Signal Processing and Machine Learning Techniques for Automatic Image-Based Facial Expression Recognition

Hayfaa Talib Hussein

Newcastle University

Newcastle Upon Tyne, United Kingdom

A thesis submitted for the degree of

Doctor of Philosophy

September 2019

To

my parents (**GOD** have mercy on them),

my supervisor **Prof. Jonathon Chambers**

my beloved husband, **Dr.Rafid**

and

my loving children

Jaafar, Nooralzahraa, & Yosif

NEWCASTLE UNIVERSITY

SCHOOL OF ENGINEERING

I, Hayfaa Talib Hussein Abugulal, hereby declare that this thesis is my own work and effort and that it has not been submitted anywhere else for any award.

Signature:

Student: Hayfaa Talib Hussein Abugulal

Date:

SUPERVISOR'S CERTIFICATE

This is to certify that the thesis entitled "Signal Processing and Machine Learning Techniques for Automatic Image-Based Facial Expression Recognition" has been prepared under my supervision at the School of Engineering / Newcastle University for the degree of PhD in Computer Engineering - Aritificial Intelligence.

Signature:

Supervisor: Prof. Jonathon A. Chambers

Date:

Signature:

Student: Hayfaa Talib Hussein Abugulal

Date:

Acknowledgements

All praise is due to Allah (God), I thank him and seek his help, guidance, and forgiveness. I am employing this opportunity to express my sincere gratitude to numerous people around me who supported me during my PhD study. It would not have been possible to write this thesis without their support and assistance to me, whom I would like to mention here particularly.

First and foremost, I wish to say thanks a lot to my supervisors, Prof. Jonathon Chambers, for his inspirational guidance, constructive criticism, and for his invaluable pieces of advice which paved the way during my study. I received unlimited help and support from him. Also, I learned how to think positively and how to be a successful researcher. I would also like to express my deepest appreciation to my supervisor, Dr. Charalampos Tsimenidis, for his invaluable advice, productive discussions, constructive suggestions, help, and support. I am sincerely grateful to all of my supervisors for being supportive and helpful since the days I started working on my project.

I would like to express my great appreciation to the Ministry of Higher Education and Scientific Research in Iraq (MOHESR) for having provided a fully-sponsored scholarship throughout my PhD study. I also would like to acknowledge the academic and technical support of the Iraqi cultural attache in London; Kufa University and Faculty of Education for their support and encouragement to gain the degree of PhD from Newcastle University.

In addition, I also thank all my friends and staff within the School of Engineering at Newcastle University for their their assistance and support since my first day in the school.

Last but not least, my heartfelt appreciation goes to my husband, Dr. Rafid, and my children, Jaafar, Nooralzahraa & Yousif for their support and great patience at all times. I stole great moments and special days from them for the sake of study. No words can express my gratitude for their unlimited love, support and encouragement in my journey to reach the highest level of education.

Abstract

In this thesis novel signal processing and machine learning techniques are proposed and evaluated for automatic image-based facial expression recognition, which are aimed to progress towards real world operation.

A thorough evaluation of the performance of certain image-based expression recognition techniques is performed using a posed database and for the first time three progressively more challenging spontaneous databases. These methods exploit the principles of sparse representation theory with identity-independent expression recognition using difference images.

The second contribution exploits a low complexity method to extract geometric features from facial expression images. The misalignment problem of the training images is solved and the performance of both geometric and appearance features is assessed on the same three spontaneous databases. A deep network framework that contains auto-encoders is used to form an improved classifier.

The final work focuses upon enhancing the expression recognition performance by the selection and fusion of different types of features comprising geometric features and two sorts of appearance features. This provides a rich feature vector by which the best representation of the spontaneous facial features is obtained. Subsequently, the computational complexity is reduced by maintaining important location information by concentrating on the crucial roles of the facial regions as the basic processing instead of the entire face, where the local binary patterns and local phase quantization features are extracted automatically by means of detecting two important regions of the face. Next, an automatic method for splitting the training effort of the initial network into several networks and multi-classifiers namely a surface network and bottom network are used to solve the problem and to enhance the performance.

All methods are evaluated in a MATLAB framework and confusion matrices and average facial expression recognition accuracy are used as the performance metrics.

Supporting Publications

The contributions of this thesis have been supported by conference and journal papers, which have been produced throughout the journey of my study as follows, only the author of the thesis and supervisor are listed:

[1] H. Hussein, J. Chambers, and M. Naqvi, “Study of sparsity-based facial expression recognition on a spontaneous database,” in 11th IMA International Conference on Mathematics in Signal Processing, 2016.

[2] H. Hussein, M. Naqvi, and J. Chambers. “Study of image-based expression recognition techniques on three recent spontaneous databases.” In Digital Signal Processing (DSP), 2017 22nd International Conference on, pp. 1-5. IEEE, 2017.

[3] H. Hussein, F. Angelini, M. Naqvi, and J. Chambers, “Deep-Learning Based Facial Expression Recognition System Evaluated on Three Spontaneous Databases”. In 9th International Symposium on Signal, Image, Video and Communications (ISIVC). IEEE 2018, Nov 27, pp. 270-275.

[4] H. Hussein, S. M. Naqvi, and J. Chambers, “Spontaneous facial expression recognition through fusion of geometric and appearance features and deep network based classification.” in 10th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2019, v.1, pp.41–46., 2019.

[5] H. Hussein, F. Angelini, and J. Chambers, “Multi-Stage Classification and Feature Fusion for Image-Based Spontaneous Facial Expression Recognition,” Submitted to APSIPA Transactions on Signal and Information Processing, 2019.

Contents

List of Figures

List of Tables

Chapter 1

Facial Expression Recognition: Motivation and Background

1.1 Motivation

People are adept at expressing themselves and understanding the emotions of others by means of non-verbal signals, for example, a glimpse of the eye, facial expressions, different hand gestures and head motions. Each of these modalities conveys crucial emotional information which is used by humans to infer some of these emotional states. Amongst these modalities, the Facial Expression (FE) has received the greatest attention by effective computing researchers and psychologists.

FE plays a vital role as a non-verbal channel of communication which is responsible for interpreting the direct attention and visual emotions of humans, where FE is a visible manifestation of human emotions, which is representative of activity, action, intention, and psychopathology of a person.

As Mehrabian indicated in [1], the facial expression of a speaker is divided into three sections: 55% of the emotional expressions of a human are translated into the facial expressions as the cognition effect, while 38% of information is translated by voice intonation and finally, only 7% of information is conveyed by the spoken word. Accordingly, the face tends to be one of the most visible channels of emotional communication. Therefore, the recognition of facial expression as a scheme is broadly utilised to measure the emotional state of human beings. Hence, in relation to facial expressions, it is possible to provide the training and test materials which are available.

In 1971, Ekman and Friesen proposed the basic facial expressions regardless of ethnicity and culture. These emotions are commonly categorised into six classes, specifically happiness, fear, sadness, disgust, anger and surprise [2], as shown in Fig. 1.1.

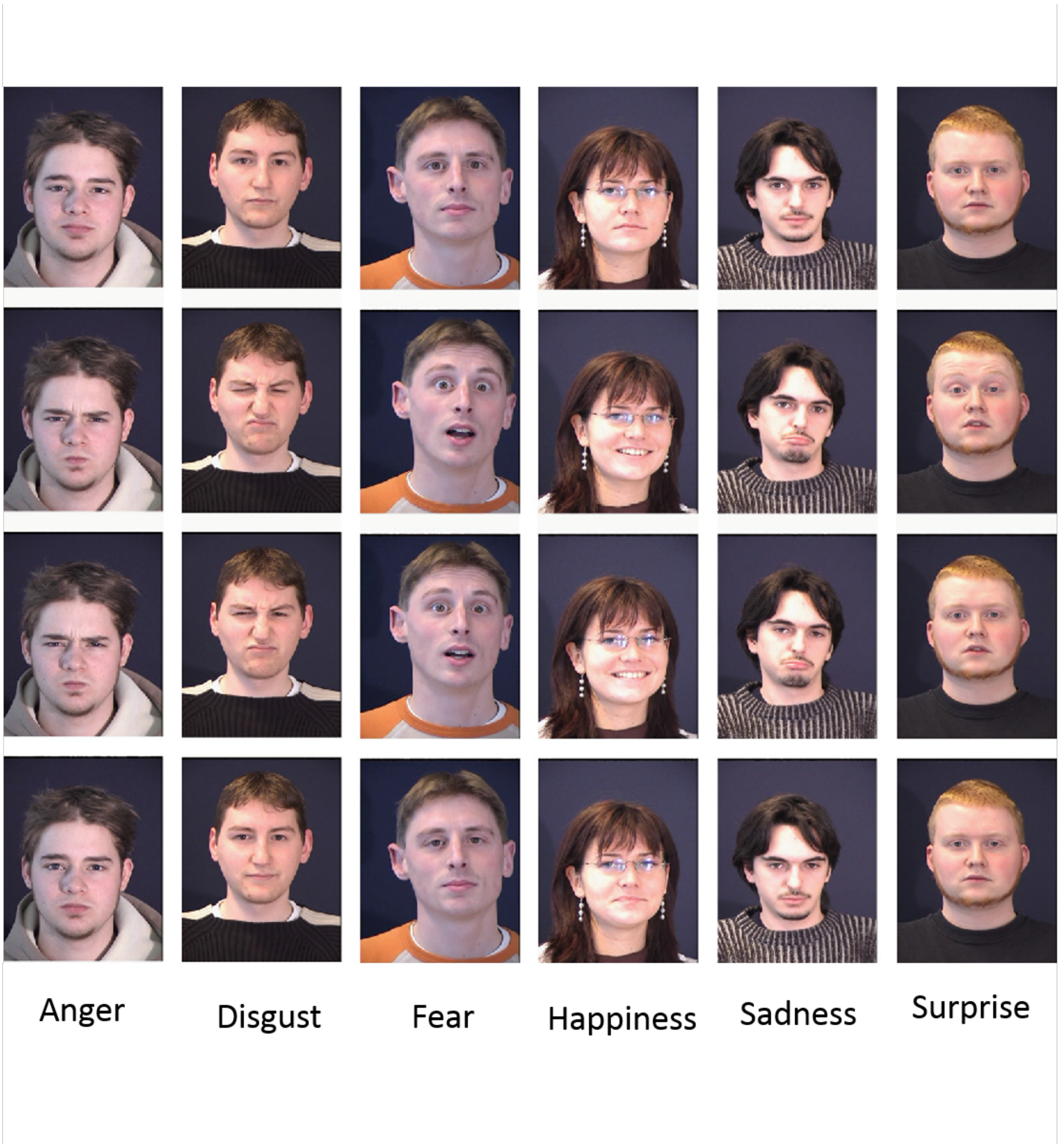


Figure 1.1: Six basic emotions which are commonly used in facial expression recognition studies.

Facial Expression Recognition (FER) is meant to determine the emotional states of humans based on facial information. However, FER faces different challenges. For example, the information presented is different in both muscle movements and relative dynamic emotions owing to the reason that facial muscle movements are very subtle. Second, different environmental conditions significantly affect the performance of FER, such as variations in illumination, size and position.

Considerable effort has been made with respect to facial expression recognition and obtaining a good classification of these basic facial expressions. This effort was distributed into a description of facial appearance and its motion, in addition to the measurement of facial geometry and classification of facial expressions.

In general, there are many questions to consider when recognising facial expressions through which the type of feature extraction is selected: should features be extracted from full coverage of the face or require analysis of specific parts of the face? Or, should appearance or geometric features be used? Are the images static or dynamic? Likewise, do the features correspond to 2D or 3D models, and is information spatial or spatio-temporal? Facial features generally extract either appearance features, which represent the texture of the specific areas of the face, including furrows, bulges and wrinkles or are based on geometric features which represent the location of landmark points of the face, such as the corners of the eye or mouth [3]. It is essential to select a robust technique for feature extraction which can provide extraordinary accuracy in various noise conditions, for instance different lighting conditions are not affected by differences in gender, age and skin colour.

1.2 Background

1.2.1 Principles of Facial Expression Recognition

Most humans are competent mind readers that can attribute states of complex emotion to other humans [4]. Although the emotional state cannot be observed directly by someone else, it can be deduced from the expressive behaviour of humans. The expressive behaviour of humans is concerned with their affective states, whether consciously or unconsciously. Expressive behaviour is frequently unintended and in some cases cannot even be controlled. Moreover, it cannot be encoded or decoded at a deliberate conscious level [4].

Generally, the human face is exceedingly complex in characteristics with a dynamic structure, which can rapidly and effectively change with time [5]. FE is considered as one of the most accurate non-verbal communication approaches. Hence, the subject of FE has been investigated for centuries; Aristotle and Stewart, however, first established FE as an area of philosophy and its underlying theoretical basis. Charles Darwin [6], published an empirical piece of work on facial analysis titled "The Expression of the Emotions in the Man and Animal". Therefore, the presentation of the similarities and differences in exhibiting emotions, concluded that state of mind is expressed in the same movements undertaken by different humans and animal species. This study received global attention from psychologists and cognitive scientists.

The FE of humans are considered as multi-spatial feature signals which contain various types of detail. In general, "the face provides three types of signals: static (such as skin colour), slowly varying (for instance permanent wrinkles) and rapidly varying (such as raising the eyebrows)"[7]. The various expressions describing emotion can possibly be understood and read by a computer. Most early research in the field of emotion analysis was based on a posed database and analysed the face via a motion extraction method or a feature extraction method [8]. These motions are commonly represented by variations in texture and shape.

Recent studies have found that human emotions can provide essential clues in relation to facial expression recognition. FER has received significant attention from researchers in recent years, mostly owing to their diverse applications, such as human-computer interactions, lie detection, surveillance, multimedia, and treatment of mentally retarded patients. High accuracy is required in recognition of facial expressions. Feature extraction plays an important role in the detection of facial expressions. Therefore, this is a significant step in the successful automatic analysis and recognition of facial expressions. So, it is imperative to study the principles of Automatic Facial Expression Recognition (AFER) in the thesis.

It should be mentioned that five recent trends have appeared in AFER:

1. Use of a variety of facial features in an attempt to increase the number of expressions that can be distinguished.

2. Facial action units and combinations are recognized rather than more global characteristics, which provide a simpler mechanism to identify emotion given a specific expression.

3. More robust systems as regards face acquisition, facial data extraction and clarification, in addition to facial expression recognition that are able to manage lighting conditions, occlusion, head poses (head motion for both in-plane and out-of-plane), low intensity expressions and capture distance, each of which are familiar in spontaneous facial

behaviour in settings that are realistic.

4. Fully automatic FERs in real-time.

5. A fusion of facial actions with other different modes, for instance, gesture, and speech.

All these developments mentioned above lead FER toward real-life applications. Many of the databases have been adopted by researchers to address FER problems and conduct comparative tests of research approaches. Databases with ground-truth labels that contain both emotion-specified expressions and action units are required for the next generation of systems which aim to study behaviour that occurs naturally, for example spontaneous and multimodal databases which are derived from the settings of real-life. Recently, work including recognition of spontaneous facial expression has been emerging that may have a significant impact on a set of theoretical and applied topics [9].

Various techniques have been employed to extract facial features. These methods can be categorised, based on the types of features extracted, into two principal types: appearance and geometric-based features methods.

1.3 Aims and Objectives of the Work

The principal aim of this thesis is to address key issues as regards how to obtain a robust image-based FER system with high performance in realistic situations by introducing novel techniques which can surmount various challenging of limitations, such as illumination variations, occlusion, and noise, by extraction of different features and different classifiers to achieve higher accuracy concerning FER, where these techniques are regarded as the basis for developing successful FER approaches. Note, for complexity reasons only image-based methods are considered in this thesis, video-based schemes are outside of the scope of the study.

The objectives of the thesis are as follows:

1. To investigate a robust FER system to mitigate spontaneous database conditions that are more challenging and essential for real-time. That is through the performance of image-based expression recognition with a posed database. The principles of sparse representation theory will be exploited together with identity-independent expression recognition.

2. To reduce the effects of the misalignment problem associated with the training images while extracting spontaneous FE features. A novel method is proposed that employs the extraction of geometric features automatically from raw data with one of the most attractive methods for classification in the field of neural networks, namely deep networks.

3. To enhance the performance of the image-based AFER system by a novel approach to finding an effective solution based on the fusion of three different types of features and using the strategy of the multi-classifier system by splitting the problem into two or more stages. Each stage is a simple classification task, whereby higher overall accuracy of facial expression recognition is obtained.

1.4 Thesis Outline

Chapter 2: introduces the general overview and background knowledge of the typical FER system by means of previous related work. The main operations in an automatic facial expression recognition system, for example face pre-processing, which consists of face detection, tracking and normalisation of the input image are introduced. Subsequently, the methods of feature extraction (appearance and geometric features) and selection are presented. Next, expression classification techniques are presented. Finally, the related FER works are thoroughly stated.

Chapter 3: gives an overview of facial expression databases, concentrating on spontaneous databases employed in evaluating the performance of the proposed methods in this thesis. Then, different metric statistics of FER are presented.

Chapter 4: presents a comparison of the performance of image-based expression in three types of recent spontaneous databases exploiting the principles of sparse representation theory with identity-independent expression recognition using difference images. Three spontaneous databases were used, namely: Man-Machine Interaction Facial Expression Database (MMI), Video Database of Moving Faces and People (VDMFP) and the Belfast Induced Natural Emotion Database. It can be observed that the accuracy of facial expression recognition largely depends on the nature of the database in terms of background and illumination.

Chapter 5: presents improved image-based facial expression recognition employing automatic geometric feature extraction from raw image data including the pre-processing steps and a deep network. Regarding realistic spontaneous databases, there is an important improvement in terms of average accuracy by exploiting the combination of extracted features with the procedures of normalization by using translation, rotation and scaling in comparison to the state-of-the-art methods. The pre-processing procedure has been applied to reduce the difference point of views between these images.

Chapter 6: introduces a novel approach to determine an effective solution for image-based spontaneous facial expression recognition based on a fusion of three types of features, including geometric and appearance features; specifically, Local Phase Quantization method (LPQ), and Local Binary Patterns method (LBP) and geometric features. Furthermore, an auto-encoder deep network is employed as the basic classifier. Moreover, a novel strategy including multiple features and a classifier method are suggested to enhance the performance, such that automatic analysis is utilised to split the training effort into two levels; surface level and bottom level, each stage, and each level has a network and a classifier. Higher accuracy for recognition of facial expressions is thereby achieved.
Furthermore, the fusion approach by multi-feature descriptors has a higher recognition rate than any single feature descriptor.

Chapter 7: provides the main conclusions of this thesis. Topics for future work will also be presented.

Chapter 2

General Overview of Facial Expression Recognition and Background for the Thesis

This chapter focuses on the background knowledge of an FER system in accordance with previous related work, it is not aimed to be exhaustive, but concentrates upon the key aspects of developing an FER system and explains related techniques [10].

In this chapter, the following points will be considered :

- General overview of human emotion,
- Typical aspects of facial expression recognition,
- Facial feature techniques in the context of a facial expression recognition system, and
- Background of the thesis.

The rest of this chapter will involve: Section 2.1, which introduces the overview of human emotion, Section 2.2, which shows facial expression recognition. Section 2.3, which presents the automatic recognition of facial expressions. Section 2.4, displays types of feature extraction. Section 2.5, which introduces the expression classification/ recognition. Section 2.2, which presents Related FE works.

2.1 Human Emotion

A vast range of literature exists with respect to emotions, although the multi-faceted nature prohibits a comprehensive review of all literature. Therefore, only what is essential in promoting this work will be reviewed. Recent discoveries [11], indicate that emotions are closely related to other tasks, for instance perception, attention, decision-making, learning, and memory. As such it may be useful for computers to recognise the human user's emotions and other related expressions to benefit research and applications areas as varied as education, medicine, behavioural science, security, marketing, and entertainment [11]. In this chapter, the focus will be on the expressive nature of emotions, particularly those expressed on the face.

2.1.1 Affective Human-computer Interaction

In many significant Human-computer Interaction (HCI) applications such as marketing, it is compulsory that computer responsiveness considers the emotional or cognitive state of the customer. Emotional communication explicitly investigates how to recognize and express emotions during the interaction between human and intelligent devices such as an iPad or iPhone. In [12], the effective communication problem is addressed, where three specific points are considered regarding numerous-applications: the first point is to capture effective information by advanced systems for facial expression recognition. The second point is the dynamics of mapping expression and the emotional state of the label. The last point is adaptive interaction that means conveying emotive response and responding to a recognized emotional state.

2.1.2 Theories of Emotion

There is little agreement concerning the definition of emotion. Many emotional theories have been proposed; while several of these theories could not be verified till recently. Regardless of the many theories, it is apparent that there is considerable variability in the way in which people presented expressions. One task which has been considerably studied is the ruling of emotions and how emotional expressions are conveyed by human observers to others, by way of the face and in the voice. Related questions are: Do these represent their real emotions? Can they be described convincingly? Can people hide their emotions? [11].

It is important to mention that the three most popular methods applied in emotional psychology research are: discrete categories, dimensional representation and appraisal-based, which are explained in the following pages. These methods are a good starting point for understanding the effects on the targets of automatic recognition, and provide information on the ways in which they have been expressed and interpreted.

A. Categorical

One of the methods is to label the emotions as a discrete category. Therefore, words must be chosen from a specific list of word labels, which are used in language in daily life. In the early 1970s, Ekman and others [13], conducted intensive studies in relation to human facial expressions. Consequently, they established a guide to support universality in facial expressions. Reference to “universal facial expressions” which include happiness, anger, disgust, sadness, fear and surprise, were developed in the same way for all humankind. Ekman et al., in [14], conducted several cross-cultural research activities in 1982. Ekman suggests that regardless of culture, the facial expressions related to basic emotions are perceived in the same way. Fig. 2.1 displays the facial expressions that represent these emotions. Facial expressions have been studied in diverse cultures to find the most common factors in relation to expression and emotion recognition on the face. However, differences in expressions have been observed as well; thus, in social contexts, a person’s facial expressions are controlled by “display rules”.

The problem with this method is that the stimuli may well consist of mixed emotions that were not meant to be an extensive record of possible affective states that an individual might demonstrate (Ekman et al., 1982) [15]. Until recently, most researchers focused on detecting specific emotions [16]. However, there has been an increasing amount of evidence revealing that these emotions are inappropriate for the purposes of affective computing, given that they do not often appear in HCI scenarios [17].

Baron-Cohen et al. (2004), developed a taxonomy that included 24 groups of 412 various emotions. This taxonomy was created via a linguistic analysis of the emotions. It includes emotions such as confusion, boredom, frustration, and interest as well as the basic emotions. The emotions that belong to these categories, for interest, thinking and confusion are more common in daily human-human and human-computer interactions [17].

Several research studies into automatic recognition have used Baron-Cohen’s taxonomy, such as El Kaliouby and Robinson, 2005 [18], Sobol-Shikler and Robinson, 2010 [19]. Additionally, Mahmoud et al. (2011) in [20], used a description of affect; however, it is not as common as

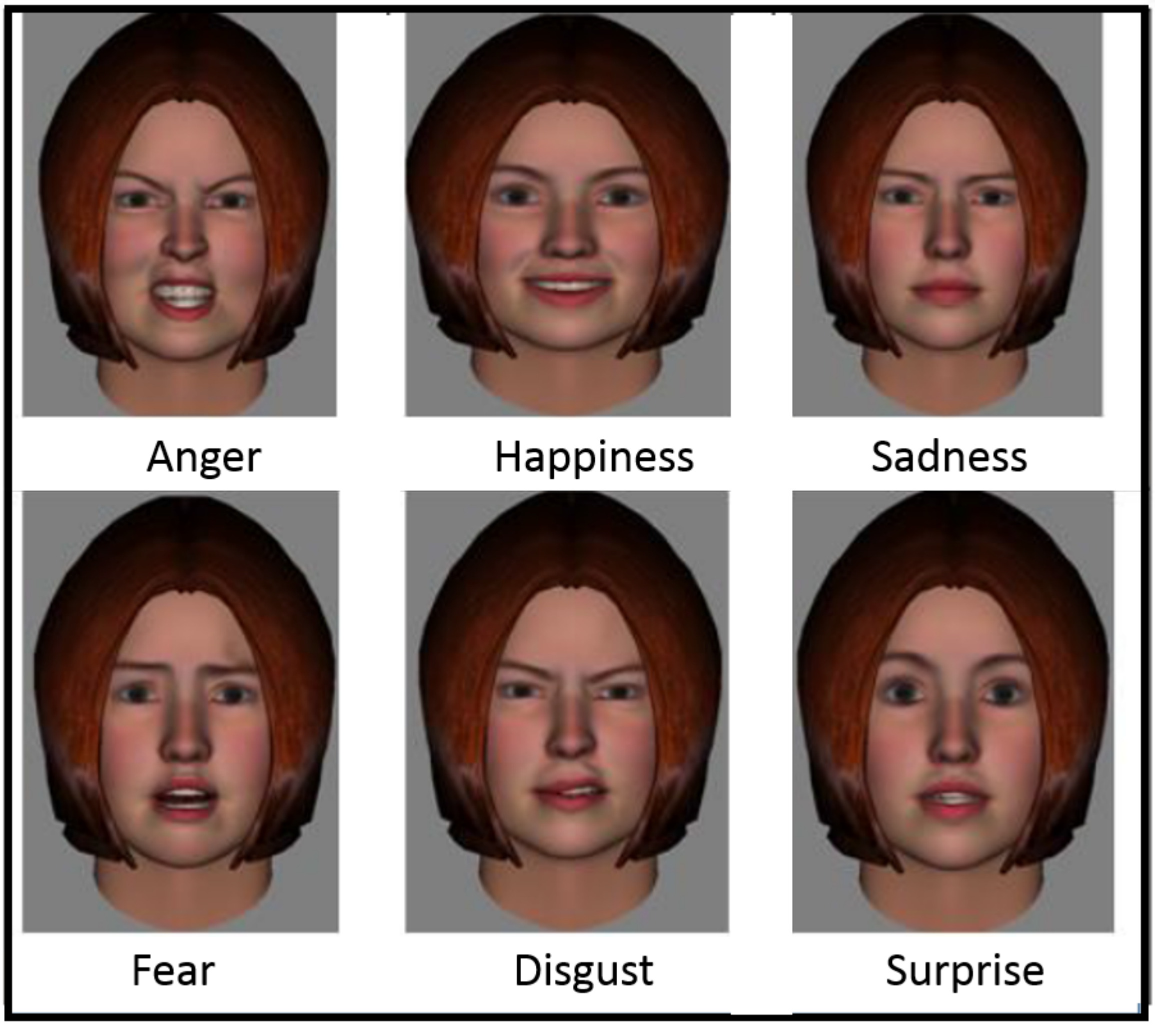


Figure 2.1: Facial expressions of the six basic emotions - anger, happiness, sadness, fear, disgust and surprise taken from Ekman and Friesen (1976).

categories related to basic emotion. Complex emotions might be a more appropriate representation; nonetheless, they lack the level of inherent psychological research in comparison with the six basic emotions.

B. Dimensions

Emotions are described by multiple dimensions or scales. Observers can denote their impression of each stimulus on several consecutive scales instead of selecting discrete labels, such as simple, complicated, pleasant, unpleasant, attention and rejection. Russell and Mehrabian used a dimensional representation in 1977 to describe the affect, whereby an affective state was characterised as a point in a multi-dimensional space, while the axes represent a small number of emotions in dimensions [21]. In contrast [22] used these dimensions to calculate similarities and variations in emotional experience. Examples of affective dimensions are: valence and arousal (pleasant vs. unpleasant) and (lower level vs. high level), activation (relaxed vs. aroused), power (sense of control, dominance vs. submission) and furthermore, expectancy (anticipation and appraisals of novelty and unpredictability). Valence and arousal are two common scales; on the one hand, valence describes a positive or pleasant state (the pleasantness of the stimuli), conversely, valence describes a negative or unpleasant condition. For instance, happiness is a positive valence, whereas, disgust is a negative valence. Arousal describes the level of emotion. Sadness is the lowest level of arousal, while surprise is the highest level of arousal. The various emotional labels can be plotted in different positions on a two-dimensional level by stretching these two axes to build a two-dimensional emotion model [23]. While, in [22], Fontaine et al. (2007) discussed these four dimensions that represent most of the distinctions between daily emotional experiences; thus, creating a good set for analysis. Fig. 2.2 displays facial expressions that can be attributed to specific points in the emotion dimensional space.

Dimensional representation has more elasticity where analysing emotions is compared to categorical representations. However, problems emerged when attempting to use several dimensions, given that some emotions cannot be distinguished when projecting high-dimensional emotional situations upon lower dimensional representations. For instance, fear cannot be distinguished from anger if only valence and activation have been used. Therefore, the representation is not intuitive but requires training for the label expressive behaviour.

Many researchers in affective computing started to explore the dimensional representation of emotion, which regularly addresses a binary classification problem as in [24]. For example, active vs. passive, positive vs. negative or it may even be considered as one of the four classes (classification to quadrants of a two-dimensional space), where, when handled as a classification problem it loses the added elasticity of this representation, several recent studies treated it as a regression [25].

C. Appraisal based

The latest approach to representing emotion is appraisal theory which is exceedingly influential among psychologists [26]. In this approach, an emotion is described by assessing the situation that triggered the emotion, which accounts for individual differences. Appraisal theories of

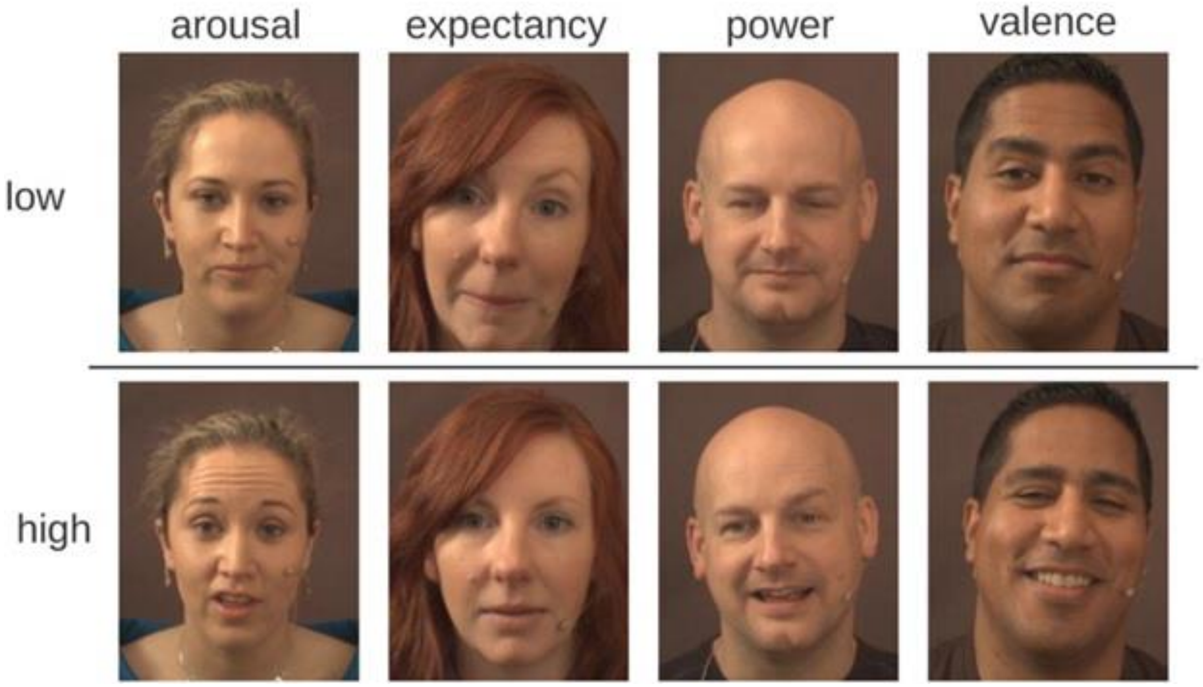


Figure 2.2: Facial expressions which can be associated with certain values in the dimensional emotion space.

emotion are theories that evaluate emotions caused by people's interpretations and explanations of their circumstances, with or without the physiological effects [27].

2.2 Facial Expression Recognition

Facial expression is one of the most powerful ways that human beings can communicate their emotions and intentions regarding natural and immediate contact. It should be noted that it was Suwa et al. (1978) who initiated research on automatic facial expression recognition systems [28]. A growing number of studies have provided plain evidence of the importance of emotions in human-human interaction. Hence, they have been considered as the basis for researchers in engineering and computer science communities, to develop automatic methods for computers to recognise an emotional expression, as a target toward realising human-computer intelligent interaction. The labelling of emotions in various states led to pattern recognition methods being applied to recognise emotions in most research [11].

Countless researchers have been inspired by Ekman's research, such as those working on facial expressions via image and video processing by means of tracking facial features and measuring the amount of facial motion. Subsequently, researchers attempted to classify various facial expressions. Research into computer quantification of facial expressions started in the 1990s. Mase et al. in [29], was considered the first researcher to use image processing techniques to recognise facial expressions. Mase suggested the possibility of using optical flow to distinguish facial expressions. Moreover, image coding by flexible shape and appearance models was used by Lanitis et al., in [30], to recognise a person's identity, gender and pose and perform facial expression recognition. In [31], non-rigid motion was recovered using local parameterized models of image motion, where these parameters were fed into a rule-based classifier to recognise the six basic facial expressions. Furthermore, [32], calculated optical flow and applied comparable rules for the category of six facial expressions. In [33], a radial basis function network was applied to classify expressions after calculating the optical flow of regions on the face. Similarly, an optical flow region-based method was applied in [34] on recognised expressions.

2.3 Automatic Recognition of Facial Expressions

Generally, the approach to the automatic analysis of image-based facial expression consists of steps in the basic processing pipeline. The input of the system is represented by the frame of an image from a video sequence of the facial expression. Each frame of facial images in a video sequence is entered into the system independently and is processed by a series of stages as follows:

- Face Pre-Processing: face detection, tracking and face registration (normalization).
- Feature Extraction: appearance and geometric features.
- Feature selection.
- Expression recognition (classification).
- System evaluation

Fig. 2.3 presents the framework of a typical FER system which consists of five main stages, which are pre-processing that includes three steps (face detection, face tracking, and normalisation), feature extraction, feature selection, classification, and evaluation of the system.

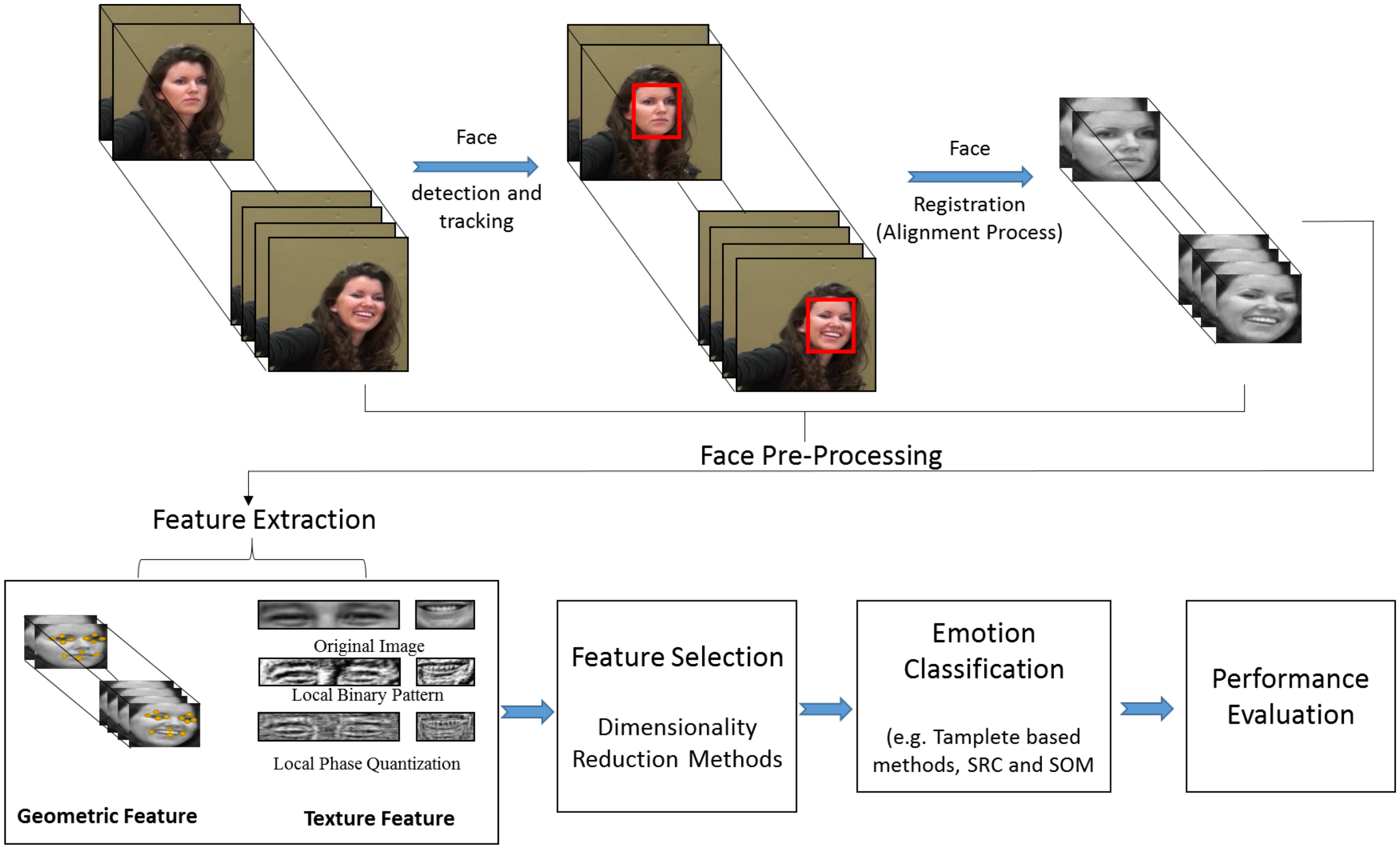


Figure 2.3: Main operations in automatic facial expression recognition system.

2.3.1 Face Pre-Processing

Pre-processing for an input image attempts to locate/track the region of the face and normalise to eliminate the influence produced by differences in face pose, position, illumination and size. Face pre-processing involves three main stages: face detection (segmentation), face tracking (video sequence only) and face registration, which comprises reducing the effects of alignment and illuminations.

2.3.1.1 Face Detection

Facial detection, the first stage to be considered in facial expression analysis that reveals the face in a particular image, or video sequence involves, locating the face and tracking it across the various frames of a video sequence and has been termed "face tracking". Considerable research is available in the field of face detection and tracking. Nevertheless, it is beyond the scope of this literature review to introduce and study all of the suggested methods. Face detection methods are classified into four categories [35] as follows: knowledge-based methods, template matching methods, feature invariant approaches and appearance-based methods.

1. ***Knowledge-based methods*** Knowledge-based methods utilise pre-defined rules to locate a face based on human knowledge. The relationships are captured from facial features by these rules. Usually, the face in an image contains the eyes, nose and mouth. Therefore, the relationship between features, capture distance and relative position are described primarily in relation to face localisation.

   In [36], the possibility of using a hierarchical knowledge-based method to detect faces is described. They suggest a three-level system as follows: the higher two levels were based on mosaic images at different resolutions. While the lower level was an improved edge detection method. In [37], the algorithm that detects the face and inclination is suggested, Adaboost learner is used concerning face detection, whereas an eye detector is used for the calculation of inclination. Thus, calculating the angle of the horizon is undertaken by identifying lines passing through the eyes. It is important to mention that this approach has a few problems. For example, it is difficult to translate human knowledge into rules. That is because some of these rules may be strict, so they may fail to detect faces that do not pass all of these specific rules. Moreover, this approach may give false positives, if the rules are very general.

Furthermore, this approach is difficult to extend to detect faces in various situations, due to the difficulty of an enumeration with regards to all possible states. In contrast, the process of heuristics on the faces works well, especially in the detection of frontal faces in scenery.

2. ***Template matching methods*** Pre-stored face templates have been applied in template matching methods that judge whether an image represents a face. According to the input image, the correlation values and the standard patterns of the face contour, mouth, nose and eyes are computed independently. Therefore, a face is determined based on the correlation values. This method contains two types of templates:

   Predefined templates: these were proposed by Sinha in 1996 to be used in a project on cognitive robotics at MIT [38].

   Deformable templates: deformable templates: the technique consists of two stages: firstly, a model is created to generate a set of reasonable representations in terms of the shape and texture of the face. The second stage is termed the segmentation phase, through which optimal parameters of the model variation are found. The segmentation phase finds the optimal parameters of variation regarding the model, with the aim of matching the shape and the texture of the face with unknown stimuli [39].

3. ***Feature invariant approaches*** The goal of the feature invariant approach is to ascertain the face structure features. Typically, these features are robust to pose and different lighting. These methods lead to low-level analysis (or early vision) of the stimuli, in order to extract discriminative features. A statistical model is constructed based on the extracted features to describe their relationships as well as verify the existence of the face. Frequently, the different extracted features are specified in the context at hand, which is constructed on edge, colour or texture.

   Face features: this method was proposed by Sirohey in 1998 [40], to detect a face from a cluttered background. This method applies an edge map that is known as a Canny detector and heuristics to remove the edges of the group that the facial contour is maintaining.

   Skin color: many studies claim that human skin colour is an effective feature in numerous applications, for example, hand tracking and face detection. In [41], an iterative approach related to skin identification by applying histogram intersection into colour space HSV (Hue, Saturation, and Value) was proposed.

Texture: can be defined as surface appearance characteristics. In an image texture, a set of measures is calculated in image processing. In [42], a method of deducing the presence of the face via the identification of face-like-textures was developed.

4. ***Appearance-based methods*** In this method, large numbers of examples are employed, such as images of faces and/or facial features, which depict different variations, for example the shape of the face, eye colour, skin colour, in addition to closed/open mouth. Face detection is viewed as a problem for pattern recognition that has two classes: specifically, face and non-face [37].

   Generally, appearance-based methods are based on statistical analysis techniques and machine learning, which find the characteristics that are relevant to face and non-face images [35]. Additionally, appearance-based methods pertaining to face detection are based on different methods, for instance eigenfaces, neural networks, support vector machines and hidden Markov models.

   Adaboost proposed by Freund and Schapire in 1995 [43], applies this algorithm to detect pedestrians by applying a Haar wavelet that is an extraction of discriminating features. This method was inspired by the algorithm proposed by Viola-Jones [44], that is used for face detection.

   It should be noted that research by Viola-Jones, which is considered the algorithm for the fastest and most accurate pattern recognition approach related to face detection is used in this study. Wherein this algorithm is used for object detection, and furthermore, detecting face and salient facial regions [44].

   The following four key points are intrinsically included in performing the Viola-Jones algorithm:

   a. Haar-like features were introduced with the Viola-Jones face detector [44] because of their intuitive and computational simplicity in extracting the features.

   b. In order to generate an integral image, Viola-Jones is recommended to calculate the Haar-like features of each pixel location (x, y) rapidly, by summing all pixels located in the top or the left. Therefore, it can obtain a rectangular in four array references based on the integral image.

   c. Viola-Jones proposed to employ the AdaBoost method [45] to identify the Haar-like features. Based on setting threshold values, training has been set by the presence of

Haar-like features. Due to a very large number of sub-windows for the total number of the Haar-like features, a set of weak classifiers are used to reduce this number, by selecting the single rectangle which has high discriminants of negative and positive cases. Numerous weak classifiers were combined to sign the weights of the positive instances, and addition rejects the negative sub-windows.

d. Viola-Jones presented an approach combining more classifiers (cascading classifiers) in a cascade form in order to increase the detection performance. It detects most positive cases by boosting classifiers and rejecting the negative sub-windows before implementing further classification processes.

2.3.1.2 Face Tracking

Face tracking is similar to face detection, which is intended to track the region of the face or fiducial facial landmarks in video sequence frames and has been successful in challenging tasks such as handling differences in a facial pose, partial occlusions, image size and illumination. Face tracking is extensively used in FER, both image-based and video based schemes, so is considered for completeness, and includes different methods such as Piecewise Bezier volume Deformation (PBVD), Candide model [46], Kanade Lucas Tomasi (KLT) [47], an Active Appearance Model (AAM) [48], Active Shape Model (ASM) [49], and Discriminative Response Map Fitting (DRMF) [53]. The PBVD and candide models were designed to track the three dimensional (3D) face and extract the facial action units. The KLT model and Particle filters are more appropriate for individual fiducial points tracking. AAM and ASM are the most exploited models in face tracking. They have numerous advantages, such as fast-tracking speed, publicly accessible code and accurate point tracking. Nevertheless, there is one major weakness in both approaches; specifically that they require 68 landmarks in every training image to be labelled manually, which is a difficult task for all of the training images for the three types of large databases. This causes loss of time and effort. Additionally, DRMF [53] is another method to detect points on the face, which is characterised by highly efficient computation. Therefore, using this method is more suitable and less complicated for the task of face tracking in this thesis. It allows for detection of the facial points on the face in real time and maintains an accurate detection of these points even if the face is partly occluded. Generally, the Discriminative Response Map Fitting (DRMF) method [53] in the training procedure involves two main steps.

The first step is aims to train a dictionary to approximate of the response map that can be employed to extract the relevant feature to learn the fitting update model. The second step consists of iteratively learning a model of the fitting update that is obtained via a modified boosting procedure.

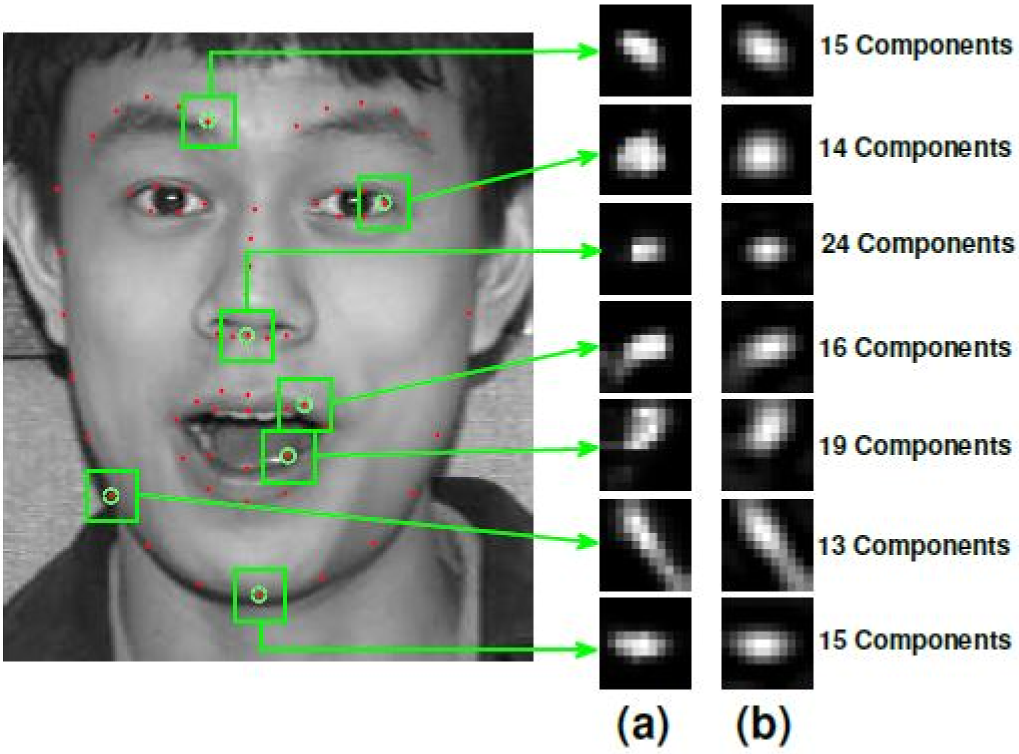


Figure 2.4: Example of a response map [53].

The procedure of eliminating and replacing the sample for each iteration has dual benefits. In the first direction, it has an important role in ensuring the progress of the functions of the fitting parameter update are trained as regards the more problematic samples that did not converge in the prior iterations. While in the second direction, it plays an important role in assisting regularising the learning procedure into correcting the samples which diverged in the prior iterations due to overfitting. Then, repeating the training procedure iteratively until all the training samples converge or reach the maximum number of desired training iterations. Fig 2.5 shows example fitting results of DRMF [53].

