AUTOMATIC AND PERSON-INDEPENDENT DETECTION OF 3D FACIAL EXPRESSIONS USING ENHANCED SUPPORT VECTOR MACHINE WITH PROBABILITY ESTIMATION

AMAL ABDULAZIZ ABDULLAH

UNIVERSITI SAINS MALAYSIA

2014

AUTOMATIC AND PERSON-INDEPENDENT DETECTION OF 3D FACIAL EXPRESSIONS USING ENHANCED SUPPORT VECTOR MACHINE WITH PROBABILITY ESTIMATION

by

AMAL ABDULAZIZ ABDULLAH

Thesis submitted in fulfilment of the requirements for the degree of Master of Science

December 2014

ACKNOWLEDGEMENTS

First of all, praise is due to Allah, the Almighty, for giving me the opportunity and strength to complete my thesis.

I would like to express my deep gratitude to my supervisors, Dr. Ibrahim Venkat, Dr. Syaheerah Lebai Lutfi, and Prof. Dr. Bahari Belaton. Their invaluable guidance and consistent encouragement throughout my study have been of a great value to me.

My profound gratitude is also expressed to my ex-supervisor, Dr. Iman Liao Yi. Her valuable advice and assistance at the beginning of my journey enlightened my research.

I also appreciate the help and support from all who were directly or indirectly involved in my research. To my friends in Artificial Intelligent lab: Junfen Chen, Tan Joi San, Bisan Salibi, and Preeta Naidu thank you for listening, offering me advice, and supporting me through this entire process.

Last but not least, I owe my success to my family, my parents, brothers, and sisters. Their high expectations, love, and well-wishes/prays keep me motivated all the time. A special and heartfelt thank to the backbone of my study, my husband Osamah, who supported me financially and spiritually. The unconditional love and patience I received from you and our son, Magd, was overwhelming and inspired me to complete my study despite all the hard moments we went through. Thank you for being by my side through the thick before the thin. I hope I have been a good mother and wife and that I have not lost too much during my study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

2D 2 Dimension

3D 3 Dimension

AFM Annotated Face Model

AC Affective Computing

AI Artificial Intelligence

AU Action Coding

BBN Bayesian Belief Network

BU-3DFE Binghamton University-3D Facial Expression

DE Differential Evolution

DPM Deformable Parts Model

FACS Facial Action Coding System

FAP Facial Animation Protocol

FDA Fisher Discriminant Analysis

FER Facial Expression Recognition

GA Genetic Algorithm

HCI Human Computer Interaction

HoG Histogram of Gradient

HoS Histogram of Shape index

ICP Iterative Closest Point

IKCCA Improved Kernel Canonical Correlation Analysis

k-NN k-Nearest Neighbours

LBP Local Binary Pattern

LDA Linear Discriminative Analysis

MLP NN Multilayer Perceptron Neural Network

MPEG-4 Moving Pictures Experts Group-4

NB Naíve Bayes

NN Neural Network

PCA Principle Component Analysis

PNN Probabilistic Neural Network

QDC Quadratic Discriminant ClassiïňĄer

RBF Radial Basis Function

SFAM Statistical Facial feAture Model

SIFT Scale Invariant Feature Transform

SVM Support Vector Machine

PENGESANAN SECARA AUTOMATIK DAN KETIDAKBERGANTUNGAN WAJAH BAGI EXPRESI WAJAH 3D MENGGUNAKAN SUPPORT VECTOR MACHINE YANG DIPERTINGKATKAN DENGAN ANGGARAN KEBARANGKALIAN

ABSTRAK

Ekspresi wajah merupakan salah satu saluran komunikasi yang berkuasa dalam kehidupan seharian kita kerana ia memperlihatkan emosi dan niat seseorang. Manusia sentiasa menunjukkan and meluahkan perasaan mereka melalui ekspresi wajah sepanjang masa, termasuk ketika berinteraksi dengan mesin. Keadaan ini menjadi pendorong bagi kajian utk menambah baik mesin supaya menjadi lebih sensitif terhadap emosi dan perasan manusia, seperti kebolehan untuk mengenali ekspresi wajah. Berbanding dengan imej wajah 2D, imej wajah 3D berkebolehan memberi isyarat lebih terperinci yang tidak terdapat dalam imej 2D. Isyarat lebih jelas yang ditunjukkan melalui imej 3D mengatasi batasan sistem 2D seperti kepekaan terhadap posisi kepala dan variasi pencahayaan. Walau bagaimanapun, satu kelemahan imej wajah 3D adalah ia memaparkan kematraan dimensi yang lebih tinggi daripada imej wajah 2D.

Cabaran tersebut diatasi di dalam tesis ini, iaitu dengan mencadangkan model pengecaman ekspresi wajah 3D secara automatik sepenuhnya yang menangani masalah kematraan dimensi yang tinggi melalui dua kaedah penyelesaian. Pertama, imej wajah 3D dipetakan ke dalam pengimejan 2D melalui kaedah geometri konformal. Kedua, kaedah pengoptimuman baru dicadangkan untuk memilih ciri-ciri wajah yang optimum daripada ciri-ciri wajah yang luas. Ciri-ciri wajah yang dipilih adalah ciri-ciri yang dapat mendiskriminasi enam ekspresi wajah universal secara umum (iaitu, *iaitu, marah, gembira, takut, sedih, terkejut, meluat*) dan *neutral*

secara automatik dan bebas dari kebergantungan kepada wajah tertentu.

Menggunakan SVM yang standard untuk pengelasan ekspresi wajah, model dengan dua kaedah penyelesaian yang dicadangkan tersebut telah dinilai menggunakan dua pangkalan data yang biasadigunakan: Bosphorus dan BU-3DFE. Model ini menghasilkan ketepatan pengecaman keseluruhan sebanyak 79% menggunakan pangkalan data Bosphorus dan 79.36% menggunakan pangkalan data BU-3DFE. Bagi meningkatkan lagi kadar ketepatan pengecaman, pergerakan otot wajah telah dianalisis dan diterjemahkan sebagai pemberat menggunakan sistem Facial Action Coding System (FACS) untuk mengoptimumkan variasi kelas antaraekspresi. Pemberat tersebut kemudian dijumlahkan dengan anggaran kebarangkalian SVM untuk meningkatkan kebarangkalian dalam mengganggar kelas dengan lebih tepat. Gabungan peningkatan anggaran kebarangkalian yang dicadangkan menghasilkan kadar ketepatan pengecaman keseluruhan sebanyak 84% dengan data Bosphorus dan 85.81% dengan data BU-3DFE. Ini adalah keputusan statistik yang signifikan ($p < 0.01$ dan $p < 0.001$, masing-masing) berbanding dengan keaedah eksploitasi ciri-ciri optimum secara individu. Pada keseluruhannya, kadar ketepatan pengecaman model yang dicadangkan mengatasi kadar ketepatan kajian-kajian sebelum ini.

