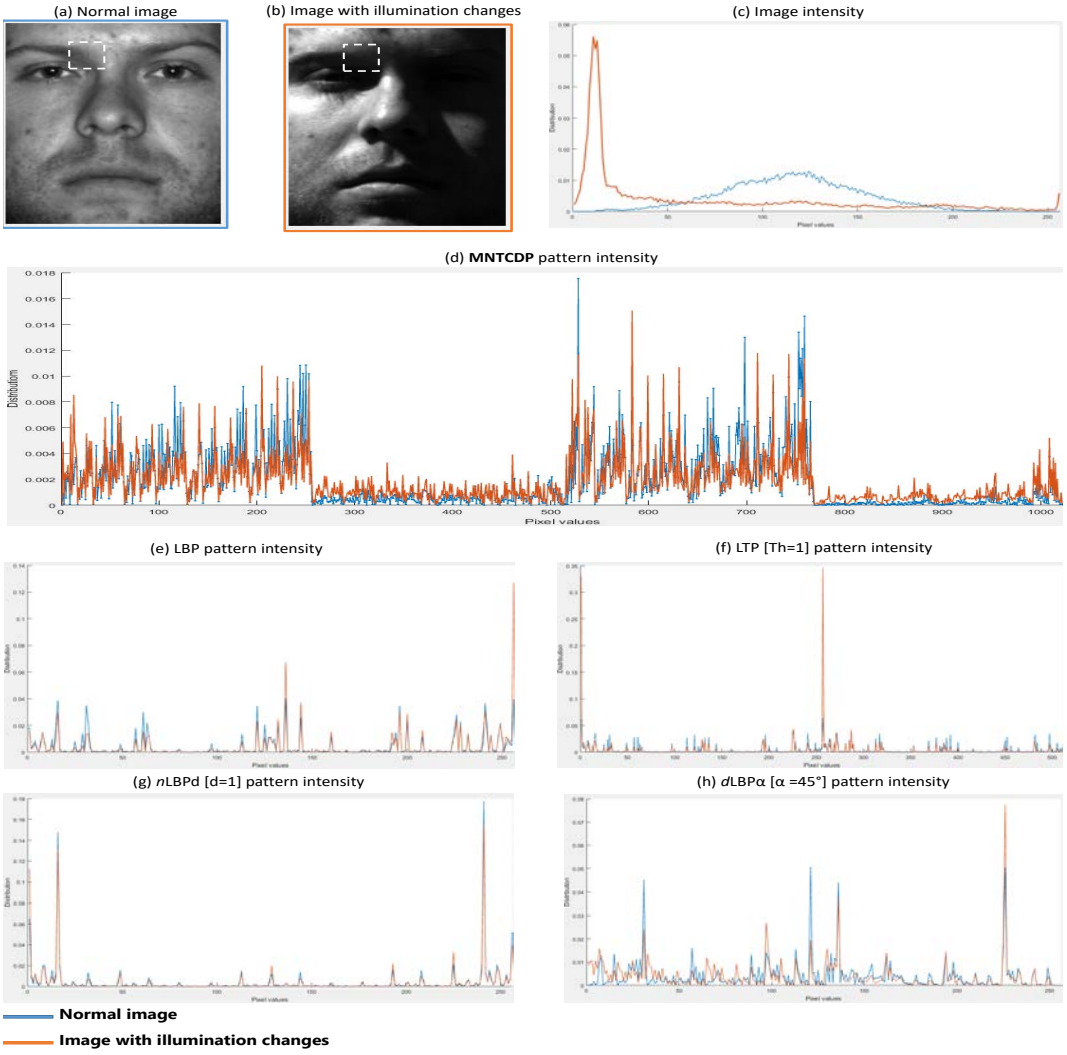

Figure 3.4: Comparing the obtained feature histograms of two images of the same person using MNTCDP, LBP, LTP, $nLBPd$ and $dLBP\alpha$ descriptors.

tween a pair of face images using metrics such as City Block and chi-squared distances or histogram intersection.

$$\mathbb{H} = \prod_{i=0}^{w^2} H_i \tag{3.10}$$

Where $\mathbb{H}$ is the final descriptor, $\prod$ is the concatenation operation, and H_i is the histogram of the MNTCDP codes for block B_i calculated using Eq. 5.9.

$$H_i = \left\langle H_i^{Ec_1}, H_i^{Ec_2} \right\rangle \tag{3.11}$$

where

$$H_i^{Ec_1}(\mathbf{k}) = \sum_{\chi \subset B_i} \delta(\mathrm{Ec}_1(\chi)), \mathbf{k}) \tag{3.12}$$

$$H_i^{Ec_2}(\mathbf{k}) = \sum_{\chi \subset B_i} \delta(\mathrm{Ec}_2(\chi)), \mathbf{k}) \tag{3.13}$$

In Equations 5.10 and 3.13, $\mathbf{k} \in [0, N_{bins}]$ is a pattern to compare to Ec_1 or Ec_2 patterns, N_{bins}=2^9 is the number of bins, χ is the set of gray-scale values of a 5×5 square neighborhood and the delta function $\delta(\cdot)$ is defined as below (cf. Eq. 5.11):

$$\delta(\mathbf{a}, \mathbf{b}) = \begin{cases} 1, & \text{if } \mathbf{a} = \mathbf{b}; \\ 0, & \text{otherwise} \end{cases} \tag{3.14}$$

3.3/ FACE RECOGNITION SYSTEM USING MNTCDP

Similar to most state-of-the-art face recognition systems, the proposed system, as shown in Figure 3.5, involves several steps. First, the images of each dataset are preliminarily divided into ten random splits generated for each number of training images. The number *l* of training images is varied from 1 to $N_{pc} - 1$ (N_{pc} is the number of images per class) and the rest ($N_{pc} - l$) is taken as the testing set. Each split contains two sub-sets, one for the training and the other for testing. The images of the training and testing sets are fed to the next stage without submitting them to any kind of preprocessing technique. Secondly, the feature images are obtained using the proposed MNTCDP operator. After that, each feature image is further divided into $w \times w$ non-overlapping sub-image blocks, and a histogram of patterns is generated for each block. The histogram bins of all blocks are chained to form the final MNTCDP descriptor of the whole image. Finally, the images of the test set are classified through a supervised image classification task. Our focus was on evaluating the discrimination power of the proposed MNTCDP descriptor, and therefore we tried to make as few assumptions as possible and chose the simple non-parametric nearest-neighbor rule (1-NN) to compute the minimum L1 distance between the probe image and the gallery images (cf. Eq. 3.15).

$$\mathfrak{D}_{\mathrm{L1}}(\mathbf{h}_i, \mathbf{h}_j) = \sum_k |\mathbf{h}_i^k - \mathbf{h}_j^k| \tag{3.15}$$

where $\mathbf{h}_i$={$\mathbf{h}_i^1$,$\mathbf{h}_i^2$,...,$\mathbf{h}_i^k$} is the query feature vector and $\mathbf{h}_j$={$\mathbf{h}_j^1$,$\mathbf{h}_j^2$,...,$\mathbf{h}_j^k$} is the target feature vector.

This kind of parameter-free classifiers is particularly suitable for feature comparison pur-

poses as they can handle a large number of classes, avoids parameter overfitting, and requires no learning/training ([71], [110]). The procedure is repeated ten times, each time with new subdivision into training and validation sets, and the accuracy obtained with each subdivision is recorded.

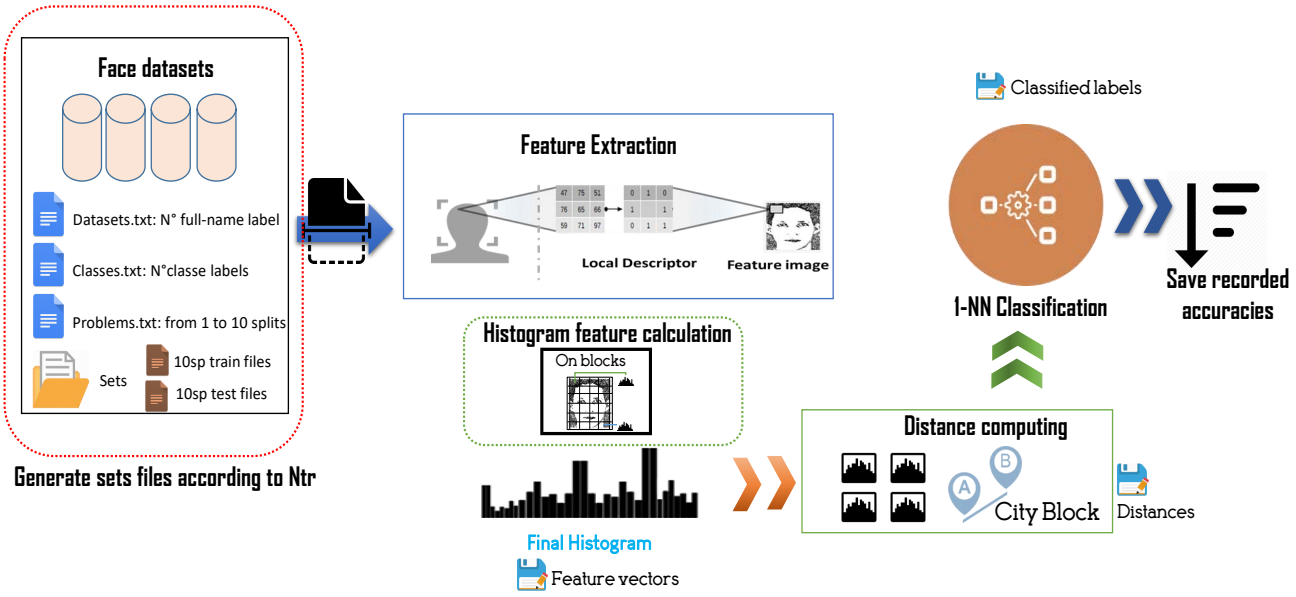

Figure 3.5: Overview of the proposed face recognition system.

3.4/ EXPERIMENTAL ANALYSIS

This section evaluates the proposed MNTCDP descriptor for face recognition to validate its performance and stability. Experiments are conducted through the framework presented in Figure 3.5 using five challenging and widely used benchmarks. A comparison is performed against the 71 LBP-variants listed in Table 2.1 and PCNANet2 and CBFD deep face features within our framework. Moreover, this section highlights the superiority of our MNTCDP-based recognition system according to the state-of-the-art ones.

3.4.1/ DATASETS

Figures 3.6, 3.7, 3.8, 3.9 and 3.10 illustrate images of subjects from ORL, YALE, Extended YALE B, FERET and AR databases, respectively, used in our experiments. The main characteristics of each database are described in the following subsections.

3.4.1.1/ ORL

The ORL[1] Database of Face [111] is composed of 400 images, covering 40 individuals with ten images per person. The size of each image is 92×112. Figure 3.6 illustrates ten images of one person in the ORL database. This database is characterized by a homogeneous dark background, slightly varying lighting conditions, and various facial expressions. The faces are in an upright position in frontal view, with a slight left-right rotation.

Figure 3.6: Images of subject from ORL database.

3.4.1.2/ YALE

The YALE Face Database[2] includes 15 individuals with 11 images per subject, illustrated in Figure 3.7, one per different facial expression or configuration: center-light, with glasses, happy, normal, sad, sleepy, surprised, and wink. This results a total of 165 grayscale images. These images are used to analyze the performance of texture models under noisy conditions, different poses, and illumination changes.

Figure 3.7: Images of subject from YALE database.

3.4.1.3/ EXTENDED YALE B

the Extended YALE Face Database B [112] is composed of 38 subjects taken under nine poses and 64 illumination conditions. The images are divided into five subsets according to the angle between the light source direction and the central camera axis (Subset

[1] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
[2] vision.ucsd.edu/content/yale-face-database

1: 12°, Subset 2: 25°, Subset 3: 50°, Subset 4: 77°, Subset 5: 90°). The Extended YALE-B dataset only contains 38 subjects but with little variability of expression, aging, etc. However, its extreme lighting conditions still make it a challenging dataset for most face recognition descriptors. Figure 3.8 shows the 64 samples of one person from the Extended Yale B dataset.

Figure 3.8: Images of subject from Extended YALE B database.

3.4.1.4/ FERET

The FERET[3] face image database [113] is a large benchmark, widely used database. The FERET database was collected in 15 sessions in 6 years. The database contains 1564 sets of images for a total of 14126 images, including 1199 individuals. We used a subset that contains 1400 images of 200 persons (7 images per person) as adopted in [114] and many other state-of-the-art recent works: [115], [116], [117]. Figure 3.9 presents the images of one class illustrating the various orientations.

Figure 3.9: Images of subject from the used subset of FERET database.

3.4.1.5/ AR

The AR Face Database[4] was created by [118], containing over 4000 face images of 70 men and 56 women. In addition to many images, sunglass and scarf occlusions make this database more challenging the classification task and evaluate the robustness and

[3]http://www.itl.nist.gov/iad/humanid/feret/feret_master.html
[4]http://www2.ece.ohio-state.edu/ aleix/ARdatabase.html

effectiveness of the described methods. A subset that contains 100 individuals (50 men and 50 women) is selected from the AR database. Each subject has 26 samples covering expression changes, illumination variations, and different disguises (sunglass and scarf). Figure 3.10 shows the samples of one class from the used subset.

Figure 3.10: Images of subject from the used subset of AR database.

3.4.2/ EXPERIMENTS CONFIGURATION

To better illustrate the advantages of MNTCDP against the evaluated state-of-the-art descriptors according to the particularities of selected datasets, six experiments are conducted:

- Experiment #1: This experiment is performed on the ORL database, considering all possible numbers of training images. For each tested descriptor, the accuracies are recorded over ten random splits. After performing a series of extensive experiments on the ORL database, we found that the best number of blocks required to obtain the best accuracy is nine blocks for all evaluated descriptors.
- Experiment #2: Like the previous one, this experiment is carried out on the adopted subset of the FERET database. It consists of investigating the performance of MNTCDP descriptor according to the multi-orientation of facial images, which is a sheer fact in face recognition. As on the ORL dataset, we calculate the accuracy over ten subdivisions in the train and test sets, and the best number of blocks to compute the histogram feature is nine blocks for all the tested descriptors.
- Experiment #3: In the objective of evaluating the performance of the proposed descriptor against noisy conditions, this experiment is performed on the Yale dataset,

composed of images with illumination changes. In this dataset, unlike the previous datasets where the evaluated descriptors requested only nine blocks to achieve their highest performance, empirical results show that dividing the encoded image into 100 non-overlapped blocks represents the best adequate block number results in high recognition accuracy. Indeed, histogram feature calculation on 100 blocks permits the feature to precisely describe the feature face images of the YALE database by identifying the difference between the face and the background and overcoming low contrast issues. The stability of the descriptor is again evaluated under all possible numbers of training images.

- Experiments #4: This experiment was designed to test the robustness of the MNTCDP descriptor on a dataset with an important number of images and massive lighting and illumination changes, which are offered by the Extended Yale B database. This experiment aims to investigate the performance according to the number of training images we recorded, for each tested descriptor, the accuracies over ten random splits and under five different train/test configurations (i.e., Train= 5/ Test= 59 images, 10/54 images, 20/44 images, 30/34 images, and 32/32 images). The histogram computation for this experiment was performed on 324 blocks.

- Experiment #5: This is a second experiment carried out on the Extended Yale B database using the evaluation protocol adopted in [52] with five subsets, where Subset 1 is used as the train set, and the remaining four subsets are considered as the validation sets. The recognition rate is calculated on each of the four validation sets without cross-validation. In this experiment, we adopted the same number of blocks as in the previous one.

- Experiment #6: The objective of this experiment is to evaluate the performance of the proposed method against sunglass and scarf occlusion configurations, which was not evaluated in the previous experiments. To achieve that, we used the adopted subset of the AR database. Like the other used datasets, we adopted the evaluation protocol based on various training images. The classification accuracy is recorded over 10 random subdivisions, under 5 Train/Test division configurations (ie. Train= 5/ Test= 21 images, 10/16 images, 13/13 images, 15/11 images and 20/6 images). The histogram feature computation is performed on 324 non-overlapping blocks.

- Implementation and Execution: The face recognition experiments have been performed on an HP ProDesk with Core i7 Processor 4.0 GHz with turbo boost technology and 16GB of RAM, running with Ubuntu 16.04 LTS (Xenial Xerus) operating system. The descriptors and the recognition system have been implemented in the MATLAB® R2016b environment.

3.4.3/ EXPERIMENTAL EVALUATION AGAINST LBP-LIKE DESCRIPTORS

Tables 3.1, 3.2, 3.3, 3.4 and 3.6 report the average accuracies (i.e., over 10 random splits) achieved by the proposed MNTCDP descriptor and the top 50 evaluated state-of-the-art descriptors on ORL (Exp#1), FERET (Exp#2), YALE (Exp#3), Extended Yale B (Exp#4) and AR (Exp#6) databases, respectively. Table 3.5 summarizes the obtained recognition rates on Extended Yale B database adopting the subsets based experimental protocol adopted in [52].

3.4.3.1/ PERFORMANCE ANALYSIS ON ORL: EXPERIMENT #1

The ORL database is widely used in the literature to test new descriptors and approaches thanks to frontal images taken in a controlled environment, equilibrated number of subjects, and the samples per each subject. The challenge herein is to achieve 100% average accuracy with a number of reference images as minimum as possible. Table 3.1 reports the average accuracy of each evaluated descriptor. The results in this Table clearly depict that MNTCDP is the top 1 under all numbers of training images, as it reaches the highest accuracies compared to the tested state-of-the-art descriptors. It is to underline that the proposed MNTCDP descriptor achieves a score of 98.64% at 3/7 configuration (training/testing sets), outperforming the rest of the descriptors. Note that realizing this recognition rate over ten random splits with few images in the train set (3 images only) is a promising result for real-time applications requiring minimum computational resources. The most important thing that can be noticed is the 100% average accuracy recorded over the half/half (Train/Test) configuration and over the ten random splits. Achieving a 100% recognition rate over ten subdivisions with only five images in the train set demonstrates both the stability and the powerful description of MNTCDP. The AECLBP-S-MxC and NI-CI-LBP descriptors came in second place and reached an encouraging result in the ORL database where they realized 100% average accuracy at 8/2 configuration while other tested descriptors recorded 100% average accuracy only for the 9/1 setup. Note that many descriptors like XCS-LBP, WLD, and RDLBP suffer from the drop and decrease in their average recognition rates when the number of training images increases. Figure 3.11 illustrates the performance evolution of the top 5 descriptors on the ORL dataset. It can be found from Figure 3.11 that the performance of the MNTCDP descriptor increased rapidly to reach 100% average accuracy compared to all the evaluated state-of-the-art descriptors. The second remarkable statement, which can be made from this Figure, is the performance gap between the MNTCDP descriptor and its competitors, mainly when the test set contains more images than in the train one.

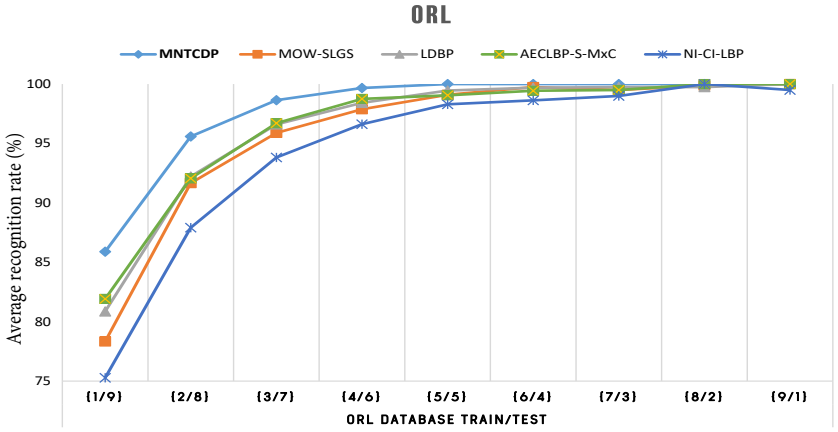

Figure 3.11: Performance evolution of top 5 descriptors on ORL dataset.

3.4.3.2/ PERFORMANCE ANALYSIS ON FERET: EXPERIMENT #2

To test the robustness of the proposed descriptor on FERET, a real-world challenge database, we set up a subset containing the frontal images and also samples with various orientations taken at different sessions. Unlike existing state-of-the-art approaches, we used facial images with the whole original background (not cropped images), negatively influencing the description process. Table 3.2 summarizes the obtained average face recognition accuracies recorded over ten splits under all possible numbers of training images. It can be seen that MNTCDP shows the best performance as it outperforms all the evaluated methods. MNTCDP achieves a higher recognition rate of 99.7% with six images in the train set over ten random splits, vs. 99.1% with DDBP, 99.05% with DCLBP and AECLBP-S, 98.99% with CRLBP-S-MxC, which are considered as the top performing descriptors (following the proposed one) in this experiment. To underline that some descriptors like LDBP and QBP which achieved good results on the ORL database, could not keep the same performance on this subset of FERET. In contrast, DDBP, which was not among the top-ranked descriptors on the ORL database, performs well on the FERET database. It would be of interest to note that facial images with orientations between 67.5° and 90° are hard to recognize, especially when the train and test sets are composed of samples with different angles, which is the case of the selected subset. This fact has prevented the proposed descriptor from reaching perfect accuracy of 100% over ten subdivisions. Moreover, achieving an average accuracy above 95% in a challenging database like FERET is considered a very satisfactory result. This is the case of the proposed descriptor, which achieves higher accuracies (above 96%) starting from 4 images in the train set. From Figure 3.12, we find that the accuracy recorded by the proposed MNTCDP method surpasses all the tested state-of-the-art descriptors and this under all numbers of training images. Furthermore, MNTCDP realizes a clear advantage against the other descriptors between the two train/test configurations 2/5 and 5/2.

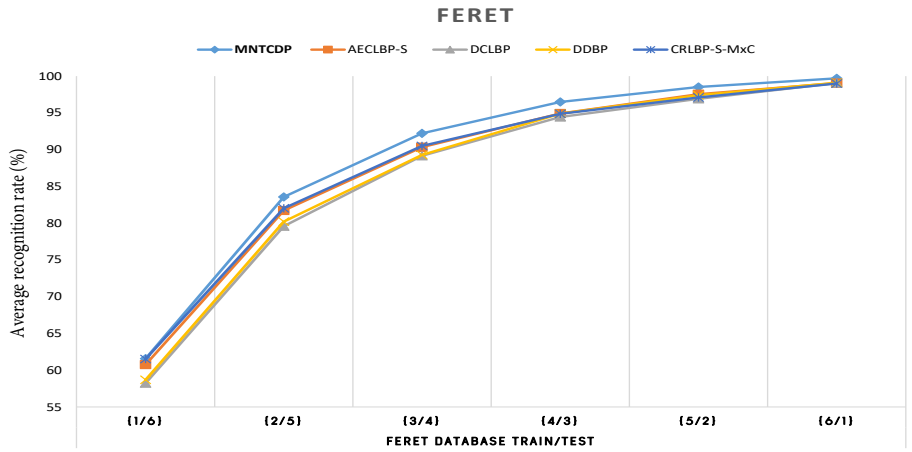

Figure 3.12: Performance evolution of the top 5 descriptors on FERET dataset.

3.4.3.3/ PERFORMANCE ANALYSIS ON YALE: EXPERIMENT #3

In this experiment, the objective is to demonstrate the performance and the effectiveness of the proposed MNTCDP operator and the state-of-the-art evaluated methods against massive illumination changes and variable background of the subjects. Table 3.3 shows the classification performance of the evaluated descriptors. It is apparent from Table 3.3 and Figure 3.13 which illustrates the performance evolution of the five top-ranked descriptors according to the different number of training images that the proposed MNTCDP descriptor proves a great success in the Yale database. Indeed, it achieves, for almost all the train/test configurations, accuracy above 91%, and it is the only descriptor able to reach 100% over ten random splits using seven or more images in the train. Note that, oppositely to all tested descriptors, MNTCDP provided rising scores from Train/Test configuration to the following one. Moreover, we remark that MNTCDP is more stable than the evaluated descriptors according to the train/test configurations. The second-best performing descriptor is dLBPα with its high score of 99.67% achieved only with two images in the test set. XCS-LBP method occupies the third rank by reaching 99.3% recognition rate average over ten splits at 7/4 configuration.

3.4.3.4/ PERFORMANCE ANALYSIS ON EXTENDED YALE B: EXPERIMENTS #4 AND #5

Due to its large number of images per subject and illumination changes, the Extended YALE B dataset is one of the most selected databases in the literature to evaluate the performance yet robustness of new descriptors and face recognition frameworks. As indicated previously in Section 3.4.1.3, we adopted two experimental protocols: the first one investigates the performance of the evaluated descriptors according to the number of training images (Experiment #4), while in the second experiment, the images are divided into five sets as in [52]: Subset 1, Subset 2, Subset 3, Subset 4 and Subset 5 respectively

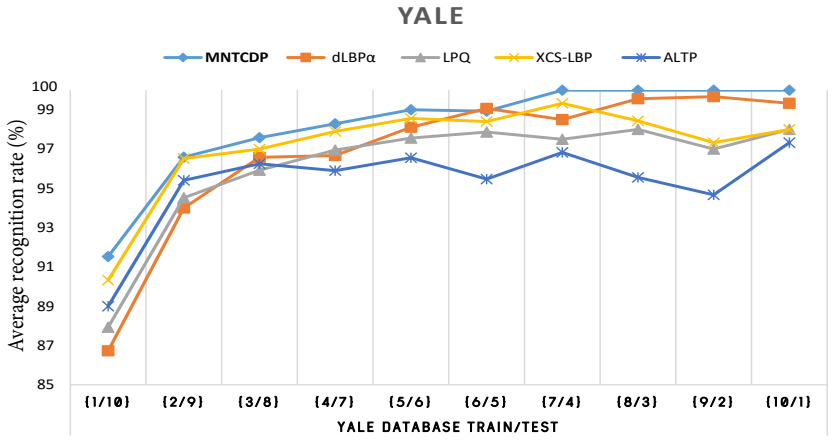

Figure 3.13: Performance evolution of the top 5 descriptors on YALE dataset.

(Experiments #5). Subset1 is used as the gallery set.

- Experiment #4 (according to the number of training images): As each subject of this dataset has 64 samples, it is computationally complicated to calculate the average accuracy over each number of training images. To study the performance stability and recognition rate evolution, we performed the classification with 5, 10, 20, 30, and 32 images in the train set. Table 3.4 reports the obtained classification results (average accuracies over ten random splits). The results in Table 3.4 indicate that the proposed MNTCDP operator has obtained the maximum recognition rates for 20/44, 30/34, and 32/32 Train/Test configurations. MNTCDP obtains the maximum score of 98.91% at half/half configuration, followed by NI-LBP and DDBP, which can be considered as the second and third top-performing descriptors on the Extended YALE B dataset, where they reached 98.08% and 98.04% average accuracies, respectively, at half/half configuration. Figure 3.14 illustrates the evolution of the top 5 descriptors' performance according to the Train/Test partitions. We considered NI-CI-LBP and CI-LBP among the top 5 ranked descriptors, thanks to their higher recognition rates recorded with fewer images in the train set. They recorded, when using the configuration 5/59, 92.39%, and 91.26%, respectively. It is clear that MLDCBBP is the descriptor that achieved the highest recognition rate average, conserving sustainable stability over Train/Test configurations.

- Experiment #5 (Subset evaluation protocol): Extended YALE B database can be divided into five subsets from slight to extreme lighting variations. Subset 1 is considered a training set; it contains images taken under nominal lighting conditions, while the other subsets are used as test sets. Subsets 2 and 3 are characterized by slight-to-moderate illumination variations, while subsets 4 and 5 depict severe illumination changes. The obtained recognition rates of the four subsets are listed in Table 3.5. As can be seen, we disclose that subset two is the easiest one where

almost all the evaluated descriptors achieved a 100% recognition rate. We can express the same remark for subset three, where many descriptors and the proposed MNTCDP descriptor manage to differentiate all classes perfectly (average accuracy equal to 100%). Note that MNTCDP succeeded to record the top 1 accuracy on Subsets 4 and 5, by reaching 99.81% and 98.52% recognition rate, respectively. ELGS is the second-best performing descriptor on Subset 4 with an accuracy of 98.67%. Meanwhile, the LPQ descriptor secured the second rank on Subset 5 by recording 98.07% accuracy. MNTCDP descriptor demonstrated, unlike all the evaluated state-of-the-art descriptors, consistent stability across the four subsets despite the challenging conditions of Subsets 4 and 5.

The remarkable performance of the proposed MNTCDP operator for the classification of face samples of the Extended YALE B database using the two evaluation protocols (Exp#4 and Exp#5) demonstrates its strength and effectiveness against massive illumination changes and the significant number of samples per subject.

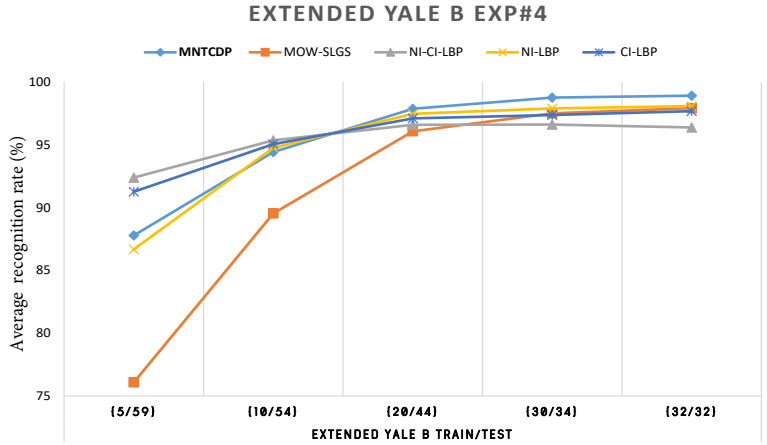

Figure 3.14: Performance evolution of top 5 descriptors on EYB dataset.

3.4.3.5/ Performance analysis on AR Face Database: Experiment #6

The experiment on the AR dataset was designed to evaluate the performance of the proposed MNTCDP method and the state-of-the-art descriptors against occlusions. The selected subset contains 2600 images of 50 men and 50 women subjects with 26 face images per individual, including sunglass and scarf occlusion situations. Table 3.6 lists the obtained results on the AR subset adopting the evaluation protocol discussed in Experiment #6. MNTCDP achieved an average accuracy of 82.33% over ten random splits using only 5 images in the train set, surpassing all the evaluated state-of-the-art handcrafted descriptors 4%. By enlarging the train set to 13 images per subject, which represents half of the subject set, the proposed descriptor attained 96.18% average accuracy, which

is a very satisfying rate on images with occlusions, outperforming all the other evaluated descriptors. The maximum average recognition rate of 99.28% is obtained again by the MNTCDP descriptor at 20/6 train/test configuration. LDN descriptor, which was not among the top five descriptors on the previous datasets, managed to be the second-best performing descriptor. It achieved an average accuracy of 99.1% by adopting 20 samples per individual in the train set. LQP descriptor, as on Yale dataset, reached the third-highest score of 99.03% at 20/6 configuration on this subset of AR database, followed by SLGS and DSLGS descriptors, which recorded 99.05% and 99.01% average accuracies at the same configuration, respectively. On the other side, WLD and LESTP handcrafted descriptors are the worst-performing methods in the experiment, and the first one attained a maximum average accuracy of 88.55% exploiting 20 samples (in the train) and testing 6 images, while the second one reached 88.9% average accuracy for the same number of testing images. Figure 3.15 illustrates the evolution of recognition rates according to the number of training images. It's clear that MNTCDP realizes the highest accuracy at each Train/Test configuration, presenting sustainable stability of its performance.

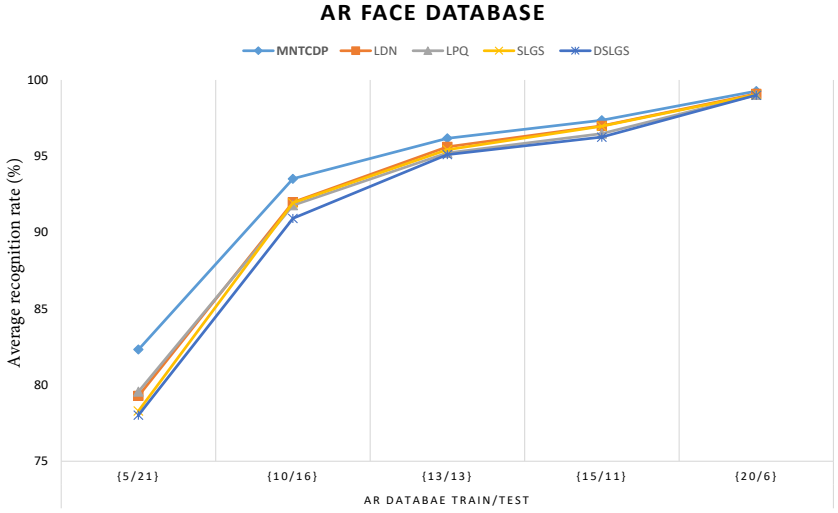

Figure 3.15: Performance evolution of top 5 descriptors on AR dataset.