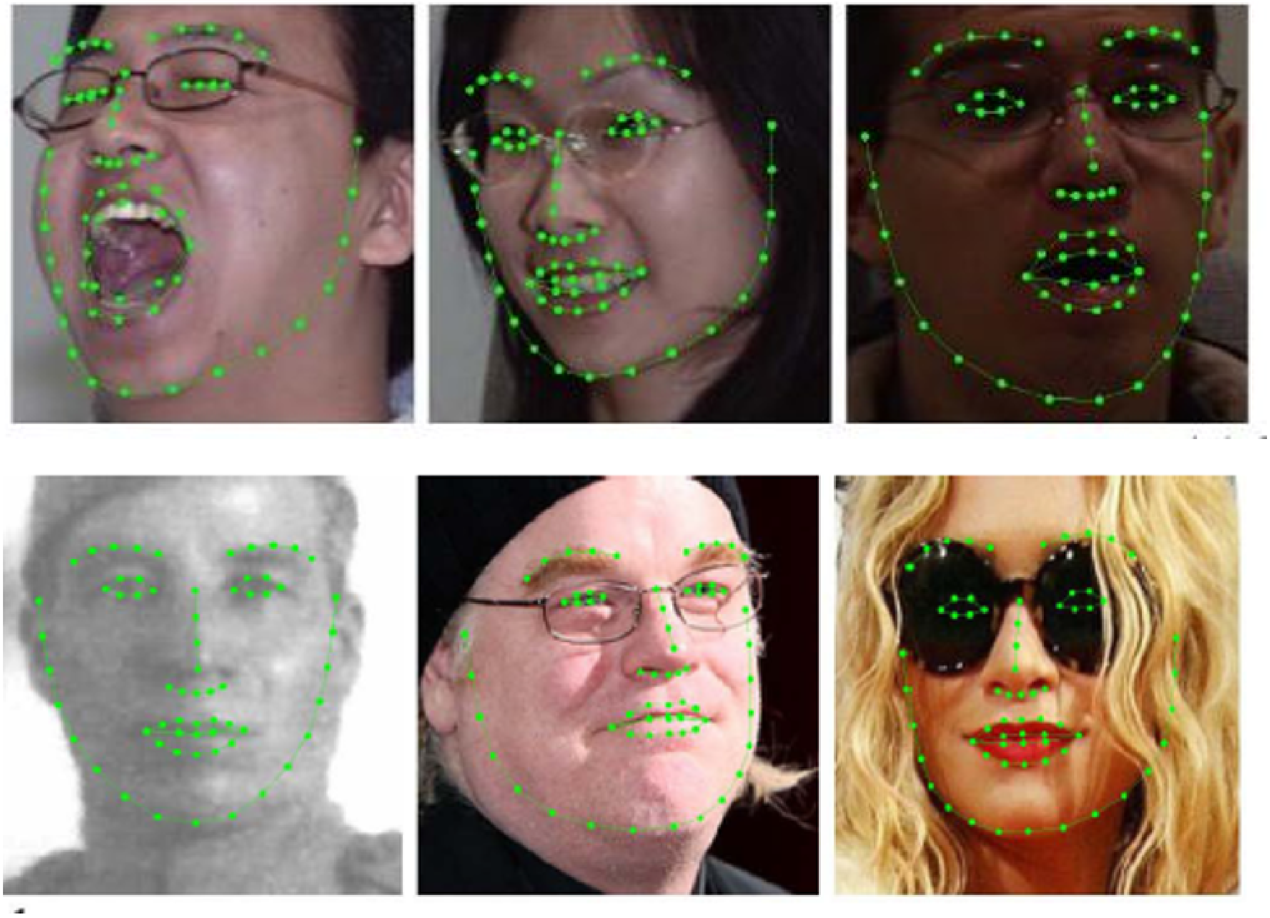


Figure 2.5: Example of DRMF Fitting Results [53].

2.3.1.3 Face Registration

Face detection or tracking algorithms return a facial region from image/video, which may include misalignment problems such as variations in illumination, size and position. In order to extract feature (appearance & geometric features), this misalignment problem might be solved by activating face/video registration.

To address misalignment problems and in order to obtain a successful FER system, several approaches are adopted. Illumination effects are addressed by using the histogram equalisation technique [50] and Gamma adjustment [51]. The histogram equalisation technique [50] handles the effect of the illumination variation on the image by mapping the certain distribution of the intensity value on another uniform distribution or adjusts the defined histogram. Gamma adjustment [51] is adopted to adjust the global brightness changes in the image, that is by changing between black and white. Also, it should be noted that the rest of the misalignment such as pose, size and position have been addressed in Chapter 5.

2.4 Feature Extraction

Feature extraction is an essential stage for a robust FER system, concerning discriminating features extraction which reveal subtle appearance changes and facial activities movements. This is an important step for success, concerning the analysis and recognition of facial expressions automatically. The optimal features should minimise differences within-class of expressions, while maximising differences between classes. The qualified classifier may fail to achieve an accurate recognition if the features are insufficient [37].

A considerable amount of the literature has been published on emotion image retrieval and facial expression recognition techniques. Recently, images of emotion related to automatic facial expression recognition have been studied intensively, and consequently, a variety of approaches based on static and dynamic facial images have been presented. The most challenging aspects of the facial expression recognition system are to obtain relevant facial features and apply the classifier which gives the best classification for those features [52, 64].

Generally, various techniques are used to extract the facial features. These techniques can be categorised into two: geometric and appearance-based feature methods. Geometric-based features introduce the shape and locations of facial components, including the eyes, brows, mouth and nose. Moreover, they require accuracy and reliability in the detection of facial features and tracking methods, although the principal drawback of this technique is its sensitivity to noise. Appearance-based features are most suitable for extracting subtle changes in the texture of the face, such as furrows and wrinkles.

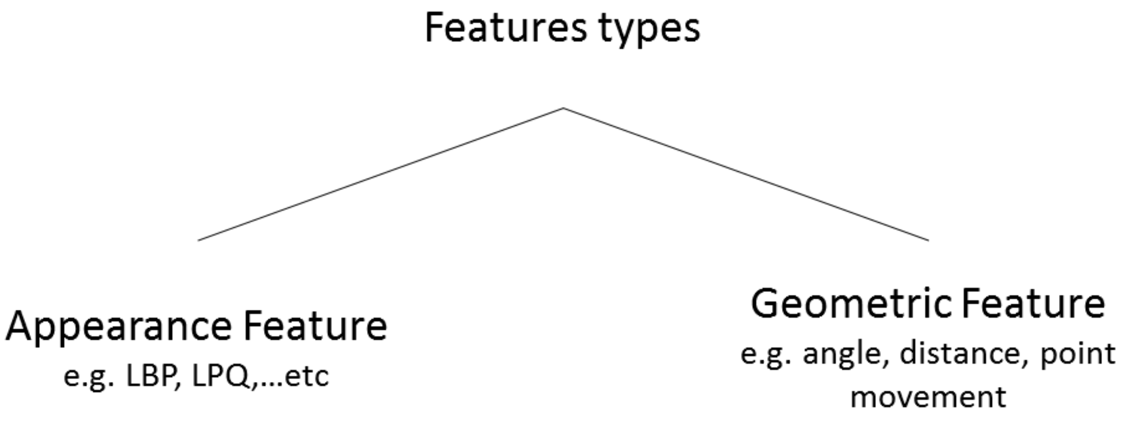


Figure 2.6: Types of Feature Extraction.

This technique is complicated to generalise to different people, although it is less sensitive to noise [3]. In addition, it contains extremely significant information regarding expression recognition, as it can encode micro patterns existing in the skin texture of the face. Nevertheless, the features of both techniques play a significant role in facial expression recognition [52, 63]. Fig. 2.6 displays two main types of features in FER; specifically appearance and geometric features.

2.4.1 Appearance based feature methods

Appearance-based features are contrary to the geometric features and represent information on the texture of the face. Thus, they are able to capture face changes caused by bulges, wrinkles and furrows more effectively [65]. The original pixels are used for the facial image as an appearance descriptor. Furthermore, several advanced features that are more suitable for analysing facial expressions have been suggested. Over time, appearance based feature methods include different methods, as follows:

A. Local Binary Patterns based Methods

LBPs were initially designed as a texture descriptor for texture classification problems [66]. LBPs are some of the most common features used in the FER literature.
Originally, when using LBP [66], each pixel of the input image is specified by a decimal number that is termed an LBP label and calculated by way of binary thresholding, where, a gray level with P neighbour is located on a circle of radius r, which is centred on the pixel. Additionally, bilinear interpolation has been exploited to calculate the neighbouring pixel values and compared with each block in a face image, for example (3×3) with the pixel on the centre. Fig. 2.7 provides an illustration of the basic LBP operator.

Ojala et al. [66], also had extended the LBP descriptor by allowing handling for different numbers of the neighbourhood points (P). The circular neighbourhoods are exploited for bilinear interpolation of pixel values, which allow any number of neighbourhood pixels to be decided at any radius (R). Fig. 2.8 represents the extended LBP descriptor that shows the distribution of the neighbourhood points (P) circularly in symmetric locations in different radius (R) on the centre pixel. Consequently, the extended LBP descriptor can then be employed to encode different varieties of micro-information such as corners, edges and spots,

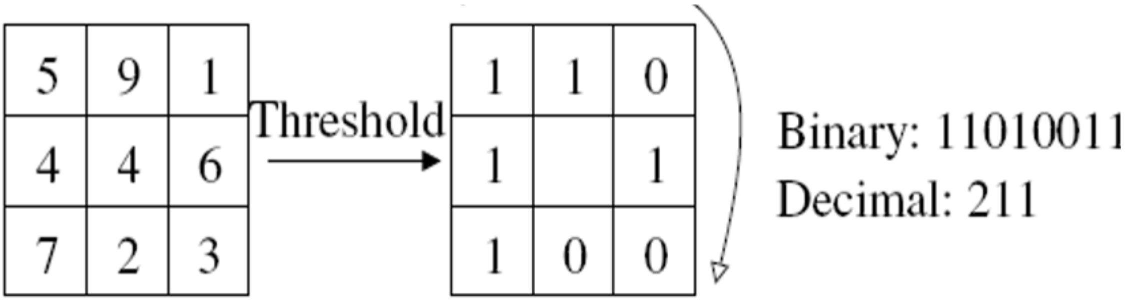


Figure 2.7: The basic LBP operator.

as displayed in Fig. 2.9. Hence, the extended LBP can be expressed for any point in the local region such as (x_c, y_c) as follows

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{p-1} s(i_p - i_c)2^P, \tag{2.1}$$

where i_c and i_p are the gray-level values of neighbours (P) with the centre pixel (x_c, y_c), and R is the radius.

The d represent the two discrimination levels of the LBP descriptor:

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

Ojala et al. [66] recorded that not all patterns (2^P) that were created from the LBP descriptor were important. For instance, if the sum of the number of neighbourhood points P is 8, that will create 256 (2^8) patterns. Hence, this leads to the selection of only an effective subset from these patterns. Therefore, selection of patterns which have at least two bitwise transitions was proposed. For example, 00000000, 00000111 and 11100001. These patterns are labelled as uniform, otherwise all other patterns are labelled as non-uniform. In this particular case, the LBP descriptor has 59 patterns which are divided into 58 uniform patterns and one nonuniform pattern. According to that, the LBP descriptor $LBP^{u2}_{P,R}$ is indicated.

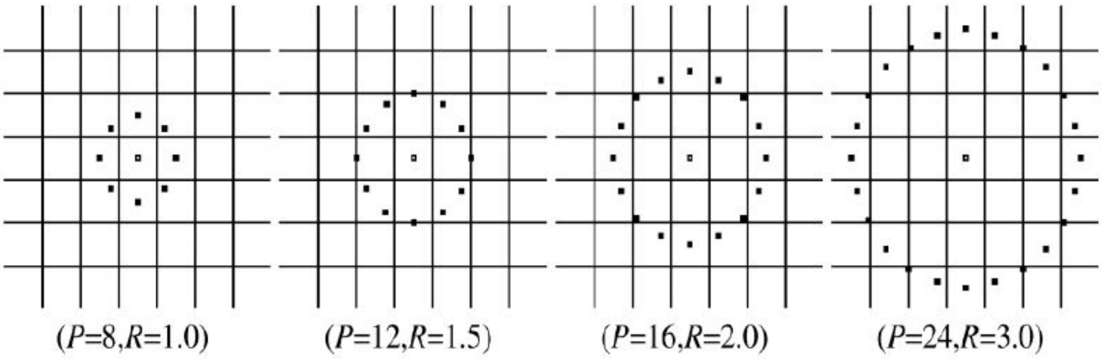


Figure 2.8: Examples of the extended LBP descriptor with symmetric neighbourhood points (P) with different radius (R) [54].

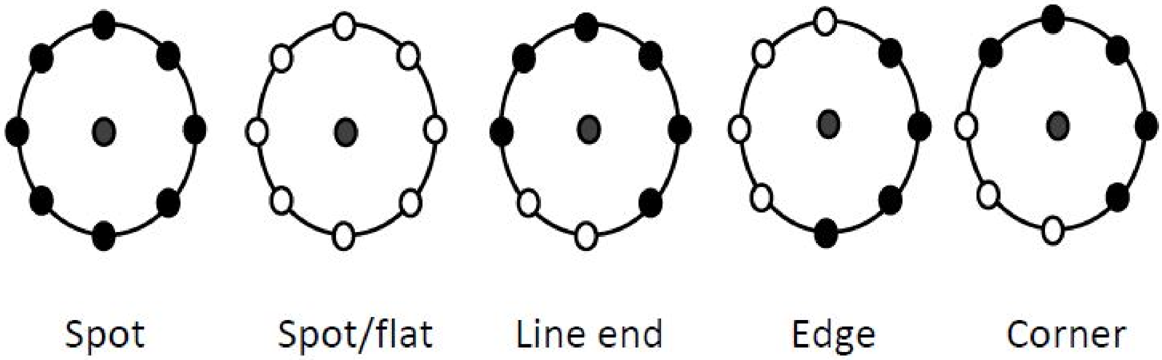


Figure 2.9: Examples of different types of micro-texture patterns represented in black and white circles indicate bits 0 and 1 in the yield of the LBP descriptor, respectively [55].

LBP is described by uniform patterns $LBP^{u2}_{P,R}$ [67] in the case that the inclination of the binary pattern is predominantly of two bit-wise transitions from 0 to 1 or vice versa in the case that the bit pattern is considered circular.

A histogram (H_i) of the labeled image $f_l(x, y)$ is expressed by

$$H_i = \sum_{x,y} I\left\{f_l(x, y) = i\right\} \tag{2.3}$$

where $i = 0, 1, ..., n - 1$, in which n is the number of various labels provided by the LBP operator and the indicator function $I(x)$ is

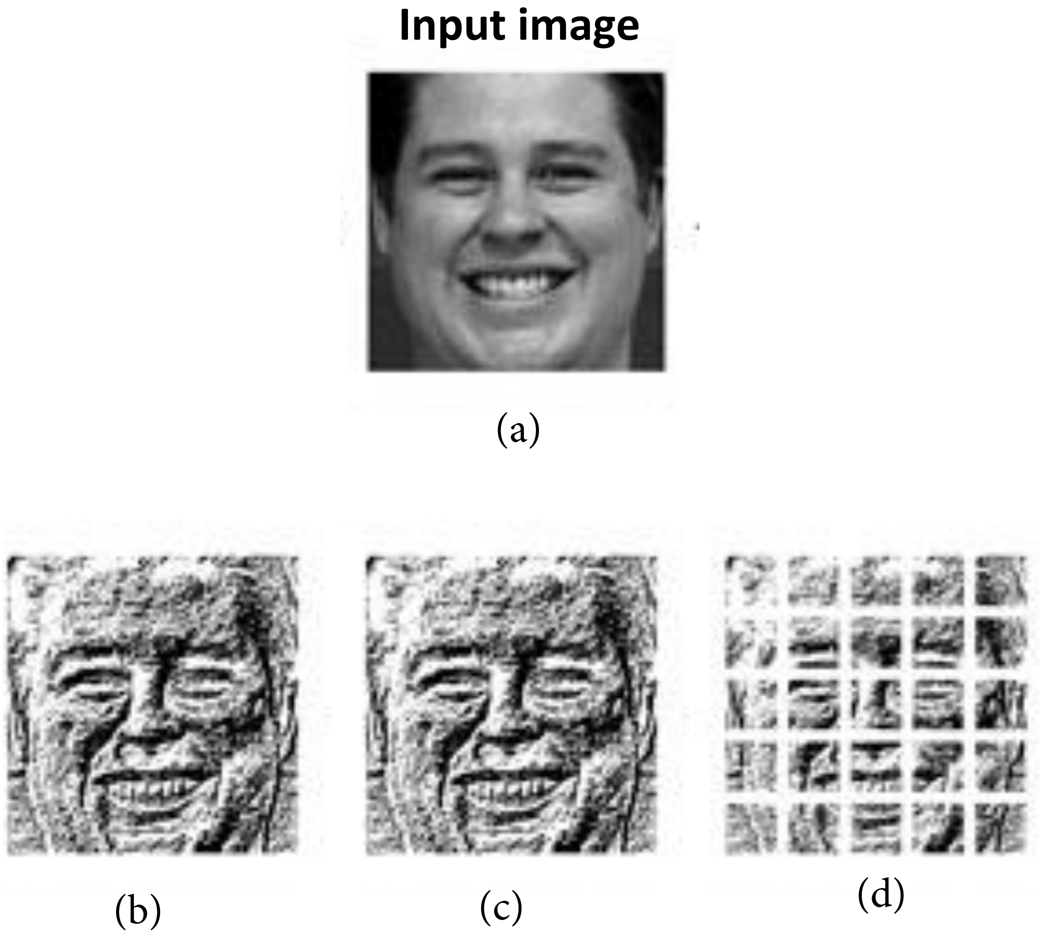


Figure 2.10: Samples of an expressive image with LBP features, (a) is original image and (b), (c), and (d) denote the expressive images after applying LBP features.

$$I(x) = \begin{cases} 1 & if \quad x \; is \; true \\ 0 & if \quad x \; is \; false \end{cases} \tag{2.4}$$

This descriptor was widely applied to face analysis and had a good achievement in the classification process, in particular, for the static and dynamic FER applications. Fig. 2.10 shows an example of basic LBP descriptor.

B. Local Phase Quantization based methods

LPQ was originally introduced by Ojansivu and Heikkila [68] for texture description. The LPQ achieves good results, especially when handling blurred images. The spatial blurring is expressed via convolution of the image intensity and a point spread function. Let us assume that $e(\mathbf{x})$ is an original image and $g(\mathbf{x})$ is the observed image, consequently, by convolution it can be represented by the discrete model for the spatially invariant blurring of $e(\mathbf{x})$ as

$$g(\mathbf{x}) = e(\mathbf{x}) \otimes h(\mathbf{x}). \tag{2.5}$$

where $h(\mathbf{x})$ is the Point Spread Function (PSF) of the blur, $\otimes$ is 2-D convolution and $\mathbf{x}$ indicates a vector of coordinates $[\mathbf{x},\mathbf{y}]^T$.

$$G(\mathbf{u}) = E(\mathbf{u}).H(\mathbf{u}). \tag{2.6}$$

where $G(\mathbf{u}), E(\mathbf{u})$ and $H(\mathbf{u})$ correspond to the Discrete Fourier Transforms (DFT) of the blurred image $g(\mathbf{x})$, the original image $e(\mathbf{x})$, and the PSF $h(\mathbf{x})$, respectively, $[\mathbf{u},\mathbf{y}]^T$ are the coordinate points of the vector. The magnitude and phase are separated by

$$|G(\mathbf{u})| = |E(\mathbf{u})| \cdot |H(\mathbf{u})\ |, \angle G(\mathbf{u}) + \angle H(\mathbf{u}). \tag{2.7}$$

The Fourier transform is usually real-valued due to the fact that the blur PSF $h(\mathbf{x})$ is centrally symmetric, $h(\mathbf{x}) = h(-\mathbf{x})$. Therefore, a consequence of phase is only a two-valued function given by

$$\angle H(\mathbf{u}) = \begin{cases} 0 & if \quad H(\mathbf{u}) \geq 0, \\ \pi & if \quad H(\mathbf{u}) < 0. \end{cases} \tag{2.8}$$

In local neighborhoods N_x of the LPQ, the phase is examined at each pixel position $\mathbf{x}$ for the image $f(\mathbf{x})$. These local spectra are calculated by utilizing a short-term Fourier transform represented by

$$F(\mathbf{u},\mathbf{x}) = \sum_{y \in N_x} f(\mathbf{x}\text{-}\mathbf{y}) e^{-j2\pi \mathbf{u}^T \mathbf{y}}, \tag{2.9}$$

where $\mathbf{x} \in \{x_1, x_2, ..., x_N\}$ and consists of a simple one dimension 1-D convolution for the rows and columns successively. In the four frequency points $\mathbf{u}_1 = [a, O]^T, \mathbf{u}_2 = [O, a]^T, \mathbf{u}_3 = [a, a]^T$ and $\mathbf{u}_4 = [a, -a]^T$, the local Fourier coefficients are calculated, in which a is a small scalar enough to satisfy $H(\mathbf{u}_i) > 0$. For every pixel position the vector of results is:

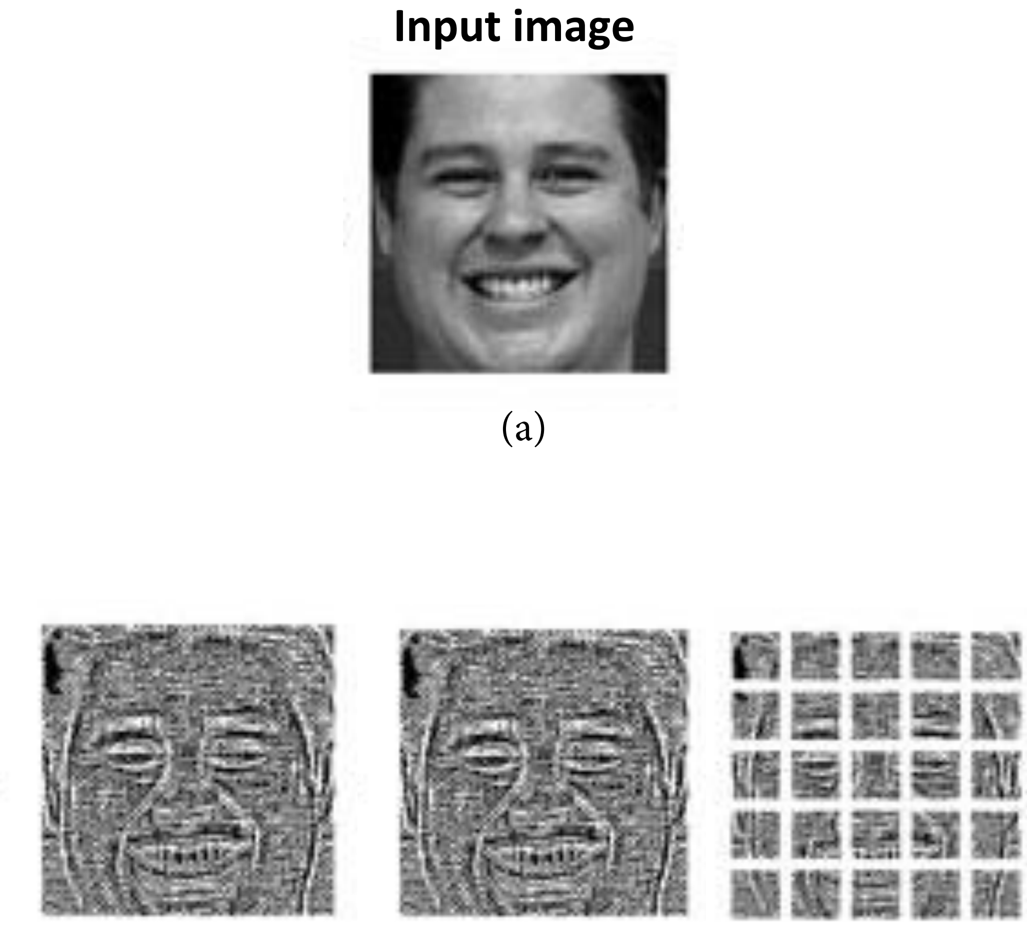


Figure 2.11: Samples of an expressive image with LPQ features, a is original image and b, c, and d denote the expressive images after applying LPQ features.

$$E(\mathbf{x}) = [E(\mathbf{u}_1, \mathbf{x}), E(\mathbf{u}_2, \mathbf{x}), E(\mathbf{u}_3, \mathbf{x}), E(\mathbf{u}_4, \mathbf{x})]. \tag{2.10}$$

The phase information is calculated by observing the signs of the real and imaginary parts of every component in $E(\mathbf{x})$.

$$q_i = \begin{cases} 1 & if \quad g_j > 0, \\ 0 & otherwise. \end{cases} \tag{2.11}$$

Fig. 2.11 illustrates an example of an expressive image with its LPQ.
In various research works [69–72] LBP and LPQ were compared in terms of overall affect recognition performance. Typically, the LPQ outperforms LBP because of the size of the local description and as LBPs are usually extracted from smaller areas which are 3 × 3 pixels diameter [73], while LPQs are extracted from a larger area, which is approximately 7 × 7 pixels [69, 70]. Furthermore, if LBPs are extracted from larger regions, it causes loss of information as they ignore the pixels which remain inside the centre region. On the contrary, LPQ integers described the regions as a whole.

C. Other methods

Several other descriptors used to extract local appearance features will be briefly reviewed.
- Gabor Wavelets Transformation is a powerful collective time-frequency tool that is employed in image analysis based on Gabor filters [74], which is one of the most important feature extraction techniques in facial expression recognition [75]. Gabor wavelets have been applied for years in many feature extraction algorithms, due to their strong similarity with the mechanism of perception in the human visual system [27], their ability to provide representations of multi-resolution/ multi-orientation [76, 77], and, as they encompass a large amount of essential visual features regarding FER.
Fig. 2.12 shows the real component of the Gabor filters parts at five scales and eight orientations in (a), and the magnitude component of the Gabor filters with five scales in (b). Fig. 2.13 demonstrates the convolution of an image with 40 Gabor wavelets.

Various researchers have used Gabor wavelets in facial expression recognition. The following is a summary of a few. In 1998, Lyons [78], completed the first study in this field and described using Gabor wavelets to code facial expressions. The results of this study indicate the possibility of building a facial expression recognition system automatically based on Gabor wavelet code, which is of importance in psychological plausibility.

Sisodia et al. [79], used a Gabor filter to detect appearance based features initially, then converted the image to grayscale. Thereafter, the face was detected and resized. Fewer features

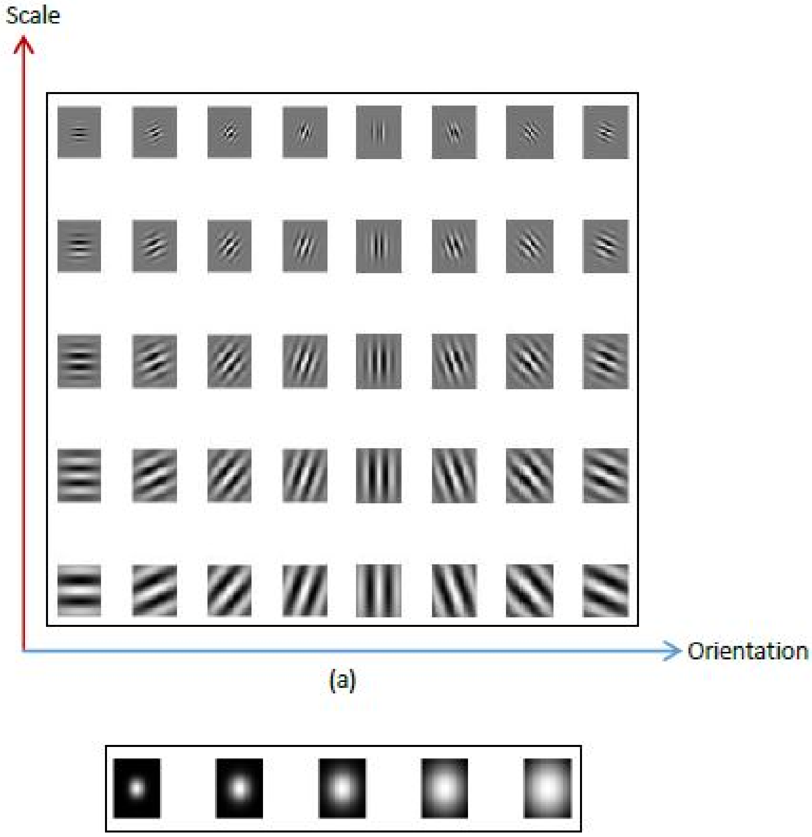


Figure 2.12: Gabor wavelets map. (a) The real components parts at five scales and eight orientations. (b) The magnitude components.

which are more representative of the feature selection method were selected in this case. Samad et al., [80] improved the algorithm to identify a minimum number of Gabor wavelet parameters for natural FER. The image was converted to grayscale and resized. Subsequently, a Gabor filter was applied after that and Gabor features down sampled. Thereafter, feature vector dimensions were reduced using PCA.

Abdulrahman et al., [81] used the Gabor wavelet transform approach in facial expression recognition that was used as a pre-processing stage to extract a feature vector representation. Additionally, in 2015, Hegde et al., [82], represented facial expression recognition by the employment of both Gabor magnitude and phase information. The dimension of both the magnitude feature vector and phase feature vector of Gabor is reduced by removing unnecessary data.

- ***Histogram of Oriented Gradients (HOGs) Method*** [83] represents images through the orientation of the edges that are contained. HOGs are used to extract local features by using gradient operators via the image and subsequently encoding their output from both gradient magnitude and angle. The HOGs method works by extracting local magnitude-angle

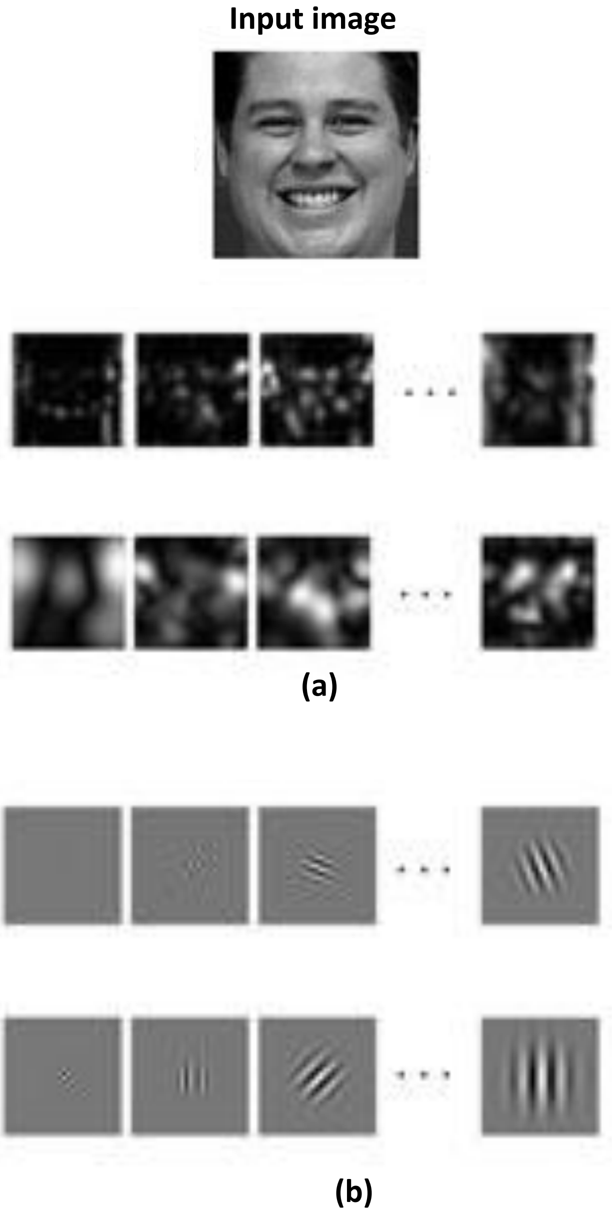


Figure 2.13: Gabor wavelets. (a) The magnitude components. b The real parts at five scales and eight orientations.

histograms from cells. Next the local histograms are grouped via larger entities (blocks); thus, overlapping the blocks increases the dimensions [83]. HOGs had been applied by way of an outstanding system in the FER emotion challenge [84]. Low-level representations (LBP and HOGs) are compared from many perspectives, in terms of sensitivity to registration errors, histograms are less sensitive [85]. Fig. 2.14 provides an example of an expressive image with its HOGs.

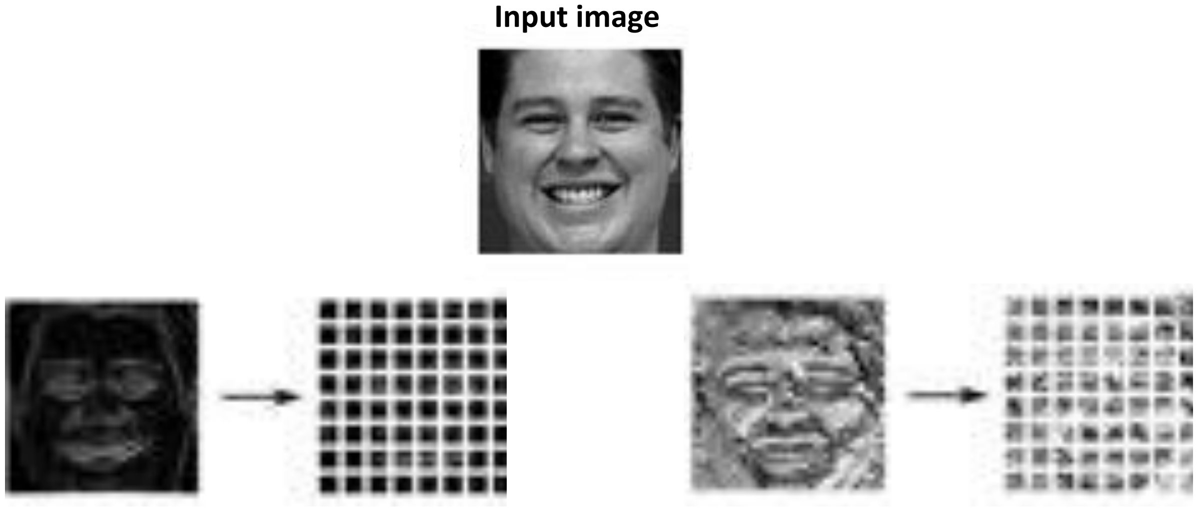


Figure 2.14: An expression image with HOG features.

- ***Scale Invariant Feature Transform (SIFT)*** [86] has been commonly used in real world applications because of its effective computations, resistance to partial occlusions and its relative insensitivity.

- Later, ***Local Ternary Pattern (LTP)*** [87] was proposed to increase the robustness of LBP of uniform and near-uniform regions by adding an additional level of discrimination to the intensity. The binary LBP value is subsequently extended to a ternary code by encoding the small pixel difference into a third state.

- Recently, the ***Sobel Operator Method*** [88] was suggested to improve the performance of LBP to enhance the edge information prior to using LBP for feature extraction. However, Sobel-LBP generates inconsistent patterns in uniform and near-uniform regions, seeing as it applies only two levels of discrimination similar to LBP.

In 2010, a different approach was proposed for texture encoding termed Local Directional Pattern (LDP) [89, 90], which used directional edge response values concerning a position, instead of gray levels. In [91], LDP achieves better performance in recognition than LBP, although LDP tended to produce inconsistent patterns in uniform and near-uniform facial regions, that is heavily based on the selection of a number of prominent edge direction parameters. Considering the limitations of the present local texture descriptors, a new texture pattern called Gradient Local Ternary Pattern (GLTP) [92] was proposed to use in person independent facial expression recognition. The GLTP operator encodes for the local texture information using three varied discrimination levels to quantize the local neighbourhood of gradient magnitude values. The GLTP encoding scheme has the ability to differentiate between limitations at high and smooth level in textured facial parts, ensuring that the formation of texture micro-patterns is consistent with the characteristics of the local image (high or textured).

2.4.2 Geometric-based feature method

Geometric facial feature techniques focus on describing the shape changes of a face structure, for instance the shape and locations of facial components which are modulated by a set of landmark facial points. For example, shrinking and widening of the eye shape, nose, lips and eyebrows. This technique primarily concentrates on detecting the angular changes and displacement of the feature points compared with a neutral expression. Generally, two main approaches have been exploited to extract the geometric features: 1) facial point displacements and 2) grid node displacements. However, these types of features can be used individually or combined.

1. Facial Point Displacement

The landmark facial points locations are considered fundamental to extracting the geometrical features of a face. After the face detection stage, diverse methods can be employed to detect the locations of landmark facial points. For example, AAM, ASM, CLM or one of the more recent such as DRMF and an Active Orientation Model (AOM) [93]. In 1998, Zhang et al., [94], used 34 fiducial points in static images to represent facial feature geometry. In [95], multi-state models were used for geometric feature extraction by using 15 parameters of geometric features in dynamic images. In 2006, Valstar and Pantic [96], derived the geometric facial features of 20 points that include (x. y) coordinate displacements. The points are tracked using a modified particle filtering scheme, where distances between points are coordinated based on the point and their first temporal derivatives. Chang et al. [97] proposed the classification of facial expressions to a low-dimensional emotion manifold by applying a modified Lipschitz function which embeds 58 facial landmarks in video sequences that represents the changes to a shape model. In [98], 68 facial points which are basically located around the contours of the eyes, eyebrows, nose, lips, inner lips and chin are used to represent the 2D face shape. Lucey et al. [99], proposed a geometrical study to classify extended CK database to six basic expressions plus a contempt emotion by using AAM, which included 68 landmark facial points. Fig. 2.15 shows an example of key points around the eyebrows, eyes and mouth, which are employed to discriminate facial expressions.

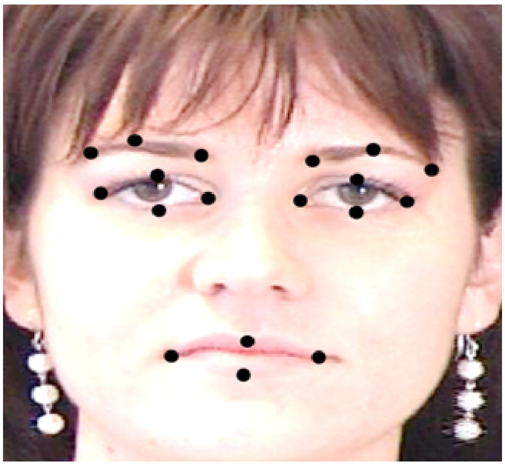


Figure 2.15: Example of key facial points.

2. Grid Node Displacements

The second method for extracting geometrical features is a wire-frame model termed Candide, which is a famous parameterised face mask employed for tracking procedures. Rydkalk [46], developed the model of human-faces, which involves approximately 104 of the nodes that form a number of triangles to describe the facial animations [56], modelled through the global and local action unit. A simple candidate version can be created to analyse facial expression by selecting important nodes around the eyes, eyebrows and mouth. Kotsia and Pitas [56] introduced a geometrical model by using the Candide wire-frame model to track the grid deformation, where the Candide wire-frame model was placed on the first frame of the video sequence frames randomly. Then, the significant nodes are manually selected (simplified Candidate model) to match the original face. Important nodes are automatically selected in several cases, by using the elastic graph matching algorithm in some cases. Subsequently, deformations of the wire-frame model were tracked along a video sequence by the KLT algorithm. Finally, in the video sequence, the displacements of the grid node are calculated between the first and apex frames. Fig. 2.16 shows the different samples of different facial expressions with Grid Node Displacements. Kotsia in [100], used geometrical displacement information pertaining to the grid node coordinates, while SVMs or FAU-based on facial expression recognition are employed.

This thesis adopts two appearance features (LBP and LPG) and one geometric feature (Facial Point Displacement) due to their documented high performance and extensive use in previous FER works. Table 2.1 gives a summary of geometric-based feature methods.

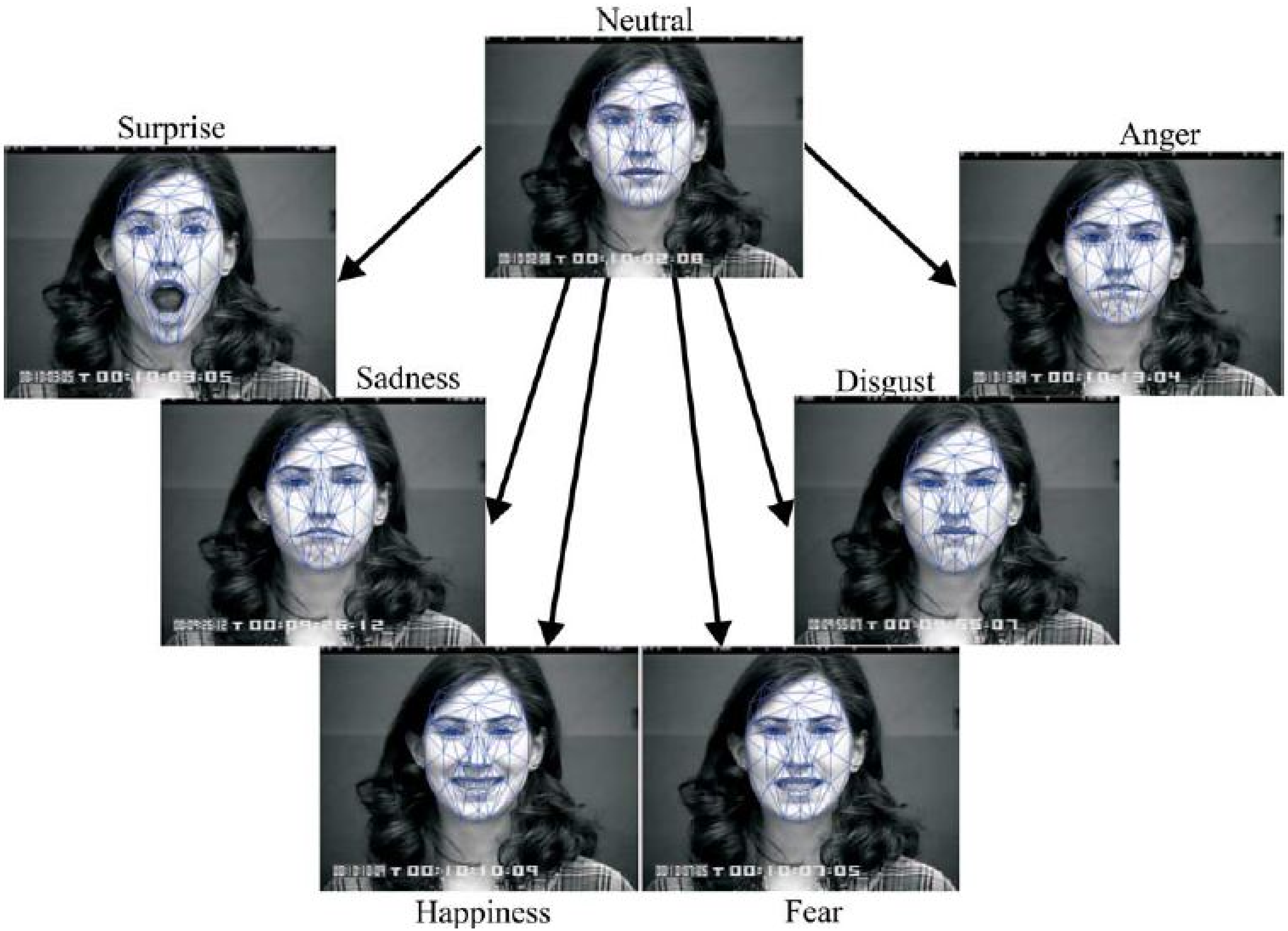


Figure 2.16: Samples of different facial expression with Grid Node Displacements [56].

Table 2.1: Summary of geometric-based feature methods

Reference	Features	Tracking method	Static/Dynamic	Classifier
Zhang et al. (1998)	34 facial feature points	Manual labeling	Static	Two-LayerPerceptron
Tian et al. (2002)	15 parameters of geometric features	Multi-state models	Dynamic	Three-Layer Neutral Network
Zhang & Ji (2005)	Geometric deformation feature of AUs	Kalman filtering	Dynamic	Dynamic Bayesian Networks
Kanaujia & Metaxas (2006)	78 facial feature points	Modified active shape model	Dynamic	Conditional Random Fields (CRF)
Kotsia & Pitas (2007)	Geometric deformation feature	Kanade-Lucas-Tomasi (KLT) tracker	Dynamic	Support Vector Machine (SVM)
Shin & Chun (2008)	18 facial feature points	Dense optical flow	Dynamic	Hidden Markov Models
Zafeirious & Petrou (2010)	Geometric deformation feature	KLT tracker	Static	Sparse Representation
Jain et al. (2011)	68 facial points	Generalized proscrustes analysis	Dynamic	Latent-Dynamic CRF
Rudovic et al. (2013)	39 facial feature points	Active appearance model	static	SVM
Majumder& Behera (2016)	22 facial feature points	data fusion using autoencoders	Dynamic	SOM-based classifier.

2.4.3 Dimensionality Reduction Methods

This section presents the most common methods in the feature selection by reducing the dimensionality of the vector of the features before feeding the classifier.

The most well-known and holistic method is Eigenfaces, known as Principal Component Analysis (PCA) [57]. The PCA method is an unsupervised learning technique. The covariance matrix is found to calculate the eigenvalues and eigenvectors, in order to obtain the Eigenfaces which account for the most variance within the set of face images by using only the best Eigenfaces that correspond to the largest eigenvalues. The correlation of the eigenvector with the greatest eigenvalue that is one of the reasons for the large variance in the image [101]. Moreover, the eigenvector associated with the smallest eigenvalue finds the least variance (further details are shown in Fig. 2.17. Eigenfaces work well with good-quality images that are captured under similar strict conditions, such as pose, light and facial expressions. However, there is a problem in relation to the linear selection of the reduction of the dimensionality of an unsupervised technique and as such, it does not include label information of the data. As the learning set is labelled, it therefore makes sense to use this piece of information to build a highly-reliable approach to dimensionality reduction in the feature space. More detail is explained in Chapter 4.

2.4.3.1 Principal Component Analysis

Principal Component Analysis (PCA) is one of the major methods to extract features discussed above, as well as reduce the dimensionality of the image by extracting key features.

Eigenfaces Method

The eigenface technique is used to reduce dimensionality in the Principal Component Analysis (PCA) method. In using PCA in facial expressions analysis, the 2D image is converted into a 1D vector via concatenating rows. An image I_i is converted into a vector of length $N = mn$ [102], as below

$$I = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1n} \\ x_{21} & x_{22} & ... & x_{2n} \\ . & . & . & . \\ . & . & . & . \\ x_{m1} & x_{m2} & ... & x_{mn} \end{bmatrix}_{m\times n} \rightarrow \begin{bmatrix} x_{11} \\ . \\ . \\ . \\ x_{1n} \\ . \\ . \\ . \\ x_{2n} \\ . \\ . \\ . \\ x_{mn} \end{bmatrix}_{1\times N} = \mathbf{x}. \tag{2.12}$$

Assume M such vectors $\mathbf{x}_i$ where $(i = 1, ..., M)$ with length N representing a matrix $\mathbf{X}$ of learning images. It is important to center the matrix to ensure the first principal component describes the maximum variance direction. Then compute the vector of mean values $\mathbf{\Psi}$ of all training images through

$$\mathbf{\Psi} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x}_i, \tag{2.13}$$

The mean centered image is calculated by

$$\phi_i = \mathbf{x}_i - \mathbf{\Psi}. \tag{2.14}$$

and is arranged to form a new training matrix of size $(N \times M)$; $\mathbf{A} = (\phi_1, \phi_2..., \phi_M)$.