AUTOMATIC AND PERSON-INDEPENDENT DETECTION OF 3D FACIAL EXPRESSIONS USING ENHANCED SUPPORT VECTOR MACHINE WITH PROBABILITY ESTIMATION

ABSTRACT

Facial expressions are powerful communication tools in our everyday life as they impart a person's emotional state and his/her intentions. Humans express their feeling through facial expressions all the time, even while interacting with machines. This motivates efforts to empower machines to become affect-aware, such as the ability to recognize facial expressions. Compared to 2D face images, 3D face images offer more granular cues not available in 2D images. The depth cues in 3D images overcome the limitations of 2D systems, such as sensitivity to head pose and variations in illumination. However, one major setback of 3D face images is that they impose a higher dimensionality than 2D face images do.

In this thesis, this challenge has been addressed by proposing a fully automatic 3D facial expression recognition model that tackles the high-dimensionality problem in a two-fold solution. First, the 3D face images are mapped into the 2D plane using conformal geometry. Secondly, a new optimization method is proposed to select the optimal facial features from the large pool of facial features. The features are selected to best discriminate between the six universal facial expressions (i.e., *anger, happiness, fear, sadness, surprise*, and *disgust*) and *neutral* in an automatic and person-independent manner.

Using support vector machine with probability estimation (SVM-PE) for classification, the model with the proposed two-fold solution has been evaluated using two common databases: the Bosphorus database and the BU-3DFE database. The model yielded an overall recognition

accuracy of 79% using the Bosphorus database and 79.36% using the BU-3DFE database. To further improve the recognition accuracy, the facial muscular movements were analyzed and interpreted into weights using the Facial Action Coding System (FACS) to maximize the inter-expression class variations. The weights were aggregated with the probability estimates of SVM-PE to enhance the probabilities of the right classes. The combination of the proposed feature selection method with the proposed probability estimate enhancement (termed SVM-EPE) yielded an overall recognition accuracy of 84% using the Bosphorus database and 85.81% using the BU-3DFE database, both of which are statistically significantly better (at $p < 0.01$ and $p < 0.001$, respectively) when compared to the individual exploit of only the optimal features. The overall recognition accuracies of the proposed model outperform the accuracies of the previous studies.

CHAPTER 1

INTRODUCTION

1.1 Background

Emotion is the language of humans' thinking and behavior. Emotional cues convey a person's affective and cognitive state and his/her intentions. In the fields of neurology, psychology, medicine and sociology, emotions are considered as the guide of human's social interaction and communication (Carroll, 2013). For example, nodding the head up and down signifies a person's agreement about the current situation, lowering eyebrows depicts his/her dissatisfaction. This significance of emotions in human-human interaction was the motivation behind the concept of *affective computing* (AC) that was first introduced by Picard (1997). It aims to advance the machine's artificial emotional intelligence to resemble the human's natural emotional intelligence so the machines become more seamless and sensitive to the human mind states. Indeed, this could be achieved by modeling the human perception of the emotions computationally so that machines are able to sense, interpret and adapt to the user's affective state automatically to become more socially intelligent.

Human emotions could be sensed through various cues. These emotional cues could be physiological cues such as brain activity, heartbeat, blood pressure, and skin sweat, or they could be audio-visual cues such as voice, facial expressions and body gesture. Among these cues, facial expressions are the most common and understandable emotional ones. Even during speaking, they contribute 55% of the speaker effect (Segal and Jaffe, 2008). Hence, automatic facial expression recognition is a substantial step towards intelligent human-computer interaction. Waves of research efforts in human computer interaction (HCI) and artificial intelligence (AI) attempt to boost the machine's emotional intelligence through facial expressions by means of advanced sensors and methods in computer graphics and vision. However, this topic is still a very active area with many problems that should be addressed.

1.2 Motivation

Human beings express their feelings through facial expressions all the time. Not only when they interact with one another, but also when they interact with machines. According to Reeves and Nass (1996), human beings have the tendency to interact with machines the same way they interact with one another. However, machines still do not have the ability to respond or adapt to the human state accordingly. This is one of the reasons that motivate the efforts to enrich machines with the ability to recognize facial expressions automatically. Such skills in machines constitute stepping stones towards a user-centered smart environment where machines are able to analyze the user's emotions and adapt the surrounding environment to his/her affective state.

Within HCI field, facial expression recognition is a key point that aims to translate the user's emotional experience into meaningful feedback in various areas. For example, automatic detection of *fatigue, stress*, and *drowsiness* is essential where the user's attention is highly required, such as in surveillance and vehicle driving (Vural et al., 2007). For example, if the machine could detect such expressions and alarm the user, the accident rates could be decreased. Furthermore, machines with automatic recognition ability could provide great assessment cues about the affective state of the users under observation. In e-learning, synthetic tutors could control the flow of the class when they know if the students are *bored, interested* or *puzzled* (Whitehill et al., 2008; Ben Ammar et al., 2010; Bahreini et al., 2014). A doctor may monitor his/her patients from a distance through tele-home care programs and be alerted when the patients are *suffering, annoyed, depressed*, or *comfortable* (Lau, 2010; Khosla et al., 2012). The business sector could further benefit from such skills by translating the customers' affective states, *satisfied* or not, into a statistical report replacing the traditional pen-and-paper method of testing customer satisfaction (Puccinelli et al., 2010).

Although humans can recognize others facial expressions effortlessly, such skills are a major challenge for machines. Establishing a smart, user-centered technology relies heavily on the computational model that is embedded in machines with the ability to analyze and respond in an automatic manner and in reliable time. Towards this goal, the fields of HCI, AI, and AC have witnessed advances during the last decade; however, the existing outcomes are still far from satisfactory (Sandbach et al., 2012).