3.4.4/ EXPERIMENTAL EVALUATION AGAINST PCANET2 AND CBFD DEEP FEATURES

After proving the superiority of the MNTCDP descriptor over the state-of-the-art local descriptors, this subsection is dedicated to prove it one again against two deep face feature extraction methods PCANet2 and CBFD. The main statements are sammurized as follows:

- ORL Database (Exp#1): MNTCDP achieved at each train/test configuration the

highest recognition rate over ten splits against PCANet2 and CBFD methods. The proposed descriptor consistently increased the recognition rate from a train/test configuration to the next one and CBFD method, while PCANet2 deep learning architecture suffered from performance drop at 5/5 and 6/4 configurations. We mention that PCANet2 achieved 100% accuracy using four images in the train set but only over one split.

- FERET Database (Exp#2): As we can see, the CBFD method surpassed MNTCDP and PCANet2 methods from 1/6 to 4/3 train/test configurations, but MNTCDP achieved the top average accuracy at the two last configurations. The maximum accuracy obtained at 6/1 configuration by all the methods is close: MNTCDP attained 99.7% over ten splits, PCANet2 method recorded the second top accuracy of 99.5% over one split, while CBFD gave the 3rd accuracy of 99% over one split. PCANet2 could not overcome the drawback of a performance drop, and it was again unstable as on ORL database.

- YALE Database (Exp#3): The effectiveness and stability of MNTCDP are again proved on YALE Database. It reached 100% average accuracy at 7/4 configuration and showed stability, while the PCANet2 deep learning method was not stable at all on this dataset, which suffered a significant performance drop of 83% (from 90.48% at 4/7 configuration to 07.78% at the following one). CBFD method presented promising performance, it achieved 99.12% using only one image in the train set and recorded 100% using 6 images in the train instead of 7 samples in the case of our proposed descriptor. However, CBFD method experienced minor performance drop several times and could not keep the 100% accuracy even over 1 split.

- Extended YALE B Database (Exp#4 and Exp#5): The performance of deep learning methods on Extended Yale B dataset according to the number of training images (Exp#4), experienced some drops at some configurations, while the proposed MNTCDP showed stability. The maximum accuracy achieved by PCANet2 was limited to 94.55% over one split at 30/34 configuration and CBFD method reached 97.5% accuracy at 32/32 configuration, while MNTCDP managed to record 98.91 over 10 random splits at 32/32 configuration. In Exp#5(based on dataset partition into subsets according to the illumination conditions from slight to severe), as we can see, all the tested methods did successfully achieve 100% accuracy on Subset 2. The performance of PCANet2 and CBFD methods on Subset 3 decreased slightly (below 100%) as they achieved 99.47% and 98%, respectively, while MNTCDP kept its performance at 100%. MNTCDP reached good accuracies on Subset 4 (99.81% accuarcy) and Subset 5 (98.52% accuarcy) outperforming the two other methods. PCANet2 achieved 92.21% on Subset 4 but its performance de-

creased to 71.11% on Subset 5. CBFD method could not maintain its effectiveness on Subsets 4 and 5 as it recorded only 70.34% and 54.03%, respectively.

- AR Database (Exp#6): On AR Database, the maximum accuracy of all the tested methods was arround 99%. MNTCDP reached 99.28% average accuracy over 10 splits at 20/6 configuration. by PCANet2 method attained 99.08% one split accuracy at 13/13 configuration, while CBFD achieved 99% over one split at 15/11 train/test configuration. Deep learning methods demonstrated their effectiveness and performance at low numbers of training images, however their stability cannot be guaranteed. Evaluating its performance over 10 random splits, MNTCDP was capable of reaching higher accuracies and providing performance stability from train/test configuration to another.

In light of the previous statements, we conclude that MNTCDP outperformed PCANet2 and CBFD methods in performance and stability in all performed experiments. PCANet2 method suffered of major performance drops and as we saw it performed badly at various configuration of training and testing sets. CBFD presented better stability than the PCANet2 method but did not outperform the proposed method on all the datasets. Moreover, the evaluated deep methods have many parameters to adjust, which require many experiments in order to reach the optimal ones. Meanwhile, the proposed descriptor has no parameter to set and gave excellent and competitive accuracies on all the tested datasets. Deep learning methods are computationally expensive and require pre-trained models and feature learning phase in case of deep feature methods.

3.5/ COMPARISON WITH STATE-OF-THE-ART SYSTEMS

This section is dedicated to comparing the proposed MNTCDP face recognition system performance recorded on each dataset against those reported in recent state-of-the-art works. It should be pointed out that each published paper considered for comparison purpose is reviewed carefully to determine the top recognition rate achieved on each dataset along with the used experimental setting, i.e., the used configuration into training/validation sets and number of splits. The extracted results from reviewed state-of-the-art papers are arranged in Table 3.8 for Experiments #1 to #4 and Experiment #6, and in Table 3.9 for Experiment #5. Based on these results, we can readily make the following statements:

- ORL: It can be inferred that MNTCDP based system significantly outperforms all reported papers by achieving 100% average accuracy over ten random splits using only 5 images in the train set, while the best performing state-of-the-art work [117]

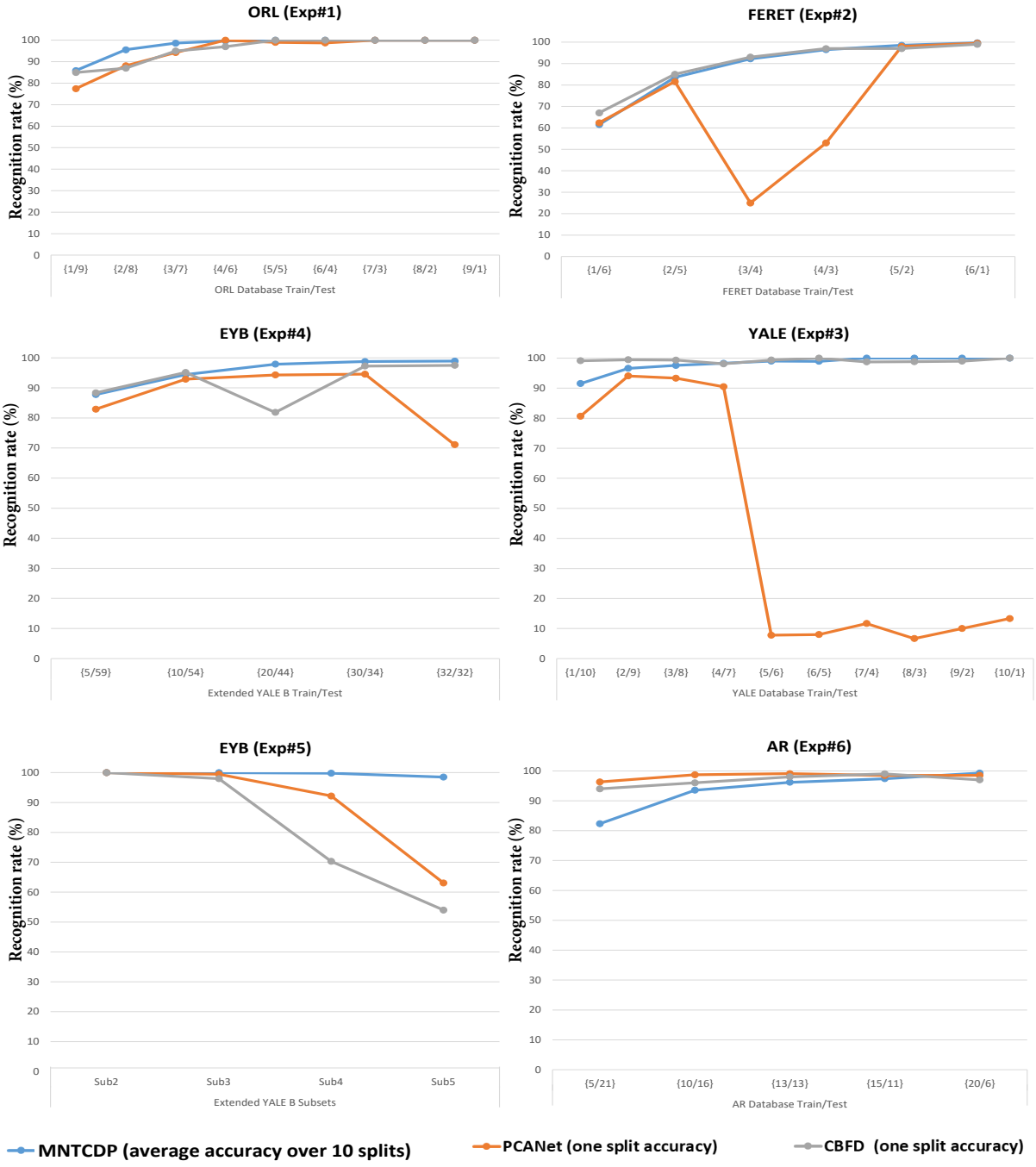

Figure 3.16: Performance evaluation of the tested deep learning methods.

recorded only 98% at the same configuration over only one split. Note that the top accuracy reported in the literature was 100% reached by [34] but with eight images in the train and 2 for the test, over only 1 split.

- FERET: The results recorded on the FERET database listed in Table 3.8 show that MNTCDP outperforms again the best performing existing approaches, especially with the two most used 5-2 configuration into training/testing sets where it reached

a score of 98.52% over ten random splits. As can be seen, many works suffer on the 5/2 configuration using only 1 split and are unable to achieve a competitive recognition rate to that obtained by the proposed MNTCDP based system. The most competitive existing recognition rate was 98% at 3/4 configuration over only 1 split, which is achieved by [128]. Note that the performance of the MNTCDP based system over 1 split achieved a score of 100% after feeding the train set with 5 images.

- YALE: It can be observed clearly from Table 3.8, that the best achieved state-of-the-art recognition rate was 96% at 5/6 configuration over only 1 split [122] vs 99% with MNTCDP based system over 10 splits (see Table 3.3). To our best knowledge, we are the first to achieve 100% average accuracy with the configuration of 7 images for the train set and 4 images as the probe set.

- Extended YALE B: The recognition accuracy reported in the literature and the one obtained by our system according to the training images (Exp#4) are listed in Table 3.8. It is easily found that the MNTCDP based system outperformed all existing state-of-the-art systems, by achieving 98.91% average accuracy over 10 splits using half/half configuration, exceeding [123], [142] and [145] by 12.36%, 5.51% and 5.88%, respectively. Table 3.9 summarizes the face recognition results on Extended YALE B using illumination subsets evaluation protocol. In light of these results, it is clear that many state-of-the-art approaches as well as the proposed MNTCDP based system manage to differentiate all classes perfectly (average accuracy equal to 100%) on Subsets 2 and 3. Furthermore, our system achieved the top accuracy of 99.81% on Subset 4 and ranked second on Subset 5 by achieving a score of 98.52% surpassed by [152] (only with only 0.37%) which reached 98.89%. Note that the method presented in [152] could not reach 100% on Subset 4 .

- AR Database: On AR dataset, the MNTCDP based system managed to outperform all reported state-of-the-art systems. It reached 99.28% average accuracy over 10 random permutations into train/test subdivisions using 20 images in the train set and 6 samples for testing. At the same time, the best result of the literature works is 99% accuracy achieved by [151] at the same configuration but considering only one split. At 20/6 train/test configuration and evaluating the performance over 10 splits, [148] reached 98.97% average accuracy. Hence, the proposed hand-crafted MNTCDP based system demonstrated its effectiveness and stability according to facial images with sunglass and scarf occlusions.

Table 3.1: Average recognition accuracy (%) over 10 splits on ORL dataset (Exp#1)

Descriptor	ORL Database Train/Test								
	1/9	2/8	3/7	4/6	5/5	6/4	7/3	8/2	9/1
MNTCDP	**85.89**	**95.59**	**98.64**	**99.67**	**100**	**100**	**100**	**100**	**100**
AECLBP-M	75.61	87.34	93.54	95.21	97.25	98.31	98.25	99	99.25
AECLBP-S	77.06	89.72	94.39	97.42	98.4	99.44	98.83	99.5	99.5
AECLBP-S-MxC	81.92	92.06	96.71	98.75	99.05	99.44	99.5	**100**	**100**
AELTP	79.72	91.38	95.43	97.87	98.65	99.5	99.25	99.5	99.75
ALTP	79.64	91.59	96.07	97.75	98.65	99.5	99.42	99.63	99.5
BGC-1	79.44	91.5	95.89	98.17	99.05	99.25	99.25	99.63	99.75
BGC-2	77.06	90.06	95.36	97.79	98.6	99	99.25	99.5	99.75
BGC-3	76.42	89.34	94.25	97.21	98.65	99.13	98.92	99.63	99.75
CI-LBP	39.69	54.69	60.43	65.33	68.45	73.31	74.58	75.25	74.75
CLBC-M	65.58	79.84	86.21	88.88	92.55	94.94	94.83	95.75	95.5
CRLBP-M	76.83	87.38	93.86	95.04	96.7	98.31	97.92	98.88	99
CRLBP-S	79.06	91.22	95.32	97.67	98.95	98.94	98.83	99.63	99.75
CRLBP-S-MxC	81.89	91.69	96.64	98.38	98.9	99.5	99.25	99.88	99.75
CSALTP	71.94	86.69	91.57	95.79	97.65	98.69	98.67	99.63	99.75
DBC	72.03	85.44	91.86	94.29	97	98.13	98.17	99.38	99.25
DCLBP	80.47	91.44	96.79	98.04	98.7	99.31	99.33	99.63	99.75
DDBP	80.39	90.53	95.29	97.83	98.85	99.19	99.17	99.5	99.75
dLBPa	73.22	85.66	91.36	94.5	96.4	97.5	98	98.88	99.25
DRLBP	65.94	80.16	87.07	90.92	92.75	95.5	95.58	96.75	97.5
DSLGS	79.5	91.06	95.96	98.13	98.6	99.44	99.08	99.5	99
ELGS	78.5	90.94	95.96	98.21	98.9	99.69	99.25	99.63	**100**
IBGC-1	81.03	92.78	97.21	98.42	99.15	99.5	99.58	99.75	**100**
ILTP	81.67	92.31	97.14	98.88	98.8	99.69	99.5	99.75	99.5
LBP	77	89.69	95.11	97.04	98.1	99	99	99.5	99.25
LCCMSP	74.58	86.25	93.93	96	97.8	98.88	99.17	99.63	99.75
LDBP	80.86	92.22	96.61	98.42	99.45	99.69	99.75	99.75	**100**
LDN	65.89	80.53	86.82	90.83	93.95	94.56	94.83	97.5	96
LECTP	65.39	77.22	83.61	88.42	90.25	92.75	92.67	96.25	93.75
LETRIST	80.81	91.25	95.21	97.96	98.2	99.25	99.67	99.75	99.5
LESTP	71.33	84.16	90.14	92.58	94.3	96.56	96.83	97.88	97.75
LGCP	68.97	83.78	89.93	94.04	96.15	98.19	98.08	99.25	99.5
LGS	77.78	90.47	95.25	97.29	98.4	99.31	98.83	99.25	98.75
LMEBP	60.22	72.97	81.61	84.58	87	90	89.83	93.13	93
LNDP	72.53	85.56	91.64	95.58	96.9	98.06	99.17	99.25	99.5
LQPAT	76.69	88.31	93.5	96.17	97.8	98.63	99.5	99.75	**100**
LPQ	79.86	92.75	96.29	98.96	99.15	99.56	99.5	99.75	99.75
LTP	81.03	92.47	96.36	98.42	98.95	99.56	99.42	99.75	**100**
MMEPOP	69.97	85.34	91.25	94.04	95.8	97.38	97.75	98.63	98
MOW-SLGS	78.33	91.66	95.89	97.88	99.1	99.75	99.5	99.88	**100**
LBP	78.83	91.16	95.5	97.63	98.45	99.13	98.67	99.5	99
NI-CI-LBP	75.28	87.91	93.82	96.62	98.3	98.63	99	**100**	99.5
NI-LBP	75.64	88.66	94.36	97	98.3	99.06	99	99.38	99.75
NI-RD-LBP	76.89	89.16	94.71	97.58	98.5	99.25	99.25	99.63	**100**
QBP	74.78	87.47	92.54	95.58	97.2	98.56	98.25	98.75	98.75
RD-CI-LBP	75.47	88	93.54	96.42	98.35	98.88	98.58	99.88	98.75
RDLBP	75.75	88.28	93.46	96.25	97.9	99	98.67	99.38	99.25
SLGS	79.5	91.06	95.96	98.13	98.6	99.44	99.08	99.5	99
SMEPOP	77.39	89.72	95.07	97.33	98.45	98.88	99.25	99.38	99.5
WLD	80.25	91.88	96.36	97.46	98.25	99.25	98.92	99.5	98.5
XCS-LBP	77.11	86.97	92.25	95.04	96.15	98	97.83	99.13	99.25

Table 3.2: Average recognition accuracy (%) over 10 splits on FERET dataset (Exp#2)

Descriptor	FERET Database Train/Test					
	1/6	2/5	3/4	4/3	5/2	6/1
MNTCDP	61.57	**83.56**	**92.21**	**96.48**	**98.52**	**99.7**
AECLBP-M	53.51	74.13	84.51	90.4	94.02	96.43
AECLBP-S	60.76	81.72	90.33	94.91	97.54	99.05
AECLBP-S-MxC	51.42	73.08	83.59	90.44	93.99	97.19
AELTP	55.57	76.57	86.86	92.33	95.8	98.14
ALTP	55.9	76.84	87.34	92.71	96.11	97.89
BGC-1	57.48	78.17	88.2	93.35	96.26	98.29
BGC-2	55.7	76.71	87.12	92.63	95.88	97.99
BGC-3	54.37	75.9	86.17	91.91	95.35	97.54
CI-LBP	31.36	44.64	52.63	58.39	64.4	67.14
CLBC-M	45.94	64.72	74.92	81.01	85.95	90.05
CRLBP-M	54.07	75.21	84.77	90.03	93.87	96.38
CRLBP-S	56.28	76.94	86.96	92.35	95.4	97.24
CRLBP-S-MxC	61.51	82.02	90.5	94.87	97.09	98.99
CSALTP	49.31	70	81.16	87.44	91.93	95.03
DBC	49.1	69.92	81.71	88.78	93.29	96.13
DCLBP	58.24	79.58	89.16	94.44	96.91	99.05
DDBP	58.69	80.21	89.28	94.89	97.41	99.1
dLBPa	57.54	77.28	86.49	91.93	94.77	96.68
DRLBP	54.92	72.58	82.5	88.11	92.14	94.42
DSLGS	57.04	78.09	88.24	93.74	96.66	98.64
ELGS	57.24	78.43	88.63	93.53	96.66	98.49
IBGC-1	58.61	79.73	89.79	94.71	97.36	99.05
ILTP	57.22	78.59	88.18	93.25	96.63	98.69
LBP	55.44	76.39	87.24	92.71	95.8	97.59
LCCMSP	63.92	82.17	91.19	94.64	96.81	97.79
LDBP	58.83	79.54	89.52	94.74	96.88	98.69
LDN	45.38	66.15	77.61	85.39	90.05	93.42
LECTP	53.14	70.25	80.24	85.73	89.57	92.51
LESTP	52.77	71.03	80.8	85.9	90.2	93.17
LETRIST	65.86	83.15	90.82	95.09	97.01	98.19
LGCP	41.48	61.78	73.48	80.6	86.38	90.05
LGS	57.5	78.08	88.27	93.25	96.26	98.44
LMEBP	51.31	68.54	77.66	84	87.86	90.5
LNDP	52.91	73.26	83.22	90.75	94.22	97.14
LPQ	55.6	76.87	86.56	92.41	95.83	98.14
LQPAT	52.48	74.14	83.94	91.96	94.97	97.24
LTP	55.6	76.87	86.56	92.41	95.83	98.14
MMEPOP	49.21	70.89	83.13	89.95	94.22	96.63
MOW-SLGS	57.19	77.89	88.13	93.57	96.58	98.54
$nLBPd$	55.64	76.69	87.19	92.48	95.8	98.14
NI-CI-LBP	53.23	74.03	83.92	89.53	93.39	95.33
NI-LBP	51.54	72.55	83.42	90.08	93.97	96.93
NI-RD-LBP	55.8	76.31	86.76	92.35	95.5	97.49
QBP	50.93	70.45	81.03	87.47	91.33	94.17
RD-CI-LBP	56.06	76.55	85.58	90.79	93.87	95.18
RDLBP	57.35	76.74	87.27	92.35	95.23	97.39
SLGS	57.04	78.09	88.24	93.74	96.66	98.64
SMEPOP	52.82	74.12	85.63	91.56	95.03	97.49
WLD	**62.64**	80.15	88.39	92.91	95.5	98.04
XCS-LBP	52.94	72.47	82.55	88.64	93.54	96.63

Table 3.3: Average recognition accuracy (%) over 10 splits on YALE dataset (Exp#3)

Descriptor	YALE Database Train/Test									
	1/10	2/9	3/8	4/7	5/6	6/5	7/4	8/3	9/2	10/1
MNTCDP	**91.53**	**96.59**	**97.58**	**98.29**	**99**	98.93	**100**	**100**	**100**	**100**
AECLBP-M	84.07	91.7	92	92	92.89	92.8	95.5	93.78	91	93.33
AECLBP-S	89.47	94.96	95	95.52	96.33	95.07	96.83	96	95.67	97.33
AECLBP-S-MxC	86.67	91.7	90.92	91.24	90.89	90.13	92	90.22	89	90.67
AELTP	88.73	94.3	93.83	94.67	94.33	93.2	96.17	94.89	93.67	96.67
ALTP	89	95.41	96.25	95.9	96.56	95.47	96.83	95.56	94.67	97.33
BGC-1	87	90.22	92	92.1	92.78	92.67	93.17	92.22	91.67	92.67
BGC-2	87.4	91.26	92.92	92.95	94.78	94.13	95.33	93.33	93.67	94.67
BGC-3	87.6	92.59	93.33	93.24	93.56	93.33	94.17	93.11	92.67	92.67
CI-LBP	70.73	81.56	85.33	85.81	86.22	86.93	89.83	88.44	90.67	88.67
CLBC-M	80.93	89.56	90.58	90.76	91.56	91.2	93	90.67	89.33	93.33
CRLBP-M	82.33	90.22	91	89.9	91.44	91.6	94.17	92	90	93.33
CRLBP-S	88.4	93.7	94.58	94.38	94.89	94.27	94.83	94.22	93.33	94.67
CRLBP-S-MxC	84.6	90.37	90.33	89.9	90.22	89.2	91	88.44	87.33	88.67
CSALTP	88.07	92.81	92.42	92.48	91.89	90.93	92.83	91.33	91.67	92.67
DBC	86.67	90.22	93.25	92.57	93.56	92.93	93.5	92	91.67	90.67
DCLBP	89.6	95.78	95.75	96.48	96.56	95.6	97.83	96.22	95.33	96
DDBP	87.87	93.41	94.58	93.81	95	93.6	94.67	92.89	93.67	96
dLBPa	86.73	94	96.58	96.67	98.11	**99.07**	98.5	99.56	99.67	99.33
DRLBP	82.6	87.56	90.83	91.52	92.22	91.6	92.17	91.56	91.33	90.67
DSLGS	87.33	92.59	92.92	93.05	93.67	93.2	94	92.22	93	92
ELGS	87.07	92.44	93.42	93.52	94.22	93.07	94.33	93.33	93.67	94.67
IBGC-1	87.47	91.48	93.33	93.14	94.67	93.47	94.17	92	93	94
ILTP	87.13	91.93	93.17	92.95	93.89	92.93	95	94.44	93.33	96.67
LBP	89.53	93.85	95.25	95.33	96.78	95.6	96.33	95.56	95	97.33
LCCMSP	73.93	80.74	84.42	82.48	84	82.93	83.83	84.44	83	84
LDBP	88.13	92.44	94.33	94.38	95.67	94.53	94.67	93.33	93	92.67
LDN	85.73	90.81	92.83	93.71	94.33	94.53	95.17	94.44	95	96.67
LECTP	70.87	80.3	82.67	84.29	84.22	85.2	86.83	85.78	85	82.67
LETRIST	71.53	83.63	89.42	89.24	89.44	90.27	90.5	92.67	92	93.33
LESTP	75	83.48	85.42	86.29	86.56	87.87	89.83	88.22	86.33	87.33
LGCP	87	92.15	91.67	92.29	92.22	91.2	92.17	91.33	91.67	92.67
LGS	86.87	92.52	93.25	93.24	93.67	92.8	93.67	92.67	92.67	93.33
LMEBP	72.67	83.56	85.33	87.71	87.56	88.4	89.83	88.89	88.67	86
LNDP	75.13	84.67	89	89.24	89.22	89.87	90.67	88.89	90	94
LPQ	87.93	94.52	95.92	96.95	97.56	97.87	97.5	98	97	98
LQPAT	80	87.41	90.83	90.57	89.67	90.8	91.33	91.56	93	92.67
LTP	87.87	91.63	92.75	92	92.89	92.67	94.5	93.78	92.33	94
MMEPOP	84.2	91.85	95.25	96.19	96.44	98	97.67	97.56	97.33	97.33
MOW-SLGS	88.2	93.26	93.5	93.71	94.22	92.93	93.83	92.44	92.33	94
nLBPd	87.53	90.96	93.08	92.86	94.11	92.93	93.83	92.67	92.67	94.67
NI-CI-LBP	78.4	87.78	89.92	90.1	90.22	90.53	92.5	91.33	92.67	90.67
NI-LBP	79.73	87.48	90.25	90.19	90.56	90.4	91.67	90.89	93	90
NI-RD-LBP	80.53	88.07	90.33	90.76	90.78	90.4	91.67	89.78	92	88.67
QBP	82.4	89.85	90.92	90.48	91.22	91.47	91.83	90.44	89	89.33
RD-CI-LBP	79.47	88.3	90.08	91.14	91.56	91.6	93	92	93.33	90.67
RDLBP	80.73	88.52	90.17	91.05	92.11	91.07	92.83	90.89	94	90.67
SLGS	87.33	92.59	92.92	93.05	93.67	93.2	94	92.22	93	92
SMEPOP	85.93	90.81	92.08	92.38	93	93.6	93.83	92.67	93	93.33
WLD	66.47	75.7	79.58	81.33	83.67	82.8	85.83	84.89	84	82.67
XCS-LBP	90.33	96.52	97	97.9	98.56	98.4	99.33	98.44	97.33	98

Table 3.4: Average recognition accuracy (%) over 10 splits on Extended Yale B dataset (Exp#4)

Descriptor	Extended YALE B Train/Test				
	5/59	10/54	20/44	30/34	32/32
MNTCDP	87.78	94.42	**97.87**	**98.76**	**98.91**
AECLBP-M	49.62	60.8	71.72	76.67	79.07
AECLBP-S	78.96	90.41	96.24	97.41	97.84
AECLBP-S-MxC	60.7	75.99	88.49	93.16	94.57
AELTP	57.77	69.9	82.1	85.89	87.49
ALTP	77.06	89.2	96.01	97.43	97.88
BGC-1	75.04	88.67	95.9	97.5	97.84
BGC-2	76.43	89.78	96.21	97.6	97.84
BGC-3	76.92	89.66	96.21	97.41	97.78
CI-LBP	91.26	95.05	97.1	97.37	97.68
CLBC-M	47.26	57.7	65.36	69.87	72.3
CRLBP-M	50.34	60.93	73.08	78	80.65
CRLBP-S	69.09	82.85	92.7	95.68	96.59
CRLBP-S-MxC	57.34	72.62	85.42	91.24	92.42
CSALTP	76.68	89.52	96.24	97.52	97.74
DBC	79.59	88.93	96.16	97.64	97.8
DCLBP	66.13	79.36	91.02	94.4	95.86
DDBP	79.16	91.41	96.77	97.64	98.04
dLBPa	82.99	92.72	96.96	97.54	97.9
DRLBP	72.08	85.39	93.61	96.21	97.35
DSLGS	76.81	90.06	96.19	97.54	97.94
ELGS	75.88	89.65	96.16	97.49	97.88
IBGC-1	75.57	89.28	95.99	97.43	97.84
ILTP	54.49	65.73	76.5	81.03	83.13
LBP	75.74	89.05	95.69	97.49	97.8
LCCMSP	57.87	75.43	89.76	94.15	94.81
LDBP	79.16	91.41	96.77	97.64	98.04
LDN	78.87	89.71	95.9	97.22	97.76
LECTP	67.74	80.63	91.69	94.9	95.82
LETRIST	58.36	73.71	87.47	92.59	93.78
LESTP	63.04	75.9	87.63	91.64	92.93
LGCP	77.17	89.98	96.16	97.47	97.84
LGS	75.8	89.73	95.99	97.49	97.82
LMEBP	68.3	81.04	91.7	94.95	95.7
LNDP	71.63	85.24	94.31	96.84	97.62
LPQ	84.11	92.53	96.7	97.66	97.92
LQPAT	79.3	91.22	96.71	97.73	97.96
LTP	58.21	70.23	82.1	86.02	87.78
MMEPOP	75.98	87.44	94.8	96.86	97.49
MOW-SLGS	76.08	89.53	96.07	97.5	97.92
$nLBPd$	75.93	89.68	96.04	97.49	97.76
NI-CI-LBP	**92.39**	**95.36**	96.59	96.61	96.38
NI-LBP	86.68	94.76	97.47	97.9	98.08
NI-RD-LBP	84.84	92.41	95.41	96.25	96.04
QBP	59.04	73.27	85.69	90.13	91.56
RD-CI-LBP	90.04	92.96	95.39	95.9	95.8
RDLBP	80.5	85.39	92.06	93.62	93.76
SLGS	76.81	90.06	96.19	97.54	97.94
SMEPOP	73.41	87.62	95.24	97.09	97.58
WLD	36.86	26.48	34.34	39.24	37.76
XCS-LBP	58.58	74.07	87.63	91.14	92.59

Table 3.5: Recognition accuracy (%) on Extended Yale B dataset using subsets evaluation protocol (Exp#5)

Descriptor	Extended YALE B Subsets			
	Sub2	Sub3	Sub4	Sub5
MNTCDP	**100**	**100**	**99.8**	**98.5**
AECLBP-M	**100**	88.92	12.74	10.67
AECLBP-S	**100**	**100**	95.63	93.04
AECLBP-S-MxC	**100**	93.67	52.66	48.59
AELTP	**100**	95.51	23.57	9.93
ALTP	**100**	**100**	94.87	89.48
BGC-1	**100**	**100**	96.01	92.44
BGC-2	**100**	**100**	97.53	96.15
BGC-3	**100**	99.74	98.48	97.19
CI-LBP	**100**	98.94	96.01	82.52
CLBC-M	**100**	85.22	29.28	10.07
CRLBP-M	**100**	86.81	12.74	11.85
CRLBP-S	**100**	99.21	79.28	61.04
CRLBP-S-MxC	**100**	91.29	43.92	33.93
CSALTP	**100**	**100**	95.82	95.41
DBC	**100**	**100**	94.49	83.11
DCLBP	**100**	95.25	79.85	66.07
DDBP	**100**	**100**	97.15	94.81
dLBPa	**100**	98.42	92.02	88.59
DRLBP	**100**	98.68	82.89	68.74
DSLGS	**100**	**100**	97.72	95.85
ELGS	**100**	**100**	98.67	96.15
IBGC-1	**100**	**100**	97.91	94.22
ILTP	**100**	92.35	23.95	7.41
LBP	**100**	99.47	93.16	87.7
LCCMSP	**100**	97.25	87.36	76.44
LDBP	**100**	**100**	94.11	90.37
LDN	**100**	99.74	96.77	91.11
LECTP	**100**	97.89	46.96	35.7
LETRIST	**100**	93.40	58.37	35.56
LESTP	**100**	96.83	27.57	20.74
LGCP	**100**	99.74	97.15	96
LGS	**100**	**100**	98.29	96
LMEBP	**100**	99.21	44.49	35.26
LNDP	**100**	99.74	86.50	73.48
LPQ	**100**	**100**	99.05	98.07
LQPAT	**100**	**100**	97.34	97.48
LTP	**100**	94.72	28.71	11.56
MMEPOP	**100**	99.47	92.59	83.7
MOW-SLGS	**100**	**100**	97.15	94.22
$nLBPd$	**100**	**100**	97.53	96.59
NI-CI-LBP	**100**	99.74	98.67	94.81
NI-LBP	**100**	99.74	97.34	96.44
NI-RD-LBP	**100**	99.47	96.01	82.96
QBP	**100**	94.46	62.55	34.67
RD-CI-LBP	**100**	99.47	96.58	85.93
RDLBP	**100**	99.21	87.26	72.89
SLGS	**100**	**100**	97.72	95.85
SMEPOP	**100**	**100**	95.63	90.96
WLD	87.17	45.12	18.63	22.22
XCS-LBP	**100**	97.36	46.96	36.89

Table 3.6: Average recognition accuracy (%) over 10 splits on AR dataset (Exp#6)

