Reconnaissance d'Expressions Faciale 3D Basée sur l'Analyse de Forme et l'Apprentissage Automatique

Ahmed Maalej

Numéro d'ordre: 40818

université Lille 1 Sciences et Technologies

Laboratoire d'Informatique Fondamentale de Lille

Ecole doctorale Sciences Pour l'Ingénieur Université Lille

Thèse

Présentée en vu d'obtenir le grade de Docteur,
spécialité Informatique

par

Ahmed Maalej

3D Facial Expressions Recognition Using Shape Analysis and Machine Learning

Thèse soutenue le 23 Mai 2012 devant le jury composé de :

M.	Liming Chen	Professeur, Ecole Centrale de Lyon, France	(Examinateur)
M.	Zhengyou Zhang	Principal Researcher, Microsoft Research Redmond, USA	(Examinateur)
M.	Bülent Sankur	Professeur, Bogazici University Istanbul, Turkey	(Rapporteur)
M.	Christophe Garcia	Professeur, INSA Lyon, France	(Rapporteur)
M.	Stefano Berretti	Professeur, University of Firenze, Italy	(Examinateur)
M.	Mohamed Daoudi	Professeur, TELECOM Lille 1 / LIFL Lille, France	(Directeur)
M.	Boulbaba Ben Amor	Maître de Conférences, TELECOM Lille 1 / LIFL Lille, France	(Co-encadrant)

Titre Reconnaissance d'Expressions Faciale 3D Basée sur l'Analyse de Forme et l'Apprentissage Automatique

Résumé La reconnaissance des expressions faciales est une tâche difficile, qui a reçu un intérêt croissant au sein de la communauté des chercheurs, et qui impacte les applications dans des domaines liés à l'interaction homme-machine (IHM). Dans le but de construire des systèmes IHM approchant le comportement humain et émotionnellement intelligents, les scientifiques essaient d'introduire la composante émotionnelle dans ce type de systèmes. Le développement récent des capteurs d'acquisition 3D a fait que les données 3D deviennent de plus en plus disponibles, et ce type de données vient pour remédier à des problèmes inhérents aux données 2D tels que les variations d'éclairage, de pose et d'échelle et de faible résolution. Plusieurs bases de données 3D du visage sont publiquement disponibles pour les chercheurs dans le domaine de la reconnaissance d'expression faciale leur permettant ainsi de valider et d'évaluer leurs approches. Cette thèse traite le problème la reconnaissance d'expressions faciale et propose une approche basée sur l'analyse de forme pour la reconnaissance d'expressions dans des cadres de données 3D statiques et 3D dynamiques. Tout d'abord, une représentation du modèle 3D du visage basée sur les courbes est proposée pour décrire les traits du visage. Puis, utilisant ces courbes, l'information de forme qui leur est liée est quantifiée en utilisant un cadre de travail basé sur la géométrie Riemannienne. Nous obtenons ainsi des scores de similarité entre les différentes formes locales du visage. Nous constituons, alors, l'ensemble des descripteurs d'expressions associées à chaque surface faciale. Enfin, ces descripteurs sont utilisés pour la classification d'expressions moyennant des algorithmes d'apprentissage automatique. Des expérimentations exhaustives sont alors entreprises pour valider notre approche. Des résultats de taux de reconnaissance d'expressions de l'ordre de 98.81% pour l'approche 3D statique, et de l'ordre de 93.83% pour l'approche 3D dynamique sont alors atteints, et sont comparés par rapport aux résultats des travaux de l'état de l'art.

Mots-clés Reconnaissance d'expressions faciale 3D, géométire riemannienne, apprentissage automatique.

Title 3D Facial Expressions Recognition Using Shape Analysis and Machine Learning

Abstract Facial expression recognition is a challenging task, which has received growing interest within the research community, impacting important applications in fields related to human computer interaction (HCI). Toward building human-like emotionally intelligent HCI devices, scientists are trying to include human emotional state in such systems. The recent development of 3D acquisition sensors has made 3D data more available, and this kind of data comes to alleviate the problems inherent in 2D data such as illumination, pose and scale variations as well as low resolution. Several 3D facial databases are publicly available for the researchers in the field of face and facial expression recognition to validate and evaluate their approaches. This thesis deals with 3D facial expression recognition (FER) problem and proposes an approach based on shape analysis to handle both 3D static and 3D dynamic FER tasks. Our approach includes the following steps: first, a curve-based representation of the 3D face model is proposed to describe facial features. Then, once these curves are extracted, their shape information is quantified using a Riemannian framework. We end up with similarity scores between different facial local shapes constituting feature vectors associated with each facial surface. Afterwards, these features are used as entry parameters to some machine learning and classification algorithms to recognize expressions. Exhaustive experiments are derived to validate our approach and recognition results of 98.81% for 3D FER and 93.83% for 4D FER are attained and are compared to the related work achievements.

Keywords 3D Facial expression recognition, Riemannian geometry, machine learning.

Acknowledgements

First of all, I would like to express my gratitude to my thesis advisor, Pr. Mohamed Daoudi, for giving me the chance to conduct my research at the Multimedia, Image, Indexing and Recognition (MIIRE) research group. He directed me to pursue my Ph.D., helped me to go into an interesting research field and whose experience and understanding added considerably to my graduate experience.

I would like to thank my co-advisor Dr. Boulbaba Ben Amor for his advice and guidance from the very early stage of this research, he has been supportive and has also provided insightful discussions about the research subject.

A special thanks is due my Ph.D. committee members, Pr. Liming Chen, Pr. Zhengyou Zhang and Pr. Stefano Berretti for taking the time to participate in this process, and a special thank to the reviewers of the manuscript, Pr. Bülent Sankur and Pr. Christophe Garcia, for having accepted this significant task and kindly accepted to sacrifice their time for the manuscript reviewing.
I would also like to thank Pr. Anuj Srivastava, and Stefano Berretti for their fruitful collaboration and for sharing their expertise, knowledge and many insightful discussions and suggestions.

Many thanks go to the people that I have been working with, the current and former members of the MIIRE research group of the Computer Science Research Laboratory (LIFL) of the Lille 1 University.

Special thank to my friends and collegues, Hassen Drira, Lahoucine Ballihi, Hedi Tabia, Halim Benhabiles and Rim Slama for their assistance,

support and their contributions to the enjoyment of my student life. I also gratefully acknowledge the institutional support that I have received while working on this thises. In particular, I thank the region of Nord-Pas-de-Calais, Lille and Telecom Institut for my thesis funding.

I would like to thank my mother, my brothers and my sisters for their support and love, and I am most grateful to my father who passed away four years ago.

I most want to thank my wife Hayet Akkari for her love, sacrifice, and kind indulgence.

Author's publications

International publications

Journal

- **Ahmed Maalej**, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava, Stefano Berretti, "*Shape Analysis of Local Facial Patches for 3D Facial Expression Recognition*", **Pattern Recognition (PR)** 44(8): 1581-1589 February 2011.

Conferences

- **Ahmed Maalej**, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava, Stefano Berretti, "*Local 3D Shape Analysis for Facial Expression Recognition*", **International Conference on Pattern Recognition (ICPR)**, Oral presentation, Istanbul, Turkey 23-26 august 2010.

- Hassen Drira, **Ahmed Maalej**, Boulbaba Ben Amor, Stefano Berretti, Mohamed Daoudi, Anuj Srivastava, "*A Dynamic Geometry-based Approach for 4D Facial Expressions Recognition*", **Submitted to the European Conference on Computer Vision (ECCV) 2012**.

Local publication

Conference

- **Ahmed Maalej**, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava, Stefano Berretti, "*Analyse locale de la forme 3D pour la reconnaissance d'expressions faciales*", **13ème édition d'ORASIS, Journées Francophones des Jeunes Chercheurs en Vision par Ordinateur, Juin 2011**.

- **Ahmed Maalej**, Boulbaba Ben Amor, Mohamed Daoudi, "*Analyse locale de la forme 3D pour la reconnaissance d'expressions faciales*", **Traitement et Analyse de l'Information Méthodes et Applications (TAIMA) 2011**.

Contents

List of Figures

Notations

Table 1 – **List of symbols and their definitions used in this thesis.**

Symbol	Definition /Explanation
S	a facial surface.
p	a point on S.
N	the number of landmarks on S.
r_l	the l^{th} landmark on S where $l \in]0, N]$.
λ	a variable for the value of the distance from the landmark r_l
β	a parametrized curve.
$\mathbb{S}^1$	the unit circle.
$q(s)$	the scaled velocity vector, $q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}$.
$\langle a, b \rangle$	the Euclidean inner product in $\mathbb{R}^3$
$SO(3)$	group of rotation in $\mathbb{R}^3$.
$[q]$	equivalence class of curves under rotation and re-parametrization.
$\mathcal{C}$	the set of all curves in $\mathbb{R}^3$, (closed curves in chapter 3 and open curves in chapter 4).
E	energy function $E[\alpha] = \frac{1}{2} \int_s \langle \dot{\alpha}(s), \dot{\alpha}(s) \rangle \, ds$.
d_c	distance on $\mathcal{C}$, $d_c(q_1, q_2) = cos^{-1} \langle q_1, q_2 \rangle$.
$\mathcal{S}$	the shape space: $\mathcal{S} \doteq \mathcal{C}/(SO(3) \times \Gamma)$.
d_s	distance on $\mathcal{S}$, $d_c([q_1], [q_2])$.
$T_v\mathcal{S}$	the space of all tangents to $\mathcal{S}$ at v
$\langle f_1, f_2 \rangle$	the Riemannian metric on $\mathcal{C}$, $\int_0^1 f_1(s) f_2(s) ds$
$\alpha(\tau)(q_1, q_2)$	a geodesic path in $\mathcal{C}$, from q_1 to q_2, parameterized by $\tau \in [0, 1]$
$\mathcal{S}^{[0,\varphi_0]}$	indexed collection of radial curves (a face).
$T_q(\mathcal{S}^{[0,\varphi_0]})$	tangent space at q on $\mathcal{S}^{[0,\varphi_0]}$

1 Introduction

1.1 Research topic

The study on facial expressions has consistently been an active topic and can be traced back to the 17th century. The English physician and philosopher John Bulwer wrote a book on physiognomy titled *Pathomyotomia* in 1649 and it turned out to be the first substantial work on the muscular basis of emotional expressions, in which the author confronted many issues concerning the nature of the emotions and their relationship to facial movements. It is reported that the same author published other books, each of which dealt with theoretical and applied aspects of various forms of nonverbal communication. Perhaps the most influential of these early theorists was Charles Darwin. In his book *The Expression of the Emotions in Man and Animals*, Darwin (1896) argued that facial expressions were universal and innate characteristics. Moving on to the 20th century, one of the important works on facial expression analysis that has a direct relationship to the modern day science of automatic facial expression recognition was the work done by the well-known psychologist Paul Ekman and his colleague Wallace Friesen since the 1970s. Ekman originally developed Facial Action Coding System (FACS) manual, which is a detailed technical guide that explains how to categorise facial behaviors based on the muscles that produce them. Dr. Paul Ekman's research along with the work of Silvan Tomkins in the study of emotions and their relation to facial expressions took Darwin's work to the next level proving that facial expressions of emotion are not culturally determined, but biological in origin and universal across human cultures. Facial expression is considered as one of the types of nonverbal communication. It helps coordinate conversation and communicate emotions and other meaningful mental, social, and psychological cues. Albert Mehrabian, who is a professor of psychology at the University of California (UCLA), has become known best by his publications on the relative importance of verbal and nonverbal messages, reported that face to face communication is governed by the 7%-38%-55% rule: three elements, often abbreviated as the 3**Vs** for Verbal, Vocal and Visual, account differently in human communication:

1. Words account for 7%
2. Tone of the voice accounts for 38%
3. Body language (e.g gesture, posture, facial expression) account for 55%

This rule point out that the non-verbal elements are particularly important for communicating feelings and attitude, even in case where the words disagree with the tone of the voice and body and facial behaviours, people tend to believe the tonality and nonverbal behaviours. Besides emotional intelligence or quotient EQ is a relatively recent concept defined as a new measurement of human intelligence in addition to the traditional intellectual quotient IQ. This recent behavioural model was introduced by the psychologist Daniel Goleman in his book *Emotional intelligence: Why it can matter more than IQ*. In this book the author emphasises on the fact that the EQ is one of the facets of human intelligence and plays an important role for successful interpersonal social and professional interactions.

Traditionally, facial expressions have been studied by clinical and social psychologists, medical practitioners. However in the last quarter of the 20th century, with the advances in the fields of robotics, computer graphics and computer vision, animators and computer scientists started showing interest in the study of facial expressions.

1.2 Definition

Facial expression is defined to be a gesture that is executed with the facial muscles, these muscles, called mimetic muscles, are innervated by facial nerves, that control facial expression throw impulses of the brain. Fasel and Luttin define facial expressions as temporally deformed facial features such as eye lids, eye brows, nose, lips and skin which result in temporary facial deformations in both facial geometry and texture. The study of facial expression cannot be done without the study of the anatomy of the face and the underlying structure of facial muscles.

1.2.1 Etymology

The English word emotion is derived from the French word émouvoir. This is based on the Latin *emovere*, where e- (variant of ex-) means *out* and movere means *move*, what suggests to associate emotion with action or move (body language, facial muscles, posture).

1.2.2 Neo-Darwinism

The evolutionary perspective has its origin in the work of Darwin [DEP02]. He studied mainly the communicative function of emotions by giving predominance to the facial expressions. Charles Darwin in 1872, was among the first to pay attention to emotional phenomena by issuing an extension of the evolutionary analysis of the living universe, a book entitled *the expression of emotions in man and animals*. For Darwin, the emotional state of an adult is a reflection of the continuity complex behavioral systems derived from other animal species [Chr98]. Darwin used three basic principles in order to explain his approach:

- Patterns associated with emotional expressions are behind acts utilities that would fulfill an adaptive function in relation to the environment;
- The antithesis emotional states are often characterized by motor manifestations antagonists;
- Direct action on the brain effect of overflow and diversion of nervous energy generated by the stimulation.

The neo-Darwinian concept have mainly focused on identifying emotions by studying the basic emotional facial expressions. The various categorisations of basic emotions proposed in the literature indicate that there are significant differences between the authors. These various theoretical approaches have in common a focus on the relationship between expressive facial configuration and a specific emotion.

1.3 Theories of emotions

Theories of emotion attempt to explain what emotions are and how they operate. Emotions are sophisticated and subtle and can be analysed from many different perspectives. Social theories explain emotions as the products of cultural and social conditioning, since that the way we, humans, express our emotions reflect our social environment. The early work on emotions was made by Darwin who had been collecting material since 1838. His intention was to show how the expressions of the emotions in man were analogous to those in animals, supporting his theory that man and animals were derived from a common ancestor.

1.3.1 James-Lange theory

The two psychologists William James [Jam07] and Carl Lange [LJH22] argued that emotional experience is largely due to the experience of bodily changes and states. According to their theory, emotions occur as a result of physiological reactions to events. You see an external stimulus that leads to a physiological reactions such as sweating, trembling and muscle tension. Your emotional reaction is dependent upon how you interpret those physical reactions. So emotion is a result of experiencing and interpreting a physical response.

For example, you are walking down a dark alley late at night. You hear footsteps behind you and you begin to tremble, your heart beats faster, and your breathing deepens. You notice these physiological changes and interpret them as your body's preparation for a fearful situation. You then experience fear.

1.3.2 Cannon-Bard Theory

The Cannon-Bard theory [Can27] also known as the thalami theory, suggests that for a given stimulus we have both physiological and emotional response simultaneously, and that neither one causes the other. These actions include changes in muscular tension, perspiration, etc. This theory challenges the James-Lange theory of emotion introduced before.

For example, you are walking down a dark alley late at night. You hear

footsteps behind you and you begin to tremble, your heart beats faster, and your breathing deepens. At the same time as these physiological changes occur you also experience the emotion of fear.

1.3.3 Schachter-Singer Theory

According to this theory [SS62], an event causes physiological arousal first. You must then identify a reason for this arousal and then you are able to experience and label the emotion.
For example, you are walking down a dark alley late at night. You hear footsteps behind you and you begin to tremble, your heart beats faster, and your breathing deepens. Upon noticing this arousal you realize that it comes from the fact that you are walking down a dark alley by yourself. This behavior is dangerous and therefore you feel the emotion of fear.

1.3.4 Lazarus Theory

In his theory [Laz91], Lazarus argued that before emotion occurs, people make an automatic, often unconscious, assessment of what is happening and what it may mean for them or those they care about. From that perspective, emotion becomes not just rational but a necessary component of survival. In other words, a thought must come before any emotion or physiological arousal. Meaning that you must first think about your situation before you can experience an emotion.
For example, You are walking down a dark alley late at night. You hear footsteps behind you and you think it may be a mugger so you begin to tremble, your heart beats faster, and your breathing deepens and at the same time experience fear.

1.3.5 Facial Feedback Theory

According to the facial feedback theory [AD08], emotion is the experience of changes in our facial muscles. In other words, when we smile, we then experience pleasure, or happiness. When we frown, we then experience sadness. It is the changes in our facial muscles that cue our brains and provide the basis of our emotions. Just as there are an unlimited number

of muscle configurations in our face, so to are there a seemingly unlimited number of emotions.

For example, you are walking down a dark alley late at night. You hear footsteps behind you and your eyes widen, your teeth clench and your brain interprets these facial changes as the expression of fear. Therefore you experience the emotion of fear.

1.4 Neurophysiology of Emotions: The limbic system

The limbic system is the part of the brain that appears to be most directly involved in human emotion generation and regulation. It is the system that is primarily responsible for our emotional life. Fig. 1.1 shows a design of the limbic system comprising a network of interconnected structures that control the emotional state. The main structures include the cingulate cortex, hippocampus, the amygdala and its extended connections with the hypothalamus and the cortex, the mammillary body the hypothalamus and prefrontal cortex. The hypothalamus is responsible for certain metabolic processes and other activities of the autonomic nervous system. It is identified as the source of motivational behaviours, that consists in a tendency to feel or desire (a.g hunger, thirst, sexual satisfaction etc.)that energizes and directs behavior in order to fulfill it. While the limbic system generates emotions. The sensory and prefrontal regions make contacts with the cingulate cortex, the hippocampus and amygdala. The last two structures make connections the hypothalamus, which in turn, by the thalamus, establishes connections with the cingulate cortex.

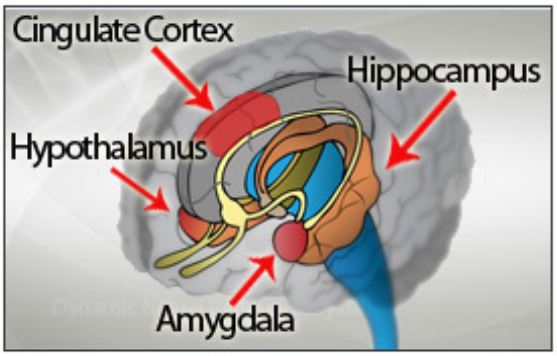

Figure 1.1 – *Limbic system.*

1.5 Types of emotions

Emotions are strong and common sequences that occur throughout our daily life and set on the tone. Emotions allow us to experience the heights of excitement, joy, and love, as well as the depths of anguish, guilt, and sorrow. Emotions color our worlds and are the foundation of our basic humanity. When you say, " I feel anxious, angry, happy...", you are interpreting the emotion by how it feels to your mind and body. We can distinguish two main types of emotions and they are listed as follows.