The covariance matrix $\mathbf{C}$ is calculated through

$$\mathbf{C} = \sum_{i=1}^{M}(\mathbf{x}_i - \mathbf{\Psi})(\mathbf{x}_i - \mathbf{\Psi})^T = \mathbf{A}\mathbf{A}^T, \tag{2.15}$$

Then, the eigenvectors $\mathbf{e}_i$ and eigenvalues λ_i, are found through solving

$$\mathbf{C}\mathbf{e}_i = \lambda_i \mathbf{e}_i. \tag{2.16}$$

The dimensions of the covariance matrix are $N \times N$ which could be massive. For instance, images of size 60×60 produce the covariance matrix of size 3600×3600. As such this method is not a practical solution to find eigenvectors of $\mathbf{C}$ directly. The eigenvectors $\mathbf{e}_i$ and eigenvalues λ_i of the covariance matrix $\mathbf{C}$ can be obtain by $\mathbf{C} = \mathbf{A}^T\mathbf{A}$ that has size of $M \times M$. Suppose $\mathbf{v}_i$ is an eigenvector of $\mathbf{A}^T\mathbf{A}$ such that

$$\mathbf{A}^T\mathbf{A}\mathbf{v}_i = \mu_i \mathbf{v}_i. \tag{2.17}$$

Multiplying both sides by $\mathbf{A}$ yields

$$\mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{v}_i = \mathbf{A}\mu_i \mathbf{v}_i,$$

$$\mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{v}_i) = \mu_i(\mathbf{A}\mathbf{v}_i),$$

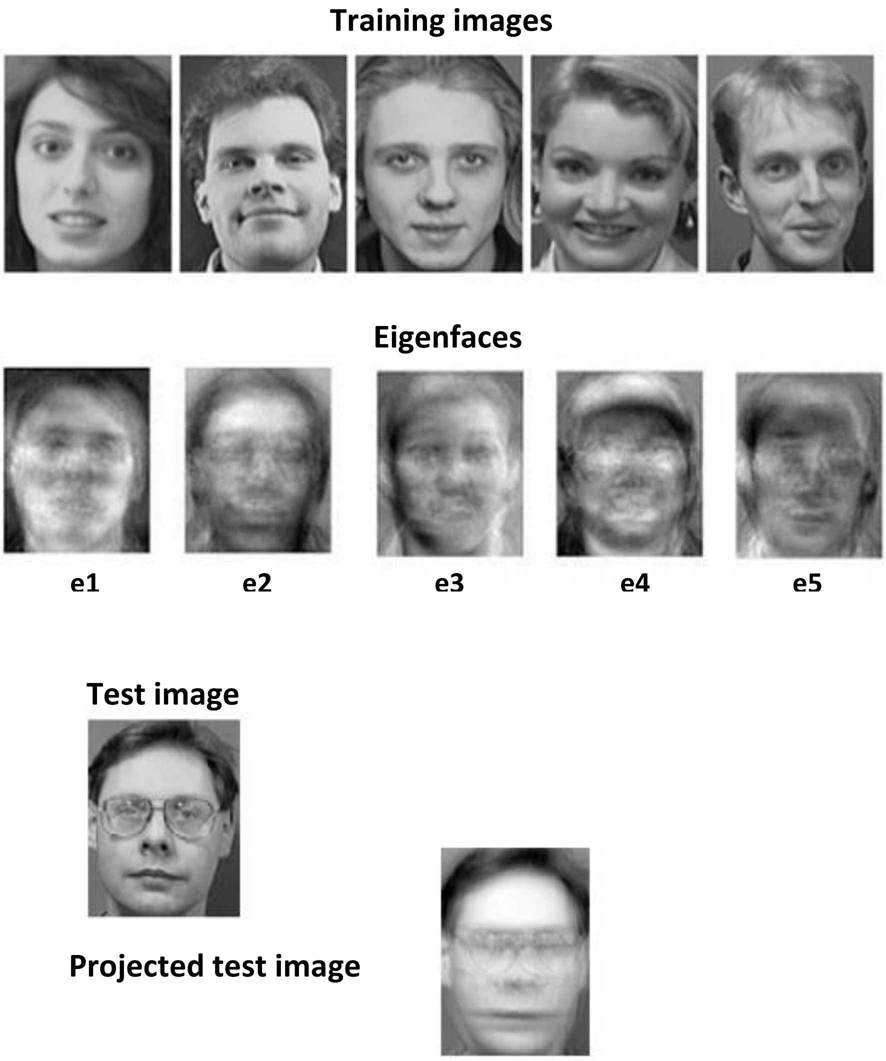


Figure 2.17: This scheme is named Eigenfaces as these eigenvectors can be reconstructed and visualized being face images [57]

where $(\mathbf{Av}_i), i = 1, ..., M$, are the eigenvectors of $\mathbf{C} \quad = \quad \mathbf{A}^T\mathbf{A}$

$$\mathbf{C}(\mathbf{Av}_i) \quad = \quad \mu_i(\mathbf{Av}_i). \qquad (2.18)$$

By comparing equations 2.16 and 2.18, it is concluded that the first $M - 1$ eigenvectors $\mathbf{e}_i$ and eigenvalues λ_i are provided by $\mathbf{Av}_i$ and μ_i, respectively. Thus, the eigenvector with the highest eigenvalue provides the highest variance while the eigenvector with the lowest eigenvalue finds the smallest variance [5].

$$\mathbf{C} = \sum_{i=1}^{M} (\mathbf{x}_i - \mathbf{\Psi})^T (\mathbf{x}_i - \mathbf{\Psi}) \quad = \quad \mathbf{A}^T \mathbf{A} \tag{2.19}$$

Consequently, the vectors are sorted via eigenvalues, in which the initial vector corresponds with the highest eigenvalue.

2.4.3.2 Fisher Linear Discriminant Analysis

Fisherface Method

Another approach employed that uses class particular linear approaches to reduce dimensionality and straightforward classifiers to reduce the feature space and obtains better recognition rates compared with the eigenface method is considered. This technique is Linear Discriminant Analysis (LDA), which is a statistical method to reduce dimensionality by using linear transformation. Therefore, the projection increases the distance between classes and reduces the variation within each class. This defines the Fisher standard, which has been maximized on all linear projections. To achieve that, suppose a D-dimensional space has a set of N_i samples $\{x_1, x_2, ..., x_n\}$ and presume that each class c contains a set of images $\{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n\}$ [103]. Then, let the between-class scatter matrix be as follows:

$$\mathbf{S}_B \quad = \quad \sum_{i=1}^{c} (\mu_i - \mu) (\mu_i - \mu)^T \tag{2.20}$$

$$\mathbf{S}_W \quad = \quad \sum_{i=1}^{c} \sum_{X_k \in X_i} (X_k - \mu_i) (X_k - \mu_i)^T \tag{2.21}$$

where: μ means the mean of the entire training set, μ_i means the mean of class $\mathbf{X}_i$, and N_i is the number of samples in the class.

The formation of the optimal projection matrix of orthogonal columns that increases the determinant ratio $\mathbf{S}_B$ to the determinant of $\mathbf{S}_W$, is denoted by $\mathbf{W}_{opt}$ and it is defined by:

$$\mathbf{W}_{opt} = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{|\mathbf{W}^T \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|} = [\mathbf{w}_1, \mathbf{w}_2, ...\mathbf{w}_m] \tag{2.22}$$

where: $\{\mathbf{w}_i | i = 1, 2, \ldots, m\}$ is the set of eigenvectors of $\mathbf{S}_B$ and $\mathbf{S}_W$ which are compatible with the m largest eigenvalues $\{\lambda_i | i = 1, 2, \ldots, m\}$. For example,

$$\mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i, \quad i = 1, ..., m \quad \Rightarrow \quad \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{w}_i \tag{2.23}$$

The practical way to obtain the optimal projection matrix for FLDA is to first apply PCA, and then find the optimal solution for FLDA. This can be given by:

$$\mathbf{W}_{opt} = \mathbf{W}_{pca} \mathbf{W}_{flda} \tag{2.24}$$

where

$$\mathbf{W}_{pca} = \underset{\mathbf{W}}{\operatorname{argmax}} |\mathbf{W}^T \mathbf{S} \mathbf{W}| \tag{2.25}$$

$$\mathbf{W}_{flda} = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{|\mathbf{W}^T \mathbf{W}_{pca}^T \mathbf{S}_B \mathbf{W}_{pca} \mathbf{W}|}{|\mathbf{W}^T \mathbf{W}_{pca}^T \mathbf{S}_W \mathbf{W}_{pca} \mathbf{W}|} \tag{2.26}$$

where $\mathbf{W}_{opt}$ which is the resulting optimal projection matrix that is used to reduce dimensionality for a given image $\mathbf{x}_k$ into the feature space.

2.5 Expression Classification/Recognition

The next phase of the automatic expression recognition system is classification/recognition. Classification (supervised learning) is labelling on a new sample or data, according to training data that contains observations or examples which define membership of the category.
Classification consists of two stages: the first stage is the training stage of input which is a set of examples that label class as one of the emotions, while the other set of feature values is reduced. The training data contain examples of all types of emotion with sufficient data for training. The second stage is the classification of the emotion into various classes such as happiness, anger, fear, sadness, disgust and surprise, which are given to a set of reduced feature vectors.
In real-time, the performance of the classification algorithm must produce greater accuracy and efficiency. Therefore, there are different methods of classification in the real-time in FER. Several of the most commonly used classification techniques are:

1. K-Nearest Neighbour [104] is commonly used as it is simple and robust, especially if the techniques that are extracting the feature before the classification would have effectively clustered the information in classes in the feature space. K-Nearest Neighbour (KNN) employs a euclidean distance to measure between the elements of the projections of the test samples with the elements of the projections of the training samples to judge their association. The projection of the test sample belongs to the same emotion class, such as the projection of the training sample to the lowest value given by the Euclidean distance.

2. Sparse Representation Classifier (SRC) is a typical combination of machine learning and compressed sensing, which is considered a learning machine, wherein the classification stage is achieved by applying signal reconstruction techniques. SRC is a nonparametric learning method that can directly designate a class label for a test sample without conducting a training process. In this way, it is similar in work to the Nearest Neighbour (NN) and nearest subspace. That means that they do not need weight vector (hypothesis functions and learn the parameters) of the hypothesis function. However, [105], [106] and [107] proved that SRC has a better classification performance than K-NN on face classification.

3. Support Vector Machines based methods (SVM) [108]: the SVM algorithm has built a model that locates new examples to one category or the other, making it the classifier

of a non-probabilistic binary linear. Furthermore, SVM represents examples of points in space as mapped, in which examples of the separate categories have been divided in an obvious gap, which as wide as possible. Consequently, new examples are mapped into the same space which are expected to belong to any category depending on which side of the gap they fall.

4. Support Vector Regression (SVR) [109] shows great potential for high-dimensional regression tasks in the input data; because its improvement does not depend on the input space dimensionality. SSVR is used with regression problems through the introduction of an alternative loss function. The distance from the correct prediction is employed to weight the actual error of the point, such that it is based on a subset of the training only to construct the model.

5. Artificial Neural Network (ANN) is one of the most widely accepted methods of artificial intelligence techniques that began in the 1990s [110]. It is a mathematical representation that is inspired in the similar way that the brain processes information. There are many ANN models and the most popular model in classification being a Restricted Boltzmann Machine (RBM) [111], which is trained to employ unsupervised learning to provide low-dimensional (2D).

6. Self-Organising Map (SOM) [112] has been broadly adopted in [113, 114] as a technique for unsupervised clustering and providing the inherent capability of topological ordering of the classes. The SOM-based classifier [112]] works in two phases; specifically training and mapping. In training, classification of new vectors is automatic, while in the mapping, the map is constructed by using input examples. In the proposed technique, the SOM is trained with all geometric features after utilising auto-encoders to obtain a closer model of the data according to the corresponding expression. The training processor SOM is complementary to the process of the auto-encoder, where it provides the final decision for the classification.

In this thesis, SRC and SOM are employed for discrete and dimensional facial expression recognition.

2.6 Related FE Works

This section focuses on the previous works of feature extraction and classification techniques and data fusion techniques.

2.6.1 Previous Work on Feature Extraction and Classification Techniques

Significant progress has been achieved in techniques for the appearance and geometric features extraction to build an automatic facial expression recognition system. Various techniques commonly used for appearance features extraction are Gabor filters [74], LBPs [66], linear discriminant analysis (LDA) [115], ojansivu2008blur, LPQ and HOG [83]. In an automatic facial expression recognition system, Gabor filters provide the best accurateness regarding the recognition [74]. However, Gabor filters handle (5×8) that are represented by five different scales and eight different orientations [116]. Therefore, it is associated with a high computational cost. For example, a face image with size (240×240) with the Gabor feature dimension became ($240 \times 240 \times 40$) that created a huge dimensional feature that adds heavily to the computational cost. Correspondingly, the LBP [54] has more advantages over Gabor filters and LDA, due to extracting a feature vector with a relatively low-dimensional feature by employing nonlinear operations. Ojala et al. [54], employed LBP successfully for facial analysis and with facial expression recognition. Shan et al. [73], proposed employing the histogram features of the uniform LBP descriptor in a comprehensive FER study, as measured by the adoption of different machine learning techniques in terms of recognition. It was acknowledged that applying boosted LBP features with a SVM classifier is better than using boosted LBP features only, where the best recognition rate with the proposed method was 91.40% and 95.10% for 7 and 6 classes, respectively.

Wang et al. [117], introduced the method for recognition of the facial expression by using LPQ then SRC. The histograms of local phase quantization are first used for facial image description, then sparse representation used for facial expression recognition.

There are two popular techniques for dimensionality reduction: PCA and LDA. In [5], the PCA method obtains a small group of very important features, which are used to describe the difference between face images. [118] presented a glance for face detection techniques based on the Viola-Jones algorithm and principal component analysis.

In 2008 [119], some of the problems that have been neglected in previous work when using Eigenfaces in PCA, such as which features are significant or not for classification are discussed. In [5], Eigenfaces and PCA are used to solve the dimensionality problems. In [120], a study to improve the complexity of (eigenfaces) in PCA, which does not have an impact on the recognition performance is considered. In [121], face recognition method using eigenfaces in PCA was proposed, which analysed facial expressions in images by emphasising different areas, for instance, the eyes and mouth, which are often affected via different facial expressions.

In 2013 Gosavi and Khot [122], used Principal Component analysis (PCA) to implement the technique of facial expression recognition. Various methods of face recognition and primarily emphasis on principal component analysis, which is undertaken for analysis and implementation in free software, Scilab. This system detects an image that is captured via a digital camera or webcam, which depends on descriptive features could check with the training image dataset. According to [123], it is suggested the eigenface approaches for the PCA algorithm used to discover the successful rate of detection and face recognition, as the experiment was conducted using Eigenfaces on 30 images of various students stored in a database training.

In [124], a novel method is proposed to solve the problem of face recognition with single training image per person by using (FLDA) to evaluate the class scatter matrix input from single training image availability. A face image can be analysed in two parts: the first part includes the smooth general appearance of the image, whereas the second part comprises a different image by using singular value decomposition (SVD). Four different methodologies for Face Recognition using PCA, FLDA, Minimum Euclidean Distance and Artificial Neural Networks have been described in [125].

The next stage in FERS is the process of selecting a suitable technique classification, which are SVMs [109], RBMs [111] and deep neural networks [110]. Zhi et al. [126] introduced sparse and graph-preserving properties used to reduce the dimensionally of the feature vector for facial expressions recognition. Lawrence et al. [113], applied a SOM and a convolutional neural network to recognise a face from local image sampling.

2.6.2 Previous Work on Data Fusion Techniques

Fusion multimodal data to extract features can result in more robust FERs as facial expressions are subjected to different characters such as colour, gender, race and lighting conditions. Many studies adopted the fusion of a variety of different features, as [127], [128], [129], [130] and [131].

Yu et al. [129] proposed a new technique for dimensionality reduction for the fusion of multiple features by way of using spectral embedding.

Zavaschi et al.[130] using LBP and Gabor features with an ensemble of classifiers, which used an ensemble of Multi-Objective Genetic Algorithms (MOGAs) based on size and accuracy. In addition, the processing had been so slow due to the use of MOGA and involvement of Gabor features which are high dimensional data. Senechal et al. [128] introduced a low level integrated represented by geometric by local Gabor binary pattern histograms with Multi-Kernel Learning (MKL) algorithm and geometric feature by AAM coefficients. In this study, the use of Gabor filters during training once again led to high computation cost. For example, each face image is generating 3×6 images represented by three spatial frequencies and six orientations. Liu et al. [127] introduced the low level represent by geometric and appearance features combined with high level features such as eyebrow gesture, head shake events and head nod within the multi-scale, spatio-temporal analysis for nonmanual grammatical markers recognition in American Sign Language. Majumder et al. [63], proposed combined appearance features using LBP and geometric features with a deep network and SOM-based classifier. Kumar et al. [132] proposed a fusion of geometric features using (ASM) and texture local binary pattern features with a multi-class of trained and tested features by way of a support vector machine (SVM).

2.7 Summary

This chapter presents background knowledge in relation to human emotion, effective human-computer interaction and processing stages in general FER research with an improvement system and also reviews related work. Moreover, the basic theories for the techniques and features employed in this thesis are presented. Appearance and geometry features reported complementary valuable information for FER. A wide variety of features and algorithms have been introduced that were applied in previous FER works. The literature related to the contribution made by this thesis has been presented from important previous works.

The following chapter will introduce types of database that will be exploited to evaluate the thesis contributions.

Chapter 3

Overview of Facial Expression Databases Used in the Thesis

This chapter concentrates upon introducing the databases used in evaluating the performance of the proposed methods. Four different types of posed and spontaneous databases have been employed:

1. Cohn-Kanade (CK) and the extended CK+ databases [99, 133].

2. MMI facial expression database [58].

3. Video Database of Moving Faces and People (VD-MFP) [59].

4. Belfast Induced Natural Emotion Database (BINED) [60].

The CK and CK+ databases include posed (deliberate) image sequences while the remainder of the databases provide spontaneous behaviour randomly collected. Each of the databases mentioned has specific characteristics and resolutions together with an increasing challenge with regard to facial expression recognition.

The organisation of the remainder of this chapter is as follows:
Section 3.1: describes the posed database characteristics. Section 3.2: explains the types of the spontaneous database characteristics, challenges, and limitations. Section 3.2: present the performance Measures criteria for the proposed FE systems. Section 3.4: summarizes the chapter.

3.1 Posed Databases

CK and CK+ Databases

Cohn-Kanade (CK), created in 2000, is a well-known Action Unit system (AUs) coded database, commonly applied in FER applications. The database contains around 97 university students from introductory psychology classes and different ethnicities. The proportion of female students, approximately 65%, was more than the male student samples. African-American students account for 15%, while Asian and Latino students account for 3% . The ages of the student volunteers range between 18-30.

Basically, 540 image sequences were selected from 97 subjects.

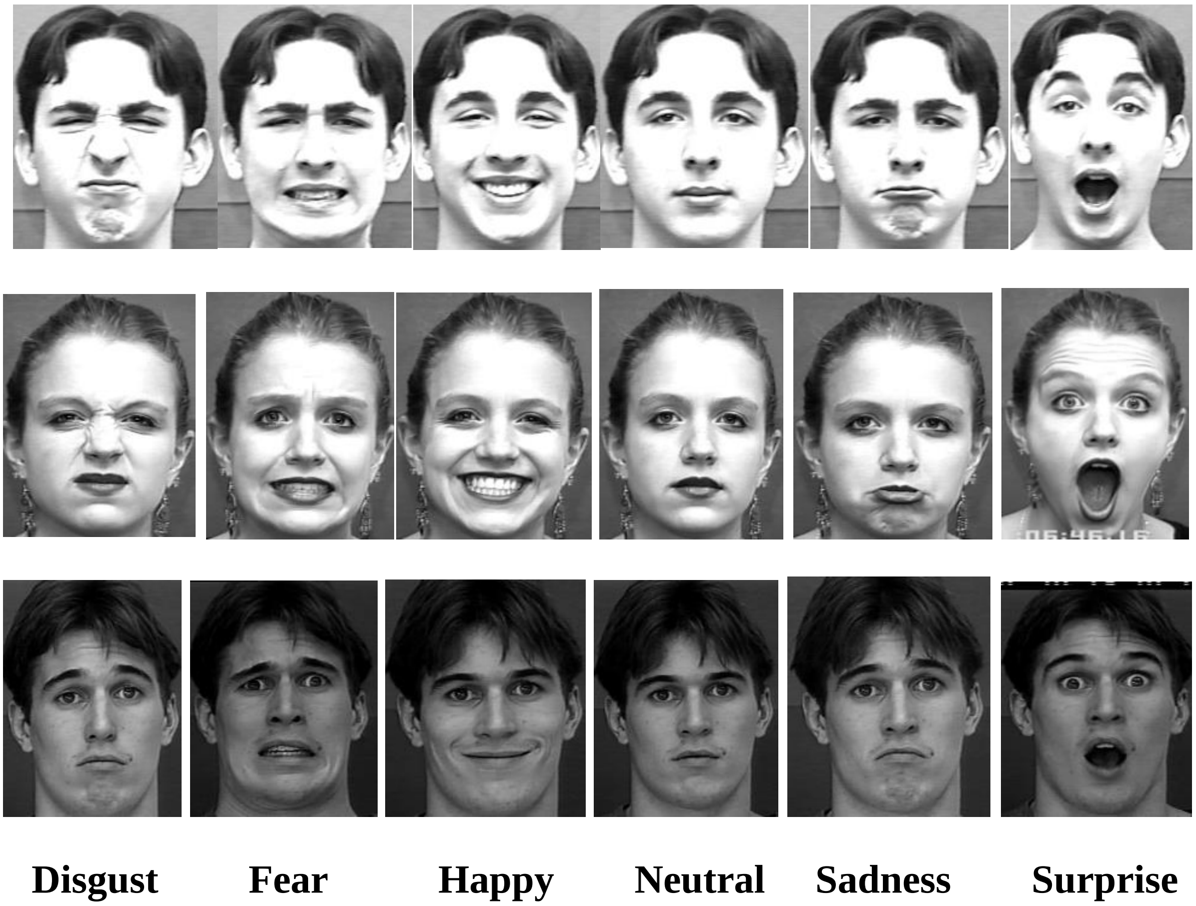


Figure 3.1: Samples of prototype FE images that were collected from the CK database. Each sample represents six specific expressions: disgust, fear, happy, neutral, sadness and surprise expressions, respectively.

The monitoring room was prepared with two Panasonic WV3230 cameras, which were connected. Additionally, a Panasonic S-VHS AG-7500 video recorder with a Horita synchronized time-code generator and a chair was provided for the subject. The cameras were in different locations in the monitoring room, one directly in front of the subject, whilst the other was positioned at 30 degrees to the right of the subject. Image data were only available from the frontal camera available at this time. Subjects were instructed via an experimenter to implement a concatenation of (23) facial expressions which included single action units, for example, AU 12 or pulling the lip corners in an oblique way, in addition to collections of action units, such as AU 1+2, which represent the inner and outer brow raisers. An experimenter depicted and constituted the required display that depends on the six basic expressions, for example, happiness, surprise, anger, fear, disgust, and sadness [133]. All frontal image sequences were digitised to dimension (640 × 490) or (640 × 480) pixels each having 8 bits grey levels or 24 bits colour values. An example of the CK database is shown in Fig. 3.1 and Fig. 3.2 illustrating samples of the CK+ database.

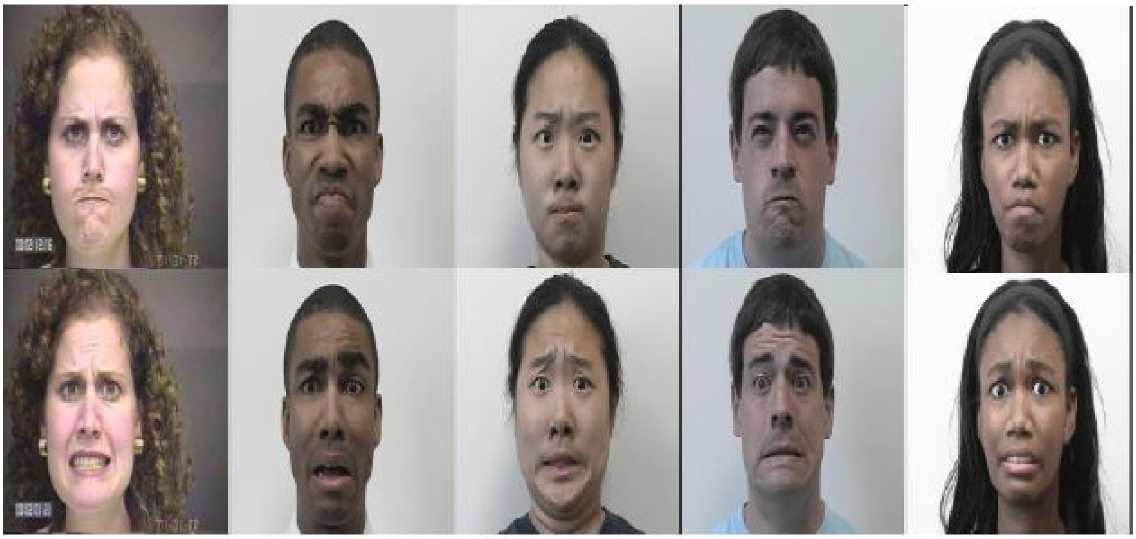


Figure 3.2: Samples of prototype FE images that were collected from the CK+ database. The upper row represents expressions of anger, whereas the lower row represents expressions of fear.

3.2 Spontaneous Databases

1. MMI Database

The MMI-Facial Expression database, which is Man-Machine Interaction (MMI) undertaken by a group at Delft University of Technology in the Netherlands was conceived by Maja Pantic, Michel Valstar and Ioannis Patras in 2002 [58], as a resource for evaluating algorithms for facial expression recognition.

Fig. 3.3 represents samples of the MMI database, which contain 335 frontal views of image sequences. They were labelled by one of the six basic prototype facial expression video sequences, which are anger, disgust, fear, happiness, sadness and surprise.

It was constructed by a series of various facial emotions, implemented by 19 different staff members and students who conduct research on faces. The database contains around 4400 females ranging in age from 19 to 62 that were selected from a European, Asian or South American ethnic background.

The subjects were requested to implement a series of up to 79 expressions consisting of either a single AU (e.g., AU2) or activations of a combination of a number of AUs (such as AU8 cannot be displayed without AU25), or a prototypic combination of AUs or performing the basic facial emotions (e.g. happiness).

All the video sequences were recorded per second at a rate of 24 frames/second using a standard PAL camera for 30 profiles-view and 750 dual-view facial expression video sequences. Dual-view images combine frontal and profile views of the face recorded using a mirror. Fig 3.4 provides an example of a profile-view, while Fig 3.5 shows an example of dual-view facial expression video sequences.

It is worth mentioning that the MMI database provides large test-bed research in the area of automated facial expression analysis.

In addition, all AU and facial emotion displays are completely annotated depending on the presence of the AUs in the sequence and partially annotated based on identifying the frame label in the temporal segments, for example, AU temporal patterns ($onset \rightarrow apex \rightarrow offset$) are certified as well.

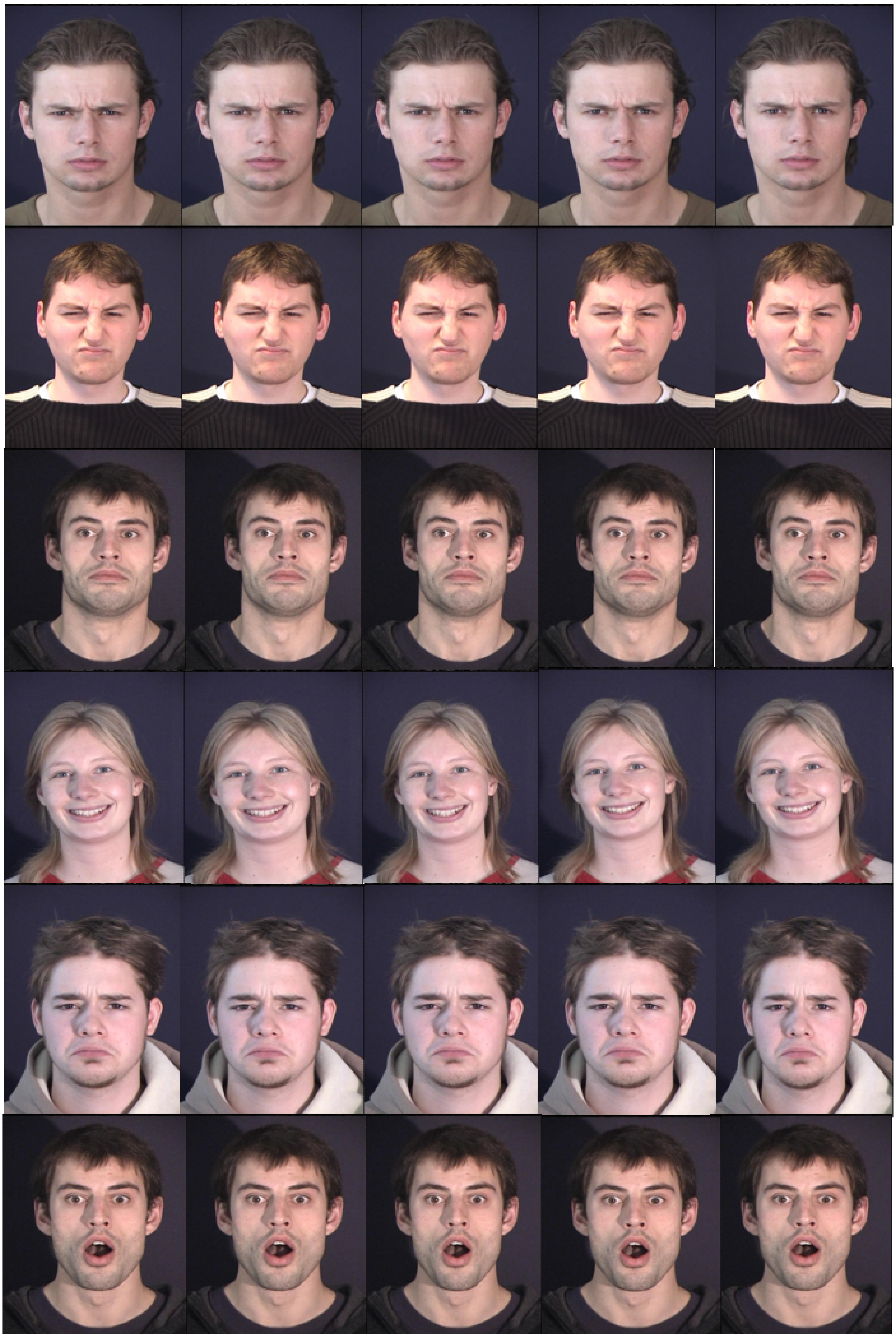


Figure 3.3: Samples of the MMI database sequences. Each row from top to bottom represents the FE sequence such as anger, disgust, fear, happiness, sadness and surprise respectively.

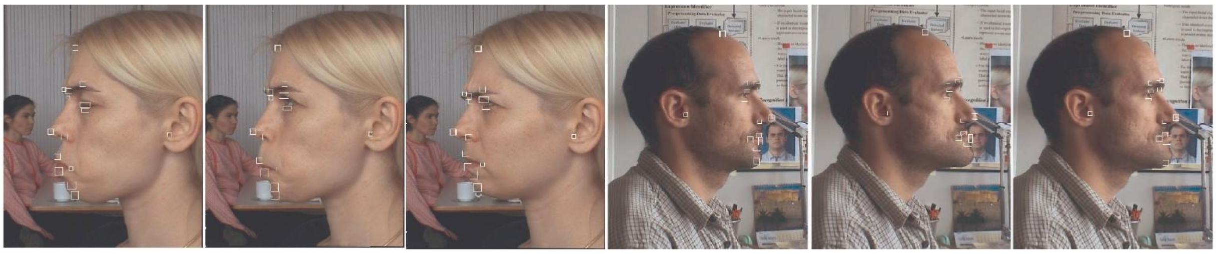


Figure 3.4: Samples of automated facial fiducial points tracking profile-face image sequences contained in the MMI Facial Expression Database [58].

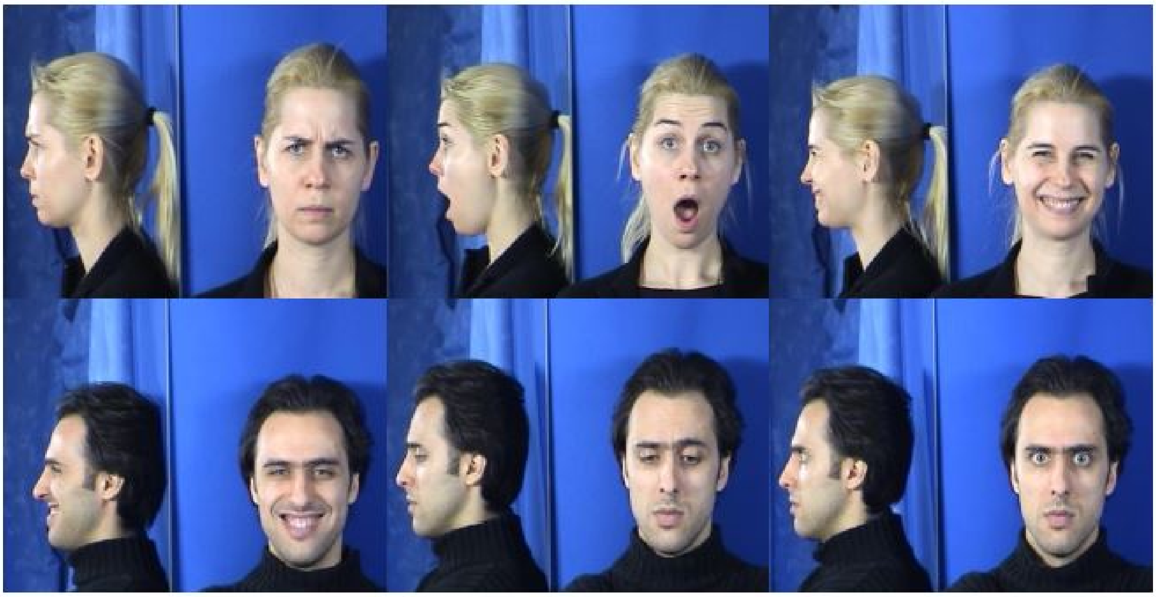


Figure 3.5: Examples of apex frames of image sequences with dual vision are represented by different expressions obtained from the MMI Facial Expression Database [58].

The MMI database has many limitations which are outlined below:

- Some of the video sequences were recorded with unwanted posed variations of the subjects.

- The database does not consider samples of partially occluded faces, such as faces covered with hands or by a long beard or moustache, except two or three videos.

- Side views of facial expressions were not included in the database (e.g., 30-degrees to the side views).

- Finally, not all the scoring of the temporal segments of AUs were labeled (only for 169 image sequences). For most sequences, only the apex frame(s) were coded.

2. Video Database of Moving Faces and People

The Moving Faces and People (VD-MFP) database was provided by the University of Texas based in Dallas, USA [59]. Fig. 3.6 displays examples of sequences of the spontaneous facial expressions of a number of subjects in the MFP database. Each subject video takes around 10 minutes. Video sequences are composed of different emotional situations that have been collected in different subtle conditions of spontaneous behaviours.

The VD-MFP database consists of 10 different groups of facial expressions, seven of these groups were described by six prototypes of facial expressions, such as anger, fear, happiness, disgust, surprise and sadness with a neutral state. The length of each video is 10 minutes.

The VD-MFP was recorded by a group consisting of 309 students from the University of Texas at Dallas as subjects. The proportion of volunteer female students was 249 about three times bigger than the male students, distributed as follows: 8 of the female subjects wear glasses 3.21%, 45 of the female subjects have multiple images 18.07% and 196 female subjects 78.71% have a single image view. There are 60 male subjects equivalent to 19.41% from the total number of subjects, distributed as follows: 55 of the male subjects have a single image view 91.66%, of the male subjects 3.33% have multiple image views and 3 wear glasses 5%. Most of the participants were Caucasians, aged between 18 - 25 years.

The MFP database has a combination of static images and video of a large number of individuals taken in a variety of contexts. A part of the database was compiled in a non-intrusive style to facilitate understanding the facial expressions under natural environments. It was collected without considering the rules of Facial Action Units that are set by Ekman and Friesen [134].

Each person has nine static shots (facial mug shots images), dynamic facial speech videos, dynamic facial expressions videos and a series of video streams. The videos consist of a moving facial mug shot, a dynamic facial speech video clip, one or more dynamic facial expression videos clips, videos of people's gaits and a conversation video taken at a moderate distance from the camera.

It is important to understand that the expressions were not categorised formally or through rigorous and strict experimental procedures. Hence, there is no ground truth regarding the video clips of facial expressions. For this reason, it is recommended that researchers carry out psychological expression steps before making claims about specific facial expressions found in the database.

The VD-MFP database has different challenges:

- The time duration of the spontaneous facial expression varies in relation to the posed facial expression. That led to challenges in determining the number of indexes for the apex frame in the video sequence.
- Various video sequences consist of several different facial expressions.
- Some video subject sequences displayed different expressions, while their content is different. For example, concerning the disgusting video sequence, the subject in the video attempts to show surprise rather than an expression of disgust.
- Particular video sequences of the subjects contain walking motions with noticeable differences as regards to moving back and forth, such as expressions of fear and surprise.

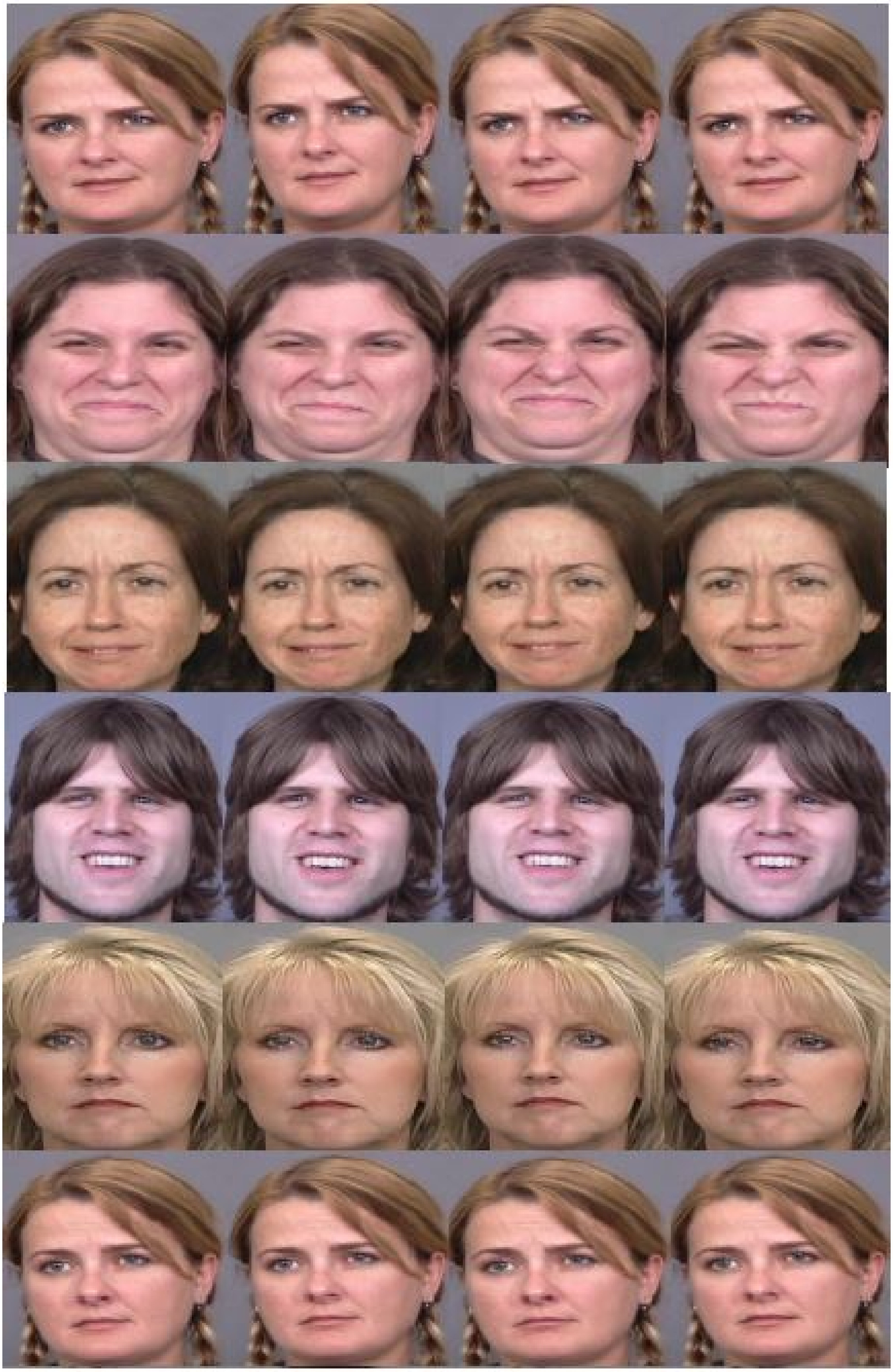
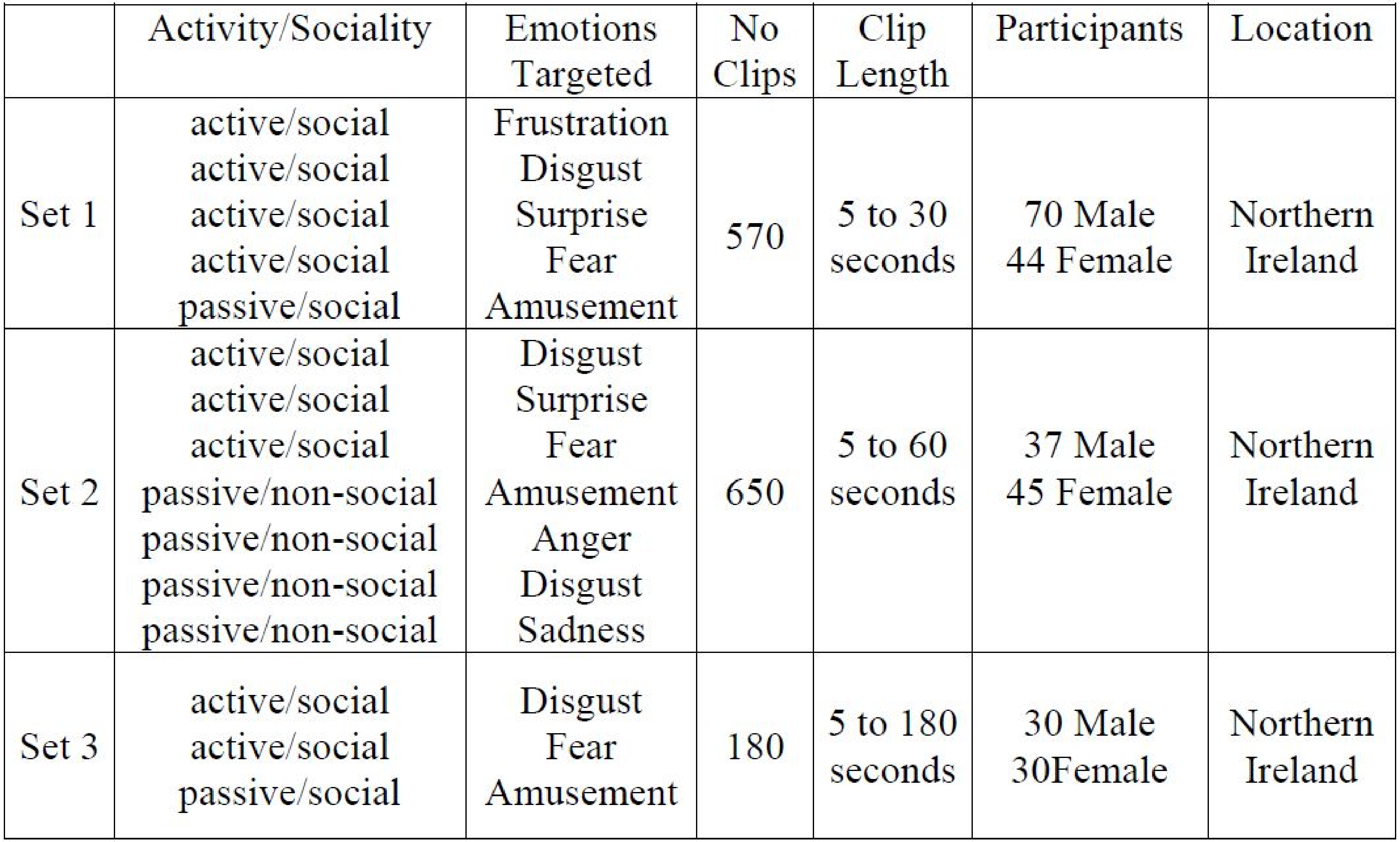


Figure 3.6: Examples of spontaneous facial expression sequences comprising 6 expressions are anger, disgust, fear, happiness, neutral and sadness, respectively [59].

Table 3.1: Summary of the Belfast Induced Natural Emotion Database Contents [60].



	Activity/Sociality	Emotions Targeted	No Clips	Clip Length	Participants	Location
Set 1	active/social active/social active/social active/social passive/social	Frustration Disgust Surprise Fear Amusement	570	5 to 30 seconds	70 Male 44 Female	Northern Ireland
Set 2	active/social active/social active/social passive/non-social passive/non-social passive/non-social passive/non-social	Disgust Surprise Fear Amusement Anger Disgust Sadness	650	5 to 60 seconds	37 Male 45 Female	Northern Ireland
Set 3	active/social active/social passive/social	Disgust Fear Amusement	180	5 to 180 seconds	30 Male 30Female	Northern Ireland

3. Belfast Induced Natural Emotion Database

The third spontaneous database, which is the Belfast Induced Natural Emotion Database, provides videos of mild to moderate emotional responses. The length of each video is short, between (5s to 60s). The effectiveness of the selected tasks is varied in inducing the expected or targeted emotion that is measured by self-reporting of the emotion, ranging from (35%) for encoders that report frustration to (92%) for the entertainment reports. However, this is predictable.

A sample of spontaneous facial expression sequences, including six expressions: anger, disgust, fear, happiness, sadness, and surprise, respectively, is shown in Fig. 3.7.

The database was divided into three broad sets depending on different chronological periods, wherein each set has different research objectives. Table 3.1 [60] presents a brief summary of the contents of the Belfast Induced Natural Emotion Database that consists of three broad sets studied from different aspects, such as activity/social, emotions targeted, no of clips, clip length, participants and location.

- **Set 1** contains 114 participants (70 males and 44 females). Most were recruited from undergraduate student encoders in Northern Ireland. Set 1 from the Belfast database consists of 570 video clips of the face and torso of both male and female encoders. The length of the video clips range from 5 to 30 seconds with a total length 237 minutes. The participants perform a series of 5 emotions inducing tasks consisting of frustration, disgust, surprise, fear and amusement. Set 1 is divided into four parts for the activity/social and one part for the passive/social.

  A Panasonic NV-GS500 digital video camera was used to record all the video clips in **Set 1**, and was placed approximately 2 metres away in front of the encoder and at a distance of 70 cm from the ground, in order to allow a view of the head and upper torso. The video was recorded on a mini DV tape cassette. Images were captured by Adobe Premiere within DV AVI Type 2 format. The screen resolution was 720 x 576 pixels.

- **Set 2** consists of 90 participants (42 males and 48 females), 69 undergraduate students, 9 postgraduate students and 12 employed professionals. All the participants are from Northern Ireland. The mean age ranges of the participants are between 23-78 years. Set 2 comprises 650 video clips recorded of the face and torso of each of the participants. A total length of 237 minutes, with the length of the video clips ranging from 5 to 60 seconds. Set

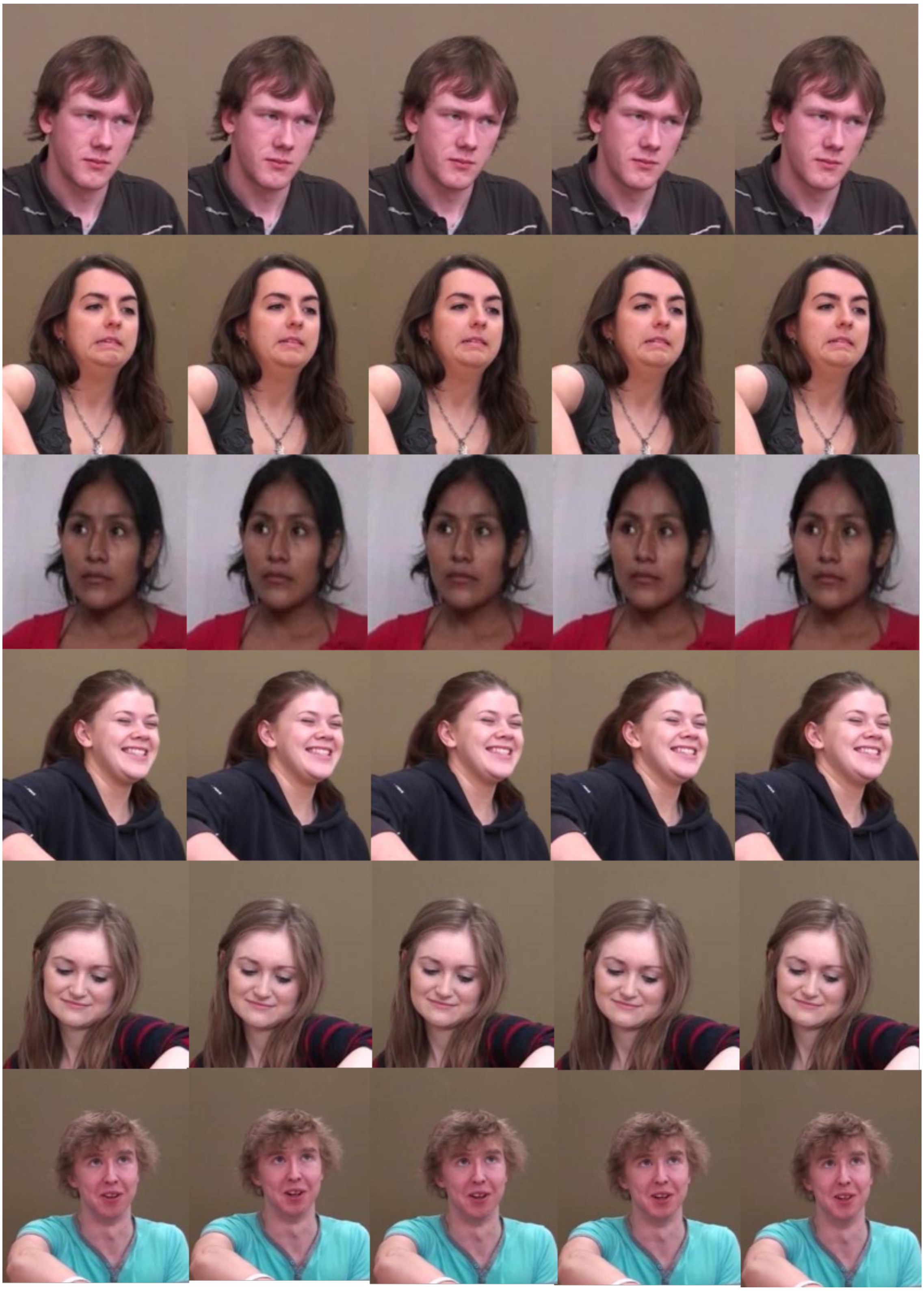


Figure 3.7: Example of spontaneous facial expression sequences with 6 expressions: anger, disgust, fear, happiness, sadness and surprise, respectively [60].