Descriptor	AR Database Train/Test				
	5/21	10/16	13/13	15/11	20/6
MNTCDP	**82.33**	**93.53**	**96.18**	**97.37**	**99.28**
AECLBP-M	69.12	84.63	90.13	92.33	96.33
AECLBP-M	69.12	84.63	90.13	92.33	96.33
AECLBP-S	77.56	91.69	95.08	97.02	99.23
AECLBP-S-MxC	62.99	81.26	88.04	91.24	96.63
AELTP	71.03	86.69	91.65	94.15	97.87
ALTP	69.58	86.04	91.17	93.91	97.77
BGC-1	75.87	90.41	94.45	96.15	98.75
BGC-2	76.26	90.74	94.69	96.47	98.98
BGC-3	74.18	89.73	93.65	95.88	98.63
CI-LBP	57.9	73.66	77.5	80.93	85.28
CLBC-M	64.03	80.56	86.48	89.45	94.28
CRLBP-M	65.46	81.39	87.05	89.89	94.15
CRLBP-S	71.78	87.91	92.68	94.8	98.32
CRLBP-S-MxC	57.98	76.83	84.27	88.23	94.25
CSALTP	67.34	84.45	89.88	92.77	97.15
DBC	72.63	86.88	91.75	94.11	97.17
DCLBP	75.35	89.96	94.19	96.14	98.75
DDBP	78.33	91.78	95.33	96.78	99.07
dLBPa	77.49	90.2	93.95	95.45	98.4
DRLBP	74.86	88.94	93.7	95.38	97.82
DSLGS	78.3	91.95	95.43	96.97	99.15
ELGS	76.99	91.28	94.92	96.75	99.1
IBGC-1	77.02	90.65	94.55	96.27	98.72
ILTP	69.91	85.49	90.92	93.32	97.1
LBP	75.58	90.17	94.32	96.15	98.75
$nLBPd$	76.27	90.67	94.61	96.49	98.97
LCCMSP	55.97	76.03	83.42	87.21	94.18
LDBP	77.72	91.52	95.15	96.66	99.02
LDN	79.29	91.99	95.62	97	99.1
LECTP	61.66	77.54	83.76	86.84	92.08
LESTP	56.48	72.96	79.29	82.51	88.9
LETRIST	59.28	77.19	84.08	87.57	92.77
LGCP	73.46	89.14	93.51	95.51	98.48
LGS	76.97	91.32	94.92	96.73	99.08
LMEBP	61.38	77.46	83.8	86.6	92.12
LNDP	77.2	90.81	94.3	96.17	98.73
LPQ	79.57	91.79	95.22	96.49	99.03
LQPAT	77.07	91.24	94.7	96.44	98.97
LTP	69.15	84.92	90.36	93.13	97.02
MMEPOP	77.33	90.09	94.02	95.43	98.43
MOW-SLGS	77.56	91.48	95.2	96.89	99.08
NI-CI-LBP	78.32	91.11	94.32	96.01	98.47
NI-LBP	78.2	91.49	94.65	96.35	98.67
NI-RD-LBP	78.34	91.8	94.81	96.57	98.92
QBP	63.52	81.45	87.06	90.32	95.47
RDLBP	77.85	91.27	94.7	96.34	98.75
RD-CI-LBP	78.04	91.04	94.18	95.99	98.43
SLGS	78.3	91.95	95.43	96.97	99.15
SMEPOP	73.08	88.77	93.16	95.2	98.55
WLD	57.73	73.42	79.84	83.25	88.55
XCS-LBP	67.25	82.79	88.12	90.84	95.33

Table 3.7: Comparison with PCANet2 and CBFD deep learning methods

Database		MNTCDP 10 Splits avg acc	PCANet2 1 Split acc	CBFD 1 Split acc
ORL Database Exp#1 configuration	1/9	85.89	77.5	85
	2/8	95.59	88.13	87
	3/7	98.64	94.29	95
	4/6	99.67	100	97
	5/5	100	99	100
	6/4	100	98.75	100
	7/3	100	100	100
	8/2	100	100	100
	9/1	100	100	100
FERET Database Exp#2 configuration	1/6	61.57	62.4	67
	2/5	83.56	81.61	85
	3/4	92.21	25	93
	4/3	96.48	53	97
	5/2	98.52	97.99	97
	6/1	99.7	99.5	99
YALE Database Exp#3 configuration	1/10	91.53	80.67	99.12
	2/9	96.59	94.07	99.45
	3/8	97.58	93.33	99.33
	4/7	98.29	90.48	98.1
	5/6	99	7.78	99.36
	6/5	98.93	8	100
	7/4	100	11.67	98.74
	8/3	100	6.67	98.82
	9/2	100	10	98.97
	10/1	100	13.33	100
Extended YALE B Exp#4 configuration	5/59	87.78	82.91	88.36
	10/54	94.42	92.89	95.12
	20/44	97.87	94.29	81.85
	30/34	98.76	94.55	97.26
	32/32	98.91	71.11	97.5
AR Database Exp#6 configuration	5/21	82.33	96.29	94
	10/16	93.53	98.69	96
	13/13	96.18	99.08	98
	15/11	97.37	98.45	99
	20/6	99.28	98.5	97
Accuracy over the test subsets, Subset 1 is used for training phase				
Extended YALE B Subsets Exp#5 [52]	Sub2	100	100	100
	Sub3	100	99.47	98
	Sub4	99.81	92.21	70.34
	Sub5	98.52	63.11	54.03

Table 3.8: Comparison with recent state-of-the-art systems on ORL, FERET, YALE, EYB(Exp#4) and AR databases.

Database	**Ref.** (Publication & Year)	**Recognition Rate** (Train-Test) [Number of splits]
ORL Exp #1	[119] (Neuro 2016)	89.5 (5-5) [1 Split]
	[120] (ESWA 2016)	97.62% (5-5) [1 Split]
	[121] (Neuro 2016)	98.52% (8-2) [1 Split]
	[115] (IEEE Access 2017)	96% (5-5) [1 Split]
	[116] (Neuro 2016)	100% (8-2) [1 Split]
	[122] (NCA 2017)	97% (5-5) [1 Split]
	[123] (IEEE TC 2016)	91.35% (5-5) [10 Splits]
	[122] (NCA 2017)	97.50% (6-4) [1 Split]
	[124] (JKSUCIS 2017)	99.58% (-) [1 Split]
	[125] (Vis.Com.Im.R 2015)	89% (5-5) [1 Split]
	[117] (PR 2016)	98% (5-5) [20 Splits]
	[126] (IEEE Access 2014)	85% (5-5) [1 Split]
	[127] (IEEE SPL 2015)	91.5% (5-5) [1 Split]
	[128] (PR 2014)	99% (6-4) [1 Split]
	[129] (KBS 2015)	97.5% (5-5) [1 Split]
	[130] (FGCS 2017)	94.60% (2-8) [1 Split]
	Proposed MNTCDP based system	**100% (5-5) [10 Splits]**
FERET Exp #2	[119] (Neuro 2016)	65.4% (2-5) [1 Split]
	[115] (IEEE Access 2017)	90.7% (6-1) [1 Split]
	[116] (Neuro 2016)	92.25% (5-2) [1 Split]
	[131] (IEEE TCE 2012)	95.40% (6-4) [1 Split]
	[132] (IEEE SPL 2013)	91.2% (3-1) [-] 250 Subjects
	[128] (PR 2014)	98% (3-4) [1 Split]
	[133] (Optik 2016)	93.33% (-) [1 Split]
	[117] (PR 2016)	73.75% (3-1) [20 Splits]
	[134] (PR 2014)	95.85% (5-2) [20 Splits]
	[135] (IEEE TCSVT 2011)	69% (4-3) [35 Splits]
	[136] (EAAI 2017)	84.50% (5-2) [1 Split]
	[129] (KBS 2015)	92.5% (4-3) [1 Split]
	Proposed MNTCDP based system	**99.7% (6-1) [10 Splits]** **98.52% (5-2) [10 Splits]**
YALE Exp #3	[133] (Optik 2016)	45.33% (1-10) [1 Split]
	[121] (Neuro 2016)	88.56% (8-3) [1 Split]
	[137] (EUVIP 2016)	75.56% (8-3) [1 Split]
	[138] (IEEE Access 2016)	75% (-) [1 Split]
	[122] (NCA 2017)	96% (5-6) [1 Split]
	[126] (IEEE Access 2014)	86% (6-5) [1 Split]
	[139] (Neuro 2012)	90.7% (6-5) [1 Split]
	[140] (Optik 2013)	83.6% (1-10) [1 Split]
	Proposed MNTCDP based system	**100% (7-4) [10 Splits]**
EYB Exp #4	[123] (IEEE TC 2016)	86.55% (32-32) [10 Splits]
	[141] (IEEE TIP 2017)	81.38% (5-49) [1 Split]
	[142] (IEEE TIP 2016)	93.4% (32-32) [1 Split]
	[143] (IEEE TM 2017)	89.65% (13-51) [1 Split]
	[144] (PR 2016)	90% (-) [1 Split]
	[145] (IEEE TNNLS 2015)	93.09% (32-32) [1 Split]
	[122] (NCA 2017)	74.34% (-) [1 Split]
	[146] (SP:ICom 2017)	96.26% (32-32) [10 Split]
	[147] (Vis.Com.Im.R 2017)	92.41% (16-48) [1 Split]
	Proposed MNTCDP based system	**98.91% (32-32) [10 Splits]**
AR DB Exp#6	[148] (PR 2017)	98.97% (20-6) [10 Splits]
	[149] (PR 2017)	98.31% (13-13) [1 Split]
	[146] (SP:ImCom 2017)	97.23% (13-13) [10 Split]
	[147] (Vis.Com.Im.R 2017)	93% (7-7) [1 Split]
	[150] (Neuro 2017)	81.39% (9-17) [3 Split]
	[151] (Neuro 2018)	99% (20-6) [1 Split]
	Proposed MNTCDP based system	**99.28% (20-6) [10 Splits]**

Table 3.9: Comparison with recent state-of-the-art systems on EYB database using subset evaluation protocol.

Ref. (Publication \& Year)	Extended YALE B Subsets			
	Sub2	Sub3	Sub4	Sub5
[52] (ESWA 2016)	**100**	**100**	94	95
[40] (IEEE TIP 2012)	**100**	**100**	-	-
[153] (Info.Sc 2016)	99.8	99.6	99.4	97.8
[154] (PR 2017)	-	98	94	93
[155] (PR 2017)	**100**	99.12	96.99	97.74
[156] (IEEE TIF&S 2016)	**100**	99.54	93.89	93.17
[157] (DSP 2015)	**100**	91.42	83.65	85.71
[152] (PR 2017)	**100**	99.12	99.44	**98.89**
[158] (IEEE SPL 2015)	**100**	**100**	94.55	91.14
[130] (FGCS 2017)	**100**	**100**	95	95
Proposed MNTCDP based system	**100**	**100**	**99.81**	98.52

3.6/ Implementation in the Human Support Robot (HSR) of UTBM

The UTBM takes part of the HSR Developers Community with an HSR named Eighty-Eight, worth a value of 60,000 €through the partnership with Toyota. To promote the autonomy of the elderly or disabled through home assistance, Toyota has joined forces with research institutes to create the HSR Developers Community - a cooperation that will accelerate the development and commissioning of the HSR robot. Thanks to its light, compact, and very maneuverable cylindrical body with a folding arm, it can grab objects on the floor or placed them on shelves and perform various tasks. In addition to its local control, the robot can be remotely controlled by family or friends, the operator's face and voice being relayed in real-time to offer fundamental human interactions while helping with daily tasks. Since its first presentation in 2012, the HSR has received several improvements based on feedback from patients and caregivers.

One of the objectives of this thesis, which we discussed in the first chapter, is embedding the developed facial analysis frameworks within end devices such as HSR. Indeed, we programed Eighty-Eight to host the MNTCDP-based face recognition system so it will be capable of recognizing our lab members. To do so, we collected our lab database of 10 subjects with 50 images each to use as references. The 50 samples are covering different backgrounds and lighting conditions. The collection was made using the HSR integrated wide camera, and the framework was developed in a python3 environment. Figure 3.17 shows a complete view of the HSR robot and its embedded cameras, while Figure 3.18 illustrates some samples from the collected database. The recognition performance on

this dataset was ideal and confirmed the one that the MNTCDP reached on state-of-the-art ones. We included some post-processing programs to make Eight-Eight capable of telling the name of the recognized person.

Figure 3.17: Camera system of Toyota HSR.

Figure 3.18: Images of subject from the collected database

.

3.7/ CONCLUSION

Based on combining two different neighborhood sampling concepts: the direction and radius, Mixed Neighborhood Topology Cross Decoded Patterns (MNTCDP), proved to

be a conceptually and computationally efficient yet straightforward descriptor for face image modeling, is introduced in this chapter. MNTCDP makes effective the use of micro-structures and relationships between pixels within 5×5 window. It encodes micro-structures of image patterns in the most informative directions within a face image. In the objective of investigating the performance of the proposed MNTCDP descriptor, a comprehensive evaluation is performed on five challenging representative widely-used face datasets, covering the experimental evaluation according to illumination changes and the number of samples used in the train set, in addition to face recognition of facial images with sunglass and scarf occlusions. Comprehensive and systematic performance comparison with a large number of state-of-the-art texture descriptors, deep learning methods, and several recent state-of-the-art face recognition systems on wide world used databases demonstrate the efficiency of the proposed MNTCDP model and its significant performance improvements in face recognition over the evaluated LBP-like descriptors, deep learning methods and existing face recognition approaches.

Furthermore, the performance of the proposed MNTCDP descriptor is recorded using only the simple nearest neighbor classifier, requiring few training data as well as low processing time. These strengths give an open window to handcrafted descriptor approach to reach new achievements by motivating the research community to develop new LBP variants and feed the need to propose potentially useful handcrafted descriptors for face recognition and other applications.

4

GAN-BASED PROFILE FACE RECOGNITION

4.1/ INTRODUCTION

Back in 2014, the deep learning community experienced the proposal of a new kind of architecture, which serves to generate fake images based on conditional inputs referred to as Generative Adversarial Networks (GAN) proposed by Goodfellow [159]. Then, the GANs have been widely used to resolve many challenges [160; 161; 162], generating newly synthesized datasets [163]. Inspired by the GAN success, we proposed a fully automatic PIFR framework based on a paired GAN translation of non-frontal input images. Afterward, the generated frontal image is fed to the face recognition network based on ResNet architecture. As shown in Figure 4.1, the overall pipeline can be divided into two subsystems: Frontalization bloc and Recognition bloc, where the first translates the input image into a frontal one and the second tries to classify the generated image to recognize the subject. To train the GANs and assess the face recognition performance of our pipeline, we collected a database from four existing benchmarks: ColorFERET, KDEF, RaFD, and FEI. The collected dataset is divided into two subsets: training the GAN and the second to train the face recognition subsystem based on ResNet architecture. Note that the two subsets are person-independent to ensure that the GANs are evaluated on unseen subjects. Moreover, the ResNet for face classification is only trained on profile images of the subjects included in the second subset. The main contributions of this work can be highlighted in the followings:

- End-to-end Pose Invariant Face recognition framework based on Generative Adversarial Networks image translation.
- The poses considered by our framework are up to $90°$ (full profile), which makes our system more robust and generic.

- Evaluation of multiple GAN architectures and Residual CNN face classifiers.
- Collection of a comprehensive dataset to train and evaluate our framework respecting the person-independent constraint. The database will be shared with the state-of-the-art for future research.

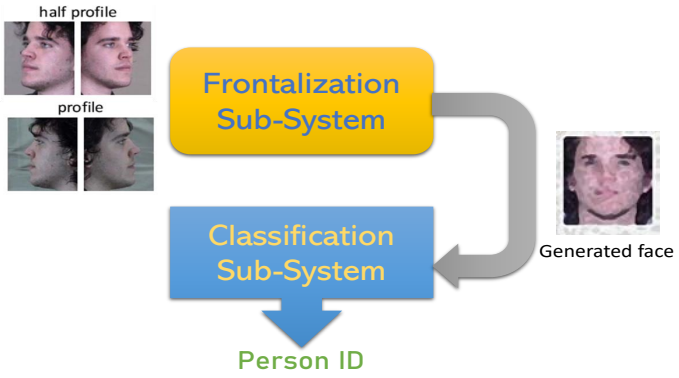

Figure 4.1: The overall pipeline of the proposed framework.

4.2/ FACE POSE TRANSLATION LITERATURE REVIEW

GAN-based face frontalization is a quite new solution to deal with PIFR challenge, and it consists of 2D face synthesis by training a deep neural network to generate front faces. However, only a few pre-print works are found in arXiv that are not fully peer-reviewed and are evaluated only for the face verification task. The first work was published in 2017 referred to as TP-GAN [164], which is adopting two generators: Global Pathway (GP) and Local Pathway (LP), devoted to global and local features that are a quite concept used previously for 2D/3D local texture warping [165; 166]. The LP generator takes as input the sub-images centered on the eyes, nose, and mouth to synthesize their corresponding sub-images in frontal view, then performs position aggregation. Meanwhile, the GP processes the whole face image, and its output is concatenated with LP output to form the final synthesized frontal image that will be fed to the discriminator network. However, this work lacks many explanations about how the LP generator works since the landmarks cannot be all detected in the case of $90°$. Therefore, the number of extracted sub-images is not consistent. Moreover, there is only one discriminator for the two pathways, which is an issue since the outputs of the two pathways are not correlated, and the local one generates more images than the global one. Also, the weights of the local generator cannot be optimized because it has no direct feedback from the discriminator. Besides, the two pathways are paired generators, and each has different inputs from the other. Hence two independent subsets should be prepared for training, and each generator must be trained independently. Also, for evaluating, the authors did

not adopt a person-independent protocol to train their TP-GAN. Hence the model will also be optimized on identities from the test set. Moreover, they adopted LightCNN [167], a state-of-the-art model for face verification, to assess the performance on the baseline images and the synthesized ones, which helps to highlight the improvement of the TP-GAN. This work has inspired other researchers to adopt the same philosophy to serve face frontalization. Both DA-GAN [168] and M2FPA [169] employed the Attention Guided GAN originally proposed in [170], which introduced the capability of GANs to focus on particular regions of the input image based on heat or binary maps. This ability has been explored by [168] and [169] in the same way computing the U-Net generator loss according to three generated masks: hair, landmarks, and skin. These masks are obtained from the generated frontal image by a state-of-the-art face parser [171]. The bottleneck of this approach is the calculated masks since the generated image it is not clear at the early epochs of the training. Hence the parser will not be capable of detecting the three masks, and as a result, the back-propagation loss would not be precise to optimize the generator, which may derive the system into the collapsing mode, a well-known weakness of GANs. Moreover, [168] and [169] adopted the same evaluation protocol and LightCNN model as [164] trying to demonstrate the improvement of their frontalization techniques in case of face verification. However, the reported results demonstrate that the performance of the LightCNN on the baseline images (original) is similar to the one reached on frontalized poses and the GAN improvement is very slight. For example on MultiPie dataset [172], the LightCNN reached 97.71% on 45° pose probe images, while [164], [168], and [169] reached 95.38%, 99.53% and 99.15%, respectively. Hence, the improvement was less than 2%, and also TP-GAN impacted the performance of the LightCNN. Furthermore, all three works lack proper comparison of the generated images with ground truth to evaluate the quality of the proposed GANs based on quantitative metrics such as MSE and SSIM that are widely used to assess the performance of GANs. Another issue observed in these works is the inconsistency of the LightCNN baseline results, however, adopting the same evaluation protocol and model weights. Indeed, on MultiPie 90° test set, the TP-GAN authors reported only 9.00%, but we found that it reached 66.08% as reported in the DA-GAN [168]. Also, no execution time evaluation was that is important for such applications imposing real-time constraint.

4.3/ PROPOSED GAN-BASED PIFR FRAMEWORK

We propose to deal with the PIFR challenge by exploring the benefits of GANs to generate identity preserving frontal images of poses reaching full profile (90°). This section is arranged into subsections to comprehensively present the overall framework, the GAN architectures, and the ResNet face-classification subsystem.

4.3.1/ OVERALL FRAMEWORK

Our system detects the pose angle of the input image then selects the appropriate GAN generator according to the estimated angle. Afterward, the generated frontal image is fed to the face recognition network based on ResNet architecture. We adopted two GANs denoted as *HP-GAN* and *FP-GAN* to deal with the poses around $45°$ (half-profile) and the ones close to $90°$ (full-profile), respectively. This makes the frontalization more fast and efficient since each GAN will be trained on fewer images but correlated ones. Also, only one generator is loaded each time depending on the estimated angle, which keeps the same computational cost of a framework with a generator for all the poses. We adopted two paired GANs (one for each pose) to maintain the person related features

Each GAN is selected to serve the frontalization of a given input image within its range. To do so, a pose angle detector is estimates the angle, we used the Dlib [173] package that proved to be highly fast and accurate. The input images with pose angle less than $20°$ are considered as frontal images since they present all the visual features revealing the person identity, angles between $20°$ and $60°$ are processed as half-profile while the remaining (from $60°$ to $90°$) are considered as full-profile. Once the input image is frontalized, we feed it to the face classifier based on ResNet architecture that is trained only on frontal samples.

The overall pipeline is illustrated in Figure 4.2. As can be seen, the system contains frontalization and classification blocks. The first one includes the Dlib based pose estimation and the two generators (HP-GAN and FP-GAN), while the second block is charged to identify the label (identity) of the person based on the generated image. The frontalization and classification steps are summarized in Algorithm **1**.

Algorithm 1 The process of the proposed PIFR framework

Input: Input Image Im
Output: Subject ID
Functions: Dlib: estimate the pose angle; **Gen**: trained Generators HP and FP; **Classifier**: trained classifier
read the input image; calculate the pose angle $\alpha = Dlib(Im)$;
if $\alpha \leq 20°$ **then**
  the image is detected as frontal classify the input image without frontalization ID =**Classifier**(Im)
if $20° < \alpha \leq 60°$ **then**
  the image is detected as half-profile half-profile frontalization Im' =**Gen**$^{HP}(Im)$ classification of the frontalized image ID =**Classifier**(Im')
if $60° < \alpha \leq 90°$ **then**
  the image is detected as full-profile full-profile frontalization Im' =**Gen**$^{FP}(Im)$ classification of the frontalized image ID =**Classifier**(Im')

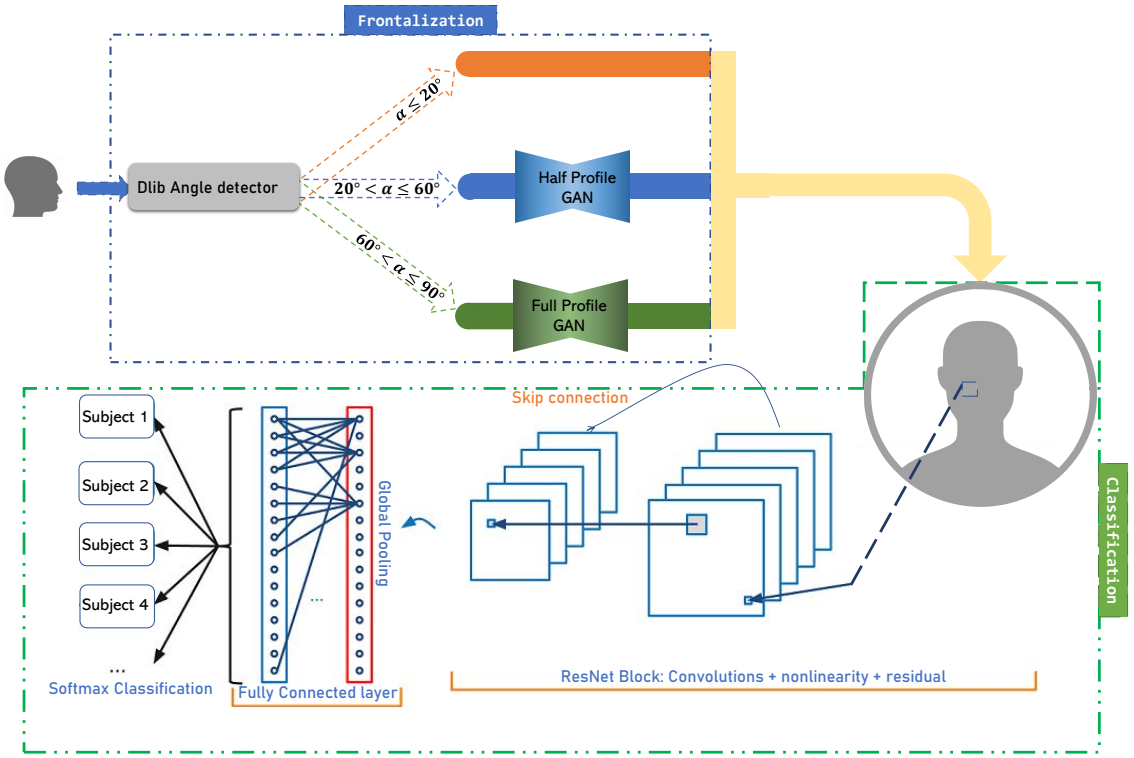

Figure 4.2: The overall architecture of the proposed PIFR framework based on GAN image translation

4.3.2/ GENERATIVE ADVERSARIAL NETWORKS

The proposed IFR framework is based on GAN image translation, which translates an input image to a target space based on a trained generator. The GAN consists of a generator and discriminator convolutional networks. The first one is an auto-encoder that tries to encode the profile input image Im^P into a reduced latent space z and then generates an image $\mathbf{G}(Im^P)$ that looks like the frontal one Im^F. The discriminator network helps the generator to synthesize images similar to Im^F. The discriminator is a pixel classification network trained to identify real and fake images and gives the feedback to the generator to optimize its weights to generate real images similar to Im^F. It supervises the generator and judges the quality of the generated image $\mathbf{G}(Im^P)$. Therefore, at each epoch (batch) the discriminator is the first to be optimized, then the generator network. We summarized in Algorithm **2** the steps of training a GAN. Note that only the trained generator is needed in the test stage, which requires low computational cost. The GAN is similar to any deep neural network in training based on the backpropagation process and objective loss function G^*. Referring to the original work [159], the generator tries to fool the discriminator by producing fake images looking like the real ones, while the discriminator seeks to differentiate the fake from the real images correctly. This process is a min-max game as Eq 4.1 formulates, where the generator tries to minimize the loss

of produced frontal image detected as fake by the discriminator, which maximizes its performance to differentiate real from fake.

$$G^* = arg \min_G \max_D ([log(\mathbf{D}(Im^F)) + log(1 - \mathbf{D}(\mathbf{G}(Im^P)))]) + \lambda ||Im^F - \mathbf{G}(Im^P)||_1 \quad (4.1)$$

λ is a regularization parameter assigned to the L1 loss included in the objective function G^* to help the generator produce less blurry images. The L1 loss is used to update the generator parameters through the adopted optimizer. In our work, we investigated four generator networks and two discriminator networks widely used in the state-of-the-art.

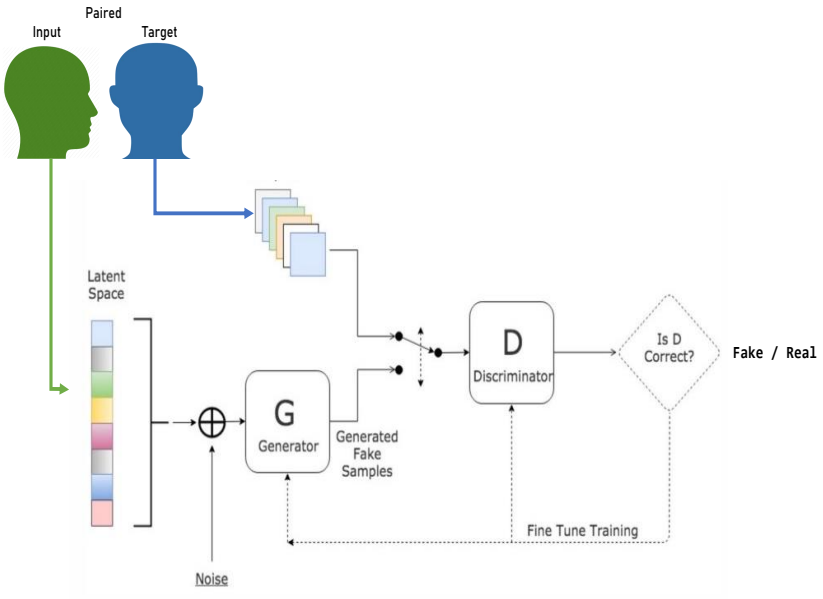

Figure 4.3: Paired GAN used for frontalization of profile image

4.3.2.1/ GENERATOR MODELS

The main autoencoder architectures used for the generator part are U-Net256, U-Net128, ResNet-6Blocks, and ResNet-9Blocks. The U-Net network was proposed in [174] for medical imaging purposes. The main idea is to supplement the downsampling (encoding) convolutional network by symmetric layers (decoding) where pooling operations are replaced by upsampling. Hence these layers increase the resolution of the encoded image with more precision thanks to the skip connections. Moreover, in U-Net, there are many feature channels in the upsampling part, which allow the network to propagate context information to higher resolution layers. Consequently, the expansive path is more or less symmetric to the encoding part and yields a u-shaped architecture. To predict the pixels in the border region of the image, the missing context is extrapolated by mirroring the input image. All the U-Net variants share the same pipeline, the only difference is the

Algorithm 2 GAN training process for the PIFR proposed framework

Input: Paired Input Images $\{Im^F, Im^P\}$
Output: Trained Generator **G**
Functions: optim: optimizer to update the networks; **D**: discriminator network;
for N epochs **do**
  for N/m steps **do**
    sample a minibatch of m frontal samples $\{Im_1^F, Im_2^F, ..., Im_m^F\}$ sample a minibatch of m profile samples $\{Im_1^P, Im_2^P, ..., Im_m^P\}$
    translate the m profile images to frontal ones $\{\mathbf{G}(Im_1^P), \mathbf{G}(Im_2^P), ..., \mathbf{G}(Im_m^P)\}$
    calculate the ascending loss gradient of the discriminator: $\nabla_{\mathbf{D}} \frac{1}{m} \sum_{i=1}^{m} [log(\mathbf{D}(Im_i^F)) + log(1 - \mathbf{D}(\mathbf{G}(Im_i^P)))]$
    such that $\mathbf{D}(Im_i^F) = true$ and $\mathbf{D}(\mathbf{G}(Im_i^P)) = fake$ update the discriminator parameters **D**=**optim**(**D**)
    calculate the descending L1 loss gradient of the generator: $\nabla_{\mathbf{G}} \frac{\lambda}{m} \sum_{i=1}^{m} \|Im_i^F - \mathbf{G}(Im_i^P)\|_1$ update the generator parameters **G**=**optim**(**G**)

supported size of the images controlled by the amount of GPU memory and most of the works adopt 256 and 128 resolutions. Figure 4.4 illustrates the architecture of U-Net256 and U-Net128 generators as they share the same u-shape with different filter sizes and numbers. On the other hand, ResNet-based generator adopts residual blocks to compute relevant features from the input image, exploring the distinguishing power of the original ResNet [175]. Hence, the input image is downsampled n times, where n is generally set to 2 or 3, then the resulted from convoluted feature maps are further processed by residual sub-network referred to as ResNet-Block. Afterward, the output of the ResNet-Blocks is upsampled to reach the specified size of the generator output. Figure 4.5 shows the architecture of ResNet-based generator with 6 and 9 blocks. The number of intermediate blocks can be adjusted according to the application needs. However, [176] reported that 6 and 9 blocks provide good generation. Also, a high number of blocks risks the gradient vanishing in addition to high computational cost.