1.5.1 Primary emotions

Primary or basic emotions are those that we feel first, as a first response to a situation and a direct result of encountering some kind of cue. For example, if you have a memory come up about losing someone you care about, the primary emotion that will likely come up is sadness. Likewise, if someone cuts you off in traffic, you will likely experience anger or irritation. Anger and irritation, in this case, would be considered a primary emotion because the emotion occurred as a direct consequence of encountering some kind of event (being cut off in traffic).

1.5.2 Secondary emotions

They are the emotional reactions and response to the primary emotions. For example, with going back to the example of anger mentioned above. You have the primary emotional response of anger. However, let's say that you were brought up to believe that it is not OK to be angry, or when you feel anger, you think you are going to lose control and do something impulsive. If you evaluate your primary emotional response of anger in this way, you are likely going to feel shame or anxiety in response to being angry. Consequently, secondary emotions don't pass that quickly. They tend to stick around for a long time. They are slow-acting emotions.

1.5.3 Tertiary emotions

Compared to secondary emotion, tertiary emotions involve something more, that is a partial loss of control, of attention and thought processes (perturbance), e.g. states of feeling humiliated, infatuated, guilty, or full of excited anticipation, where attempts to focus attention on urgent or important tasks can be difficult or impossible, because attention is drawn back to the focus of the humiliation or infatuation, etc. Sloman [Slo02] argues that tertiary emotions can exist with no physiological involvement and they arise from conflicts among cognitive processes, not from any emotion mechanism. In Tables 1.1 and 1.2 we summarize different types of emotions into a short tree structure as categorised by Plutchik [Plu91, Plu03].

Primary Emotion	Secondary Emotion	Tertiary Emotion
Love	Affection	Adoration, affection, love, fondness, liking, attraction, caring, tenderness, compassion, sentimentality
	Lust	Arousal, desire, lust, passion, infatuation
	Longing	Longing
Joy	Cheerfulness	Amusement, bliss, cheerfulness, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria
	Zest	Enthusiasm, zeal, zest, excitement, thrill, exhilaration
	Contentment	Contentment, pleasure
	Pride	Pride, triumph
	Optimism	Eagerness, hope, optimism
	Enthrallment	Enthrallment, rapture
	Relief	Relief
Surprise	Surprise	Amazement, surprise, astonishment
Anger	Irritation	Aggravation, irritation, agitation, annoyance, grouchiness, grumpiness
	Exasperation	Exasperation, frustration
	Rage	Anger, rage, outrage, fury, wrath, hostility, ferocity, bitterness, hate, scorn, spite, vengefulness, dislike, resentment
	Disgust	Disgust, revulsion, contempt, loathing
	Envy	Envy, jealousy
	Torment	Torment

Table 1.1 – *Types of emotions (1).*

Primary Emotion	Secondary Emotion	Tertiary Emotion
Sadness	Suffering	Agony, suffering, hurt, anguish
	Sadness	Depression, despair, hopelessness, gloom, glumness, sadness, unhappiness, grief, sorrow, woe, misery, melancholy
	Disappointment	Dismay, disappointment, displeasure
	Shame	Guilt, shame, regret, remorse
	Neglect	Alienation, isolation, neglect, loneliness, rejection, homesickness, defeat, dejection, insecurity, embarrassment, humiliation, insult
	Sympathy	Pity, sympathy
Fear	Horror	Alarm, shock, fear, fright, horror, terror, panic, hysteria, mortification
	Nervousness	Anxiety, nervousness, tenseness, uneasiness, apprehension, worry, distress, dread

Table 1.2 – *Types of emotions (2).*

1.6 Emotion Representation

The very first issue encountered when studying facial expressions, is the emotion representation task. There is a need to find a formalism that is consistent with existing psychological studies dealing with emotions, while allowing easy handling for human computer interaction (HCI) systems. For such systems, the goal is to assign category labels that identify emotional states. However, such labels are poor descriptions, especially since humans use an overwhelming number of labels to describe emotion. Therefore there is a need to incorporate a more transparent, as well as continuous representation, that matches closely our conception of what emotions are or, at least, how they are expressed and perceived. There are a number of choices available for representing [PH06]. In the following, a selection of such representations is presented.

1.6.1 Categorical representation

Categorical representation is a discrete approach, claiming that basic emotions are universal and can therefore be found in all cultures [Plu80, ED94]. Several psychologists have suggested a different number of these, ranging from 2 to 18 categories, but there has been considerable agreement on the following six: anger, disgust, fear, happiness, sadness and surprise. Several arguments for the existence of these categories have been provided, like distinct universal facial signals, distinct universals in antecedent events, presence in other primates etc. Ekman (and also other researchers) based his assumptions mainly on the facial expression of emotions. In his studies, facial expressions of emotions were recognised by people from very different cultures. Table 1.3 presents different emotions as categorised by several Psychologists.

1.6.2 Dimensional representation

The dimensional representation uses dimensions rather than discrete categories to describe the structure of emotions. This approach, assumes the existence of two or more major dimensions which are able to describe different emotions and to distinguish between them [Rus80]. According to a

Psychologists	Categories of emotions
Ekman et al. [EFE72]	Anger, disgust, fear, happy, sad, surprise.
Izard [Iza77]	Anger, contempt, disgust, anxiety, fear, shame, interest, happy, surprise, guilt.
Plutchik [Plu80]	Acceptation, anger, anticipation, disgust, fear, happy, sad, surprise.
Tomkins [Tom62, Tom63]	Anger, interest, contempt,disgust, anxiety, fear,happy, shame, surprise.
Watson [Wat59]	Fear, love, rage
Mowrer [Mow60]	Pain, pleasure
James [Jam07]	Fear, grief, love, rage
Frijda [Fri86]	Desire, happiness, interest, surprise, wonder, sorrow

Table 1.3 – *Emotion categorization.*

dimensional view, all emotions are characterised by their two dimensions named valence and arousal. The dimension of valence ranges from highly positive to highly negative, whereas the dimension of arousal ranges from calming or soothing to exciting or agitating. Fig. 1.2 illustrates the dimensional representation of emotions as seen by Russell [Rus80].

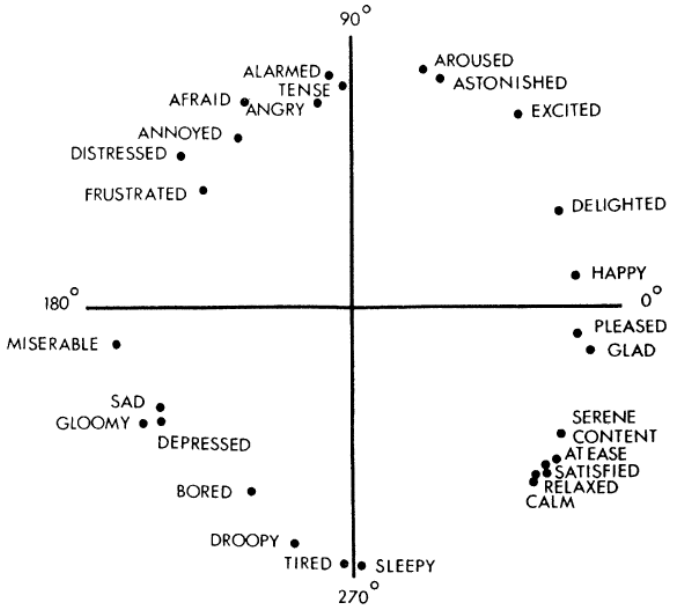

Figure 1.2 – *Circumplex model of affect (taken from Russell [Rus80]).*

Plutchik believed that there were 8 basic categories of emotions and stated that all other emotions evolved from these 8 basic emotions. He illustrated the relationships of one emotion to another in a wheel diagram that shows his multidimensional model, see Figure 1.3. Where eight primary emotions are presented on a section in the horizontal plane. On a vertical section, are reported different intensities of the same primary emotion (eg, apprehension, fear and terror). Far from being opposed, both approaches, categorical and dimensional, are complementary for the study of emotions.

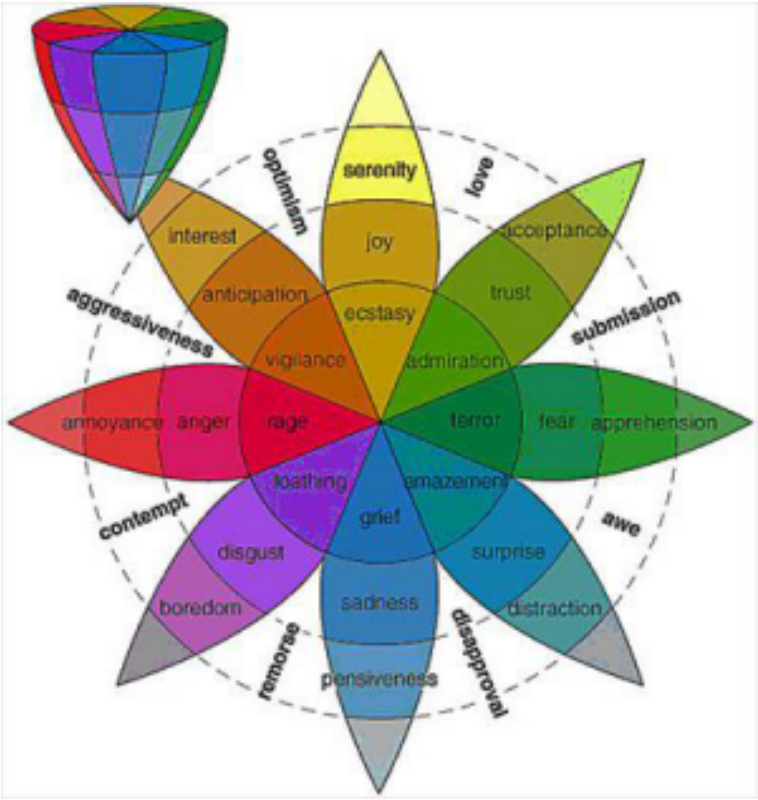

Figure 1.3 – *Wheel diagram showing multidimensional model for representing emotions (taken from Plutchik [Pluo3]).*

1.6.3 Action units representation

In their manual called FACS, Ekman and Friesen described visually distinguishable facial movements that are due to facial expressions. Their system is based on the enumeration of all "'action units"' (AUs) of a face. These AUs refer a measuring of specific facial muscle movements, and are anatomically related to their contraction. Thus the anatomy of the face is

significant for understanding the behaviors and appearances of the face and head. The anatomy of facial muscles is most directly related to facial expression, as the muscles underlie these appearance changes. The Figure 1.4 depict the muscles of the face and head that are involved in a dynamic process to convey different expressions.

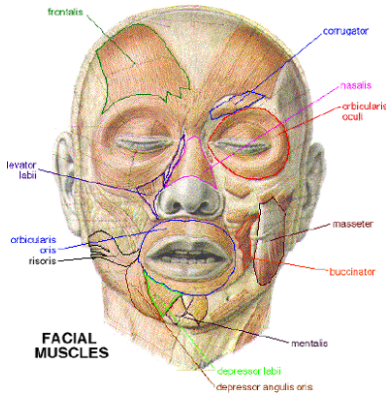

Figure 1.4 – *Facial Muscles.*

Listed below are the facial muscles that are subcutaneous (just under the skin) and that control facial expression. For a deeper study we refer you to the DataFace free to view web site [Hag], a creation of Joseph C. Hager, where anyone can learn the scientific basics of the face and facial expression, as well as its mythology and folklore.

1.7 Emotion Components

We can distinguish three basic components to define emotion: (1) the physiological component, (2) the behavioral component and (3) the cognitive component. As for the physiological aspect, the mental state that arises spontaneously rather than through conscious effort is often accompanied by active changes in the body physically. Consequently a neurophysiological activity emerges and establishes a adaptive function attributed to the emotion. Hypothalamic center that is a part of the sympathetic nervous system controls vegetative activation. The measurement of the latter can be very valuable in the study of emotions and the brain connections leading to that. As for the behavioral component, emotion can be revealed by a set of behavioral traits through which it reveals itself, voice

Upper face muscles	Eye muscle	Lower face muscles
frontalis: the forehead corrugator: the brow nasalis: the nose	obicularis oculi: around the eye	levator labii: raises the upper lip masseter: closes the jaw Obicularis oris: purses the lips risoris: draws the lips in a smile buccinator: pulls the lips wide and tight depressor labii: lowers the lower lips depressor anguli oris: lowers the bottom corner of the lips levator anguli oris (not shown): raises the upper corner of the lips pterigoid (not shown): pulls jaw back or shut mentalis: pulls chin down

Table 1.4 – *Facial anatomy.*

tone and the facial expressions are examples of these traits. The main function of emotional expression is a language to generate detectable by other individuals. As for the cognitive component, it emphasizes on the importance of thoughts, beliefs, and expectations in determining the type and intensity of emotional response. The cognition attached to a situation determines which emotion is felt in response to physiological arousal.

1.8 Applications

As machines are becoming more and more involved in everyday human life and take part in both his living and work spaces, they need to become more intelligent in terms of understanding the human's moods and emotions. Embedding these machines with system capable of recognizing human emotions and mental state is precisely what the Human Computer Interaction research community is focusing on in the Affective Computing and humane machine interaction communities. The automatic facial expression recognition systems find applications in several interesting areas:

- Human-machine interface: facial expressions is a way of communication as many other ways (e.g. speech signal). The emotional detection is natural for humans but it is very difficult task for machines. Therefore the purpose of emotion recognition system is to use emotion related knowledge in such a way that human machine communication can be improved and make machines and robots more human-like.

- medical care and cure field: facial expressions are the direct means to identify when specific mental processes (e.g., pain, depression) are occurring.

- psychological field: where expressions detection are tremendously useful for the analysis of the human psychology.

- security field: where decoding the language of micro-expressions is crucial for establishing or detracting from credibility, and to determine any deception from suspects during interrogations. This is due

to the fact that micro-expression is a momentary involuntary facial expression that people unconsciously display when they are hiding an emotion.

- education field: pupils facial expressions inform the teacher of the need to adjust the instructional message.

1.9 Thesis contributions

Endowing machines with systems that are capable of detecting and recognizing expressions has been, for decades, a challenging task for a large variety of scientific communities, such as in signal processing, image processing, artificial intelligence, computer vision, robotic, etc. This PhD thesis brings two main contributions; the first one is related to the static 3D facial expression recognition and the second one is directed toward dynamic 3D facial expression recognition.

1. Static 3D facial expression recognition: we were interested in this special task and we have proposed a new approach for facial expression recognition using 3D static data. The main idea of this approach is to apply a Riemannian framework to derive statistical analysis on shapes of facial features. Using these features, that are actually local 3D surfaces extracted from 3D face scan and designated as patches, we were able to quantify the shape of these local patches. Taking two corresponding patches extracted from the same region belonging to two face models, a similarity/dissimilarity measure can be computed to characterize their shapes. These measures are then used as entry parameters to a machine learning techniques in order to derive classification procedures and determine the appropriate expression.

2. Dynamic 3D facial expression recognition: we proposed somehow a different approach from the one proposed for static 3D FER, to tackle the dynamic 3D FER problem, this approach tends to alleviate limitations of the first one, and could be more suitable for real

world applications. In this approach, we adopted a slightly different representation of the 3D facial surface. Using this representation we applied differential geometry in order to compute scalar fields that capture deformations due to expressions generated from two consecutive frames of a 3D image sequences. Hidden Markov Models are then applied to train and learn these deformations along the sequences. The performance of our method were evaluated and expression classification results were reported.

Related Work

2

2.1 Introduction

Facial expression recognition has been extensively studied over the past decades especially in 2D domain (e.g., images and videos) resulting in a valuable enhancement. Existing approaches that address facial expression recognition can be divided into two categories: (1) *static* vs. *dynamic* (2) *2D* vs. *3D*. To analyze facial expression there are, in general, two stage process to derive:

- facial feature detection and extraction
- facial expression classification

Facial feature extraction attempts to find the most appropriate representation as well as valuable information of the face images for expression recognition. Feature components are used as feature vectors for recognizing facial expression. Different approaches have been proposed to extract these facial features from images or video sequences [HL01]. A description of these approaches is given as follows.

2.1.1 Local Approaches

Local-based approaches use geometric information such as key points relative positions, distances and sizes of the face components or regions like eyes, nose, mouth eyebrows, and other potential facial characteristic points (FCPs) as features measures. These measures are computed, and are usually normalized using facial parameters. Thus geometric features present the shape and locations of facial components, which are extracted to form a feature vector that represents the face geometry. Many approaches in facial feature detection [YKA02] and eye detection [RKS10] have get encouraging results. However, the Local-based approaches usually requires accurate and reliable facial feature detection and tracking, which is difficult to accommodate in many situations.

2.1.2 Holistic Approaches

In holistic approaches, template can be a pixel image, a 3D face model or a feature vector obtained after processing the face image as a whole. These

approaches are based on developing a generic face model or template. The construction of such a representation involves three major steps; shape and texture data acquisition, data normalization and statistical analysis through principal component analysis. The appearance-based face model needs to be represented by a compact coding. For this purpose, statistical redundancy reduction principles are used. Among these principles we can mention the unsupervised learning techniques such as principal component analysis (PCA), independent component analysis (ICA), kernel principal component analysis, local feature analysis, and probability density estimation, as well as supervised learning techniques such as average, multi layer perceptron (MLP), multi-linear analysis, linear discriminant analysis (LDA) and kernel discriminant analysis (KDA). The Gabor wavelets are also used to extract the facial feature vector. Most of these techniques projects the data images into eigenspace that encodes the variation among known face images. This gives eigenvectors of the set of faces, which they do not necessarily correspond to isolated features such as eyes, ears, and noses. The face template can be generated from face images that have to be aligned in orientation and size in a preprocessing step. The designed standard face pattern template is then used to match with the located face components in facial images. This usually uses appropriate energy function. The best match of a template in the facial image will yield the minimum energy. This face template matching constitutes one of the techniques being used in detecting and tracking significant features given a series of image frames, thus it is commonly used for dynamic facial expression recognition. Although this method is relatively simple to implement and do not need extensive a priori information, it is prone to failure when large variations in pose or scale exist [YKA02]. However, the above problem can be tackled by deformable models, like Active shape model (ASM) and active appearance model (AAM) first described by Cootes and Taylo [CT99]. The ASM and AAM are two popular shape and appearance models for object localization, and are statistical model based approaches that have attracted interest for many years and numerous extensions and improvements have been proposed. The AAM enables

to fully automatically create a model of a face depicted in an image. The created models are realistic looking faces, closely resembling the original. The AAM extends the functionality of ASM capturing texturing information along with shape information. Methods similar to ASM employing a point distribution model to fit the shapes, expand into 3-dimensional problems [NC08]. Blanz and Vetter [BV03] proposed a 3D morphable face model generated from a vector space representation of 3D facial scans. The construction of the morphable model is made through a convex combination of shape and texture vectors of a set of realistic 3D human faces. This generic model is then fitted to 2D face images for 3D shape reconstruction, which includes a fitting algorithm for shape and texture parameter optimization. Thus, similarity measurement can be obtained for face and facial expression recognition.