2 is divided into three parts in relation to active emotion, with emotions inducing targeted fear, disgust and surprise, and four parts with regards to passive emotion consisting of anger, disgust, sadness and amusement.

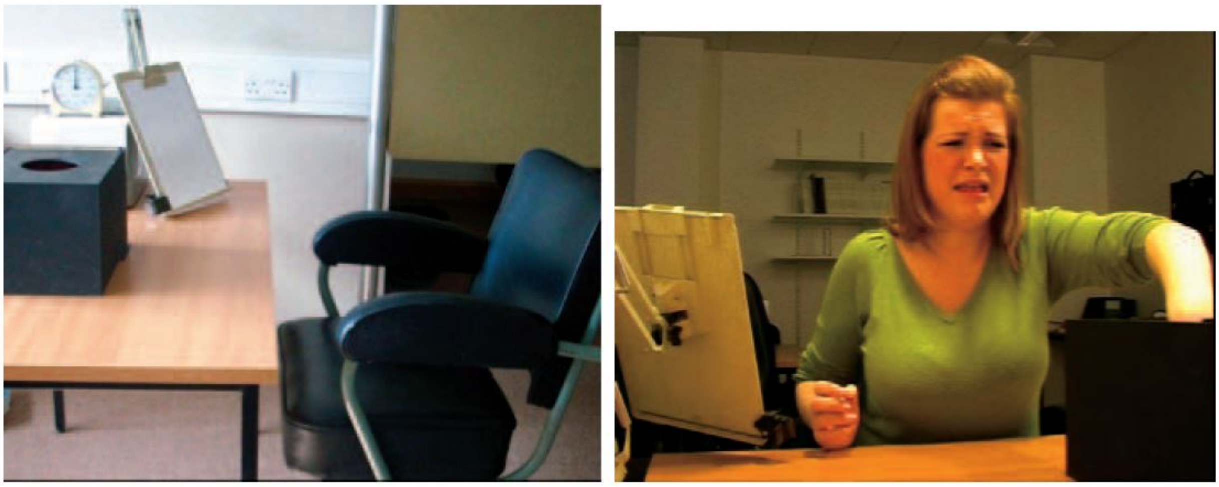


Figure 3.8: The Set 1 disgust task [60].

Table 3.2: Self-reported levels to consist of the target emotion and other additional elicited emotions.

Task	Target Emotion	Male	Female	Total	Other emotion elicited
Spaghetti	Disgust	3.55	4.23	4.19	Interest 5.62 Surprise 5.57 Amusement 5.02
Alarm	Surprise	6.74	6.91	7.60	
Imaginary Spider	Fear	3.45	4.92	4.60	Anxiety 5.15
Neutral	Relaxed	5.62	6.15	6.48	
Disgust Video	Disgust	4.40	6.00	5.76	
Sad Video	Sadness	4.76	6.52	6.23	
Anger Video	Anger	5.02	6.48	6.41	Sadness 5.49 Disgust 6.53
Amusing Video	Amusement	5.79	6.56	6.87	Relaxation 5.82 Happiness 6.58

The recorded video clips in **Set 2** contain a different mix to **Set 1**. The tasks used in Set 2 used active induction techniques to obtain fear, disgust and surprise, and passive induction by using viewing for a film to obtain anger, disgust, sadness, and amusement. In this set, the remaining recording details and the general procedures were as explained in Set 1.

- **Set 3** consists of 60 participants divided equally between males and females and collected from two countries. As a result, there are 30 participants from Northern Ireland and 30 participants from Peru, to represent different cultures. The participants from Peru have various occupations ranging from students and factory workers, to domestic and agricultural workers, and the average age ranges between 32-54 years. The Northern Ireland sample is the same as those mentioned in Set 1. This set consists of 180 video clip recordings that contain the faces and torsos of male (n=15) and female (n=15) encoders from Northern Ireland and males (n=15) and females (n=15) from Peru. The length of the video clips ranges from 5 to 180 seconds with a total length 90 minutes. The participants perform a series of 5 emotions inducing tasks consisting of disgust, fear and amusement. Set 3 was divided into two parts: the activity/socialite and passive/social. The video recording in Set 3 was based on modified versions of three tasks from Set 1; specifically, disgust, fear and amusement.

  A Sony HDR-CX105E camera was used to record video clips in **Set 3**, which was placed approximately 2 metres directly away in front of the encoder and at a distance of 70 cm from the ground, to allow a view of the head and upper torso. The video was recorded in High Definition Full HD AVCHD. The H.264/MPEG-4 (x264) codec was then used in order to compress the video, which was compressed by 7,949 kbit/s (resolution 1920 x 1080 pixels).

The Belfast Induced Natural Emotion Database has a different set of challenges:

- During the eliciting procedures in this database and despite the situations being fixed, the encoders are free to respond as they see fit.

- In laboratory-based experimental conditions, the extent of the individual's interaction is entirely spontaneous and open to question.

- The extent to which the encoder was led to believe that the principal focus of the research was on something other than the facial expression of emotion varied from task to task.

Figure 3.9: Samples of Set 3 presents fear and amusement [60].

- Although there are various approaches investigating the degree of conscious control that encoders show in their behavioural responses in the different tasks, there may be no way to confirm that in lab conditions this action reflects the natural environment exactly.

3.3 Performance Measures

To measure the performance of the system, two basic points should be considered: a suitable database; then a set of statistical measures based on the confusion matrix (CM) should be employed. For the databases, as mentioned above, four databases have been employed in this work: one is a posed database CK and the extended CK+ databases [99]. The remainder of the databases which are spontaneous, namely (MMI [58], VD-MFP [59], and BINED [60]), each become progressively more challenging.

A total of 2160 different facial expression images were employed from all the databases and distributed evenly over the four types of databases mentioned above, where each person contributed ten images.

Regarding the Performance Measures for the confusion matrix which is formed, the four outcomes yielded as a result of binary classification with a two-class problem, 2×2 CM, can be represented by counting the number of the four outcomes (predictions) for the class label with the actual class in P or N which represents a Positive or Negative label respectively, based on the use of a binary classifier.

Typically, a binary classifier predicts with all samples of a test dataset as to either positive or negative. This prediction or classification results in four different types of outcomes – true positive, false positive, true negative, and false negative [61], as shown in Fig. 3.10.

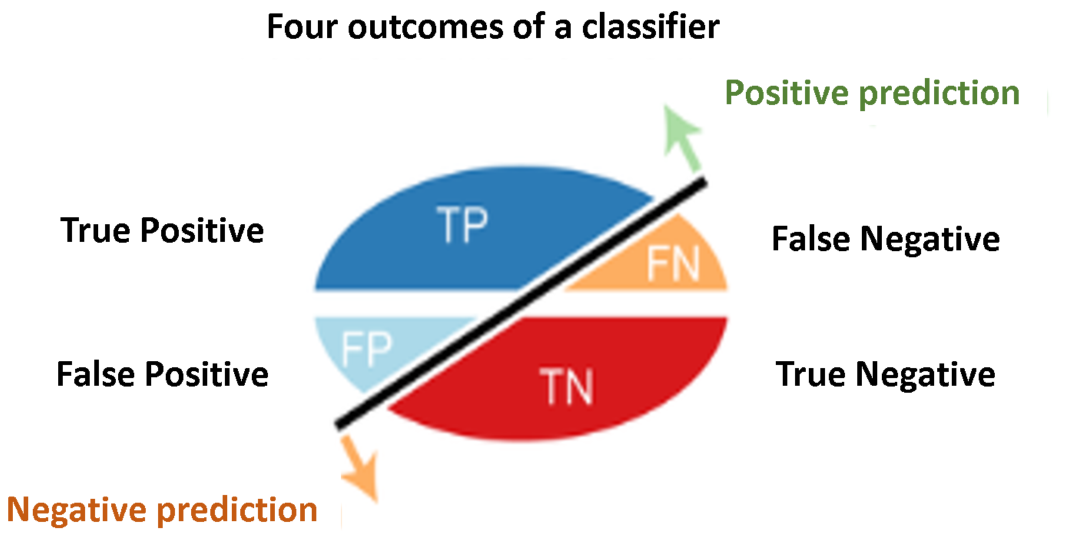
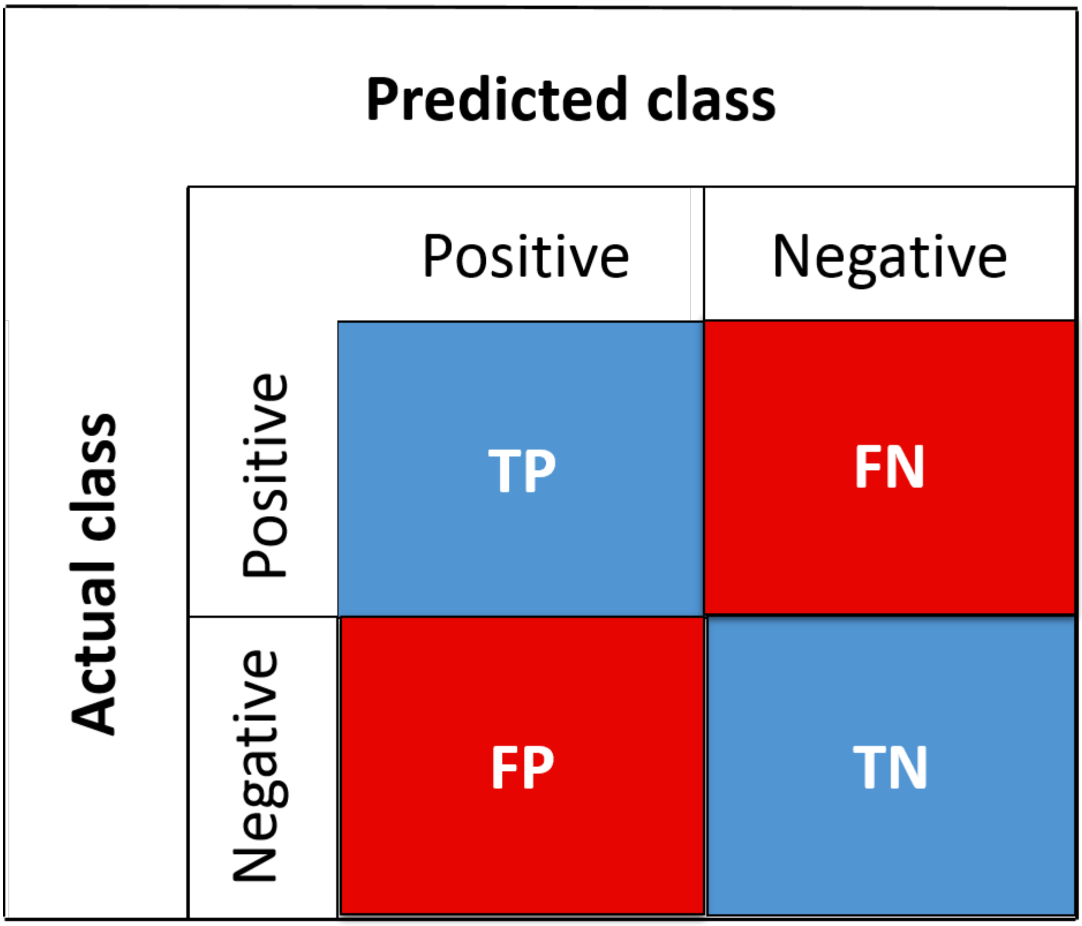


Figure 3.10: Classification or prediction of a test dataset produces four outcomes TP, FP, TN, and FN [61].

- **True Positive (TP)**: if sample P is classified as P or True in class P (correct positive prediction).
- **False Positive (FP)**: if the sample N is classified as P (incorrect positive prediction).
- **True Negative (TN)**: if sample N is classified as N in class N (correct negative prediction).
- **False Negative (FN)**: if sample P is classified as N in class P (incorrect negative prediction).

Table 3.3 shows an example of 2 × 2 of the CM, which is employed for the Performance Measures of a binary classifier.

Table 3.3: Example of the confusion matrix represents the outcome (prediction) for the class label of the actual class. Blue indicates the correct predictions, whilst red denotes the incorrect predictions.



		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

There are different measures that can be derived from a confusion matrix.

- **Error Rate (ERR)** is calculated by dividing the sum of the number of incorrect predictions over the total number of samples in the dataset [61]. For example, the best rate of error is (0.0); while, the worst is (1.0).

$$ERR = \frac{FP + FN}{TP + TN + FN + FP} = \frac{FP + FN}{P + N} \tag{3.1}$$

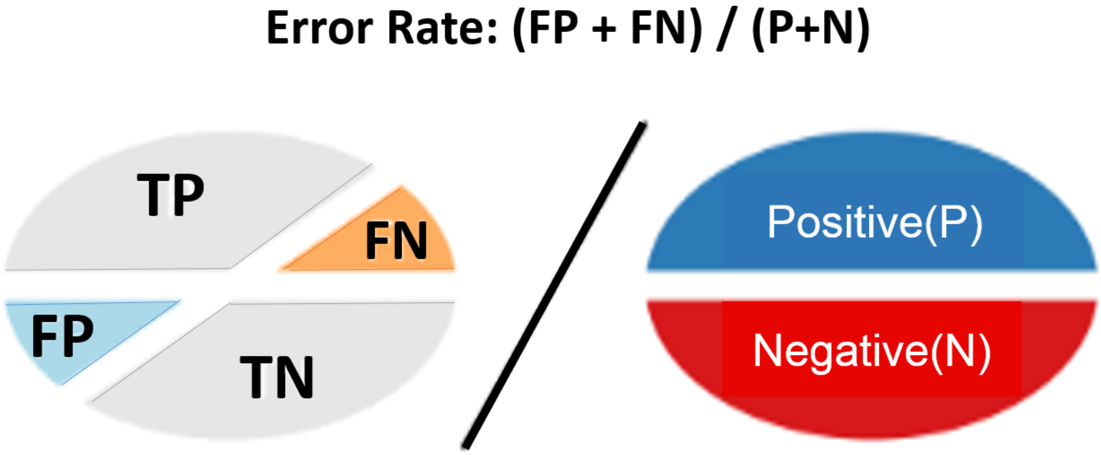


Figure 3.11: Error Rate is computed by dividing the summation of (FN + FP) over the total number of a dataset (P + N).

- **Accuracy (ACC)** is calculated by dividing the summation of the number of correct predictions over the total number of a dataset. For instance, the best accuracy to be at 1.0, whereas the worst to be at 0.0. It can also be calculated by (1 – ERR).

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} = \frac{TP + TN}{P + N} \tag{3.2}$$

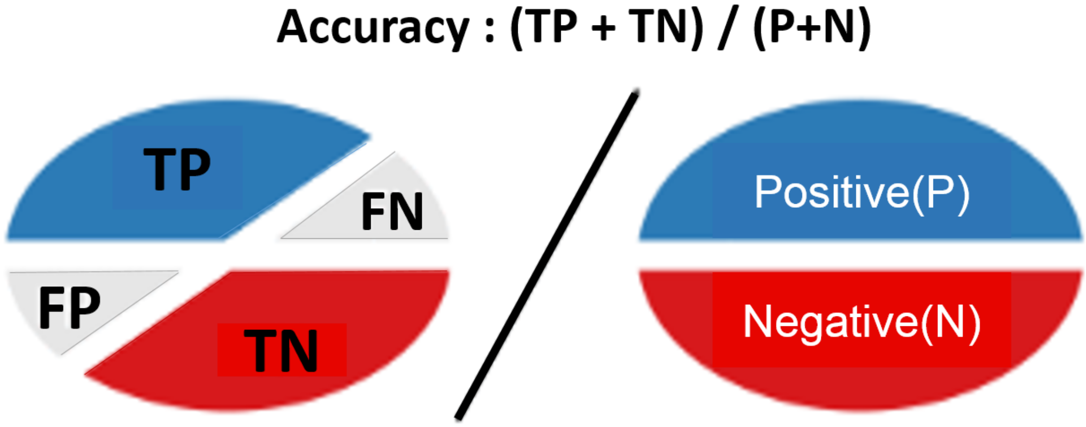


Figure 3.12: Accuracy is calculated by dividing the summation of (TP + TN) over the total number of a dataset (P + N) [61].

- **Sensitivity (Recall or True positive rate)**:

  Sensitivity (SN) is calculated by dividing the number of correct positive predictions over the total number of positives. It is also known as the True Positive Rate (TPR) or Recall (REC). The best result for sensitivity is (1.0), while the worst is (0.0).

$$SN = \frac{TP}{TP + FN} \tag{3.3}$$

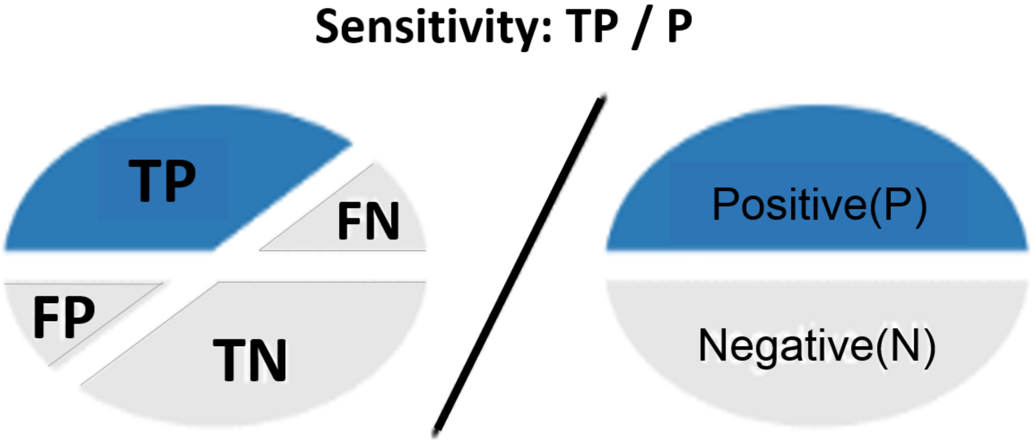


Figure 3.13: Sensitivity is calculated by dividing (TP) over the total number of positives (P) [61].

- **Specificity (True negative rate)**:

  Specificity (SP) also termed True Negative Rate (TNR) is calculated by dividing the number of correct negative predictions over the total number of negatives. The best result obtained with regard to specificity is 1.0, whereas the worst to be 0.0.

$$SP = \frac{TN}{TN + FP} \tag{3.4}$$

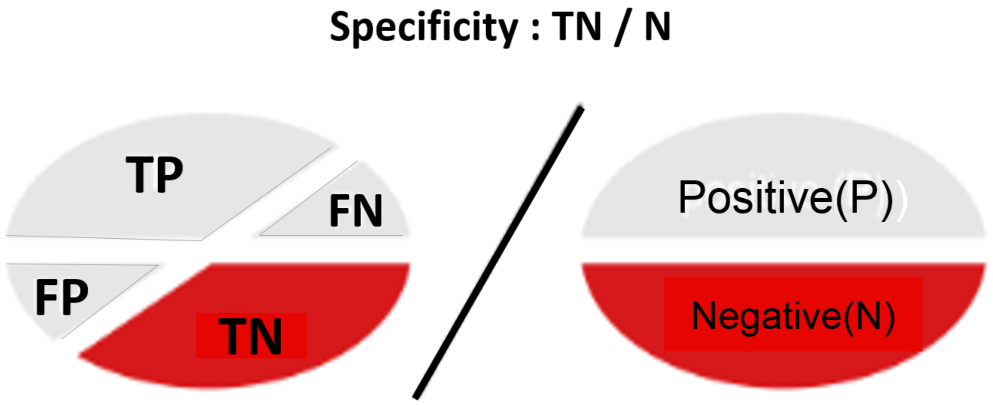


Figure 3.14: Specificity is calculated by dividing (TN) over the total number of negatives (N) [61].

- **Precision (Positive predictive value)**:

  Precision (PREC) also termed Positive Predictive Value (PPV) is calculated by dividing the number of correct positive predictions over the total number of positive predictions. The best result achieved in relation to precision is 1.0, while the worst to be 0.0.

$$PREC = \frac{TP}{TP + FP} \tag{3.5}$$

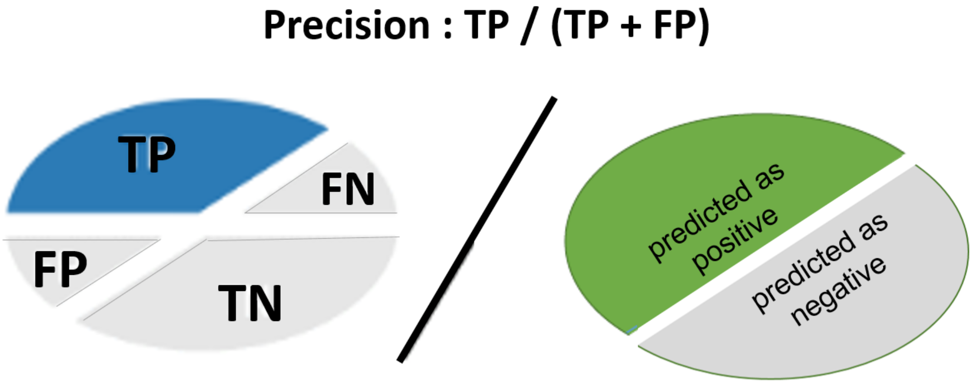


Figure 3.15: Precision is calculated by dividing (TP) over the total predictions (TP+FP) [61].

- **False Positive Rate (FPR)** is calculated by dividing the number of incorrect positive predictions over the total number of negatives. The best result for the false positive rate is 0.0, whereas the worst to be 1.0. It can also be calculated by (1 – specificity).

$$FPR = \frac{FP}{TN + FP} = 1 - SP \tag{3.6}$$

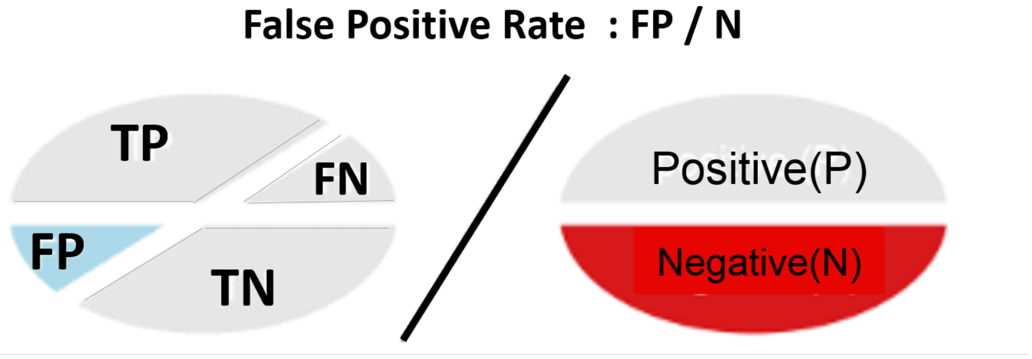


Figure 3.16: False Positive Rate is calculated by dividing (FP) over the total number of negatives (N) [61].

3.4 Summary

In this chapter, all the databases that were adopted to evaluate the performance of the proposed work have been reviewed and summarised.

In addition, several performance measures that were exploited to evaluate the performance of the proposed system were discussed.

The following three chapters will present the contributions of the thesis by employing the FER performance measurements with the databases which are mentioned in this chapter.

Table 3.4: An Overview of the Effect of Facial Expression Recognition Datasets.

Dataset	Accessi -bility	Basic emotions	Type	AU Coded	Subjects	Videos	Images	Frame by frame labels	Results	Regist -ration	Repr -esen
CK [133]	Yes	6+N	Posed	Apex frame only	97	486	-	-	-	-	-
CK+ [99]	Yes	6+N	Dynamic &Posed	Apex frame only	123	593	-	-	✓	✓	-
MMI [58]	Yes	6+N	Dynamic & Posed & Spontaneous	Apex frame only	75	2420	484	temp. phas.	-	-	-
VD-MFP[59]	Yes	6	Dynamic & Spontaneous	No	309	n/a	2016	-	-	-	-
BINED [60]	Yes	6+N	Dynamic & Spontaneous	No	125	298	-	✓	-	-	-

Chapter 4

Identity-Independent Expression Recognition

4.1 Introduction

In computer vision and human-computer interaction, there is growing interest in automatic Facial Expression Recognition (FER) [135]. Facial expression is generally the most accurate channel that communicates human emotions, among the various channels such as the textual content of voice, gestures and facial expressions [2, 136]. Thus, automatic systems for facial emotion recognition may be effective in understanding the activity of a person's emotion. There have been various research studies which have investigated emotion analysis based on posed and spontaneous databases and analyzed the facial characteristics via a motion extraction method or a feature extraction method [135], [5], [124] and [62]. However, many problems need to be solved according to the type of databases, namely posed or spontaneous databases and in terms of head-pose variations, illumination conditions, occlusions and nature of expressions [137]. Much recent research in the literature has used the differences in the appearance of a face in a single image and performed evaluation in a variety of facial expressions by testing different people [135].

In this chapter, approaches are proposed to obtain the best average accuracy in facial expression recognition by:

- Study of a form of identity-independent expression recognition problem by using difference images in order for creation of the decomposition of expressive images.

- Comparison of the performance of such image-based expression recognition techniques on posed and three progressively more challenging spontaneous databases, thereby exploiting the principles of sparse representation theory [138].

The organization of the remainder of the chapter is as follows: Section 4.2 illustrates the procedure of facial expression recognition. Section 4.3, explains the extraction of difference images. Section 4.4, provides the texture feature extraction: Section 4.5, explains the sparse representation classifier. Section 4.6, presents the evaluation of results and discussions. Finally, Section 4.7, summarizes the chapter.

4.2 The procedure of the facial expression recognition

Various methods have been proposed to enhance the FER system performance to solve the problems of the complexity of facial expressions and thereby achieve higher accuracy of facial expression recognition. The first problem that is addressed is to study a form of identity-independent expression recognition using the difference images with the intensity values of the neutral image from the corresponding facial expression image in order to create the decomposition of expressive images.

4.2.1 Pre-processing Stage

Pre-processing in the analysis of facial expression plays a significant role to improve the rate of facial expression recognition. The major target of the pre-processing is to remove the irrelevant information and maintain the reality and salutary information in the images; also, to improve the detection ability and simplify data information maximally in order to enhance the feature extraction reliability, image segmentation, recognition, and matching. Ordinarily, pre-processing steps on images involve several operations, such as image scaling, illumination effects as contrast adjustment, image brightness, histogram equalization, detecting and other processes of image enhancement. Therefore, the illumination normalization and histogram equalization have been applied to each image before extracting difference Images. An example is illustrated in Fig. 4.1, step (a) represents the original images as in CK+ database. Then, step (b) represents a preliminary pre-processing step in this work namely facial detection; wherein, it is necessary to detect the facial regions from an image sequence by using one of the most popular methods which is the so-called Viola–Jones face detector. This method has a good performance in terms of speed and accuracy compared with other algorithms [139]. Then, the apex image of each sequence is selected for further processing. After that, all frames are converted to grayscale. Next, step (c) shows the distribution of the image density by using the histeq function. Finally, in step (d), the desired facial area is extracted, and the background is removed to obtain better recognition; while the areas of the upper part of the forehead, ears, and below the chin are cropped automatically. Then, unification of all sizes of images to one size was achieved.

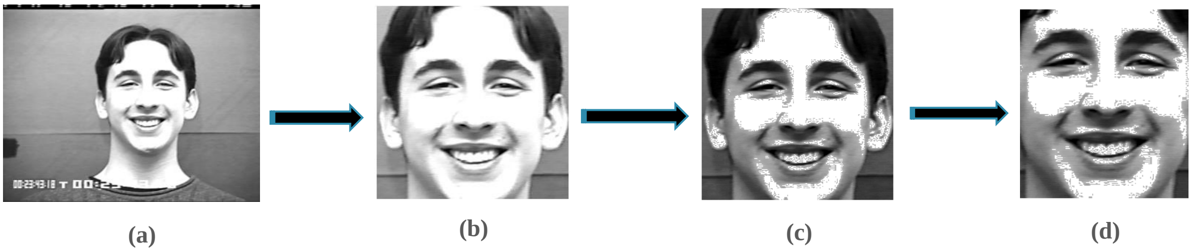


Figure 4.1: Pre-processing for image: (a) Image is original as in the CK+ database; (b) Processed face detection; (c) Distribution of image density by using histeq function; (d) Image after cropping the image for excess parts by using the imcrop function.

4.3 Extracting Difference Images

Let $\mathbf{E}_{(K,I)}$ represent the training images of the subjects with different expressions, where $I = 1, 2, ..., 6$ denotes the six basic expressions and K is the subjects index $= 1, 2, ..., S$, where S is a total number of subjects. $\mathbf{N}_{(K,I)}$ represent the images with neutral expressions. Afterwards, the difference image $\mathbf{D}_{(K,I)}$ can be calculated by subtracting the neutral image intensity values from the corresponding intensity values of the full facial expression as shown in Eq. 4.1.

$$\mathbf{D}_{(K,I)} = \mathbf{E}_{(K,I)} - \mathbf{N}_{(K,I)} \tag{4.1}$$

The process of constructing the difference image is illustrated in Fig. 4.2. The difference image should remove the identity of the subject in the facial image [62], and provides a person-independent representations for the facial expressions. Fig 4.2 shows the difference images (c) calculated by subtracting the intensity values of the neutral image (b) from the corresponding facial expressive images (a) in order to create the decomposition of expressive images. Different methods to analyse this representation are discussed in the following sections.

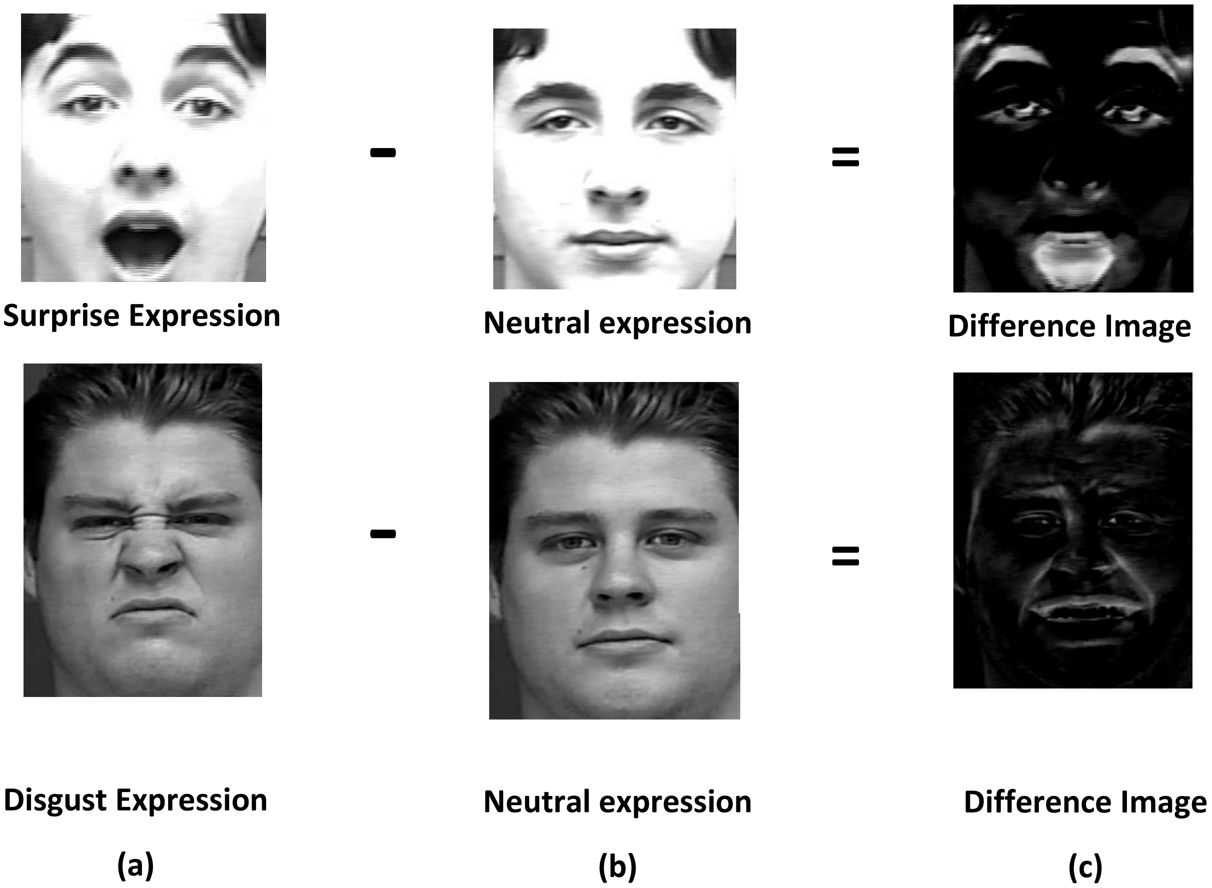


Figure 4.2: A sample of images for the differences images is represented in (c) that correspond to the expressive images in (a) and (b) [62].

4.4 Texture Feature Extraction Method

Increasing interest has been focused upon feature extraction which is able to recognize facial expression automatically; for instance, one approach is to extract the component features from the facial expression. This requires movement of the eyes, and the mouth, to be extracted which are the major facial features of expression [140]. In the Principal Component Analysis (PCA) method, the covariance matrix is found to calculate the eigenvalues and eigenvectors, in order to obtain the eigenfaces which account for the most variance within the set of face images through using only the 'best' eigenfaces that correspond to the largest eigenvalues. The highest variance reflects the eigenvector associated with the highest eigenvalue [141], while the eigenvector associated with the smallest eigenvalue finds the least variance. In the Fisher Linear Discriminant Analysis (FLDA) method, eigenvalues and eigenvectors need to be calculated, after that the within-class scatter and between-class scatter are calculated, the largest value of the eigenvalue and eigenvector are selected to build the 'best' Fisherface, as in the PCA method.

4.5 Sparse Representation Classifier

A facial expression recognition system works by solving the problem of determining to which class the test sample belongs with the help of training samples [138]. The Sparse Representation-based Classification (SRC) algorithm concept for facial expression recognition is consistent with the structure of the problem and is described next.

In the SRC the columns of matrix $\mathbf{A}_i$ are the training set of the subject.

$$\mathbf{A}_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, ..., \mathbf{v}_{i,n_i}] \in \mathbb{R}^{m \times n_i}, \tag{4.2}$$

where vector $\mathbf{v} \in \mathbb{R}^m$. Each new test image $\mathbf{y} \in \mathbb{R}^m$ belongs to one of the classes in the training set. Hence, for expression i, we can calculate $\mathbf{y}$ as:

$$\mathbf{y} = \alpha_{i,1}\mathbf{v}_{i,1} + \alpha_{i,2}\mathbf{v}_{i,2} + ... + \alpha_{i,n_i}\mathbf{v}_{i,n_i}, \tag{4.3}$$

where: $\alpha_{i,j} \in \mathbb{R}, j = 1, ..., n_i$.

Initially, the expression i of the test sample is unknown, therefore, a new matrix $\mathbf{A}$ is defined for the training set.

$$\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_k] = [\mathbf{v}_{1,1}, \mathbf{v}_{1,2}, ..., \mathbf{v}_{k,n_k}]\,. \tag{4.4}$$

where k is the number of object classes. Thus, the linear representation of the test sample $\mathbf{y}$ can be rewritten in terms of each training image as:

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 \in \mathbb{R}^m, \tag{4.5}$$

where: $\mathbf{x}_0 = [0, ..., 0, \alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,n_i}, 0, ..., 0]^T \in \mathbb{R}^n$ is a coefficient vector whose inputs are zero except concerning those correlated by the i_{th} class. The vector entries of $\mathbf{x}_0$ encodes the identity of the test image $\mathbf{y}$ and attempts to obtain it through the linear system solution of equations:

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{4.6}$$

Clearly, if $m > n$, the system of equations $\mathbf{y} = \mathbf{A}\mathbf{x}$ becomes overdetermined, and a non-sparse solution for $\mathbf{x}_0$ can be found. However, in the problem of robust facial expression recognition, usually, the system of equations $\mathbf{y} = \mathbf{A}\mathbf{x}$ becomes underdetermined, and thus, the solution is not unique. Traditionally, the solution is found by choosing the minimum l^2-norm solution:

$$l^2 : \quad \hat{\mathbf{x}}_2 = \text{argmin} \left\|\mathbf{x}\right\|_2 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{4.7}$$

A correct test image $\mathbf{y}$ should only be represented by using the training images that belong to the same class. Obviously, the more sparse the recovered $\mathbf{x}_0$, the easier it is to determine the identity of the test images $\mathbf{y}$ accurately. This is a motivation to search for the sparsest solution to $\mathbf{y} = \mathbf{A}\mathbf{x}$, i.e. to solve the following optimization problem:

$$l^0 : \quad \hat{\mathbf{x}}_0 = \text{argmin} \left\|\mathbf{x}\right\|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{4.8}$$

where the l^0-norm is indicated by $\left\|.\right\|_0$, which computes the number of nonzero coefficients entries in the vector $\mathbf{x}$. In the presence of noise the equality constraint can be replaced by a bound an the error $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|$. In practice, the l^0-norm is approximated by the l^1-norm [138].

4.6 Results and Discussions

Evaluation results are presented in the form of confusion matrices (CM) which demonstrate the results of original and difference images. Experiments are performed on both databases: the posed (CK+), and for the first time with three types of spontaneous databases that are represented by the MMI Facial Expression Database, Video Database of Moving Faces and People (VDMFP), and Belfast Induced Natural Emotion Database. The aforementioned databases are proposed to cover various imaging conditions each having different challenges as mentioned in Chapter 3.

4.6.1 Experimental Setup

The simulations were carried out by using Matlab (version R2015a) on a PC with Intel core i5-4690 CPU @ 3.50 GHz processor and 16.0 GB RAM. The subjects have been randomly selected from the posed and spontaneous databases.

Each image was cropped automatically around the face using a *facedetection* function while the areas of the upper part of the forehead, ears, and below the chin have been cropped manually using the *imcrop* function. Thereafter, all images were converted to grayscale, and unification of all sizes of images to one size was achieved by using the *imresize* function. Afterwards, the difference images have been calculated by subtracting the original image from the corresponding facial expression image and each negative pixel value was converted to zero as in the Fig.4.2. Then, using statistical methods to extract features and reduce dimensions namely: Principal Component Analysis (PCA) and Fisher Linear Discriminant Analysis (FLDA). The Sparse Representation Classifier (SRC) is also used for classification. Leave-one-out strategy has been used for cross-validation.

4.6.2 Experimental Results

4.6.2.1 System Evaluation of the Posed Database (CK+)

In this study, 540 was the total number of images used that include six classes of facial expressions; each class has 90 images from the CK+ database. The expressions are distributed as follows, AN, DI, FE, HA, SA and SU, while the neutral (NE) class does not include the learning set in order to trivialize the facial expression recognition problem and to increase recognition rates. A full description of recognition accuracies can be displayed by using Confusion Matrices (CMs) which consist of six classes of expressions by using the four sets of databases.

When using the PCA and FLDA combined with the SRC based-classifier by using the images of (CK+ database) supply rightly harmonious recognition averages across the six classes and an acceptable average recognition rate of the CK databases. For the experiments, a leave-one-out strategy is used for cross-validation. Wherein one expressive image is left out to be used as a test and the remaining are used as training.

Table 4.1: Confusion matrix showing evaluation results of original images by using PCA + SRC with the CK+ database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**85.57**	3.33	1.11	3.33	3.33	3.33
Disgust	3.33	**78.91**	5.55	4.44	2.22	5.55
Fear	3.33	4.44	**68.90**	2.22	4.44	16.66
Happy	2.22	1.11	4.44	**83.35**	4.44	4.44
Sadness	1.12	4.44	6.66	3.33	**78.90**	5.55
Surprise	2.22	2.22	4.44	3.33	6.66	**81.13**

- Table 4.1 shows the confusion matrix in terms of evaluation results for the original images by using PCA with SRC for posed CK+ database. Average recognition accuracy of (79.45%) is obtained with the highest recognition accuracy of the best class being (85.57%) in case of anger and the lowest accuracy of (68.90%) with fear class. Table 4.2 displays the confusion matrix of the result of original images by using FLDA with SRC for the CK+ database. The results give an average recognition accuracy of (90.93%) with the highest recognition accuracy of (97.78%) for anger and (76.67%) in case of surprise.

Table 4.2: Confusion matrix showing evaluation results of original images by using FLDA + SRC with the CK+ database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**97.78**	2.22	0	0	0	0
Disgust	3.33	**94.45**	0	1.11	0	1.11
Fear	4.44	1.11	**85.57**	3.33	4.44	1.11
Happy	3.33	0	0	**94.45**	1.11	1.11
Sadness	2.22	1.11	0	0	**96.67**	0
Surprise	5.55	5.55	2.22	4.44	5.55	**76.67**

- Table 4.3 displays the confusion matrix evaluation results of difference images by using PCA with SRC with the posed database CK+ database. In terms of the average recognition accuracy of different images, it is (82.97%) that is including the highest recognition accuracy of (87.79%) for the best class being in the case of surprise the lowest accuracy of (76.69%) with fear class. Table 4.4 the overall recognition rate is improved by using the difference images with FLDA together with SRC for CK+ database. The results yield an average recognition accuracy of (94.63%) with the highest recognition accuracy of (97.78%) for both classes anger and disgust; and (87.79%) is the lowest accuracy in Surprise.

The best performance for the PCA and the SRC classifiers with the CK+ databases show that using the difference images, where average accuracy was equal to (82.97%) better than the original images as Tables 4.1 & 4.3, respectively. Similarly, the results of the confusion matrix when using the difference images with the FLDA & SRC under the same conditions were better than the original images as shown in Tables 4.2 & 4.4.

Table 4.3: Confusion matrix showing evaluation results of difference images by using PCA + SRC with the CK+ database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**81.13**	3.33	2.22	2.22	5.55	5.55
Disgust	1.11	**82.23**	3.33	1.11	2.22	10
Fear	3.33	2.22	**76.69**	5.55	4.44	7.77
Happy	0	4.44	4.44	**84.46**	2.22	4.44
Sadness	2.22	0	0	2.22	**85.56**	10
Surprise	0	3.33	1.11	2.22	5.55	**87.79**

Table 4.4: Confusion matrix showing evaluation results of difference images by using FLDA + SRC the CK+ database

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**97.78**	2.22	0	0	0	0
Disgust	2.22	**97.78**	0	0	0	0
Fear	2.22	0	**93.34**	3.33	1.11	0
Happy	2.22	0	3.33	**94.45**	0	0
Sadness	2.22	0	0	2.22	**96.67**	10
Surprise	3.33	2.22	0	0	1.11	**87.79**

The results of FLDA combined with the SRC classifier for different images to CK+ database outperformed the rest of results such as PCA was combined with SRC with both original and difference images and FLDA with SRC with original images, in which it provides consistent recognition rates, and a reasonable average recognition rate of the six classes.
Fig. 4.3 shows a comparison of the recognition results percentage of the class between original and difference images of the posed databases (CK+) by using different methods (PCA + SRC and FLDA + SRC).

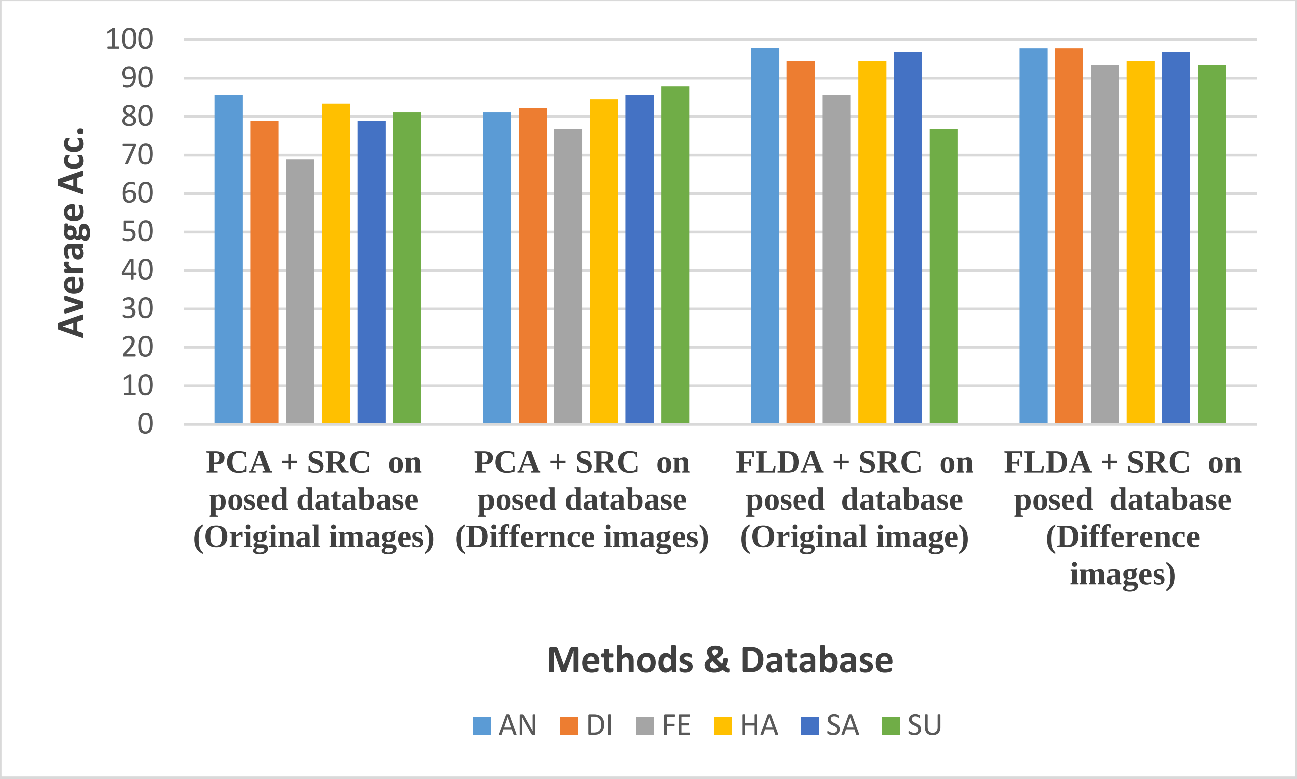


Figure 4.3: The recognition results percentage of the number/class between original and different images of the posed databases using different methods

4.6.2.2 System Evaluation of the Spontaneous database

In this study, the same total number of images in the posed database are used and are selected subjects randomly with the same six classes of facial expressions. Thus, for each expression class there are 90 images. Consequently, the selected three types of databases (MMI, VD-MFP, and Belfast) databases are categorized into 540 images for each database. The expressive images (apex images) from the video are picked to represent the facial expression. The expressions are distributed as follows, AN, DI, FE, HA, SA and SU, while the neutral (NE) class does not include the learning set in order to trivialize the facial expression recognition problem and to increase recognition rates.

Table 4.5: Confusion matrix showing evaluation results of original images by using PCA + SRC with MMI database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**66.67**	2.22	7.77	6.66	4.45	11.13
Disgust	12.24	**64.44**	5.55	10.00	2.22	5.55
Fear	2.22	4.45	**66.67**	7.77	11.13	6.66
Happy	7.77	12.23	2.23	**58.89**	8.88	11.12
Sadness	5.56	2.23	20.23	2.23	**63.33**	0.00
Surprise	7.77	2.23	4.45	7.77	7.78	**70.00**

MMI database:

- Table 4.5 represents full descriptions of the confusion matrix with average recognition rates using the original images of facial expression recognition with PCA and SRC for the MMI database. The average recognition accuracy of the original images is (65.00%)

Table 4.6: Confusion matrix showing evaluation results of original images by using FLDA + SRC with MMI database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**70.00**	12.5	6.26	0	7.77	2.22
Disgust	6.25	**74.44**	0	6.25	7.77	6.25
Fear	12.50	6.25	**57.78**	6.25	12.50	0
Happy	6.25	6.25	0	**71.11**	2.23	12.50
Sadness	6.25	0	12.50	0	**73.33**	6.25
Surprise	6.25	2.23	12.50	6.25	0	**71.11**

with highest recognition accuracy of the best class being (70.00%) in the case of surprise and the lowest accuracy of (58.89%) with the happy class.

- Table 4.6 illustrates the evaluation results of the original images by using FLDA with SRC for the MMI database in terms of the average accuracy, it is (69.63%) that is including the highest recognition accuracy of (74.44%) for the best class being in the case of disgust and (57.78%) is the lowest accuracy in fear.

- Average recognition rates with the full descriptions of confusion matrix with the proposed method for difference images are shown in Tables 4.7 and 4.8 respectively. In Table 4.7, the average accuracy using PCA with SRC is (67.59%), and highest recognition accuracy is (73.33%) with disgust and the lowest accuracy being (58.89%) for the fear class. In contrast, in Table 4.8, the average accuracy using FLDA and SRC provides (72.52%), with highest recognition accuracy of the best class being (75.45%) in disgust cases and the lowest accuracy of (68.89%) with fear class.