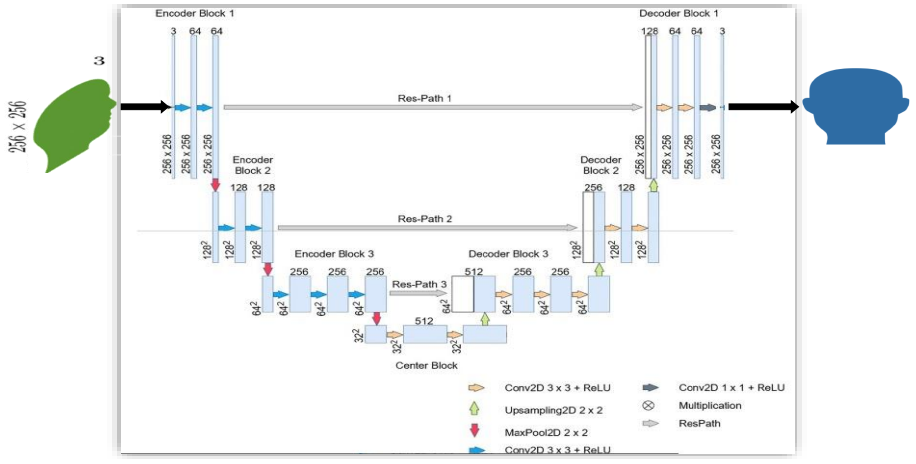

Figure 4.4: The architectures of U-Net generators as reported in [174]

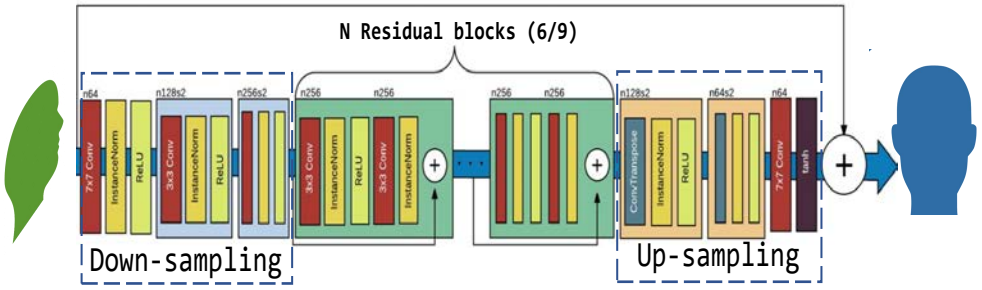

Figure 4.5: The generic architecture of ResNet-based generators

4.3.2.2/ DISCRIMINATOR MODELS

The discriminator is charged to supervise the generator network and differentiate the real images from the fake ones produced by the generator. Therefore, the discriminator is a binary pixel classification network including few convolution layers. The discriminator model adopted in the Pix2Pix GAN [176] is implemented as PatchGAN or Pixel discriminator; both discriminators share the same architecture that is illustrated in Figure 4.6. In contrast, the first classifies 70×70 patches, and the second classifies each pixel of the input image (generated/real). Hence in general, the discriminator has a $N \times N$ classification network that is processed convolutionally across the image to calculate the loss of each patch (non-overlapping blocks/pixel wise in case $N = 1$). All the responses are averaged to provide the overall loss, which is considered to update the network weights through the optimizer. According to [177], a smaller PatchGAN has fewer parameters, runs faster, and can be applied to arbitrarily large images.

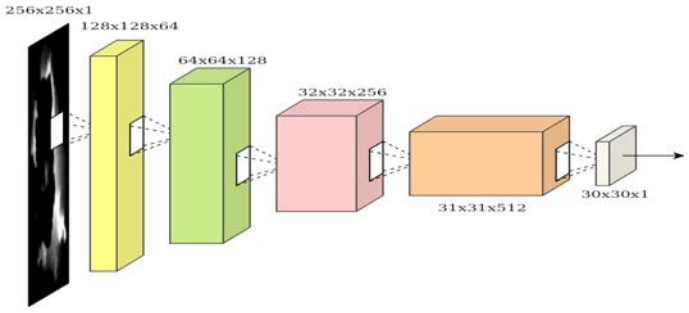

Figure 4.6: The layers used to build the discriminator

4.3.3/ CNN BASED FACE CLASSIFICATION

Our second sub-system is dedicated to identifying the subject of the frontalized input image. We adopted a CNN-based classification architecture that is very basic and common with the state-of-the-art to assess the improvement of the frontalization sub-system. With the development of deep learning, CNN (Convolutional Neural Network) has become the

main method adopted in the field of face recognition. CNN typically consists of convolutional layers, pooling layers, and fully connected layers. Convolutional layers are core building blocks of CNN. The size of the last fully connected layer is fixed by the number of the classes composing the database, which is, in our case, the number of persons of the training dataset as can be shown in Figure 4.2, that illustrates the overall pipeline of a Deep-based face recognition framework. Many CNN models have been proposed in the literature, however, residual-based ones proved to be the most effective models with low computational cost compared to the rest. Therefore, we adopted to evaluate different ResNet depths and the DenseNet model, which uses the residual connection in a similar way to ResNet.

Figure 4.7 shows the difference between ResNet and DenseNet in terms of skip connection, where the DenseNet network concatenates the outputs of the convolution layers all together before proceeding to the classification layer. This fact could lead to a heavy amount of data to deal with and hence more computational cost. On the other hand, ResNet sums these outputs that may reduce the feature extraction power since many sum combinations can have close values.

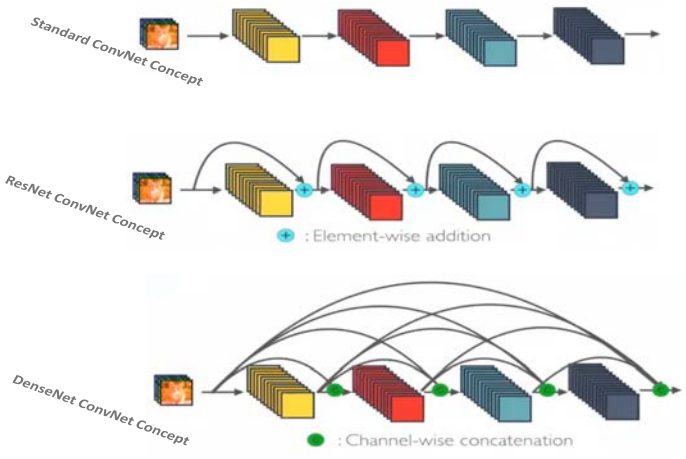

Figure 4.7: Comparison of ResNet and DenseNet residual connections

4.4/ EXPERIMENTAL ANALYSIS AND DISCUSSIONS

This section discusses the evaluation of the proposed PIFR framework based on GAN facial image frontalization. It presents the used database for training and evaluation through a comprehensive person-independent protocol, evaluating the different GANs models based on MSE and SSIM metrics and the overall improvement on the recognition performance. Moreover, this section includes computation time analysis.

4.4.1/ DATABASE

The performance of deep architecture highly depends on the dataset used for training. However, the proposed PIFR datasets are very few, and most of them are no longer available such as the MultiPIE dataset, which is the widely used one. Therefore, we were faced with finding another alternative to train and evaluation our PIFR framework. We proposed to combine 4 state-of-the-art databases so we can build a consistent and challenging benchmark referred to as the Combined-PIFR database. We adopted ColorFERET, FEI, RaFD, and KDEF databases, which provide frontal, Half-Profile, and Full-Profile samples. In the following, we present the properties of each database.

- **ColorFERET database [113]:** The Facial Recognition Technology (FERET) database is known as the first benchmark used to evaluate face recognition frameworks. Regarding the resources and materials devoted to design this database, the creators managed to simulate illumination, head pose, and aging challenges. The images were collected in 15 sessions between August 1993 and July 1996, accumulating a total of 14,126 images representing 1199 individuals.

- **FEI Face database [178]:** The Brazilian FEI face database has been collected between June 2005 and March 2006 at the Artificial Intelligence Laboratory of FEI in Brazil. It includes a set of 2800 images covering 200 individuals, each represented over 14 images. All images, which have a common white homogeneous background, are colorful and have a resolution of 640×480 with a variable scale of 10%. The main challenge of this database is the profile rotation, which varies from 0 to $180°$. Moreover, this database offers age and gender variability by including students and staff at FEI lab with equal female/male partition, between 19 and 40 years old with a distinct appearance, hairstyle, and adorns.

- **KDEF Face database [179]:** The Karolinska Directed Emotional Faces (KDEF) database contains 70 individuals, each is displaying 7 different emotional expressions. The changes in the facial landmarks can affect enough the appearance of the face, hence the person will not be correctly recognized, which challenges the face recognition systems and assesses their robustness to the appearance changes. Also, the database provides frontal, half-profile, and full-profile samples for all the subjects.

- **RaFD [180]:** The Radboud Faces Database (RaFD) is composed of 67 individuals, including Caucasian, Moroccan, Dutch adults, and Caucasian children, both boys and girls, and displaying 8 emotional expressions across multiple head poses (frontal to full profile). The challenge of this database relies on the variety of the subjects and their ages.

After presenting the properties of the considered datasets for making the Combined-PIFR database, we discuss the evaluation protocol that we followed to train and evaluate the proposed PIFR framework, which is composed of a GAN-frontalization subsystem and a CNN based classifier. These two subsystems must be trained respecting the person-independent constraint so the GAN can fairly generate frontal samples of unseen subjects. Figure 4.8 explains in-depth how we performed the dataset partitioning and the amount of data dedicated to train each subsystem and to evaluate the overall framework. For training the GANs, we selected only one pair of images per subject for each angle ($45°$ and $90°$) to make the GANs more generic and less dependent on the persons. The *HP-GAN* and *FP-GAN* are trained on 1000 pairs each, whereas the face classification network is trained and evaluated on a set of 155 subjects not belonging to the subset used to train the GANs. For each subject, we took only 4 frontal images for training and from 2 to 8 images (depending on the original database) for testing that cover half and full profile angles leading to a total of 1004 probe images, which is big enough to fairly assess the performance of the framework.

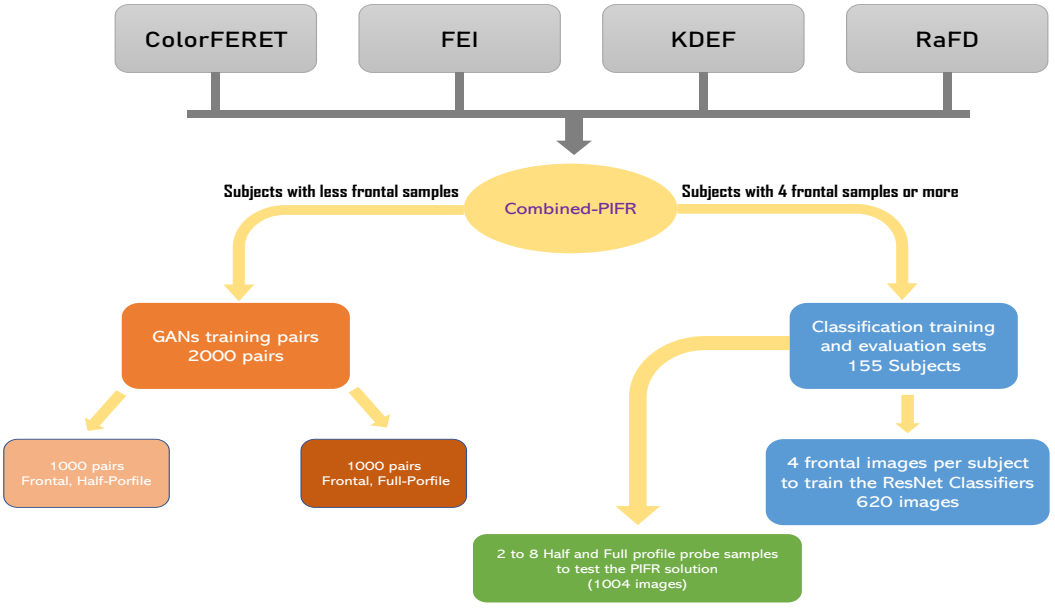

Figure 4.8: Construction and partitioning of the Combined-PIFR database

4.4.2/ GAN'S ARCHITECTURES EVALUATION

The adopted GAN in our PIFR framework supports 4 generator's models and 2 models for the discriminator that are presented previously. In this subsection, we investigate the performance of all the possible combinations of these networks (8 combinations) for *HP-GAN* and *FP-GAN* by calculating the MSE and SSIM values between the generated images and their ground-truths.

- Mean Squared Error (MSE): As 2D-image is a digital signal, we can assess the quality based on representing the average of the square errors between an image and its target (frontalized image and frontal ground-truth).

$$MSE = \frac{1}{m\ n\ c} \sum_{k=1}^{c} \sum_{i=1}^{m} \sum_{j=1}^{n} \|\mathbf{G}(Im^P(i,j,k)) - Im^F(i,j,k)\|^2 \tag{4.2}$$

  Where m, n, and c serve as the number of rows, columns, and channels, respectively. $\mathbf{G}(Im^P(i,j,k))$ is the pixel value of the frontalized profile input that is compared to the ground-truth one $Im^F(i,j,k)$.

- Structure Similarity (SSIM): This measure is a perceptual metric that quantifies image quality degradation caused by processing such as data compression or by losses in data reproduction such as our PIFR proposed solution. SSIM incorporates important structural information (luminance and contrast), meaning that the nearby pixels have strong inter-dependencies and carry information about the structure of the objects in the visual scene. Luminance tends to be less visible in bright regions, while contrast becomes less visible where there is significant activity in the image. SSIM ranges from 0 to 1, the higher the better.

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{x,y} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{4.3}$$

Where $\{\mu_x\ ,\sigma_x\ \}$, $\{\mu_y\ ,\sigma_y\ \}$ are the {mean, variance} values of the frontalized profile input $\mathbf{G}(Im^P)$ and ground-truth Im^F, respectively while $\sigma_{x,y}$ is the covariance. c_1 and c_2 are two constants serving to stabilize the fraction as $c_1 = (0.01 \times L)^2$ and $c_2 = (0.03 \times L)^2$, where L is the range of pixels that typically takes 255.

Table 4.1 lists the calculated MSE and SSIM values. It appears that the Pixel discriminator allowed the majority of the generators to achieve the highest results and then good generated images compared to the 70×70 PatchGAN, which is quite straightforward since the Pixel discriminator is more precise than the PatchGAN by checking the generation quality of each pixel. Moreover, the full-profile $(90°)$ challenged the GAN more as the metrics are bit lower than the ones of $45°$ with 62.41% SSIM and $4.187\ e4$ MSE vs. 71.19% SSIM and $3\ e3$ MSE.

From Figure 4.9, which illustrates the frontalization quality of the 8 evaluated GAN combinations, it appears that U-Net 256 served more clear generation compared to ResNet-9B and ResNet-6B, however, their MSE and SSIM metrics were close. Moreover, the PatchGAN discriminator struggled and could not help each of the generators to produce a complete shape of the face. Therefore, the U-Net 256 with Pixel discriminator managed to be the best performing combination.

Table 4.1: Similarity-based comparison of the evaluated GAN models

Pose Angle GAN	Generator	Discriminator			
		PatchGAN		Pixel	
		SSIM	MSE	SSIM	MSE
45	U-Net 256	0.678	3.19 e3	0.7109	3.1634 e3
	U-Net 128	0.6353	3.1417 e3	0.6442	3.114 e3
	ResNet 9B	0.7104	2.7748 e3	0.7119	3.0229 e3
	ResNet 6B	0.6956	2.9238 e3	0.7114	3.0006 e3
90	U-Net 256	0.6057	4.385 e3	0.6144	4.4067 e3
	U-Net 128	0.5594	4.3302 e3	0.5714	4.3097 e3
	ResNet 9B	0.6032	4.2446 e3	0.6241	4.2144 e3
	ResNet 6B	0.6069	4.2372 e3	0.6039	4.187 e4

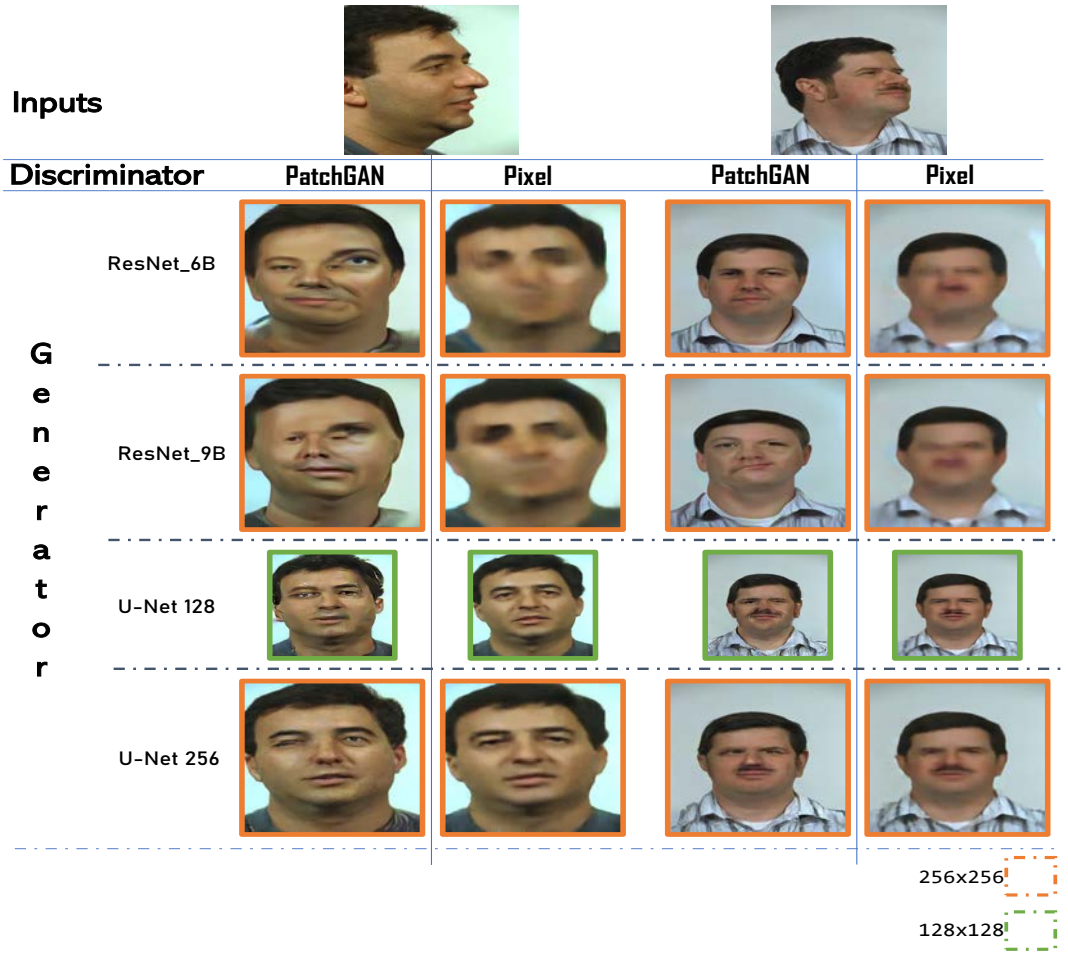

Figure 4.9: Comparison of the produced frontal images by the evaluated GAN's architectures

4.4.3/ EVALUATION OF THE PROPOSED PIFR FRAMEWORK

After evaluating the quality of the frontalized probe images against their ground truths, we assess in this section the overall performance of the proposed PIFR framework by calculating the recognition rate on the probe images. The output of the frontalization subsystem is the input of the Residual-CNN classifiers (ResNet-18, ResNet-50, ResNet-101, DenseNet-201) trained only on the frontal faces to predict the label of the person. To measure the improvement gained by the GAN-based frontalization, we need to calculate the baseline rates so we can compare the performance on the original and frontalized images. Table 4.2 lists the recorded recognition rate of the 8 discussed GANs combinations along with the four adopted ResNet face classifiers. The rates highlighted with green color are the higher and with red the lowest ones, which makes the understanding and ranking of the accuracies easier. Therefore, it can be seen that the combination of U-Net 256 and ResNet101 performed as the best compared to the rest since it reached 82.97% in the case of using the Pixel discriminator and 78.785% with PatchGAN. Moreover, the maximum baseline results was only 49.4%, which proved the effectiveness of the proposed PIFR framework with an improvement of 33.57%. ResNet18, along with the PatchGAN, delivered the lowest performance overall the evaluated architectures with an accuracy of 19.955%, which is even lower than the baseline performance of ResNet18 (31.08%). It can also be inferred that the ResNet_6B generator completely outperforms the 9 blocks one, and the difference is quite big in the case of PatchGAN discriminator (4% to 18%) depending on the face classifier and between 2% and 5% when they are combined with Pixel discriminator. On the other hand, the performance of U-Net 256 and 128 can be seen as close since each architecture outperforms the other, but U-Net 256 allowed the top accuracy overall of 82.97% against 78.685% for the 128 architecture. Therefore, the U-Net generators generally outperformed the ResNet-based ones. For the ranking of the face classifiers, the ResNet101 topped all the rest, and its minimum accuracy was around 52% with the GAN composed of ResNet 9B as generator and PatchGAN as the discriminator, which proves the classification power of ResNet101. ResNet50 classifier also demonstrated good performances, especially when classifying the frontal images produced by U-Nets models, but its performance suffered some drops with the ResNet-based generator, and the same behavior occurs with the DenseNet201 classifier. However, the top accuracy of ResNet50 and DenseNet201 looks close (74.305% and 73.955%, respectively). The ResNet50 classifier demonstrated good performance also in the case of adopting PatchGAN discriminator by reaching 71.81% with the GAN composed of {U-Net 128, PatchGAN} vs. 67.38% for DenseNet201 with the GAN composed of {U-Net 256, PatchGAN}. We can also notice that the more the face classification is deep, the better results we reach. However, the DenseNet network couldn't compete with the ResNets. The recorded accuracies proved that the Combined-PIFR dataset is very tough and challenged the ResNet face classifiers since the maximum baseline result

did not exceed 50% on 1000 probe images. In addition, there still room to improve the frontalization solutions as our proposed PIFR framework reached around 83%.

Table 4.2: Recognition rate of the proposed framework on the Combined-PIFR database test set

Discriminator	Generator	Face Classifier			
		ResNet18	ResNet50	ResNet101	DenseNet
PatchGAN	ResNet_6B	23.505	56.47	66.835	49.3525
	ResNet_9B	19.955	38.845	51.895	40.5875
	U-Net 128	30.48	71.81	77.29	56.77
	U-Net 256	34.065	70.615	78.785	67.3825
Pixel	ResNet_6B	48.805	48.705	62.05	52.04
	ResNet_9B	44.625	44.425	60.655	47.5575
	U-Net 128	52.39	74.305	78.685	61.6525
	U-Net 256	71.015	74.1	82.97	73.955
Baseline		31.08	41.61	49.4	39.05
Frontalization Improvement		39.935	32.695	33.57	34.905

We point out that the handcrafted face recognition framework based on the MNTCDP descriptor did not deliver good results on the frontalized images by the GAN. This fact is due to the convolutional generation of the GAN as the pixels variations are less compared to the original image, which can be stated through the blur effect. On the other hand, LBP-like methods rely the high variability of the pixels within small neighborhoods to compute relevant features.

4.4.4/ Implementation and execution time analysis

The execution time is one of the key performance indicators considered for evaluating face recognition frameworks as it should respect the real-time constraint. The training of the frontalziation and classification sub-systems is performed offline, so only the execution time to identify one probe image is enough to assess the efficiency of our proposed PIFR framework. Therefore, we run an experiment to calculate the elapsed time for each stage: Angle detection, frontalization, and classification. The recorded times are illustrated in Table 4.3. Note that the angle detection time is not included in this table since it is common to all the possible combinations of frontalization and classification architectures proposed and discussed in this paper. Dlib takes around 0.48 ms to estimate the pose angle. Moreover, the discriminator networks are not used in the evaluation stage, and only the generators are used. Therefore, we have only four architectures for the frontalization stage along with the four face classifiers. It can be inferred that the computation time of the top-performing configuration, which is the U-Net 256 generator with ResNet101 classifier, took only 9.72 ms to predict the identity of the probe image. Hence, by adding

did not exceed 50% on 1000 probe images. In addition, there still room to improve the frontalization solutions as our proposed PIFR framework reached around 83%.

Table 4.2: Recognition rate of the proposed framework on the Combined-PIFR database test set

Discriminator	Generator	Face Classifier			
		ResNet18	ResNet50	ResNet101	DenseNet
PatchGAN	ResNet_6B	23.505	56.47	66.835	49.3525
	ResNet_9B	19.955	38.845	51.895	40.5875
	U-Net 128	30.48	71.81	77.29	56.77
	U-Net 256	34.065	70.615	78.785	67.3825
Pixel	ResNet_6B	48.805	48.705	62.05	52.04
	ResNet_9B	44.625	44.425	60.655	47.5575
	U-Net 128	52.39	74.305	78.685	61.6525
	U-Net 256	71.015	74.1	82.97	73.955
Baseline		31.08	41.61	49.4	39.05
Frontalization Improvement		39.935	32.695	33.57	34.905

We point out that the handcrafted face recognition framework based on the MNTCDP descriptor (see Chapter 3) did not deliver good results on the frontalized images by the GAN. This fact is due to the convolutional generation of the GAN as the pixels variations are less compared to the original image, which can be stated through the blur effect. On the other hand, LBP-like methods rely the high variability of the pixels within small neighborhoods to compute relevant features.

4.4.4/ Implementation and execution time analysis

The execution time is one of the key performance indicators considered for evaluating face recognition frameworks as it should respect the real-time constraint. The training of the frontalziation and classification sub-systems is performed offline, so only the execution time to identify one probe image is enough to assess the efficiency of our proposed PIFR framework. Therefore, we run an experiment to calculate the elapsed time for each stage: Angle detection, frontalization, and classification. The recorded times are illustrated in Table 4.3. Note that the angle detection time is not included in this table since it is common to all the possible combinations of frontalization and classification architectures proposed and discussed in this paper. Dlib takes around 0.48 ms to estimate the pose angle. Moreover, the discriminator networks are not used in the evaluation stage, and only the generators are used. Therefore, we have only four architectures for the frontalization stage along with the four face classifiers. It can be inferred that the computation time of the top-performing configuration, which is the U-Net 256 generator with ResNet101 classifier, took only 9.72 ms to predict the identity of the probe image. Hence, by adding

keeping a real-time prediction in less than 10ms. On the other hand, the Combined-PIFR database proved to be challenging and presents big room for improvement.

5

FACIAL EXPRESSION RECOGNITION

5.1/ INTRODUCTION

This chapter presents our contributions to the facial expression recognition task. Two popular approaches have been proposed in the literature for decoding facial expressions. The first is geometric-based feature extraction. This approach relies on encoding geometric information such as the position, distance, and angle on the facial landmark points that a landmark detector should first identify and then extracts the feature vectors. The second approach is the appearance-based technique, which characterizes the appearance textural information resulting from the emotion classes' facial movements. Therefore, a set of features is extracted and is expected to contain relevant discriminating information to classify the different classes. The appearance-based approach utilizes many techniques for feature extraction. Moreover, the automatic FER task is further categorized into static and dynamic approaches depending on the input configuration. Static FER relies on using only one image to detect the dominant emotions. On the other hand, a dynamic framework requires many observations (samples) of the same person representing the evolution of the facial expression. We contributed to both cases by two independent frameworks. We proposed a hybrid framework for static person-independent FER that combines geometric and appearance concepts by extracting the textural and shape features from facial landmarks. The proposed combination is expected to promote an enhanced performance for person-independent FER because we consider geometric and appearance information that carries sufficient relevant features to describe the emotional classes. The dynamic person-independent FER is based on a Deep LSTM-CNN network to compute discriminant filter responses and encode the temporal information through the LSTM gates. Our network includes a dedicated residual sub-network "R-SubNet" to each input image for extracting the features. An LSTM block joins the R-SubNets and predicts the dominant emotion on the input images. The evaluation of both contributions is performed on widely used benchmarks with the leave-one-subject-out (LOSO) protocol, which is adopted to ensure the person-independent constraint. The achieved results

are compared to much state-of-the-art work to prove the superiority and improvement of our proposed frameworks.

5.2/ STATIC PERSON-INDEPENDENT FER

We propose an entirely new framework for person-independent FER based on combining textural and shape features from 49 detected landmarks in an input facial image. The geometric representation is obtained by interpolating 49 keypoints (landmarks) detected in the input image generating a binary patch, which is exploited to compute the shape features using the HOG method. By contrast, the appearance description is also extracted based on the detected landmarks instead of the whole face image, making our proposed FER framework able to fulfill the person-independent constraint. The appearance features are extracted from 32 pixel $\times 32$ pixel sub-images centered on each landmark using a brand new handcrafted descriptor, referred to as orthogonal and parallel-based directions–generic query map binary patterns (OPD-GQMBP). The OPD-GQMBP handcrafted descriptor is based on orthogonality and parallelism geometries for selecting the most prominent neighbors. It adopts an $n \times n$ neighborhood region to extract four feature maps based on four defined thresholding structures for each central pixel. The four feature maps are then decoded into one histogram. Afterward, the 49 feature vectors are concatenated to form the final appearance feature. During the classification step, we use the SVC library preprocessed by the PCA technique to reduce the dimensionality of the feature vectors. To highlight the contributions of our study and describe the workflow of our system in detail, we first describe the new textural handcrafted descriptor referred to as OPD-GQMBP. We then present how it is combined with the HOG shape descriptor to build the overall workflow.

5.2.1/ OPD-GQMBP: NEW HANDCARFTED DESCRIPTOR FOR FER

As discussed in the earlier chapter, the LBP operator is extremely flexible, and many of its aspects can be employed to develop enhanced descriptors for specific tasks. In our case, we propose a new LBP variant, referred to as OPD-GQMBP, which is based on new neighborhood topologies leading to four discriminant feature maps and adopts the LBP original kernel function that outputs low computational codes. The motivation behind the OPD-GQMBP descriptor relies on selecting orthogonal and parallel neighboring pixels that are believed to present the most information within a sub-block. In mathematics, orthogonality is defined as the generalization of the perpendicularity notion adopted by [181], who proposed a reduced LBP version referred to as OC-LBP, which considers two sets of four pixels located on the orthogonal lines. Thus, it produces only a feature

histogram with a 2×2^3 feature histogram. The OPD-GQMBP descriptor is generic and adjustable depending on the needs of the considered application. It adopts a $n\times n$ sub-block neighborhood, where n is an odd integer (3,5,7,9,...), to maintain symmetric neighborhoods. The concept behind this is the selection of prominent pixels within this neighborhood. Given a central pixel I_c, as shown in Figure 5.1, we define four-pixel groups, each of which contains $n \times 2$ pixels forming two lines. Two sampling groups are based on orthogonality $\{SG_1{}^{Ort}, SG_2{}^{Ort}\}$, whereas the two others are based on the concept of parallelism, i.e., $\{SG_1{}^{Par}, SG_2{}^{Par}\}$. Therefore, each sampling group SG is defined on two lines $SG_k{}^t(I_c) = \{L^t{}_{k,1}(I_c), L^t{}_{k,2}(I_c)\}$ where t stands for the type (Ort/Par) of the sampling group and k the group number $(1/2)$. Figure 5.2 shows a Cartesian coordinate system centered on the central pixel I_c to encode the position of each pixel considered in each sampling group within a 7×7 neighborhood. The sampling groups are defined as follows:

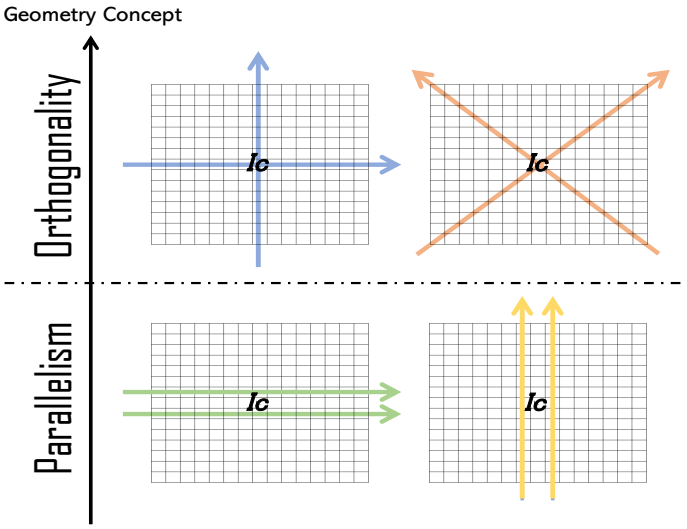

Figure 5.1: OPD-GQMBP neighborhood topologies.

$$SG_1{}^{Ort}(I_c) = \begin{bmatrix} L^{Ort}{}_{1,1}(I_c) = \{(x,0)/x\epsilon T\} \\ L^{Ort}{}_{1,2}(I_c) = \{(0,y)/y\epsilon T,\} \end{bmatrix} \tag{5.1}$$

$$SG_2{}^{Ort}(I_c) = \begin{bmatrix} L^{Ort}{}_{2,1}(I_c) = \{(x,x)/x\epsilon T\} \\ L^{Ort}{}_{2,2}(I_c) = \{(x,-x)/x\epsilon T\} \end{bmatrix} \tag{5.2}$$

$$SG_1{}^{Par}(I_c) = \begin{bmatrix} L^{Par}{}_{1,1}(I_c) = \{(x,1)/x\epsilon T\} \\ L^{Par}{}_{1,2}(I_c) = \{(x,-1)/x\epsilon T,\} \end{bmatrix} \tag{5.3}$$

$$SG_2{}^{Par}(I_c) = \begin{bmatrix} L^{Par}{}_{2,1}(I_c) = \{(1,y)/y\epsilon T\} \\ L^{Par}{}_{2,2}(I_c) = \{(-1,y)/y\epsilon T\} \end{bmatrix} \tag{5.4}$$

where

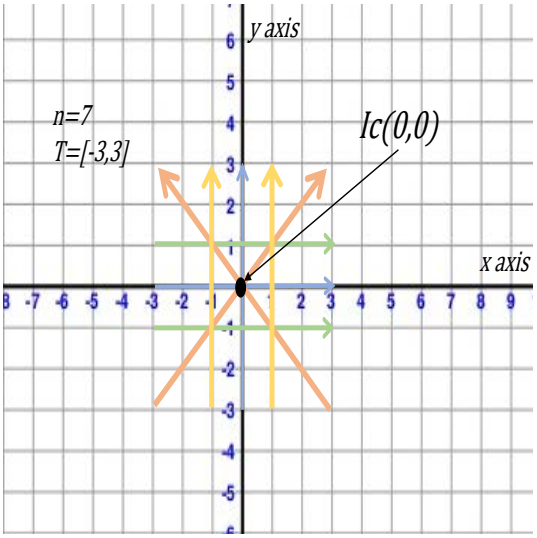

Figure 5.2: Cartesian system used to identify the pixel coordinates for each line and group.

$$T = \left[-\frac{n-1}{2}, \frac{n-1}{2}\right] \tag{5.5}$$

Here, T is the interval of values defining the coordinates (x, y) of the pixels constructing the two lines $\{L^t{}_{k,1}, L^t{}_{k,2}\}$ of each sampling group $SG_k{}^t(I_c)$.