Traditionally, studies in facial expression recognition have mainly utilized 2D face images to build recognition models. Although good accuracies in 2D facial expression recognition have been reported in a number of different studies (Valstar et al., 2012), the accuracy gradually decreases when head pose and illumination variations occur. Nowadays, the technological advancement in 3D scanners and their considerable price reduction facilitates researchers to carry out facial expression analysis in the 3D domain thereby overcoming the limitations imposed by the 2D systems. 3D face images remarkably offer more granular cues that are not available in 2D face images. 2D images only offer texture cues, but 3D images also provide the depth cues that best describe the face shape. This shape information mitigates the illumination and head pose variations, which greatly affect the texture information. These cues motivate the researcher to utilize 3D face images for 3D facial expression recognition.

1.3 Research Questions

Incorporating 3D face images in affect-recognition systems in the context of HCI would contribute towards an enhanced recognition capability. The leap from 2D to 3D facial expression recognition is rather recent, and the work done in this area is still in the infant stage. However, this leap comes with a great challenge: although 3D face images offer detailed information on shape deformation that would improve expression recognition, they impose higher dimensionality than 2D images do.

Towards a real-life facial expression recognition system that incorporates 3D face images as input, several questions are raised in the attempt to address the problem of 3D facial expression recognition:

1. How to deal with the dimensionality problem without losing the valuable information provided by the 3D images?
2. How to identify the best facial features that could automatically be extracted from any probe face regardless of personal facial variations?
3. How to improve discriminating among facial expression classes that leads to a state-of-the-art recognition accuracy?

1.4 Thesis Objectives and Scope

In general, affect-sensitive applications encompass three layers, namely: the sensing layer, the computational modeling layer, and the application layer. This thesis concerns the computational modelling of emotion recognition through 3D facial expressions. Although people express a wide array of emotions through facial expressions, most of the colouring of emotions falls under a set of basic emotions (i.e. *anger, disgust, fear, happiness, sadness, surprise*, and *neutral*) which this thesis explores in the static 3D faces. These expressions are considered as universal facial expressions and are widely addressed in this field. Facial expressions could be detected by machines using static images or in image sequence/video. This thesis focuses on 3D facial expression recognition in static images.

Following the problems stated in Section 1.3, this thesis has the following specific objectives:

1- To investigate how the high dimensionality inherent within the 3D data could be adequately reduced so that the valuable information provided by the 3D domain can be well preserved.

2- To develop an automatic and person-independent computational model for 3D facial expression recognition with a good accuracy compared to state-of-the-art recognition accuracies.

3- To evaluate the proposed computational model on standard benchmark databases that are commonly used in the field.

1.5 Contributions

This thesis presents a novel 3D facial expression recognition model that is grounded on several contributions to solve the problems mentioned in Section 1.3:

- First, the recognition problem is transferred from the 3D space into the 2D space to reduce the dimensionality by means of advanced conformal geometry method. Conformal mapping has recently been introduced to the field of expression recognition and continues to be investigated due to its efficiency in transferring the problem from one space to another conformally.

- Second, an algorithm for facial feature selection based on differential evolution is proposed to select the optimal facial feature set having the ability to discriminate among the seven facial expressions. The algorithm is able to search for the minimum number of person-independent features that could be detected on any face automatically.

- Third, an enhanced classification method is proposed to improve classification accuracy. Facial muscular movements are analyzed and integrated with the optimal features to maximize inter-expression class variations.

- Fourth, all of the above mentioned contributions are integrated into one fully automatic and person-independent 3D facial expression model of 3D faces. This model is evaluated using two commonly used 3D databases and yields comparable results to state-of-the-art results in this field.

1.6 Thesis Outlines

This thesis is organized as follows:

Chapter 2 first discusses how expression recognition study in the computer science community have been inspired by some psychological theories in the facial expression domain. The most-followed psychological theories and systems are presented. Next, the Chapter reviews the state-of-the-art studies in the field of 3D facial expression recognition. These studies are categorized into different approaches discovering their advantages and limitations. The previous studies in facial feature selection and customized classification for facial expression recognition are further reviewed in this Chapter to provide more insight into these two issues.

Chapter 3 introduces the architecture of the proposed framework for 3D facial expression recognition. How the 3D face images have been mapped into the 2D plane is presented at first, followed by introducing the proposed facial feature selection method that combines three well-known methods in one algorithm. Then, an enhanced method for improving the expression classification accuracy is introduced. How the facial muscular movements are analyzed and integrated with the classification decisions of the optimal features is illustrated in detail.

Chapter 4 demonstrates framework evaluation. First, the results of the proposed facial feature selection algorithm and its efficiency are presented. Next, the performance of the model is evaluated using two databases in two stages: (i) using the optimal features and an existing machine-learning method for classification, and (ii) using the optimal features and the proposed classification enhancement method. The results are further discussed throughout this Chapter.

Chapter 5 summarizes the work presented in this thesis, shows its limitations, and indicates the future directions that could improve this framework.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

When it comes to analyzing the facial expressions of emotions, the first question comes to the mind is *"How facial expressions are perceived and processed to reflect a person's emotion?"*, *i.e.* what are the basic elements of expressions, how do they form a specific emotion, and how would they be classified?. This question first arose in psychology and different theories have been proposed to answer this question. A few of these theories have inspired the researchers in the field of computer science and established a strong foundation for facial expression analysis, since the first work by Parke (1974). This Chapter starts with a brief overview of the most related and common psychological theories that have been followed by computer scientists. Next, the state-of-the-art studies in 3D facial expression recognition with different approaches are discussed in detail.

2.2 Facial Expression Modelling

Emotions and facial expressions have a long history of dispute since the first work in understanding the facial expressions by Darwin *"The expression of the Emotions in Man and Animals"* in 1872. Darwin was the first psychologist who studied the facial expressions across different cultures. Posterior to this study, only a few researchers in psychology focused on the factors that influence the expressions of emotions till the mid of the twentieth century when this field became a quite active research area (Goldstein, 1983). Two major streams have evolved to model the facial expressions of emotions with distinct views: the categorical emotions view and the dimensional emotions view. These two views dominated the study of analyzing the facial expressions in the computer science.