Tables 4.5 & 4.6 show the performance of original images in the MMI database when using PCA & SRC or FLDA & SRC which have a lower recognition accuracy (65.00%, 69.63%)

Table 4.7: Confusion matrix showing evaluation results of difference images by using PAC + SRC with MMI database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**68.89**	12.50	0	6.25	12.50	6.25
Disgust	6.25	**73.33**	0	12.50	6.25	0
Fear	5.56	0	**58.89**	0	18.87	12.50
Happy	6.25	12.50	2.23	**64.44**	0	6.25
Sadness	6.25	2.23	18.75	0	**70.00**	2.23
Surprise	2.23	1.12	18.75	6.25	0	**70.00**

Table 4.8: Confusion matrix showing evaluation results of difference images by using FLDA + SRC with MMI database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**72.22**	18.75	6.25	0	6.25	0
Disgust	6.25	**75.45**	6.25	12.50	6.25	0
Fear	12.50	0	**68.89**	0	6.25	18.75
Happy	12.50	6.25	0	**74.34**	0	12.50
Sadness	6.25	6.25	12.50	0	**74.34**	0
Surprise	6.25	0	12.50	6.25	0	**69.89**

compared to the difference image as shown in Tables 4.7 & 4.8 are (67.59% and 72.52%). For the MMI database experiment, the robustness and stability of this work obviously appears for the emotions when using the difference images with both PCA and FLDA combined with the SRC based classifier, in order for the creation of the decomposition of expressive images. The difference images emphasize the expressive areas in the face while eliminating the irrelevant parts; in this way, the identity of the facial image is removed and the identity-independent expression recognition problem is addressed.

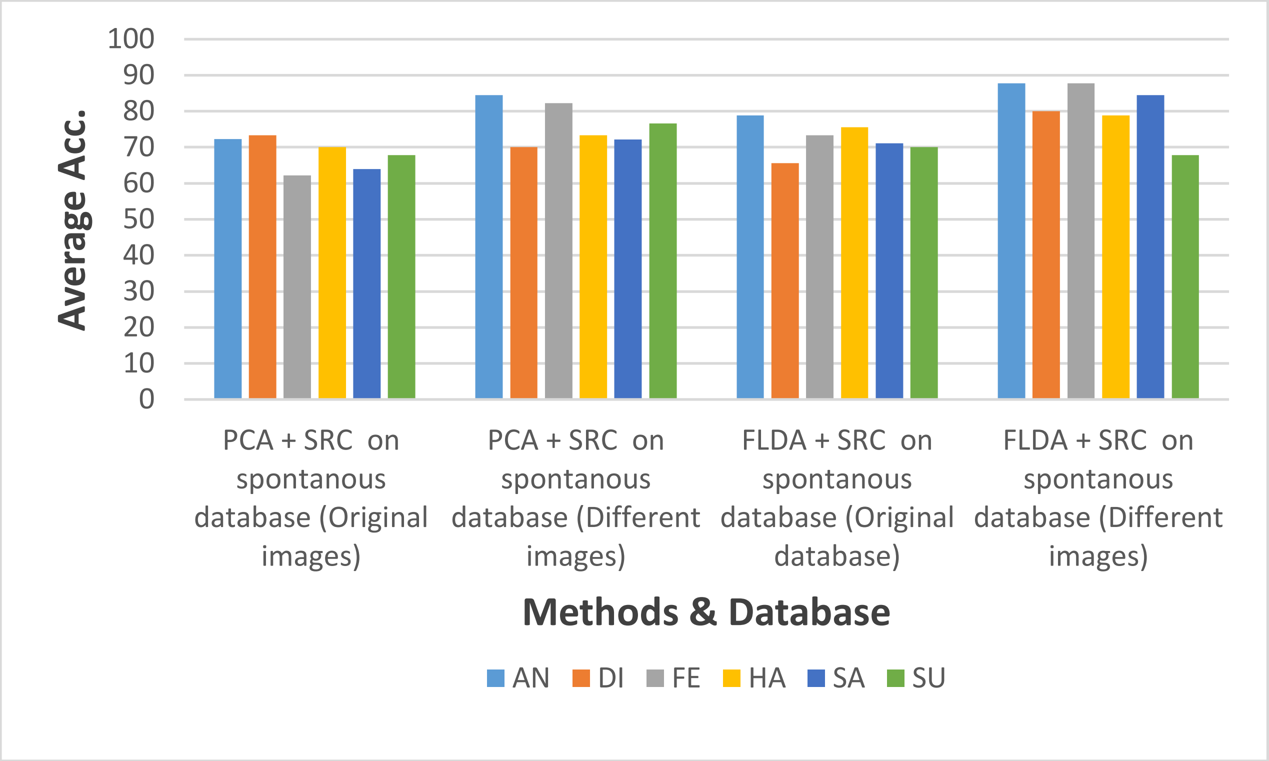


Figure 4.4: Compare the recognition results percentage of the number/class between original and different images of the MMI database using different methods (PCA + SRC and FLDA + SRC)

VD-MFP database:

- Table 4.9 represents a full description of the average accuracy for facial expression recognition of original images by using PCA with SRC with the VD-MFP database. The average recognition accuracy of the original images is (68.34%) with highest recognition accuracy of the best class being (73.35%) in the case of disgust and the lowest accuracy of (62.24%) with fear class.

- Table 4.10 illustrates the evaluation results of the original images by using FLDA with SRC for the VD-MFP database in terms of the average accuracy, it is (72.42%) that is including the highest recognition accuracy of (78.91%) for the best class being in the case of angry and (65.65%) is the lowest accuracy in disgust.

Table 4.9: Confusion matrix showing evaluation results of original images by using PCA + SRC with VD-MFP database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**72.24**	5.55	2.22	4.44	10	5.55
Disgust	4.44	**73.35**	3.33	4.44	4.44	10
Fear	6.66	4.44	**62.24**	6.66	5.55	14.45
Happy	6.67	10	4.44	**70.00**	1.12	7.77
Sadness	10	12.23	5.55	2.22	**64.45**	5.55
Surprise	3.33	5.55	11.13	7.77	4.44	**67.78**

- Table 4.11 shows the confusion matrix of average accuracy by using PCA with SRC for difference images. The accuracy is (76.48%), and highest recognition accuracy is (84.45%) with angry and the lowest accuracy being (70.00%) for the disgust class. While, in Table 4.12 difference images with FLDA provide (81.11%) average accuracy, with highest recognition accuracy of the best class being (87.79%) in both the cases of anger and fear and the lowest accuracy of (67.78%) with surprise class.

Table 4.10: Confusion matrix showing evaluation results of original images by using FLDA + SRC with VD-MFP database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**78.91**	7.77	1.11	2.22	7.77	2.22
Disgust	11.13	**65.56**	4.44	5.55	8.88	4.44
Fear	8.88	1.11	**73.36**	3.33	5.55	7.77
Happy	6.66	4.44	2.22	**75.58**	7.77	3.33
Sadness	8.88	7.78	8.88	2.23	**71.12**	1.11
Surprise	4.45	6.66	8.88	2.23	7.78	**70.00**

Tables 4.11 & 4.12 display the performance of different images in the VD-MFP database when using PCA & SRC or FLDA & SRC, which have recognition accuracy (76.48%, 81.11%) respectively, better than compared to the original image as shown in Tables 4.9 & 4.12 (68.34%, 72.42%) respectively.

The results of FLDA combined with the SRC classifier for different images to VD-MFP database outperformed the rest of the results such as PCA was combined with SRC with both original and difference images and FLDA with SRC with original images, in which it provides consistent recognition rates reasonable, and a reasonable average recognition rate of the six classes.

Table 4.11: Confusion matrix showing evaluation results of difference images by using PCA + SRC with VD-MFP database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**84.45**	2.22	1.11	2.22	3.34	6.67
Disgust	1.11	**70.00**	4.44	4.44	5.55	14.46
Fear	2.22	0	**82.22**	1.11	3.33	11.12
Happy	3.34	5.55	4.44	**73.34**	5.55	7.78
Sadness	5.55	3.33	4.44	3.33	**72.23**	11.12
Surprise	8.89	1.11	5.55	5.55	2.23	**76.67**

Table 4.12: Confusion matrix showing evaluation results of difference images by using FLDA + SRC with VD-MFP database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**87.79**	3.33	1.11	1.11	3.33	3.33
Disgust	8.89	**80.00**	3.33	2.22	1.11	4.45
Fear	3.33	1.11	**87.78**	1.11	2.22	4.45
Happy	4.44	4.44	3.33	**78.90**	2.22	6.67
Sadness	5.56	3.33	3.33	3.33	**84.45**	0
Surprise	11.13	2.22	7.77	6.66	4.44	**67.78**

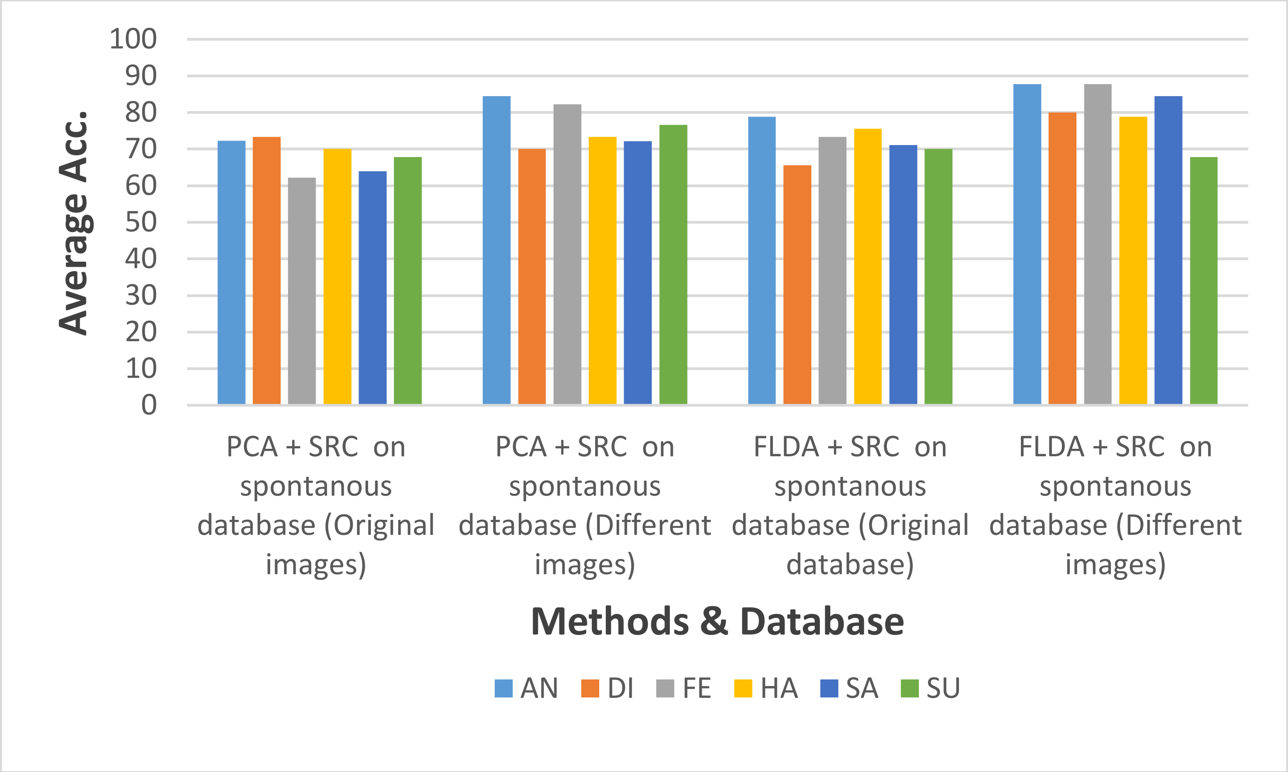


Figure 4.5: Compare the recognition results percentage of the number/class between original and different images of the database using different methods (PCA + SRC and FLDA + SRC)

Belfast database:

- Table 4.13 represents the average accuracy of facial expression recognition of original images by using PCA with SRC for the Belfast database. The average recognition accuracy of the original images is (57.59%) with highest recognition accuracy of the best class being (65.56%) in the case of surprise and the lowest accuracy of (47.78%) with anger class.

- Table 4.14 shows the results of the original images by using FLDA with SRC for the Belfast database in terms of the average accuracy, it is (64.07%) that is including the highest recognition accuracy of (73.33%) for the best class being in the case of disgust and (50.00%) is the lowest accuracy in happy.

Table 4.13: Confusion matrix showing evaluation results of original images by using PCA + SRC with Belfast database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**47.78**	3.33	1.11	1.11	3.33	3.33
Disgust	8.89	**64.44**	3.33	2.22	1.11	4.45
Fear	3.33	1.11	**60.00**	1.11	2.22	4.45
Happy	4.44	4.44	3.33	**48.89**	2.22	6.67
Sadness	5.56	3.33	3.33	3.33	**58.89**	0
Surprise	11.13	2.22	7.77	6.66	4.44	**65.56**

Table 4.14: Confusion matrix showing evaluation results of original images by using FLDA + SRC with Belfast database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**61.11**	3.33	1.11	1.11	3.33	3.33
Disgust	8.89	**73.33**	3.33	2.22	1.11	4.45
Fear	3.33	1.11	**71.11**	1.11	2.22	4.45
Happy	4.44	4.44	3.33	**50.00**	2.22	6.67
Sadness	5.56	3.33	3.33	3.33	**57.78**	0
Surprise	11.13	2.22	7.77	6.66	4.44	**71.11**

Table 4.15: Confusion matrix showing evaluation results of difference images by using PCA + SRC with Belfast database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**55.56**	3.33	1.11	1.11	3.33	3.33
Disgust	8.89	**64.44**	3.33	2.22	1.11	4.45
Fear	3.33	1.11	**58.89**	1.11	2.22	4.45
Happy	4.44	4.44	3.33	**53.33**	2.22	6.67
Sadness	5.56	3.33	3.33	3.33	**52.22**	0
Surprise	11.13	2.22	7.77	6.66	4.44	**73.33**

Table 4.16: Confusion matrix showing evaluation results of difference images by using FLDA + SRC with Belfast database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**62.22**	3.33	1.11	1.11	3.33	3.33
Disgust	8.89	**74.44**	3.33	2.22	1.11	4.45
Fear	3.33	1.11	**71.11**	1.11	2.22	4.45
Happy	4.44	4.44	3.33	**62.22**	2.22	6.67
Sadness	5.56	3.33	3.33	3.33	**58.89**	0
Surprise	11.13	2.22	7.77	6.66	4.44	**64.44**

- Table 4.15 illustrates the confusion matrix of average accuracy by using PCA with SRC with difference images in Belfast database. The accuracy is (59.62%), and highest recognition accuracy is (73.33%) with surprise and the lowest accuracy being (52.22%) for the sadness class. In contrast, in Table 4.16 difference images with FLDA produce (65.55%) average accuracy, with highest recognition accuracy of the best class being (74.44%) in disgust class and the lowest accuracy of (58.89%) with sadness class.

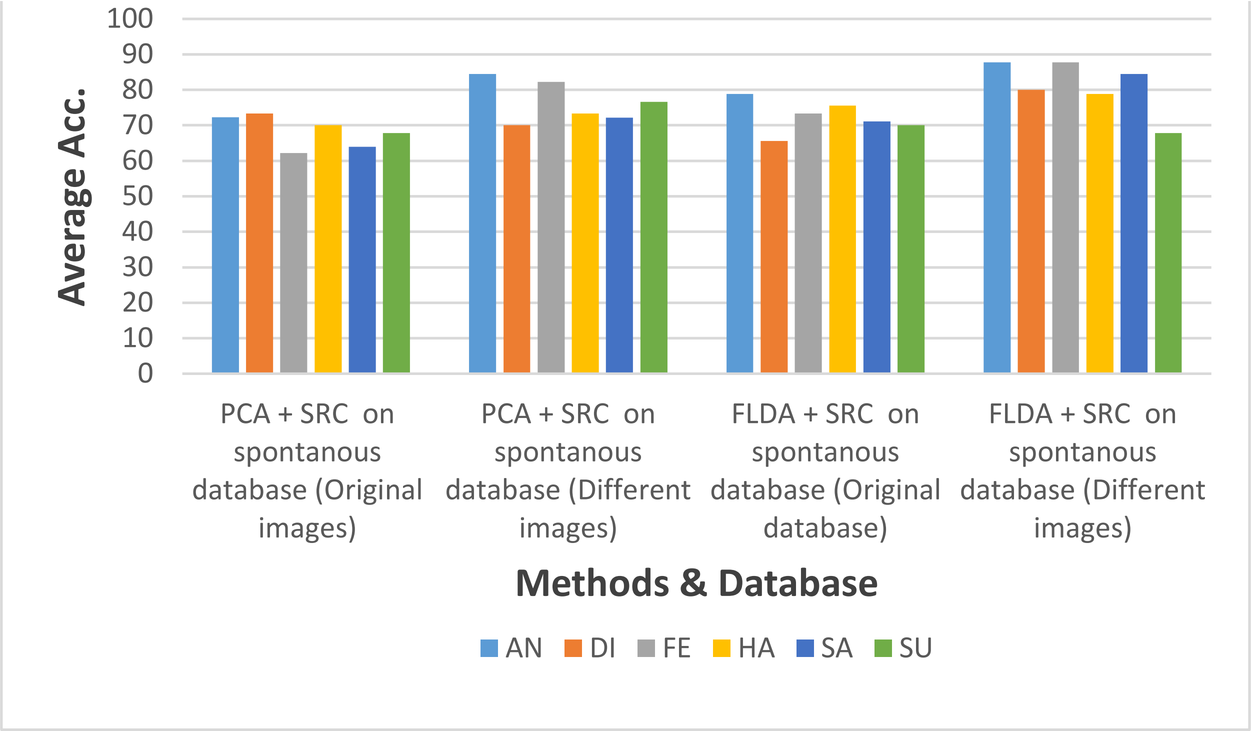


Figure 4.6: Compare the recognition results percentage of the number/class between original and different images of the spontaneous databases using different methods (PCA + SRC and FLDA + SRC)

The results of FLDA combined with the SRC classifier of different images for Belfast database exceeded the results with PCA was combined with SRC for both original and difference images, as shown in Fig. 4.6.

4.6.3 Performance Analysis

By studying the percentage of the recognition accuracy in Fig. 4.3 and Fig. 4.4, it is clear that the recognition accuracy for the different images outperformed the results of the original images. The core reason is that the difference images are better than the original image because using the difference image emphasises the expression areas in the face while eliminating the irrelevant parts.

It can be clearly seen that the best results are given when using the posed database according to the results of the average accuracy in Table 4.1 to Table 4.16 and Fig.4.7. The posed database is more convenient because each expression image is taken under controlled environment in terms of background and illumination. On the other hand, the spontaneous database is taken under more challenging conditions and the expression is less expressive due to the overlap among the expressions at a time.
Comparison of results are presented in terms of average accurecy for three types of spontaneous databaes. The Belfast Induced Natural Emotion Database is more challenging than the VDMFP and MMI databases.

4.7 Summary

In this chapter, the study of a form of identity-independent image-based expression recognition problem has been presented. The Sparse Representation Classifier (SRC) with Principal Component Analysis (PCA) and Fisher Linear Discriminant Analysis (FLDA) methods are used with the posed database CK+ and the three types of spontaneous database (MMI, VD-MFP, and Belfast). The results confirm that the performance with the posed database was better than with the spontaneous database in terms of average accuracy.

In addition, it was also observed that the recognition rate results when using the difference images were higher than with the original images. The core reason is that the difference images are better than the original image because using the difference image emphasizes the expression areas in the face while eliminating the irrelevant parts. Moreover, a comparison of performance for image-based expression in three types of recent spontaneous databases was presented by using principles of sparse representation theory. It can be seen that the accuracy of facial expression recognition largely depends on the nature of the database in terms of background and illumination. The Belfast Induced Natural Emotion Database is taken under more challenging conditions, and the expression is less expressive due to the overlap among the expressions at a time. While with the MMI and Video Database of Moving Faces and People (VDMFP) to better average accuracy can be achieved.

The next chapter will introduce a novel method for FER with the spontaneous databases to extract the geometric features from the image rather than the texture features and using deep network based classification.

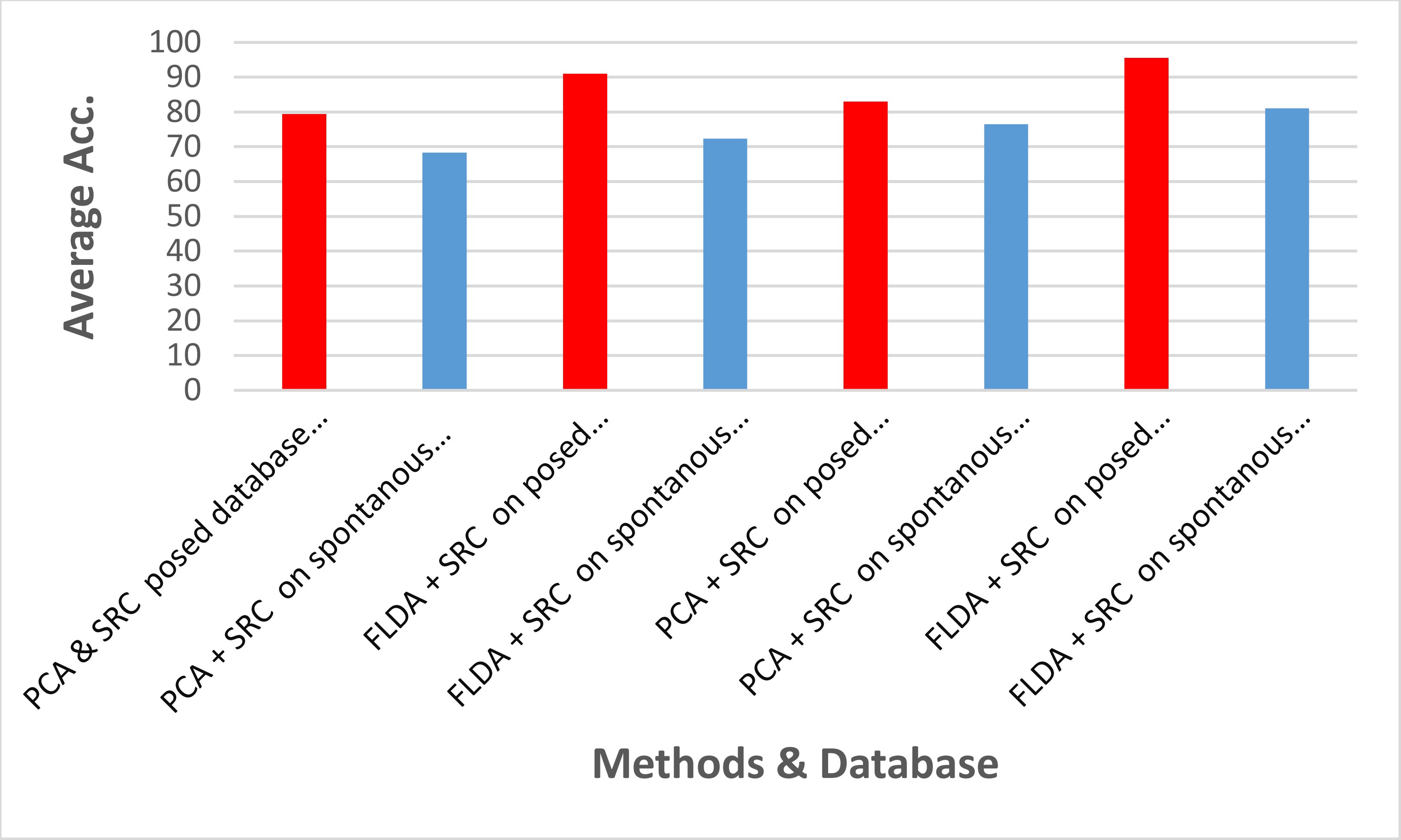


Figure 4.7: Comparison of recognition results percentage of the number of original and different images between posed and spontaneous databases using different methods (PCA + SRC and FLDA + SRC)

Chapter 5

Improving Image-Based Facial Expression Recognition

5.1 Introduction

Atric-based feature methods. Appearanceutomatic facial expression recognition systems require robust techniques for feature extraction and classification [142]. The basic processes of facial expression recognition are: (1) face detection; (2) extraction of expressive features and (3) final classification of expression. In the three processes mentioned, the second process plays a critical role in improving expression recognition results. Various techniques have been utilized to extract facial features which can be categorized into two principal types: appearance and geome-based features include changes in the texture of the face, such as furrows and wrinkles. It is hard to generalize these techniques to different people, however, they are less sensitive to noise compared with using geometric features [3]. Also, they contain extremely significant information regarding expression recognition, and micro-patterns exist in the skin texture of the face which can be encoded. Nevertheless, techniques for both types of features play a significant role in facial expression recognition [52, 63]. Different methods and algorithms have been developed and tested for facial expression recognition. Recently, there has been a research trend to exploit fusion of different features to improve the recognition rate of facial expressions [63, 143, 144].

Fusion of features contained in image spatial and frequency domains can produce significant information that cannot be found when using only one type of feature. The nature of the features and their distribution play a crucial role in the accuracy of facial expression recognition [145].

In Chapter 4, appearance features were investigated with two types of methods to extract features and reduce the dimensionality with SRC based-classifiers with original and difference images in order to obtain best average accuracy of the facial expression recognition. However, the average accuracy obtained was between 79% to 95% with a posed database and 50% to 75% with spontaneous databases for both original and difference images. Therefore the performance needs to be improved to achieve a higher accuracy of expression recognition. This is possible by using geometric features and a different type of classifier to obtain the highest average accuracy for the spontaneous databases.

In this chapter, approaches are proposed to obtain the best average accuracy of facial expression recognition:

- Pre-processing which is alignment, translation, rotation, and scaling (zooming in or out on the face in the camera view) to calculate displacement information for facial landmarks.
- Using another type of feature extraction method called geometric-based features that are more accurate and reliable when applied to the same three types of databases.
- A deep network is proposed that contains auto-encoders which have a simple architecture yet provide the basis for a more sophisticated classifier.
- The proposed classifier consists of a Self Organizing Map (SOM)-based classifier.

The structure of the remaining parts of this chapter is as follows: Section 5.2 explains the proposed method. Section 5.3 describes the proposed geometric feature extraction. Their evaluations, experimental results and performance analysis are given in Section 5.4. Finally, the summary of the overall chapter is represented in Section 5.5.

5.2 Proposed Method

In this work, an image-based automatic facial expression recognition system which exploits the geometric features of facial expressions with relatively low complexity is presented. It is necessary to compare every frame with a neutral facial expression which is considered to be the reference frame for measurement. Unfortunately, in previous work [63] the spatial normalization applied is not sufficient to ensure the eyes in all faces are aligned correctly because some of the spontaneous databases come from highly unconstrained environments or from the wild. To address this problem, the region of the face is aligned by normalization of the rotation about the center point in the horizontal dimension to correct most possible geometric issues, such as translation, rotation and scaling (zooming in or out of the face in the camera view).

5.2.1 Real-time Detection

A process to extract the features accurately is a vital step to obtaining the best performance in recognition of facial expression. Detection is performed in two important stages:
1) Facial Detection: The first step is pre-processing for facial detection from an image sequence. In this work, one of the most common detection algorithms was used, that is the Viola-Jones face detector [139]. This algorithm is described by high-grade performance in terms of accuracy and speed compared with other algorithms. The procedure for extracting geometric features from an image starts as follows [63]:

- Detection of the face to determine the height of the face (H_{face}) by using the Viola Jones [139] object detection algorithm.

- Detection of both eye regions, including the eyebrow to measure the height of the eye region. Moreover, the centers of the eyes are determined which are denoted as (x_1, y_1) and (x_2, y_2) by estimating the region that consists of the eyes and eyebrows in order to obtain eye region height (H_{eye}).

- Detection of the mouth region to derive its height which occupies 1.5 of the (H_{face}) where $(H_{lips}) = 1.5 \times (H_{face})$.

- Using the center of the eyes, locate the mouth from the face height, and estimate the nose region that occupies 1/3 of the (H_{face}). Calculate the nose height by $(H_{nose}) = 1/3 \times (H_{face})$. The estimated region of the eyes, also includes the eyebrow regions. The located centers of the eyes are (x_1, y_1) and (x_2, y_2). According to the eye center, the nose was estimated where nose width by computing the distance between $x_2 - x_1$. Fig 5.1 displays the extraction of four regions of the face utilizing facial geometric information.

2) Detection of Landmarks: There are many possible geometric or fiducial point detection techniques, the most famous exploit two models that are an Active Appearance Model (AAM)

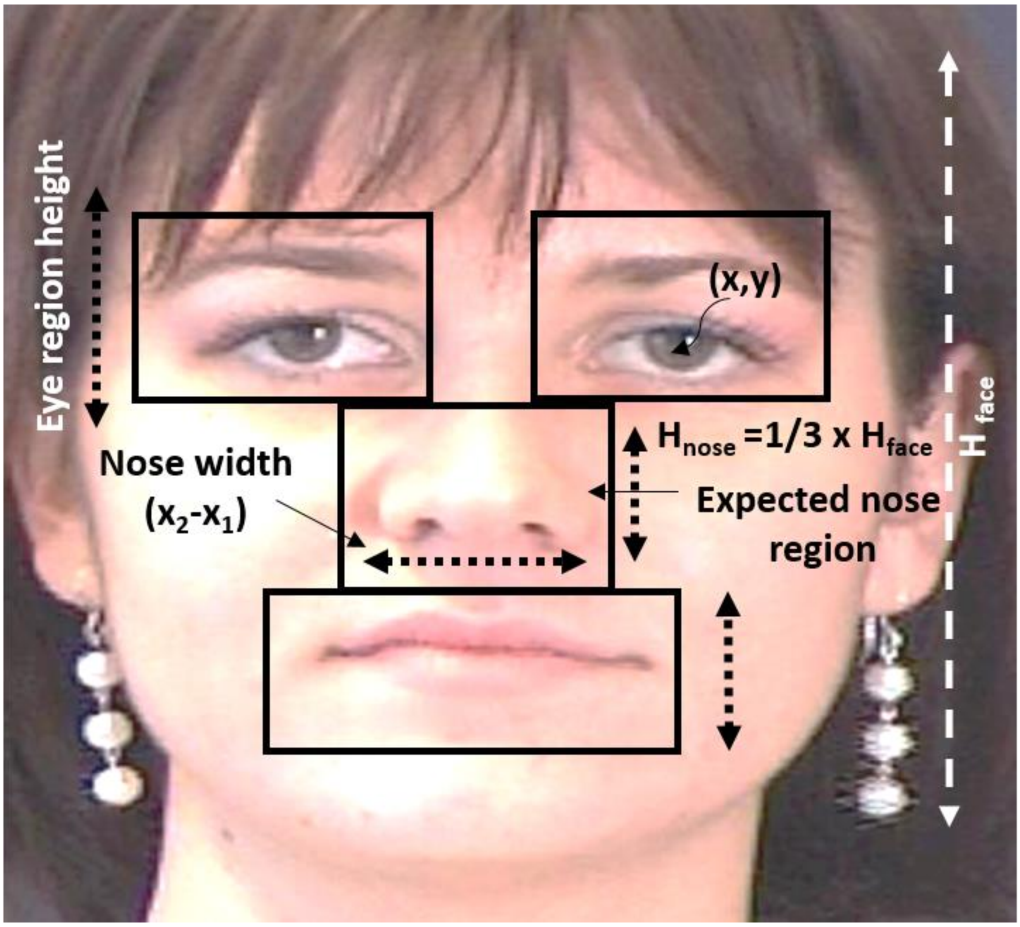


Figure 5.1: The four regions used for facial expression recognition, namely face height, eyes and mouth location.

[48] and Active Shape Model (ASM) [49]. Nevertheless, there is one major weakness in both approaches, namely that they require 68 landmarks in every training image to be labelled manually, which is a difficult task for all of the training images for the three types of large databases. This causes loss of time and effort. Discriminative Response Map Fitting (DRMF) [53], is another method to detect points in the face, which is characterized by highly efficient computation. It allows for detection of the facial points in the face in real time and maintains an accurate detection of these points even if the face is partly occluded.

In each frame of the successive frames, the facial regions are detected by using Viola-Jones [139]; then the landmark points are extracted from those regions by applying the DRMF algorithm [53] by selecting 22 landmarks as in Fig 5.2. In the case that the algorithm fails to detect the face or key facial regions, the information of the previous frame is used to estimate the location of the required regions in the current frame, this method improves the detection accuracy of the main facial regions.

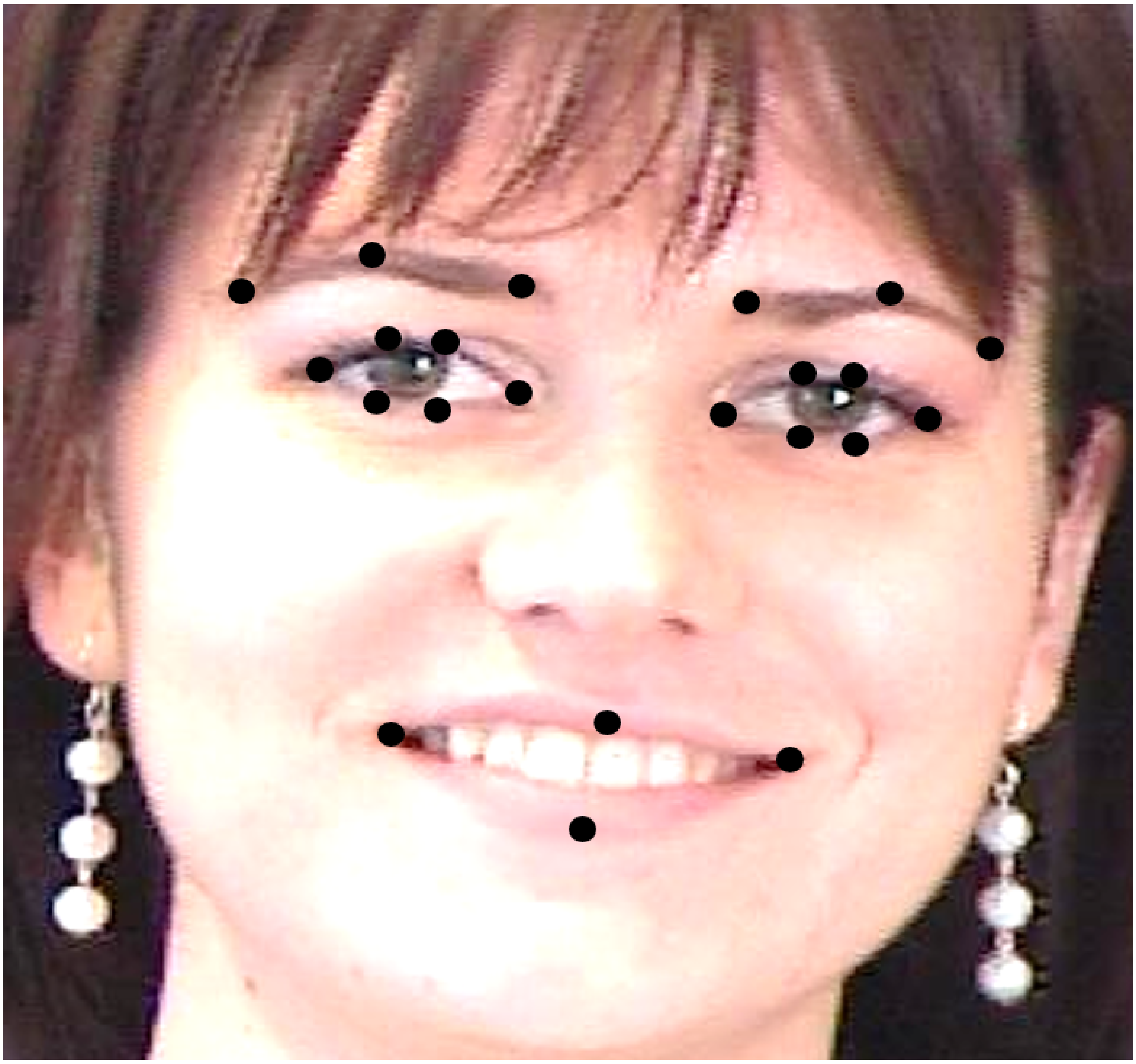


Figure 5.2: Selective DRMF landmarks to extract geometric feature vector from three databases.

5.2.2 Alignment Process

Let $R = \{\mathbf{r}_i\}_{i=1}^{n}$ and $E = \{\mathbf{e}_i\}_{i=1}^{n}$ be respectively the 2-dimensional column-wise landmarks for the reference and expression images, where n is the total number of landmarks considered. Let $\{\bar{\mathbf{r}}_r, \bar{\mathbf{r}}_l\} \subset R$ and $\{\bar{\mathbf{e}}_r, \bar{\mathbf{e}}_l\} \subset E$ be respectively the right and left inner corner points for the reference image and the right and left inner corner points for the expression image. According to this notation, let $\mathbf{m}_{ref} = \frac{1}{2}(\bar{\mathbf{r}}_r + \bar{\mathbf{r}}_l)$ and $\mathbf{m}_{exp} = \frac{1}{2}(\bar{\mathbf{e}}_r + \bar{\mathbf{e}}_l)$ be the middle points between each couple of inner points, as shown in Fig. 5.3.

1. *Translation*: with simple algebra, R' and E' are defined as the sets of translated points in R and E, such that $\mathbf{m}'_{ref} = \underline{0}$ and $\mathbf{m}'_{exp} = \underline{0}$.

2. *Rotation*: let $\theta_r = sign(\bar{\mathbf{r}}'_r \cdot \mathbf{u}_2) \arccos\left(\frac{\bar{\mathbf{r}}'_r \cdot \mathbf{u}_1}{|\mathbf{r}'_r|}\right)$ where $\mathbf{u}_1$ and $\mathbf{u}_2$ represent respectively the unit vectors along the x and y axes [146]. Let θ_e also be similarly defined for the expression image.

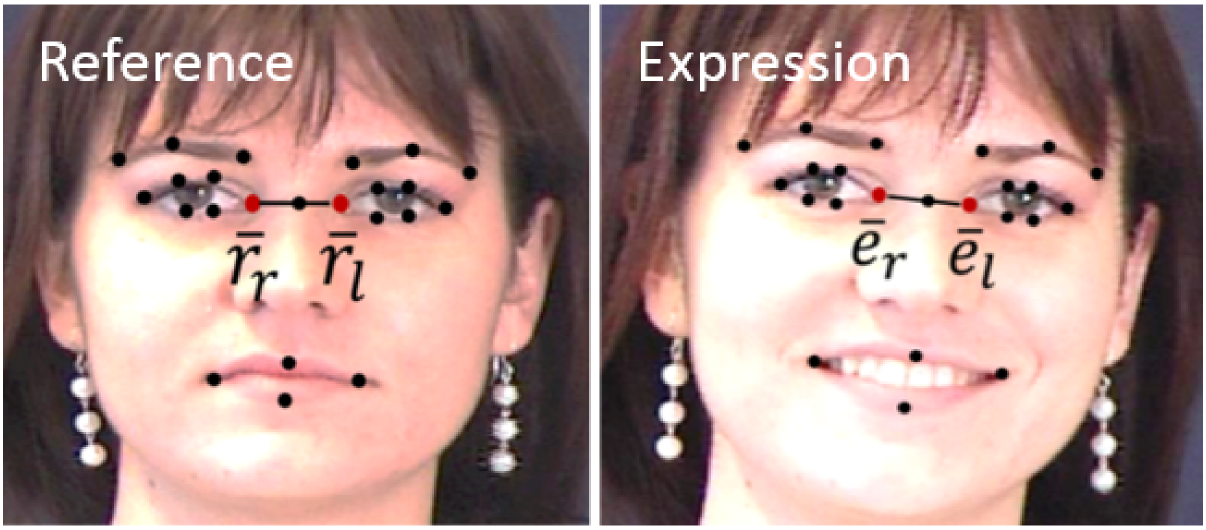


Figure 5.3: Middle points between the eyes using detected right and left inner corner points in the reference image and expression images.

Defining the transformation matrix

$$T(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}. \tag{5.1}$$

$\mathbf{r}''_i = T(\theta_r)\mathbf{r}'_i$ and $\mathbf{e}''_i = T(\theta_e)\mathbf{e}'_i$ have been computed for all $i = 1, \ldots, n$, in order to align $\bar{\mathbf{r}}''_r$ and $\bar{\mathbf{e}}''_r$ with the x-axis. Thus, the sets R'' and E'' remain defined.

3. The final step consists of defining E''' as the set of points obtained by scaling E'' in order to match the $\bar{\mathbf{e}}'''_l$ and $\bar{\mathbf{e}}'''_r$ respectively with $\bar{\mathbf{r}}''_r$ and $\bar{\mathbf{r}}''_r$. Due to symmetry constraints, defining

$$\sum = \begin{bmatrix} \frac{||\mathbf{r}'_r||}{||\mathbf{e}'_r||} & 0 \\ 0 & \frac{||\mathbf{r}'_r||}{||\mathbf{e}'_r||} \end{bmatrix}. \tag{5.2}$$

$\mathbf{e}'''_i = \Sigma \mathbf{e}''_i$ are computed for all $i = 1, \ldots, n$.

In conclusion, E''' and R'' have been defined such that the expression inner points match with the correspondent reference inner points, solving the misalignment problem. For simplicity, it will be assumed that $E = E'''$ and $R = R''$ for the rest of the chapter.

Algorithm 1 provides a brief explanation of the pre-processing.

Algorithm 1 Pre-Processing for Reference and Expression Frames

```
for frame $i$ do
    Face detection applying Viola Jones [139]
    if $\{\bar{\mathbf{r}}_r, \bar{\mathbf{r}}_l\} \subset R$ and $\{\bar{\mathbf{e}}_r, \bar{\mathbf{e}}_l\} \subset E$ are detected then
        the medium points are calculate by
        $\mathbf{m}_{ref} = \frac{1}{2}(\bar{\mathbf{r}}_r + \bar{\mathbf{r}}_l)$ and $\mathbf{m}_{exp} = \frac{1}{2}(\bar{\mathbf{e}}_r + \bar{\mathbf{e}}_l)$
        if the frame is reference image then
            Translation
            Rotation
        else
            Scaling
        end if
    end if
end for
```

5.3 Geometric Feature Extraction

Having a distinct and robust set of features is essential for each classification task. More importantly a group of features that are suitable for the problem at hand should be predefined. Predominately, feature selection is used to obtain a specific and an optimal subset of features from a predefined larger set of features to provide the robust classification. However, the features to be used can be learnt autonomously in a data-driven manner, instead of predefining a feature group. Therefore, geometric features are proposed to address and apply to improve the performance and obtain the highest accuracy for the facial expression recognition.

In order to extract the geometric features of facial expressions, it is necessary to compare every video sequence with a neutral facial expression which is considered to be the reference frame for measurement.

The 22-D geometric features were selected from the 66 landmark points detection based on the DRMF algorithm [53], where the geometric feature vector for facial expressions is $\mathbf{x_g} \in \mathbb{R}^{22}$. This vector includes directional displacement information concerning facial landmarks. The mean distance between the inner corner points of the right and left eyes is used as the reference

for the alignment and rotation of head pose and back dilation (zooming in or out on the face in the camera view) to calculate displacement information for facial landmarks.
Then, after segmenting the key facial regions, geometric features are extracted by detecting facial regions by calculating the displacement for facial landmarks points features as follows:

I. The features of the eye are extracted by determining six landmarks for each eye as in Fiq 5.2, that consist of two points in the inner and outer corners of the eye as well as two points in the upper eyelid and two points in the lower eyelid. Then the mean is extracted for every two points at the upper eyelid $\mathbf{m}_{ue} = \frac{1}{2}(\mathbf{x}_{ue} + \mathbf{y}_{ue})$ and lower eyelid $\mathbf{m}_{le} = \frac{1}{2}(\mathbf{x}_{le} + \mathbf{y}_{le})$ to define the center of the eye.

II. Three points are located for each eyebrow, one point on the left corner and another at the right corner of the eyebrow, and the last point in the middle of the eyebrow.

III. Finally, four landmarks on the mouth are located, that is a landmark point at the left corner and another at the right corner, one point in the upper middle and one point in the lower middle of the pair of lips. After that, the projected ratio of vertical (V_{Proj}) and the horizontal (H_{Proj}) for the four new points in the eyes and mouth is calculated by (H_{Proj}/V_{Proj}) to estimate the degree of eye and mouth opening and shape their features, as shown in Fig. 5.4. Also, these features help to maintain the direction and location information for the key facial points.

In addition to obtaining displacement information about each key facial points, these features help to maintain the direction and location information for those features. Often each sequence in the three types of databases starts with a neutral facial expression, and then expression begins and ends with the peak expressions. Therefore, the first frame is considered to be the reference frame. The following concepts are utilized in this work. The landmarks of the first frame are determined as the reference frame, then the 22-D geometric features are extracted from the first frame and all of the consecutive images as in Fig 5.2.

5.3.1 Auto-encoder Deep Network

The auto-encoder concept has recently been more broadly applied to learning generative models of data [147], it is an unsupervised learning algorithm for efficient coding [148] using a backpropagation technique to approximate the target values in the input feature vector [63].

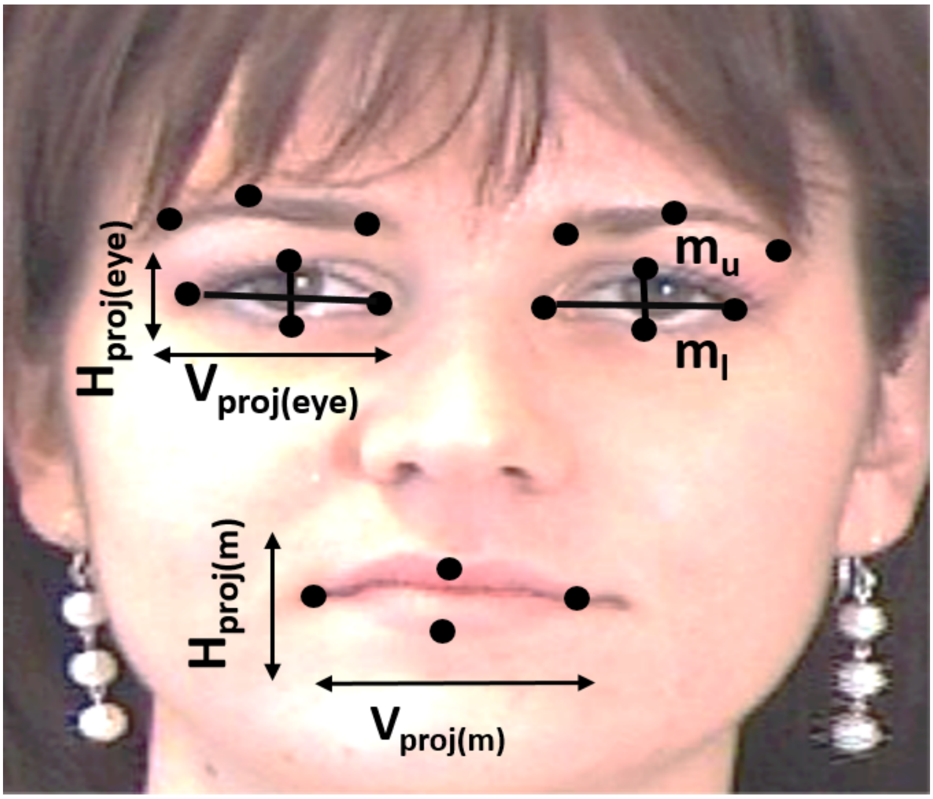


Figure 5.4: The 22 geometric features with an estimate of the degree of eye and mouth opening.

It has three layers which are input, hidden (encoding), and decoding layers as shown in Fig. 5.5. For a particular input vector $\mathbf{x} = [x_1, x_2, ..., x_5]$, the auto-encoder attempts to learn a function so that the output of the network $h_{w,b}(\mathbf{x}) \approx \mathbf{x}$. This means that it is attempting to learn an approximation of the identity function, so that, the output $\hat{h}(\mathbf{x})$ is similar to the input vector $h(\mathbf{x})$.

As a particular example, assume that the input data vector $\mathbf{x}$ has intensity values of 100 pixels from 10 × 10 images, and so n = 100, and the number of hidden nodes in level 2 is $l2 = 50$, and also $\mathbf{y} \in R^{100}$. Note that there are 50 hidden units, thus the neural network is attempting to learn a compressed representation of the input, i.e., presented just with the hidden unit vector of activations $\mathbf{a}^{(2)} \in R^{50}$. Therefore, it requires attempting to reconstruct the 100-pixel input x. If the input data were completely random, then the compression process would be very complicated.