Because all pixels within the $n \times n$ neighborhood are identified, we can proceed to the thresholding process. In this step, we generate for each sampling group $SG_k{}^t$ its feature map $\Im_k{}^t$, and obtain four feature maps:

$$\Im(I_c) = \begin{Bmatrix} \Im_1{}^{Ort}(I_c) = \Gamma(SG_1{}^{Ort}(I_c)) \\ \Im_2{}^{Ort}(I_c) = \Gamma(SG_2{}^{Ort}(I_c)) \\ \Im_1{}^{Par}(I_c) = \Gamma(SG_1{}^{Par}(I_c)) \\ \Im_2{}^{Par}(I_c) = \Gamma(SG_2{}^{Par}(I_c)) \end{Bmatrix} \tag{5.6}$$

with

$$\Gamma(SG_k{}^t)(I_c) = \Delta(L^t{}_{k,1}(I_c), L^t{}_{k,2}(I_c)) \tag{5.7}$$

Where Δ is the Heaviside function, which was initially used in the LBP operator defined in Eq 5.8, and applied the two lines of the same group to the threshold element by element. Thus, the length of the generated binary code for each feature map is the size (n) of the neighborhood, and the number of possible produced patterns is 2^n. Thus, by concatenating the patterns produced by all feature maps, we generate 4×2^n possible patterns. After encoding each pixel in the input image and obtaining the four feature maps, we transform them into a histogram vector as the final descriptor for the image, as defined in

$$\Delta(x,y) = \begin{cases} 1 & , x \geq y \\ 0 & , x < y \end{cases} \tag{5.8}$$

$$H(F) = \left\langle H^{\mathfrak{I}_1{}^{Ort}}, H^{\mathfrak{I}_2{}^{Ort}}, H^{\mathfrak{I}_1{}^{Par}}, H^{\mathfrak{I}_2{}^{Par}} \right\rangle \tag{5.9}$$

where

$$H^{\mathfrak{I}_k{}^t}(\mathbf{p}) = \sum_{\chi \in F} \delta(\mathfrak{I}_k{}^t(\chi), \mathbf{p}) \tag{5.10}$$

In Eq 5.10, $\mathbf{p} \in [0,\ 2^n - 1]$ is a pattern used to compare to the patterns $\mathfrak{I}_k{}^t(\chi)$, χ is the gray-scale value of the computed feature image F, and the delta function $\delta(\cdot)$, which is defined as follows (see. Eq. 5.11):

$$\delta(\mathbf{a},\mathbf{b}) = \begin{cases} 1, & \text{if } \mathbf{a} = \mathbf{b}; \\ 0, & \text{otherwise} \end{cases} \tag{5.11}$$

To include more spatial information into the OPD-GQMBP descriptor, the feature image is spatially divided into $w\times$ w small non-overlapping blocks B_i. All corresponding histograms $H(B_i)$ extracted from all blocks are concatenated to form the final holistic image representation through Eq. 5.12.

$$\mathbb{H} = \prod_{i=1}^{w^2} H(B_i) \tag{5.12}$$

where $\mathbb{H}$ is the final descriptor, $\prod$ is the concatenation operation, and $H(B_i)$ is the histogram of the OPD-GQMBP descriptor computed on the i^{th} block. Note that each elementary histogram $H(B_i)$ has a length of 4×2^n, whereas the dimensionality of $\mathbb{H}$ is $4 \times 2^n \times w^2$.

5.2.2/ Overall FER framework

After defining the neighborhood topology and thresholding kernel of the OPD-GQMBP descriptor, we now present the overall view of our proposed system for the FER task. The idea behind this framework is to combine the shape and appearance information to provide a more accurate FER, whereas most of the state-of-the-art proposed FER systems rely only on one piece of information, either geometric or appearance-based. To do so, we computed the textural and shape features based on the location of 49 detected

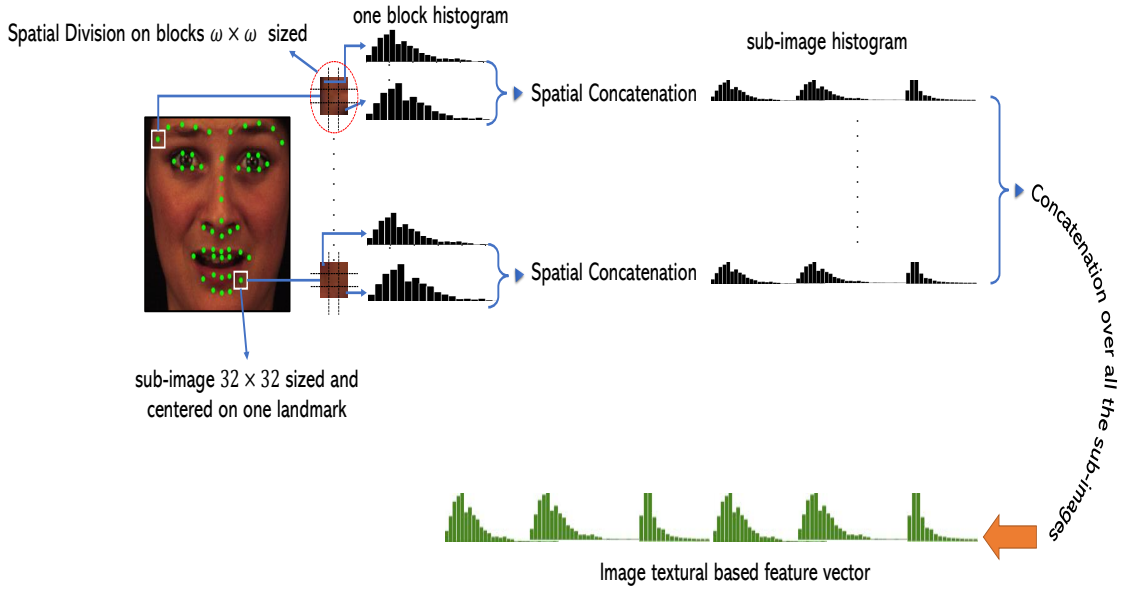

Figure 5.3: Texture feature extraction workflow based on the proposed OPD-GQMBP operator.

key points (landmarks) on the input face image. The shape representation is obtained by interpolating the 49 landmarks to form curves to be further analyzed using the HOG operator, whereas the appearance representation is based on texture analysis by applying the proposed OPD-GQMBP operator on specific sub-images of the input image. To make the FER system more able to fulfill the person-independent constraint, the appearance features are extracted from sub-images with a pixel resolution of 32×32 and centered on each landmark carrying sufficient and relevant information about the expressed emotion and less irrelevant information of the person's face. Figure 5.4 illustrates the overall pipeline of the proposed FER system. First, the input image is fed to the dlib landmarks extractor to locate the 49 points (green color). These locations are then interpolated to generate a binary patch of the expressed emotion, upon which the HOG operator is applied to compute the shape feature vector. Meanwhile, the OPD-GQMBP descriptor (or state-of-the-art descriptors for comparison) was used to extract the textural features from each 32×32 sized sub-image centered on one landmark leading to a set of 49 histograms (49 landmarks) that are further concatenated together to construct the appearance feature vector. Note that spatial division was adopted to compute the OPD-GQMBP feature vector by dividing each sub-image into non-overlapping blocks of size $w \times w$, as illustrated in Figure 5.3. The number of spatial blocks that divide the sub-image depends on the considered dataset and is related to the camera resolution and image blur. Indeed, blurred images require fewer blocks than clear images, which present more details to be detected. At the end of the feature extraction stage, the HOG and OPD-GQMBP computed histogram vectors are concatenated to compose the final image descriptor that is further fed to a dimensionality reduction using the PCA method before proceeding to the

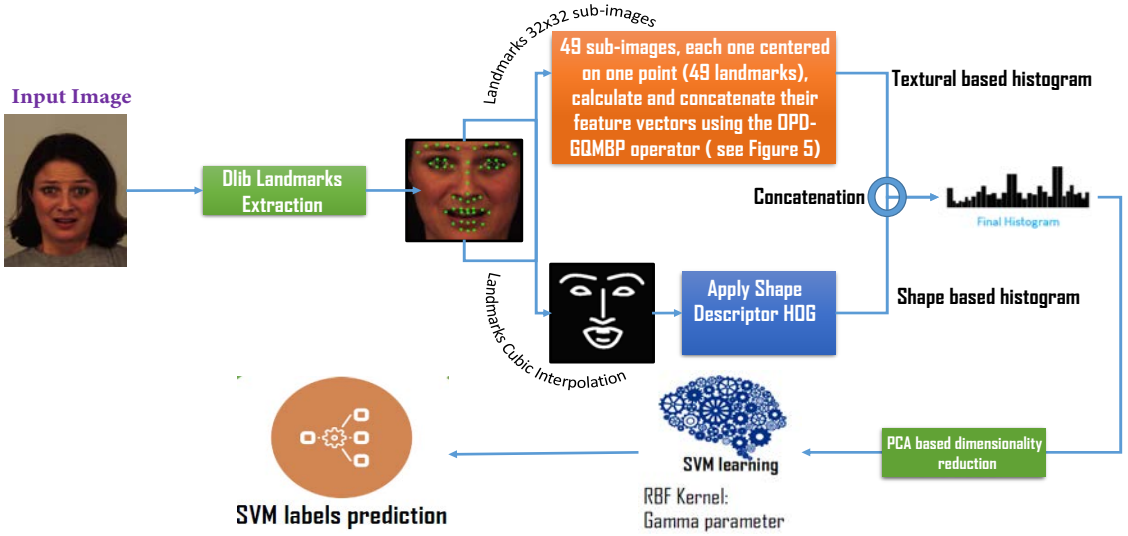

Figure 5.4: Overall view of the proposed FER framework.

classification phase based on an SVM. We used the LIBLINEAR 2.30 library as a multiclass kernel-based vector machine implementation for MATLAB/Python environments. This library provides many classification and regression solvers. We chose the support vector classification based on the Cramer and Singer solver $(Kernel = 4)$ as a simplified multi-class SVM. Furthermore, this kernel allows optimized training and takes less time compared to the LibSVM library implementation.

5.2.3/ Experimental Analysis and Discussions

To show the effectiveness of our proposed framework for person-independent Facial Expression Recognition, we performed extensive experiments on 5 well-known and widely used benchmarks of the literature: KDEF, CK+, RaFD, JAFFE, and OuluCasia. To ensure person independence in our testing, we set up a leave-one-subject-out cross-validation (LOSO) protocol, where all the samples of one person are excluded from the training set and used for testing. The process is repeated for N-persons, and no prior person information is included in the training stage. As presented in the previous subsection, our contribution to static person-independent FER relies on the FER framework itself and the OPD-GQMBP handcrafted descriptor. To highlight the results of each one, we firstly evaluate four possible configurations of the OPD-GQMBP descriptor, then compare its performance against LBP-like handcrafted methods and 10 deep features of the state-of-the-art within our FER framework, keeping the same evaluation protocol and conditions. The deep models are pretrained on the FER2013 large dataset. Afterward, the performance of the proposed FER framework is compared to the ones presented in state-of-the-art works, published in highly indexed and well-known journals of the literature.

5.2.3.1/ EXPERIMENTAL DATASETS

- The Japanese Female Facial Expression (JAFFE) dataset has 213 facial expression images, representing the seven basic emotions: Anger (30 images), Disgust (29 images), Fear (32 images), Happiness (31 images), Sadness (31 images), Surprise (30 images), and Neutral (30 images). Figure 5.5 illustrates some examples of the facial expressions. This database is very challenging regarding the particularity of Japanese females that have similar face features generating more inter-classes visual features.

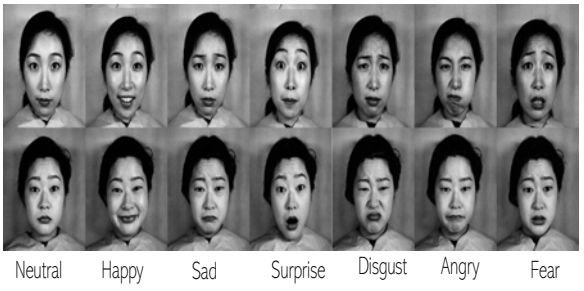

Figure 5.5: Samples of two subjects from JAFFE database

- The Karolinska Directed Emotional Faces (KDEF) is a widely used dataset for evaluating FER methods. It includes 70 individuals (50% Men, 50% Women) expressing uniformly the basic emotions over two sessions leading to a total of 980 images. In our experiments, we considered only one session to have only one observation per emotion for each person (490 images), which resulted 70 samples per class. Figure 5.6 shows the observation of each class over a female from this database.

Figure 5.6: Samples of a subject from KDEF database

- Cohn-Kanade v2 database (CK+) is a sequence based database. It contains 593 image sequences from 123 subjects. The first image of each sequence represents the neutral state of the subject, while the peak of the emotion is represented at the end of the sequence. In our experiment, we selected only the last frame to

construct the sets of the six emotions, while the neutral class is constructed by the first frame from each sequence. The Angry class has 45 samples, Disgust: 59; Fear: 25; Happy: 69; Neutral: 45; Sad: 28 and Surprise is represented over 82 images. Some of these samples are shown in Figure 5.7.

Figure 5.7: The seven emotions of one person from CK+ database

- The Oulu-CASIA NIR & VIS expression database is also a sequence based dataset, including 80 subjects (south asian and caucasian) with the six typical expressions. The videos are recorded with two imaging systems, NIR (Near Infrared) and VIS (Visible light). Only the last frame from each sequence of VIS the database, is considered and the neutral is represented by the first frame. Therefore a dataset of 560 images is obtained (80 samples per class). As can be seen in Figure 5.8, the images are a bit blur and not clear, in addition to the similarity of visual features of south asian individuals that make this dataset also a challenging one.

Figure 5.8: Samples of two subjects from OuluCasia database

- The Radboud Faces Database (RaFD) is composed of 67 individuals (including Caucasian, Moroccan, Dutch adults, and Caucasian children, both boys and girls) displaying 8 emotional expressions. In addition to the seven basic emotion expressions, this database includes the Contempt facial expression, which can be similar to angry and disgust emotions, but expresses the feeling of dislike for and superiority over another person, and/or his actions. Moreover, the Contempt emotion is not symmetric and occurs only on one side of the face. Figure 5.9 displays the 8 facial expressions of a person from RaFD database.

Figure 5.9: Samples of one subject from RaFD database

5.2.3.2/ EVALUATION OF OPD-GQMBP NEIGHBORHOOD SIZE CONFIGURATION

The proposed OPD-GQMBP descriptor is a generic method defined by the neighborhood size n, which can be seen as a user-specified parameter depending on the needs of the considered application. To find the best value for FER, we run an experiment evaluating the performance of four configurations: n=3, 5, 7, and 9. For each, we evaluated the FER framework on the five datasets using the LOSO protocol. The smaller neighborhood sizes provide less computational cost but with weak discriminative power, and higher ones enhance the discriminative power, but they require more resources to store and classify the extracted features. For example, a neighborhood size of 3 ($OPD-GQMBP^3$) generates only $4 \times 2^3 = 32$ possible patterns, while a neighborhood size of 5 ($OPD-GQMBP^5$), 7 ($OPD-GQMBP^7$), and 9 ($OPD-GQMBP^9$) produce 128, 512, and 2048 patterns, respectively. Table 5.1 shows the obtained recognition rates resulting from this experiment. It can be concluded that with a higher neighborhood size, we obtain more discriminative feature extraction. The most effective configuration is $n = 7$, which managed to reach the top performance on 4 databases. Thus, the 512 generated patterns proved to be enough for characterizing the seven emotional classes. $OPD-GQMBP^9$ achieved top accuracy of 97.53% on CK+ database outperforming $OPD-GQMBP^7$, but it suffered performance drop on the other datasets. Indeed, in some cases, methods that generate a high number of patterns may cause performance drop due to pattern redundancy. It's clear that $OPD-GQMBP^3$ configuration could not outperform the other ones. However, the recorded accuracies remain prominent regarding the low computation (32 patterns only). The performance of $OPD-GQMBP^3$ and $OPD-GQMBP^5$ are very similar with slight variations. We acknowledge that we could not evaluate neighborhood sizes greater than 9 due to the required computation resources (out of memory). Therefore, we adopt the $OPD-GQMBP^7$ in the rest of the experiments since it corresponds to the best configuration among the tested ones.

Table 5.1: Average FER rate of each OPD-GQMBP configuration (neighborhood size), overall databases

N_{Size} Config	CK+	JAFFE	KDEF	OuluCasia	RaFD
$OPD-GQMBP^3$	95.74	73.33	88.57	73.93	96.08
$OPD-GQMBP^5$	96.01	73.33	87.96	74.82	95.9
$OPD-GQMBP^7$	96.48	78.57	90.2	77.32	97.39
$OPD-GQMBP^9$	97.53	71.9	87.96	75.95	96.08

5.2.3.3/ COMPARATIVE ANALYSIS AGAINST STATE-OF-THE-ART HANDCRAFTED AND DEEP FEATURE METHODS

This comprehensive analysis aims to compare the performance of the proposed OPD-GQMBP descriptor to the ones recorded in the literature of feature extraction methods, including handcrafted and deep-based approaches. We evaluated the top 10 deep learning models proposed so far in the state-of-the-art. These models were initially trained on the FER2013 database, and each of them reached a validation accuracy above 60% on 25000 images, which can be considered very significant. Then, we use transfer learning to extract the features of the five databases adopted in this experiment. We respected the same evaluation protocol (LOSO) to provide a fair and systematic analysis. Table 5.2 lists the performance reached by each method or model. We provide two metrics, the first one is the average of the accuracies recorded for all the runs of each database depending on the number of individuals, where JAFFE has 10 runs, CK+ 106, KDEF 70, OuluCasia 80, and 67 runs for the RaFD database. The second metric is the maximum accuracy reached overall the runs per database. We highlight with the green color the top 3 average values.

It can be seen from the average accuracies that the proposed OPD-GQMBP descriptor managed to score the top performance on all tested datasets. On the CK+ database, the OPD-GQMBP descriptor achieved 96.48% on 106 runs with a maximum of 100% recognition accuracy, keeping more than a 2% gap to the following top method, which is the handcrafted MNTCDP descriptor that managed to secure an average accuracy of 94.36%. All the evaluated methods were capable of reaching 100% Max accuracy on the CK+ dataset. Moreover, most handcrafted methods performed above 87.81%, reached by DCP descriptor, while the deep features reached a maximum average of 88.76% with VGG19 only 54.18% (AlexNet) as minimum average on the CK+. Based on Table 5.1, we found that $n = 9$ configuration of the OPD-GQMBP method scored 97.53% on CK+, which improves by 1% compared to $OPD-GQMBP$ with $n = 7$ adopted in this evaluation. The OPD-GQMBP reached 77.62% on JAFFE with a lead of 3% to the rest of the methods since the second highest avg accuracy is 74.76% recorded by the DCP

Table 5.2: Average and Maximum accuracies recorded on the five datasets by each method

	Method	JAFFE		KDEF		CK+		OuluCasia		RaFD	
		Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
Deep Features	VGG16	56.67	76.19	76.53	100	86.34	100	56.96	100	82.84	100
	VGG19	53.81	76.19	76.73	100	88.76	100	58.57	100	81.53	100
	ResNet18	35.24	57.14	72.24	100	72.67	100	52.32	100	83.02	100
	ResNet50	44.76	61.9	75.71	100	77.32	100	56.96	100	84.51	100
	ResNet101	52.86	71.43	70.82	100	76.79	100	53.57	100	78.17	100
	AlexNet	43.81	66.67	66.94	100	54.18	100	29.64	71.43	67.54	100
	DenseNet	48.1	80.95	70.2	100	76.15	100	50.36	100	77.8	100
	GoogLeNet	47.14	66.67	71.63	100	69.98	100	44.64	85.71	79.66	100
	Inceptionv3	39.52	57.14	65.31	100	71.22	100	44.46	85.71	70.52	100
	InceptionResNetv2	47.62	66.67	76.33	100	80.17	100	55	100	81.72	100
Handcrafted LBP Variants	ELGS	73.81	95.24	85.71	100	92.59	100	70	100	95.34	100
	DSLGS	60.95	85.71	78.57	100	91.2	100	64.64	100	91.23	100
	MNTCDP	69.05	85.71	85.51	100	94.36	100	55.36	85.71	95.15	100
	QBP	62.38	85.71	79.39	100	91.99	100	66.43	100	91.79	100
	DRLBP	70	85.71	84.49	100	90.35	100	65.36	100	96.08	100
	LNDP	69.52	95.24	86.94	100	92.86	100	68.75	100	96.08	100
	DCP	74.76	100	69.18	100	87.81	100	45.89	100	87.31	100
	DC	66.67	95.24	84.69	100	89.87	100	74.64	100	95.52	100
	LCCMSP	71.43	95.24	86.12	100	92.28	100	69.29	100	97.01	100
	LOOP	64.76	85.71	85.92	100	91.7	100	69.64	100	96.27	100
	LDTP	70	90.48	88.37	100	93.3	100	67.5	100	97.01	100
	ARCS-LBP	65.24	95.24	85.51	100	91.45	100	75.5	100	96.64	100
	OPD-GQMBP	**77.62**	**100**	**90.2**	100	**96.48**	100	**77.32**	100	**97.2**	100

descriptor followed by ELGS as the third best performing method (73.81%). According to the maximum accuracy on the JAFFE database, only OPD-GQMBP and DCP methods managed to score 100%. The Japanese females' high facial similarities make the JAFFE a very tough benchmark to the FER systems. The results demonstrated that the JAFFE database is very challenging, especially for deep features where their top average accuracy was limited to 56.67% obtained by VGG16, while the handcrafted ones granted an average accuracy above 64.76% reached by the LOOP descriptor. For the KDEF database, the first remark to be concluded is that the OPD-GQMBP descriptor is the only method that breaks the 90% average performance ceiling and the best state-of-the-art method reached 88.37% (LDTP). The handcrafted methods' performance and deep ones vary from 78.57% to 88.37% and from 65.31% to 76.73%, respectively. Again, the leadership gap on this database was around 2% for the favor of OPD-GQMBP descriptor, which proves its discriminative power for FER application. Despite the quality of the recorded images in terms of resolution, lighting conditions, and uniform background, the KDEF database is also challenging in face to the CK+. It presents fewer individ-

uals (fewer runs), who express the seven emotions in different manners making more intraclass similarities and approaching the spontaneous facial expressions. OuluCasia can be considered the toughest among the adopted benchmarks due to the blur images and the south Asian persons composing this database. The OPD-GQMBP descriptor scored 77.32% as the overall best performing method, followed by ARCS-LBP (75.5%) and DC (74.64%) descriptors. The lowest accuracy was recorded by AlexNet deep network reaching only 29.64%. All the methods performed 100% Max accuracy except the MNTCDP, AlexNet, GoogLeNet, and Inceptionv3. Moreover, only ELGS, DC, ARCS-LBP, and OPD-GQMBP methods exceeded 70% average accuracy, while the DCP descriptor was the second top-ranked on JAFFE, found its performance limited to 45.89% as the lowest average accuracy over all the handcrafted LBP-like methods. The RaFD database is collected by a set of well-trained individuals who clearly express eight emotions (basic emotions + Contempt). Indeed, the peak average accuracy exceeded 97% by the proposed OPD-GQMBP descriptor with a minor lead (0.19%) against LDTP and LCCMSP methods (97.01%). Moreover, all the handcrafted descriptors reached above 90% except DCP (87.31%). On the other hand, ResNet50 was the best among deep feature methods with an average accuracy of 84.51%. In terms of stability, the OPD-GQMBP descriptor performed well on the five datasets, always reaching the top average accuracies and 100% max all the time. ELGS method also presented stable performance across all databases. On the other hand, DCP suffered a performance drop on KDEF and OuluCasia datasets. For the deep feature methods, VGG16 can be considered as the best performing deep feature method.

The deep learning networks did not perform well on the five benchmarks despite reaching a validation accuracy above 60% on 25,000 images of the FER2013 database. The problem is that the deep learning methods should be fine-tuned on each dataset before extracting the features to get satisfying results. The subject application for this contribution is person-independent, and to ensure that the probe images of a given person are unseen by the framework, we should perform fine-tuning on each run and for each deep method. Hence, since we have 333 persons on the five datasets and considered 10 deep learning models, we need 3330 fine-tunes to expect satisfying results from these models, which is time and resources consuming. Moreover, such frameworks are intended for real implementations and deployments, and fine-tuning is not always a possible option; besides, the probe images will have generally different characteristics than the train ones.

5.2.3.4/ COMPARISON AGAINST STATE-OF-THE-ART FER SYSTEMS

Now, we compare the results obtained by our proposed FER framework to those achieved by previous works in the field of facial expression recognition. Tables 5.3, 5.4, 5.5, 5.6 and 5.7 list the highest accuracies on the five datasets reported in well-indexed journals and

conferences of the literature. We tried to collect the maximum of the works that followed the same adopted evaluation protocol (person independent).

Table 5.3: State-of-the-art person independent FER accuracies on CK+ database

Methods	Type	Avg accuracy
LPDP [105]	Handcrafted	94.5
DCNN [182]	Deep	94.44
DNN [183]	Deep	93.52
CNN+AFM [184]	Deep	89.84
AlexNet+SVM [184]	Deep	86.83
GoogLeNet [184]	Deep	85.71
STM-ExpLet [185]	Deep	94.13
CER [186]	Handcrafted	92.34
SRC+ICV [187]	Handcrafted	90.5
MSR [188]	Handcrafted	91.4
Gabor+SRC [186]	Handcrafted	82.82
3DCNN-DAP [189]	Deep	92.4
LTeP+SVM [190]	Handcrafted	94.93
LPQ+SLPM+NN [191]	Handcrafted	94.61
WPLBP [192]	Handcrafted	91.72
Proposed	Handcrafted	**97.53**

As can be found in Table 5.3, the proposed FER framework outperforms all the listed systems, including both handcrafted and deep based ones on the CK+ database. We reached 96.48% (97.53% with the OPD-GQMBP neighborhood size $n = 9$), while the best accuracy of the state-of-the-art is 94.96% achieved using LPQ with SVM classifier. Moreover, the majority of the published works performed between 90% and 94%. On the JAFFE database, the state-of-the-art accuracies are low compared to CK+, where the maximum reported is 76.46% scored by CFER based framework outperformed by our FER framework (77.62%). The proposed framework managed to surpass with significant leads all the works on the KDEF dataset, except for WCFN (89.55%) and AlexNet (89.33%) based systems where the margin is small (0.65% and 0.87% to our framework performance, respectively). On the OuluCasia dataset, the proposed framework (77.32%) outperformed all the state-of-the-art methods, where the top accuracy was limited to 75.52% reached by Atlases. On the RaFD database, our proposed framework obtained 97.2%. Many works of the literature performed nearly to 97%. However, the majority were applied only on 7 emotion classes. Overall, we conclude that the proposed facial expression recognition framework managed to outperform all the tested state-of-the-art ones.

5.2.3.5/ CONFUSION MATRIX-BASED ANALYSIS FOR THE FER

The confusion matrix allows analyzing the performance of the recognition according to each label (each emotion in our case). Through this chart, we are capable of analyzing the recognition rate of each emotion and what are the easiest and hardest ones to be recognized. Also, this analysis allows identifying which emotion is affecting the others.

Table 5.4: State-of-the-art person independent FER accuracies on JAFFE database

Methods	Type	Avg accuracy
LTeP+SVM [190]	Handcrafted	67.14
LPQ+SLPM+NN [191]	Handcrafted	67.61
EDR-PCANet [193]	Deep	69.4
C-classLDA-NN [194]	Deep	74.73
LBP based LDA [195]	Handcrafted	73.4
BDBN [196]	Deep	68
CFER [197]	Handcrafted	76.46
Features fusion [198]	Handcrafted	70
Proposed	Handcrafted	**77.62**

Figures 5.10, 5.11, 5.12, 5.13 and 5.14 illustrate the confusion charts generated from the results of our proposed FER framework for CK+, JAFFE, KDEF, OuluCasia and RaFD, respectively.

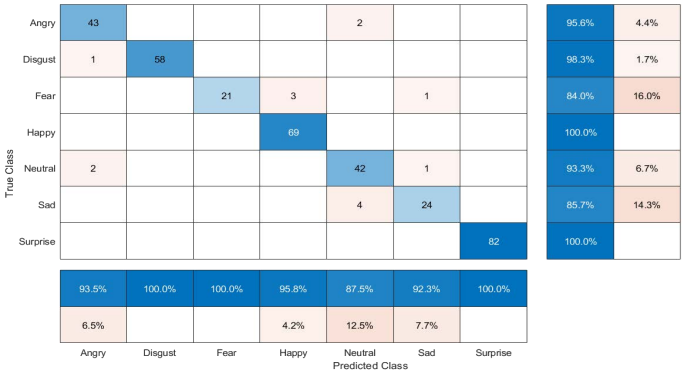

Figure 5.10: Confusion matrix of the seven emotions of CK+

Figure 5.11: Confusion matrix of the seven emotions of JAFFE

On the CK+ database, Happy and Surprise classes were perfectly-recognized, while the Fear and Sad experienced the highest misclassification error (16.0% and 14.3% respectively). The Fear emotion was confused with Happy (3 times) and once with Sad. The Neutral emotion was the most affecting one (with 12.5% false-negative rate) and predicted 4 times in the case of Sad and 2 for Angry. For the JAFFE database, Happy and

Table 5.5: State-of-the-art person independent FER accuracies on KDEF database

Methods	Type	Avg accuracy
AlexNet+FC6+LDA [199]	Deep	89.33
HOG+SRC [200]	Handcrafted	78
VGG-Face Deep [201]	Deep	72.55
SCAE [202]	Deep	86.73
CNN Caffe-ImageNet [203]	Deep	59.15
Geometric Features [204]	Handcrafted	79.69
DFD [205]	Handcrafted	82.24
HPBSVM [206]	Handcrafted	81.84
WCFN [207]	Deep	89.55
MobileNet [208]	Deep	73.74
EDR-PCANet [193]	Deep	80.61
Proposed	Handcrafted	**90.2**

Table 5.6: State-of-the-art person independent FER accuracies on OuluCasia database

Methods	Type	Avg accuracy
STM-ExpLet [185]	Deep	74.59
LBP+Gabor+SVM [209]	Handcrafted	74.37
HOG 3D [210]	Handcrafted	70.63
AdaLBP [211]	Handcrafted	73.54
Atlases [208]	Deep	75.52
Proposed	Handcrafted	**77.32**

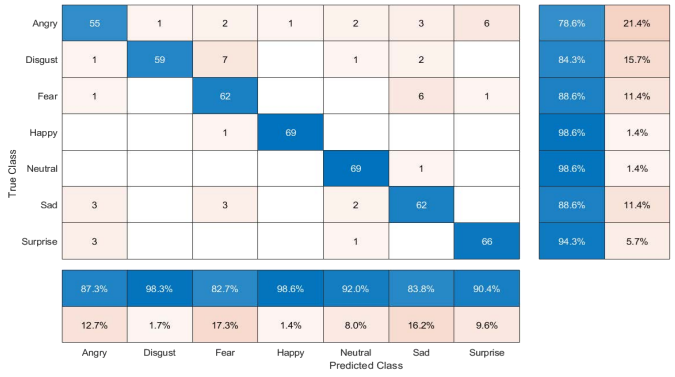

Figure 5.12: Confusion matrix of the seven emotions of KDEF

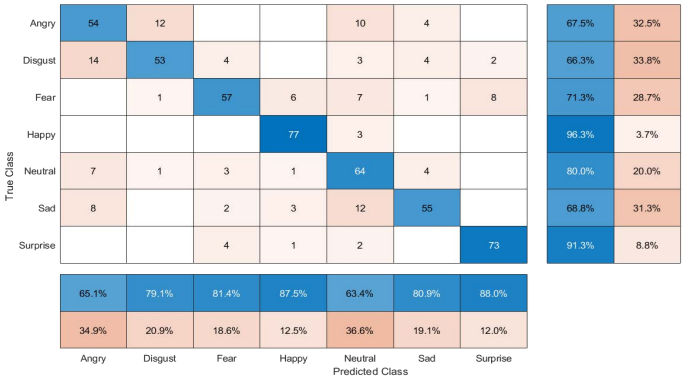

Figure 5.13: Confusion matrix of the seven emotions of OuluCasia

Table 5.7: State-of-the-art person independent FER accuracies on RaFD database

Methods	Type	Avg accuracy
Visual Attention CNN [212]	Deep	95.2
DS+FE+GEM+SVM [213]	Handcrafted	90.8
LPQ+FE+GEM+SVM [213]	Handcrafted	94.4
LBP+FE+GEM+SVM [213]	Handcrafted	94.5
DS+SVM [214]	Handcrafted	79
Metric Learning [215]	Deep	95.95
BAE-BNN-3 [216]	Deep	96.93
W-CR-AFM [184]	Deep	96.27
Net1-Net2 [217]	Deep	93.41
Proposed	Handcrafted	**97.2**

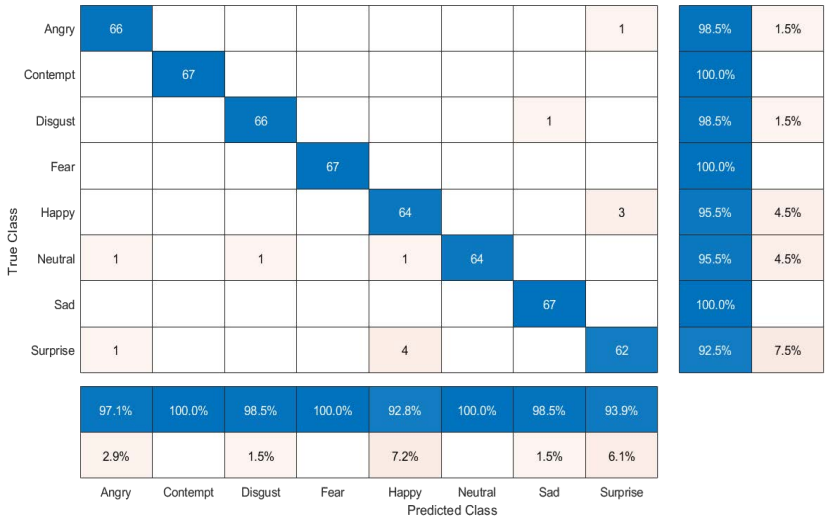

Figure 5.14: Confusion matrix of the eight emotions of RaFD

Neutral were the easiest to identify with an accuracy of 86.7% and the hardest is Fear (60%), which was confused with all the emotions: 3 times with Disgust, Neutral, Surprise, and once with the rest. Sad class presented the highest false-negative rate (34.4%). The females composing this database express the Angry and Sad emotion in similar ways since 5 Angry samples were identified as Sad. On KDEF, the errors were less compared to JAFFE. Happy and Neutral once again were the highly recognized emotions (98.6%), but Angry is the challenging class with only a 78% rate. Our FER framework confused Disgust 7 times with Fear and 6 times Fear with Sad. Happy and Disgust emotions were not affecting any other emotion except 1 time for each with Angry only. As expected on OuluCasia, the misclassification errors were very high. The Happy emotion is the only one to be highly recognized with an accuracy of 96.3%, followed by Surprise with 91.3%. The rest accuracies were between 66.3% and 80%. Also, Angry and Neutral dramatically influenced the other emotions with 34.5% and 36.6%, respectively. On the RaFD database, all the rates were high with the perfect recognition for three classes (Contempt, Fear, and Sad) in addition to Angry and Disgust that were misclassified only once. However, there is a mutual confusion between Happy and Surprise since 3 Happy samples

were identified as Surprise and 4 times in the opposite direction (Surprise to Happy). Overall the databases, we disclose that Happy and Neutral emotions are the most recognized ones. In addition to that, Neutral is very perturbing the framework as it presents high false-negative rates on many benchmarks.