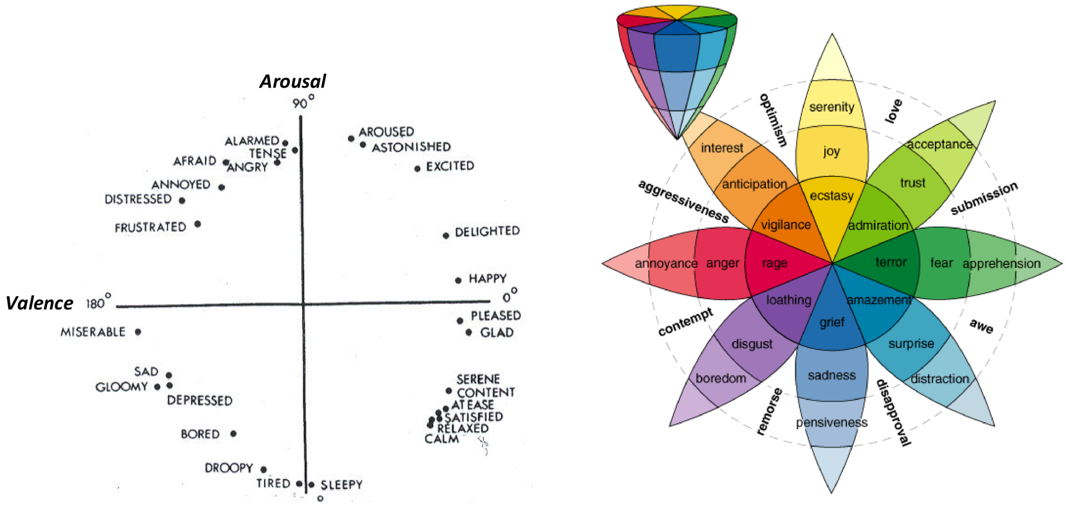
2.2.1 The Categorical Emotions

According to the categorical emotions view, facial expressions are mutually independent and could be categorized into a set of distinct emotions, called "basic emotions". Each category of emotion reflects a distinct affective state as a response to a relevant stimuli and could be identified by particular signals. For instant, winning a competition leads to a state of *joy*, whereas losing it leads to state of being *sad*.

The term "basic" has been used by various theorists to identify their set of emotion categories as basic emotions. Each theory has been derived from different concept with evidences about the meaning of "basic emotions". For example, Frijda (1986) introduced six expressions -*desire, happiness, interest, surprise, wonder*, and *sorrow*- as basic emotions as they form the action of readiness. Mowrer (1960) argued that only *pain* and *pleasure* are the basic emotions as they are unlearned emotions. More interestingly, Ekman (1971) found that there is a set of facial expressions that could be recognized universally, namely: *anger, disgust, fear, happiness, sadness* and *surprise*. Although their universality has been argued by some researchers, such as (Baron-Cohen, 1996) and (Jack et al., 2012), this theory still is widely accepted in the scientific community due to the systematic methods followed in their extensive research.

2.2.2 The Dimensional Emotions

The dimensional emotions view in contrast does not focus on emotion categories, but it rather posits that emotion is a form of a continuous trajectory in n-dimensional emotion space (see Schlosberg (1952); Osgood (1975); Feldman (1995)). In the emotion space, the similar facial expressions are clustered together whereas the different expressions are far apart. Thus, it takes into consideration the emotional colours and subtle emotions as well. The two major theories



(a) The Russell's 2D emotion space (Russell, 1980)

(b) The Plutchik's wheel of emotions (Plutchik, 1980)

Figure 2.1: The dimensional emotions

that support this view are (Plutchik, 1980) and (Russell, 1980) theories.

Russell's study (Russell, 1980) is one of the most significant studies in the field of dimensional emotions. He argued that there is no set of basic emotions, but all the emotions are rather drawn as a circle in a 2D emotion space, as shown Figure 2.1a. His emotion space consists of two dimensions: *valence* and *arousal*. The *valence* dimension varies from *miserable* to *glad* and the *arousal* dimension varies from *sleepy* to *aroused*. Conversely, Robert Plutchik argued that there are eight basic emotions -*joy, trust, fear, surprise, sadness, disgust, anger* and *anticipation*- that could be represented as a wheel in a multi-dimensional space forming different colours of emotions within each category, as illustrated in Figure 2.1b (Plutchik, 1980). He believed that these emotions are biologically adaptive by human beings and any other emotion is a mix of these basic emotions. In Plutchik's 3D circumplex model, the arrange of emotions designed to smooth the transition flow between the different emotions. Every basic emotion has a polar opposite with another emotion, *i.e. trust-disgust, joy-sadness, anticipation-surprise* and *anger-fear*. Furthermore, the intensity of the emotion increases when we move towards the center of the emotion wheel.

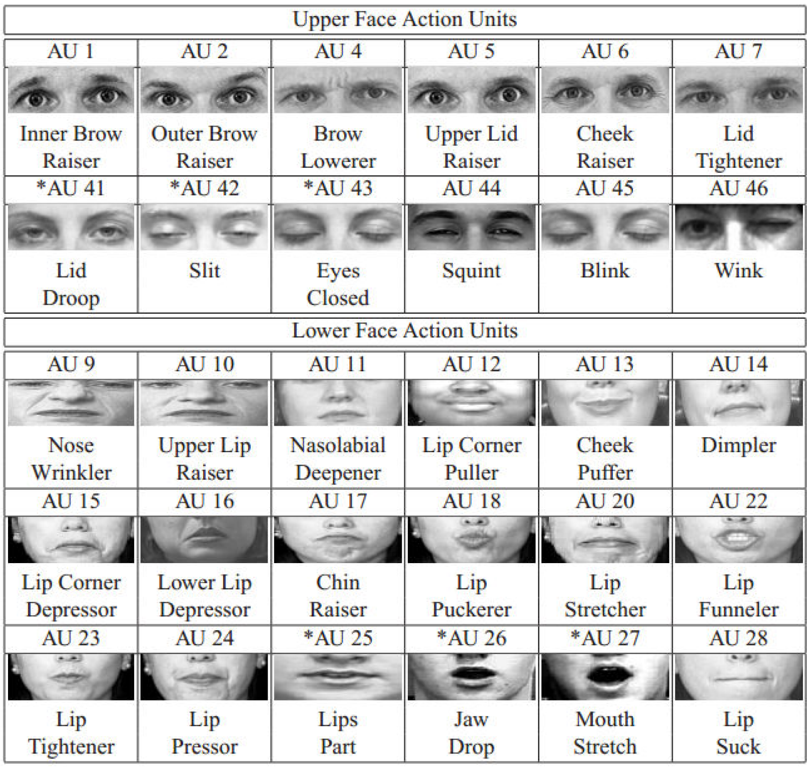


Figure 2.2: Facial Action Coding System (FACS), taken from Tian et al. (2005). AUs start with '*' have been changed later to indicate AUs intensities.