As in the discussion above, the number of hidden nodes adopted in $l2$ is small. However, even if the number of hidden nodes is even greater than the number of input data pixels, it is possible to discover interesting structures through prescribing other constraints for the network. The hidden nodes representation of $\mathbf{x}$ are provided by $\hat{h}_j(\mathbf{x})$

$$\hat{h}_j(\mathbf{x}) = f(a_j(\mathbf{x})) \quad \text{where} \;\; a_j(x) = b_j + \sum_k W_{jk} x_k \tag{5.3}$$

$\hat{h}_j(\mathbf{x})$ is the yield of the hidden layer j after implementing the activation function $f(\cdot)$. In general, the activation function $f(\cdot)$ is a sigmoid function b which generates an output between (0 to 1), such that the reconstructed yield $(\hat{\mathbf{x}})$ is obtained by applying a decoding function [63]. It can be presented as follows:

$$\hat{\mathbf{x}} = g(\hat{a_k}) \quad \text{where} \;\; \hat{a_k} = c_k + \sum_j W^*_{jk} \hat{h}_j(\mathbf{x}). \tag{5.4}$$

The function $g(\cdot)$ has often been selected as a sigmoid function because it is particularly used for models that have a high probability of being predicted as outputs. It is noted empirically [149] that, in accordance with training a Restricted Boltzmann Machine (RBM), W^* becomes equivalent to W^T where $(\cdot)^T$ is the transpose operation. In the decoding section of the auto-encoder, W^* is set to W^T which is comparable to the RBM [150]. This approach does not allow the network to learn the function of trivial identity. The choice of $W^* = W^T$ provides best results even if the number of the hidden nodes is greater than the input nodes number, such that the parameters b_j and c_k are set to unity.

In the suggested technique, the model is trained with geometric features utilizing auto-encoders. The features represented in the first layer are concatenated so as to consist of 22 features, which are provided as the input to the second layer represented by the Self-Organizing Map(SOM)-based classifier, that is associated with the auto-encoder as a final layer. The output of the nodes in the final layer is considered to represent the final output features. Fig. 5.5 shows an example of an auto-encoder with a single hidden layer. Fig. 5.6 illustrates using a technique based on auto-encoders to encode geometric features.

In the first step the geometric features as the input feature $\mathbf{x}_g \in R^{22}$ are passed through the first layer auto-encoder, then the outputs of the hidden layers $(\hat{h}^g)$ of the auto-encoder are concatenated and passed through the second layer of the auto-encoder. The output of the hidden layer j at the first layer for the geometric feature vector $\mathbf{x}_g$ can be calculated as follows:

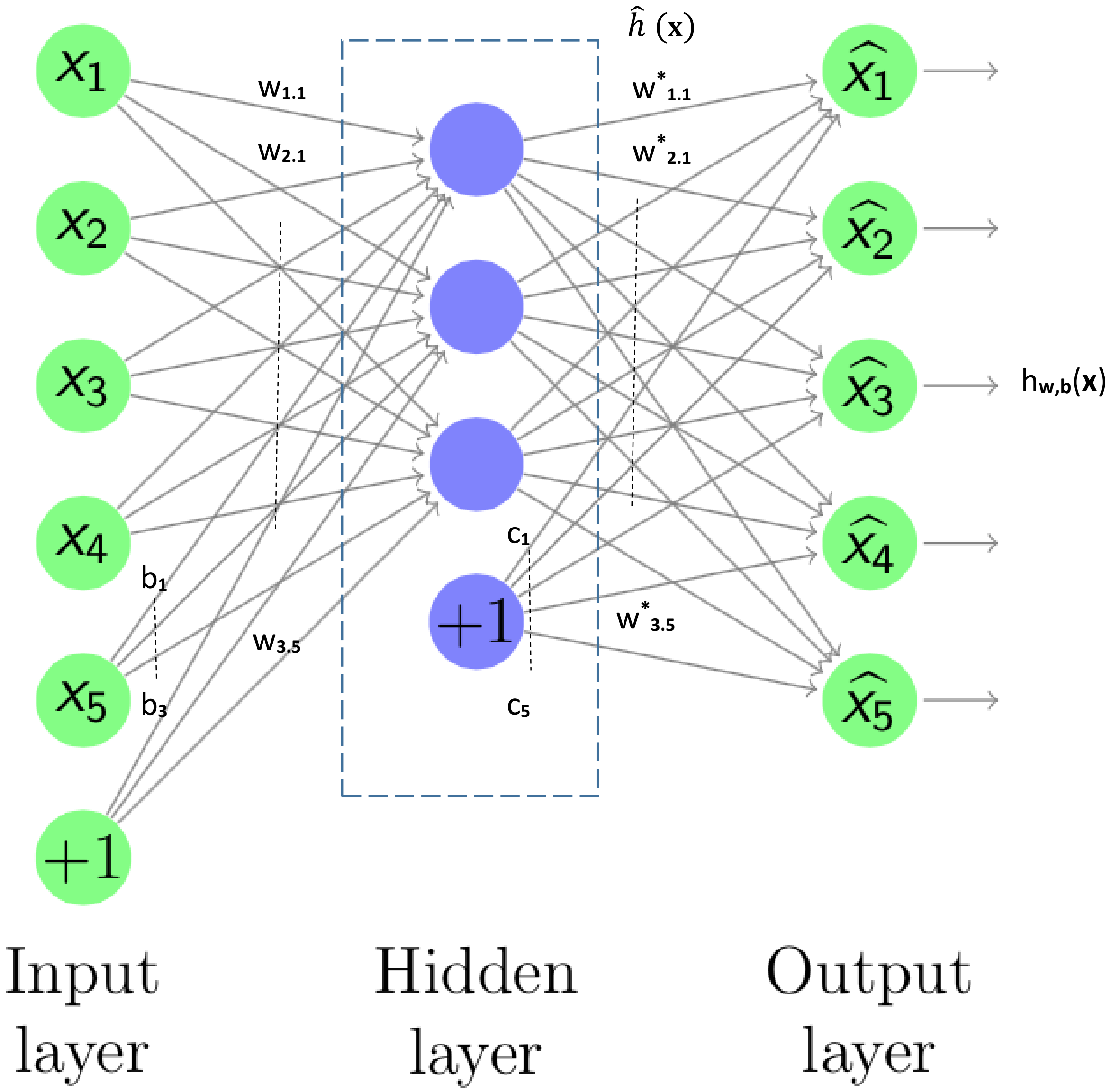


Figure 5.5: Architecture of an auto-encoder with single hidden layer [63].

$$\hat{h}_j^z \mathbf{x} = \frac{1}{1 + e^{-a_j^Z}(\mathbf{x})} \tag{5.5}$$

where

$$a_j^Z(\mathbf{x}) = b_j^Z + \sum \mathbf{W}_j^{Zk} \mathbf{x}_k \tag{5.6}$$

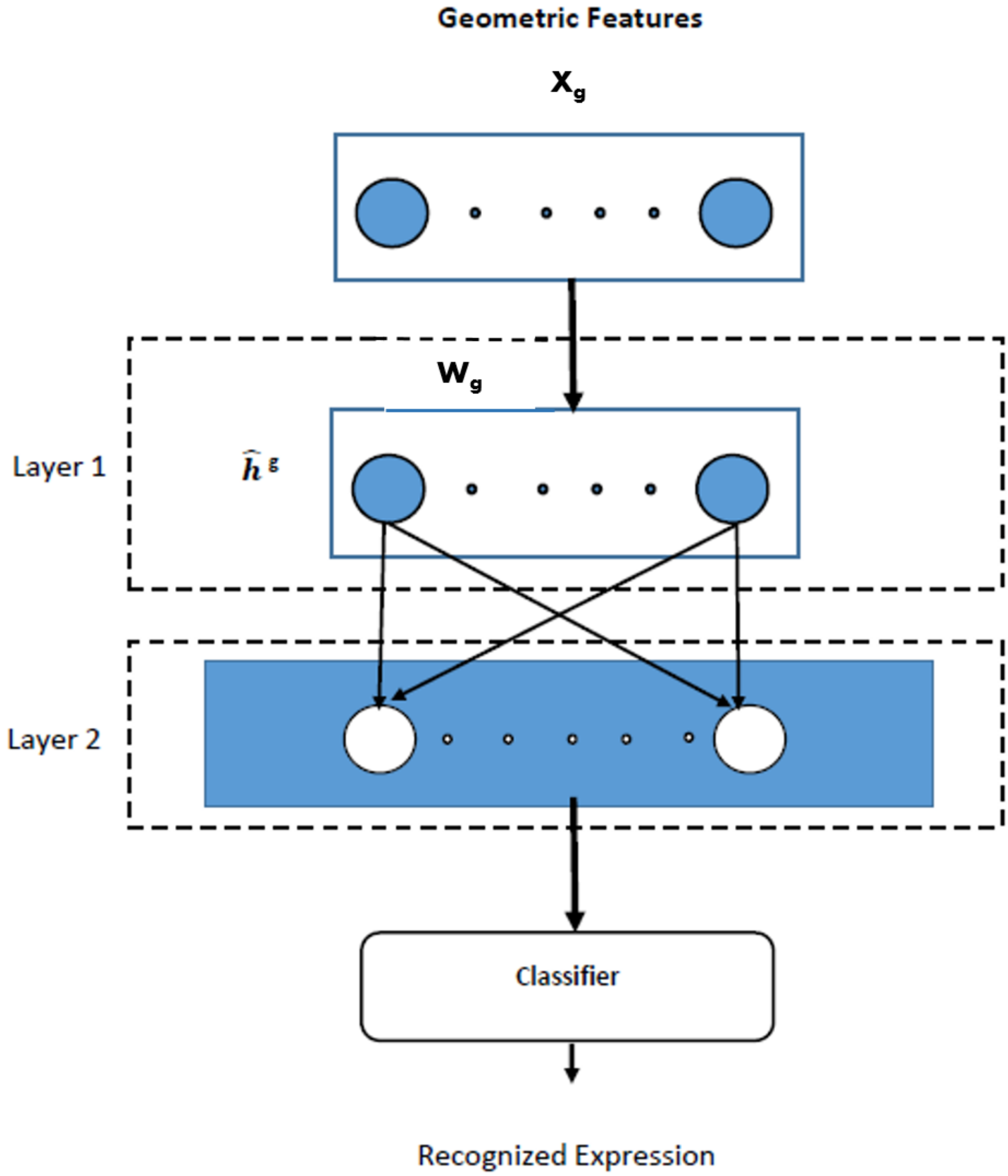


Figure 5.6: A technique based on auto-encoders is applied to an input feature vector $\mathbf{x}_g$ that consists of the geometric features; $\hat{\mathbf{h}}^g$ is the output of the hidden layer and $\mathbf{w}_g$ is the updated weight vector.

and a^z is the output of the hidden layer at node j before using the activation function, Z is the data type which is geometric features, and b_j^z is a bias term for the feature vector $\mathbf{x}$.

$\mathbf{x}_g$ is the input feature vector of the geometric features. Regarding $\hat{h}^g \in R^M$ which is the hidden layer output for the geometric feature vector $\mathbf{x}_g$ after using the activation function, hence the connected feature vector of the hidden layer $\mathbf{x}_2$ be an input vector for the second layer is computed by

$$\mathbf{x}_2 = [\hat{h}^g_1, \hat{h}^g_2, ...\hat{h}^g_M] \tag{5.7}$$

where M is the overall number of hidden nodes of the auto-encoders deep learning for geometric features. The weights $\mathbf{W}$ and bias $\mathbf{b}$ in the auto-encoder are updated by applying a stochastic gradient descent algorithm. The update rules of the weight and bias in layer l of auto-encoder network are determined for a training sample $\mathbf{x}^{(i)}, \mathbf{y}^{(i)}$ where $i \in (1, 2, 3, ..., M)$ is presented as

$$\mathbf{W}(i+1) = \mathbf{W}(i) - \alpha \frac{\partial}{\partial \mathbf{W}(i)} J(\mathbf{W}, \mathbf{b}; \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \tag{5.8}$$

$$\mathbf{b}(i+1) = \mathbf{b}(i) - \alpha \frac{\partial}{\partial \mathbf{b}(i)} J(\mathbf{W}, \mathbf{b}; \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \tag{5.9}$$

where, $J(\mathbf{W}, \mathbf{b}; \mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, which is the cost function in relation to a single training example $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, is defined as

$$J(\mathbf{W}, \mathbf{b}^{(1)}, ; \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = \frac{1}{2} \mid\mid \mathbf{W}^{(l)T}\mathbf{h}(\mathbf{x}) - \mathbf{x} \mid\mid^2 - \frac{\lambda}{2} \sum_j \sum_k (w^{(l)}_{jk})^2 \tag{5.10}$$

where the sum of squares error term is the first term of the error between the $\hat{\mathbf{x}} = \mathbf{W}^T h(\mathbf{x})$ and $\mathbf{x}$, at the input at layer $l-1$. A regularization or weight decay term is the second term that tends to decrease the magnitude of the weights, and helps prevent overfitting. The parameter λ refers to the weight decay parameter, and more details on the parameter updating rule can be found elsewhere [150].

The output feature vector is classified into six basic expressions by applying the Self-Organizing Map (SOM)-based classifier [11]. The SOM-based classifier is described in the next section.

5.3.2 SOM-Based Classifier

The SOM-based classifier[112] works in two phases namely training and mapping. In training, classification of new vectors is automatic, while in the mapping, the map is constructed by using input examples. In the proposed technique, the SOM is trained with all geometric features after utilizing auto-encoders to obtain a closer model of the data according to the corresponding expression. The training processing SOM is complementary to the process of the auto-encoder, where it is providing the final decision of the classification.

Let $\mathbf{x}$ be a feature vector that is corresponding to a facial image, $\mathbf{y}$ is the yield of the classifier which describes the emotions classes (6-D vector) for example, anger, disgust, fear, happiness, sadness, and surprise. These 6-D vectors are labeled as ON or OFF; for example, anger is the first class to be labeled as 1 (ON), while all other classes are labeled as -1 (OFF), as follows $\mathbf{y} = [1 - 1 - 1 - 1 - 1 - 1]^T$. Therefore, mapping of the input and output for features of the facial expression can be mathematically expressed as follows:

$$\mathbf{y} = f(\mathbf{x}), \quad \mathbf{y} \in \mathbb{R}^6; \quad \mathbf{x} \in \mathbb{R}^{22} \tag{5.11}$$

The highly nonlinear mapping can be linearized utilizing the first order Taylor concatenation expansion of each neuron in the network. Considering the γth neuron in the network is correlated by a vector $\mathbf{w}_\gamma \in R^{22}$ in the network, the linear model related to these neurons is expressed as follows:

$$\mathbf{y}_\gamma^{out} = \mathbf{y}_\gamma + \mathbf{A}_\gamma(\mathbf{x} \mid \mathbf{w}_\gamma). \tag{5.12}$$

where

$$\mathbf{A}_\gamma = \frac{\partial f}{\partial x} \mid \mathbf{x} = \mathbf{w}_\gamma \in \mathbb{R}^{6\times 22}. \tag{5.13}$$

$$\mathbf{y}_\gamma = f(\mathbf{w}_\gamma). \tag{5.14}$$

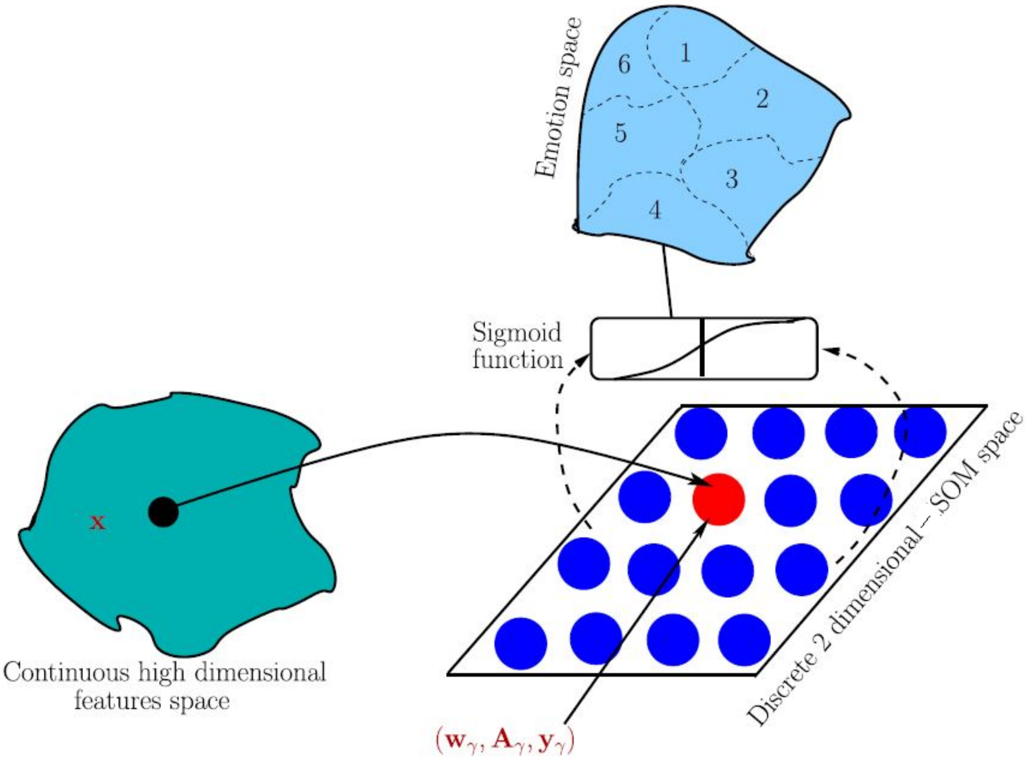

Emotion space
1
2
3
4
5
6
Sigmoid function
x
Continuous high dimensional features space
Discrete 2 dimensional– SOM space
$(\mathbf{w}_\gamma, \mathbf{A}_\gamma, \mathbf{y}_\gamma)$

Figure 5.7: A classification technique based on the SOM, every node describes the local properties of the input feature space, such that the parameters for the node γ are associated with matrix $\mathbf{A}_\gamma$, weight vector $\mathbf{w}_\gamma$ and bias $\mathbf{x}_\gamma$.

During applying the SOM-based classifier despite the winning neuron being dominant, it allows neighboring neurons to participate in decision making.

The output $\mathbf{y}$ is provided by the collective response of the network as follows:

$$\mathbf{y}^{out} = \frac{\sum_{\gamma=1}^{M\times N} h_{i,\gamma}\mathbf{y}_\gamma^{out}}{\sum_{\gamma=1}^{M\times N} h_{i,\gamma}}. \tag{5.15}$$

where $h_{i,\gamma}$ represents the neighbourhood function to the best matching unit (BMU) for i and the neuron γ.

Let $\mathbf{x}$ be the topological position node in the lattice of the SOM, $\mathbf{x}_i$ is the feature vector for nodes, where $(i = 1, 2, ..., n)$. Every node is connected to the weight vector of the last layer of the auto-encoder; $\mathbf{w}$ is the weight vector, of n dimensions,$[w_1, w_2, w_3, ..., w_n]$. All units of the SOM of the input vector are used and a measure of the similarity between them is used to identify the best winner node,

$$i = \underset{\gamma}{\operatorname{argmin}} \left\| \mathbf{x} - \mathbf{w}_{\gamma} \right\| . \tag{5.16}$$

where i is the winning node, and γ is each neuron that is associated with the locally valid linear models to provide a collective response model.

In this technique, the collective response model equation (5.15) has been expressed utilizing locally valid linear models correlated with every neuron γ. Hence, through learning local linear mapping, model (5.15) approaches the global nonlinear mapping, $\mathbf{y} = f(\mathbf{x})$. This method has the advantages of rapid learning and high accuracy. The problem of learning is to update the parameters $\mathbf{w}_{\gamma}, \mathbf{y}_{\gamma}, \mathbf{A}_{\gamma}$ based on the training pairs $\left\{\mathbf{x}, \mathbf{y}^{d}\right\}$. The neighbourhood function $hi, \gamma(n)$ and the spread $\sigma(n)$ for n_{th} repetition are computed as in equation (5.15) and (5.19) [151]. A sigmoid activation function is utilized to classify the input feature vector into six fundamental expressions at each node of the network, rather than to specify a hard threshold value for the detached outputs from the network. In fact, the highest value (Gr) for all the sigmoid function (tan hyperbolic) outputs are found. The node achieves the (Gr) which is set to 1, while the rest are set to -1. Typically, the node is chosen which provides the maximum activation as the recognized class. The sigmoid function output of the kth class is indicated by $\mathbf{y}_{k}^{sig}$

$$\mathbf{y}_{k}^{sig} = \frac{e^{\mathbf{y}_{k}^{out}} - e^{-\mathbf{y}_{k}^{out}}}{e^{\mathbf{y}_{k}^{out}} + e^{-\mathbf{y}_{k}^{out}}}. \tag{5.17}$$

$$\text{if} \qquad \mathbf{y}_{k}^{sig} = Gr \quad \text{then} \quad \mathbf{y}_{k} = 1$$

$$\text{else} \qquad \mathbf{y}_{k} = -1 \tag{5.18}$$

That can be described as a winner-takes-all policy. The hard threshold has been estimated to be 0.5 [151], which is the response of multiple output nodes which passed the hard threshold on many occasions. According to equation (5.18), the model guarantees that just one response is sent to the output nodes which will be ON while the rest will be OFF. The feature of the SOM-based classifier is necessary as it takes into account that each facial image in three-DBs is labeled with just one category of expression. It has been experimentally noted that the parameter which are updated based on the minimization of least squares error produces the highest accuracy, and the sum squares cost function can be considered as follows:

$$\mathbf{E} = \sum_{k=1}^{6} \mathbf{E}_k = \frac{1}{2} \sum_{k=1}^{6} (\mathbf{y}_k^d - \mathbf{y}_k^{sig})^2 \tag{5.19}$$

The update of the parameter $\mathbf{w}_\gamma$ is preserved in a similar way as in SOM, and the steps below show the updates of the parameters $\mathbf{A}_\gamma$ and $\mathbf{y}_\gamma$. The derivatives of the cost function concerning the parameters for every output node k are provided as:

$$\frac{\partial \mathbf{E}}{\partial \mathbf{y}_\gamma} = -\frac{(\mathbf{y}_k^d - \mathbf{y}_\mathbf{k}^\mathbf{sig}) h_{i,\gamma}}{\sum_{\gamma=1}^{M \times N} h_{i,\gamma}} \left(1 - (\mathbf{y}_k^{sig})^2\right) \tag{5.20}$$

$$\frac{\partial \mathbf{E}}{\partial \mathbf{a}_\gamma} = -\frac{(\mathbf{y}_k^d - \mathbf{y}_\mathbf{k}^\mathbf{sig}) h_{i,\gamma} (\mathbf{x} - \mathbf{w}_\gamma)}{\sum_{\gamma=1}^{M \times N} h_{i,\gamma}} \left(1 - (\mathbf{y}_k^{sig})^2\right) \tag{5.21}$$

where the kth row of the matrix $\mathbf{A}$ at neuron γ into the SOM network is represented by $\mathbf{a}_\gamma$. For each neuron γ the parameter vector $\mathbf{w}_\gamma$ is updated within the neighbourhood $h_{i,\gamma}$, so that it is achieved using an unsupervised method, whereas the update of parameters $\mathbf{y}_\gamma$ and $\mathbf{A}_\gamma$ utilize both supervised and unsupervised terms.

5.4 Experimental Results

In the experiments, the same three databases used in Chapter 4 were adopted, which include difficult challenges [152] and have been selected to cover the various conditions of imaging, where some of these databases are closer to the wild. The three databases are the Video Database of Moving Faces and People (VDMFP) [59], MMI facial expression database [58] and Belfast Induced Natural Emotion Database (BINED) [60] as mentioned in Chapter 3. The first frame is considered to be the reference frame, and the last frame is considered to be the apex frame. The face images are normalized in the three databases to a standard scale of (240×240) pixels.

The average accuracy for facial expression recognition is calculated by using 22-D geometric and appearance feature vectors with the auto-encoder deep network which is a deep-network-based on automatic facial expression recognition system, where the first two layers of the auto-encoders process the geometric features for better description of the facial data. In addition, the third layer combines the benefits of both supervised and unsupervised learning algorithms by the SOM-based classifier.

The PCA method for comparison has been applied to obtain the average recognition accuracy for the aforementioned databases with the same deep network and classifier with the same number of 22-D appearance features. However, these features provide worse results, with accuracy less than (50%). So, the number of appearance features has been gradually increased to 31-D appearance features which provided the best average accuracy.

The comparison of best recognition rate in terms of average accuracy between geometric and appearance features are provided in Tables 5.1- 5.6. The details of experimental results for the three spontaneous databases are presented below.

MMI database:

- Table 5.1 represents full descriptions of the confusion matrix with average accuracy for facial expression recognition by 22-D geometric features with the deep network when using the SOM-based classifier. The average recognition accuracy of the geometric features is (97.0%) with a highest recognition accuracy of the best class being (100%) in the case of Anger and the lowest accuracy being (93.3%) with disgust class.

Table 5.1: Confusion matrix showing evaluation results of 22D-geometric features with MMI Database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**100**	0	0	0	0	0
Disgust	1.11	**93.3**	0	2.22	2.22	1.11
Fear	1.11	0	**96.7**	0	2.22	0
Happy	0	1.11	0	**97.8**	0	1.11
Sadness	0	0	3.33	0	**96.7**	0
Surprise	0	1.11	0	1.11	0	**97.8**

- Table 5.2 represents full descriptions of the confusion matrix with average accuracy for facial expression recognition 31-D appearance features with the deep network when using the SOM-based classifier. The average recognition accuracy of the appearance features is (91.9%) with the highest accuracy being (100%) of happiness and surprise classes and (70%) of disgust class for the lowest, which is confused with anger class giving (86.7%).

It can be seen that the results of the proposed system in terms of average recognition accuracy of (97.0%) and (91.9%) outperform in terms of average recognition accuracy of (91.17%) and (91.27%) the MMI databases with separately geometric and appearance features as reported in [63], as shown in no. 1 from section Experimental Results on MMI Database and Tables 5.1 and 5.2 from this reference. The details of the performance of average recognition accuracy for geometric and appearance features are as in Fig 5.8.

Table 5.2: Confusion matrix showing evaluation results of 31-Appearance features by PCA with MMI Database

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**86.7**	13.33	0	0	0	0
Disgust	30	**70**	0	0	0	0
Fear	0	0	**98.99**	0	1.11	0
Happy	0	0	0	**100**	0	0
Sadness	0	0	4.44	0	**95.6**	0
Surprise	0	0	0	0	0	**100**

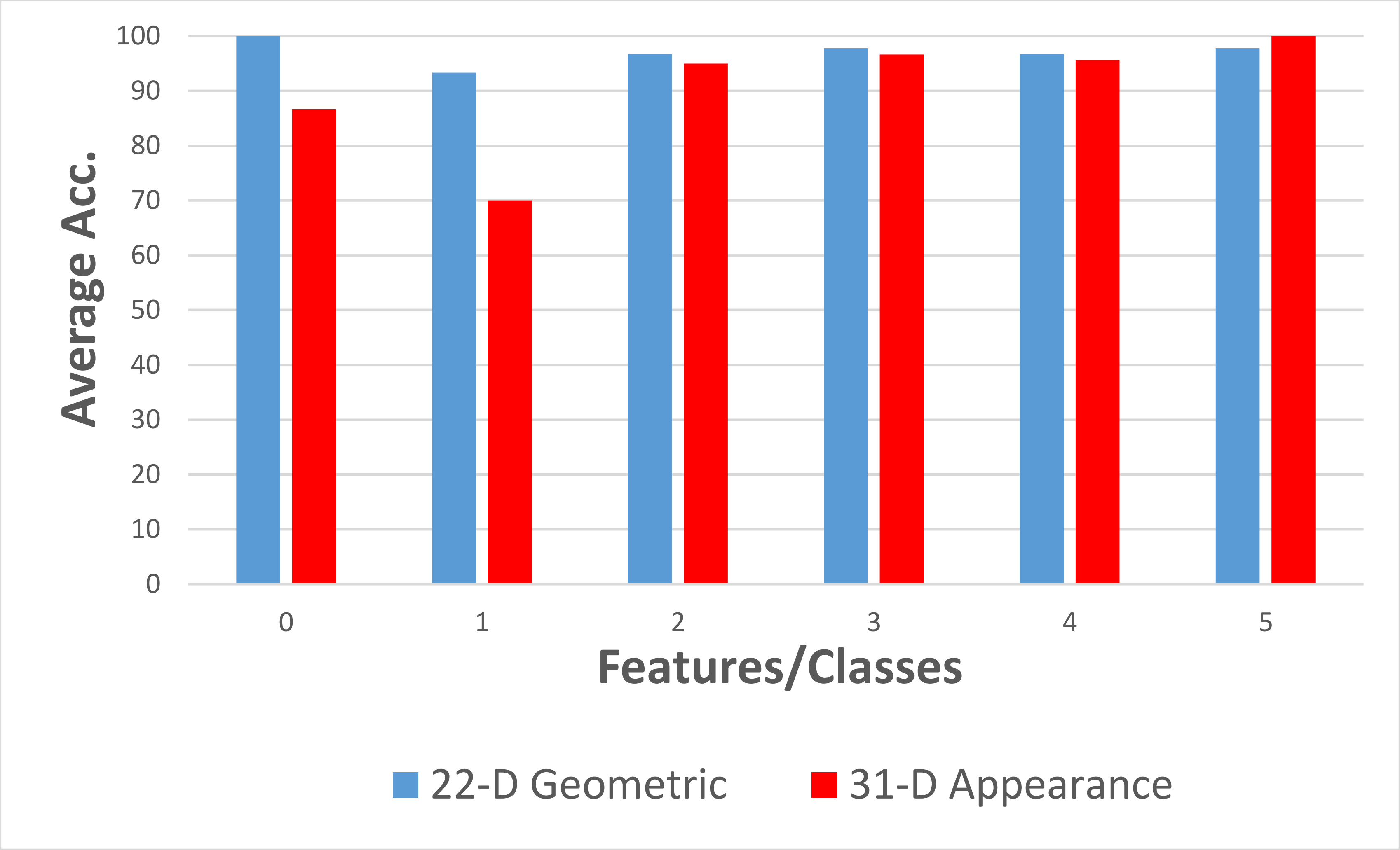


Figure 5.8: Comparison in terms of average accuracy for facial expression recognition by 22-D geometric and 31-D appearance features with deep network when using the SOM-based classifier MMI Database.

VD-MFP database:

- Table 5.3 displays the contrast in terms of average accuracy for facial expression recognition by 22-D geometric facial features with the deep network when using the SOM-based classifier. In terms of the average accuracy for geometric features, it is (94.1%), that is including the highest recognition accuracy of (98.9%) in happiness class and the lowest accuracy of (90%) with disgust class, which is confused with fear class giving (91.1%).

Table 5.3: Confusion matrix showing evaluation results of 22D-geometric features with VD-MFP Database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**95.6**	2	4.44	0	0	0
Disgust	0	**90**	4.44	3.33	1.11	1.11
Fear	2.22	1.11	**91.1**	0	4.44	1.11
Happy	0	0	0	**98.9**	1.11	0
Sadness	0	1.11	3.33	0	**92.2**	3.33
Surprise	0	1.11	2.22	0	0	**96.7**

- In contrast Table 5.4 represents the 31-D appearance features with the deep network when using the SOM-based classifier yield (90.2%) with the highest recognition accuracy of the best class being (100%) in the case of happiness class and the lowest accuracy of (80%) with the disgust class.

Table 5.4: Confusion matrix showing evaluation results of 31-Appearance features by PCA with VD-MFP Database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**83.33**	8.88	1.11	0	6.66	0
Disgust	15.11	**80.0**	1.11	1.11	2.22	0
Fear	0	0	**92.2**	0	6.66	1.11
Happy	0	0	0	**100**	0	0
Sadness	3.33	0	5.55	0	**91.12**	0
Surprise	0	0	5.55	0	0	**94.4**

The results of geometric features combined with the deep network when using the SOM-based classifier for VD-MFP database outperformed the rest of the results such as when appearance features were combined with the same deep network as shown on Fig 5.9.

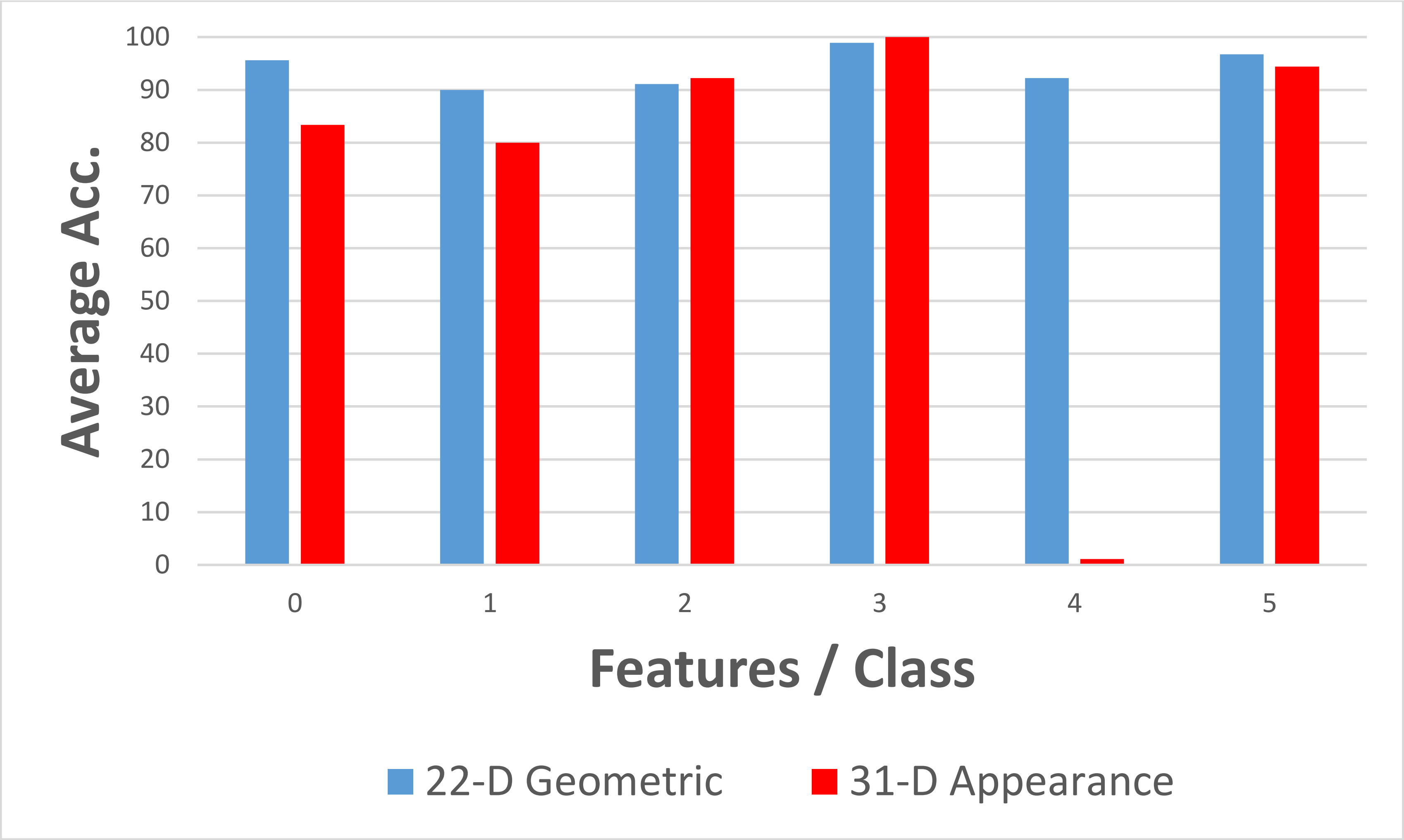


Figure 5.9: Comparison in terms of average accuracy for facial expression recognition by 22-D geometric and 31-D appearance features with deep network when using the SOM-based classifier for VD-MFP Database.

Belfast database:

- Table 5.5 illustrates full descriptions of the confusion matrix with an average accuracy of facial expression recognition by 22-D geometric features with deep networks when using the SOM-based classifier. The average recognition accuracy of the geometric features is (93.7%) with the highest recognition accuracy of the best class being (96.7%) in case of surprise class and the lowest accuracy of (91.11%) with the sadness class.

Table 5.5: Confusion matrix showing evaluation results of 22D-geometric features with Belfast Induced Natural Emotion Database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**92.22**	0	2.22	1.11	0	0
Disgust	2.22	**94.44**	2.22	0	0	1
Fear	0	2.22	**93.33**	0	4	1
Happy	2.22	0	0	**94.44**	0	2.22
Sadness	1.11	3.33	2.22	1.11	**91.11**	1.11
Surprise	2.22	1.11	0	5.55	3.33	**96.7**

- Table 5.6 shows full descriptions of the confusion matrix with an average accuracy of facial expression recognition by 31-D appearance facial features with deep networks when using the SOM-based classifier. The appearance features provide (92.8%) average accuracy with the highest recognition accuracy of the best class being (100%) in case of happy class and the lowest accuracy of (80.00%) with the disgust class.

Table 5.6: Confusion matrix showing evaluation results of 31-Appearance features by PCA with Belfast Induced Natural Emotion Database.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**98.9**	0	1.11	0	0	0
Disgust	0	**80.0**	0	0	16.67	3.33
Fear	0	0	**92.2**	0	0	0
Happy	7.77	0	0	**100**	0	0
Sadness	0	1.11	3.33	0	**95.6**	0
Surprise	0	0	0	0	10	**90.00**

The experimental results of Tables 5.1-5.6 show that the geometric features provide the best performance with the MMI database. Also, the geometric features are better than appearance features in terms of recognition rate because of their higher accuracy in detecting expressions as shown in Fig 5.10.

The three types of databases provide better results in both geometric and appearance features due to exploiting of pre-processing and deep network techniques compared with the state-of-the-art as shown in Table 5.7.

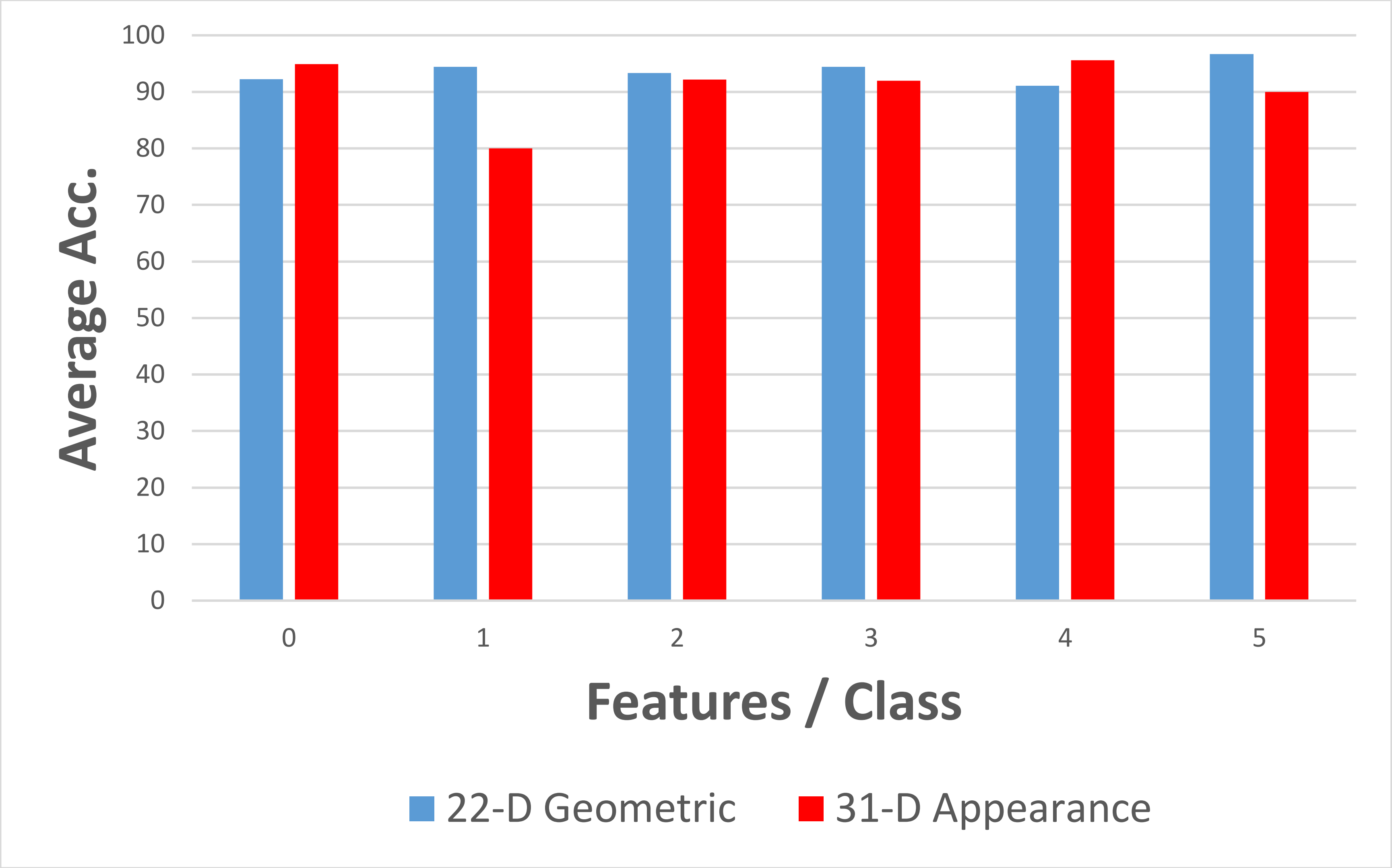


Figure 5.10: Comparison in terms of average accuracy for facial expression recognition by 22-D geometric and 31-D appearance features with deep network when using the SOM-based classifier for Belfast Database.

Table 5.7: Comparison in terms of average recognition accuracy with geometric and appearance features with different techniques

Methods	**Database**	**Appearance**	**Geometric**
[73]	MMI	86.90%	-
[153]	CK+	-	83.01%
[154]	MMI	71.83%	
[63]	MMI	91.27	91.17
Proposed Method	**MMI**	**91.9**	**97.0**
	VD-MFP	**90.2**	**94.1**
	Belfast (BINED)	**92.8**	**93.7**

5.5 Summary

In this chapter, the focus was improving image-based facial expression recognition using automatic geometric feature extraction from raw image data with the deep network and pre-processing steps. It is observed on realistic spontaneous databases that there is a significant improvement in terms of average accuracy by the combination of extracted features with the procedures of the normalization by using the translation, rotation and scaling in comparison with the state-of-the-art methods. The pre-processing procedure has been applied to reduce the difference point of view between these images. The proposed method has been validated for the first time on three databases and two types of features. It is evident that the recognition results using geometric-based features are more accurate than those for appearance-based features. Moreover, in terms of average accuracy, geometric feature extraction and deep network achieved significant performance enhancements in the facial expression recognition.

Although many current studies have achieved improved results in recognition of facial expressions, a challenge remains due to the complexity of human expressions which causes difficulty in the robust discrimination of the particular expression from the extracted features and determination of the type of classifier. The next chapter will introduce a novel method for FER with the spontaneous databases to find an effective solution to reduce the complexity of expressions based on a fusion of three different types of features and multi-stage classification.

Chapter 6

Multi-Stage Classification for Image-Based Spontaneous Facial Expression Recognition

6.1 Introduction

In this chapter, the focus is to overcome the problems of spontaneous image-based facial expression recognition that are mentioned above, in order to obtain a high recognition accuracy. Therefore, a novel system is proposed for spontaneous facial expression recognition using individual video frames in real-time applications. It also employs various spontaneous databases having progressively more significant challenges. To solve the problem of the complexity of facial expressions, the fusion of different functions of features that are mentioned in Chapter 5 is exploited. Also, multi-stage classification is used to boost the final performance; the output is fused for each stage of the classifier together with a hierarchical multi-classifier method to determine the final decision. In summary, the main contributions of the proposed system are:

- The fusion of different types of features comprising geometric features and two sorts of appearance features to provide a rich feature vector by which the best representation of the spontaneous facial features is obtained.
- Reduce the burden of computing by maintaining important location information by concentrating on the crucial roles of the facial regions as the basic processing instead of

the entire face, where the LBP and LQP features are extracted automatically by means of detecting two important regions of the face.

- A deep network consisting of autoencoders and the SOM-based classifier, which employs backpropagation in its training, together with an automatic method for splitting the training effort of the initial network into several networks and multi-classifiers (Surface Network and Bottom Network) are used, in order to solve the problem and to enhance the performance.

- Evaluation on three progressively more challenging databases is performed; namely, MMI, a video database of moving faces and people (VD-MFP) and the Belfast induced natural emotion database (BINED), as described in Chapter 3.

The remainder of the chapter is structured as follows. Section 6.2 introduces the pre-processing system that consists of automatic extraction of regional appearance features and appearance feature extraction. Section 6.3 shows the experimental results. Section 6.4 introduces the detailed explanation of the novel method which is hierarchical cascade and addresses the basic idea of the effectiveness of the multi-stage classification to find a solution to the complexity of facial expression recognition. Section 6.5 presents the experimental results. The summary of the overall chapter is represented in Section 6.6.

6.2 Pre-processing system

Automatic Extraction of Regional Appearance Features: In this chapter, the geometric landmark points have been adopted to crop the two key regions of the facial expression as essential components to extract appearance features. Through focusing on important facial regions that have most expressions and removing all the features that are worthless better performance is expected. According to the geometric landmark points, facial expression is divided into two principal parts containing most expressions to cover all the sub-regions in the regions for facial expression recognition, such as the eyes, eyebrows and mouth as follows:

1. ***The region around the eyes***: depending on the $R = \{\mathbf{r}_i\}_{i=1}^{n}$ in Fig. 6.1. Let us assume that the points in the middle of the left and right eyebrows are $\{\mathbf{E}_r, \mathbf{E}_l\} \subset \mathbb{R}^2$. Accordingly, the middle point between them is calculated by $\mathbf{M}_e = \frac{1}{2}(\mathbf{E}_r + \mathbf{E}_l)$. Assume

the right inner corner points and left for the eyes in the expression image are $\{\mathbf{I}_r, \mathbf{I}_l\} \subset \mathbb{R}^2$ respectively. According to this notation, $\mathbf{M}_i = \frac{1}{2}(\mathbf{I}_r + \mathbf{I}_l)$. Let us assume the outer corner points in both eyes are $\{\mathbf{O}_r, \mathbf{O}_l\} \subset \mathbb{R}^2$, respectively.