2.2 2D Facial expression recognition

Due to the huge amount of 2D data and their availability to various kind of research and applications, and the fact that systems like HCI and robots are generally equipped with 2D cameras, the majority of facial expression recognizers have been developed in 2D environment resulting in a valuable enhancement. Indeed, most of the databases for facial expression analysis are made up of 2D images, like the cohn-kanade DB, the MMI DB, the Japanese Female Facial Expression (JAFFE) DB and the Belfast Naturalistic DB. Comprehensive surveys in this area include those by Fasel and Luettin [FL03], Pantic et al. [PR00] and and Zeng et al. [ZPRH09]. Table 2.1 shows representative sample of the existing 2D databases.

Databases	Formats	Number of Images/Sequences	Number of Expressions	Colour/Gray	Resolution	Number of subjects	Year
FERRET	2D Static	14,051	2	Gray	256*384	1199	1996
CMU-PIE	2D Static	41,368	4	Color	384*286	68	2000
Multi-PIE	2D Static	750,000	4	Color	3072*2048	337	2009
MMI	2D Static	200	6	Gray	720*576	52	2005
JAFFE	2-D Static	213	7	Gray	256*256	10	1998
Cohn-Kanade	2D Dynamic	486	6	Color	640*490	97	2000
MPI	2D Dynamic	60	4	Color	450*400	8	2003
DaFEx	2D Dynamic	1,008	6	Color	360*288	8	2005
FG-NET	2-D Dynamic	399	6	Color	320*240	18	2006

Table 2.1 – *Representative sample of the existing 2D Facial databases.*

2.2.1 Static 2D FER

The static facial expression recognition classifies a single static image or a frame in a video sequence to one of the facial expression categories based on the feature extraction resulting from that 2D data. Wang and Yin [WY07] proposed an automatic facial expression reading and understanding system for recognizing facial expression from a static image. They developed a topographic modeling approach based on facial expression descriptor called topographic context (TC). This approach applies topographic analysis that treats the image as a 3D terrain surface where each pixel is assigned one type of topographic label based on the terrain surface structure.

The work of Abboud et al.[ADD04] on facial expressions belongs to appearance based approaches, they introduced two methods for building the model of facial appearance, the first one is a standard AAM, the second is a modified AAM. The construction is based on three or only one PCA, and is performed through statistical analysis of the normalized shape and texture vectors from still images data of the CMU expressive face database. The facial expression recognition is performed in the space of the AAM parameters considering six basic emotional categories in addition to the neutral expression. The obtained recognition rate is about 84%.

Zhang et al. [ZLSA98] propose a method to combine the two types of approaches listed below in order to address FER. The first type is the calculation of the geometric positions of a set of fiducial points on a face. The second is the computation of a set of multi-scale and multi-orientation Gabor wavelet coefficients extracted from the face image at the fiducial points.The two types of features are fed in the input units of a two-layer perceptron. The recognition performance is experimented based on the two approaches independently and jointly and and the generalized recognition rates are 73.3% with geometric positions alone, 92.2% with Gabor wavelet coefficients alone, and 92.3% with combined information.

Several FER approaches have been inspired by the work of Turk and Pentland [TP91] for face recognition. They propose the decomposition of face images into a weighted sum of basis images, called eigenfaces, using

PCA. Using similar techniques, Padgett and Cottrell [PC96] compare the generalization performance of three distinct representation schemes for facial emotions using a single classification strategy based on Neural Network. All three representations were based on PCA applied on aligned faces and features. The first representation is made for the whole face (eigenfaces), the second one is more localised and made for the eyes and the mouth (eigeneyes and eigenmouths), and the third one is of the eyes and mouth that makes use of basis vectors obtained by principal components of random image blocks. Neural Network architecture is then applied to achieve 86% of average recognition rate. Cottrell and Metcalfe [CM90] use holistic representations based on principal components to extract whole-face features they call *holons*. These holons are given to a back propagation networks that are trained for emotion classification.

Lekshmi and Sasikumar [LSN08] address facial expression recognition using both Gabor wavelets transform and Discrete cosine transform (DCT). The generated coefficients are then given to a neural network to classify the expression among one of the four expressions chosen for the analysis which are happy, normal, surprise and anger. they report an average recognition rate of 89.11% on JAFFE database.

Shan et al. [SGM09] empirically evaluate facial representation based on Local Binary Patterns as statistical local features. They apply different machine learning methods on several databases. After extensive experiments they report that LBP features are effective and efficient for facial expression recognition.

Lanitis et al. [LTC97] propose a unified approach to face image coding and interpretation problems. They use a set of training images to build two parametrised deformable templates. The first one is a Point Distribution Model (PDM) that describes the shape of any instance of the face and which is analogous to the ASM. The PDM is generated by statistical analysis of the positions of a set of landmark points defined in the training images. The second one is a gray-level appearance model. The same training images are used to extract the gray-level profiles perpendicular to the boundary of the shape model points. Experimenting their approach to

derive 2D facial expression recognition they achieve 74% of correct classifications. Tian et al. [lTKC01] explored both 2D and 3D modalities from single images, and have investigated on automatic detection of 25 action units through 2D mapping of 3D facial scans, for feature extraction they employed one model-driven technique, Gabor wavelets, and two data-driven methods, Independent Component Analysis (ICA) and Non-Negative Matrix Factorization with Sparseness Constraints (NMFSC). To classify and detect AUs they have used AdaBoost classifier, Support Vector Machines (SVMs),and Naïve Bayes classifiers. In a combined implementation, AdaBoost is used as an effective feature selector. This is followed by a SVM or a Naïve Bayes classifier, which is trained with these AdaBoost-selected features. In Table 2.2 we summarize some of the large number of existing approaches that addressed static 2D facial expression recognition problem.

References	Holistic (H) *vs* Local (L) based approach	Expression types	Database	Classifier	Average recognition rate (%)
Wang and Yin [WY07]	L: Topographic context	6	MMI & Cohn–Kanade	LDA	82.61
Abboud et al. [ADD04]	H: AAM	7	CMU	LDA	84
Zhang et al. [ZLSA98]	L: Gemoetric position of featur points & Gabor wavelets	7	private	two-layer perceptron	92.3
Padgett and Cottrell [PC96]	L: eigen features using PCA	6	private	Neural Network	86
Cottrell and Metcalfe [CM90]	H: hollons using PCA	8	private	Neural Network	-
Lekshmi and Sasikumar [LSN08]	H: Gabor wavelet filter	4	JAFFE	SVM	89.11
Shan et al. [SGM09]	L: Local Binary Patterns (LBP)	6	Cohn-Kanade	SVM	88.4
Lanitis et al. [LTC97]	H: PDM & gray level profile	7	private	Mahalanobis Distance	74
Hu et al. [HZY*08]	L: Point displacement	6	private	kernel associative memory model	86
Zhi et al. [ZFRK09]	H: Graph-preserving sparse non-negative matrix confusion	6	Cohn-Kanade	Nearest Neighbor	93.5
He et al. [HWYW09]	H: Gabor wavelet	7	JAFFE	HMM	96.16
Li et al. [LIK09]	L: SIFT,PHOG, Hist of Edges	6	Cohn-Kanade	Nearest Neighbor	96.3
Tian et al. [lTKC01]	L: Gabor wavelet & ICA & NMFSC	25 AUs	Bosphorus & Cohn-Kanade	AdaBoost & SVM & Naïve Bayes	92.4

Table 2.2 – *Sample of static 2D Facial expression recognition approaches.*

2.2.2 Dynamic 2D FER

In the case of 2D videos, there has been tremendous interest in tracking and recognizing facial expressions over time, it is suggested that the dynamics of facial expressions provides important cues about the underlying emotions that are not available in static images. This is because that the position and shape of these components and/or landmarks are often detected in the first frame and then tracked throughout the sequence.

The majority of work on facial expression recognition has focused on facial motion analysis through optic flow estimation. The first step towards the automatic recognition of facial expressions from a sequence of images was taken in 1978 by Suwa et al. [MSF78]. In their work, a preliminary investigation was made by tracking the motion of twenty identified spots on a frame sequence. Although this system was proposed in 1978, researchers did not pursue this line of study till the early 1990s. Since then considerable progress has been made in building computer systems that attempt to automatically analyze and recognize facial motions from videos. Still in an early exploration of facial expression recognition, Mase [K.M91] was the first to use optic flow to estimate the skin movement in a subset of the facial muscles defined in muscle windows, and each of which defines one primary direction. He showed an accuracy rate of nearly 80% for recognizing four types of expressions: happiness, anger, disgust, and surprise. Essa and Pentland [EP97] extended this approach, using optic flow coupled with a dynamic model of motion to estimate activity in a detailed anatomical and physical model of the face. Motion estimates from optic flow were refined by the physical model in a recursive estimation and control framework and the estimated forces were used to classify the facial expressions. Donato et al. [DBH*99] explore and compare techniques for automatically recognising facial actions in sequences of images. In their approach they take profit from the FACS developed by Ekman that provides objective means for measuring the facial muscle contractions involved in a facial expression. They focus on facial motion analysis through multiple methods; optic flow estimation employing a correlation-based technique, holistic spatial analysis, such as principal component analysis,

independent component analysis, local feature analysis (LFA), and linear discriminant analysis and methods based on the outputs of local filters, such as Gabor wavelet representations and local principal components. They applied all these methods for face image representation in order to classify facial actions and thus facial expression. An evaluation of all these techniques is presented as well as their performance are compared using a single dataset. Yacoob and Davis [Yac94] propose an approach for analyzing and representing the dynamics of facial expression from image sequences. Their system consists of locating prominent facial features based on the description of the epic of facial expressions from static face images, and tracking of the facial patterns associated with the edges of the mouth, eyes and eyebrows. For the tracking phase, two sets of parameters are calculated around the rectangle enclosing the extracted features, the centroid of points having high gradient value within each rectangle and the window that determines the scaling of the rectangle. The optical flow analysis is employed to enhance the tracking. As a result they arrive at elaborating a dictionary that is a motion-based feature description of facial actions due to facial expressions. The overall modeling of the approach characterizes the expressions using a beginning-epic-ending trilogy. Facial expression recognition rate is presented for each expression of a set of seven expressions namely happiness (86%), surprise (94%), anger (92%), disgust (92%), fear (86%), sad (80%) and blink (65%). Rosenblum et al. [RYD96] expanded this system to model the full temporal profile of facial expressions with radial basis functions, from initiation, to apex, and relaxation. Huang and Huang [Hua97] proposed a system that applies both point distribution model (PDM) and gray level model to describe facial features. Then the position variations of certain designated points on the facial feature are described by 10 action parameters (APs). Rosenblum et al. [RYD96] extended the work of [Yac94] by using a connectionist architecture. Individual emotion networks were trained by viewing a set of sequences of one emotion for many objects. The trained neural network was then tested for emotion recognition. However, their works are limited by the motion in six predefined and hand initialized rectangular regions

on a face that is not fully automatic. Black and Yacoob [BY97] explored a middle ground between the template-based approaches and the optical flow-based approaches. They proposed to use local parameterized models of image motion for recovering the nonrigid motions of the human face which provides a concise description of facial motions in terms of a small number of parameters. Motions of the face and some regions such as mouth, eyebrows, and eyes are modeled using image flow models with only a handful of parameters, and are estimated over an image sequence using a robust regression scheme. They reported that the recovered motion parameters are stable under adverse conditions such as motion blur, saturation, loss of focus, etc. They carried out a large set of experiments to verify and evaluate the performance of the recognition procedure. The overall success rate for the system was 93% while the accuracy was 89%. However, they did not address the problem of locating the various facial features. The effectiveness of the extracted features will affect the accuracy of the parameters estimation.

Kumano et al. [KOY*09] proposed a variable-intensity template composed of three components: a rigid shape model a set of interest points, and an intensity distribution model. A set of interest points, precisely paired points, are sparsely defined around the edges of facial parts (i.e. eyebrows, eyes, noise and mouth). The intensity of each interest point is observed along the video sequence, and its variation due to small position shifts is calculated using error in the shape model, fitted to each frame, as well as the errors in the intensity adjustment. Such variation is represented as a normal distribution that describes how interest point intensity varies for different facial expressions. Using this template in a framework of particle filter they achieved a recognition rate of 90% for a certain range of the head pose variations. Akakın and Sankur [cAS10] proposed two feature extraction schemes for 2D dynamic facial expression recognition; the first scheme consists in detecting 17 landmark points that are tracked over successive video frames to determine their trajectories, features are then extracted with applying ICA on the estimated trajectories. In the second scheme the features extracted from spatiotemporal face cubes using

global 3D DCT transform. Whatever the extracted features they used a modified Nearest Neighbor (MNN) classifier to classify the six basic emotions. Their approach was evaluated on CohnKanade database and using a fusion scheme of several classifier parameter settings they reported an average classification rate of 95.34%. In Table 2.3 we summarize some of the large number of existing approaches that addressed dynamic 2D facial expression recognition problem.

References	Holistic (H) vs Local (L) -based approach	Expression types	Database	Classifier	Average recognition rate (%)
Vretos et al. [VNP09]	L: Candide landmarks	7	Cohn-Kanade	SVMs (RBF, polynomial)	90.22
Tong et al. [TCJ10]	L: Gabor wavelet	14 AUs	Cohn-Kanade	Baysian Network	85.8
Uddin et al [ULK09]	H: Texture	7	Cohn-Kanade	HMMs	93.23
Mase [K.M91]	L: Tracking active points	4	private	SVM	80
Essa and Pentland [EP97]	H: Topologically invariant physics-based model	6	private	2D motion energy models	98
Donato et al [DBH*99]	T: Gabor wavelet	12 AUs	private	Nearest Neighbor	95.5
Yacoob and Davis [Yac94]	L: Optical flow of high gradient points	7	private	Rule based system	85
Huang and Huang [HH97]	H: PDM & gray level model	6	private	Distance-based	84.41
Black and Yacoob [BY97]	H: parameterized model	6	private	Parametric flow models	89
Kumano et al [KOY*09]	H: variable-intensity template	6	private & Cohn-Kanade	Likelihood functions	90
Akakın and Sankur [cAS10]	L: Landmark trajectories & spatiotemporal face cubes	6	Cohn-Kanade	Modified Nearest Neighbor	95.34

Table 2.3 – *Sample of dynamic 2D Facial expression recognition approaches.*

2.3 3D Facial expression recognition

In 3D domain the amount of 3D facial expression databases is not as huge as the one in 2D domain. There has been a lack of publicly available 3D facial expression databases, resulting in making the approaches proposed for 3D FER are much fewer than those proposed for 2D FER. In Table 2.4 we give a list of the existing 3D databases. In this section we will present almost all approaches that addressed 3D FER problem.

Databases	Formats	Number of Images/Sequences	Number of Expressions	Color/Gray	Resolution	Number of subjects	Year
BU-3DFE	3-D Static	2,500	7	Color	1040*1329	100	2006
Bosphorus	3-D Static	4,666	6	Color	1600*1200	105	2008
ZJU-3DFE	3-D Static	360	4	Color	-	40	2006
ADSIP	3-D Dynamic	210	7	Color	601*549	10	2009
BU-4DFE	3-D Dynamic	606	6	Color	1024*681	101	2008
Hi4D-ADSIP	3-D Dynamic	3,360	14	Color	1024*1728	80	2011

Table 2.4 – *Representative sample of the existing 3D Facial databases.*

2.3.1 Static 3D FER

Different approaches were proposed to address the problem of static 3D facial expression recognition, we will try to describe briefly the techniques developed within these approaches in what follows:

Curvature

The first work related to this issue is presented by Wang et al. [WYWS06]. They investigate the importance and usefulness of 3D facial geometric shapes to represent and recognize facial expressions using 3D facial expression range data. They propose a novel approach to extract primitive 3D facial expression features, these features are geometric based features in the 3-dimensional Euclidean space, and are estimated using the principle curvature information calculated on the 3D triangulated mesh model. A surface labeling scheme is applied to classify the 3D primitive surface features into twelve basic categories (such as ridge, ravine, peak, pit, saddle, concave hill, convex hill, etc.). In order to classify the specific expression, they partition the face surface into a number of expressive regions, and conduct the statistics of the surface primitive label distribution on each region separately. The statistic histograms of the surface labels of all these regions is combined to construct the specific facial expression feature. They report that the highest average recognition rate obtained is equal to 83%. They evaluate their approach on both frontal and non frontal-view facial images generated from the BU-3DFE DB, and tested its robustness to the latter case.

Euclidean distances between landmarks

Euclidean distances are used as features to describe expressions, these distances are measured between fiducial points localized on the face area. These measures were conducted on both neutral and expressive scans of the same subject, and the differences in term of Euclidean distances, considering the same set of points, from a neutral state to an expressive state are evaluated and stored. A normalisation step is compulsory for alleviating the scaling problem, and this is undertaken through the exploration

of facial action parameters (FAPs) defined in the MPEG-4 specifications.
Soyel and Demirel [SD07] propose to make use of the symmetry of the face, and select a small set of 3D key points (11 key points), from which six characteristic Euclidean distances were computed. These are: eye opening distance, eyebrow height, mouth opening, mouth height, lip stretching and normalization (distance between outermost points on left/right face contour). These distances serve as input for a multilayer-perceptron-based neural network classifier. They derive experiments on 60 subjects of the BU-3DFE DB, showing seven expressions (basic expressions + neutral). For evaluation, this dataset is arbitrarily divided into training set (54 subjects) and test set (6 subjects) ten times and for each fold, classification is done. Results of this system show an average recognition accuracy of 91,3%. The same authors continue work on this approach, changing some parameters of the system. Hence, 23 facial landmark 3D points are used to compute six distance features as above, with ''height of mouth'' replacing ''width of mouth'' and ''openness of jaw'' replacing ''normalization''. Again, the BU-3DFE is used for experiments, using the same setup as above. This time, a probabilistic neural network is used for classification, which shows an average recognition rate of 87,8%. We can notice that with fewer landmark points, only 11 landlmarks for the first trial, the proposed method achieved better results than the second trial where they use more landmaks (23 landmarks). This can be explianed by the fact that in their first work the authors justified their choice based on the symmetric propoerty of the front view face. And when it comes to the classification process the fewer the number of samples to classify (six Euclidean distances) is, the better performance the classifier shows. In contrast, for the second work they used a larger set of features making much more difficult for the classification process to select the discriminative features.
In a similar approach Tang and Huang [TH08] propose an automatic feature selection computed from the normalized Euclidean distances between two picked landmarks from 83 possible ones. Using regularized multi-class AdaBoost classification algorithm, they report an average recognition rate of 95.1%, and they mention that the surprise expression is recognized

with an accuracy of 99.2%. Alghough these approaches have attained high recognition rates, their main drawbacks remains the dependency on a set of accurately located landmarks and the non-use of the shape (range) information that is provided with 3D data.

Automatic keypoints

Berretti et al. [BDP*10], proposed full automatic solution that first detects a set of facial keypoints, and then exploits the local characteristics of the face around a set of sample points automatically derived from the facial keypoints. SIFT descriptors are computed around the sample points of the face and used as a feature vector to represent the face. Feature selection permits to identify the best subset of features to feed a set of classifiers based on *Support Vector Machines* (SVM). State of the art results are obtained on the BU-3DFE. Although this automatic approach solve partially the problem of keypoints detection, the detected keypoints are not always as well defined as the manually annotated points.