The dimensional emotion view may be successful if used as the underlying idea to model the human beings' feeling; however, the layout of facial expressions could not be encoded in a multi-dimensional space (Young et al., 1997). Moreover, these theories did not provide a systematic method about how these emotions correspond with our facial expressions. Not all what we feel could be conveyed through our facial expressions, and not all the facial expressions reflect a specific emotion (*e.g.* moving eyebrows when we see someone is greeting rather than expressing a feeling). Hence, dimensional emotion theories are more convenient to classify human's feelings rather than facial expressions.

2.2.3 Facial Action Coding System

In 1978, Ekman and Friesen introduced the facial action coding system (FACS). They analyzed the facial activities in terms of facial muscular movements, where every movement is called an action unit (AU), as shown in Figure 2.2. For instant, raising the inner portion of the eyebrows

EMOTION	PROTOTYPES	MAJOR VARIANTS
Surprise	1+2+5B+26 1+2+5B+27	1+2+5B 1+2+26 1+2+27 5B+26 5B+27
Fear	1 +2+4+5*+20*+25, 26, or 27 1+2+4+5*+25, 26, or 27	1+2+4+5*+L or R20*+25, 26, or 27 1+2+4+5* 1+2+5Z, with or without 25, 26, 27 5*+20* with or without 25, 26, 27
Happy	6+12* 12C/D	
Sadness	1+4+11+15B with or without 54+64 1+4+15* with or without 54+64 6+15* with or without 54+64	1+4+11 with or without 54+64 1+4+15B with or without 54+64 1+4+15B+17 with or without 54+64 11+15B with or without 54+64 11+17
	25 or 26 may occur with all prototypes or major variants	
Disgust	9 9+16+15, 26 9+17 10* 10*+16+25, 26 10+17	
Anger	4+5*+7+10*+22+23+25,26 4+5*+7+l0*+23+25,26 4+5*+7+23+25, 26 4+5*+7+17+23 4+5*+7+17+24 4+5*+7+23 4+5*+7+24	Any of the prototypes without any one of the following AUs: 4, 5, 7, or 10.

Table note: * means in this combination the AU may be at any level of intensity.

Figure 2.3: The six universal facial expressions coding using AUs in FACS (Ekman and Friesen, 1977)

is AU1. They identified a set of 46 unique AUs such that any expression could be encoded by a subset of AUs, e.g. *happiness* could be encoded as AU6+AU12. Figure illustrates how every facial expression is described using AUs.

Notably, FACS is developed based on systematic and quantitative description methods. It provides a powerful and objective measurement tool not only for analyzing the expressions of emotion in the face, but also to study the facial appearance changes in general. For this reason, the FACS dominated the major wave of research of facial expression recognition in static and dynamic face images over other coding systems, such as facial animation parameters (FAPs) introduced by Moving Pictures Experts Group MPEG-4 (ISO/IEC14496-1:1999, 1999; Pandzic and Forchheimer, 2002). This is due to its psychological background and the systematic analysis of the FACS compared to MPEG-4 FAPs, which purely invented by computer scientist and originally introduced for face animation rather than expression classification. MPEG-4 FAPs defines a set of 84 facial points to capture 66 facial displacements. One facial muscle movement could trigger more than one FAP.

2.3 Facial Expression Recognition Systems

Mainly, one can categorize the different facial expression recognition approaches based on the feature extraction method. In this context, the term *feature* refers to the facial descriptor vector that captures the discriminative information of the facial deformation under the undertaken expressions for a good classification accuracy. Categorizing feature extraction methods could be seen from different points of view: the information used, shape or texture; the extracted patterns, global or local patterns; etc.

In this thesis, the facial expression recognition systems are divided into three categories: geometric-based approach, local feature-based approach, and global feature-based approach. Geometric-based approach and local feature-based approach impose extracting a set of landmarks of interest to assist the feature extraction procedure. While geometric-based approach studies the geometric attributes of the landmarks, local feature-based approach studies the regions around them. On the other hand, global feature-based approach analyzes the whole face as one pattern so landmarks are only required for alignment or face cropping when applied.

A number of fundamental issues have been emerged regarding to facial expression recognition in the 3D space. First, the dimensionality increase in the 3D space over the 2D space leads up to increase in the required storage and the computation cost as well. Additionally, traditional methods based on algebraic operations are difficult to be applied directly to the 3D images (Venkatesh et al., 2012), as they have been designed to be applied in the 2D space. Such

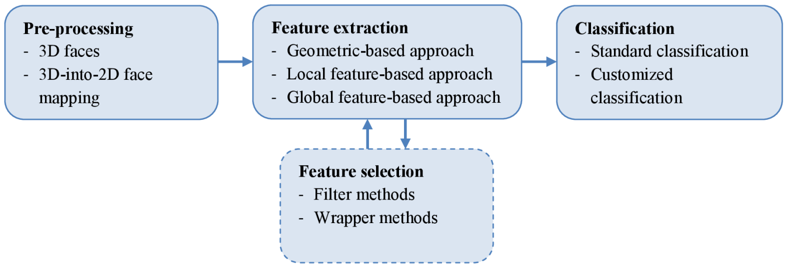


Figure 2.4: The general pipeline of 3D facial expression recognition systems.

issues were the motivations behind the idea of transferring the recognition problem from the 3D space into the 2D space by mapping the 3D face images into the 2D plane. Consequently, the existing 2D methods could be applied to the 2D mapped face images. Hence, feature extraction approaches are further divided into two categories: using 3D face images and using 3D-into-2D mapped face images.

Figure 2.4 illustrates the general pipeline of 3D facial expression recognition systems, which consists of three main phases: pre-processing phase, feature extraction phase, and classification phase. Several studies followed the three-phase system pipeline; notwithstanding, feature selection became an essential step in the pattern recognition field, especially when high dimensional data are used. Two different techniques have been applied in this context: filter methods and wrapper methods. Accordingly, only the selected features are only fed to the classification phase instead of all the features extracted in the feature extraction phase.

In the following sections, the existing 3D face databases are first presented, Section 4.2.1 followed by detailed discussion about the different approaches applied in the four phases: Section 2.3.2: pre-processing phase, Section 2.3.3: feature extraction phase, Section 2.3.4: feature selection phase, and Section 2.3.5: classification phase.