5.2.3.6/ IMPLEMENTATION AND EXECUTION TIME

The presented static FER experiments have been performed on an Alienware Aurora R8 with Core i7-8th Processor 4.6GHz Boost, 12 Threads, and 48 GB of RAM, running with Ubuntu 18.04.2 LTS (Bionic Beaver) operating system and equipped with 2 GPU GTX1080Ti. The developed framework is coded using Python3.7 and Matlab2019b environments.

The computational cost is one of the key performance indicators considered in such machine learning applications. Therefore, we run an experiment that calculates the elapsed time to predict the label of a given input image that has a resolution of 762 by 562 pixels, highlighting the execution time of each step of the proposed framework:

- Dlib landmarks detection.
- Shape feature extraction based on HOG descriptor.
- Appearance feature extraction using the proposed OPD-GQMBP as well as the state-of-the-art handcrafted methods denoted as "getFeatures".
- PCA dimensionality reduction.
- Label prediction by the SVM library.

We excluded the deep-learning methods from this evaluation since they require GPU units to perform the feature extraction. Thus, it will be unfair to compare CPU-based methods with the GPU ones regarding the computation power of GPUs. The CPU can be used to calculate the feature vector by a deep model, but it takes around 4 seconds to compute it for an input image with a size of only 224 by 224 (3 times smaller than the size of the original image), which is very high compared to the handcrafted ones. The obtained computation times are illustrated graphically in Figure 5.15. As can be seen, the elapsed times for the landmarks detection, HOG shape feature extraction, and SVM prediction did not change across the evaluated methods, where they recorded 15.9 ms, 19.58 ms, and 2.306 ms respectively. The process of extracting the appearance features demonstrated to be the most time taking one within our framework (more than 50% of the total time). The fastest handcrafted method is DRLBP that extracted the features from the 49 landmarks in 31.5 ms, while the LDTP took 311.6 ms to extract these features

judged thus as the heaviest one. However, the DRLBP did not perform well in terms of classification accuracy, whereas the proposed OPD-GQMBP descriptor managed to offer an execution time of 90.34 ms only, which is so beneficial regarding its high performance, as demonstrated earlier. Moreover, all the best-performing descriptors took more than 100 ms to compute the appearance features. Although the PCA stage is common to all the descriptors for dimension reduction, its execution time was variable and affected by the number of generated patterns of each handcrafted method as a high number of patterns leads to more computation. Nevertheless, it can be remarked that the PCA computation times are similar for the methods sharing the same size of generated patterns. The methods producing 256 patterns such as LOOP, QBP, DC, ARCS-LBP, and DRLBP recorded a PCA computation time around 11 ms. The PCA process of the proposed OPD-GQMBP descriptor as well as those producing 512 patterns, took around 20 ms, while LDTP and LNDP methods generating 1024 patterns took about 36 ms. On the other hand, the PCA-based dimensionality reduction process related to the LCCMSP (2048 patterns) descriptor was performed in 83.5 ms. Overall, we can disclose that the proposed framework, along with the OPD-GQMBP method, managed to predict the label of an image with a resolution of 762 by 562 pixels in less than 150 ms, allowing to process 7 frames per second, which is considered real-time feedback according to the specifications of person-independent FER systems [218; 219].

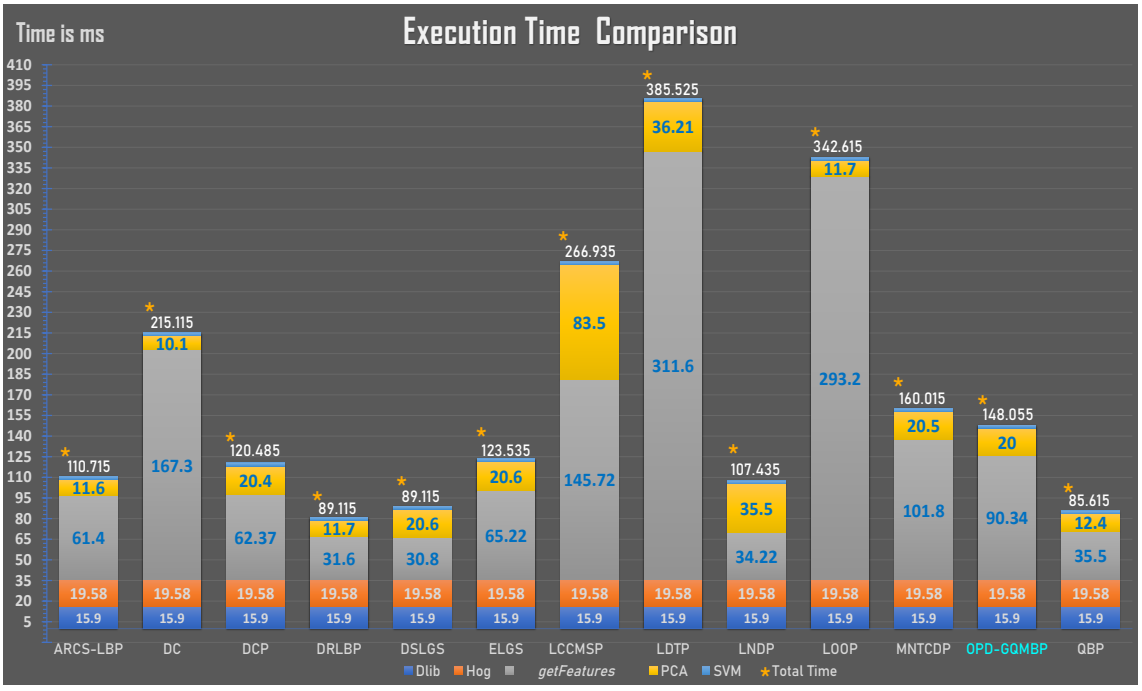

Figure 5.15: Comparison of the overall execution time elapsed to predict the label of an image

5.3/ Dynamic Person-Independent FER

The dynamic Person-Independent FER relies on processing multiple samples to detect the dominant emotion through its evolution. We propose a deep CNN-LSTM network considering four consecutive samples by extracting the deep features based on dedicated CNN streams, then the use of an LSTM network to decode the temporal information and predict the facial expression class. The CNN networks that can be used as deep-features extractors all already introduced in Section 2.11, and in the following subsubsections, we introduce the concept of the LSTM and the overall architecture for dynamic Person-Independent FER.

5.3.0.1/ Longer Short Term Memory

LSTM network is a recurrent neural network RNN that is popular for handling time series data. For each timestep, the RNN considers the new input data and its output from the previous iteration, known as the hidden state. In this way, RNNs have a basic form of short-term memory and are better at decoding short-term patterns in the data compared to plain feed-forward networks.

The LSTM network is composed of special memory units in the recurrent hidden layer. These memory blocks contain neurons with self-connections to save the current cells state in addition to special activation units called gates to control the flow of inputs to consider. Each memory unit is controlled by input and output gates, where the input manages the input activations connected to the memory cell input and the output gate controls the flow of cell activations for the rest of the LSTM network. Later, a forget gate was added to the memory block to prevent them from processing continuous input streams that are not segmented into subsequences, which was a collapsing mode of traditional LSTM. The forget gate scales the current state of the cell before adding it as input to the cell through the self-recurrent connection of the cell, therefore adaptively forgetting or resetting the cell's memory. Figure 5.16 shows the internal architecture of an LSTM unit with forget gate.

Based on Figure 5.16, the kernel function linking the input and output of an LSTM memory block can be expressed as follows:

$$\begin{aligned}
f_t &= \sigma\left(W_{fh}h_{t-1} + W_{fx}x_t + b_f\right) \\
i_t &= \sigma\left(W_{ih}h_{t-1} + W_{ix}x_t + b_i\right) \\
\tilde{c}_t &= \tanh\left(W_{\tilde{c}h}h_{t-1} + W_{\tilde{c}x}x_t + b_{\tilde{c}}\right), \\
c_t &= f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t, \\
o_t &= \sigma\left(W_{oh}h_{t-1} + W_{ox}x_t + b_o\right), \\
h_t &= o_t \cdot \tanh\left(c_t\right).
\end{aligned} \tag{5.13}$$

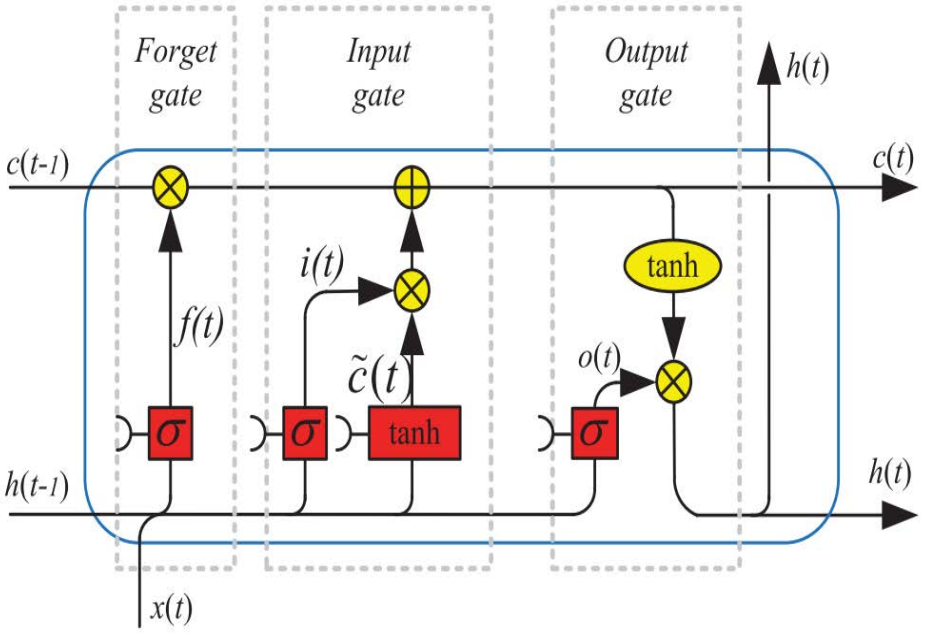

Figure 5.16: Internal configuration of the LSTM Cell

where c_t denotes the cell state of LSTM. W_i, $W_{\tilde{c}}$, and W_o are the weights, and the operator '.' denotes the point-wise multiplication of two vectors. When updating the cell state, the input gate can decide what new information can be stored in the cell state, and the output gate decides what information can be output based on the cell state. The forget gate can decide what information will be thrown away from the cell state. When the value of the forget gate, f_t, is 1, it keeps this information; meanwhile, a value of 0 means it gets rid of all the information

In our contribution, we used Deep-LSTM architecture that is simply stacking multiple LSTM layers. Moreover, we defined the LSTM to also serve features compression by a factor of 2 that helps to smoothly predict the emotion by the final fully connected layer.

5.3.0.2/ DEEP CNN-LSTM FOR DYNAMIC FER

In this thesis, we proposed an End-to-End deep architecture for dynamic person-independent FER from four sequence samples, where the fourth represents the peak of the facial expression. As shown in Figure 5.17, the proposed architecture incorporates different techniques divided into three main steps.

The preprocessing step of our proposed framework is to keep only the visual features related to the facial expression and ignore those describing the person's identity, which will help fulfill the person-independent constraint. The Dlib package, as used in the previous contribution, detects with high accuracy 49 landmarks on the human face. These

49 regions are believed to carry enough information to describe the emotional state and less person-related features. We developed a method to make the CNN encoders focus mainly on these regions by building the attention map of each input image. The attention map is achieved by inserting the landmarks on the input image then computing an Edge Glow filter that suppresses the textural information. Therefore, the CNN dedicated streams are extracting only the facial expression filters. We adopted dedicated CNN streams for the feature extraction step to guarantee an efficient extraction of the deep features from each input attention map. Moreover, dedicated streams offer flexible training by adapting the weights to each input rather than finding one suitable configuration for all of them. Also, the extraction is performed in parallel and not in sequence. Sequence computing makes the network unable to track the patterns between the inputs and will be influenced by the last input fed. Each stream outputs N-filter$^{3\times3}$ bank that is averaged and pooled into an N sized vector. The four vectors are concatenated to form the input of the classification sub-network, which is the third step of our network. Our network architecture is flexible to all models of deep CNN that can be used for extracting the features. We will present a comprehensive analysis of the CNN models presented earlier (Section 2.11). The third step classifies the extracted features through deep LSTM cells and fully connected layers as their configurations are mentioned in Figure5.17. The flatten filters from the dedicated CNN are firstly scaled by a first fully connected layer $FC1$ that has the exact size of the pooled feature vector. The need for feature adaption comes from the LSTM cell, which will process the outputs of $FC1$ as the values should be normalized, representing a homogeneous sequence. Also, FC1 is motivating the non-linearity, hence discovering more patterns. The main corps of our architecture is the Deep-LSTM, which is learning the temporal information across the four concatenated deep features. An LSTM gate processes its corresponding bin from the feature vector, then its output is fed to the next one. In addition, the first LSTM cell takes 256 activations as input performs feature reduction with a factor of two, leading to 128 activations that are further processed with a second LSTM cell. The final output computed from the Deep-LSTM is 64 activation neurons that represent the temporal patterns. The activations compression makes the prediction more efficient and avoids gradient loss when the number of classes is low, like in FER that has six or seven classes. Therefore, we used two LSTM cells to compress the activations of $FC1$ efficiently in addition to temporal processing. Moreover, Deep-LSTMs are more stable than one LSTM layer and increase the temporal correlation between the input activations. The final layer of our network is fully connected for predicting the emotion corresponding to the four input time samples. $FC2$ incorporates $Nclasses$ neurons with softmax classification layer. The 64 activations computed by Deep-LSTMs are discriminant enough to detect the correct emotion, which will be proved through an in-depth experimental analysis.

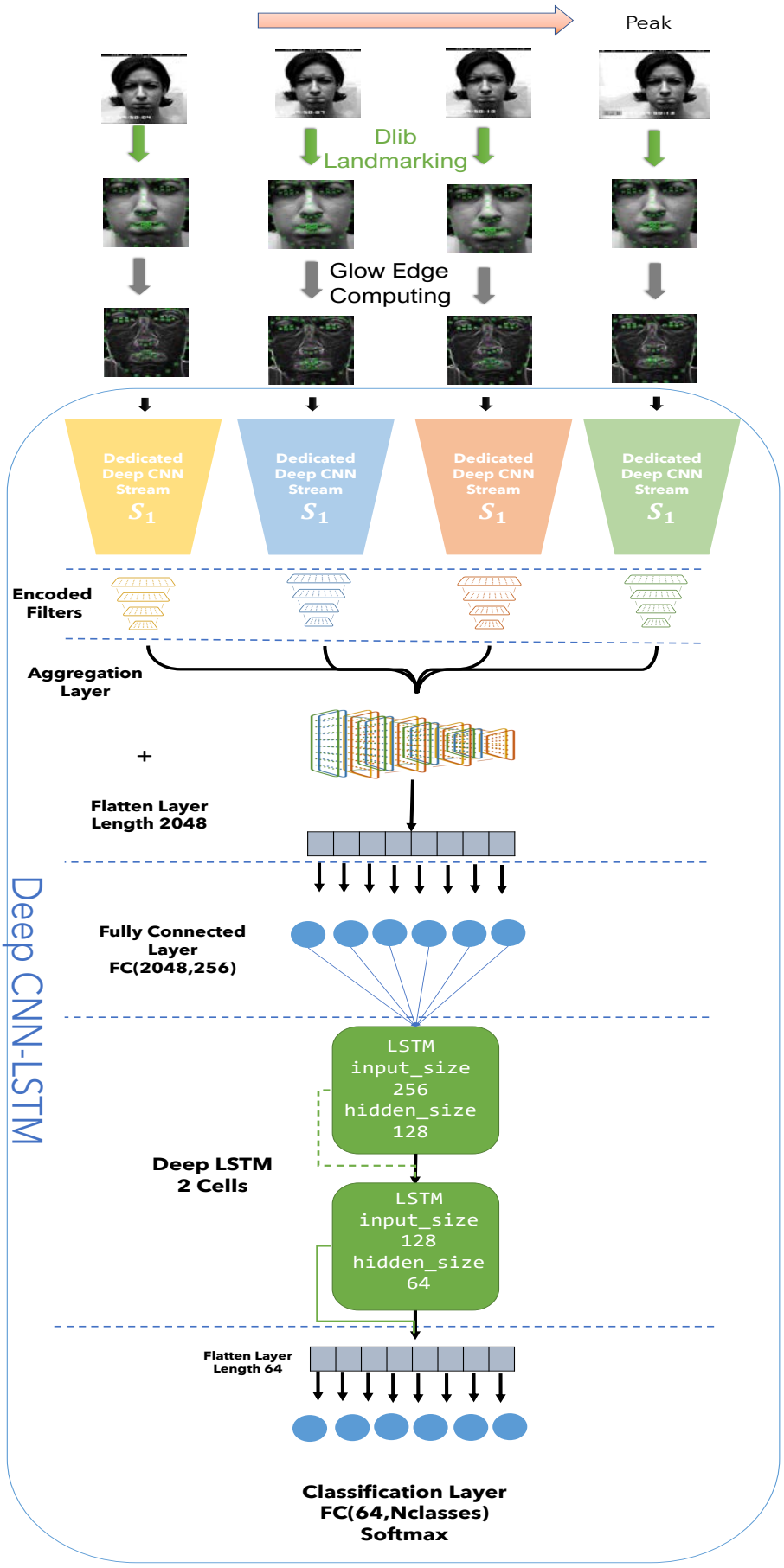

Figure 5.17: The overall pipeline of the proposed CNN-LSTM for dynamic FER model

5.3.1/ Experimental Analysis

This subsection is devoted to evaluating our proposed CNN-LSTM for dynamic person-independent facial expressions recognition. We adopted three benchmarks from the state-the-art that include sequence-based posed facial expressions. We set up a hard LOSO evaluation protocol by selecting one sequence of four samples per emotion, and we divided the individuals into ten folds, which guarantees the person-independent constraint. The conducted experiment considers different CNN deep models as feature extractors to find the ultimate configuration for our framework on the three benchmarks. This subsection presents the setup of the three benchmarks, the recorded results of the deep CNN, and comparative analysis against state-of-the-art works.

5.3.1.1/ Dynamic FER datasets

We adopted CK+, OuluCasia, and MMI benchmarks for evaluating our developed CNN-LSTM network. These datasets are sequence-based facial expressions. Hence, we present in the following the configuration adopted for each dataset.

- MMI: The MMI database includes a total of 213 sequences from 32 subjects labeled with six basic expressions (excluding "contempt"), where 205 sequences are recorded in frontal view. The sequence starts with a neutral expression and reaches the peak in the middle before returning to the neutral. In addition, MMI presents challenging conditions, i.e., there is large interpersonal variation as subjects do not perform the same expression uniformly, and many of them wear accessories (e.g., glasses, mustache). For experiments, the first frame (neutral face) and the three peak frames in each frontal sequence are usually selected to perform person-independent 10-fold cross-validation.

- CK+ and OuluCasia: The two benchmarks were already introduced in the earlier contribution related to static person-independent FER. We selected the first sample representing the neutral and the last three frames of the peak expression for the dynamic use case. The LOSO 10 folds cross-validation protocol is adopted similar to the MMI benchmark by dividing the subjects into 10 folds.

For the three benchmarks, the 10 folds do not share the same amount of images (subjects); also, we selected only one instance per subject for hard training of our network. Figure 5.18 shows an example of three individuals performing the same facial expression from the three benchmarks.

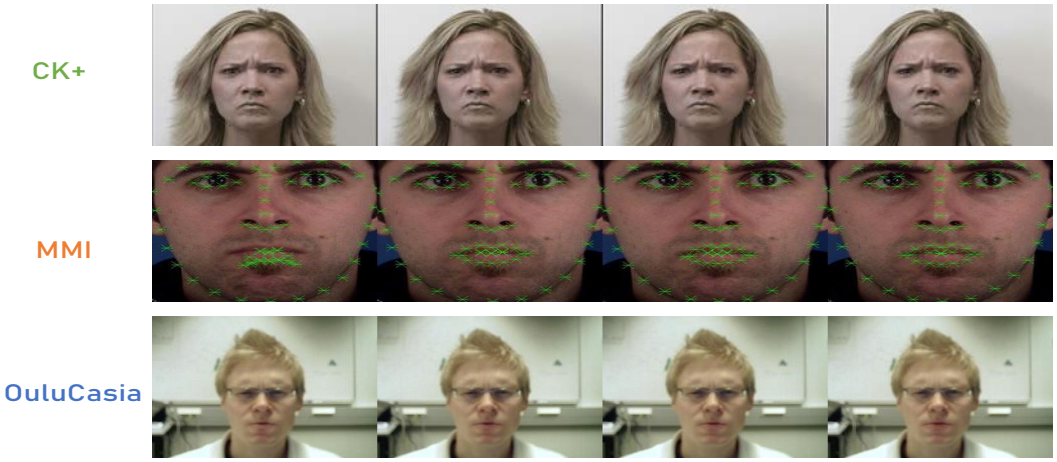

Figure 5.18: Examples of the angry sequence from the three datases

5.3.1.2/ EVALUATION OF THE PROPOSED CNN-LSTM

This experimental study evaluates the proposed CNN-LSTM network for dynamic person-independent FER regarding different CNN models for dedicated features encoding. We considered 12 well-known CNN models, including AlexNet, VGG16, VGG19, ResNet18, ResNet50, ResNet101, GoogLeNet, Inception.v2, Inception.v3, InceptionResNet.v1, InceptionResNet.v2, and DenseNet200. The comparative analysis relies on calculating the 10 folds average accuracy and the max accuracy that can be achieved on one fold. The recorded results are listed in Table 5.8.

It can be inferred from the recorded results that the achieved overall performance on the three datasets is very satisfying as the lowest average accuracy is above 81%. This fact proves that the proposed CNN-LSTM architecture is robust to person-independent constraints and grants sustainable recognition performance across the three benchmarks. Moreover, it also highlights that the architecture is generic and does not rely on any specific CNN feature encoder for the achieved minimum performance. Therefore, it is concluded that this performance is mainly thanks to the architecture that we proposed.

On the CK+ dataset, the best-performing configuration is ResNet18, as it reaches 99.41% average accuracy, while the 50 layers ResNet version comes second with 98.72% and the third best-performing model is Inception.v3. The worst-performing model is Inception.v2 with 96.19%, and the rest of the CNN models achieved close results averaged around 97%. According to the max accuracy, all the models reached a 100% rate on at least one fold. The CK+ dataset particularity relies on the uniformity of expressed emotion across all the individuals justifying the high average accuracies. On the other hand, the use of dynamic four samples as inputs improved the performance compared to the static case from 97.53% to 99.41%.

On the OuluCasia dataset, ResNet18 and 50 models managed to be the two top-

Table 5.8: Evaluation of the proposed CNN-LSTM according to different CNN models on the three benchmarks

Network CNN-LSTM	CK+		OuluCasia		MMI	
	Avg	Max	Avg	Max	Avg	Max
AlexNet	97.74	100	84.76	85.94	92.60	96.3
DenseNet200	96.85	100	84.33	85.94	92.17	95.96
GoogLeNet	97.85	100	84.48	85.94	92.32	95.96
Inception.v2	96.19	100	85.02	85.94	91.75	95.96
Inception.v3	98.52	100	84.88	85.94	92.75	95.96
InceptionResNet.v1	96.63	100	81.48	82.81	89.32	92.71
InceptionResNet.v2	97.30	100	81.91	82.81	89.75	92.71
ResNet101	98.36	100	84.91	87.5	93.03	96.3
ResNet18	99.41	100	86.62	89.06	93.46	96.3
ResNet50	98.72	100	85.19	87.5	93.03	96.3
VGG16	97.30	100	82.91	85.94	90.75	92.71
VGG19	97.74	100	84.05	85.94	91.89	92.71

performing by reaching 86.62% and 85.19%, respectively. The third best-performing model is Inception.v2, that controversy outperformed Inception.v3 on the OuluCasia as they scored 85.02% and 84.88%, respectively. The fourth-ranked model is the ResNet101 model by 84.91%. The lowest recorded average accuracy is 81.48% by InceptionResNet.v1, which was also among the worst-models on the CK+. Unlike on the CK+ benchmark, none of the evaluated models could reach 100% as max accuracy. Indeed, only the ResNet18 reached 89.06% and 87.5% for ResNet50, while the rest did not exceed 85.94%. The advantage of using four temporal samples is confirmed once again on the OuluCasia as the person-independent recognition rate increased from 77.32% in the static use case to 85.62%.

The MMI benchmarks recorded rates were higher than the OuluCasia ones. The ResNet models successfully ran to be ranked as the three top-performing models and were the only ones that reached above 93% average accuracy. The fourth-ranked model is Inception.v3 with 92.75% average accuracy, followed by the AlexNet model scored with 92.60%. InceptionResnet version 1 and 2 achieved below 90% as the lowest average accuracies (90.75%) on the MMI benchmark, while the rest of the models performed competitively, and their rates ranged are around 91% and 92%. In terms of maximum accuracy, only the ResNet and AlexNet architectures reached 96.3%. This fact proves that the MMI benchmark is challenging due to the large interpersonal variation as the facial expressions for the same emotion varies from a subject to another.

The ResNet models demonstrated outstanding performance compared to the rest of evaluated models. Indeed, the ResNet models are efficient feature extractors thanks to their

residual connections that link each layer with the previous one and conserve the features till the classification layers. Moreover, we state that the deep models do not outperform the light ones. For example, ResNet18 is always outperforming ResNet50 and 101 versions. Also, DenseNet200 could not be ranked among the top-performing ones on any dataset. The light models performed well thanks to the dedicated streams as few layers are required to find the optimal configuration and extract the relevant features.

5.3.1.3/ COMPARISON AGAINST STATE-OF-THE-ART

Table 5.9 illustrates the state-of-the-art work that we outperformed on the three considered benchmarks for dynamic person-independent FER based on the FER survey published in [220].

Table 5.9: State-of-the-art results on the three benchmarks

Datasets	Methods	Evaluation Protocol	Average Accuracy(%)
CK+	Peak-Piloted Deep Network [221]	10 folds	6 classes: 99.3
	Dynamic Geometrical Image Network [222]	10 folds	7 classes: 97.93
	ExpNet [223]	10 folds	6 classes: 97.28
	DTAGN(Weighted Sum) [224]	10 folds	7 classes: 96.94
	DTAGN(Joint) [224]	10 folds	7 classes: 97.25
	Supervised Scoring Ensemble [225]	10 folds	7 classes: 98.47
	Part-Based Hierarchical Bidirectional RNN [226]	10 folds	7 classes: 98.50
	Proposed ResNet18-LSTM	10 folds	6 classes: 99.44
MMI	Kim et al. 17 [66]	LOSO	6 classes: 78.61
	Dynamic Geometrical Image Network [222]	10 folds	6 classes: 81.53
	Hasani et al. 17 [112]	5 folds	6 classes: 77.50
	Hasani et al. 17 [55]	5 folds	6 classes: 78.68
	Part-Based Hierarchical Bidirectional RNN [226]	10 folds	6 classes: 81.18
	ExpNet [223]	10 folds	6 classes: 91.46
	Proposed ResNet18-LSTM	10 folds	6 classes: 93.46
OuluCasia	Peak-Piloted Deep Network [17]	10 folds	6 classes: 84.59
	Deeper Cascaded Peak-piloted Network [227]	10 folds	6 classes: 86.23
	DTAGN(Weighted Sum) [224]	10 folds	6 classes: 74.38
	DTAGN(Joint) [224]	10 folds	6 classes: 81.46
	Part-Based Hierarchical Bidirectional RNN [226]	10 folds	6 classes: 86.25
	Proposed ResNet18-LSTM	10 folds	6 classes: 86.62

It can be inferred from the state-of-the-art comparison that the proposed CNN-LSTM architecture with the ResNet18 configuration successfully outperformed many works in the context of person-independent FER. On the CK+ benchmark, all the methods perform well, reaching more than 90% but only two scored above 99%, including our proposed that scored the best accuracy of 99.44%. The MMI benchmark is more challenging than

CK+ as the top state-of-the-art average accuracy, concerning 10 folds cross-validation, is 91.46% reached by ExpNet [223]. Our ResNet18-LSTM managed to outperform it and scored 93.46% with a clear improvement of 2%. The rest literature works were stuck and could not bypass 82% average accuracy. The OuluCasia records highlight a similar performance of many methods, including ours. We underline that only our proposed model, in addition to the PPDN [221] and its deeper version [227] architecture, managed to guarantee an average accuracy above 86%. The Deeper Cascaded Peak-piloted Network [227] slightly outperforms PPDN with 0.02%.

The Peak-Piloted Deep Network (PPDN) [221] is one of the competitive works to our approach on the CK+ and OuluCasia benchmarks. The PPDN is based on a special-purpose back-propagation procedure referred to as peak gradient suppression (PGS) for network training. It considers two inputs representing the peak and non-peak samples encoded throw one CNN, and then the extracted features are fed to different classification layers to compare the cross-entropy of the peak and non-peak. ExpNet [223], which is the most competitive model on the MMI dataset, is a 3D approach for FER relying on computing 29D vectors characterizing the facial expression. The classification is done using the nearest neighbor rule with $k = 5$.