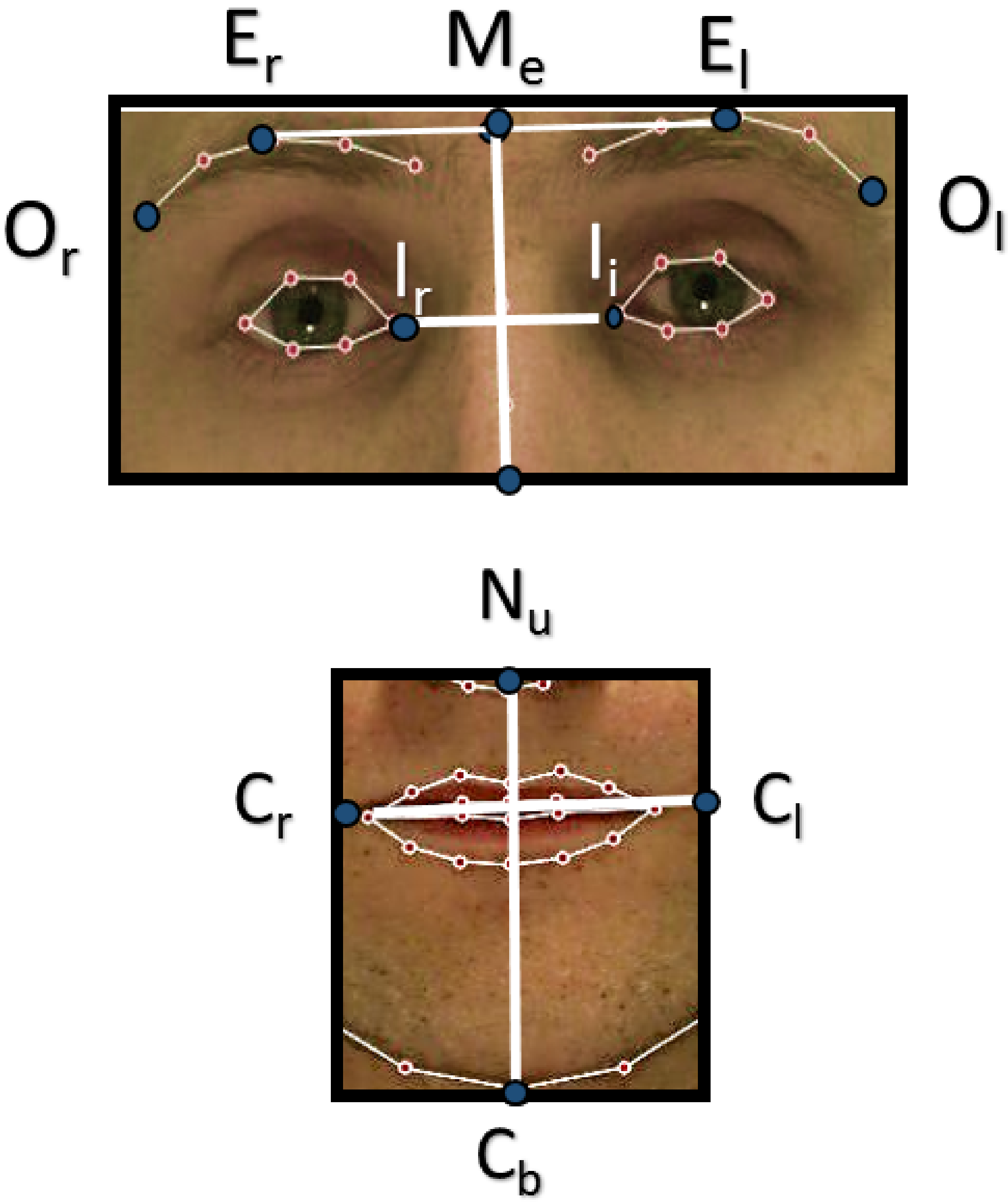


Figure 6.1: Important expressive parts around the eyes and the mouth are automatically cropped

Accordingly, the region surrounding the eyes is cropped according to the following equations:

$$(x_1, y_1) = (x_{O_r},\ y_{M_e}). \tag{6.1}$$

$$(x_2, y_2) = (x_{O_l},\ y_{M_e}). \tag{6.2}$$

$$(x_3, y_3) = (x_{O_r},\ y_{Mi} - |y_{Me} - y_{M_i}|). \tag{6.3}$$

$$(x_4, y_4) = (x_{O_l},\ y_{M_i} - |y_{M_e} - y_{M_i}|). \tag{6.4}$$

2. ***The region around the mouth***: Let us assume $\{\mathbf{C}_r, \mathbf{C}_l, \mathbf{N}_u, \mathbf{C}_b\} \subset R^2$, on the assumption that the right and left corner points of lips are $\mathbf{C}_r, \mathbf{C}_l$, the middle point of the chin is $\mathbf{C}_b$, and finally, the point under the nose is $\mathbf{N}_u$. We choose the specified points to cover the entire region surrounding the mouth, as shown in Fig. 6.1. Accordingly, the region around the mouth is computed as follows:

$$(x_1, y_1) = (x_{C_r},\ y_{N_u}). \tag{6.5}$$

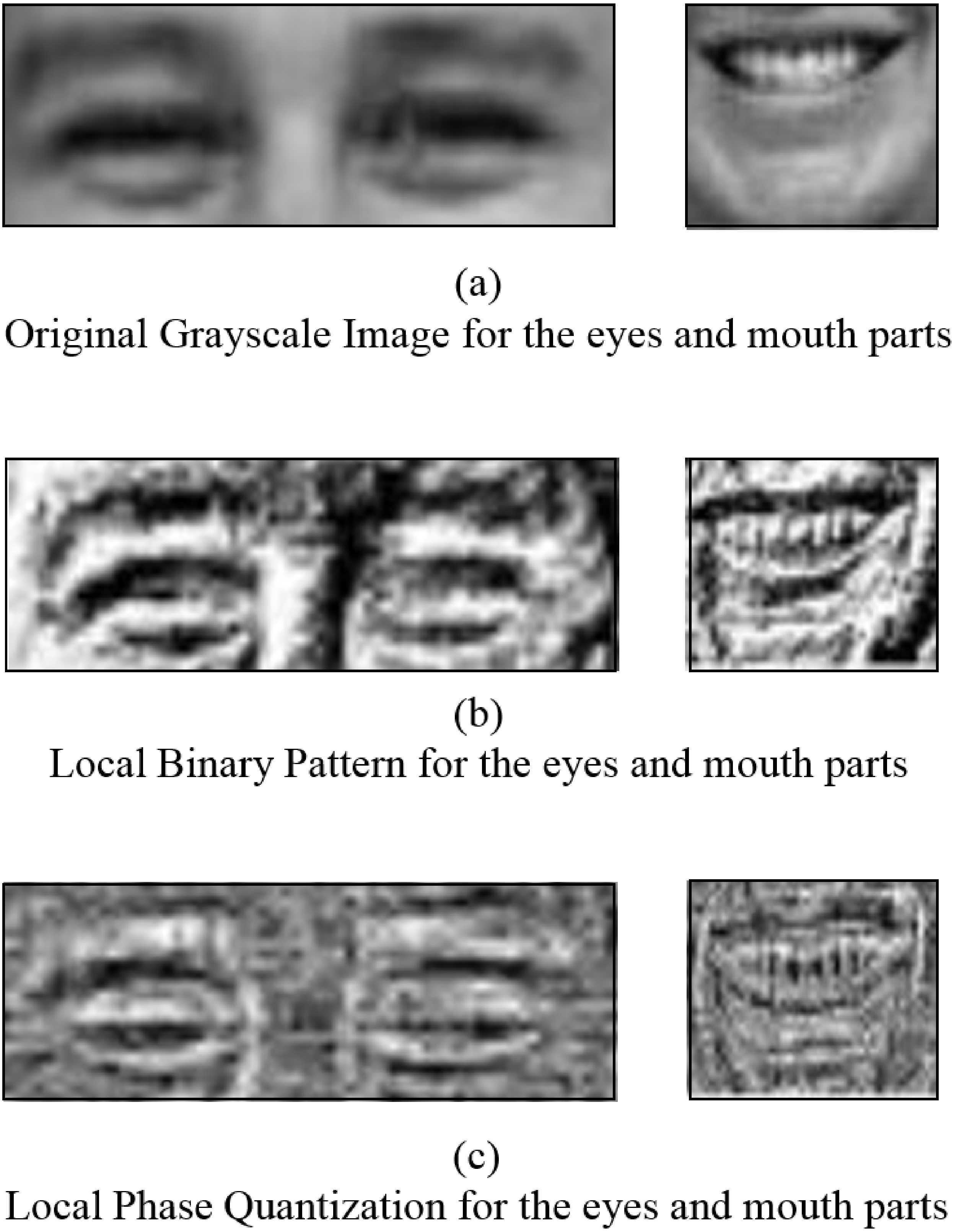
$$(x_2, y_2) = (x_{C_l},\ y_{N_u}). \tag{6.6}$$

$$(x_3, y_3) = (x_{C_r},\ y_{C_b}). \tag{6.7}$$

$$(x_4, y_4) = (x_{C_l},\ y_{C_b}). \tag{6.8}$$

6.2.1 Appearance Feature Extraction

There are different methods to extract local facial features. Two types of local feature-based approaches, namely, local binary patterns (LBP) and local phase quantization (LPQ) are used, which these methods have been explained in detail in Chapter 2. Fig. 6.2 provides an examples of resultant LBP and LPQ images of two basic facial regions. The first row (a) shows the facial regions; second row (b) displays the corresponding LBP images and the third row presents the corresponding LPQ image.



(a)
Original Grayscale Image for the eyes and mouth parts

(b)
Local Binary Pattern for the eyes and mouth parts

(c)
Local Phase Quantization for the eyes and mouth parts

Figure 6.2: Examples of resultant LBP and LPQ images of two basic facial regions. The first row (a) shows the facial regions; second row (b) displays the corresponding LBP images and the third row presents the corresponding LPQ image.

6.3 Experimental Results

In the experiments, three different types of spontaneous databases are selected that include difficult challenges to cover the diverse conditions of imaging, in which, some of these are closer to the wild, namely, MMI facial expression database [58], Video Database of Moving Faces and People (VDMFP) [59], and Belfast Induced Natural Emotion Database (BINED) [60], as described in Chapter 3. The three databases have six basic facial expressions which are anger (AN), fear (FE), happiness (HA), sadness (SA), surprise (SU) and disgust (DI). In the three databases, the alignment problem is addressed for both the geometric and the appearance features by normalizing the rotation around the center point in the horizontal dimension to correct most possible geometric issues, such as translation, rotation, and scaling, which means zooming in or out of the face in the camera view. Then the face images are normalized to a standard scale of 240 × 240 pixels.

The proposed data fusion approach is used to obtain fused feature vector among three types of features: geometric feature and two types of appearance features (LBP and LPQ with auto-encoder deep network). Moreover, the third layer that combines the benefits of both algorithms of the supervised and unsupervised learning by the SOM-based classifier. The average accuracy of facial expression recognition is calculated using 22-D geometric and 256-D appearance feature vectors with auto-encoder deep network which is a deep-network-based automatic facial expression recognition system, where the first two layers of auto-encoders effect on geometric features for better description of the facial data. In addition, the third layer combines the benefits of both supervised and unsupervised learning algorithms by the SOM-based classifier.

In the three spontaneous databases, evaluation results are presented in the form of confusion matrices. The details of the experimental results are presented below.

6.3.1 Experimental Results of the Three Databases

MMI Dataset:

- Table 6.1 shows a comparison in terms of average accuracy for facial expression recognition with three types of fusion with deep network when using the SOM-based classifier as shown in Fig. 6.3. Three methods of fusion are used: the first row includes 22-D geometric and 256-D regional LPQ appearance features, the second, 22-D geometric and 256 -D regional LBP appearance features, then, 22-D geometric with 256-D regional LBP and LPQ appearance features. The best average recognition accuracy fusion of geometric with regional LBP and LPQ appearance features is (98.1%) with highest recognition accuracy of the best class being (100%) in the three cases of anger, disgust and surprise, the lowest accuracy being (95.3%) with happiness class. While the average accuracy of other fusion methods is (97.6%) and (81.1%) for fusion of geometric and appearance (LBP) features, and geometric and appearance (LPQ) features, respectively.

Table 6.1: Accuracy results of the MMI database with 22-D geometric and 256-D appearance features.

Features	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**	**Av.**
GF	100	93.3	96.7	97.8	96.7	96.8	97.0
GF+LPQ	92.8	92.8	72.3	72.3	78.2	78.2	81.1
GF+LBP	100	100	96.3	92.8	96.3	100	97.6
GF+LBP+LPQ	100	100	97.1	95.2	96.3	100	98.1

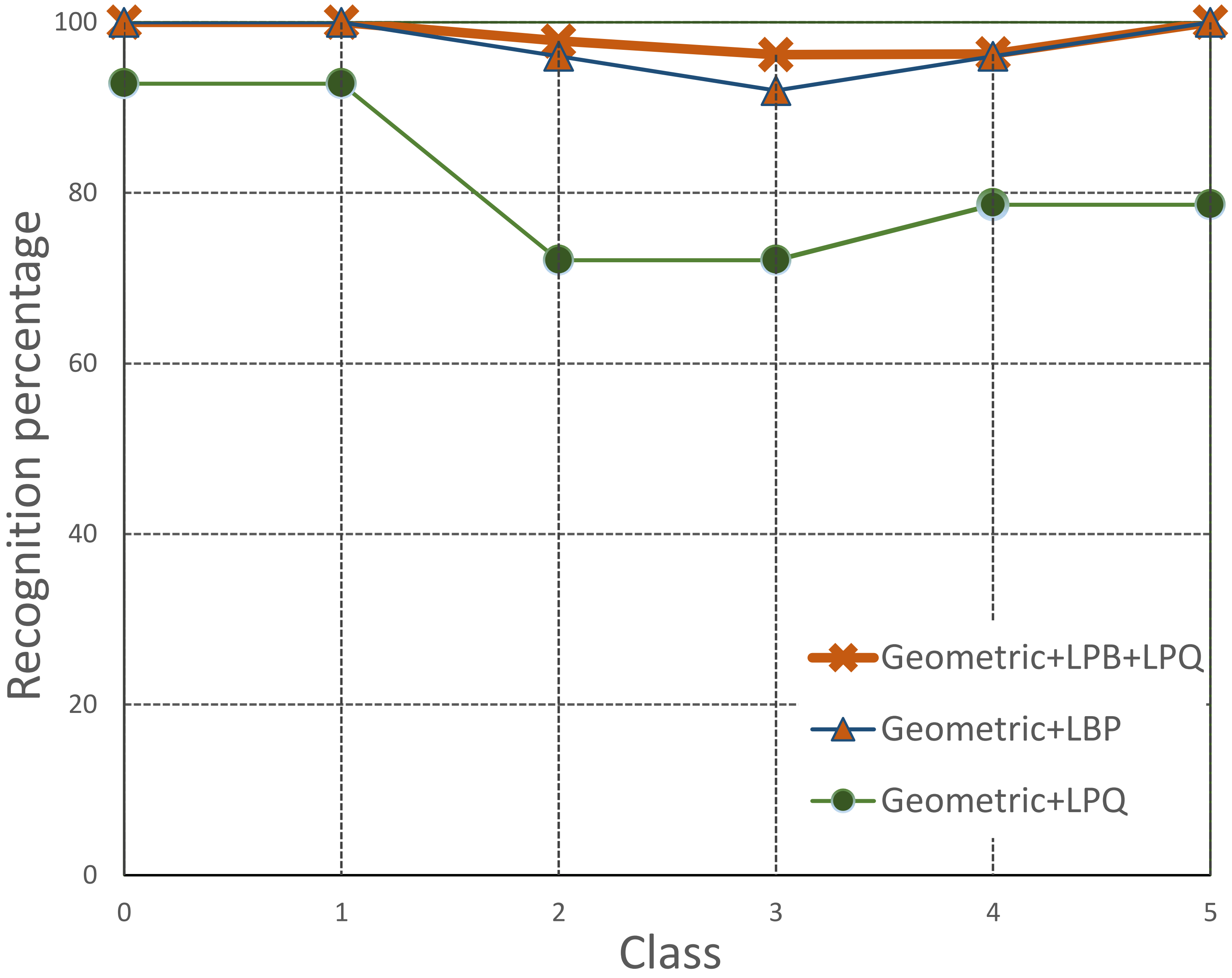


Figure 6.3: Comparison between recognition results of the MMI DB for three different types of fusion features.

VD-MFP Database

- Regarding VD-MFP Database: Table 6.2 displays the contrast in terms of average accuracy for facial expression recognition with the same three types of fusion with deep network when using the SOM-based classifier, as shown in Fig. 6.4. The best average accuracy for fusion of 22-D geometric features with 256-D regional LBP and LPQ appearance features is (96.1%), that includes the highest recognition accuracy of (100%) in anger class and the lowest accuracy of (92.8 %) with sadness class. In contrast, the average accuracy yields (95.1%) and (79.8%) with fusion of 22-D geometric and 256-D regional LBP features and fusion between 22-D geometric and 256-D regional LPQ appearance features, respectively, with the highest accuracy being (98.9%) of anger class and (92.8%) in both classes anger and disgust, in two methods LBP and LPQ respectively.

Table 6.2: Accuracy results of the VD-MFP database with 22-D geometric and 256-D appearance features.

Features	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**	**Av.**
GF	95,6	90	91.1	98.9	92.2	97.6	94.1
GF+LPQ	92.8	92.8	64.6	64.6	78.2	85.7	79.8
GF+LBP	98.8	92.8	93.2	96.3	93.2	96.3	95.
GF+LBP+LPQ	100	95.2	96.3	96.3	92.8	96.3	96.1

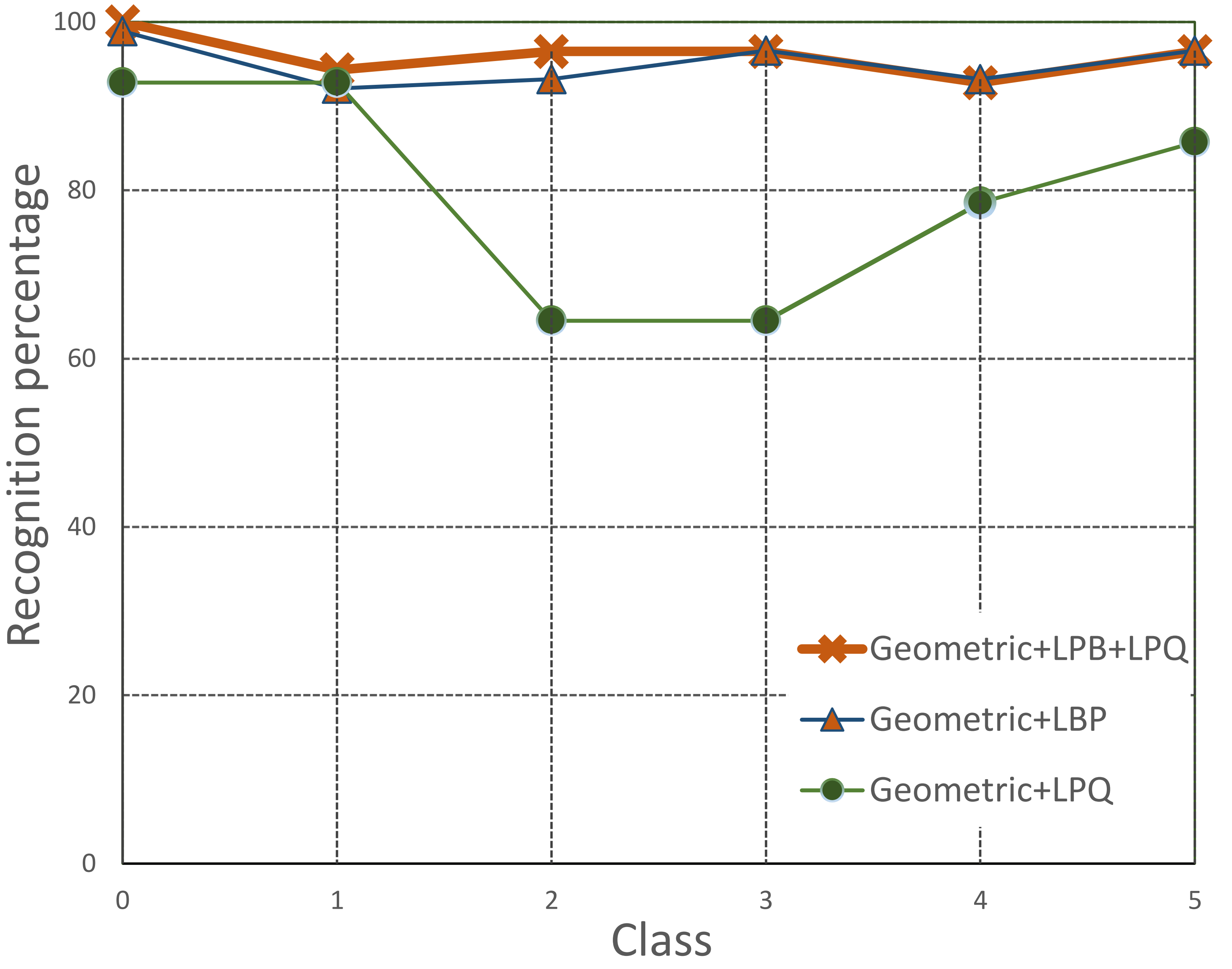


Figure 6.4: Comparison between recognition results of the VD-MFP DB for three different types of fusion features.

Belfast Database:

- Table 6.3 illustrates the variation of average accuracy by using three types of fusion method with the same network and classifier, as shown in Fig. 6.5. The accuracy of fusion of 22-D geometric with 256 -D regional LBP and LPQ appearance features, is (94.8%) and highest recognition accuracy is (100%) with happiness and the lowest accuracy being (92.8%) for the anger class. While fusion of 22-D geometric and 256 -D regional LBP features and fusion between 22-D geometric and 256-D regional LPQ appearance features provide (93.9%) and (80.9%) average accuracy.

Table 6.3: Accuracy results of the Belfast database with 22-D geometric and 256-D appearance features.

Features	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**	**Av.**
GF	92.2	94.4	93.3	94.4	91.1	96.7	93.7
GF+LPQ	85.7	85.7	77.1	77.1	82.8	77.1	80.9
GF+LBP	92.8	94.8	91.7	96.3	92.8	95.2	93.9
GF+LBP+LPQ	92.8	94.8	92.8	100	93.2	95.3	94.8

Although the recognition rate in the fusion of the geometric and regional LBP features has almost a satisfactory result, the work confirmed the robustness and accurate stability for recognition. Superior results are obtained by the proposed system in terms of average recognition accuracy from the fusion of geometric with two types of appearance features (LBP and LPQ) in which the average accuracy rates reach 98.1%, 96.1% and 94.8% in MMI, VD-MFP and Belfast databases, respectively. The experimental results of Tables 6.1-6.3 show that the fusion of geometric features with two types of appearance features provides the best performance with the three types of databases in terms of recognition rate because of their higher accuracy in detecting expressions. It can be seen that the three types of fusion for the MMI database shows the best results in terms of average accuracy of (98.1%) better than the average accuracy percentages of (97.55%) with fusion of two types of methods obtained in [63].

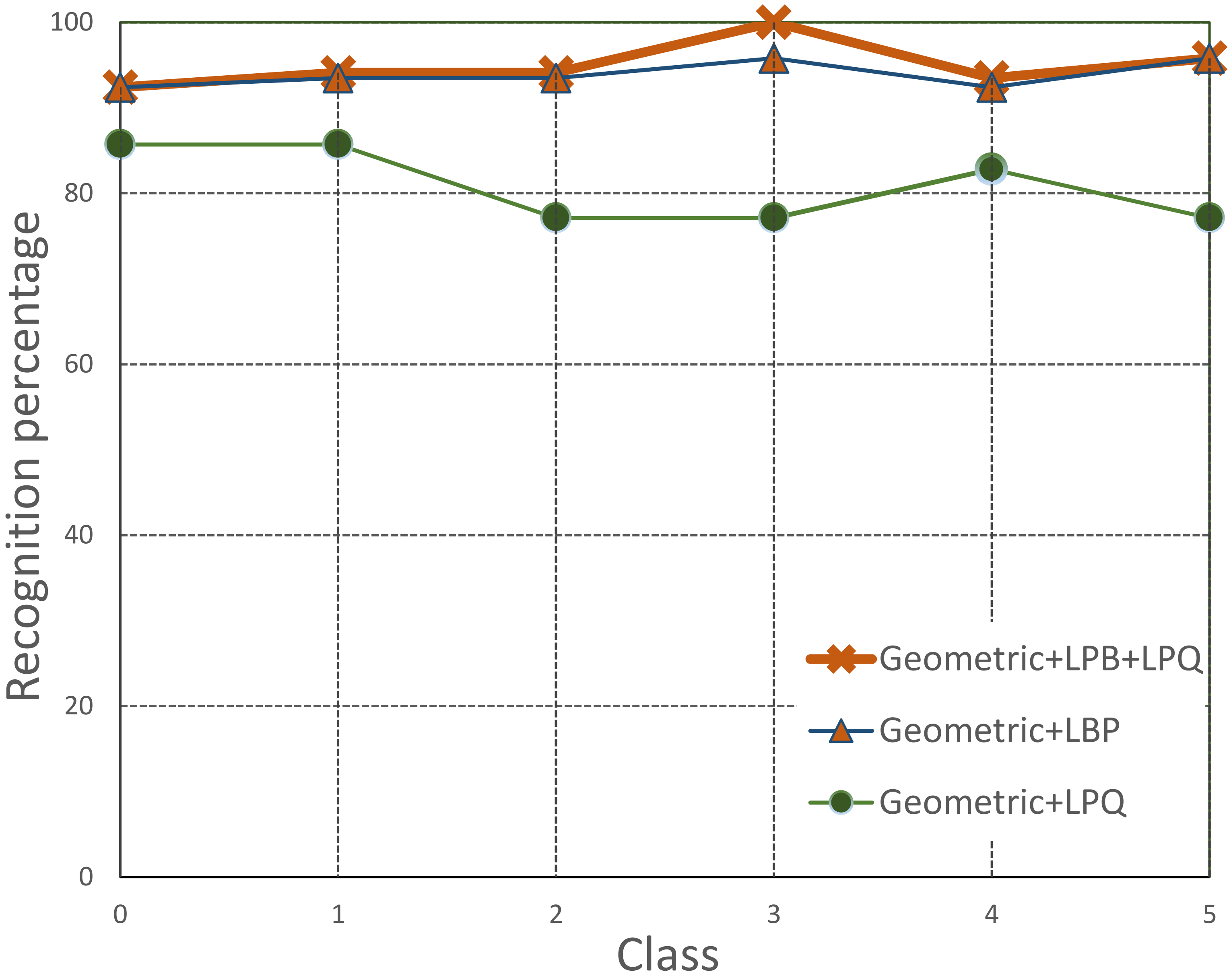


Figure 6.5: Comparison between recognition results of the Belfast DB for three different types of fusion features.

6.4 Problem Statement

Let $\mathbb{D} = \{s_i, l_i\}_{i=1}^{n}$ be the expression dataset containing n samples s_i, each of them associated with an expression label $l_i \in \mathcal{L}$, where $\mathcal{L}$ is the set of considered facial expressions. Let $\mathbb{T}$ and $\mathbb{V}$ respectively be the training and testing subset of $\mathbb{D}$, such that $\mathbb{T} \cap \mathbb{V} = \emptyset$. The final goal of the proposed algorithms is to exploit $\mathbb{T}$ to train a model to automatically predict facial expression labels for samples in $\mathbb{V}$.

6.4.1 Data Fusion Using Autoencoders

The auto-encoder has been widely used in the learning algorithm for the generative information model. It is an unsupervised machine learning algorithm which is used with efficient coding [148]. Moreover, it employs a backpropagation approach to approximate the target values for the input vector. The principal idea of the auto-encoder is learning the best representation of a feature set in a compressed form. The auto-encoder has three layers represented in Fig 6.6: an input layer, a hidden (encoding) layer and a decoding layer. The network is trained to attempt for learning to reconstruct its input vector $\mathbf{x} = [x_1, x_2, ..., x_5]$, such that it coerces the hidden layer $h_{w,b}(\mathbf{x})$ to attempt to learn the best representations of the input feature vector. The hidden layer representation of x is produced by

$$\hat{h}_j(\mathbf{x}) = f(a_j(\mathbf{x})) \text{ where } a_j(x) = b_j + \sum_k W_{jk} x_k. \tag{6.9}$$

where $\hat{h}_j(\mathbf{x})$ is the hidden layer output j subsequent implementing the activation function $f(\cdot)$, which is a sigmoid function that is generating the output (0 to 1). The reconstructed $(\hat{\mathbf{x}})$ which is acquired by way of using a decoding function as follows:

$$\hat{\mathbf{x}} = g(\hat{ak}) \text{ where } \hat{ak} = Ck + \sum_j W^*_{jk} \hat{h}_j(\mathbf{x}). \tag{6.10}$$

The function $g(\cdot)$ regularly chosen again as a sigmoid function, is used specifically with models that have a high probability of being predicted as outputs. It is experimentally observed [149], after training processing a Restricted Boltzmann Machine (RBM), W^* that converts the equivalent to W^T where $(\cdot)^T$ is the transpose operation. Regarding the network decoding level of the network, W^* is set to W^T, where it is comparable to the RBM [150]. Therefore, the selection of $W^* = W^T$ produces the most appropriate outcomes. Even if the number of these hidden nodes is greater than the number of input nodes, so the parameters represented by b_j and c_k are set to 1.

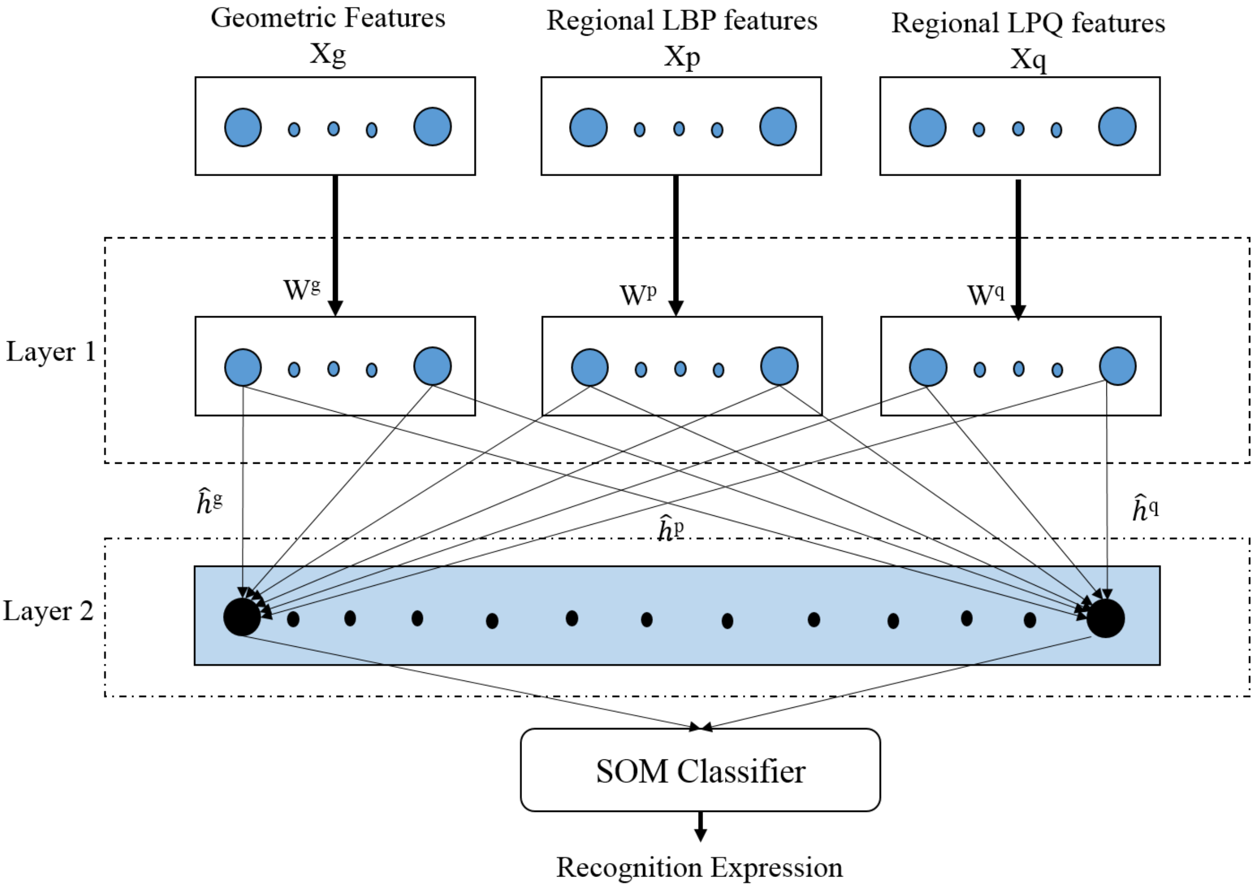


Figure 6.6: Architecture of an auto-encoder with single hidden layer [63].

The proposed system model is trained by three types of features; specifically, geometric features, regional LPQ features and regional LBP features in relation to using auto-encoders. The features of the first layer are concatenated as input into the hidden layer of the auto-encoder. Thus, the auto-encoder attempts to learn the identity of the function approximately, in which the output is similar to the vector of the input. The hidden nodes x are provided by the output of the hidden layer representing the final features of the second layer that is respected as the fusion feature.

Next, in the last phase of the network, classification is used by way of SOM-Based Classifier which is a network for the preservation of topology, such that the neural network discretizes the entire space of features into a specified number of small discrete zones, where each zone is expressed by a single neuron [112]. In the proposed system, the SOM is trained with all features subsequently using auto-encoders to acquire a closer model of the data according to the corresponding expression. The process of the training for the SOM is complementary to the auto-encoder process, such that it produces the final decision in relation to the classification.

6.5 Hierarchical Cascade

The algorithm proposed in this section provides an automatic analysis for splitting the training effort into several networks and multi-classifiers. In particular, this method intends to split the problem into a number of easier problems, defining one network to solve each of them. Thus, the overall processing will be organized as a hierarchical cascade.

Let $\mathcal{S} = \{L_j\}_{j=1}^{m}$ be a set of disjoint subsets of $\mathcal{L}$ such that $\mathcal{L} = \bigcup_{j=1}^{m} L_j$. Let h_j be a unique, temporary label associated with L_j. Therefore, we define $\mathcal{H} = \{h_j\}_{j=1}^{m}$ as the set of temporary labels for the dataset $\mathbb{D}$.

To explicitly define $\mathcal{S}$, let t be a threshold such that

$$t = \max_{p=1,\dots,n} c_{p,p}, \quad c_{p,q} \in CM \tag{6.11}$$

where CM is the confusion matrix obtained by the network $Net_{\mathcal{L}}$ trained on the classification problem based on $\mathcal{L}$. Thus, by defining $C\bar{M} = \frac{CM+CM'}{2}$, a binary matrix B remains defined as follows:

$$B = \begin{cases} 0, & \text{if } \ \bar{c}_{p,q} > 100 - t \\ 1, & \text{otherwise} \end{cases} \qquad \bar{c}_{p,q} \in C\bar{M} \tag{6.12}$$

Since B is a square, symmetric and binary matrix, an undirected graph $G(B)$ can be defined. The graph $G(B)$ is responsible for the explicit definition of the set $\mathcal{S}$. In particular, we define the elements of $\mathcal{S}$ as the m connected components of the graph $G(B)$. In the rare case of single connected component, e.g. when $c_{p,p} = c_{q,q}$ for all $p \neq q$, therefore the $\mathcal{S}$ is defined by using the *max-flow min-cut* algorithm [155].

In practical cases, as those presented in Section 6.6, each element L_j of $\mathcal{S}$ can contain a single label or multiple labels. The case of L_j containing only one label ($|L_j| = 1$), encodes that the $Net_{\mathcal{L}}$ can relatively well distinguish the correspondent expression class from the rest of the classes in $\mathcal{L}$. As opposite, when L_j contains more than one class ($|L_j| > 1$), the $Net_{\mathcal{L}}$ is confusing the action classes within L_j.

Algorithm 2 Proposed hierarchical processing for the classification task.

```
Input: Trained $Net_H$, $Net_{L_j}$ $\forall j = 1, \ldots, m$ such that $|L_j| > 1$, testing samples and threshold $t$.
Output: Estimated labels $l_{est}$ for testing samples.
Initialisation:
while testing samples are available do
    Read current testing sample
    Estimate temporary label $h_j$ by using $Net_H$
    if $|L_j| = 1$ then
        $l_{est} = h_j$
    else if $|L_j| > 1$ then
        Estimate labels $y$ with $Net_{L_j}$
        $l_{est} = y$
    end if
end while
return $l_{est}$
```

According to the above-mentioned arguments, we propose to split the classification problem into a so-called *surface* level and up to m *bottom* levels as follows:

1. *Surface level*: the surface level represents the classification problem defined by labels X. For this problem, a neural network Net_H is trained.

2. *Bottom levels*: up to m levels are defined, one for each set of labels L_j such that $|L_j| > 1$. Therefore, up to m additional networks Net_{L_j} are trained accordingly.

Therefore, the hierarchical processing is defined as in Algorithm 2: Hierarchical-Processing. The idea of the hierarchical cascade is to split the effort undertaken for the classification among several networks, according to the definition of $\mathcal{S}$.

In particular, part of the classification problem is filtered out by Net_H, providing the temporary labels to input to the bottom levels. This implies that the bottom level problems are more focused on a specific subset of classes. Thus, bottom level networks are not necessarily required to be as deep as Net_H or $Net_{\mathcal{L}}$ in terms of hidden layers. This allows reduction in complexity for the bottom networks when such complexity is not required, preventing overfitting [156] to training data.

6.6 Experimental Results

Experiments are performed on three kinds of spontaneous databases that include complex challenges to cover the diverse conditions of imaging. Several of these that are closer to the wild are the MMI facial expression database [58], Video Database of Moving Faces and People (VDMFP) [59], and Belfast Induced Natural Emotion Database (BINED) [60]. Fig. 6.7 provides an example of apex frames of image sequences in $\mathbb{D}$ for different expressions of the BINED database. The total number of images used are 1620, which divided equally upon the three databases with six basic expressions defined previously. In three databases (DBs), the alignment problem is addressed for frames of reference and expression images by normalising the rotation around the centre point in the horizontal dimension to correct possible geometric issues, such as translation, rotation and scaling. Thus the facial images are normalized via a standard scale of 240 × 240 pixels. All the experiment upon three databases which mentioned above are performed by using tenfold cross-validation. Each dataset is partitioned into ten sets with equal sized, where, one set is as testing and nine sets as a training set, then repeated this process for all the ten sets. The details of the experimental results for three DBs are presented below.

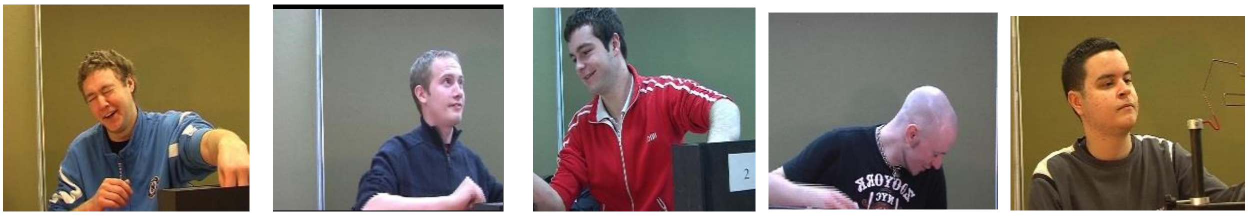


Figure 6.7: Samples of apex frames of image sequences are represented in (a, b, c, d, and e) for different expressions of the Belfast Induced Natural Emotion Database [60]

6.6.1 Experimental Results of the Three DBs

The technique of data fusion presented in Section 6.4.1 has been employed to fusion three types of features to obtain a fused feature vector. The number of hidden units in the first layer of the autoencoder is set to 20 for geometric feature, and 250 for both LBP and LPG for three types of databases mentioned above while the number of hidden units in the second layer is set to 260 for three types of databases. The evaluation results are presented in the form of confusion matrices that demonstrate the results with one classifier and multi-classifier using the aforesaid methods in the above sections on three types of spontaneous DBs, as follows:

- **MMI Dataset:** The confusion matrix of the MMI DBs in Table 6.4. shows the recognition result of (98.1%) by fusion of the three types of features consisting of the 22-D geometric with 256-D regional LBP and LPQ appearance features by auto-encoder deep network with the SOM-based classifier. The proposed algorithm provides an automatic analysis to split the training effort into two levels; specifically, the surface level and bottom level, in each stage. Each level has a network and a classifier. Fig. 6.8. reveals the split of training effort for the MMI DB to the Surface Network and Bottom Network. The steps of the hierarchical algorithm are as follows:

Table 6.4: Confusion matrix showing evaluation results of of the 22-D geometric with 256-D regional of LBP and LPQ appearance features by using the auto-encoder with the SOM for over all MMI DB.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**100**	0	0	0	0	0
Disgust	0	**100**	0	0	0	0
Fear	0	1.45	**97.10**	0	1.45	0
Happy	0	3.35	0	**95.2**	1.45	0
Sadness	0	1.45	2.25	0	**96.3**	0
Surprise	0	0	0	0	0	**100**

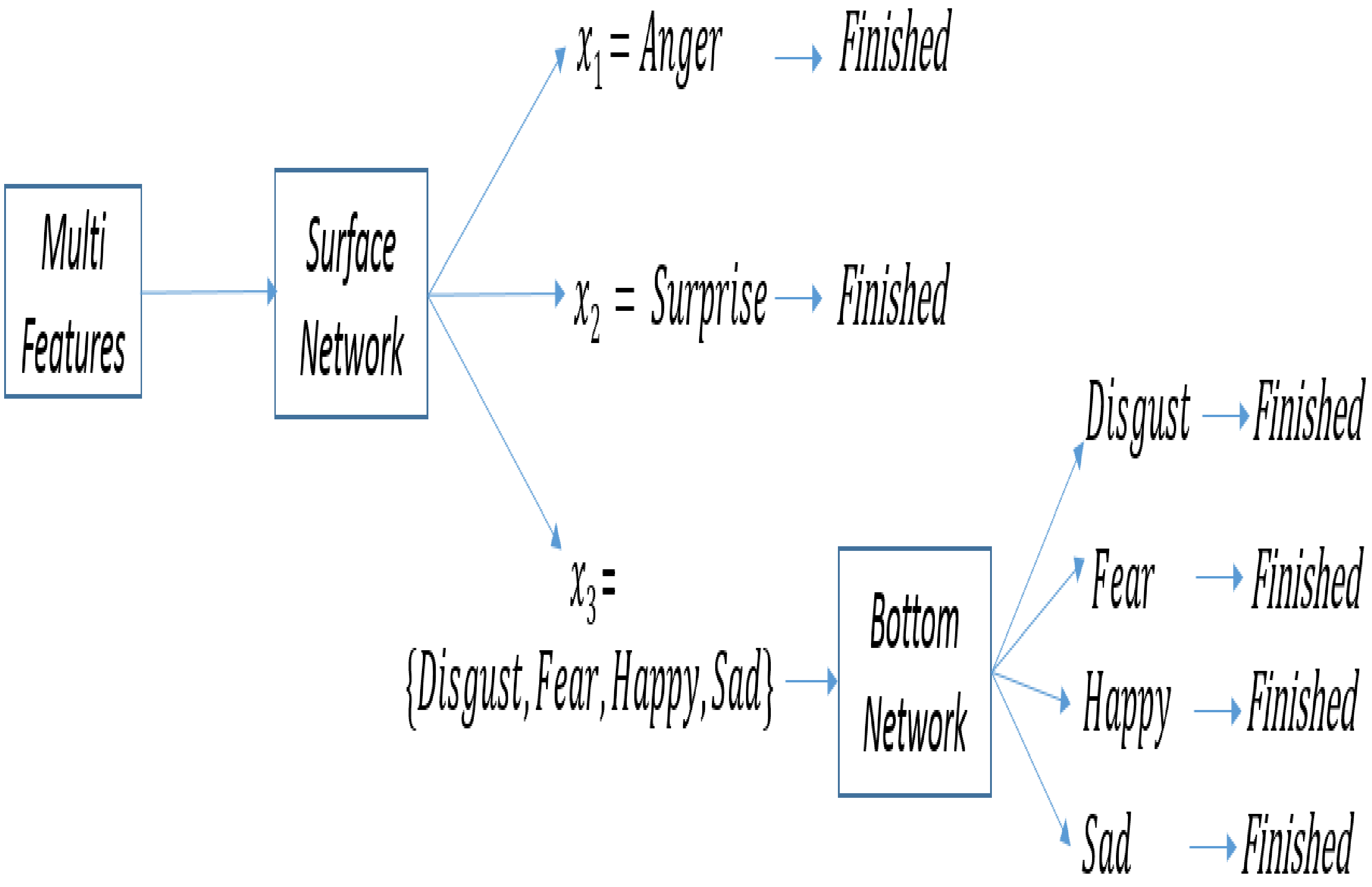


Figure 6.8: The split of training effort for the MMI DB to the Surface Network and Bottom Network.

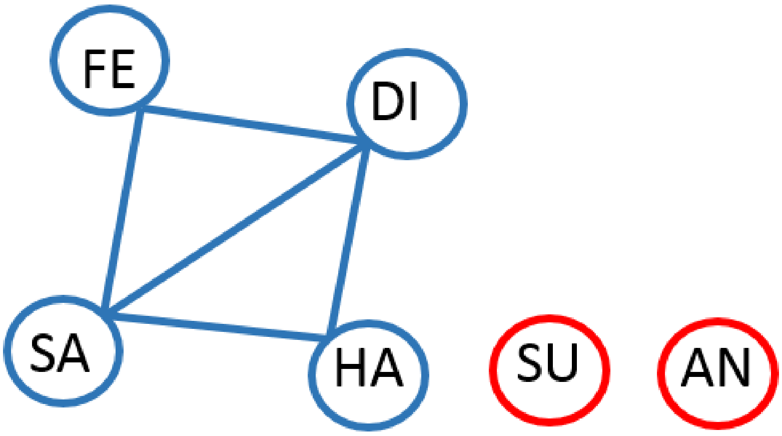


Figure 6.9: The split case of training effort for MMI DB to Surface Network.

Table 6.5: The CM of the surface level of the MMI DB using the first network of the auto-encoder with the SOM-based classifier.

Surface Level	x_1	x_2	x_3
x_1	**100**	0	0
x_2	0	**100**	0
x_3	0	0	**100**

[**1.1.**] ***Surface level***: the Hierarchical Cascade in Section 6.5. suggests defining

$$L_1 = \{\text{Anger}\} \to x_1$$
$$L_2 = \{\text{Surprise}\} \to x_2$$
$$L_3 = \{\text{Disgust}, \text{Fear}, \text{Happiness}, \text{Sadness}\} \to x_3$$
$$X = \{x_1, x_2, x_3\}$$

Table 6.4. shows the CM for $Net_{\mathfrak{L}}$ for the MMI DB and Fig. 6.9. displays the subsequent $G(B)$. Table 6.5. illustrates the Net_H CM which represents the surface level for this dataset.

[**1.2.**] ***Bottom Level:*** since only L_3 contains more than one class, only one bottom level is required. Therefore, a second classifier Net_{L_3} is trained on classes $\{\text{Disgust}, \text{Fear}, \text{Happiness}, \text{Sadness}\}$. Table 6.6. illustrates the obtained CM.

Table 6.6: The CM of the bottom level of the MMI DB using the second network of the auto-encoder with the SOM-based classifier.

Bottom Level	**Disgust**	**Fear**	**Happy**	**Sadness**
Disgust	**100**	0	0	0
Fear	0	**96.67**	0	3.33
Happy	0	0	**100**	0
Sadness	0	0	0	**100**

Fig. 6.8. shows the Hierarchical Cascade for the MMI DB. The CM obtained by the Hierarchical Cascade is shown in Table 6.7. and represents the final form of the six expressions based processing. The Hierarchical Cascade improves the baseline performance in terms of average accuracy from 98.10% to 98.96%.

Table 6.7: The final CM of the MMI DB after apply the Hierarchical Cascade.

Expressions	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	***100***	0	0	0	0	0
Disgust	0	***100***	0	0	0	0
Fear	0	0	***98.55***	0	1.45	0
Happy	0	0	0	***96.67***	3.33	0
Sad	0	0	1.45	0	***98.55***	0
Surprise	0	0	0	0	0	***100***

2. **VD-MFP DB:** Concerning the VD-MFP DB: Table 6.8. shows the confusion matrix of the VD-MFP DB with average recognition accuracy of (96.16%) for the fusion features of the 22-D geometric with 256-D regional LBP and LPQ appearance features using the auto-encoder deep network and the SOM-based classifier.

Table 6.8: Confusion matrix showing evaluation results of of the 22-D geometric with 256-D regional of LBP and LPQ appearance features by using the auto-encoder with the SOM for over all VD-MFP DB.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**100**	0	0	0	0	0
Disgust	0	**95.22**	0	3.33	0	1.45
Fear	0	0	**96.30**	0	2.25	1.45
Happy	0	1.45	0	**96.30**	0	2.25
Sadness	0	1.45	5.75	0	**92.80**	0
Surprise	0	0	1.45	2.25	0	**96.30**

When we apply the proposed algorithm to split the training effort for the VD-MFP DB, we obtain the following results.

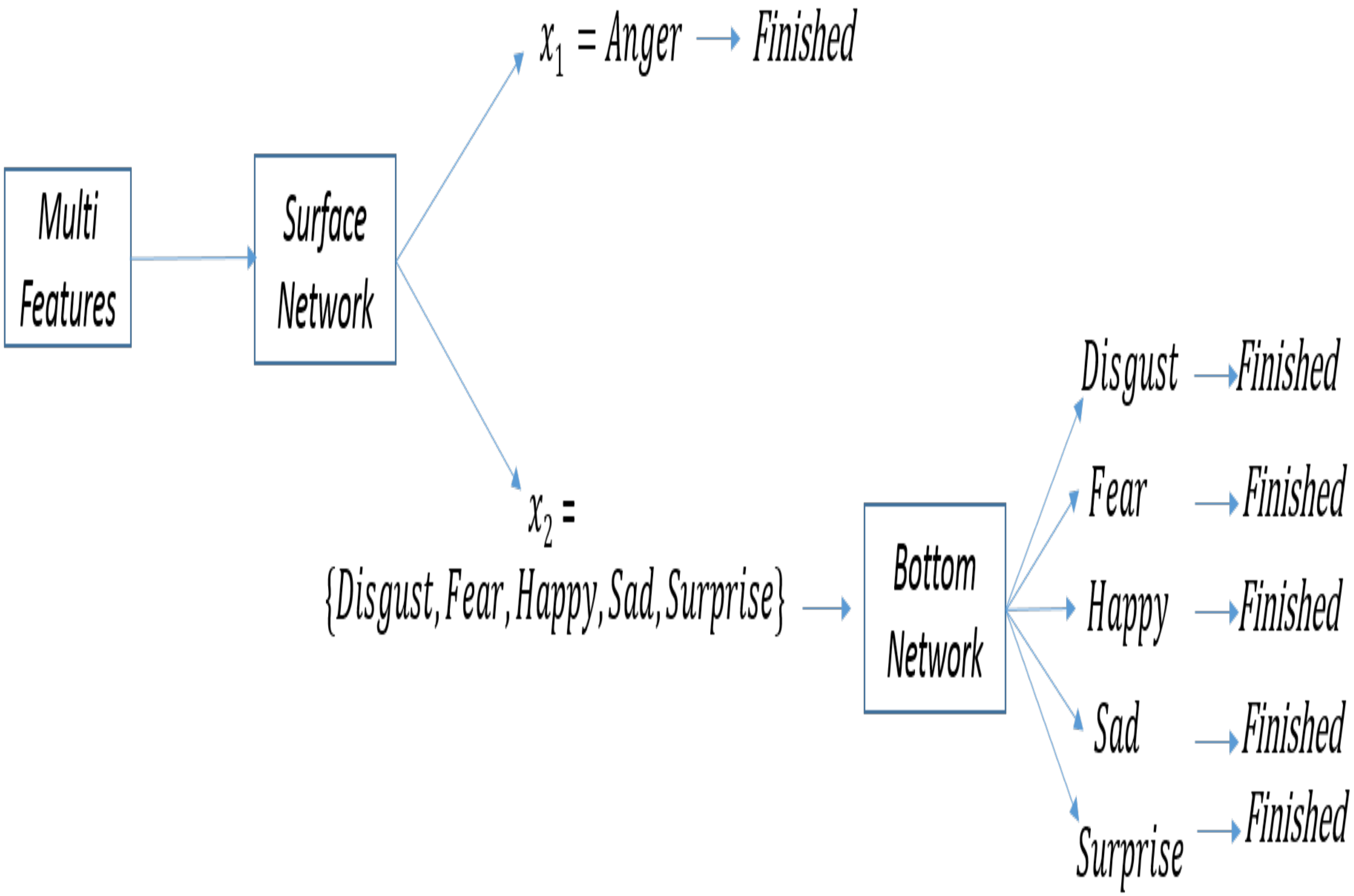


Figure 6.10: The split of training effort for the VD-MFP DB to the Surface Network and Bottom Network.