2.3.1 Databases

With the increased attention given to 3D face analysis during the last two decades, several databases have been developed for 3D face analysis, modelling, reconstruction, and recognition. A few of these databases have been designed especially for 3D face analysis with *neutral* faces (Blanz and Vetter, 1999; Beumier and Acheroy, 2001) or with only *happiness* expression (Moreno and Sanchez, 2004; Gupta et al., 2010), but not for facial expression analysis purpose. Thus, they do not contain a wide variety of face images with different expressions.

Several 3D facial expression databases have been constructed to facilitate the research in this area. However, not all of them found their ways to be exploited in different studies. For example, some of them are not available for public such as the databases created by Tsalakanidou and Malassiotis (2010) and Benedikt et al. (2010). Moreover, some databases are constrained by the race, age or gender like BJUT database (Baocai et al., 2009), which contains only Chinese subjects. The variety in the subjects origin, age and race along with the number of the subjects in the database are the major critical issues in establishing a reliable facial expression database. For the purpose of studying the six universal facial expressions, as could be noticed in Table 2.1, only three databases are available: the BU-3DFE database (Yin et al., 2006), the Bosphorus database (Savran et al., 2008) and the ICT-3DRFE database (Stratou et al., 2011).

The ICT-3DRFE database is one of the most recent databases that contains 15 expressions (i.e. the six universal expressions, two *neutral*, scrunched face, two eyebrow expressions and four gaze expressions) posed by twenty three subjects. The database is mainly designed to tackle the problem of illumination variation by providing very high resolution 3D face models. The UPM-3DFE database (Rabiu et al., 2012) is further a recent database contains the six expressions, no *neutral*, of 50 subjects. However, the numbers of the subjects in the two databases are significantly less compared to the number in the BU-3DFE database (100) and

Table 2.1: The available databases for 3D facial expression analysis.

Database	#Subjects	Expressions	#Landmarks	Intensities
GAVDB (Moreno and Sanchez, 2004)	61	*Smile, laugh, a prob expression*	NA	1
FRGC (Phillips et al., 2005)	466	*happy, anger, sad* *surprise, disgust, puffy*	NA	1
BU-3DFE (Yin et al., 2006)	100	the 6 expressions+*neutral*	83	4
CASIA (Zhong et al., 2007)	123	*Smile, laugh, anger* *surprise, closed eyes*	NA	1
Bosphorus (Savran et al., 2008)	105	the 6 expressions+*neutral*	24	1
ND-2006 (Faltemier et al., 2008)	888	*neutral, happy, sad, surprise* *disgust, a prob expression*	NA	1
York (Heseltine et al., 2008)	350	*neutral, happy, anger,* *closed eyes, raised eyebrows*	NA	1
Texas (Gupta et al., 2010)	105	*neutral, smiling,* *talking face with open or closed eyes*	25	1
ICT-3DRFE (Stratou et al., 2011)	23	the 6 expressions+ 2 *neutral* + scrunched face	NA	1
UPM-3DFE (Rabiu et al., 2012)	50	the 6 expressions	32	1

the Bosphorus database (105).

2.3.2 Pre-processing

Commonly, 3D face databases provide the face models in mesh form along with their primitive point clouds and 2D face texture. 3D faces, either point cloud or mesh, require different pre-processing processes before starting the feature extraction. Holes and spikes are the main problems in 3D faces that should be undertaken in the pre-processing step. Interpolating the vertices around the holes is the common solution to close any holes in the face. Gaussian filter is further applied to smooth the 3D face meshes and remove spikes.

When the 3D faces are smooth and closed surfaces, 3D faces are either directly used to extract 3D features or they are first mapped into the 2D plane and then extract the features. The common 3D into 2D mapping method is computing the depth image/range image from the 3D mesh or the point cloud directly, such that every point in the 2D plane $P(u,v)$ is assigned to its correspondence z value in the 3D space. Shape descriptor could then be extracted from the depth images.

Although the depth image offers a simple representation, it provides a perpendicular projection of the face shape from one view angle only, commonly the frontal view. On the contrary, conformal mapping methods provide a one-to-one and angle preserving mapping with minimum distortion. The idea of mapping is could be visualized if we imagine the 3D face as an elastic rubber sheet and their boundaries are then tightened to the boundaries of another surface with minimum energy (Eck et al., 1995). The produced 2D mapped images implicate more and precise shape cues compared to the depth images.

2.3.3 Feature extraction

This section presents the three feature extraction approaches: geometric-base approach, local feature-based approach, and global feature-based approach. The state-of-the-art studies in each category are reviewed and discussed in details.

I. Geometric-based Approach

The first and explicit approach is geometric-based approach, which exploits the geometrical attributes (i.e. coordinates, distances, displacements, angles, or slops) of a selected set of facial landmarks. Geometric attributes can capture the facial shape deformation of the face components directly. The accuracy of this approach relies substantially on the precision of the landmark positions. Without doubt, manual facial landmarking offers the precise locations compared to the automatic work, and consequently yield better recognition accuracy (Zhao et al., 2010). Therefore, the major studies that followed this approach have been conducted on the BU-3DFE database, which provides a set of 83 manually located facial points as shown in Figure 2.5. All the existing systems, that are listed in Table 2.2, are manual systems except the work introduced by Venkatesh et al. (2009). All the 83 facial points or a subset of them have been utilized in these studies.

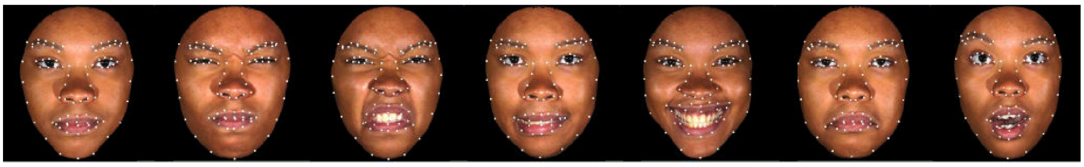
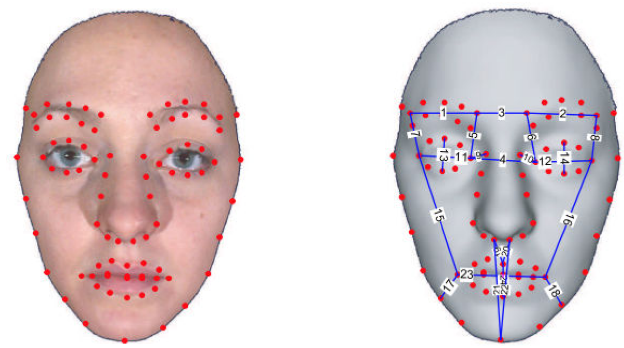