5.4/ CONCLUSION

In this chapter, we presented our contributions related to person-independent facial expressions recognition covering both static and dynamic scenarios. We proposed a framework for static person-independent FER based on a new textural features extractor referred to as referred to as orthogonal and parallel-based directions–generic query map binary patterns (OPD-GQMBP). OPD-GQMBP is applied on 49 patchs of 32×32 pixel size and the computed 32 vectors are concatenated to form the textural feature. The latter is further combined with shape-based one computed by the HOG method on a binary map representing the landmarks of the emotion's regions of interest. Hence, these landmarks helped our framework to fulfill the person-independent constraints. The proposed architecture was evaluated on 5 widely used benchmarks with respect the LOSO evaluation protocol to guarantee the person-independent use case. The evaluation demonstrated the outstanding performance of our framework against state-of-the-art deep and handcrafted methods. For dynamic person-independent FER, we proposed a deep CNN-LSTM network that considers 4 temporal samples leading to the peak of the emotion. We inserted 49 emotion-related landmarks to highlight the region of interest, then computing the egde filter to remove person-related visual features. Then, we fed each input sample to a dedicated CNN encoder to compute deep features. The 4 calculated features are concatenanted and forwarded to an LSTM block of two cells encoding the tempo-

ral patterns. The achieved 10 folds cross-validation metrics proved the superiority of the proposed CNN-LSTM network against the state-of-the-art, especially when used with ResNet18 model as features extractor.

6

CONCLUSION

6.1/ THESIS SUMMARY

This thesis was devoted to developing new and efficient handcrafted and deep-learning-based frameworks for face and facial expression recognition through image analysis covering all the steps of an image-based classification pipeline, starting with preprocessing, feature extraction, and classification. It has focused mainly on enhancing the recognition performance under delicate situations and strict evaluation protocols. For face recognition, we treated the severe lighting condition changes between the reference and probe samples. Also, face occlusion situations like wearing scarves and sunglasses have been addressed. On the other hand, we considered facial expression recognition with respect to the person-independent evaluation protocol with fewer training samples. The proposed architectures within the context of this thesis fulfilled important constraints related to performance stability, real-time inference, compatibility with end devices such as robots, and flexibility to other implementations and applications.

Through the first two chapters, we discussed the challenges related to face and emotion image-based analysis and how computer vision and machine learning managed to push this analysis forward similar to human perception. Also, we presented the typical configuration adopted to develop such frameworks and the methods used for feature extraction step and classification one. Moreover, we pointed out that the state-of-the-art still lacks efficient systems that reply to the mentioned constraints. Our main contributions, as presented in this thesis, covered the proposal of a new face descriptor referred to as Mixed Neighborhood Topology Cross Decoded Patterns (MNTCDP); a Pose Invariant Face Recognition (PIFR) system exploring the strength of Generative Adversarial Networks (GAN) in image translation; and two person-independent facial expression recognition systems to cover static and dynamic use cases with emotion-related landmarks attention maps.

The proposed MNTCDP-based face recognition framework incorporates the Nearest-

Neighbor classification rule with the city-block distance, which kept a low computational complexity and high matching power. The MNTCDP face descriptor incorporates a new neighborhood topology of 5×5 block size and selects only the prominent pixels instead of all the 25 ones. The selected pixels are further encoded through a discriminant kernel function that compares the gray-level values to well-defined local means. The conducted experiments on five benchmarks proved an outstanding and stable performance as compared to handcrafted and deep-based methods and models. The five benchmarks were chosen to cover all the discussed face recognition challenging conditions. The results highlighted the robustness of the MNTCDP descriptor to lighting changes, especially on the Extended Yale B database and face occlusions simulated over the AR face database. The MNTCDP contribution was further validated by an actual implementation within the UTBM Human Support Robot (Toyota HSR 88), allowing this robot to recognize the members of our laboratory according to a reference base that we collected containing few samples per individual.

The second contribution presented in this manuscript handled the challenge of Pose Invariant Face Recognition (PIFR), which is the most complex problem that a face classification framework can face. We presented how the state-of-the-art is dealing with it where the frameworks are supposed to include profile samples of the individuals so the probe ones could be recognized, which is hard to be satisfied in real-world implementations. What we proposed is based on image translation through Generative Adversarial Networks (GAN). Our PIFR framework takes the profile probe input and generates its frontal one based on the GAN person-independent learning, then classifies it through residual-CNN module trained only on frontal images. The experiment we conducted on a benchmark that we created from the literature ones, by assuring no person-overlapping between the train and test sets, proved the improvement bought by the GAN frontalization technique. The experiment highlighted an improvement of nearly 40% and concluded that the ResNet101 for classification and the U-Net generator offer the highest frontalization quality, keeping an inference time less than 10ms for the overall PIFR framework.

On the other hand, we proposed two systems for facial expression recognition covering both static and dynamic-based analysis. We considered the person-independent evaluation constraint, which is the most challenging scenario for FER and motivates the generic implementations within end devices. In the static analysis, only one sample is considered to output the emotion label. We built a system based on combining textural and shape features extracted from 49 emotion-related landmarks detected by the Dlib. To efficiently encode the textural information, we proposed a new local descriptor referred to as Orthogonal and Parallel-based Directions–Generic Query map Binary Patterns (OPD-GQMBP). It supports user-specified neighborhood size (3,5,7,9..) and considers the pixels in linear distributions that captures more discriminant features. The textural features extracted from local regions of 32×32 pixels centered on the 49 landmarks are combined

with shape-based one computed using HOG descriptor. We interpolated the locations of the 49 landmarks forming a binary map describing the dominant emotion. All the features are fed to the PCA dimensionality reduction and SVM classifier predicting the facial expression. The best kernel size configuration for OPD-GQMBP as facial expression descriptor is seven, leading to the highest recognition rates on five widely used benchmarks for person-independent FER. The recorded results proved the superiority of the proposed FER framework, as it outperformed the handcrafted descriptors and deep models of the literature and reached state-of-the-art accuracy on the five benchmarks. As a second contribution in facial expression recognition, we proposed a deep model for dynamic-based FER that considers four temporal samples finishing with the peak of the expression. We applied a landmark-based edge mask on each sample to promote the person-independent analysis and reduce the person-related visual features. Our deep model includes four CNN dedicated streams to extract the features from the four samples; then, it concatenates the four features before feeding them to an LSTM block that manages to encode the temporal patterns across the four samples. The recorded results on three state-of-the-art benchmarks for dynamic facial expression recognition proved that the ResNet18-LSTM configuration outperformed the evaluated CNN configurations and surpassed the accuracies reported in the literature.

6.2/ FUTURE WORKS AND PERSPECTIVES

We are working on the first perspective related to performing more ablation studies on the dynamic person-independent FER. We will run experiments to investigate the effect of the number of input samples and other attention maps that can enhance more the person-independent performance. Also, we will extend the adopted benchmarks by considering the wild ones collected from the internet. The wild datasets present more challenges to FER frameworks as the expressions are not uniform between the individuals, and the background also presents massive changes from a scene to another. Moreover, the processing time should be analyzed to find the optimal configuration leading to the best performance and response time. Finally, we started writing a paper about this contribution to send it to a high-impact journal for publication.

The accuracy recorded by the developed FER system combined with the proposed OPD-GQMBP descriptor was the highest compared to those reported by the existing state-of-the-art FER systems that adopted the same protocol (person-independent LOSO). Although our system indeed managed to outperform many state-of-the-art systems, it needs improvement on databases containing Asian individuals (e.g., JAFFE and Oulu CASIA) because they tend to present similar facial features, which confuse the classifier in the case of the LOSO protocol. We believe that the ultimate solution is to reduce the

number of extracted appearance features and focus on the binary patch calculated based on the detected landmarks. This patch should incorporate more information, not only the landmark location, and handcraft a shape descriptor to obtain the most prominent information. Moreover, this proposal will help develop generic person-independent FER systems because the input images will be coded into a standard patch. Even though our framework, along with the proposed descriptor, performs efficiently in computational time, we think there is room for improvement using other dimension reduction strategies and/or landmark selections. We are also investigating an enhancement of the performance of our framework by utilizing other sophisticated classifiers than an SVM and combining learnable features with the proposed OPD-GQMBP descriptor. In addition, we intend to extend the set of the studied emotion classes with the compound classes, reaching 20 different classes. We also considered creating a mixed database from existing databases to gather all challenges in a unique benchmark, offering more challenging testing and evaluation.

Similar to implementing the MNTCDP-based face recognition system within the Human Support Robot, we also intend to embed the static facial expression framework on the HSR. Therefore, the robot will recognize the person and its emotional state through two independent frameworks. To increase the FER performance, we will change the setup of the training database, making it person-dependent since the environment and the subjects interacting with the robot will not change. Hence, only one reference database is enough for the two tasks that also will optimize the computational resources. Collecting this database is the only step to perform, we imagine a standard recording protocol similar to the state-of-the-art benchmarks simulating various lighting conditions and backgrounds.

Moreover, we intend to improve the PIFR performance on the Combined-PIFR database by enhancing the quality of the frontalized profile images. This could be achieved through new GAN architectures and/or adversarial loss functions by including perceptual and qualitative losses. Furthermore, we aim to apply image translation based on GAN to generate 3D models conditioned by the profile image so the classification can be more accurate. Furthermore, we plan to generalize the frontalization process to be generic, and then it can be evaluated on cross databases and real-life probe images. Also, the classification sub-system could be addressed to get the maximum from the frontalized images. We can rely on an ablation study to conduct an efficient training of the CNNs. In addition, the frontalization technique can be included in other facial applications suffering from pose variations, such as facial expression recognition, gender classification, and age estimation. This technique reduces the computational complexity by building an efficient system dealing with the standard view (frontal) and translate any other view before processing it.

The proposed face recognition frameworks can be improved by performing a verification step that checks if the predicted identity matches the input image. We will build a two-inputs verification deep CNN model taking the probe image and the reference sample from the predicted class. The verification model will compare the two inputs as one inference and decide if they belong to the same class or no. Such verification systems should reach a high-performance rate and likely improve the overall classification. If the predicted class matches the probe image, the algorithm outputs this predicted class as the final output. If the probe and reference images are not matching, the verification is done using other reference samples of the same predicted class. If no match is found, the probe image is reclassified, ignoring the previously predicted classes until finding the verification match.

6.3/ PUBLICATIONS

Scientific Journals:

- "Mixed neighborhood topology cross decoded patterns for image-based face recognition". Kas, Mohamed; El-merabet, Youssef; Ruichek, Yassine; Messoussi, Rochdi;; Expert Systems with Applications- Elsevier; 2018 Impact Factor: 4.292

- "Local feature extraction based facial emotion recognition: A survey". Slimani, Khadija; Kas, Mohamed; El Merabet, Youssef; Ruichek, Yassine; Messoussi, Rochdi;; International Journal of Electrical and Computer Engineering- IAES Institute of Advanced Engineering and Science; 2020 Impact Factor: 0.32 SJR

- "A comprehensive comparative study of handcrafted methods for face recognition LBP-like and non-LBP operators". Kas, Mohamed; El-merabet, Youssef; Ruichek, Yassine; Messoussi, Rochdi;; Multimedia Tools and Applications- Springer; 2020 Impact Factor: 2.101

- "Repulsive-and-attractive local binary gradient contours: New and efficient feature descriptors for texture classification". El Khadiri, Issam; Kas, Mohamed; El Merabet, Youssef; Ruichek, Yassine; Touahni, Raja;; Information Sciences- Elsevier; 2018 Impact Factor: 5.524

- "Multi-level directional cross binary patterns: New handcrafted descriptor for SVM-based texture classification". Kas, Mohamed; El-merabet, Youssef; Ruichek, Yassine; Messoussi, Rochdi;; Engineering Applications of Artificial Intelligence- Elsevier; 2020 Impact Factor: 3.526

- "New framework for person-independent facial expression recognition combining textural and shape analysis through new feature extraction approach". Kas, Mohamed; El-merabet, Youssef; Ruichek, Yassine; Messoussi, Rochdi;; Information Sciences- Elsevier; 2021 Impact Factor: 5.524
- "Saliency Heat-Map as Visual Attention for Autonomous Driving Using Generative Adversarial Network (GAN)". Lateef, Fahad; Kas, Mohamed; Ruichek, Yassine;; IEEE Transactions on Intelligent Transportation Systems- IEEE; 2021 Impact Factor: 7.42
- "Generative Adversarial Networks for 2D-based CNN Pose-Invariant Face Recognition". Kas, Mohamed; El-merabet, Youssef; Ruichek, Yassine; Messoussi, Rochdi;; Expert Systems with Applications- Elsevier; Under Review Impact Factor: 4.292
- "Coarse-to-Fine SVD-GAN based Framework for Enhanced Frame Synthesis". Kas, Mohamed; Kajo, Ibrahim; Ruichek, Yassine; Engineering Applications of Artificial Intelligence- Elsevier; Under Review Impact Factor: 3.526

International Conferences:

- Kas, Mohamed, et al. "Survey on Local Binary Pattern Descriptors for Face Recognition." Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society. ACM, 2019.
- Kas, Mohamed, et al. "Comprehensive Experimental Analysis of Handcrafted Descriptors for Face Recognition." 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT). IEEE, 2018.
- Mohamed Kas, Youssef El Merabet, Yassine Ruichek, Rochdi Messoussi (11-13 December 2017) Local Directional Multi Radius Binary Pattern: Novel Descriptor for Face Recognition Application; the 9th International Conference on Soft Computing and Pattern Recognition- Advances in Intelligent Systems and Computing - Springer - MIR Labs USA Marrakech Maroc
- Slimani, K., Kas, M., El Merabet, Y., Messoussi, R., Ruichek, Y. (2018, March). Facial emotion recognition: A comparative analysis using 22 LBP variants. In Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence (pp. 88-94). ACM.

BIBLIOGRAPHY

[1] J. Morton, M. H. Johnson, Conspec and conlern: a two-process theory of infant face recognition., Psychological review 98 (2) (1991) 164.

[2] L. T. Likova, M. Mei, K. N. Mineff, S. C. Nicholas, Learning face perception without vision: Rebound learning effect and hemispheric differences in congenital vs late-onset blindness, Electronic Imaging 2019 (12) (2019) 237–1.

[3] N. Kanwisher, G. Yovel, Face perception, Handbook of neuroscience for the behavioral sciences.

[4] T. Farroni, E. Menon, S. Rigato, M. H. Johnson, The perception of facial expressions in newborns, European Journal of Developmental Psychology 4 (1) (2007) 2–13.

[5] V. Bruce, A. Young, Understanding face recognition, British journal of psychology 77 (3) (1986) 305–327.

[6] M. J. Peltola, J. M. Leppänen, S. Mäki, J. K. Hietanen, Emergence of enhanced attention to fearful faces between 5 and 7 months of age, Social cognitive and affective neuroscience 4 (2) (2009) 134–142.

[7] S. M. Crouzet, H. Kirchner, S. J. Thorpe, Fast saccades toward faces: face detection in just 100 ms, Journal of vision 10 (4) (2010) 16–16.

[8] S. Caharel, M. Ramon, B. Rossion, Face familiarity decisions take 200 msec in the human brain: Electrophysiological evidence from a go/no-go speeded task, Journal of Cognitive Neuroscience 26 (1) (2014) 81–95.

[9] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, R. von der Heydt, A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization., Psychological bulletin 138 (6) (2012) 1172.

[10] M. J. Farah, K. D. Wilson, M. Drain, J. N. Tanaka, What is" special" about face perception?, Psychological review 105 (3) (1998) 482.

[11] B. Rossion, Picture-plane inversion leads to qualitative changes of face perception, Acta psychologica 128 (2) (2008) 274–289.

[12] B. Rossion, The composite face illusion: A whole window into our understanding of holistic face perception, Visual Cognition 21 (2) (2013) 139–253.

[13] J. W. Tanaka, M. J. Farah, Parts and wholes in face recognition, The Quarterly journal of experimental psychology 46 (2) (1993) 225–245.

[14] J. Sergent, An investigation into component and configural processes underlying face perception, British journal of psychology 75 (2) (1984) 221–242.

[15] J. W. Tanaka, J. A. Sengco, Features and their configuration in face recognition, Memory & cognition 25 (5) (1997) 583–592.

[16] D. J. Felleman, D. C. Van Essen, Distributed hierarchical processing in the primate cerebral cortex., Cerebral cortex (New York, NY: 1991) 1 (1) (1991) 1–47.

[17] J. V. Haxby, E. A. Hoffman, M. I. Gobbini, The distributed human neural system for face perception, Trends in cognitive sciences 4 (6) (2000) 223–233.

[18] N. Kanwisher, J. McDermott, M. M. Chun, The fusiform face area: a module in human extrastriate cortex specialized for face perception, Journal of neuroscience 17 (11) (1997) 4302–4311.

[19] B. Rossion, S. Caharel, Erp evidence for the speed of face categorization in the human brain: Disentangling the contribution of low-level visual cues from face perception, Vision research 51 (12) (2011) 1297–1311.

[20] V. Ojansivu, J. Heikkilä, Blur insensitive texture classification using local phase quantization, in: International conference on image and signal processing, Springer, 2008, pp. 236–243.

[21] U. Raghavendra, U. R. Acharya, A. Gudigar, R. Shetty, N. Krishnananda, U. Pai, J. Samanth, C. Nayak, Automated screening of congestive heart failure using variational mode decomposition and texture features extracted from ultrasound images, Neural Computing and Applications 28 (10) (2017) 2869–2878. doi:10.1007/s00521-017-2839-5.
URL https://doi.org/10.1007/s00521-017-2839-5

[22] M. Maktabdar Oghaz, M. A. Maarof, M. F. Rohani, A. Zainal, S. Z. M. Shaid, An optimized skin texture model using gray-level co-occurrence matrix, Neural Computing and Applicationsdoi:10.1007/s00521-017-3164-8.
URL https://doi.org/10.1007/s00521-017-3164-8

[23] L. Moraru, S. Moldovanu, L. T. Dimitrievici, A. S. Ashour, N. Dey, Texture anisotropy technique in brain degenerative diseases, Neural Computing and Applications 30 (5) (2018) 1667–1677. doi:10.1007/s00521-016-2777-7.
URL https://doi.org/10.1007/s00521-016-2777-7

[24] M. H. Kayyal, J. A. Russell, Language and emotion: certain english–arabic translations are not equivalent, Journal of Language and Social Psychology 32 (3) (2013) 261–271.

[25] Z.-H. Huang, W.-J. Li, J. Shang, J. Wang, T. Zhang, Non-uniform patch based face recognition via 2d-dwt, Image Vis Comput 37 (2015) 12–19.

[26] T. Abhishree, J. Latha, K. Manikantan, S. Ramachandran, Face recognition using gabor filter based feature extraction with anisotropic diffusion as a pre-processing technique, procedia Computer Science 45 (2015) 312–321.

[27] N. Guan, D. Tao, Z. Luo, B. Yuan, Nenmf: An optimal gradient method for nonnegative matrix factorization, IEEE Transactions on Signal Processing 60 (6) (2012) 2882–2898.

[28] M. El Aroussi, M. El Hassouni, S. Ghouzali, M. Rziza, D. Aboutajdine, Local appearance based face recognition method using block based steerable pyramid transform, Signal Processing 91 (1) (2011) 38–50.

[29] P. Secchi, S. Vantini, P. Zanini, Hierarchical independent component analysis: a multi-resolution non-orthogonal data-driven basis, Computational Statistics & Data Analysis 95 (2016) 133–149.

[30] M. Bereta, W. Pedrycz, M. Reformat, Local descriptors and similarity measures for frontal face recognition: a comparative analysis, Journal of Visual Communication and Image Representation 24 (8) (2013) 1213–1231.

[31] A. Fathi, P. Alirezazadeh, F. Abdali-Mohammadi, A new global-gabor-zernike feature descriptor and its application to face recognition, Journal of Visual Communication and Image Representation 38 (2016) 65–72.

[32] G. D. Cavalcanti, T. I. Ren, J. F. Pereira, Weighted modular image principal component analysis for face recognition, Expert Systems with Applications 40 (12) (2013) 4971–4977.

[33] G.-F. Lu, J. Zou, Y. Wang, Incremental complete lda for face recognition, Pattern Recognition 45 (7) (2012) 2510–2521.

[34] W. Yang, Z. Wang, B. Zhang, Face recognition using adaptive local ternary patterns method, Neurocomputing 213 (2016) 183–190.

[35] M. Topi, O. Timo, P. Matti, S. Maricor, Robust texture classification by subsets of local binary patterns, in: Pattern Recognition, 2000. Proceedings. 15th International Conference on, Vol. 3, IEEE, 2000, pp. 935–938.

[36] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, IEEE Trans Image Process 19 (6) (2010) 1635–1650.

[37] T. Jabid, M. H. Kabir, O. Chae, Local directional pattern (ldp) for face recognition, in: Consumer Electronics (ICCE), 2010 Digest of Technical Papers International Conference on, IEEE, 2010, pp. 329–330.

[38] M. Heikkilä, M. Pietikäinen, C. Schmid, Description of interest regions with center-symmetric local binary patterns, in: ICVGIP, Vol. 6, Springer, 2006, pp. 58–69.

[39] X. Fu, W. Wei, Centralized binary patterns embedded with image euclidean distance for facial expression recognition, in: Fourth International Conference on Natural Computation, 2008, Vol. 4, 2008, pp. 115–119.

[40] A. R. Rivera, J. R. Castillo, O. O. Chae, Local directional number pattern for face analysis: Face and expression recognition, IEEE Trans Image Process 22 (5) (2013) 1740–1752.

[41] H. Zeng, J. Chen, X. Cui, C. Cai, K.-K. Ma, Quad binary pattern and its application in mean-shift tracking, Neurocomputing 217 (2016) 3–10.

[42] Y. El merabet, Y. Ruichek, Local concave-and-convex micro-structure patterns for texture classification, Pattern Recognition.

[43] I. E. Khadiri, M. Kas, Y. E. Merabet, Y. Ruichek, R. Touahni, Repulsive-and-attractive local binary gradient contours: New and efficient feature descriptors for texture classification, Information Sciences.

[44] I. E. Khadiri, A. Chahi, Y. E. merabet, Y. Ruichek, R. Touahni, Local directional ternary pattern: A new texture descriptor for texture classification, Computer Vision and Image Understanding 169 (2018) 14–27.

[45] T. P. Nguyen, N.-S. Vu, A. Manzanera, Statistical binary patterns for rotational invariant texture classification, Neurocomputing 173 (2016) 1565–1577.

[46] S. Chakraborty, S. K. Singh, P. Chakraborty, Local quadruple pattern: A novel descriptor for facial image recognition and retrieval, Computers & Electrical Engineering 62 (2017) 92–104.

[47] M. Verma, B. Raman, Local neighborhood difference pattern: A new feature descriptor for natural and texture image retrieval, Multimedia Tools and Applications 77 (10) (2018) 11843–11866.

[48] L. Liu, P. Fieguth, Y. Guo, X. Wang, M. Pietikäinen, Local binary features for texture classification: Taxonomy and experimental study, Pattern Recognition 62 (2017) 135–160.

[49] B. Heisele, T. Serre, T. Poggio, A component-based framework for face detection and identification, International Journal of Computer Vision 74 (2) (2007) 167–181.

[50] S. K. Vipparthi, S. K. Nagar, Local extreme complete trio pattern for multimedia image retrieval system, International Journal of Automation and Computing 13 (5) (2016) 457–467.

[51] F. Ouslimani, A. Ouslimani, Z. Ameur, Rotation-invariant features based on directional coding for texture classification, Neural Computing and Applications (2018) 1–8.

[52] R. Mehta, K. Egiazarian, Dominant rotated local binary patterns (drlbp) for texture classification, Pattern Recognition Letters 71 (2016) 16–22.

[53] S. K. Vipparthi, S. Murala, S. K. Nagar, A. B. Gonde, Local gabor maximum edge position octal patterns for image retrieval, Neurocomputing 167 (2015) 336–345.

[54] K. Song, Y. Yan, Y. Zhao, C. Liu, Adjacent evaluation of local binary pattern for texture classification, Journal of Visual Communication and Image Representation 33 (2015) 323–339.

[55] J. Sun, G. Fan, L. Yu, X. Wu, Concave-convex local binary features for automatic target recognition in infrared imagery, EURASIP Journal on Image and Video Processing 2014 (1) (2014) 23.

[56] Y. Zhao, W. Jia, R.-X. Hu, H. Min, Completed robust local binary pattern for texture classification, Neurocomputing 106 (2013) 68–76.

[57] M. S. Islam, et al., Local gray code pattern (lgcp): A robust feature descriptor for facial expression recognition, International Journal of Science and Research (IJSR), India Online ISSN (2013) 2319–7064.

[58] M. Subrahmanyam, R. Maheshwari, R. Balasubramanian, Local maximum edge binary patterns: a new descriptor for image retrieval and object tracking, Signal Processing 92 (6) (2012) 1467–1479.

[59] R. Gupta, H. Patil, A. Mittal, Robust order-based methods for feature description, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 334–341.

[60] X. Wu, J. Sun, An extended center-symmetric local ternary patterns for image retrieval, in: International Conference on Computer Science, Environment, Ecoinformatics, and Education, Springer, 2011, pp. 359–364.

[61] C.-I. Chang, Y. Chen, Gradient texture unit coding for texture analysis, Optical Engineering 43 (8) (2004) 1891–1903.

[62] Y. Ruichek, et al., Attractive-and-repulsive center-symmetric local binary patterns for texture classification, Engineering Applications of Artificial Intelligence 78 (2019) 158–172.

[63] T. Chakraborti, B. McCane, S. Mills, U. Pal, Loop descriptor: Local optimal-oriented pattern, IEEE Signal Processing Letters 25 (5) (2018) 635–639.

[64] H. K. Bashier, L. S. Hoe, L. T. Hui, M. F. Azli, P. Y. Han, W. K. Kwee, M. S. Sayeed, Texture classification via extended local graph structure, Optik-International Journal for Light and Electron Optics 127 (2) (2016) 638–643.

[65] S. Rajput, D. J. Bharti, A face recognition using linear -diagonal binary graph pattern feature extraction method, International Journal in Foundations of Computer Science andTechnology (IJFCST 6 (2) (2016) 55–65.

[66] Y. Kaya, Ö. F. Ertuğrul, R. Tekin, Two novel local binary pattern descriptors for texture analysis, Applied Soft Computing 34 (2015) 728–735.

[67] S. Dong, J. Yang, C. Wang, Y. Chen, D. Sun, A new finger vein recognition method based on the difference symmetric local graph structure (dslgs), International Journal of Signal Processing, Image Processing and Pattern Recognition 8 (10) (2005) 71–80.

[68] C. Silva, T. Bouwmans, C. Frélicot, An extended center-symmetric local binary pattern for background modeling and subtraction in videos, in: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP 2015, 2015.

[69] S. Dong, J. Yang, Y. Chen, C. Wang, X. Zhang, D. S. Park, Finger vein recognition based on multi-orientation weighted symmetric local graph structure., KSII Transactions on Internet & Information Systems 9 (10).

[70] M. F. A. Abdullah, M. S. Sayeed, K. S. Muthu, H. K. Bashier, A. Azman, S. Z. Ibrahim, Face recognition with symmetric local graph structure (slgs), Expert Syst Appl 41 (14) (2014) 6131–6137.

[71] A. Fernández, M. X. Álvarez, F. Bianconi, Texture description through histograms of equivalent patterns, Journal of mathematical imaging and vision 45 (1) (2013) 76–102.

[72] M. Sayeed, I. Yusof, H. K. Bashier, M. Hossen, M. F. A. Abdullah, et al., Plant identification based on leaf shape and texture pattern using local graph structure (lgs), Aust J Basic Appl Sci 7 (11) (2013) 29–35.

[73] A. Fernández, M. X. Álvarez, F. Bianconi, Image classification with binary gradient contours, Optics and Lasers in Engineering 49 (9-10) (2011) 1177–1184.

[74] H. Zeng, X. Dong, X. Wang, Improved center-symmetric local binary pattern descriptor for local feature region description, Energy Procedia (11) (2011) 1032–1038.

[75] B. Zhang, L. Zhang, D. Zhang, L. Shen, Directional binary code with application to polyu near-infrared face database, Pattern Recognit Lett 31 (14) (2010) 2337–2344.

[76] B. Xu, P. Gong, E. Seto, R. Spear, Comparison of gray-level reduction and different texture spectrum encoding methods for land-use classification using a panchromatic ikonos image, Photogrammetric Engineering & Remote Sensing 69 (5) (2003) 529–536.

[77] F. J. Madrid-Cuevas, R. M. Carnicer, M. P. Villegas, N. L. F. García, A. C. Poyato, Simplified texture unit: a new descriptor of the local texture in gray-level images, Lecture notes in computer science (2003) 470–477.

[78] Y. Zhao, R.-G. Wang, W.-M. Wang, W. Gao, Local quantization code histogram for texture classification, Neurocomputing 207 (2016) 354–364.

[79] R. Mehta, K. O. Egiazarian, Rotated local binary pattern (rlbp)-rotation invariant texture descriptor., in: ICPRAM, 2013, pp. 497–502.

[80] Y. Zhao, D.-S. Huang, W. Jia, Completed local binary count for rotation invariant texture classification, IEEE transactions on image processing 21 (10) (2012) 4492–4497.

[81] L. Nanni, S. Brahnam, A. Lumini, A local approach based on a local binary patterns variant texture descriptor for classifying pain states, Expert Systems with Applications 37 (12) (2010) 7888–7894.

[82] H. Jin, Q. Liu, H. Lu, X. Tong, Face detection using improved lbp under bayesian framework, in: Image and Graphics (ICIG'04), Third International Conference on, IEEE, 2004, pp. 306–309.

[83] T. Ahonen, E. Rahtu, V. Ojansivu, J. Heikkila, Recognition of blurred faces using local phase quantization, in: 19th International Conference on Pattern Recognition, 2008, pp. 1–4.

[84] D.-C. He, L. Wang, Unsupervised textural classification of images using the texture spectrum, Pattern Recognition 25 (3) (1992) 247–255.

[85] D.-C. He, L. Wang, Texture unit, texture spectrum, and texture analysis, IEEE transactions on Geoscience and Remote Sensing 28 (4) (1990) 509–512.

[86] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, Pcanet: A simple deep learning baseline for image classification?, IEEE Transactions on Image Processing 24 (12) (2015) 5017–5032.

[87] J. Lu, V. E. Liong, X. Zhou, J. Zhou, Learning compact binary face descriptor for face recognition, IEEE transactions on pattern analysis and machine intelligence 37 (10) (2015) 2041–2056.

[88] F. Lateef, Y. Ruichek, Survey on semantic segmentation using deep learning techniques, Neurocomputing 338 (2019) 321–348.

[89] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM transactions on intelligent systems and technology (TIST) 2 (3) (2011) 27.

[90] A. A. Fathima, S. Ajitha, V. Vaidehi, M. Hemalatha, R. Karthigaiveni, R. Kumar, Hybrid approach for face recognition combining gabor wavelet and linear discriminant analysis, in: 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), IEEE, 2015, pp. 220–225.

[91] F. Juefei-Xu, K. Luu, M. Savvides, Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios, IEEE Transactions on Image Processing 24 (12) (2015) 4780–4795.

[92] Y. Yan, H. Wang, D. Suter, Multi-subregion based correlation filter bank for robust face recognition, Pattern Recognition 47 (11) (2014) 3487–3501.

[93] K. Simonyan, O. M. Parkhi, A. Vedaldi, A. Zisserman, Fisher vector faces in the wild., in: BMVC, Vol. 2, 2013, p. 4.

[94] C. Ding, D. Tao, Robust face recognition via multimodal deep face representation, IEEE Transactions on Multimedia 17 (11) (2015) 2049–2058.

[95] R. Sharma, M. Patterh, A new pose invariant face recognition system using pca and anfis, Optik 126 (23) (2015) 3483–3487.

[96] M. Moussa, M. Hmila, A. Douik, A novel face recognition approach based on genetic algorithm optimization, Studies in Informatics and Control 27 (1) (2018) 127–134.

[97] A. Mian, M. Bennamoun, R. Owens, An efficient multimodal 2d-3d hybrid approach to automatic face recognition, IEEE transactions on pattern analysis and machine intelligence 29 (11) (2007) 1927–1943.

[98] S. Karanwal, M. Diwakar, Od-lbp: Orthogonal difference-local binary pattern for face recognition, Digital Signal Processing 110 (2021) 102948. doi:https://doi.

org/10.1016/j.dsp.2020.102948.
URL https://www.sciencedirect.com/science/article/pii/S1051200420302931

[99] H. Cho, R. Roberts, B. Jung, O. Choi, S. Moon, An efficient hybrid face recognition algorithm using pca and gabor wavelets, International Journal of Advanced Robotic Systems 11 (4) (2014) 59.

[100] J. K. Sing, S. Chowdhury, D. K. Basu, M. Nasipuri, An improved hybrid approach to face recognition by fusing local and global discriminant features, International Journal of Biometrics 4 (2) (2012) 144–164.

[101] P. Kamencay, M. Zachariasova, R. Hudec, R. Jarina, M. Benco, J. Hlubik, A novel approach to face recognition using image segmentation based on spca-knn method, Radioengineering 22 (1) (2013) 92–99.

[102] C. Shan, S. Gong, P. W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, Image and vision Computing 27 (6) (2009) 803–816.

[103] B. Zhang, G. Liu, G. Xie, Facial expression recognition using lbp and lpq based on gabor wavelet transform, in: 2016 2nd IEEE International Conference on Computer and Communications (ICCC), IEEE, 2016, pp. 365–369.