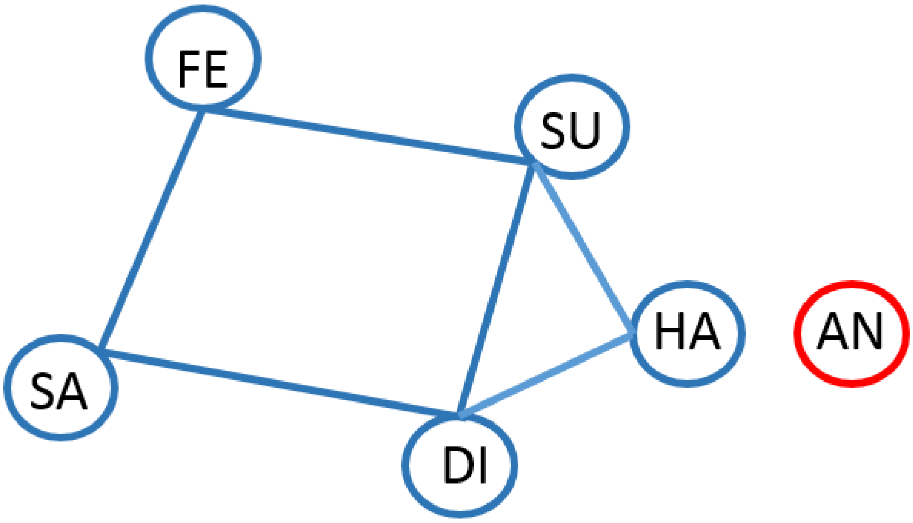


Figure 6.11: The split case of training effort for the MMI database to Surface Network.

[2.1.] ***Surface level*: the Hierarchical Cascade algorithm proposed defining**

$$L_1 = \{\text{Anger}\} \rightarrow x_1$$

$$L_2 = \{\text{Disgust}, \text{Fear}, \text{Happiness}, \text{Sadness}, \text{Surprise}\} \rightarrow x_2$$

$$X = \{x_1, x_2\}$$

Table 6.8. displays the CM for $Net_{\mathfrak{L}}$ for the VD-MFP DB and Fig. 6.11. shows the subsequent $G(B)$. Table 6.9. illustrates the Net_H CM which represents the surface level for this dataset.

Table 6.9: The CM of surface level for the VD-MFP BD using the first network of the auto-encoder with the SOM-based classifier.

Surface Level	x_1	x_2
x_1	**100**	0
x_2	0	**100**

[**2.2.**] ***Bottom Level:*** given that L_2 contains more than one class, therefore, only one bottom level is required. Accordingly, a second classifier Net_{L_2} is trained on classes $\{\text{Disgust}, \text{Fear}, \text{Happiness}, \text{Sadness}, \text{Surprise}\}$.

Table 6.10: The CM of the bottom level for the VD-MFP DB.

Bottom Level	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Disgust	***96.67***	0	3.33	0	0
Fear	0	***98.55***	0	1.45	0
Happy	1.45	0	***98.55***	0	0
Sadness	0	5.75	0	***94.25***	0
Surprise	0	0	0	0	***100***

Table 6.10. displays the bottom level in CM for the VD-MFP DB. Fig. 6.10. demonstrates the Hierarchical Cascade with the training effort for the classifier to the Surface Network and Bottom Network for the VD-MFP database. Thus, we note that the CM obtained by the Hierarchical Cascade in terms of accuracy of multi-classifying is displays in Table 6.7. and represents the final form of the six expressions based processing. The Hierarchical Cascade is a significant improvement compared with the baseline performance in terms of average accuracy from 96.16% to 95.64%.

Table 6.11: The final CM of the VD-MFP DB after apply the Hierarchical Cascade.

Expressions.	***Angry***	***Disgust***	***Fear***	***Happy***	***Sad***	***Surprise***
Ange	***100***	0	0	0	0	0
Disgust	0	***95.55***	0	4.45	0	0
Fear	0	0	***96.67***	0	3.33	0
Happy	0	2.25	0	***97.75***	0	0
Sad	0	0	6.25	0	***93.75***	0
Surprise	0	0	0	1.45	0	***98.55***

3. ***Belfast Database*:** Finally, we evaluate the third database which is termed the Belfast Database: Table 6.12. presents the confusion matrix for the recognition result of (94.80%) by the fusion of the 22-D geometric with 256-D regional LBP and LPQ appearance features for the Belfast database by way of employing auto-encoder deep network and the SOM-based classifier.

Table 6.12: The CM of the Belfast DB with 6 class expression recognition of the 22-D geometric with 256-D regional of LBP and LPQ appearance features (%) using the auto-encoder with the SOM-based classifier over all.

Expression	**Angry**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Angry	**92.80**	5.75	1.45	0	0	0
Disgust	3.75	**94.80**	0	0	1.45	0
Fear	0	1.45	**92.80**	0	5.57	0
Happy	0	0	0	**100**	0	0
Sadness	2.25	0	4.75	0	**93.00**	0
Surprise	0	0	4.75	0	0	**95.25**

By applying the proposed algorithm, the following suggested results are obtained.

[**3.1**] ***Surface level:*** the Hierarchical Cascade algorithm proposed based on the original result of the CM for the Belfast DB.

$$L_1 = \{\text{Happiness}\} \rightarrow x_1$$
$$L_2 = \{\text{Anger}, \text{Disgust}, \text{Fear}, \text{Sadness}, \text{Surprise}\} \rightarrow x_2$$
$$X = \{x_1, x_2\}$$

Table 6.12. shows the CM for $Net_{\mathcal{L}}$ for the Belfast DB and Fig. 6.13. shows the subsequent $G(B)$. Table 6.13. illustrates the Net_H CM which represents the surface level for this dataset.

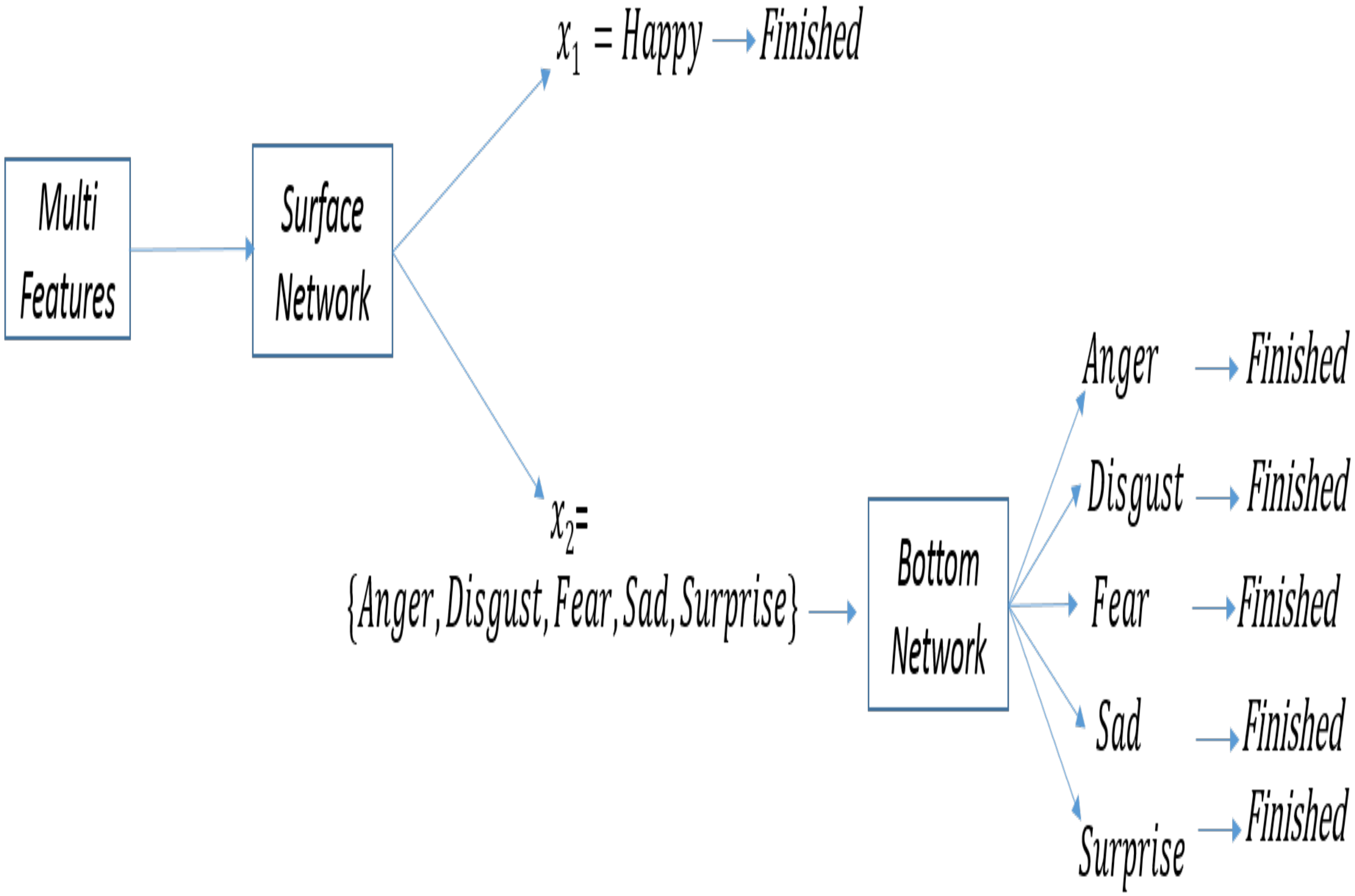


Figure 6.12: The split of training effort for the Belfast database to the Surface Network and Bottom Network.

By applying the proposed algorithm, the following suggested results are obtained.

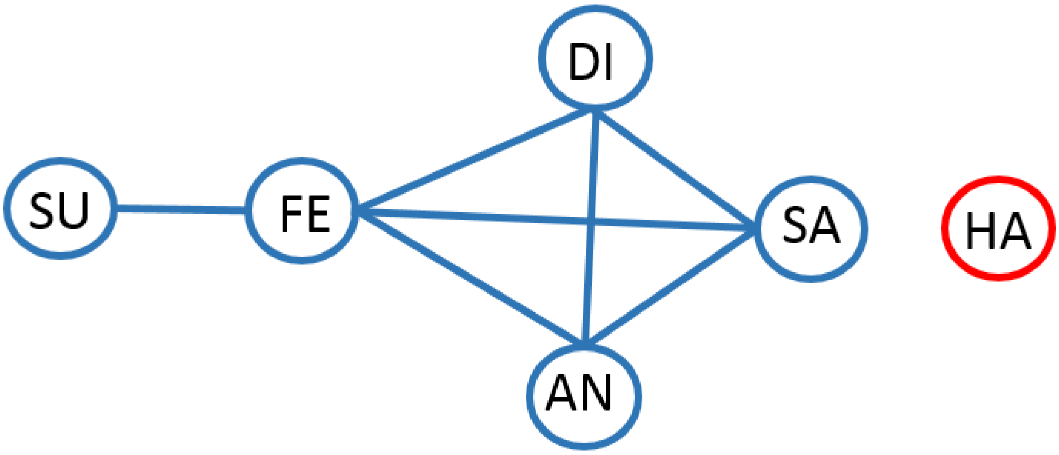


Figure 6.13: The split case of training effort for the Belfast database to Surface Network.

[3.2] ***Bottom Level:*** considering that L_2 contains more than one class, such that only one bottom level is required. Thus, a second classifier Net_{L_2} is trained on classes $\{\text{Anger}, \text{Disgust}, \text{Fear}, \text{Sadness}, \text{Surprise}\}$. Table 6.14. pretensions the bottom level in CM for the Belfast DB.

Table 6.13: The CM of surface level for the Belfast BD using the first network of the auto-encoder with the SOM-based classifier.

Surface Level	x_1	x_2
x_1	**100**	0
x_2	0	**100**

Experimentally, we evaluate the system by applying the Hierarchical Cascade algorithm proposed for the test images of the Belfast database by passing it through the Surface and Bottom Networks. We obtain the following recognition results (100%) and (96.16%) respectively. Fig. 6.12. proves the split of training effort of the Belfast database to the Surface Network and Bottom Network. The recognition result is achieved in terms of the confusion matrix with average accuracy being (95.64%) for all six classes, as shown in Table 6.15.

Table 6.14: The CM of the bottom level for the Belfast DB using the second network of the auto-encoder with the SOM-based classifier.

Bottom Level	**Angry**	**Disgust**	**Fear**	**Sadness**	**Surprise**
Angry	**94.42**	3.33	2.25	0	0
Disgust	3.33	**96.67**	0	0	0
Fear	0	0	**94.25**	5.75	0
Sadness	2.25	0	2.25	**95.50**	0
Surprise	0	0	0	0	**100**

Table 6.15: The final CM of the Belfast DB after apply the Hierarchical Cascade.

Expressions	Angry	Disgust	Fear	Happy	Sadness	Surprise
Angry	***93.75***	6.25	0	0	0	0
Disgust	4.75	***95.25***	0	0	1.45	0
Fear	0	0	***93.75***	0	6.25	0
HA	0	0	0	***100***	0	0
SA	0	2.25	0	3.33	***94.42***	0
SU	0	0	0	3.33	0	***96.67***

Table 6.16: Comparison in terms of average recognition accuracy with geometric and appearance features with different techniques.

Reference	**MMI Database**	**VD-MFP Database**	**Belfast Database**
[157]	70.31%	-	-
[63]	97.55%	-	-
[158]	98.1%	96.1%	94.9%
Proposed Method	**98.96**%	**97.05**%	**95.64**%

The practical results of the recognition accuracy rate for the fusion of the three types of features are geometric and regional LBP features provide a satisfactory result in three types of databases. The experimental results of our proposed system confirmed the robustness of the novel algorithm and the recognition accuracy, which illustrates that overall performance improves when we enhance the single classifier performance by splitting to multi-classifiers. Superior results are acquired with our proposed system in terms of average accuracy of recognition by splitting the training effort of the classifier into multi classifiers according to the required case, where the average recognition accuracy rates reach in Tables (4, 8 and 12) 98.96%, 97.05% and 95.64% in MMI, VD-MFP and Belfast databases, respectively. Our system provides better results in term of fusion and multi classifiers for three types of databases compared with the state-of-the-art, as shown in Table 6.16.

6.7 Summary

This chapter has presented a novel strategy to determine an effective solution for image-based spontaneous facial expression recognition based on a fusion of three types of features, including geometric and appearance features which are local phase quantization (LPQ), and local binary pattern (LBP) and geometric features. Likewise, an auto-encoder deep network is used as the basic classifier. In addition, a novel strategy with multiple typical features and classifiers system is proposed to improve the performance, such that automatic analysis is used to split the training effort into two levels; namely, the surface level and bottom level, in each stage, and each level has a network and a classifier. Higher accuracy of recognition of facial expressions is thereby achieved.

Furthermore, the fusion approach with multi-feature descriptors has a higher recognition rate than any single feature descriptor. Each part of the face has a different contribution with various facial expressions. Consequently, the algorithm based on the multi classifiers provides the most appropriate solution, as well as more stable facial expression recognition. The experiment results confirm that the strategy of multi-classification outperforms the strategy of single classification and moreover, attains higher recognition performance for facial expression methods with a comparison of some state-of-the-art methods, including a deep network solution.

Chapter 7

Conclusions and Future Work

This chapter presents a summary of the major contributions in the thesis followed by discussions about limitations and suggestions for future work.

7.1 Conclusion and Main Findings

In this thesis three issues are mainly investigated regarding building robust image-based FER systems with high performance:

- Exploring the issue of image-based expression recognition by identity-independent images and comparing the performance by using different types of databases (posed database CK+ and three types of spontaneous databases) each having different challenges. The study also aimed to show which types of spontaneous conditions are more challenging in terms of system accuracy.

- Improving image-based facial expression recognition employing automatic extraction of the geometric features from raw image data with pre-processing steps and a deep network for the classifier, and also comparing the performance of different types of features (appearance and geometric features).

- Finding an effective solution for image-based spontaneous facial expression recognition based on a fusion of three types of features and multi-stage classification.

Accordingly, three main contributions were derived from those investigations:

1. Recent work in recognition of naturalistic expressions, which is also known as spontaneous facial expression recognition, has attracted researchers' attention due to its

importance in different behavioural and clinical applications. The main design challenges in the area of emotion computing for automatic recognition of spontaneous facial expression are the face pose, capture distance, illumination variation, head rotation, and occlusion. Therefore, designing a robust system to mitigate these challenges is essential for real-time applications.

Thus, the target of facial expression recognition is to analyse a specific image (the focus of this thesis) or a set of frames of video to detect an individual's emotion, thereby producing more natural and smarter interaction between human beings and computers. A study was presented of the identity-independent image-based expression recognition problem.

The experimental results were obtained using a Sparse Representation Classifier (SRC) with Principal Component Analysis (PCA) and Fisher Linear Discriminant Analysis (FLDA) methods with the posed database CK+ and, for the first time, three progressively more challenging spontaneous databases (MMI, VD-MFP and Belfast). The study yielded two main findings. First, the recognition results with the difference images were higher than with the original images because the difference image emphasises the expression regions in the face and eliminates the unnecessary parts.

Therefore, training on difference images rather than the original images yields better accuracy in both posed and spontaneous databases.

The evaluation results of the posed database CK+ using difference images with PCA + SRC and FLDA + SRC were (82.97%, 94.63%) respectively, better than with the original images by the same methods and the database which were (79.45%, 90.93%).

For the MMI database experiment, the robustness and stability of this work obviously appear for the emotions when using the difference images with both PCA and FLDA combined with the SRC based classifier. The performance of the original images in the MMI database when using PCA & SRC or FLDA & SRC, had a lower recognition accuracy (65.00%, 69.63%) compared to using the difference images, which gave (67.59% and 72.52%) respectively.

While, for the VD-MFP database, the performance of different images by using the PCA & SRC or FLDA & SRC, had recognition accuracy (76.48%, 81.11%) respectively, which were better than using the original image, as shown in Tables 4.9 & 4.12 (68.34%, 72.42%) respectively.

For the results of the Belfast database, the performance of the difference images with PCA & SRC were (59.62%) and with FLDA & SRC, it was (65.55%), which were therefore better than the performance of the original image with the same methods (57.59% and 64.07%) respectively.

Next, for the different types of spontaneous databases, the accuracy of facial expression recognition was found to depend on the nature of the database in terms of illumination and background found in a comparison of performance for image-based expressions with three varieties of spontaneous databases.

2. Feature extraction and selection are significant operations to improve the recognition accuracy of facial expression systems. The distribution of geometric features and their number plays a decisive role in the quality of the process of image matching, particularly for some databases which have more challenges in terms of system accuracy. This study exploited a robust system to mitigate these challenges as this is essential for real-time applications.

   Geometric feature extraction automatically from raw data was concentrated on one of the most attractive methods for classification in the field of neural networks, namely a deep network. The improved system consisted of the following: solving the misalignment problem of the training images, in which a significant improvement in terms of average accuracy was observed on realistic and spontaneous databases by using the combination of extracting features and the normalisation processes, which consisted of mitigating translation, rotation, and scaling. The pre-processing procedure was applied to reduce the difference viewpoint between the images.

   Subsequently, geometric features were extracted with low complexity by employing the DRMF algorithm. Finally, a deep network represented by auto-encoders was introduced based on the AFER system. The auto-encoder has three layers: an input layer, a hidden (encoding) layer and a decoding layer. The network is trained to learn to reconstruct its input vector thereby forcing the hidden layer to attempt to learn the best representations of the input feature vector. The performance of the image-based expression recognition was evaluated on the same three spontaneous databases with geometric and appearance-based features for comparison. A deep network with its high-level feature representation outperformed state-of-the-art techniques.

   The performance of the proposed system in terms of average recognition accuracy of (97.0%) and (91.9%) was outperformed in terms of average recognition accuracy of

(91.17%) and (91.27%) for the MMI databases with separate geometric and appearance features as reported in [63]. In addition, recognition performance for the Video Database of Moving Faces and People (VD-MFP) database of (94.1%) and the Belfast Induced Natural Emotion Database (BINED) of (93.7%).

3. In view of the challenge of achieving high accuracy in enhancing recognition, the selection of the most informative features and classifiers remains an important issue. In this study, therefore, a novel approach to finding an effective solution based on a fusion of three different types of features was presented. An automatic facial expression recognition system evaluated on the same three types of spontaneous databases was presented by using the auto-encoder deep network framework with the SOM based classifier. The geometric features were extracted with lower complexity, after solving the misalignment issue of the training images. Then, the regions containing expressions in the face were automatically cropped based on geometric landmarks. Appearance features were extracted by using two types of methods LBP and LPQ.

   Next, the fusion of geometric features and two types of appearance features was a novel idea to obtain a better representation of facial attributes. The fusion of geometric features, together with two different types of appearance features based on local phase quantisation (LPQ) and local binary pattern (LBP), provides a rich feature vector. An effective fusion of three different features using an auto-encoder was considered the first of a kind. In addition, a better result in all databases has also been obtained by the fusion of three features.

   Finally, a multi-classifier system was used to enhance the performance, where the problem was split into two or more stages, thereby splitting the training effort into two levels, namely, the surface level and bottom level, in each stage. Each level has a network and a classifier, each of which is a more straightforward classification task, thereby achieving a higher accuracy of recognition of facial expressions. Experimental results displayed a significant improvement in the average recognition accuracy compared with previous approaches when using the MMI database. Moreover, the proposed method is clearly superior to recognition using state-of-the-art methods when applied to three types of spontaneous databases, with a recognition performance of 98.96% for the MMI database, 97.05% for the VD-MFP database and 95.64% for the BINED database.

The experimental results confirmed that the strategy of multi-classification outperforms the strategy of single classification and moreover, attains higher recognition performance for facial expression methods with a comparison of some state-of-the-art methods, including a deep network solution.

7.2 Future Work

Although the proposed system has made many improvements upon the current state-of-the-art in the automatic image-based facial expression recognition field with the work introduced in this thesis, several problems remain unsolved and need to be addressed in future work.
The findings of this thesis have a number of important implications which can be studied in the future. Accordingly, some of these findings can provide new insights for future research.

- On the one hand, the proposed system in this thesis, the AFER, only considers the view change in image-based facial expression images. In order to apply the system in practice, a number of issues still require to be solved. A possible solution to the recognition of the dynamic facial expression of arbitrary views is to use a multi-view model. As most of the databases (posed and spontaneous) have been collected and used for the frontal FER, in future work, it is, therefore, proposed to use Multi-PIE [196], which is a multi-view FER database.

- In the case of facial segmentations more effort will be demanded to generate a robust ground truth. The currently proposed ground truth is based on the basic key points of facial expression (eyebrow, eye and mouth). It is observed that these key points cannot cover all the facial expression patterns. Thus, justified ground truth for the facial expression segmentation depends on a region that has furrows, bulges and wrinkles in addition to the eye, eyebrow and mouth which can be established and then provided for researchers.

- Additional geometric feature extraction approaches with lower complexity can be proposed, and selecting an extra/reduced landmark point. For appearance extraction, it is recommended to introduce a feature to improve the LBP method by adaptive horizontal and vertical weights, which may contribute a good basis to calculate the adaptive weights.

- On the other hand, the proposed novel approach in Section 6.4 was successful in improving recognition performance of facial expression methods with a comparison of some state-of-the-art methods. A possible solution may use different types of classifiers with deep learning.

- Finally, the proposed system in this thesis can possibly be used as a solution to obtain high accuracy in terms of recognition of facial expression in live video/multi-frames to achieve unconstrained interaction.

Appendix A

AFER Analysis based on Geometrical Model

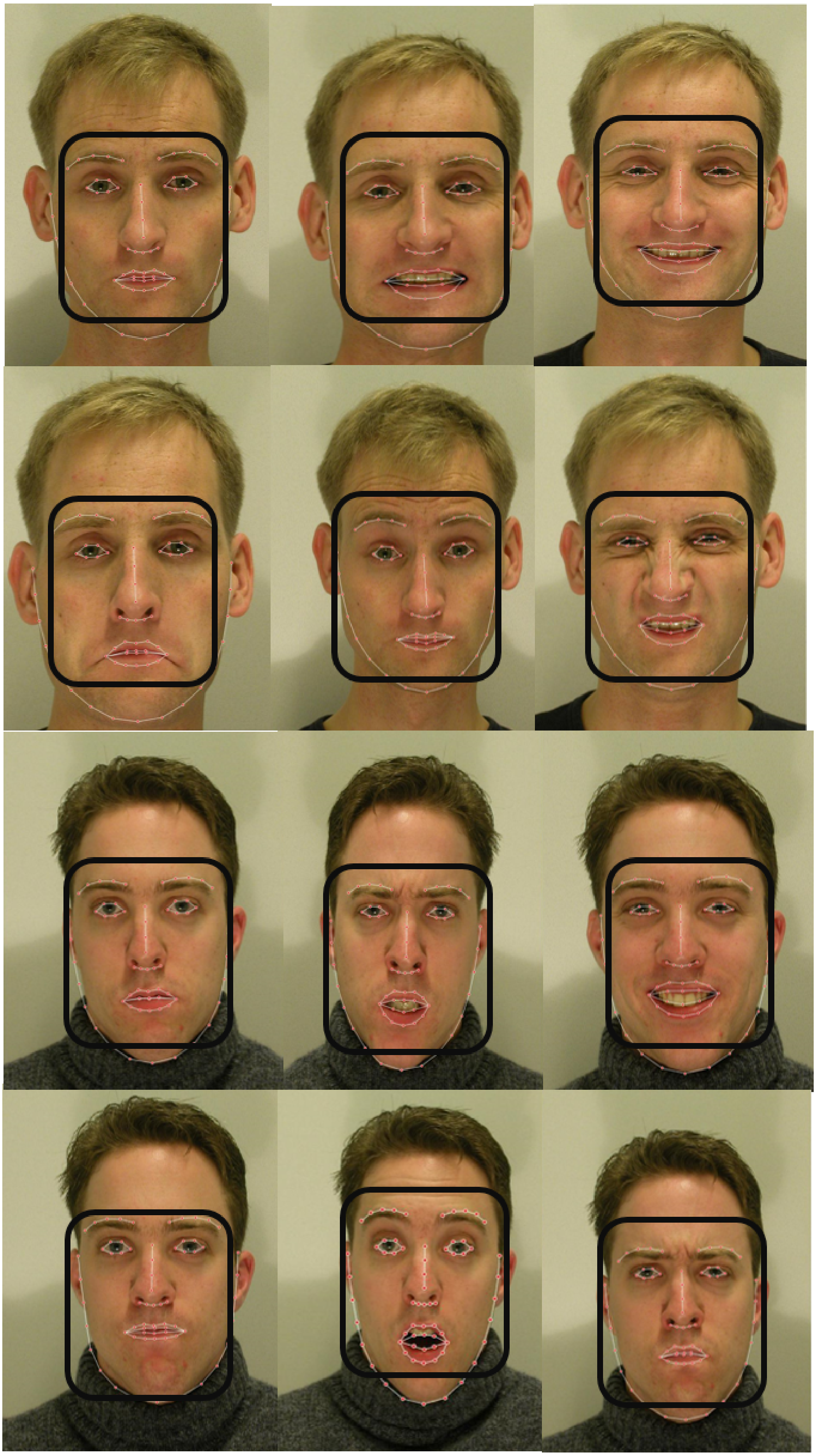


Figure A.1: Automatic detecting and tracking frames for facial expression analyses by 65 landmarks points.

References

[1] A. Mehrabian, "Communication without words," *Communication Theory*, pp. 193–200, 2008.

[2] P. Ekman, *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life.* Macmillan, 2007.

[3] A. K. Jain and S. Z. Li, *Handbook of face recognition.* Springer, 2011.

[4] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." *Psychological bulletin*, vol. 111, no. 2, p. 256, 1992.

[5] M. Slavković and D. Jevtić, "Face recognition using eigenface approach," *Serbian Journal of Electrical Engineering*, vol. 9, no. 1, pp. 121–130, 2012.

[6] C. Darwin and P. Prodger, *The expression of the emotions in man and animals.* Oxford University Press, USA, 1998.

[7] P. Ekman and W. Friesen, "Unmasking the face a guide to recognizing emotions from facial clues," 2003.

[8] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[9] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," in *Handbook of Face Recognition.* Springer, 2011, pp. 487–519.

[10] H. Hussein, F. Angelini, M. Naqvi, and J. A. Chambers, "Deep-learning based facial expression recognition system evaluated on three spontaneous databases," in *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC).* IEEE, 2018, pp. 270–275.

[11] N. Sebe, *Machine learning in computer vision.* Springer Science & Business Media, 2005, vol. 29.

[12] N. Bianchi-Berthouze and C. L. Lisetti, "Modeling multimodal expression of user's affective subjective experience," *User Mdeling and User-Adapted Interaction*, vol. 12, no. 1, pp. 49–84, 2002.

[13] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *California Mental Health Research Digest*, vol. 8, no. 4, pp. 151–158, 1970.

[14] P. Ekman, "Averaged gabor

lter features - studies in emotion interaction," cambridge university press, second edition."

[15] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *Journal of Nonverbal Behavior*, vol. 6, no. 4, pp. 238–252, 1982.

[16] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[17] S. D'Mello and R. A. Calvo, "Beyond the basic emotions: what should affective computing compute?" in *CHI'13 Extended Abstracts on Human Factors in Computing Systems.* ACM, 2013, pp. 2287–2294.

[18] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-Time Vision for Human-Computer Interaction.* Springer, 2005, pp. 181–200.

[19] T. Sobol-Shikler and P. Robinson, "Classification of complex information: Inference of co-occurring affective states from their expressions in speech," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1284–1297, 2010.

[20] M. S. Mahmoud, *Decentralized control and filtering in interconnected dynamical systems.* CRC Press Boca Raton, 2011, vol. 10.

[21] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[22] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[23] A. Lang, J. Harro, A. Soosaar, S. Kõks, V. Volke, L. Oreland, M. Bourin, E. Vasar, J. Bradwejn, and P. T. Männistö, "Role of n-methyl-d-aspartic acid and cholecystokinin receptors in apomorphine-induced aggressive behaviour in rats," *Naunyn-Schmiedeberg's Archives of Pharmacology*, vol. 351, no. 4, pp. 363–370, 1995.

[24] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[25] T. Baltrušaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on.* IEEE, 2013, pp. 1–8.

[26] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Networks*, vol. 18, no. 4, pp. 317–352, 2005.

[27] E. Aronson, T. D. Wilson, R. M. Akert *et al.*, "Social psychology," *Attitudes and Attitude Change: Influencing Thoughts and Feelings*, pp. 199–235, 2005.

[28] M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," in *International Joint Conference on Pattern Recognition*, vol. 1978, 1978, pp. 408–410.

[29] K. Mase and A. Pentland, "Automatic lipreading by optical-flow analysis," *Systems and Computers in Japan*, vol. 22, no. 6, pp. 67–76, 1991.

[30] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic face identification system using flexible appearance models," *Image and Vision Computing*, vol. 13, no. 5, pp. 393–401, 1995.

[31] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion," in *Computer Vision, 1995. Proceedings., Fifth International Conference on.* IEEE, 1995, pp. 374–381.

[32] Y. Yacoob and L. S. Davis, “Recognizing human facial expressions from long image sequences using optical flow,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 636–642, 1996.

[33] M. G. Rosenblum, A. S. Pikovsky, and J. Kurths, “Phase synchronization of chaotic oscillators,” *Physical Review Letters*, vol. 76, no. 11, p. 1804, 1996.

[34] I. A. Essa and A. P. Pentland, “Coding, analysis, interpretation, and recognition of facial expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 757–763, 1997.

[35] M.-H. Yang, D. J. Kriegman, and N. Ahuja, “Detecting faces in images: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[36] G. Yang and T. S. Huang, “Human face detection in a complex background,” *Pattern Recognition*, vol. 27, no. 1, pp. 53–63, 1994.

[37] H. Mekami and S. Benabderrahmane, “Towards a new approach for real time face detection and normalization,” in *Machine and Web Intelligence (ICMWI), 2010 International Conference on.* IEEE, 2010, pp. 455–459.

[38] P. Sinha, “Perceiving and recognizing three-dimensional forms,” Ph.D. dissertation, Massachusetts Institute of Technology, 1995.

[39] T. Cootes, A. Hill, C. Taylor, and J. Haslam, “The use of active shape models for locating structures in medical images,” in *Information Processing in Medical Imaging.* Springer, 1993, pp. 33–47.

[40] S. A. Sirohey, “Human face segmentation and identification,” Tech. Rep., 1998.

[41] D. Saxe and R. Foulds, “Toward robust skin identification in video images,” in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on.* IEEE, 1996, pp. 379–384.

[42] M. F. Augusteijn and T. L. Skufca, “Identification of human faces through texture-based feature recognition and neural network technology,” in *Neural Networks, 1993., IEEE International Conference on.* IEEE, 1993, pp. 392–398.

[43] Y. Freund and R. E. Schapire, “A desicion-theoretic generalization of on-line learning and an application to boosting,” in *European Conference on Computational Learning Theory.* Springer, 1995, pp. 23–37.

[44] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[45] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[46] M. Rydfalk, *CANDIDE: A parameterised face.* Linköping Univ., 1987.

[47] H. Tao and T. S. Huang, “Explanation-based facial motion tracking using a piecewise bezier volume deformation model,” in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 1. IEEE, 1999, pp. 611–617.

[48] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[49] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active shape models-their training and application,” *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[50] S. M. Lajevardi and M. Lech, “Averaged gabor filter features for facial expression recognition,” in *2008 Digital Image Computing: Techniques and Applications.* IEEE, 2008, pp. 71–76.

[51] Y. Tie and L. Guan, “Automatic face detection in video sequences using local normalization and optimal adaptive correlation techniques,” *Pattern Recognition*, vol. 42, no. 9, pp. 1859–1868, 2009.

[52] A. Majumder, L. Behera, and V. K. Subramanian, “Local binary pattern based facial expression recognition using self-organizing map,” in *Neural Networks (IJCNN), 2014 International Joint Conference on.* IEEE, 2014, pp. 2375–2382.

[53] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Computer Vision and Pattern Recognition (CVPR).* IEEE, 2013, pp. 3444–3451.

[54] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 971–987, 2002.

[55] A. Hadid, M. Pietikainen, and T. Ahonen, "A discriminative feature space for detecting and recognizing faces," in *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.

[56] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2006.

[57] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[58] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Multimedia and Expo. ICME 2005. IEEE International Conference on*, 2005.

[59] A. J. Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi, "A video database of moving faces and people," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 5, pp. 812–816, 2005.

[60] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The Belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2012.

[61] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[62] S. Zafeiriou and M. Petrou, "Sparse representations for facial expressions recognition via l1 optimization," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, p. 32.

[63] A. Majumder, L. Behera, and V. K. Subramanian, “Automatic facial expression recognition system using deep network-based data fusion,” *IEEE Transactions on Cybernetics*, vol. 25, pp. 165–177, 2016.

[64] S. Li and W. Deng, “Deep facial expression recognition: A survey,” *IEEE Transactions on Affective Computing*, 2019.

[65] S. Z. Li and A. K. Jain, “Encyclopedia of biometrics: I-z.” vol. 1, 2009.

[66] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[67] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face recognition with local binary patterns,” *Computer Vision-eccv 2004*, pp. 469–481, 2004.

[68] V. Ojansivu and J. Heikkilä, “Blur insensitive texture classification using local phase quantization,” in *International Conference on Image and Signal Processing.* Springer, 2008, pp. 236–243.

[69] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, “A dynamic appearance descriptor approach to facial actions temporal modeling,” *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 161–174, 2014.

[70] B. Jiang, M. F. Valstar, and M. Pantic, “Action unit detection using sparse appearance descriptors in space-time video volumes,” in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on.* IEEE, 2011, pp. 314–321.

[71] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, “Avec 2013: the continuous audio/visual emotion and depression recognition challenge,” in *Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge.* ACM, 2013, pp. 3–10.

[72] P. Yang, “Facial expression recognition and expression intensity estimation,” Ph.D. dissertation, Rutgers University-Graduate School-New Brunswick, 2011.

[73] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[74] D. Gabor, “Theory of communication. part 1: The analysis of information,” *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[75] T. Barbu, “Gabor filter-based face recognition technique,” *Proceedings of the Romanian Academy*, vol. 11, no. 3, pp. 277–283, 2010.

[76] J. G. Daugman, “Complete discrete 2-d gabor transforms by neural networks for image analysis and compression,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1169–1179, 1988.

[77] T. S. Lee, “Image representation using 2d gabor wavelets,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[78] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on.* IEEE, 1998, pp. 200–205.

[79] P. Sisodia, A. Verma, and S. Kansal, “Human facial expression recognition using gabor filter bank with minimum number of feature vectors,” *Int. J. Appl. Inf. Syst*, vol. 5, no. 9, pp. 9–13, 2013.

[80] R. Samad and H. Sawada, “Extraction of the minimum number of gabor wavelet parameters for the recognition of natural facial expressions,” *Artificial Life and Robotics*, vol. 16, no. 1, pp. 21–31, 2011.

[81] M. Abdulrahman, T. R. Gwadabe, F. J. Abdu, and A. Eleyan, “Gabor wavelet transform based facial expression recognition using pca and lbp,” in *Signal Processing and Communications Applications Conference (SIU), 2014 22nd.* IEEE, 2014, pp. 2265–2268.

[82] G. Hegde, M. Seetha, and N. Hegde, “Facial expression recognition using entire gabor filter matching score level fusion approach based on subspace methods,” in *International Conference on Mining Intelligence and Knowledge Exploration.* Springer, 2015, pp. 47–57.

[83] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[84] M. Dahmane and J. Meunier, "Continuous emotion recognition using gabor energy filters," *Affective Computing and Intelligent Interaction*, pp. 351–358, 2011.

[85] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning, "Local features based facial expression recognition with face registration errors," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on.* IEEE, 2008, pp. 1–8.

[86] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[87] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," in *International Workshop on Analysis and Modeling of Faces and Gestures.* Springer, 2007, pp. 168–182.

[88] F. Ahmed, "Gradient directional pattern: a robust feature descriptor for facial expression recognition," *Electronics Letters*, vol. 48, no. 19, pp. 1203–1204, 2012.

[89] T. Jabid, M. H. Kabir, and O. Chae, "Robust facial expression recognition based on local directional pattern," *ETRI journal*, vol. 32, no. 5, pp. 784–794, 2010.

[90] ——, "Local directional pattern (ldp) for face recognition," in *Consumer Electronics (ICCE), 2010 Digest of Technical Papers International Conference on.* IEEE, 2010, pp. 329–330.

[91] F. Ahmed and M. H. Kabir, "Directional ternary pattern (dtp) for facial expression recognition," in *Consumer Electronics (ICCE), 2012 IEEE International Conference on.* IEEE, 2012, pp. 265–266.

[92] F. Ahmed and E. Hossain, "Automated facial expression recognition using gradient-based ternary texture patterns," *Chinese Journal of Engineering*, vol. 20, 2013.

[93] G. Tzimiropoulos, J. Alabort-i Medina, S. P. Zafeiriou, and M. Pantic, "Active orientation models for face alignment in-the-wild," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2024–2034, 2014.

[94] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on.* IEEE, 1998, pp. 454–459.

[95] Y.-l. Tian, T. Kanade, and J. F. Cohn, "Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on.* IEEE, 2002, pp. 229–234.

[96] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06).* IEEE, 2006, pp. 149–149.

[97] Y. Chang, C. Hu, R. Feris, and M. Turk, "Manifold based analysis of facial expression," *Image and Vision Computing*, vol. 24, no. 6, pp. 605–614, 2006.

[98] S. Jain, C. Hu, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on.* IEEE, 2011, pp. 1642–1649.

[99] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on.* IEEE, 2010, pp. 94–101.

[100] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2007.

[101] T. T. Do and T. H. Le, "Facial feature extraction using geometric feature and independent component analysis," in *Pacific Rim Knowledge Acquisition Workshop.* Springer, 2008, pp. 231–241.

[102] V. Perlibakas, "Face recognition using principal component analysis and wavelet packet decomposition," *Informatica*, vol. 15, no. 2, pp. 243–250, 2004.

[103] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468).* Ieee, 1999, pp. 41–48.

[104] C. F. F. Costa Filho, T. d. A. Falcão, M. G. F. Costa, and J. R. G. Pereira, "Proposing the novelty classifier for face recognition," *Revista Brasileira de Engenharia Biomédica*, vol. 30, no. 4, pp. 301–311, 2014.

[105] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry, "Feature selection in face recognition: A sparse representation perspective," *submitted to IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 2, 2007.

[106] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[107] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, 2018.

[108] V. N. Vapnik and V. Vapnik, *Statistical learning theory.* Wiley New York, 1998, vol. 1.

[109] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, 1997, pp. 155–161.

[110] M. A. Razi and K. Athappilly, "A comparative predictive analysis of neural networks (nns), nonlinear regression and classification and regression tree (cart) models," *Expert Systems with Applications*, vol. 29, no. 1, pp. 65–74, 2005.

[111] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[112] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[113] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113, 1997.

[114] W. Huang and H. Yin, "Visom for dimensionality reduction in face recognition," in *International Workshop on Self-Organizing Maps.* Springer, 2009, pp. 107–115.

[115] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[116] W. Liu and Z. Wang, "Facial expression recognition based on fusion of multiple gabor features," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3. IEEE, 2006, pp. 536–539.

[117] Z. Wang and Z. Ying, "Facial expression recognition based on local phase quantization and sparse representation," in *2012 8th International Conference on Natural Computation.* IEEE, 2012, pp. 222–225.

[118]

[119] R. Manjhi, J. Abbas, and A. Priyam, "Face recognition using eigenface," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 7, pp. 625–627, 2013.

[120] M. Abdullah, M. Wazzan, and S. Bo-Saeed, "Optimizing face recognition using pca," *arXiv preprint arXiv:1206.1515*, 2012.

[121] V. Praseeda, M. Sasikumar, and S. Naveen, "Analysis of facial expressions from video images using pca," in *Proceedings of the World Congress on Engineering*, vol. 1, 2008, pp. 2–4.

[122] A. Gosavi and S. Khot, "Facial expression recognition using principal component analysis," *International Journal of Soft Computing Engineering (IJSCE) ISSN*, pp. 2231–2307, 2013.

[123] A. Rahman, A. R. Armanadurni, H. Seyal, and N. Kamarudin, "Facial recognition using eigenfaces approach," in *2014 International Conference on Global Economy, Commerce and Service Science (GECSS-14).* Atlantis Press, 2014.

[124] Q. Gao and D. Zhang, L.and Zhang, "Face recognition using flda with single training image per person," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 726–734, 2008.

[125] T. Shereena and A. Babu, “Face recognition system by integrating pca, flda, artificial neural networks and minimum euclidean distance.”

[126] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, “Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38–52, 2010.

[127] J. Liu, B. Liu, S. Zhang, F. Yang, P. Yang, D. N. Metaxas, and C. Neidle, “Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions,” *Image and Vision Computing*, vol. 32, no. 10, pp. 671–681, 2014.

[128] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, “Facial action recognition combining heterogeneous features via multikernel learning,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 993–1005, 2012.

[129] K. Yu, Z. Wang, M. Hagenbuchner, and D. D. Feng, “Spectral embedding based facial expression recognition with multiple features,” *Neurocomputing*, vol. 129, pp. 136–145, 2014.

[130] T. H. Zavaschi, A. S. Britto Jr, L. E. Oliveira, and A. L. Koerich, “Fusion of feature sets and classifiers for facial expression recognition,” *Expert Systems with Applications*, vol. 40, no. 2, pp. 646–655, 2013.

[131] J.-H. Kim, B.-G. Kim, P. P. Roy, and D.-M. Jeong, “Efficient facial expression recognition algorithm based on hierarchical deep neural network structure,” *IEEE Access*, vol. 7, pp. 41 273–41 285, 2019.

[132] N. Kumar and D. Bhargava, “A scheme of features fusion for facial expression analysis: A facial action recognition,” *Journal of Statistics and Management Systems*, vol. 20, no. 4, pp. 693–701, 2017.

[133] T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on.* IEEE, 2000, pp. 46–53.

[134] P. Ekman, W. V. Friesen, and J. C. Hager, “Facial action coding system (facs),” *A Technique for the Measurement of Facial Action. Consulting, Palo Alto*, vol. 22, 1978.

[135] S. H. Lee, H. Kim, Y. M. Ro, and K. N. Plataniotis, “Using color texture sparsity for facial expression recognition,” in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on.* IEEE, 2013, pp. 1–6.

[136] S. Happy, P. Patnaik, A. Routray, and R. Guha, “The indian spontaneous expression database for emotion recognition,” *IEEE Transactions on Affective Computing*, November 2015.

[137] E. Sariyanidi, H. Gunes, and A. Cavallaro, “Automatic analysis of facial affect: A survey of registration, representation, and recognition,” *IEEE*, vol. 37, no. 6, pp. 1113–1133, 2015.

[138] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[139] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[140] M. Nixon, *Feature extraction & image processing.* Academic Press, 2008.

[141] K.-C. Chung, S. C. Kee, and S. R. Kim, “Face recognition using principal component analysis of gabor filter responses,” in *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 1999. Proceedings. International Workshop on.* IEEE, 1999, pp. 53–57.

[142] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, “Facial expression recognition with convolutional neural networks: coping with few data and the training sample order,” *Pattern Recognition*, vol. 61, pp. 610–628, 2017.

[143] J. Chen, Z. Chen, Z. Chi, and H. Fu, “Facial expression recognition in video with multiple feature fusion,” *IEEE Transactions on Affective Computing*, 2016.

[144] Y. Liu, Y. Li, X. Ma, and R. Song, “Facial expression recognition with fusion features extracted from salient facial areas,” *Sensors*, vol. 17, no. 4, p. 712, 2017.

[145] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1940–1954, 2010.

[146] R. Fenn, *Geometry*, ser. Springer undergraduate mathematics series. New York: Springer, 2000.

[147] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[148] C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou, "Autoencoder for words," *Neurocomputing*, vol. 139, pp. 84–96, 2014.

[149] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, pp. 1–40, 2009.

[150] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted Boltzmann machines for collaborative filtering," in *Proceedings of the 24th International Conference on Machine Learning.* ACM, 2007, pp. 791–798.

[151] A. Majumder, L. Behera, and V. K. Subramanian, "Emotion recognition from geometric facial features using self-organizing map," *Pattern Recognition*, vol. 47, no. 3, pp. 1282–1293, 2014.

[152] H. Hussein, M. Naqvi, and J. Chambers, "Study of image-based expression recognition techniques on three recent spontaneous databases," in *Digital Signal Processing (DSP), 2017 22nd International Conference on*, pp. 1–5.

[153] A. Saeed, A. Al-Hamadi, R. Niese, and M. Elzobi, "Frame-based facial expression recognition using geometrical features," *Advances in Human-Computer Interaction*, vol. 2014, p. 4, 2014.

[154] A. C. Cruz, B. Bhanu, and N. S. Thakoor, "Vision and attention theory based sampling for continuous facial emotion recognition," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 418–431, 2014.

[155] G. Dantzig and D. R. Fulkerson, "On the max flow min cut theorem of networks," *Linear inequalities and related systems*, vol. 38, pp. 225–231, 2003.

[156] T. Dietterich, “Overfitting and undercomputing in machine learning,” *ACM Computing Surveys (CSUR)*, vol. 27, no. 3, pp. 326–327, 1995.

[157] H. Kaya, F. Gürpınar, and A. A. Salah, “Video-based emotion recognition in the wild using deep transfer learning and score fusion,” *Image and Vision Computing*, vol. 65, pp. 66–75, 2017.

[158] H. Hussein, M. Naqvi, and J. Chambers, “Spontaneous facial expression recognition through fusion of geometric and appearance features and deep learning based classification.”