Action Units (AUs)

Other approaches propose to recognize AUs in order to analyse facial expressions. Savran et al. [SSB12] proposed to recognize a wide variety of expressions through the detection of 25 AUs that are found singly or in combination, in their approach they derived the detection of these AUs observed in a 3D facial surface that is then mapped into 2D single image to derive detection using luminance information. A set of 3D surface features (i.e., Shape index, Gaussian curvature, Mean curvature, depth, etc.) and Gabor magnitude features were computed and used singly or in combination as entry parameters to statistical learning techniques (i.e., AdaBoost, LinearSVM, RBFSVM and Naïve Bayes) and a recognition rate of 96.9% is reached on the Bosphorus database. In [SSB11] the same authors propose a novel method for AU intensity estimation applied to 2D luminance and/or 3D surface geometry images to analyse facial expressions. For feature extraction stage, Gabor wavelets are extracted separately from both 2D luminance and 3D surface geometry images. For the geo-

metric information various types of local 3D shape indicators have been considered, such as mean curvature, Gaussian curvature, shape index and curvedness, as well as their fusion. Then for the classification stage, they investigate the potential of fusing different types of features as they may contain complementary information. So they implement the feature fusion by AdaBoost feature selection. Then SVM is applied to recognize a set of 25 AUs.
Knowing that the problem of AUs recognition is a challenging task for both computer vision and pattern recognition research communities, the proposed approaches still remain limited in term of the number of AUs that can be detected.

Texture and shape informations

Zhao et al. [ZHDC10] proposed to combine Bayesian Belief Net (BBN) and Statistical facial feature models (SFAM) to develop an automatic FER approach. They were able to learn both global and local variations by studying respectively face landmark configuration (morphology) and texture and shape information around automatically detected fiducial points. The report a recognition classification rate of 82% on the BU-3DFE database.

Template

Templates are used extensively especially in computer graphics. These techniques are also explored in FER problems. They are used to simulate physical process of generating expressions through adjustment of some controlled parameters. Different terms are proposed to designate templates; active shape model (AAM), active appearance model (AAM), deformable model, elastically deformable model, morphable model, etc. In [GWLT09a], the shape of an expressional 3D face is approximated as the sum of a basic facial shape component (BFSC), representing the basic face structure and neutral-style shape, and an expressional shape component (ESC) that contains shape changes caused by facial expressions. The two components are separated by first learning a reference face for each input non-neutral 3D face then, based on the reference face and the orig-

inal expressional face, a facial expression descriptor is constructed which accounts for the depth changes of rectangular regions around eyes and mouth.

Mpiperis et al. [MMS08] propose an approach for a joint 3D face and facial expression recognition using bilinear models. They begin with constructing a prototypic 3D facial surface using a low-dimensional face eigen space. This generic 3D model, termed subdivision surface, is used to represent a neutral expression and an average identity. Then an energy minimization technique is used to derive automatic point to point correspondence of the subdivision surface to the realistic 3D face model. Once point to point correspondence is established they apply bilinear models to encode identity and expression in independent control parameters in order to be able to perform joint expression-invariant facial identity recognition and identity-invariant facial expression recognition. Finally they build a facial expression classifier on the basis of an asymmetric bilinear model fitted to a training set of faces of the BU-3DFE DB. The subjects are divided randomly in two sets: 1) a training set consisting of 90 subjects and 2) a test set consisting of the remaining ten subjects, thus ensuring the independence on a subject identity. They reported an average recognition rate of 90.5%. They also reported that the facial expressions of disgust and surprise are well identified with an accuracy of 100%.

The downside of these approaches is first their computational cost which is heavy and expensive. The second is the fitting process that requires existence of reliable landmarks for initialization, and which hardly converge in case of opening of the mouth in facial expressions. Still with the fitting, despite the fact that this process can handle non-rigid registraion between the template and the real facial 3D model, there exists no results to quantify the plausibility of such generic model of being a real face.

All the presented approaches that tend to recognize facial expression do not explore the shape information it self, and in some cases where the shape information is used, this information is extracted based on calculation of geometrical feature such as curvature and gradient information. Although such feature definitions are intuitively meaningful, the com-

putation of such features involves numerical approximation of second derivatives and is very susceptible to noisy observations and are numerically unstable. In our case we will present in chapter 2 an novel approach that is based on the analysis of shapes through comparison of shapes of 3D surfaces themselves.

2.3.2 Dynamic 3D FER

Although several of these outlined methods exploited the motion between frames, and even some temporal information, none of them aimed to model the temporal dynamics of the expression for recognition purposes. To the best of our knowledge there exits only two approaches, so far, that exploit 3D facial expression dynamics. The first approach was proposed by Sun and Yin [SY08] proposed a spatio-temporal descriptor for facial expression classification. They combined a spatial HMM and a temporal HMM to investigate the 3D spatio-temporal facial behavior in the 4D domain. They applied a generic model to track range models frame by frame. The developed spatio-temporal descriptors was built based on two stages:

Temporal and spatial feature extraction

Tracking Vertex flow In [SY08] the authors proposed to adapt and track a generic model to each frame of the 3D model sequence and establish the vertex flow estimation as follows: First, they establish correspondences between 3D meshes using a set of 83 pre-defined key points. This adaptation process is performed to establish a matching between the generic model (or the source model) and the real face scan (or the target model). Second, once the generic model is adapted to the real face model, it will be considered as intermediate tracking model for finding vertex correspondences across 3D dynamic model sequences. The vertex correspondence across 3D model sequences provides a set of motion trajectories (vertex flow) of 3D face scans. The vertex flow can be depicted on the adapted generic model (tracking model) through estimation of displacement vector from the tracked points of the current frame to the corresponding points of the

first frame with a neutral expression. The vertex flow is described by the facial motion vector $U = [u_1, u_2, \ldots, u_n]$, where n is the number of vertices of the adapted generic model.

Surface labelling An automatic labelling approach was used to label the surface geometric features as one of the eight label types on the tracked locations of the range models. As a result each range model in the sequence can be represented by a vector $G = [g_1, g_2, \ldots, g_n]$, where g_i represents one of the primitive shape labels and $n$equals the number of vertices of the facial region on the adapted model. Such a shape descriptor provides a robust facial surface representation.

HMM models for classification

Based on the fact that individual facial characteristics are represented by not only the temporal change (inter-frame) but also the spatial change (intra-frame). After extracting both the motion features and the primitive surface features, the second stage consists of training and testing different HMMs in both spatial (S-HMM) and temporal (THMM) domains and their combination for each subject.

1. T-HMM: applied to explore the temporal dynamics of 3D facial surface along a time sequence. The T-Hmm takes each frame of a face sequence as one observation, where each frame is the surface labeled 3D face model transformed into an optimized feature space using LDA.

2. S-HMM: applied also to each sequence frame, where the 3D face model is subdivided to 6 regions from top to bottom. The subdivision was based on the feature points that have been tracked. A one-dimensional HMM models were constructed and 6 states were used to build both T-HMM and S-HMM models.

3. Combined Spatio-Temporal HMM (C-HMM): to model both spatial and temporal information of 3D face sequences, they combined the S-HMM and the T-HMM to construct a pseudo 2D HMM(P2D-

HMM). The final decision is based on both spatial decision and temporal decision.

The model adaptation and the vertex correspondence between the generic model and the range scans permitted to describe the facial surface features and their motion trajectories on the adapted generic model. And they applied this tracked information to the task of facial expression recognition. To do so they constructed a spatio-temporal face model descriptor, which consists of two parts: a facial expression label map and a set of vertex flow. The geometric surface feature labelling resulting in a 3D expression model's label distribution was formulated by the facial expression label map (FELM) vector: $e = \left[\frac{n_1}{n}, \frac{n_1}{n}, \ldots, \frac{n_c}{n}\right]$, where c denotes the number of label types and n the total number of vertices. In addition the vertex flow represents the temporal expressive information. Concatenating both e and u a unique spatial-temporal facial expression descriptor F was constructed for each individual frame where $F = [e, u]$. Thus a rich spatio-temporal descriptor is build through combination of template (generic 3D model) and geometrical feature (curvature). The average classification rate achieved by this approach attained 90.44%.
Alghough the authors presented a comparison of their 3D model-based HMM approach against static 2D/3D-based approaches and outlined that their solution has shown better perfermance than other techniques. Unfortunately, there were no information about the computational cost of the proposed method, especially that they adapted a generic 3D model to each frame in the sequences. Such technique needs for a reliable fitting process that can hardly converge in certain cases.

Sandbach et al. [SZPR12] proposed an alternative method for dynamic facial expression analysis. They used Free-Form Deformations (FFD) to captured motion between consecutive frames in a sequence, rather than fitting a deformable model to each mesh. Second, they used quad-tree decomposition of several motion fields to extract motion features especially features reflecting dense percentage of motion present in each part of the image. And, finally, they used HMM models for temporal modeling of

the full expression sequence to be represented by 4 segments which are neutral-onset-apex-offset expression segments. Features selection is then derived using GentleBoost technique applied on the onset and offset temporal segments of a given expression, the average correct classification results for three basic expressions(i.e., happy, angry and surprise) achieved 81.93%. Only tree expressions out of six were considered for the classification process, namely happiness, anger and surprise.

The third work on 4D facial expression recognition is proposed by Le et al. [LTH11]. They proposed a level curve-based approach to capture the shape of 3D facial models. The level curves are parameterized using the arclength function. The Chamfer distances is applied to measure the distances between the corresponding normalized segments, partitioned from these level curves of two 3D facial shapes. These measures are then used as spatio-temporal features to train Hidden Markov Model (HMM), and since the training data were not sufficient for learning HMM, the authors propose to apply the universal background modeling (UBM) to overcome the over-fitting problem. Using the BU-4DFE database to evaluate their approach, they reached an overall recognition accuracy of 92.22% for only three prototypic expressions (i.e., happy, sad and surprise). We notice that the curve-based representation was generated from 2D range images, hence the extracted curves are planar curves, and in case of pose variations the feature extraction can be affected.

Recently Fang et al. [FZSK11] proposed a fully automatic pipeline to classify expression from 4D data. Their pipline is set to start with a robust registration of each pair of consecutive frames of a given sequence. They developped a two-step technique to derive a mesh matching process; the first step consists in estableching vertex correspondence between two meshes with providing two alternative methods, one is based on spin images similarities and the other based on the Euclidean distance between MeshHOG descriptors [ZBVH09]. Then RANdom SAmple Consensus (RANSAC) is applied to alleviate the problem of nosy point correspondence due to outliers that might be generated the vertex correspondence step. A deformable face model (AFM) is then applied to generate a fitted sequence

from a 4D dataset. After that Local Binary Patterns (LBP) descriptors to compute what they termed flow image to represent the deformation vectors in a subsequent frame.In the final stage of their proposed pipeline for 4D FER, they applied support vector machines with a Radial Basis Function kernel for classification and outperformed previous work by achieving 95.75 % average classification rate on the BU-4DFE database. Tables 2.5and 2.6 summarise the existing approaches in static 3D FER and dynamic 3D FER respectively.

We note that for most the presented approaches there were no information provided about the computational time needed to run these methods. There exist different stages that are time consuming, such as the preprocessing stage, the feature extraction stage, in case of local approaches, the fitting process, in case of global approaches, and the classification stage. And the reported methods that are described as automatic approaches do not present a quantification of the time needed to run their techniques and obtain results. Besides, these approaches there were no experiments conducted to reveal to what extent they are robust to pose variations constraints, and where we are in presence of missing data problem. These limitations are interesting to mention and need for further studies to make these techniques as suitable as possible for real-world applications.

References	Holistic (H) vs Local (L) -based approach	Landmarks	Expression types	Database	Classifier	Average recognition rate (%)
Wang et al. [WYWS06]	L: Primitive Surface Feature	64 manual	6	BU-3DFE	Linear Discriminant Analysis	83.6
Soyel and Demirel [SD07]	L: Six Eucleadian Distances	11 manual	7	BU-3DFE	Neural Network	91.3
Tang and Huang [TH08]	L: Normalized Euclidean Distances	83 manual	6	BU-3DFE	multi-class AdaBoost	95.1
Mpiperis et al. [MMS08]	H: Bilinear Models	auto	6	BU-3DFE	Maximum Likelihood	98.1
Gong et al. [GWLT09a]	H: Facial and Expressional shape models	auto	6	BU-3DFE	SVM	76.22
Berretti et al. [**?**]	L: SIFT	27 manual	6	BU-3DFE		77.5
Zhao et al. [ZHDC10]	L& H: BBN & SFAM	auto	6	BU-3DFE	Neural Network	96.9
Savran et al. [SSB12]	L: 3D surface curvatures & Gabor magnitude	auto	6	Bosphorus	SVM-RBF	96.9

Table 2.5 – *Existing approaches in static 3D Facial expression recognition.*

References	Holistic (H) vs Local (L) -based approach	Landmarks	Expression types	Database	Classifier	Average recognition rate (%)
Tsalakanidou and Malassiotis [TM10]	L: Geometric,Appearance and Surface Curvature measurements	81 auto	5/11 AUs (subject dependent)	private	rule-based	85.0/83.6
Sun et al. [SRY08]	H: AAM	83 auto	6/8 AUs	BU-4DFE	HMM	80.9/87.1
Sun et al. [SY08]	H: Generic Shape Model	83 auto	6	BU-4DFE	HMM	90.44
Sandbach et al. [SZPR12]	H: Free-Form Deformations (FFDs)	auto	3	BU-4DFE	HMM	
Le et al. [LTH11]	H: Facial Level Curves	auto	3	BU-4DFE	HMM	92.22
Fang et al. [FZSK11]	H: Local Binary Patterns	auto	6	BU-4DFE	SVM	74.63

Table 2.6 – *Existing approaches in dynamic 3D Facial expression recognition.*

2.4 Conclusions

In this chapter, approaches related to facial expression recognition problem were outlined. The majority of existing state-of-the-art approaches are directed toward 2D data domain, whether it is static 2D images (i.e., single image) or dynamic 2D images (i.e., image sequences). Although the fact that 2D approaches has demonstrated their effectiveness, they still inherent challenges of pose, illumination, expression, and scale. Three-dimensional facial data comes to show the potential to alleviate the problems encountered by 2D-based approaches. Thus we presented a survey of approaches developed for 3D facial expression recognition and that are closely related to our research interest. In this survey we made the choice to categorize existing approaches into two categories: local *vs* Holistic and static *vs* dynamic, and key properties of some 2D-based approaches and practically all 3D-based approaches were summarized in Tables 2.2, 2.3, 2.5 and 2.6. In the next chapters we will present our own approach to tackle this task. Results through evaluation of our method will be presented and comparison with related works will be reported.

3 Static 3D Facial Expression Recognition

3.1 Introduction

Unlike the identity recognition task that has been the subject of many papers, only few works have addressed 3D facial expression recognition. This could be explained through the challenge imposed by the demanding security and surveillance requirements. Besides, there has long been a shortage of publicly available 3D facial expression databases that serve the researchers exploring 3D information to understand human behaviors and analyze emotions. The main task is to classify the facial expression of a given 3D model, into one of the six prototypical expressions, namely *Happiness*, *Anger*, *Fear*, *Disgust*, *Sadness* and *Surprise*. It is stated that these expressions are universal among human ethnicity as described in [EHSH92] and [EF71]. In this chapter we present a novel approach for 3D identity-independent facial expression recognition based on a local shape analysis. The main contributions of our approach are as follows:

1. We propose a new process for representing and extracting patches on the facial surface scan that cover multiple regions of the face,

2. We apply a framework to derive 3D shape analysis in order to quantify similarity measures between corresponding patches on different 3D facial scans. Thus, we combine a local geometric-based shape analysis approach of 3D faces and several machine learning techniques to perform such classification.

The remainder of this chapter is organized as follows. First, we describe the BU-3DFE database designed to explore 3D information and improve facial expression recognition in Section 3.2. In section 3.3, we present the adopted 3D facial surface representation for our study. In Section 3.4, we summarize the shape analysis framework applied earlier for 3D curves matching by Joshi et al. [JKSJ07], and discuss its use to perform 3D patches analysis. This framework is further expounded in section 3.5, so as to define methods for shapes analysis and matching. In section 3.6 a description of the feature vector and used classifiers is given. In section 3.7, experiments and results of our approach are reported, and the average recognition rate over 97% is achieved using machine-learning algorithms

for the recognition of facial expressions such as Multi-boosting and SVM. Finally, discussion and conclusion are given in sections **??** and 3.8.

3.2 Database Description

BU-3DFE is one of the very few publicly available databases of annotated 3D facial expressions, collected by Yin et al. [YWS*06] at Binghamton University. It was designed for research on 3D human face and facial expression and to develop a general understanding of the human behavior. Thus the BU-3DFE database is beneficial for several fields and applications dealing with human computer interaction, security, communication, psychology, etc. There are a total of 100 subjects in the database, 56 females and 44 males. A neutral scan was captured for each subject, then they were asked to perform six expressions namely: Happiness (HA), Anger (AN), Fear (FE), Disgust (DI), Sad (SA) and Surprise (SU). The expressions vary according to four levels of intensity (low, middle, high and highest or 01-04). Thus, there are 25 3D facial expression models per subject in the database. A set of 83 manually annotated facial landmarks is associated to each model. These landmarks are used to define the regions of the face that undergo to specific deformations due to single muscles movements when conveying facial expression [EF78a]. In Fig. 3.1, we illustrate examples of the six universal facial expressions 3D models including the highest intensity level.

3.3 Facial Surface Representation

In recent years, due to the advances in 3D sensing technology, 3D surface data have become widely available and characterized by increasingly fine details. And throughout this development, a lot of research focused on finding an appropriate digital representation for three dimensional real-world objects, mostly for use in computer graphics. The primary concerns for a surface representation in computer graphics focus on enabling the user to observe the information, and reduce the time taken to understand the underlying data, find relationships, and get the information

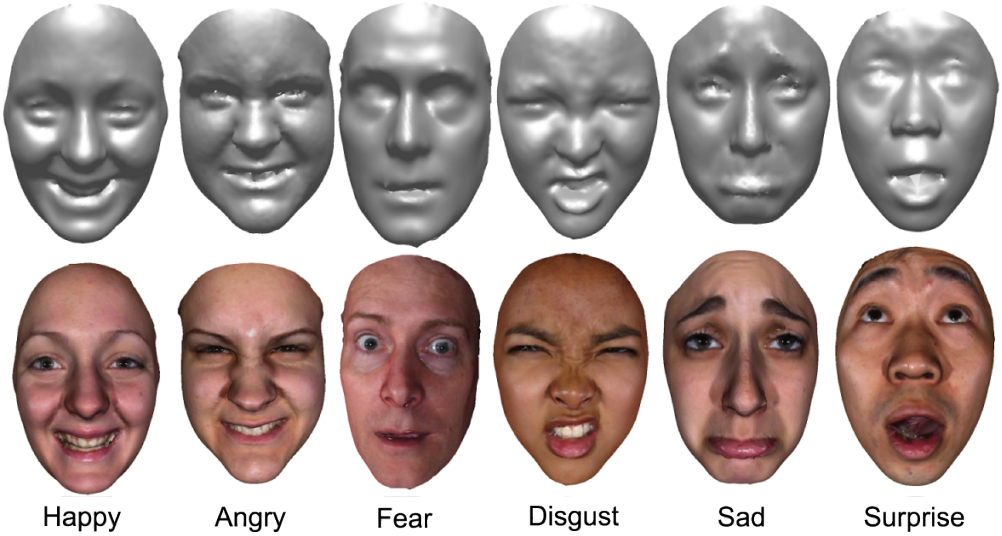

Figure 3.1 – *Examples of 3D facial expression models (first row 3D shape models, second row 3D textured models) of the BU-3DFE database.*

searched and also making the process on dedicated computer graphics hardware(GPUs) fast. In computer vision, in contrast, the concerns for surface representation are quite different. There is a real need to analyze detailed information which is designed for use in the specific vision applications such face and facial expression recognition. This information need to be accurate, concise and useful for statistical modelling.