[104] K. Lekdioui, R. Messoussi, Y. Ruichek, Y. Chaabi, R. Touahni, Facial decomposition for expression recognition using texture/shape descriptors and svm classifier, Signal Processing: Image Communication 58 (2017) 300–312.

[105] F. Makhmudkhujaev, M. Abdullah-Al-Wadud, M. T. B. Iqbal, B. Ryu, O. Chae, Facial expression recognition with local prominent directional pattern, Signal Processing: Image Communication 74 (2019) 1–12.

[106] M. Shin, M. Kim, D.-S. Kwon, Baseline cnn structure analysis for facial expression recognition, in: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE, 2016, pp. 724–729.

[107] Z. Yu, C. Zhang, Image based static facial expression recognition with multiple deep network learning, in: Proceedings of the 2015 ACM on international conference on multimodal interaction, 2015, pp. 435–442.

[108] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2983–2991.

[109] C. Ding, J. Choi, D. Tao, L. S. Davis, Multi-directional multi-level dual-cross patterns for robust face recognition, IEEE transactions on pattern analysis and machine intelligence 38 (3) (2016) 518–531.

[110] L. Liu, L. Zhao, Y. Long, G. Kuang, P. Fieguth, Extended local binary patterns for texture classification, Image and Vision Computing 30 (2) (2012) 86–99.

[111] F. S. Samaria, A. C. Harter, Parameterisation of a stochastic model for human face identification, in: Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on, IEEE, 1994, pp. 138–142.

[112] K.-C. Lee, J. Ho, D. J. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Transactions on pattern analysis and machine intelligence 27 (5) (2005) 684–698.

[113] P. J. Phillips, H. Moon, S. A. Rizvi, P. J. Rauss, The feret evaluation methodology for face-recognition algorithms, IEEE Transactions on pattern analysis and machine intelligence 22 (10) (2000) 1090–1104.

[114] P. Huang, G. Gao, C. Qian, G. Yang, Z. Yang, Fuzzy linear regression discriminant projection for face recognition, IEEE Access 5 (2017) 4340–4349.

[115] P. Huang, G. Gao, C. Qian, G. Yang, Z. Yang, Fuzzy linear regression discriminant projection for face recognition, IEEE Access 5 (2017) 4340–4349.

[116] W. Yang, Z. Wang, B. Zhang, Face recognition using adaptive local ternary patterns method, Neurocomputing 213 (2016) 183–190.

[117] P. Karczmarek, A. Kiersztyn, W. Pedrycz, M. Dolecki, An application of chain code-based local descriptor and its extension to face recognition, Pattern Recognition 65 (2017) 26–34.

[118] L. Ding, A. M. Martinez, Features versus context: An approach for precise and detailed detection and delineation of faces and facial features, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (11) (2010) 2022–2038.

[119] T. Liu, J.-X. Mi, Y. Liu, C. Li, Robust face recognition via sparse boosting representation, Neurocomputing 214 (2016) 944–957.

[120] R. S. Perumal, P. C. Mouli, Dimensionality reduced local directional pattern (dr-ldp) for face recognition, Expert Syst Appl 63 (2016) 66–73.

[121] S. Huang, L. Zhuang, Exponential discriminant locality preserving projection for face recognition, Neurocomputing 208 (2016) 373–377.

[122] N. Zhou, A. Constantinides, G. Huang, S. Zhang, Face recognition based on an improved center symmetric local binary pattern, Neural Computing and Applications (2017) 1–7.

[123] Y. Wang, Y. Y. Tang, L. Li, Correntropy matching pursuit with application to robust digit and face recognition, IEEE transactions on cybernetics 47 (6) (2017) 1354–1366.

[124] M. A. Muqeet, R. S. Holambe, Local appearance-based face recognition using adaptive directional wavelet transform, Journal of King Saud University-Computer and Information Sciences.

[125] M. Verma, B. Raman, Center symmetric local binary co-occurrence pattern for texture, face and bio-medical image retrieval, Journal of Visual Communication and Image Representation 32 (2015) 224–236.

[126] G. Ghinea, R. Kannan, S. Kannaiyan, Gradient-orientation-based pca subspace for novel face recognition, IEEE Access 2 (2014) 914–920.

[127] N. Piao, R.-H. Park, Face recognition using dual difference regression classification, IEEE Signal Processing Letters 22 (12) (2015) 2455–2458.

[128] V. H. Gaidhane, Y. V. Hote, V. Singh, An efficient approach for face recognition based on common eigenvalues, Pattern Recognit. 47 (5) (2014) 1869–1879.

[129] Z. Zhang, L. Wang, Q. Zhu, S. K. Chen, Y. Chen, Pose-invariant face recognition using facial landmarks and weber local descriptor, Knowledge-Based Systems 84 (2015) 78–88.

[130] A. Pillai, R. Soundrapandiyan, S. Satapathy, S. C. Satapathy, K.-H. Jung, R. Krishnan, Local diagonal extrema number pattern: A new feature descriptor for face recognition, Future Generation Computer Systems.

[131] R. Atta, M. Ghanbari, An efficient face recognition system based on embedded dct pyramid, IEEE Transactions on Consumer Electronics 58 (4).

[132] S.-M. Huang, J.-F. Yang, Linear discriminant regression classification for face recognition, IEEE Signal Processing Letters 20 (1) (2013) 91–94.

[133] L. Li, J. Gao, H. Ge, A new face recognition method via semi-discrete decomposition for one sample problem, Optik-International Journal for Light and Electron Optics 127 (19) (2016) 7408–7417.

[134] Y. Yan, H. Wang, D. Suter, Multi-subregion based correlation filter bank for robust face recognition, Pattern Recognition 47 (11) (2014) 3487–3501.

[135] Y. Xu, D. Zhang, J. Yang, J.-Y. Yang, A two-phase test sample sparse representation method for use with face recognition, IEEE Transactions on Circuits and Systems for Video Technology 21 (9) (2011) 1255–1262.

[136] L. Dora, S. Agrawal, R. Panda, A. Abraham, An evolutionary single gabor kernel based filter approach to face recognition, Engineering Applications of Artificial Intelligence 62 (2017) 286–301.

[137] M. Belahcene, M. Laid, A. Chouchane, A. Ouamane, S. Bourennane, Local descriptors and tensor local preserving projection in face recognition, in: 2016 6th European Workshop on Visual Information Processing, 2016, pp. 1–6.

[138] J. Pan, X.-S. Wang, Y.-H. Cheng, Single-sample face recognition based on lpp feature transfer, IEEE Access 4 (2016) 2873–2884.

[139] S.-R. Zhou, J.-P. Yin, J.-M. Zhang, Local binary pattern (lbp) and local phase quantization (lbq) based on gabor filter for face representation, Neurocomputing 116 (2013) 260 – 264. doi:http://dx.doi.org/10.1016/j.neucom.2012.05.036.

[140] T. Gao, X. Feng, H. Lu, J. Zhai, A novel face feature descriptor using adaptively weighted extended lbp pyramid, Optik-International Journal for Light and Electron Optics 124 (23) (2013) 6286–6291.

[141] S. Yuan, X. Mao, L. Chen, Multilinear spatial discriminant analysis for dimensionality reduction, IEEE Trans Image Process 26 (6) (2017) 2669–2681.

[142] D. Tian, D. Tao, Coupled learning for facial deblur, IEEE Trans Image Process 25 (2) (2016) 961–972.

[143] J. Jiang, C. Chen, J. Ma, Z. Wang, Z. Wang, R. Hu, Srlsp: A face image super-resolution algorithm using smooth regression with local structure prior, IEEE Transactions on Multimedia 19 (1) (2017) 27–40.

[144] H. Li, C. Y. Suen, Robust face recognition based on dynamic rank representation, Pattern Recognition 60 (2016) 13–24.

[145] F. Zhang, J. Yang, J. Qian, Y. Xu, Nuclear norm-based 2-dpca for extracting features from images, IEEE transactions on neural networks and learning systems 26 (10) (2015) 2247–2260.

[146] W. Zhu, Y. Yan, Y. Peng, Pair of projections based on sparse consistence with applications to efficient face recognition, Signal Processing: Image Communication 55 (2017) 32–40.

[147] F. Cao, X. Feng, J. Zhao, Sparse representation for robust face recognition by dictionary decomposition, Journal of Visual Communication and Image Representation 46 (2017) 260–268.

[148] X. Wu, Q. Li, L. Xu, K. Chen, L. Yao, Multi-feature kernel discriminant dictionary learning for face recognition, Pattern Recognition 66 (2017) 404–411.

[149] J. Zhao, Y. Lv, Z. Zhou, F. Cao, A novel deep learning algorithm for incomplete face recognition: Low-rank-recovery network, Neural Networks 94 (2017) 115–124.

[150] H. Nguyen, W. Yang, B. Sheng, C. Sun, Discriminative low-rank dictionary learning for face recognition, Neurocomputing 173 (2016) 541–551.

[151] B. Chen, J. Li, B. Ma, G. Wei, Discriminative dictionary pair learning based on differentiable support vector function for visual recognition, Neurocomputing 272 (2018) 306–313.

[152] J.-Y. Zhu, W.-S. Zheng, F. Lu, J.-H. Lai, Illumination invariant single face image recognition under heterogeneous lighting condition, Pattern Recognition 66 (2017) 313–327.

[153] L. Liu, P. Fieguth, G. Zhao, M. Pietikäinen, D. Hu, Extended local binary patterns for face recognition, Information Sciences 358 (2016) 56–72.

[154] W. Huang, H. Yin, Robust face recognition with structural binary gradient patterns, Pattern Recognition 68 (2017) 126–140.

[155] Y.-F. Yu, D.-Q. Dai, C.-X. Ren, K.-K. Huang, Discriminative multi-layer illumination-robust feature extraction for face recognition, Pattern Recognition 67 (2017) 201–212.

[156] H. Roy, D. Bhattacharjee, Local-gravity-face (lg-face) for illumination-invariant and heterogeneous face recognition, IEEE Transactions on Information Forensics and Security 11 (7) (2016) 1412–1424.

[157] Z. Yang, Y. Wu, W. Zhao, Y. Zhou, Z. Lu, W. Li, Q. Liao, A novel illumination-robust local descriptor based on sparse linear regression, Digital Signal Processing 48 (2016) 269–275.

[158] N. Piao, R.-H. Park, Face recognition using dual difference regression classification, IEEE Signal Processing Letters 22 (12) (2015) 2455–2458.

[159] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[160] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[161] M. Sato, K. Hotta, A. Imanishi, M. Matsuda, K. Terai, Segmentation of cell membrane and nucleus by improving pix2pix., in: BIOSIGNALS, 2018, pp. 216–220.

[162] S. M. Gang, J. J. Lee, Depth map extraction from the single image using pix2pix model, Journal of Korea Multimedia Society 22 (5) (2019) 547–557.

[163] H. Zou, H. Zhang, X. Li, J. Liu, Z. He, Generation textured contact lenses iris images based on 4dcycle-gan, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 3561–3566.

[164] R. Huang, S. Zhang, T. Li, R. He, Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2439–2448.

[165] T. Hassner, S. Harel, E. Paz, R. Enbar, Effective face frontalization in unconstrained images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4295–4304.

[166] X. Zhu, Z. Lei, J. Yan, D. Yi, S. Z. Li, High-fidelity pose and expression normalization for face recognition in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 787–796.

[167] X. Wu, R. He, Z. Sun, T. Tan, A light cnn for deep face representation with noisy labels, IEEE Transactions on Information Forensics and Security 13 (11) (2018) 2884–2896.

[168] Y. Yin, S. Jiang, J. P. Robinson, Y. Fu, Dual-attention gan for large-pose face frontalization, arXiv preprint arXiv:2002.07227.

[169] P. Li, X. Wu, Y. Hu, R. He, Z. Sun, M2fpa: A multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10043–10051.

[170] H. Tang, D. Xu, N. Sebe, Y. Yan, Attention-guided generative adversarial networks for unsupervised image-to-image translation, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[171] P. Luo, X. Wang, X. Tang, Hierarchical face parsing via deep learning, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2480–2487.

[172] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image and Vision Computing 28 (5) (2010) 807–813.

[173] D. E. King, Dlib-ml: A machine learning toolkit, The Journal of Machine Learning Research 10 (2009) 1755–1758.

[174] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[175] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[176] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[177] R. Xu, Z. Zhou, W. Zhang, Y. Yu, Face transfer with generative adversarial network, arXiv preprint arXiv:1710.06090.

[178] C. E. Thomaz, G. A. Giraldi, A new ranking method for principal components analysis and its application to face image analysis, Image and vision computing 28 (6) (2010) 902–913.

[179] M. G. Calvo, D. Lundqvist, Facial expressions of emotion (kdef): Identification under different display-duration conditions, Behavior research methods 40 (1) (2008) 109–115.

[180] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, A. Van Knippenberg, Presentation and validation of the radboud faces database, Cognition and emotion 24 (8) (2010) 1377–1388.

[181] O. Barkan, J. Weill, L. Wolf, H. Aronowitz, Fast high dimensional vector multiplication face recognition, in: Proceedings of the IEEE international conference on computer vision, 2013, pp. 1960–1967.

[182] S. Ouellet, Real-time emotion recognition for gaming using deep convolutional network features, arXiv preprint arXiv:1408.3750.

[183] A. Mollahosseini, D. Chan, M. H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: 2016 IEEE Winter conference on applications of computer vision (WACV), IEEE, 2016, pp. 1–10.

[184] B.-F. Wu, C.-H. Lin, Adaptive feature mapping for customizing deep learning based facial expression recognition model, IEEE access 6 (2018) 12451–12461.

[185] M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1749–1756.

[186] S. H. Lee, W. J. Baddar, Y. M. Ro, Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos, Pattern Recognition 54 (2016) 52–67.

[187] S. H. Lee, K. N. K. Plataniotis, Y. M. Ro, Intra-class variation reduction using training expression images for sparse representation based facial expression recognition, IEEE Transactions on Affective Computing 5 (3) (2014) 340–351.

[188] R. Ptucha, G. Tsagkatakis, A. Savakis, Manifold based sparse representation for robust expression recognition without neutral subtraction, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, pp. 2136–2143.

[189] M. Liu, S. Li, S. Shan, R. Wang, X. Chen, Deeply learning deformable facial action parts model for dynamic expression analysis, in: Asian conference on computer vision, Springer, 2014, pp. 143–157.

[190] F. Bashar, A. Khan, F. Ahmed, M. H. Kabir, Robust facial expression recognition based on median ternary pattern (mtp), in: 2013 International Conference on Electrical Information and Communication technology (EICT), IEEE, 2014, pp. 1–5.

[191] C. Turan, K.-M. Lam, Histogram-based local descriptors for facial expression recognition (fer): A comprehensive study, Journal of visual communication and image representation 55 (2018) 331–341.

[192] L. Du, H. Hu, Weighted patch-based manifold regularization dictionary pair learning model for facial expression recognition using iterative optimization classification strategy, Computer Vision and Image Understanding 186 (2019) 13–24.

[193] Z. Sun, Z. Hu, M. Zhao, Automatically query active features based on pixel-level for facial expression recognition, IEEE Access 7 (2019) 104630–104641.

[194] M. Kyperountas, A. Tefas, I. Pitas, Salient feature and reliable classifier selection for facial expression classification, Pattern Recognition 43 (3) (2010) 972–986.

[195] C. Shan, S. Gong, P. W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, Image and vision Computing 27 (6) (2009) 803–816.

[196] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1805–1812.

[197] Y. Sun, G. Wen, Cognitive facial expression recognition with constrained dimensionality reduction, Neurocomputing 230 (2017) 397–408.

[198] A. Poursaberi, H. A. Noubari, M. Gavrilova, S. N. Yanushkevich, Gauss–laguerre wavelet textural feature fusion with geometrical information for facial expression identification, EURASIP Journal on Image and Video Processing 2012 (1) (2012) 17.

[199] Z. Fei, E. Yang, D. D.-U. Li, S. Butler, W. Ijomah, X. Li, H. Zhou, Deep convolution network based emotion analysis towards mental health care, Neurocomputing.

[200] A. St, Emotion recognition: The influence of texture's descriptors on classification accuracy, in: Beyond Databases, Architectures and Structures. Towards Efficient Solutions for Data Analysis and Knowledge Representation: 13th International Conference, BDAS 2017, Ustroń, Poland, May 30-June 2, 2017, Proceedings, Vol. 716, Springer, 2017, p. 427.

[201] M. V. Zavarez, R. F. Berriel, T. Oliveira-Santos, Cross-database facial expression recognition based on fine-tuned deep convolutional network, in: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, 2017, pp. 405–412.

[202] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, V. Palade, Stacked deep convolutional auto-encoders for emotion recognition from facial expressions, in: 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, 2017, pp. 1586–1593.

[203] M. Shin, M. Kim, D.-S. Kwon, Baseline cnn structure analysis for facial expression recognition, in: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE, 2016, pp. 724–729.

[204] Y. Yaddaden, A. Bouzouane, M. Adda, B. Bouchard, A new approach of facial expression recognition for ambient assisted living, in: Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, 2016, pp. 1–8.

[205] Z. Sun, Z.-P. Hu, M. Wang, S.-H. Zhao, Discriminative feature learning-based pixel difference representation for facial expression recognition, IET Computer Vision 11 (8) (2017) 675–682.

[206] A. Samara, L. Galway, R. Bond, H. Wang, Affective state detection via facial expression analysis within a human–computer interaction context, Journal of Ambient Intelligence and Humanized Computing 10 (6) (2019) 2175–2184.

[207] Y. Ye, X. Zhang, Y. Lin, H. Wang, Facial expression recognition via region-based convolutional fusion network, Journal of Visual Communication and Image Representation 62 (2019) 1–11.

[208] S. Guo, L. Feng, Z.-B. Feng, Y.-H. Li, Y. Wang, S.-L. Liu, H. Qiao, Multi-view laplacian least squares for human emotion recognition, Neurocomputing 370 (2019) 78–87.

[209] L. Zhao, Z. Wang, G. Zhang, Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and gabor multiorientation fusion histogram, Mathematical Problems in Engineering 2017.

[210] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, 2008.

[211] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE transactions on pattern analysis and machine intelligence 29 (6) (2007) 915–928.

[212] W. Sun, H. Zhao, Z. Jin, A visual attention based roi detection method for facial expression recognition, Neurocomputing 296 (2018) 12–22.

[213] A. Moeini, H. Moeini, Multimodal facial expression recognition based on 3d face reconstruction from 2d images, in: International Workshop on Face and Facial Expression Recognition from Real World Videos, Springer, 2014, pp. 46–57.

[214] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE transactions on pattern analysis and machine intelligence 31 (2) (2008) 210–227.

[215] B. Jiang, K. Jia, Robust facial expression recognition algorithm based on local metric learning, Journal of Electronic Imaging 25 (1) (2016) 013022.

[216] W. Sun, H. Zhao, Z. Jin, An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks, Neurocomputing 267 (2017) 385–395.

[217] W. Sun, H. Zhao, Z. Jin, A complementary facial representation extracting method based on deep learning, Neurocomputing 306 (2018) 246–259.

[218] D. Duncan, G. Shine, C. English, Facial emotion recognition in real time, Stanford University.

[219] P. Michel, R. El Kaliouby, Real time facial expression recognition in video using support vector machines, in: Proceedings of the 5th international conference on Multimodal interfaces, 2003, pp. 258–264.

[220] S. Li, W. Deng, Deep facial expression recognition: A survey, IEEE Transactions on Affective Computing (2020) 1–1doi:10.1109/TAFFC.2020.2981446.

[221] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, S. Yan, Peak-piloted deep network for facial expression recognition, in: European conference on computer vision, Springer, 2016, pp. 425–442.

[222] W. Li, D. Huang, H. Li, Y. Wang, Automatic 4d facial expression recognition using dynamic geometrical image network, in: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), 2018, pp. 24–30. doi:10.1109/FG.2018.00014.

[223] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, G. Medioni, Expnet: Landmark-free, deep, 3d facial expressions, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 122–129.

[224] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2983–2991. doi:10.1109/ICCV.2015.341.

[225] P. Hu, D. Cai, S. Wang, A. Yao, Y. Chen, Learning supervised scoring ensemble for emotion recognition in the wild, in: Proceedings of the 19th ACM international conference on multimodal interaction, 2017, pp. 553–560.

[226] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Transactions on Image Processing 26 (9) (2017) 4193–4203. doi:10.1109/TIP.2017.2689999.

[227] Z. Yu, Q. Liu, G. Liu, Deeper cascaded peak-piloted network for weak expression recognition, The Visual Computer 34 (12) (2018) 1691–1699.

List of Figures

1.1 Examples of face-like objects 2

1.2 Lower and Upper parts missalignement effect on face percetion 3

1.3 Left: fMRI image of the FFA, Right: ERPs responses of the FFA to Face/Object [19] 4

1.4 Facial Expression Recognition difficulty levels 10

2.1 Generic architecture of an image-based facial analysis framework 14

2.2 CNN-based architecture for image facial analysis 16

2.3 Example of extracting global features using wavelet transformations 17

2.4 The LBP pixel transformation process. 18

2.5 The circular (8,1),(16,2) and (8,2) neighborhoods. 18

2.6 The PCANet2 learnable feature extraction method proposed in [86] 22

2.7 CBFD workflow for face feature extraction as proposed in [87] 23

2.8 Deep-feature based transfer learning approach for face and facial expression recognition. 26

2.9 Example of Nearest Neighbor matching with $K = 1$ and $K = 3$ 26

2.10 Feature vectors representation both in the input feature spaces. 28

2.11 Neural Networks-based classification architecture. 29

3.1 Local sampling topology of MNTCDP. 37

3.2 The two groups of sampling adopted in MNTCDP descriptor. 38

3.3 The overall framework of the MNTCDP feature vector calculation. 40

3.4 Comparing the obtained feature histograms of two images of the same person using MNTCDP, LBP, LTP, $nLBPd$ and $dLBP\alpha$ descriptors. 41

3.5 Overview of the proposed face recognition system. 43

3.6 Images of subject from ORL database. 44

3.7 Images of subject from YALE database. 44
3.8 Images of subject from Extended YALE B database. 45
3.9 Images of subject from the used subset of FERET database. 45
3.10 Images of subject from the used subset of AR database. 46
3.11 Performance evolution of top 5 descriptors on ORL dataset. 49
3.12 Performance evolution of the top 5 descriptors on FERET dataset. 50
3.13 Performance evolution of the top 5 descriptors on YALE dataset. 51
3.14 Performance evolution of top 5 descriptors on EYB dataset. 52
3.15 Performance evolution of top 5 descriptors on AR dataset. 53
3.16 Performance evaluation of the tested deep learning methods. 56
3.17 Camera system of Toyota HSR. 67
3.18 Images of subject from the collected database 67

4.1 The overall pipeline of the proposed framework. 70
4.2 The overall architecture of the proposed PIFR framework based on GAN image translation 73
4.3 Paired GAN used for frontalization of profile image 74
4.4 The architectures of U-Net generators as reported in [174] 75
4.5 The generic architecture of ResNet-based generators 76
4.6 The layers used to build the discriminator 76
4.7 Comparison of ResNet and DenseNet residual connections 77
4.8 Construction and partitioning of the Combined-PIFR database 79
4.9 Comparison of the produced frontal images by the evaluated GAN's architectures 81

5.1 OPD-GQMBP neighborhood topologies. 89
5.2 Cartesian system used to identify the pixel coordinates for each line and group. 90
5.3 Texture feature extraction workflow based on the proposed OPD-GQMBP operator. 92
5.4 Overall view of the proposed FER framework. 93
5.5 Samples of two subjects from JAFFE database 94

5.6 Samples of a subject from KDEF database 94
5.7 The seven emotions of one person from CK+ database 95
5.8 Samples of two subjects from OuluCasia database 95
5.9 Samples of one subject from RaFD database 96
5.10 Confusion matrix of the seven emotions of CK+ 101
5.11 Confusion matrix of the seven emotions of JAFFE 101
5.12 Confusion matrix of the seven emotions of KDEF 102
5.13 Confusion matrix of the seven emotions of OuluCasia 102
5.14 Confusion matrix of the eight emotions of RaFD 103
5.15 Comparison of the overall execution time elapsed to predict the label of an image 105
5.16 Internal configuration of the LSTM Cell 107
5.17 The overall pipeline of the proposed CNN-LSTM for dynamic FER model . . 109
5.18 Examples of the angry sequence from the three datases 111

List of Tables

2.1 summary of texture descriptors tested. 21

2.2 Examples of Distance Metrics for Classification 27

3.1 Average recognition accuracy (%) over 10 splits on ORL dataset (Exp#1) . 58

3.2 Average recognition accuracy (%) over 10 splits on FERET dataset (Exp#2) 59

3.3 Average recognition accuracy (%) over 10 splits on YALE dataset (Exp#3) . 60

3.4 Average recognition accuracy (%) over 10 splits on Extended Yale B dataset (Exp#4) 61

3.5 Recognition accuracy (%) on Extended Yale B dataset using subsets evaluation protocol (Exp#5) 62

3.6 Average recognition accuracy (%) over 10 splits on AR dataset (Exp#6) . . 63

3.7 Comparison with PCANet2 and CBFD deep learning methods 64

3.8 Comparison with recent state-of-the-art systems on ORL, FERET, YALE, EYB(Exp#4) and AR databases. 65

3.9 Comparison with recent state-of-the-art systems on EYB database using subset evaluation protocol. 66

4.1 Similarity-based comparison of the evaluated GAN models 81

4.2 Recognition rate of the proposed framework on the Combined-PIFR database test set 83

4.3 Computation time analysis overall the architectures of the proposed PIFR framework 84

5.1 Average FER rate of each OPD-GQMBP configuration (neighborhood size), overall databases 97

5.2 Average and Maximum accuracies recorded on the five datasets by each method 98

5.3 State-of-the-art person independent FER accuracies on CK+ database . . 100

5.4 State-of-the-art person independent FER accuracies on JAFFE database . 101
5.5 State-of-the-art person independent FER accuracies on KDEF database . . 102
5.6 State-of-the-art person independent FER accuracies on OuluCasia database 102
5.7 State-of-the-art person independent FER accuracies on RaFD database . . 103
5.8 Evaluation of the proposed CNN-LSTM according to different CNN models on the three benchmarks 112
5.9 State-of-the-art results on the three benchmarks 113

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — https://github.com/gallandarakhneorg/tex-templates
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/

école doctorale **sciences pour l'ingénieur et microtechniques**

Title: Development of handcrafted and deep based methods for face and facial expression recognition

Keywords: Machine learning, Classification, Deep learning, Deep neural networks, CNN, Facial image Analysis

Absract

The research objectives of this thesis concern the development of new concepts for image segmentation and region classification for image analysis. This involves implementing new descriptors, whether color, texture, or shape, to characterize regions and propose new deep learning architectures for the various applications linked to facial analysis. We restrict our focus on face recognition and person-independent facial expressions classification tasks, which are more challenging, especially in unconstrained environments. Our thesis lead to the proposal of many contributions related to facial analysis based on handcrafted and deep architecture. We contributed to face recognition by an effective local features descriptor referred to as Mixed Neighborhood Topology Cross Decoded Patterns (MNTCDP). Our face descriptor relies on a new neighborhood topology and a sophisticated kernel function that help to effectively encode the person-related features. We evaluated the proposed MNTCDP-based face recognition system according to well-known and challenging benchmarks of the state-of-the-art, covering individuals' diversity, uncontrolled environment, variable background and lighting conditions. The achieved results outperformed several state-of-the-art ones. As a second contribution, we handled the challenge of pose-invariant face recognition (PIFR) by developing a Generative Adversarial Network (GAN) based image translation to generate a frontal image corresponding to a profile one. Hence, this translation makes the recognition much easier since most reference databases include only frontal face samples. We made an End-to-End deep architecture that contains the GAN for translating profile samples and a ResNet-based classifier to identify the person from its synthesized frontal image. The experiments, which we conducted on an adequate dataset with respect to person-independent constraints between the training and testing, highlight significant improvement in the PIFR performance. Our contributions to the facial expression recognition task cover both static and dynamic-based scenarios. The static-based FER framework relies on extracting textural and shape features from specific face landmarks that carry enough information to detect the dominant emotion. We proposed a new descriptor referred to as Orthogonal and Parallel-based Directions Generic Query Map Binary Patterns (OPD-GQMBP) to efficiently extract emotion-related textural features from 49 landmarks (regions of 32 by 32 pixels). These features are combined with shape ones computed by using Histogram of Oriented Gradients (HOG) descriptor on a binary mask representing the interpolation of the 49 landmarks. The classification is done through the SVM classifier. The achieved Person-Independent performance on five benchmarks with respect to Leave One Subject Out protocol demonstrated the effectiveness of the overall proposed framework against deep and handcrafted state-of-the-art ones. On the other hand, dynamic FER contribution incorporates Long Term Short Memory (LSTM) deep network to encode the temporal information efficiently with a guiding attention map to focus on the emotion-related landmarks and guarantee the person-independent constraint. We considered four samples as inputs representing the evolution of the emotion to its peak. Each sample is encoded through a ResNet-based stream, and the four streams are joined by an LSTM block that predicts the dominant emotion. The experiments conducted on three datasets for dynamic FER showed that the proposed deep CNN-LSTM architecture outperforms the state-of-the-art.

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France

SPIM
école doctorale **sciences pour l'ingénieur et microtechniques**

Titre : Développement de méthodes à base de descripteurs locaux et apprentissage profond pour la reconnaissance du visage et des émotions faciales
Mots-clés : Apprentissage machine, Classification, Apprentissage profond, Réseaux de neurones profonds, CNN, analyse d'images faciales

Résumé

Cette thèse porte sur le développement de nouveaux concepts de segmentation d'images et de classification de régions pour l'analyse d'images. Il s'agit de mettre en œuvre de nouveaux descripteurs, qu'ils soient de couleur, de texture ou de forme et proposer de nouvelles architectures d'apprentissage profond pour des applications liées à l'analyse faciale. Nous nous concentrons sur la reconnaissance faciale et la classification des expressions faciales. Notre thèse a débouché sur la proposition de nombreuses contributions liées à l'analyse faciale couvrant des architectures classiques et profondes. Nous avons contribué à la reconnaissance faciale tout d'abord par la proposition d'un descripteur local appelé Mixed Neighborhood Topology Cross Decoded Patterns. Notre descripteur de visage repose sur une nouvelle topologie de voisinage et une fonction de noyau avancée permettant en encodage efficace des caractéristiques liées à la personne. Nous avons évalué le système de reconnaissance faciale proposé à base du MNTCDP sur des bases de données connues de l'état de l'art, disposant de challenges, couvrant la diversité des individus, environnement non contrôlé, des conditions de fond et d'éclairage variables. Les résultats obtenus ont dépassé plusieurs résultats de l'état de l'art. Pour la deuxième contribution, nous avons relevé le défi de la reconnaissance faciale invariante aux poses (PIFR) en développant une méthode de génération d'images basée sur le Generative Adversarial Network, afin de générer une image frontale correspondant à une image en profil. Cette transformation rend la reconnaissance beaucoup plus facile puisque la plupart des bases de données de référence n'incluent que des échantillons de face frontale. Nous avons créé une architecture profonde de bout en bout, composé du GAN pour la génération des échantillons de profil et un classificateur basé sur ResNet pour l'identification de la personne à partir de son image frontale synthétisée. Les expériences, que nous avons menées sur une base de données adéquate, notamment en termes de chevauchement des individus entre la base de l'apprentissage et celle de l'évaluation, mettent en évidence une grande amélioration des performances du PIFR grâce à la génération d'images frontales du GAN. Nos contributions à la tâche de reconnaissance des expressions faciales (FER) couvrent à la fois des scénarios statiques (une image) et dynamiques (plusieurs images). Notre architecture pour FER avec le mode statique repose sur l'extraction de caractéristiques de texture et de forme à partir de points de repère du visage spécifiques qui présentent suffisamment d'informations pour détecter l'émotion dominante. Nous avons proposé un descripteur appelé Orthogonal and Parallel-based Directions Generic Query Map Binary Patterns pour extraire efficacement les caractéristiques texturales liées aux émotions à partir de 49 points de repère. Ces caractéristiques sont combinées avec celles à base de forme calculées à l'aide du descripteur HOG sur un masque binaire représentant l'interpolation des 49 points. La classification est réalisée via le SVM. Less performances obtenues sur cinq bases de données avec le protocole Leave One Subject Out ont démontré l'efficacité de l'architecture proposée par rapport à l'état de l'art. D'autre part, notre contribution relative à la FER avec le mode dynamique intègre un réseau LSTM pour encoder avec précision les informations temporelles avec un masque d'attention permettant de se concentrer sur les repères liés aux émotions et garantir la robustesse de la reconnaissance. Nous avons considéré quatre échantillons comme entrées représentant l'évolution de l'émotion jusqu'à son pic. Chaque échantillon est codé via une branche CNN et les quatre branches sont jointes par un bloc LSTM qui prédit l'émotion dominante. Les expériences menées sur trois bases de données pour FER dynamique ont montré que l'architecture CNN-LSTM profonde proposée dépasse l'état de l'art.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ
Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France