Figure 2.5: The 83 facial points marked on seven 3D face models with seven expressions that belong to the same subject (Yin et al., 2006)

Geometric-based approach is akin to the way the human beings understand others' expressions. For example, the increase in the distance between the mouth corners may imply *happiness*, and the decrease in the distance between the inner eyebrows may imply *anger*. However, such attributes could be misleading in many cases. Large distance between mouth corners could be due to personal appearance (i.e. wide mouth), but not face shape deformation under an expression. Many studies eliminated the personal variation by incorporating the neutral face of the same subject as a reference (Tang and Huang, 2008a,b; Srivastava and Roy, 2009; Venkatesh et al., 2009; Ocegueda et al., 2011). Although the studies that use such technique achieved better accuracies, it leads up to a person-dependent expression recognition system, which is a constrain for establishing a real time recognition system.



(a) The 83 BU-3DFE facial points

(b) A set of 24 manually selected features, Euclidean distances

Figure 2.6: The geometric features used by Tang and Huang (2008a)

With a person-dependent system, Tang and Huang (2008a) achieved the best recognition accuracy (95.10%) overall other studies that followed this approach. They utilized the 83 BU-

3DFE landmarks to extract two types of features: manually selected features and automatically selected features. First, a set of 24 movement parameters from MPEG-4 are selected manually, as shown in Figure 2.6, to form the feature descriptors to classify the six expressions. Second, a feature selection method has been proposed, as described in Section 2.3.4, to select the optimal features automatically. The features are then sent to Adaboost algorithm with three different weak classifiers: *k*-nearest neighbours (*k*-NN), naíve bayes (NB), and linear discriminative analysis (LDA). Using the three types of classification schemes respectively, they obtained average recognition accuracies of 93.6%, 93.8%, and 91.8% using the manual features, and 94.8%, 90.8%, and 95.1% using the automatic features. Only the best recognition accuracy achieved in the two experiments are reported in Table 2.2. The high accuracy could be attributed to the use of the features in the neutral face for normalization, which significantly reduces the intra-class variation due to the personal face shape variation. The automatically selected features yielded better recognition accuracies which indicates the efficiency of their feature selection method.

Furthermore, Tang and Huang (2008b) extended their study to investigate the discrimination power of the slop of the line segments between the facial points. Using the same manual features in (Tang and Huang, 2008a), they form a new feature descriptor for every 3D face composed of the distance features and slops. However, they obtained a lower recognition accuracy of 87.10% compared to their previous work.

Several solutions have been proposed to overcome the person-dependent limitation. Soyel and Demirel (2007) selected 11 facial points to extract six characteristic distances, namely: eye opening, eyebrow height, mouth opening, mouth height, lip stretching, and face width. While the first five distances describe the shape deformation, the sixth distance is used to normalize the first five to eliminate the personal shape variation. The calculated distances are then used to train a multilayer perceptron neural network (MLP NN) with backpropagation learning method to classify the six expressions and *neutral*. In a continuous study, Soyel and Demirel (2008) extracted the same first five characteristic distances and replaced the sixth one by jaw opening, using 23 facial points inspired by FACS. They first transferred all the faces into a unified space with the same origin to avoid the personal variation and skip the normalization characteristic. Probabilistic neural network (PNN) is used for expression classification, which yielded less recognition accuracy of 87.80% compared to the first study using MLP NN, 91.30%. Later, they improved the recognition accuracy of PNN to 93.70% using coarse-to-fine classification (Soyel and Demirel, 2010). Decision tree PNN (DT-PNN) is established to first classify the seven expressions into three groups and then another fine classification is done within each group to classify the expressions into the seven classes.

Similarly, Tekgüç et al. (2009) normalized the distances between all the unique pairs connecting the 83 BU-3DFE facial points by the outer eye corners. They achieved less recognition accuracy of 88.18%, which almost equals to the accuracy achieved by Yurtkan and Demirel (2014) (88.28%) using the coordinates of 71 facial points without any normalization. However, Yurtkan and Demirel (2014) evaluated their system on the highest expressions intensity only whereas most of the studies in this literature review used at least the two highest intensities.

Srivastava and Roy (2009) argued that computing the residues/displacements is more efficient than computing the Euclidean distance between facial points. In their experiment, they calculated the displacements of the 83 facial points in the expressive 3D faces using the neutral face of the same subject as a reference. Even though they achieved a quite good recognition accuracy of 91.70%, which may support their argument, the proposed framework application is limited to person-dependent area.

Angles are another geometric attribute that can capture the shape changes between specified points. Rabiu et al. (2012) calculated the distances and the angles between a set of 29 facial points. For every 3D face, they formed one feature descriptor of 16 characteristic distances and 27 angles. Using support vector machine (SVM), they obtained 92.20% as overall recognition accuracy for the seven expressions. Moreover, they tested their framework on another dataset, which composed of 30 subjects from the BU-3DFE database and 30 subjects from the UPM-3DFE database (Habibu et al., 2012). The recognition accuracy obtained in the second experiment, 88.9%, is less than the accuracy obtained using the BU-3DFE subset. In Table 2.2, only the result obtained using the BU-3DFE database is reported for a fair comparison with other studies.

The lowest recognition accuracy achieved using the geometric-based approach is the accuracy reported by Sha et al. (2011), 76.75%. They studied the discriminative power of all the distances between the 83 BU-3DFE facial points and found that only 16 distances, which are normalized by the distance between the outer eyes corners, are discriminative to distinguish between the six expressions. Their experimental setup could be the reason behind the low recognition accuracy. They conducted 1000 independent experiments and the final average accuracy is the mean of the 1000 accuracies. With the aid of the 2D texture image, Venkatesh et al. (2009) automatically located 68 facial points on the 3D faces. Using different detection algorithms, i.e. active contour and Canny edge detection, the face components are located on the 2D images and the points around them are sampled. Only 68 out of all the possible sampled points are selected. As the 2D images and the 3D shapes of the same face model are in correspondence, the location of the selected points on the 3D shape could be acquired directly. Then, the displacements of the selected points in the expressive faces are calculated using the neutral face of the same subject. Modified PCA is used to construct six subspaces corresponding to the six expressions where any new face feature vector could be projected to these subspaces and the expression subspace with the maximum likelihood is chosen as the predicted expression.