3D surface representation can be divided into two categories: Explicit representation and implicit representation. In an explicit representation, one explicitly writes down the points that belong to the surface. As examples of explicit representation we can distinguish: Point clouds, Contour and profile curves, Polygon meshes, depth maps, spherical harmonics, etc. In an implicit surface representation, implicit functions are defined as the iso-level of a certain scalar function that defines 3D objects implicitly as a particular iso-surface. Implicit surfaces have two important classes depending on the way the implicit function: Radial Basis Function (RBF) and algebraic surfaces.

For our study, we choose to represent facial surface by a set of profile curves called also level curves. This representation is sparse and permits to approximate the surface as good as wanted. The purpose is to capture the shape of a 3D surface using a set of these curves. Depending on the extraction criterion, different types of contour curves can be defined. We

were interested in extracting iso-radius curves, these curves are contour curves obtained as an intersection of the 3D facial surface with a sphere with radius $r = \sqrt{x^2 + y^2 + z^2}$, with the z-axis parallel to the gaze direction and the y-axis parallel to the longitudinal axis of the face. In the following section we explain how level curves are extracted to represent the local facial parts of a 3D face model.

3.4 3D Facial Patches-based Representation

Most of the earlier work in 3D shape analysis use shape descriptors such as curvature, crest lines, shape index (e.g., ridge, saddle, rut, dome, etc.). These descriptors are defined based on the geometric and topological properties of the 3D object, and are used as features to simplify the representation and thus the comparison for 3D shape matching and recognition tasks. Despite their rigorous definition, such features are computed based on numerical approximation that involves second derivatives and can be sensitive to noisy data. In case of 3D facial range models, the facial surface labeling is a critical step to describe the facial behavior or expression, and a robust facial surface representation is needed. In Samir et al. [SSDK09] the authors proposed to represent facial surfaces by an indexed collections of 3D closed curves on faces. These curves are level curves of a surface distance function (i.e., geodesic distance) defined to be the length of the shortest path between a fixed reference point (taken to be the nose tip) and a point of the extracted curve along the facial surface. This being motivated by the robustness of the geodesic distance to facial expressions and rigid motions. Using this approach they were able to compare 3D shapes by comparing facial curves rather than comparing corresponding shape descriptors.

In our work we intend to further investigate on local shapes of the facial surface. We are especially interested in capturing deformations of local facial regions caused by facial expressions. Using a different solution, we compute curves using the Euclidean distance which is sensitive to deformations due to expressions. To this end, we choose to consider N reference points (landmarks) $\{r_l\}_{1 \leq l \leq N}$ (Fig.3.4 (a)) and associated sets

of level curves $\{c_\lambda^l\}_{1\leq\lambda\leq\lambda_0}$ (Fig.3.4 (b)). These curves are extracted over the patches centered at these points. Here λ stands for the value of the distance function between the reference point r_l and the point belonging to the curve c_λ^l, and λ_0 stands for the maximum value taken by λ.

Accompanying each facial model there are 83 manually picked landmarks, these landmarks are practically similar to the MPEG-4 feature points and are selected based on the facial anatomy structure. Given these points the feature region on the face can be easily determined and extracted. We were interested in a subset of 68 landmarks laying within the face area, discarding those marked on the face border. Contrary to the MPEG-4 feature points specification that annotates the cheeks center and bone, in BU-3DFE there were no landmarks associated with the cheek regions. Thus, we add two extra landmarks at both cheeks, obtained by extracting the middle point along the geodesic path between the mouth corner and the outside eye corner.

We propose to represent each facial scan by a number of patches centered on the considered points. Let r_l be the reference point and P_l a given patch centered on this point and localized on the facial surface denoted by S. Each patch will be represented by an indexed collection of level curves. To extract these curves, we use the Euclidean distance function $\|r_l - p\|$ to characterize the length between r_l and any point p on S. Indeed, unlike the geodesic distance, the Euclidean distance is sensitive to deformations, and besides, it permits to derive curve extraction in a fast and simple way. Using this function we defined the curves as level sets of:

$$\|r_l - .\| : c_\lambda^l = \{p \in S \mid \|r_l - p\| = \lambda\} \subset S, \quad \lambda \in [0, \lambda_0]. \tag{3.1}$$

Each c_λ^l is a closed curve, consisting of a collection of points situated at an equal distance λ from r_l.

To give more information about how we derive curve extraction and some related properties, we can describe a practical example of the extraction of the level curve associated with $\lambda = 1mm$ around a given landmark r. First we define a sphere whose center is set to be the point $r = (x_r, y_r, z_r)$ and radius $rd = 1mm$. This sphere will be used as a function to slice-through the 3D facial surface. The result of this cutting is a set a points lying on

the facial surface and situated at a distance $\lambda = 1mm$ from r. We obtain as many points as the number of edges connecting the landmark to his neighbours, illustration of this level curve is given in Fig. 3.2. Second, a Kochanek interpolating spline function [KB84] is applied to go through a smoothing process of the level curve. This technique allows to add a large number of points and do up-sampling controlled by a certain subdivision parameter. Last a uniform down-sampling process is applied to reduce the sampling rate of the smoothed curve and hence obtain the level curve needed for our study that will be denoted by β and characterized by a fixed number of points Fig. 3.3.

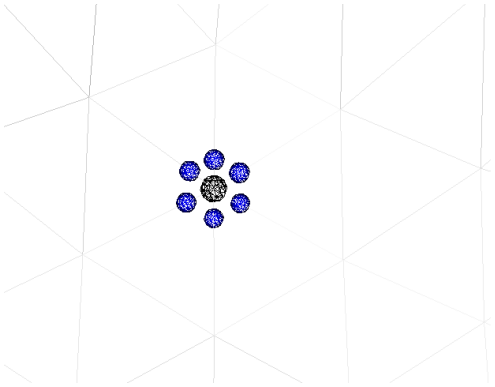

Figure 3.2 – *Illustration of the level curve points (blue points) around a given landmark (black point) resulting from a cutting of a sphere function centered on this point and of radius rd* $= 1mm$ *with the facial surface (wire frame representation).*

The curve extraction process is extended to produce the patches extraction process around the chosen landmarks. Fig. 3.4 shows the scheme of all patches extracted on a 3D facial surface.

3.5 Framework for 3D Shape Analysis

Once the patches are extracted, we aim at studying their shape and design a similarity measure between corresponding ones on different scans under different expressions. This is motivated by the common belief that people smile, or convey any other expression, the same way, or more appropriately certain regions taking part in a specific expression undergo practically the same dynamical deformation process. We expect that certain corresponding patches associated with the same given expression will

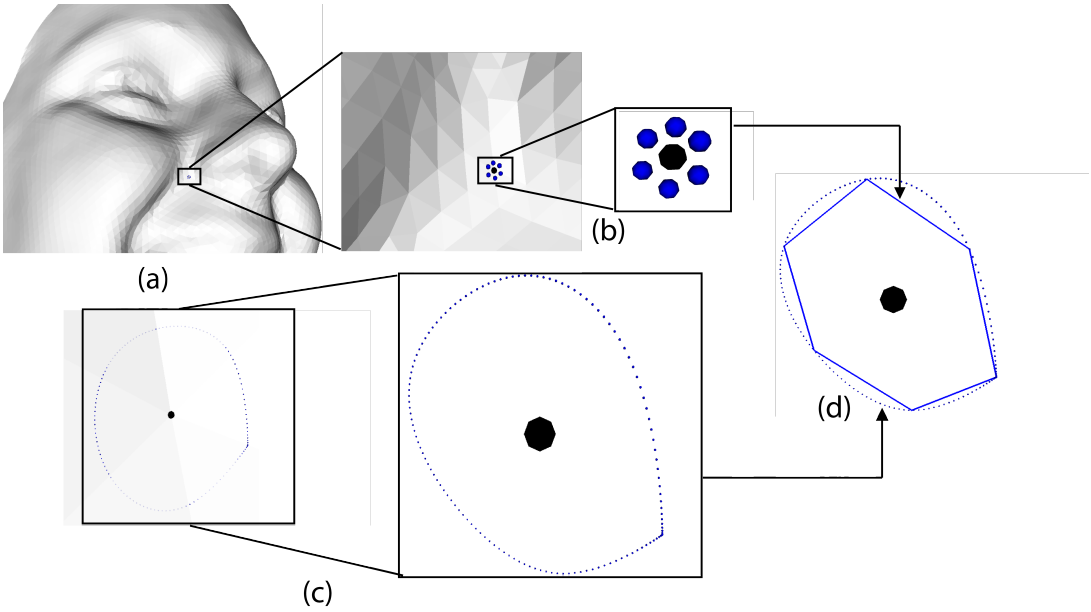

Figure 3.3 – *Extraction of the level curve associated with $\lambda = 1mm$, (b) zoomed rendering of (a), (c) point representation of the level curve resulting from a spline fitting and a down-sampling pipeline characterized by a total number over than 100 points, (d) both illustration of polyline representation of the initial extracted curve and the processed curve needed for our study.*

be deformed in a similar way, while those associated with two different expressions will deform differently. The following sections describe the shape analysis of closed curves in $\mathbb{R}^3$, initially introduced by Joshi et al. [JKSJ07], and its extension to analyze shape of local patches on facial surfaces.

3.5.1 Elastic metric

Let B be the set of all parametrized closed curves in $\mathbb{R}^3$, and β an element of B with $\beta : \mathbb{S}^1 \to \mathbb{R}^3$, where $\mathbb{S}^1$ is the unit circle and hence the domain of parametrization. β is supposed to be continuous and whose derivative is $\dot{\beta}(t)$, exists almost everywhere and never vanishes: $\dot{\beta}(t) \neq 0, \forall t$.
$\dot{\beta}$ can be seen as: $\dot{\beta}(t) = \exp(\phi(t))v(t)$, where ϕ represents the log-speed and $v(t)$ represents the direction vector as: $\phi(t) = log(\|\dot{\beta}(t)\|)$ and $v(t) = \frac{\dot{\beta}(t)}{\|\dot{\beta}(t)\|}$. Clearly $v(t)$ and ϕ completely specify $\dot{\beta}$ and the curve is seen as element in $\Phi \times \mathrm{Y}$, where $\Phi = \{\phi : \mathbb{S}^1 \to \mathbb{R}\}$ and $\mathrm{Y} = \{v : \mathbb{S}^1 \to \mathbb{S}^{n-1}\}$. Intuitively, ϕ tells us the (log of the) speed of traversal of the curve, while v tells us direction of the curve at each time t. In order to quantify the

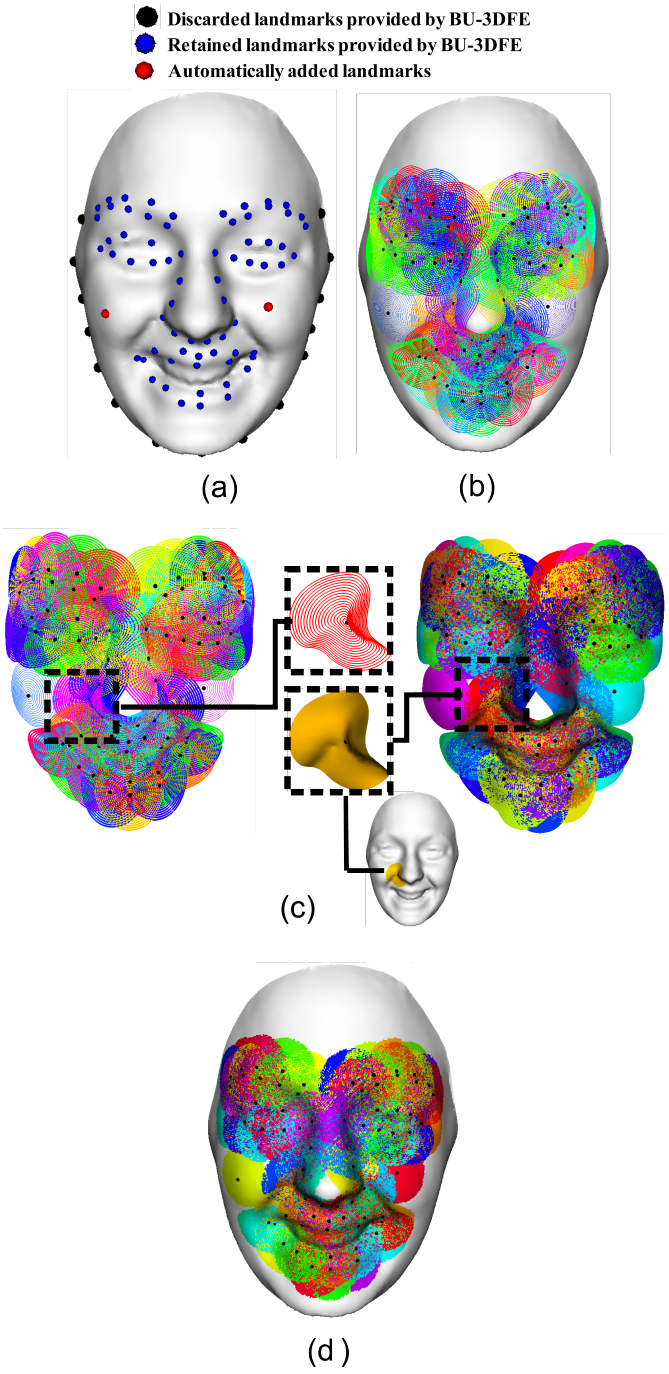

Figure 3.4 – *(a) 3D annotated facial shape model (70 landmarks); (b) 3D closed curves extracted around the landmarks; (c) 3D curve-based patches composed of 20 level curves with a size fixed by a radius $\lambda_0 = 20mm$; (d) Extracted patches on the face.*

magnitudes of perturbations of β, a metric on $\Phi \times \mathrm{Y}$ should be defined. First we note that the tangent space of $\Phi \times \mathrm{Y}$ at any point (ϕ, v) is given by

$$T_{(\phi,v)}(\Phi, \mathrm{Y}) = \{(u, v) : u \in T_\phi \Phi, v \in T_v \mathrm{S}^{n-1}\} \tag{3.2}$$

with $T_\phi\Phi = \Phi$. as it is a linear space. Y is a hypersphere in the Hilbert space $\mathbb{L}^2(\mathbb{S}^1, \mathbb{R}^3)$ and its tangent space is given by:

$$T_v\mathrm{Y} = \{f, f : \mathbb{S}^1 \to \mathbb{R}^3 \ , \forall t, < f(t), v(t) >= 0\} \tag{3.3}$$

Suppose (u^1, f^1) and (u^2, f^2) are both elements of $T_{(\phi,v)}(\Phi, \mathrm{Y})$ and let a and b be positive numbers.

Definition ***(Elastic Metric)***. For every point $(\phi, v) \in (\Phi \times \mathrm{Y})$, define an inner product on the tangent space $T_{(\phi,v)}(\Phi, \mathrm{Y})$ as:

$$< (u^1, f^1), (u^2, f^2) >= a^2 \int_{\mathbb{S}^1} u^1(t)u^2(t) \exp(\phi(t)) \, \mathrm{d}t + b^2 \int_{\mathbb{S}^1} < f^1(t), f^2(t) > \exp(\phi(t)) \, \mathrm{d}t \tag{3.4}$$

Note that $< ., . >$ in the second integral on the right denotes the standard dot product in $\mathbb{R}^n$. This elastic metric has the interpretation that the first integral measures the amount of *stretching*, since u^1 and u^2 are variations of the log speed ϕ of the curve, while the second integral measures the amount of *bending*, since f^1 and f^2 are variations of the direction v of the curve. Therefore, the choice of weights a and b determines relative penalty on bending and stretching and a family of elastic metric is formed. The use of this metric to compare radial facial curves is motivated by the fact that the groups $SO(n)$ and Γ both act by isometries [Dri11]. Let $O \in SO(n)$ acts on a facial curve β by $(O, \beta)(t) = O\beta(t)$ and $\gamma \in \Gamma$ acts on β by $(\gamma, \beta)(t)$. $O \in SO(n)$ acts on (ϕ, v) by $(O, (\phi, v)) = (\phi, Ov)$ and $\gamma \in \Gamma$ acts on (ϕ, v) by $(\gamma, (\phi, v)) = (\phi \circ \gamma + ln \circ \dot{\gamma}, v \circ \gamma)$. $O \in SO(3)$ acts by the restriction of a linear transformation on the tangent space of $\Phi \times \mathrm{Y}$: $(O, (u, f)) = (u, Of)$, where $(u, f) \in T_{(\phi,v)}(\Phi, \mathrm{Y})$ and $(u, Of) \in T_{(\phi,Ov)}(\Phi, \mathrm{Y})$.

The action of γ given in the above formula is affine linear, because of the term $ln \circ \gamma$. This make its action on the tangent space the same, but without this additive term: $(\gamma, (u, f)) = (u \circ \gamma, v \circ \gamma)$, where $(u, f) \in T_{(\phi,v)}(\Phi, \mathrm{Y})$ and $(u \circ \gamma, v \circ \gamma) \in T_{(\gamma,(\phi,v))}(\Phi, \mathrm{Y})$. Combining the action of $SO(3)$ and Γ with the inner product presented in equation 3.4 on (Φ, Y), it is easy to verify that these actions are by isometries, *ie*,
$< (O, (u_1, f_1)), (O, (u_2, f_2)) >_{(O,(\phi,v))} =< (u_1, f1), (u_2, f_2) >_{(\phi,v)}$

$$< (\gamma,(u_1,f_1)),(\gamma,(u_2,f_2)) >_{(\gamma,(\phi,v))} =< (u_1,f1),(u_2,f_2) >_{(\phi,v)}$$

We note that regardless of the values of a and b, both the groups $SO(3)$ and Γ act by isometries. An important question is: Is there some particular choice of weights a and b to make calculus easier?
We propose to use the SRV representation for its simplicity of calculus and for the belief that it finds its potential for elastically match facial curves.

3.5.2 Square Root Velocity representation SRV

In term of (ϕ, v), SRV is given by

$$q(t) = \exp(\frac{1}{2}\phi(t))v(t)$$

The tangent vector to $\mathbb{L}^2(\mathbb{S}^1,\mathbb{R}^n)$ at q is given by a simple derivation calculus as: $h = \frac{1}{2}\exp(\frac{1}{2}\phi)uv + \exp(\frac{1}{2}\phi)f$.
Let (u_1,f_1) and (u_2,f_2) denote two elements of $T_{(\phi,v)}(\Phi,\mathrm{Y})$, and let h_1 and h_2 denote the corresponding tangent vectors to $\mathbb{L}^2(I,\mathbb{R}^n)$ at q. The $\mathbb{L}^2$ inner product of h_1 and h_2 is given by:

$$< h_1,h_2 >= \int_{\mathbb{S}^1} < \frac{1}{2}\exp(\frac{1}{2}\phi)u_1v + \exp(\frac{1}{2}v)f_1, \frac{1}{2}\exp(\frac{1}{2}phi)u_2v + \exp(\frac{1}{2}v)f_2 > \mathrm{d}t \tag{3.5}$$

$$< h_1,h_2 >= \int_{\mathbb{S}^1} (\frac{1}{4}\exp(\phi)u_1u_2 + \exp(\phi) < f_1,f_2 >)\,\mathrm{d}t \tag{3.6}$$

In this computation, $v(t)$ is an element of the unit sphere hence the fact $< v(t),v(t) >= 1$ was used to reduce the formula. It was used also that $< v,f_i(t) >= 0$ since each $f_i(t)$ is a tangent vector to the unit sphere at $v(t)$.

This expression illustrates a particular elastic metric: for $a = \frac{1}{2}$ and $b = 1$. Therefore, the $\mathbb{L}^2$ metric on the shape of SRV representations corresponds to the elastic metric on $\Phi \times \mathrm{Y}$ and this makes the calculus simpler. Actually, expressed in terms of SRV, the $\mathbb{L}^2$ metric does not depend on the point at which these tangent vectors are defined. Finally, the inner

product is simply given by:

$$< h_1, h_2 >= \int_{S^1} < h_1(t), h_2(t) > \, \mathrm{d}t \tag{3.7}$$

In term of β, the SRV map is defined as: $SRV : B \rightarrow \mathbb{L}^2(\mathbb{S}^1, \mathbb{R}^3)$

$$q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}} \, . \tag{3.8}$$

if $\dot{\beta}(t) \neq 0$ and 0 otherwise.

3.5.3 3D Curve Shape Analysis

To analyze the shape of β, we shall represent it mathematically using a square-root velocity function (SRVF) defined as $q : \mathbb{S}^1 \longrightarrow \mathbb{R}^3$, where:

$$q(t) \doteq \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}} \, . \tag{3.9}$$

Where t is a parameter of $\mathbb{S}^1$ and $\|.\|$ is the Euclidean norm in $\mathbb{R}^3$.
Here q is a special function that captures the shape of β and is particularly convenient for shape analysis, as we describe next. While there are several ways to analyze shapes of closed curves, an elastic shape analysis of the parametrized curves is particularly appropriate in 3D curves analysis. This is because (1) such analysis uses a square-root velocity function representation which allows us to compare local facial shapes in presence of elastic deformations, (2) this method uses a square-root representation under which the elastic metric is reduced to the standard $\mathbb{L}^2$ metric and thus simplifies the analysis, (3) under this metric the Riemannian distance between curves is invariant to the re-parametrization.
What we mean by elastic shape analysis is basically an analysis that results when the Riemannian metric is a specific metric called the elastic metric. Mio et al. [MSJ07] presented a family of elastic metrics that quantified the bending and stretching needed to deform shapes into each other. the elastic metric allows for optimal matching of features, and is better suited for shape analysis of curves, as it is the only metric that remains invariant under re-parameterizations. The classical elastic metric for comparing shapes of curves becomes the $\mathbb{L}^2$-metric under the SRVF representation [SKJJ10]. This point is very important as it simplifies the calculus of elastic metric

to the well-known calculus of functional analysis under the $\mathbb{L}^2$-metric. We note that the squared $\mathbb{L}^2$-norm of q is given by:

$$\|q\|^2 = \int_{S^1} < q(t), q(t) > dt = \int_{S^1} \|\dot{\beta}(t)\| dt \tag{3.10}$$

which is also the length of β. Therefore, the $\mathbb{L}^2$-norm is convenient to analyze curves of specific lengths.

In order to restrict our shape analysis to closed curves, we define the set:

$$\mathcal{C} = \{q : \mathbb{S}^1 \longrightarrow \mathbb{R}^3 | \, \|q\| = 1 \int_{S^1} q(t)\|q(t)\| dt = 0\} \subset \mathbb{L}^2(\mathbb{S}^1, \mathbb{R}^3) \tag{3.11}$$

$\mathcal{C}$ denotes the set of all closed curves in $\mathbb{R}^3$, each represented by its SRVF. We note that there are two constraints for this set, the first one is given by the term $\|q\| = 1$ meaning that the elements of $\mathcal{C}$ have the same length (equal to 1). As for the second one it is revealed by the term $\int_{S^1} q(t)\|q(t)\| dt = 0\}$ that denotes the total displacement in $\mathbb{R}^3$ as a particle travel along the curve from the starting point to the final one. Setting this term equal to zero is equivalent to having a closed curve, it is the closure constraint.

We note also that $\mathbb{L}^2(\mathbb{S}^1, \mathbb{R}^3)$ denotes the set of all functions from $\mathbb{S}^1$ to $\mathbb{R}^3$ that are square integrable. Given two elements q_1 and q_2 of the space $\mathcal{C}$, we would like to quantify the similarities and dissimilarities between their corresponding shapes. It is important to remind that these quantifications should not depend on the rotation, placement and other transformations that can change curves but do not change their shapes. A more formal question is: How can we define a metric space of shapes and to compute distances between shapes as elements of this space?

Due to a non-linear closure constraint on its elements, $\mathcal{C}$ is a non-linear manifold, and there is a way to endow $\mathcal{C}$ with a Riemannian metric. Informally, a Riemannian metric is an inner product defined on the tangent spaces of the manifold. Thus we introduce $T_q(\mathcal{C})$ as the tangent space at the element q of the manifold $\mathcal{C}$. $\mathcal{C}$ becomes a Riemannian manifold if we impose an inner-product on its tangent spaces. There are several possibilities for the metric A.1.2, we adopt the $\mathbb{L}^2$ metric: so for any $u, v \in T_q(\mathcal{C})$, we define:

$$< u, v >= \int_{S^1} < u(t), v(t) > dt \ . \tag{3.12}$$

Next, we introduce some tools from covariant calculus, the calculus dealing with differentiation of tangent vector fields along paths on manifolds. More specifically, we will define covariant derivatives of vector fields. A vector field w along α implies a collection of tangent vectors along α:

$$\left\{ w(t) \in T_{\alpha(t)}(M), t \in [0,1] \right\}.$$

Covariant Derivative Let $\mathcal{H}$ be the set of all differentiable paths in $\mathcal{C}$. For a given path $\alpha \in \mathcal{H}$ and a vector $w \in T_\alpha(\mathcal{H})$ (a vector field along α), define the covariant derivative of w along α, denoted $\frac{Dw}{dt}$, to be the vector field obtained by projecting $\frac{dw}{dt}(t)$ onto the tangent space $T_{\alpha(t)}(\mathcal{C})$, for all t. Note that given $w \in T_\alpha(\mathcal{H})$, we also have $\frac{Dw}{dt} \in T_\alpha(\mathcal{H})$. Since $M \subset V$ is not a linear submanifold, the vector $\frac{Dw}{dt}$, for any t, is an element of V, but not necessarily of $T_{\alpha(t)}(\mathcal{C})$. The projection ensures that the resulting set of vectors form a vector field along α which is tangent to $\mathcal{C}$ at each point.

3.5.4 Parallel translation

A vector field $\tilde{w}$ is called the forward parallel transport of a tangent vector $w \in T_{\alpha(t)}(\mathcal{C})$, along α, if and only if $\tilde{w}(0) = w$ and $\frac{D\tilde{w}(t)}{dt} = 0$ for all $t \in [0,1]$. Similarly, $\tilde{w}$ is called the backward parallel translation of a tangent vector $w \in T_{\alpha(1)}(\mathcal{C})$, along α, if and only if $\tilde{w}(1) = w$ and $\frac{D\tilde{w}(t)}{dt} = 0$ for all $t \in [0,1]$.

So far we have described a set of closed curves and have endowed it with a Riemannian metric. Next we consider the issue of representing the *shapes* of these curves. It is easy to see that a subset of elements of $\mathcal{C}$ can represent curves with the same shape. For example, if we rotate a curve in $\mathbb{R}^3$, we get a different SRVF but its shape remains unchanged. Another similar situation arises when a curve is re-parametrized; a re-parameterization changes the SRVF of curve but not its shape. In order to handle these variabilities, we define orbits of the rotation group $SO(3)$ and the re-parameterization group Γ as the equivalence classes in $\mathcal{C}$. For

example, for a curve $\beta : \mathbb{S}^1 \longrightarrow \mathbb{R}^3$ and a function $\gamma : \mathbb{S}^1 \to \mathbb{S}^1$, $\gamma \in \Gamma$, the curve $\beta \circ \gamma$ is a re-parameterization of β. The corresponding SRVF changes according to $q(t) \mapsto \sqrt{\dot{\gamma}(t)}q(\gamma(t))$. We set the elements of the orbit:

$$[q] = \{\sqrt{\dot{\gamma}(t)}Oq(\gamma(t))|O \in SO(3), \ \ \gamma \in \Gamma\}\ , \tag{3.13}$$

to be equivalent from the perspective of shape analysis. The set of such equivalence classes, denoted by $\mathcal{S} \doteq \mathcal{C}/(SO(3) \times \Gamma)$ is called the *shape space* of closed curves in $\mathbb{R}^3$. $\mathcal{S}$ inherits a Riemannian metric from the larger space $\mathcal{C}$ due to the quotient structure.

The main ingredient in comparing and analysing shapes of curves is the construction of a geodesic between any two elements of $\mathcal{S}$, under the Riemannian metric given in Eq.(3.12). Given any two curves β_1 and β_2, represented by their SRVFs q_1 and q_2, we want to compute a geodesic path between the orbits $[q_1]$ and $[q_2]$ in the shape space $\mathcal{S}$.

However, due to the existence of the closure constraint, this introduces a strong nonlinearity in the formulation system, and manifolds resulting from the closure constraint become relative more complicated. The explicit expressions for geodesics are not available and we have to resort to numerical approaches for constructing geodesics. This challenging task has been resolved using a *path-straightening approach* which was introduced in [KS06]. The basic idea here is to connect the two points q_1 and q_2 by an arbitrary initial path α and to iteratively update this path using the negative gradient of an energy function $E[\alpha] = \frac{1}{2}\int_{\mathbb{S}^1} < \dot{\alpha}(s), \dot{\alpha}(s) > ds$, more details about the definition of E is given in Annex A.1. The interesting part is that the gradient of E has been derived analytically and can be used directly for updating α. As shown in [KS06], the critical points of E are actually geodesic paths in $\mathcal{S}$. Thus, this gradient-based update leads to a critical point of E which, in turn, is a geodesic path between the given points. In the remainder of this chapter, we will use the notation $d_{\mathcal{S}}(\beta_1, \beta_2)$ to denote the length of the geodesic in the *shape space* $\mathcal{S}$ between the orbits q_1 and q_2, to reduce the notation.

3.5.5 3D Patches Shape Analysis

Now, we extend ideas developed in the previous section from analyzing shapes of curves to the shapes of patches. As mentioned earlier, we are going to represent a number of l patches of a facial surface S with an indexed collection of the level curves of the $\|r_l - .\|$ function (Euclidean distance from the reference point r_l). That is, $P_l \leftrightarrow \{c_\lambda^l, \lambda \in [0, \lambda_0]\}$, where c_λ^l is the level set associated with $\|r_l - .\| = \lambda$. Through this relation, each patch has been represented as an element of the set $\mathcal{S}^{[0,\lambda_0]}$. In our framework, the shapes of any two patches are compared by comparing their corresponding level curves. Given any two patches P_1 and P_2, and their level curves $\{c_\lambda^1, \lambda \in [0, \lambda_0]\}$ and $\{c_\lambda^2, \lambda \in [0, \lambda_0]\}$, respectively, our idea is to compare the patches curves c_λ^1 and c_λ^2, and to accumulate these differences over all λ. More formally, we define a distance $d_{\mathcal{S}^{[0,\lambda_0]}}$ given by:

$$d_{\mathcal{C}^{[0,\lambda_0]}}(P_1, P_2) = \int_0^{\lambda_0} d_{\mathcal{C}}(c_\lambda^1, c_\lambda^2) d\lambda \ . \tag{3.14}$$

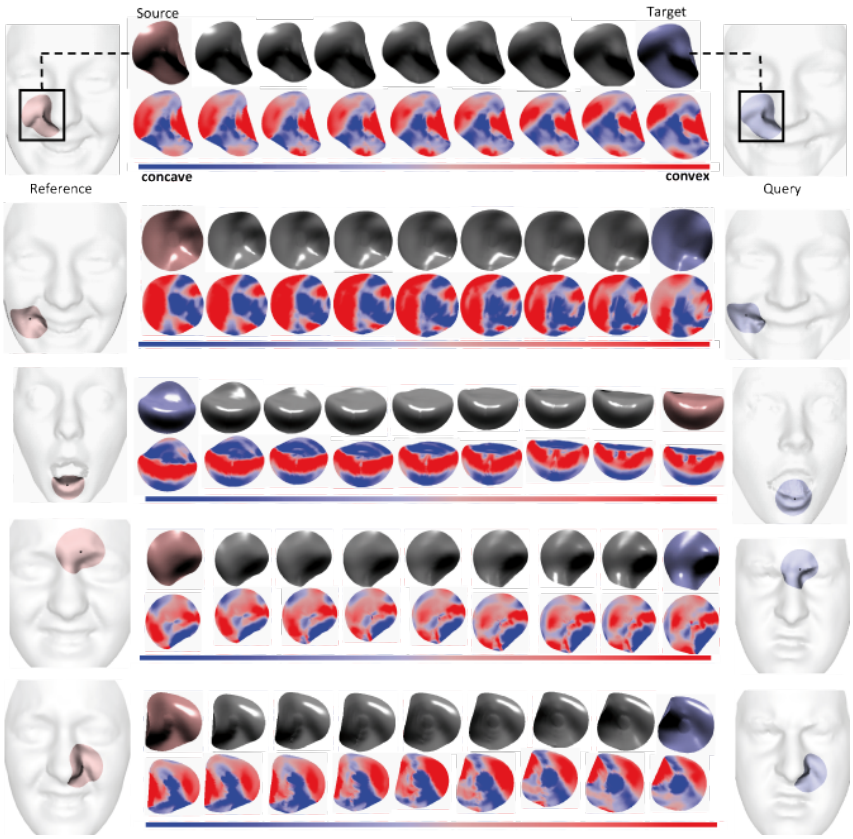

Figure 3.5 – *Examples of intra-class (same expression) geodesic paths with shape and mean curvature mapping between corresponding patches.*

In addition to the distance $d_{\mathcal{S}^{[0,\lambda_0]}}(P_1, P_2)$, which is useful in biometry and other classification experiments, we also have a geodesic path in $\mathcal{S}^{[0,\lambda_0]}$ between the two points represented by P_1 and P_2. This geodesic corresponds to the optimal elastic deformations of facial curves and, thus, facial surfaces from one to another. Fig. 3.5 shows some examples of geodesic paths that are computed between corresponding patches associated with shape models sharing the same expression, and termed *intra-class geodesics*. In the first column we illustrate the source, which represents scan models of the same subject, but under different expressions. The third column represents the targets as scan models of different subjects. As for the middle column, it shows the geodesic paths. Each of the geodesic paths is obtained through the computation of a set of geodesic paths between correspondent level curves forming the source and target patches. The geodesic paths between curves were computed using the path-straightening approach. It is an energy optimization method based on the gradient descent method along the steepest descent direction from the source curve to the target curve. Each element of this path is a q-element that is used to reconstruct the curve β in $\mathbb{R}^3$. Each element of the overall geodesic path is a set of level curves, this set is employed to reconstruct the surface of an intermediate patch by applying Delaunay triangulation. Along the geodesic we illustrate some intermediates patches corresponding at some chosen steps of the gradient descent technique.
In each row of the geodesic path, we have both the shape and the mean curvature mapping representations of the patches along the geodesic path from the source to the target. The mean curvature representation is added to identify concave/convex areas on the source and target patches and equally-spaced steps of geodesics. This figure shows that certain patches, belonging to the same class of expression, are deformed in a similar way. In contrast, Fig. 3.6 shows geodesic paths between patches of different facial expressions. These geodesics are termed *inter-class geodesics*. Unlike the intra-class geodesics shown in Fig. 3.5, these patches deform in a different way.

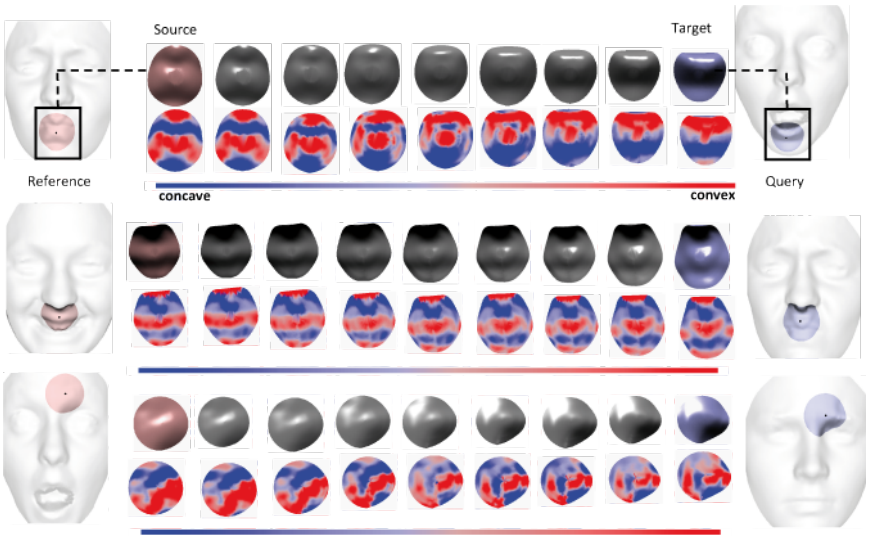

Figure 3.6 – *Examples of inter-class (different expressions) geodesic paths between source and target patches.*

3.6 Feature Vector Generation for Classification

In order to classify expressions, we build a feature vector for each facial scan. Given a candidate facial scan of a person j, facial patches are extracted around facial landmarks. For a facial patch P_j^i, a set of level curves $\{c_\lambda\}_j^i$ are extracted centered on the i^{th} landmark. Similarly, a patch P_{ref}^i is extracted in correspondence to landmarks of a reference scans *ref*. The length of the geodesic path between each level curve and its corresponding curve on the reference scan are computed using a Riemannian framework for shape analysis of 3D curves (see Sections 3.5.3 and 3.5.5). The shortest path between two patches at landmark i, one in a candidate scan and the other in the reference scan, is defined as the sum of the distances between all pairs of corresponding curves in the two patches as indicated in Eq. (3.14). The feature vector is then formed by the distances computed on all the patches and its dimension is equal to the number of used landmarks $N = 70$ (i.e., 68 landmarks are used out of the 83 provided by BU-3DFED and the two additional cheek points). The i^{th} element of this vector represents the length of the geodesic path that separates the relative patch to the corresponding one on the reference face scan. All feature vectors computed on the overall dataset will be labeled and used as input data to machine learning algorithms such as Multi-boosting A.2.2 and

SVM A.3, where Multi-boosting is an extension of the successful Adadoost A.2.1 technique.

3.7 Recognition Experiments

To investigate facial expression recognition, we have applied our proposed approach on a dataset that is appropriate for this task. In this Section, we describe the experiments, obtained results and comparisons with related work.