Table 2.2: Geometric-based 3D facial expression recognition systems using the BU-3DFE database. All the systems are manual except (Venkatesh et al., 2009). PI: person-independent, #Exp: #expressions, #F: #extracted features, RA: recognition accuracy.

Reference	**Feature type**	**PI**	**#Exp**	**#F**	**Classifier**	**Experimental setup**	**RA**
Soyel and Demirel (2007)	Euclidean distance	yes	7	6	MLP NN	(54-6) subjects	91.30%
Soyel and Demirel (2008)	Euclidean distance	yes	7	6	PNN	(54-6) subjects	87.80%
Tang and Huang (2008a)	Euclidean distance	no	6	24	Adaboost with NB	(54-6) subjects-10 fold	93.80%
	Euclidean distance	no	6	NA	Adaboost with LDA	(54-6) subjects-10 fold	95.10%
Tang and Huang (2008b)	Euclidean distance and slops	no	6	96	SVM	(54-6) subjects-10 fold	87.10%
Tekgüç et al. (2009)	Euclidean distance	yes	7	NA	PNN	(48-12) subjects	88.18%
Srivastava and Roy (2009)	Displacement	no	6	83×3	SVM	(50-10) subjects	91.70%
Venkatesh et al. (2009)	Displacement	no	6	68	Modified PCA	(48-12) subjects	81.67%
Soyel and Demirel (2010)	Euclidean distances	yes	7	NA	DT-PNN	(48-12) subjectsj	93.70%
Ocegueda et al. (2011)	Coordinates	no	6	1	linear regression	(90-10) subjects-10 folds	90.40%
	Normals						84.40%
Sha et al. (2011)	Euclidean distances	yes	6	16	k-NN	(54-6) subjects-1000 times	76.75%
Rabiu et al. (2012)	Euclidean distance and angles	yes	7	43	SVM	(54-6) subjects-10 folds	92.20%
Yurtkan and Demirel (2014)	coordinates	yes	6	71	2 levels SVM	(90-10) subjects	88.28%

Compared to other studies followed the geometric approach, they archived lower recognition accuracy of 81.67%; however, the work is fully automatic.

II. Local Feature-based Approach

The common pipeline of the local feature approach includes detecting a set of facial points that assists extracting the salient regions/patches of interest. The regions are either located between the detected facial points or centered at these points. The feature descriptor of one face is a concatenation of all the extracted features from the regions of interest. This approach possesses some advantages over the two other approaches. As the features are extracted from regions, minor displacement of the detected facial points could be tolerated, while in geometric-based approach such minor displacements could affect the recognition accuracy significantly. Moreover, compared to the global feature-based approach, local feature-based approach is less sensitive to the global variations like personal variations or the appearance of distinctive parts as scars or mustache. However, similar to geometric-based approach, the major disadvantage is the requirement of locating a set of facial points manually or automatically. As stated earlier, the studies in this approach and the global feature-based approach are going to be categorized into two categories: using 3D face images and using 3D-into-2D mapped face images.

One of the most popular methods in this approach has been to use the curvature analysis on the facial surface as it can efficiently describe the local facial shape deformation (Tanaka et al., 1998). Wang et al. (2006), who are the first researchers utilized the BU-3DFE database for 3D facial expression recognition, considered the face as a terrain map, which could be described with a set of 12 topographic descriptors/primitive labels, namely: peak, pit, flat, ravine, ridge, ridge saddle, ravine saddle, convex hill, concave hill, concave saddle hill, convex saddle hill and slope hill. The face is segmented into 7 expressive regions with the aid of a set of 64 manually located facial points. In each segment, the two principle curvatures are first calculated for every vertex in the face mesh so the primitive labels could be calculated. Next, the distributions of the labels in the 7 expressive regions are concatenated to form one feature vector for each face. Four different classifiers have been evaluated: quadratic discriminant classifier (QDC), LDA, NB , and SVM which yielded recognition accuracies of 74.5%, 83.6%, 71.7%, and 77.8% respectively.

Similarly, Sha et al. (2011) described the face as a terrain map, but with fewer primitive labels and more face segments, eight shape primitives and 67 triangular regions. The segmentation is done using 43 facial points. However, they obtained less recognition accuracy of 70.00% compared to the accuracy achieved by Wang et al. (2006). The recognition accuracy is then improved by combining the geometric features, as described in Section 2.3.3, and the primitive label descriptors in one feature vector. Using LDA the accuracy increased up to 83.5%.

Li et al. (2011) computed the shape index and the gradient at each vertex in the regions of interest using two different methods: the normal cycle theory (Alliez et al., 2003) and local cubic fitting method (Goldfeather and Interrante, 2004). The histogram of gradient (HoG) and the histogram of shape index (HoS) are extracted from regions around 60 selected landmarks and then sent to two types of SVM: linear SVM and kernel SVM. The classification using HoG with linear SVM yielded the lowest recognition accuracy of 76.4%, whereas the combined features HoG+HoS with the linear SVM yielded the highest accuracy of 82.01%. However, the results showed that there is no significant difference between the discriminative powers of the features extracted using normal cycle theory or local cubic fitting method.

Maalej et al. (2011) applied a novel curvature-based shape descriptor introduced by (Samir et al., 2009). The method defines a height function, based on the z coordinates, to identify level sets of 3D closed curves centered at every selected facial landmark, such that all the points on one curve has a similar distance from the center. Maalej et al. (2011) extracted the 3D curves around 70 landmarks: 68 points from the facial point set provided by the BU-3DFE database and 2 are calculated as the middle point between the mouth corner and the outer eye corner. In the training phase, first, six expressive faces belong to one subject are fixed as references for shape analysis. Then, the corresponding curves in all level sets extracted from the 3D faces in the training set and the references are compared by calculating the shortest path between them. The sum of the distances of all the curves in the same region formed the feature of the corresponding landmark. The final feature vector calculated from all the landmarks are then sent to SVM and multi-Adaboost algorithm with three classifiers: LDA, NB and k-NN, which achieved recognition accuracies of 97.75%, 98.81%, 98.76% and 98.07% respectively. Adaboost with LDA achieved the highest recognition accuracy; however, the difference between