3.7.1 Experimental Setting

For the goal of performing identity-independent facial expression recognition, the experiments were conducted on the BU-3DFE static database. A dataset captured from 60 subjects were used, half (30) of them were female and the other half (30) were male, corresponding to the high and highest intensity levels 3D expressive models (03-04). These data are assumed to be scaled to the true physical dimensions of the captured human faces. Following a similar setup as in [GWLT09a], we randomly divided the 60 subjects into two sets, the training set containing 54 subjects (648 samples), and the test set containing 6 subjects (72 samples).

To drive the classification experiments, we arbitrarily choose a set of six reference subjects with its six basic facial expressions. We point out that the selected reference scans do not appear neither in the training nor in the testing set. These references, shown in Fig. 3.7, with their relative expressive scans corresponding to the highest intensity level, are taken to play the role of representative models for each of the six classes of expressions. For each reference subject, we derive a facial expression recognition experience.

3.7.2 Discussion of the Results

Several facial expression recognition experiments were conducted with changing at each time the reference. Fig. 3.7 illustrates the selected references (neutral scan). Using the *Waikato Environment for Knowledge Analysis*

(Weka) [HFH*09], we applied the Multiboost algorithm with three weak classifiers, namely, Linear Discriminant Analysis (LDA), Naive Bayes (NB), and Nearest Neighbor (NN), to the extracted features, and we achieved average recognition rates of 98.81%, 98.76% and 98.07%, respectively. We applied the SVM linear classifier as well, and we achieved an average recognition rate of 97.75%. We summarize the resulting recognition rates in Table 3.1.

Classifier	Multiboost-LDA	Multiboost-NB	Multiboost-NN	SVM-Linear
Recognition rate	**98.81%**	98.76%	98.07%	97.75%

Table 3.1 – *Classification results using local shape analysis and several classifiers.*

We point out that these recognition rates are obtained by 10-fold cross validation and the highest accuracy is achieved by Multiboost-LDA classifier. We note that different selections of the reference scans do not affect significantly the recognition results and there is no large variations in recognition rates values. The reported results represent the average over the six runned experiments. The Multiboost-LDA classifier achieves the highest recognition rate and shows a better performance in terms of accuracy than the other classifiers. This is mainly due to the capability of the LDA-based classifier to transform the features into a more discriminative space and, consequently, result in a better linear separation between facial expression classes.

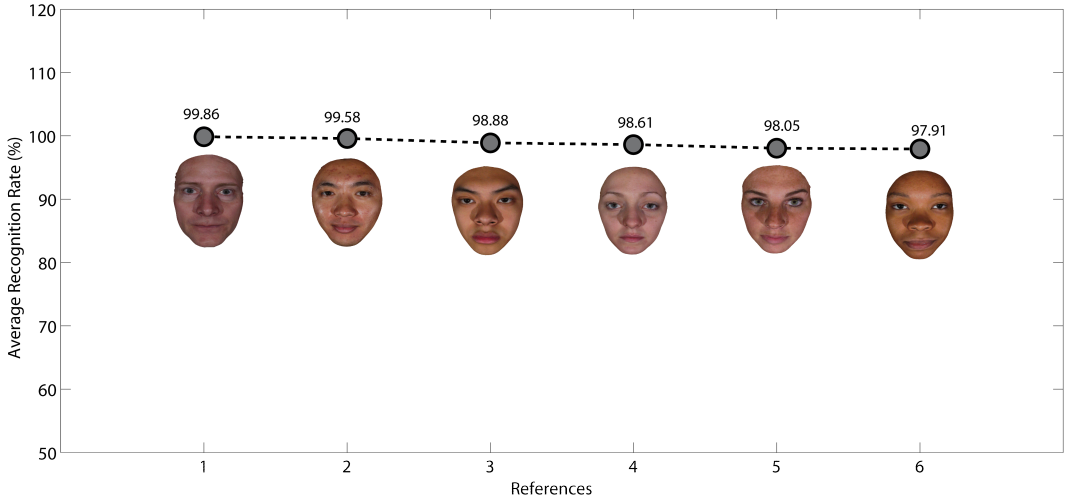

Figure 3.7 – *Different facial expression average recognition rates obtained using different reference subjects (using Multiboost-LDA).*

The average confusion matrix relative to the the best performing classification using Multiboost-LDA is given in Table 3.2.

%	AN	DI	FE	HA	SA	SU
AN	**97.92**	1.11	0.14	0.14	0.69	0.0
DI	0.56	**99.16**	0.14	0.0	0.14	0.0
FE	0.14	0.14	**99.72**	0.0	0.0	0.0
HA	0.56	0.14	0.0	**98.60**	0.56	0.14
SA	0.28	0.14	0.0	0.0	**99.30**	0.28
SU	0.14	0.56	0.0	0.0	1.11	**98.19**

Table 3.2 – *Average confusion matrix given by Multiboost-LDA classifier.*

In order to better understand and explain the results mentioned above, we apply the Multiboost algorithm on feature vectors built from distances between patches for each class of expression. In this case, we consider these features as weak classifiers. Then, we look at the early iterations of the Multiboost algorithm and the selected patches in each iteration.

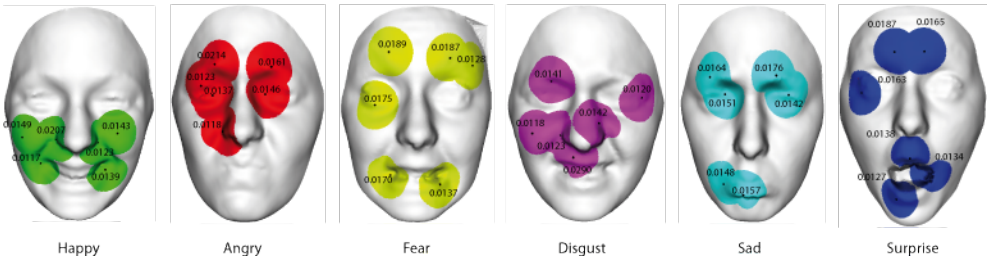

Figure 3.8 – *Selected patches at the early few iterations of Multiboost classifier for the six facial expressions (Angry, Disgust, Fear, Happy, Sadness, Surprise) with their associated weights.*

Fig. 3.8 illustrates for each class of expression the most relevant patches. Notice that, for example, for the Happy expression the selected patches are localized in the lower part of the face, around the mouth and the chin. As for the Surprise expression, we can see that most relevant patches are localized around the eyebrows and the mouth region. It can be seen that patches selected for each expression lie on facial muscles that contribute to this expression.

3.7.3 Comparison with Related Work

In Table 3.3 results of our approach are compared against those reported in [TH08], [SD07], and [WYWS06], on the similar experimental dataset

composed of 60 subjects of the BU-3DFE database. The differences between approaches should be noted: Tang et al. [TH08] performed automatic feature selection using normalized Euclidean distances between 83 landmarks, Soyel et al. [SD07] calculated six distances using a distribution of 11 landmarks, while Wang et al. [WYWS06] derived curvature estimation by locally approximating the 3D surface with a smooth polynomial function. In comparison, our approach capture the 3D shape information of local facial patches to derive shape analysis. For assessing how the results of their statistical analysis will generalize to an independent dataset, in [WYWS06] a 20-fold cross-validation technique [1] was used, while in [TH08] and [SD07] the authors used 10-fold cross-validation [2] to validate their approach. For example, within the 10-fold cross-validation experiment, we split data (60 subjects) into 10 sets (of size 6). Then, the classifier is trained on 9 subsets and tested on the remaining subset. Finally, the experiment is repeat 10 times and one takes the mean accuracy. Thus, the 10 subsets are used exactly once as the test data.

Table 3.3 – *Comparison of this work with respect to previous work [TH08], [SD07] and [WYWS06].*

Cross-validation	*This work*	*Tang et al. [TH08]*	*Soyel et al. [SD07]*	*Wang et al. [WYWS06]*
10-fold	**98.81%**	95.1%	91.3%	-
20-fold	**92.75%**	-	-	83.6%

However, as pointed out in [GWLT09b], since from experiment to another, the accuracies can vary, in order to permit a fair generalization and obtain stable accuracies, we run 1000 independent experiments and averaged the results (100$\times$ 10-fold cross-validation). Accordingly, our approach achieved 98.04$\pm$1.65% of average recognition rate (1.65% represent the standard deviation). Following similar protocol, Gong et al. [GWLT09b], Berretti et al. [BBADdB11] and Li et al. [LMC11] reported results given in table 3.4.

[1] 57 subjects for training *vs.* 3 subjects for testing runned 20 times

[2] 54 subjects for training *vs.* 6 subjects for testing

Table 3.4 – *Comparison of this work with respect to previous work [GWLT09b], [BBADdB11] and [LMC11].*

	This work	*Gong et al. [GWLT09b]*	*Berretti et al. [BBADdB11]*	*Li et al.[LMC11]*
1000 experiments	**98.04±1.65%**	76.22%	77.54%	82.01%

3.7.4 Non-frontal View Facial Expression Recognition

In real world situations, frontal view facial scans may not be always available. Thus, non-frontal view facial expression recognition is a challenging issue that needs to be treated. We were interested in evaluating our approach on facial scan under large pose variations. By rotating the 3D shape models in the y-direction, we generate facial scans under six different non-frontal views corresponding to 15°, 30°, 45°, 60°, 75° and 90° rotation. We assume that shape information is unavailable for the occluded facial regions due to the face pose. For each view, we perform facial patches extraction around the visible landmarks in the given scan. In cases where a landmark is occluded, or where the landmark is visible, but the region nearby is partially occluded, we treat it as a missing data problem for all faces sharing this view. In these cases, we are not able to compute the geodesic path between corresponding patches. The corresponding entries in the distance matrix are blank and we fill them using an imputation technique [BM03]. In our experiments we employed the mean imputation method, which consists of replacing the missing values by the means of values already calculated in frontal-view scenario obtained from the training set. Let $d_{ijk} = d_{\mathcal{S}^{[0,\lambda_0]}}(P_i^k, P_j^k)$ be the geodesic distance between the k^{th} patch belonging to subjects i and j $(i \neq j)$. In case of frontal view (fv), the set of instances X_i^{fv} relative to the subject i need to be labeled and is given by:

$$\mathbf{X_i^{fv}} = \begin{pmatrix} d_{i11} & \dots & d_{i1k} & \dots & d_{i1N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{ij1} & \dots & d_{ijk} & \vdots & d_{ijN} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{iJ1} & \dots & d_{iJk} & \dots & d_{iJN} \end{pmatrix}$$

where N is the number of attributes. In case of non-frontal view (nfv), if an attribute k is missing, we replace the k^{th} column vector in the distance matrix X_i^{nfv} by the mean of geodesic distances computed in the frontal-view case, with respect to the k^{th} attribute and given by: $m_k^{fv} = \frac{\sum_{j=1}^{J} d_{ijk}}{J}$, where J is the total number of instances.

$$\mathbf{X_i^{nfv}} = \begin{pmatrix} d_{i11} & \dots & m_k^{fv} & \dots & d_{i1N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{ij1} & \dots & m_k^{fv} & \vdots & d_{ijN} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{iJ1} & \dots & m_k^{fv} & \dots & d_{iJN} \end{pmatrix}$$

To evaluate the robustness of our approach in a context of non-frontal views, we derive a view-independent facial expression recognition. Error recognition rates are evaluated throughout different testing facial views using the four classifiers trained only on frontal-view facial scans. The Fig. 3.9 shows the average error rates of the four classification methods. The Multiboost-LDA shows the best performance for facial expression classification on the chosen database. From the figure, it can be observed that the average error rates increase with the rotation angle (values from 0 to 90 degrees of rotation are considered), and the Multiboost-LDA is the best performing methods also in the case of pose variations. As shown in this figure, recognition accuracy remains acceptable, even only 50% of data (half face) are available when we rotate the 3D face by 45 degree in y-direction.

3.7.5 Sensitivity to Landmarks Mis-localization

It is known that the automatic 3D facial feature points detection is a challenging problem. The most difficult task remains the localization of points around the eyebrow regions, which appear to play an important role in the expression of emotions. The effect of the mis-localization of the landmarks has been addressed in a specific experiment. We considered the eyebrow regions in that the points in these regions are expected to be the most difficult to detect automatically. In these regions, we added noise to the landmarks provided with the BU-3DFED. In particular, we added noise to

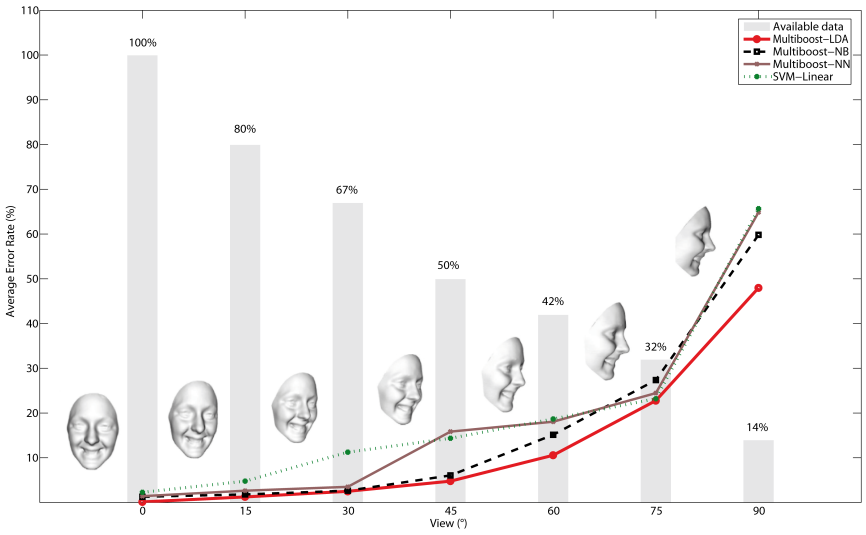

Figure 3.9 – *The average error rates of six expressions with different choice of views corresponding to the best reference and using different classifiers.*

the position of the landmarks by moving them randomly in a region with a radius of *10mm*, as illustrated Fig. 3.10 by blue circles. Then we performed expression recognition experiments with such noisy landmarks. The results are reported in Fig. 3.10. We notice that using the Multiboost-LDA classifier the recognition rate decreases but still shows better performance than the other classifiers, and even with a recognition rate equal to 85.64% the result still outperforms the one reported in Wang et al [WYWS06].

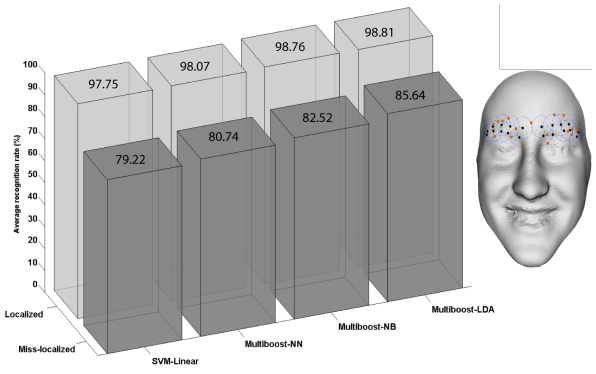

Figure 3.10 – *Recognition experiment performed adding noise to the eyebrow landmarks (random displacement).*

3.8 Conclusions

In this chapter we presented a novel approach for identity-independent facial expression recognition from 3D facial static images. Our idea was to describe the changes due to facial expressions as a deformation in the vicinity of facial patches in 3D shape scan. An automatic extraction of local curve based patches within the 3D facial surfaces was proposed. These patches were used as local shape descriptors for facial expression representation. A Riemannian framework was applied to compute the geodesic path between corresponding patches. Qualitative (inter and intra-geodesic paths) and quantitative (geodesic distances) measures of the geodesic path were explored to derive shape analysis. The geodesic distances between patches were labeled with respect to the six prototypical expressions and used as samples to train and test machine learning algorithms. Using Multiboost algorithm for multi-class classification, we achieved a 98.81% average recognition rate for six prototypical facial expressions on the BU-3DFE database. We demonstrated the robustness of the proposed method to pose variations. In fact, the obtained recognition rate remain acceptable (over 93%) even half of the facial scan is missed.

Although the results of our approach outperforms state-of-the-art related methods, it still has some limitations that require a thorough analysis and future work. The major limitation of our approach is that it relies on a large set of landmark points that were provided by the BU-3DFE database creators. The dependence to a large number as well as an accurate and well defined points make our approach semi-automatic. One possible future work could investigate on key points detection technique [ccAS08] that will make our approach fully automatic. In the next chapter we will address the dynamic 3D facial expression recognition problem and we will present a somehow different approach, which is fully automatic, that will alleviate the mentioned limitation and offer the possibility to be turned to real-world application.

4 Dynamic 3D Facial Expression Recognition

4.1 Introduction

Automatic recognition of facial expressions has emerged as an active research field with applications in several different areas, such as human-machine interaction, psychology, computer graphics, facial animation of 3D avatars, etc. The first systematic studies on facial expressions date back to the late 70s with the pioneering work of Ekman [EF78b]. In these studies, it is evidenced that, apart the *neutral* expression, the *prototypical* facial expressions can be categorized into six classes, representing *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*. This categorization of facial expressions has been proved to be consistent across different ethnicities and cultures, so that these expressions are in some sense "universally" recognized. In his studies, Ekman also evidenced that facial expressions can be coded through the movement of face points as described by a set of *action units*. In fact, there is the awareness that facial expressions are highly dynamical processes and looking at sequences of face instances, rather than to still images, can help to improve the recognition performance. More properly, facial expressions can be seen as dynamical processes that involve the 3D space and the temporal dimension (3D plus time, referred to as 4D), rather than being just a static or dynamic 2D behavior. In addition, 3D face scans are expected to feature less sensitivity to lighting conditions and pose variations. These considerations motivated a progressive shift from 2D to 3D in performing facial shape analysis, with the research on 3D facial expression recognition gaining a great impulse thanks to the recent availability of new databases, like the *Binghamton University* BU-3DFE [YWS*06], and the *Bosphorus* database [SAD*08]. Now, the introduction of appropriate data sets, such as the BU-4DFE developed at *Binghamton University* [YCS*08], makes also possible the study and recognition of facial expressions from 4D data. This trend is also inspired by the revolution of inexpensive acquisition devices such as the consumer 3D cameras.In such conditions, the extension of traditional methods developed for expression recognition from 2D videos or from 3D static models can be not effective or even possible and new solutions are required.

The main goal of this work is to propose and experiment an innovative

solution for 4D facial expression recognition. In order to motivate our approach and relate it to the state-of-the-art, in the following we provide an overview of existing solutions for 3D and 4D facial expression recognition.

4.1.1 Methodology and Contributions

Large part of existing works on 3D facial expression recognition rely on the presence of landmarks accurately identified on the face surface. The fact that several landmarks are not automatically detectable, and the precision required for their positioning, demand for manual annotation. This limits the applicability of many approaches and makes them difficult to be extended to the 4D case. At the same time, solutions specifically tailored for 4D are still preliminary, being semi-automatic or capable to discriminate just between a subset of expressions.

We propose an automatic facial expression recognition approach that exploits the motion extracted from 3D facial videos. An overview of the proposed approach is given in Fig. 4.1. In the preprocessing stage, the 3D mesh in each frame is first aligned to the previous one and then cropped. The 3D motion is then captured based on a dense scalar field that represents the 3D deformation between two successive frames. Then, we use Linear Discriminant Analysis (LDA) to transform derived feature space to reduce the dimension of the feature space. Given the selected features, they will be used to train HMMs. Temporal modeling of each expressive 3D frame sequences is performed via neutral-onset-apex-offset HMMs. Expression classification is then performed by using HMMs trained with the time variations of the extracted features. Experimental results show that the proposed approach is capable to improve state of the art performance on the BU-4DFE.

The main contributions of our approach are as follows:

i. A novel Scalar Field defined on radial curves of 3D faces; the scalar field grounds on Riemannian shape analysis and is capable to accurately capture the deformations occurring between 3D faces represented as sets of radial curves;

ii. A new approach for facial expression recognition from 3D dynamic

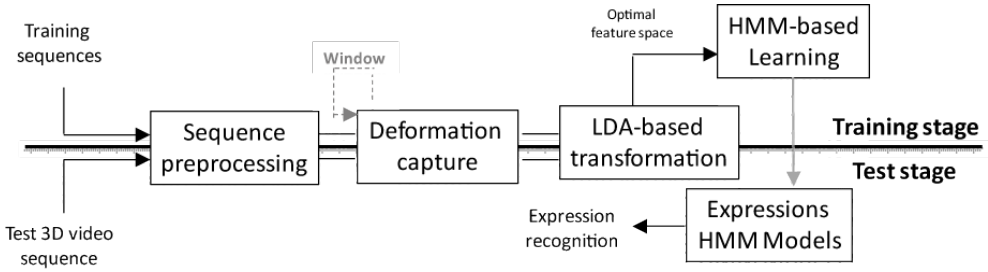

Figure 4.1 – *Overview of the proposed approach in training and test stages, including preprocessing, 3D deformation capture, dimension reduction, and HMM-based classification.*

sequences, which combines the high descriptiveness of scalar field extracted from successive 3D faces of a sequence with the temporal modeling and classification provided by HMMs;

iii. A thorough experimental evaluation that compares the proposed solution with state of the art methods on a common data set and setting.

The rest of the chapter is organized as follows: In Sect. 4.2, a face representation model is proposed that captures facial features relevant to categorize expression variations in 3D dynamic sequences. In Sect. 4.3, the HMM based classification of the selected features is addressed. Experimental results and comparative evaluation obtained on the BU-4DFE are reported and discussed in Sect. 4.4. Finally, conclusion is presented in Sect. 4.5.

4.2 3D Shape Motion Analysis

4.2.1 3D Facial Surface Representation

We have emphasized in earlier section 3.3 on the importance of surface representations of three dimensional data. A number of these representations have been listed and they all describe the data structures in a digital format that affects the way that these data are employed. It can affect retrieval and processing times, degree of accuracy, scalability, and ease with which data may be manipulated, modified, and otherwise enhanced for greater applicability. The research community dealing with structured

data recognizes the impact of such data have on the ability to meet the demands of a particular application.

Likewise the approach we have proposed for static 3D facial expression recognition in chapter 3, in which we adopted the level curves representation, we propose to apply curve-based representation to deal with dynamic 3D FER. But this time, we adopt the radial curve representation of facial surfaces (see Fig. 4.3(a)). Using indexed collection of parameterized radial curves allows to approximate facial surfaces and to capture the shape information. The other practical uses of such representation lies in accelerating the processing times, since that it relies on only one key point, which is the nose tip, and in permitting simplified application of our framework.

4.2.2 Shape-based Deformation Capture

One basic idea to capture facial deformation across 3D video sequences is to track densely meshes' vertices along successive 3D frames. To do so, as the meshes resolutions vary across 3D video frames, establishing a dense matching on consecutive frames is necessary. Sun et al. [SY08] proposed to adapt a generic model (a tracking model) to each 3D frame. However, a set of 83 predefined key-points is required to control the adaptation based on radial basis function. The main limitation is that the 83 points are manually annotated in the first frame of each sequence. A second solution is presented by Sandbach et al. [SZPR12], where the authors used an existing non-rigid registration algorithm (FFD) based on B-splines interpolation between a lattice of control points. The dense matching is a step of preprocessing stage to estimate a motion vector field between frames t and $t-1$. However, the results provided by the authors are limited to three facial expressions: *happy*, *angry* and *surprise*.

Here we are are going use differential geometry and materials that have been defined in section 3.5. But this time we are dealing with radial curve representation, hence there will modify some of the previous mathematical representations and use more explicit calculus.

Let B be the set of all parametrized open curves in $\mathbb{R}^3$. and β an

element of B with $\beta : I \to \mathbb{R}^3$, since β is an open curve $I = [0,1]$. β keeps all properties we have seen (i.e. continuous, whose derivative exists and that never vanishes). β can be represented by the SRVF $q(t)$, given by: $q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}$. We emphasize on the fact that this specific representation has the advantage of capturing the shape of the curve β. Let $\mathcal{C}$ be the space of square-root velocity functions (or the space of all open curves) defined by: $\mathcal{C} = \{q : I \to \mathbb{R}^3, \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^3)$, where $\|.\|$ implies the $\mathbb{L}^2$ norm. Since the elements of this space are open curves, the closure constraint is omitted and no longer compulsory for this specific study.

In previous chapter we have seen that the shortest path α^* on the manifold $\mathcal{C}$ between the two closed curves q_1 and q_2 is a critical point of some energy $E = \frac{1}{2}\int < \dot{\alpha}(t), \dot{\alpha}(t) > dt$ where α denotes a path on the manifold $\mathcal{C}$ between q_1 and q_2, $\dot{\alpha} \in T_\alpha(\mathcal{C})$ is a tangent vector field on the path $\alpha \subset \mathcal{C}$ and $< . >$ denotes the $\mathbb{L}^2$ inner product on the tangent space. Unlike closed curves the shape analysis of open curves is much simpler. In our case, as the elements of $\mathcal{C}$ have a unit $\mathbb{L}^2$ norm, $\mathcal{C}$ is a Hypersphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^3)$. Using the Riemannian structure we can write explicit forms for geodesics between any two open curves $q_1, q_2 \in \mathcal{C}$ and it is simply given by the minor arc of great circle connecting them on this Hypersphere, $\alpha^* : [0,1] \to \mathcal{C}$, given by the following expression:

$$\alpha^*(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2\right) , \tag{4.1}$$

where $\tau \in [0,1]$ and $\theta = \cos^{-1}\langle q_1, q_2\rangle$. And the length of the geodesic path (geodesic distance) is given by:

$$d_{\mathcal{C}}(q_1, q_2) = \cos^{-1}\langle q_1, q_2\rangle \tag{4.2}$$

We point out that $sin(\theta) = 0$ if the distance between the two curves is null, in other words $q_1 = q_2$. In this case, for each τ, $\alpha^*(\tau) = q_1 = q_2$. The tangent vector field on this geodesic is then given as $\frac{d\alpha^*}{d\tau} : [0,1] \to T_\alpha(\mathcal{C})$:

$$\frac{d\alpha^*}{d\tau} = \frac{-\theta}{\sin(\theta)} \left(\cos((1-\tau)\theta)q_1 - \cos(\theta\tau)q_2\right) . \tag{4.3}$$

Knowing that on geodesic path, the covariant derivative 3.5.3 of its tangent vector field is equal to 0, $\frac{d\alpha^*}{d\tau}$ is parallel along the geodesic α^* and

we shall represent it with $\frac{d\alpha^*}{d\tau}|_{\tau=0}$. Accordingly, Eq. (4.3) becomes:

$$\frac{d\alpha^*}{d\tau}|_{\tau=0} = \frac{\theta}{\sin(\theta)}\left(q_2 - \cos(\theta)q_1\right) \ , \tag{4.4}$$

with $\theta \neq 0$.

We illustrate a geodesic path between two elements of the sphere in Fig. 4.2.g, the covariant derivative of this vector field corresponds to the tangential component of the tangent vector. Thus, $\frac{d\alpha^*}{d\tau}|_{\tau=0}$ is sufficient to represent this vector field, the remaining vectors can be obtained by parallel transport 3.5.4 of $\frac{d\alpha^*}{d\tau}|_{\tau=0}$ along the geodesic α^*.

In the practice, the first step to capture the deformation between two given 3D faces F_1 and F_2 is to extract the radial curves. Let β^1_φ and β^2_φ denote the radial curves that make an angle φ with the vertical plan. The tangent vector field $\dot{\alpha_\varphi}^*$ that represents the energy E that is needed to deform β^1_φ to β^2_φ is then calculated for each index φ. We consider the magnitude of this vector field at each point k of the curve for building a scalar vector field on the facial surface $V^k_\varphi = ||\dot{\alpha}^*_\varphi|_{(\tau=0)}(k)||$, where φ denotes the angle to the vertical radial curve and k denotes a point on this curve. This scalar field quantifies the local deformation between the faces F_1 and F_2. Algorithm 1 summarizes the steps of the proposed approach.

Algorithm 1 Scalar field computation.

Input: Facial surfaces F_1 and F_2, K: number of points on a curve; φ_0: angle between successive radial curves.

Output: $V = V^k_\varphi$: the scalar field between the two faces.

$\varphi = 0$;

while $\varphi < 2\Pi$ **do**

  for $i \leftarrow 1$ **to** 2 **do**

    Extract the curves β^1_φ and β^2_φ;

    Compute the corresponding SRVFs $q^i_\theta(t) \doteq \frac{\dot{\beta}^i_\theta(t)}{\sqrt{\|\dot{\beta}^i_\theta(t)\|}} \in \mathcal{C}$;

  Compute the deformation vector $\frac{d\alpha^*}{d\tau}|_{\tau=0}$ using Eq. (4.4);

  for $k \leftarrow 1$ **to** K **do**

    Compute the local deformation V^k_φ as the module of $\frac{d\alpha^*}{d\tau}|_{\tau=0}(k)$;

  $\varphi = \varphi + \varphi_0$;

Fig. 4.2 illustrates the proposed idea. A neutral mesh is reported on the

left. The scalar field is computed between the neutral face and apex frames of each expression. The values of the scalar field needed to be applied on that face to convey the different 6 universal expressions are reported using range of colors. In particular, black colors represent the highest deformations whereas the lower scalar values are represented in blue. It can be observed, as the regions with high deformation lie in different parts of the face for different expressions. For example, as intuitively expected, the mouth and the cheeks are mainly deformed for happiness expression.

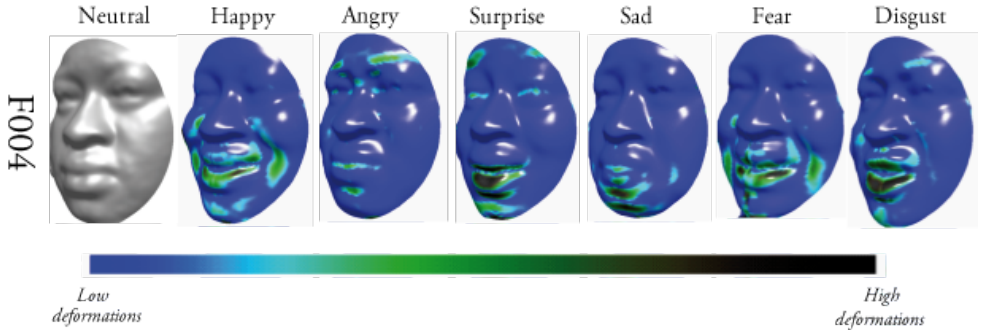

Figure 4.2 – *Deformation maps computed using a neutral face, considered as a source face, and six expressive faces, where each face, considered as a target face, shows one of the six basic emotions.*

In Fig. 4.3.a, we illustrate the face conveying happy expression with extracted radial curves. Fig. 4.3.b and Fig. 4.3.c illustrates two correspondent radial curves on neutral and happy faces respectively. These curves are reported together in Fig. 4.3.d, one can easily see the deformation between them although they lie at the same angle and belong to the same person. The amount of the deformation between the two curves is calculated using Eq. (4.4) and the plot of the module of this vector at each point of the curve is reported in Fig. 4.3.e.

4.2.3 Motion Features Extraction

We aim at learning facial deformations due to expression across a 3D video sequence. Such deformations are known to be characterized by subtle variations induced mainly by facial feature points motions. These subtle changes are important and present difficult issues for analyzing facial motions. The comparison between two faces is obtained by a pairwise com-

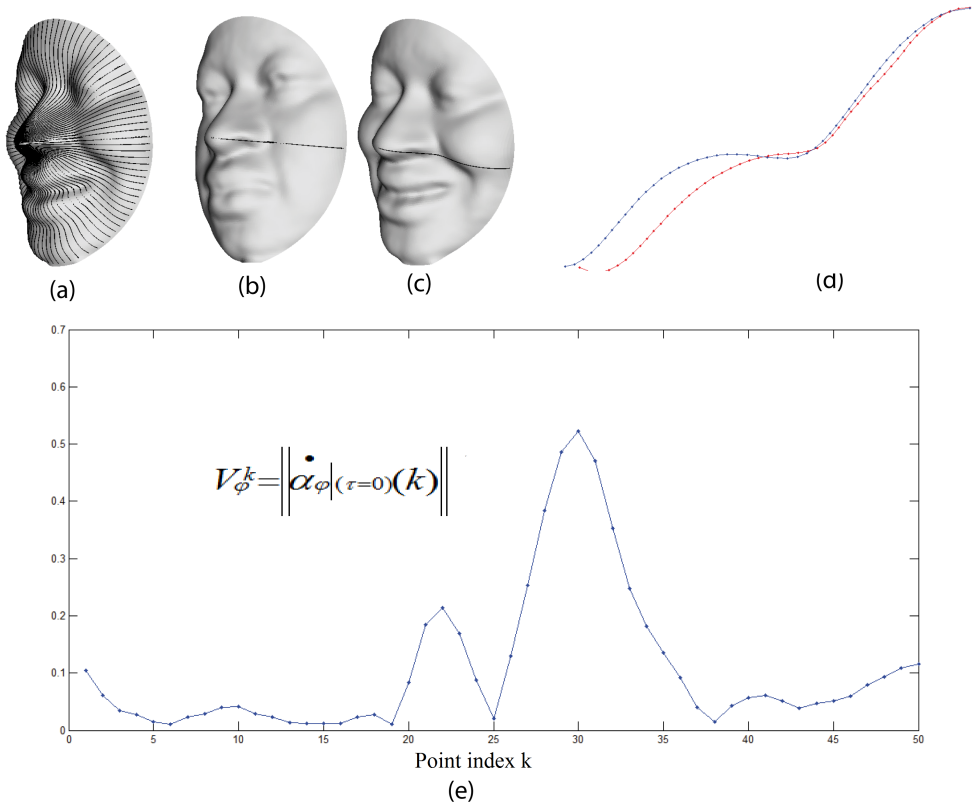

Figure 4.3 – *(a) Extraction of radial curves, (b) a radial curve extracted from neutral face, (c) the correspondent radial curve extracted from the same face but showing happy expression,(d) the two curves are reported together and (e) a plot of the trade-off between points on the curve and values of the magnitude of* $\frac{d\alpha^*}{d\tau}|_{\tau=0}(k)$.

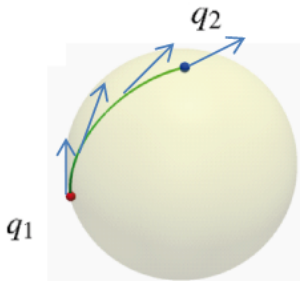

Figure 4.4 – *Illustration of the parallel vector field across the geodesic between* q_1 *and* q_2 *in the space of curves* $\mathcal{C}$.

parisons of correspondent curves. Based on this comparison, we calculate a scalar field V_α^k that can be calculated between two given 3D faces. Based on this measure, we are able to quantify the motions of feature points and thus capture the changes in facial surface geometry. Fig. 4.5 illustrates a direct application of the V_α^k and its effectiveness in capturing deformation

from one facial surface to another belonging to two consecutive frames in a video sequence. This figure shows two subsequences extracted from videos in the BU-4DFE database. In the left, we have both texture and 3D range images selected from a video captured for a female subject while conveying a happy expression. In the right, we illustrate the scalar field V_α^k computed between consecutive frames (i.e., the current frame and the previous one) in the subsequence. This scalar field is interpreted by applying an automatic labeling scheme that includes only two colors, the red color is associated with high V_α^k values and corresponds to facial regions affected by high deformations, while the blue color is associated with regions that remain stable from one frame to another. The remaining rows of the figure illustrate the same visual information, for the same subject, for the surprise expression.

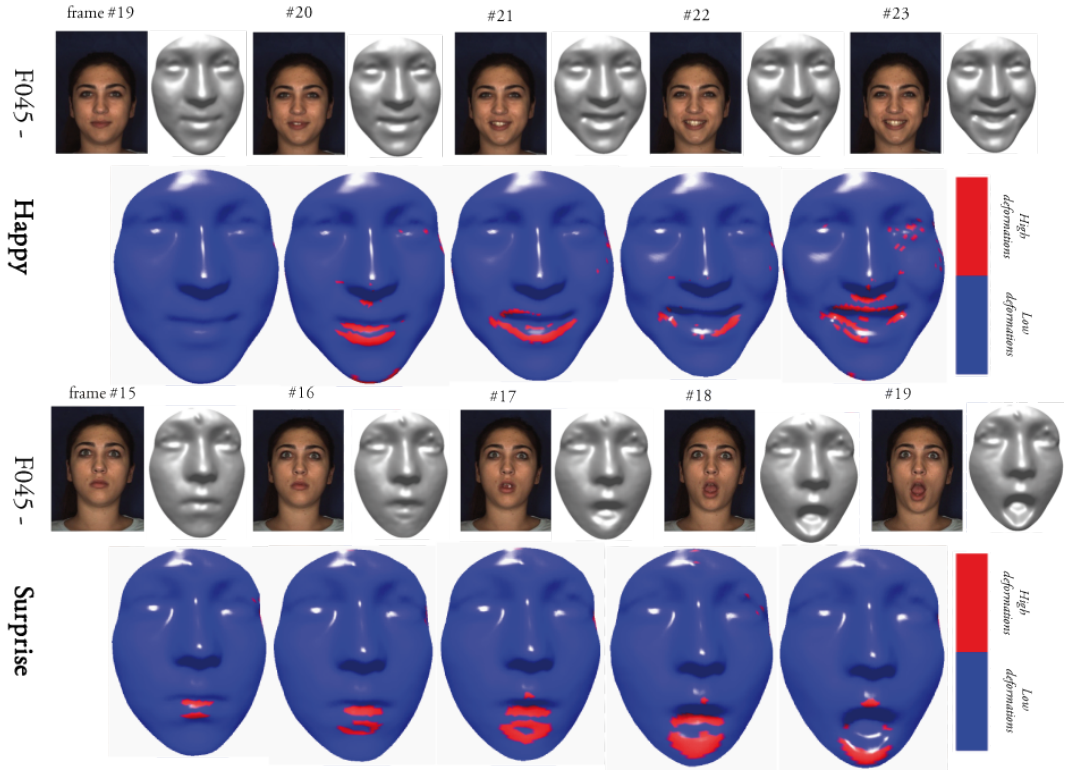

Figure 4.5 – *Examples of motion extraction from 3D video sequences computed by the proposed method. The first example illustrates motion capture of Happy expression whereas the second example gives deformations arising from surprise.*

The feature extraction process begins with a first step which characterizes each 3D range frame by a scalar field computed regarding the

previous one in a given video sequence. We obtain as many fields as the number of frames in the sequence, and where each field contains as many scalar values as the number of points composing the collection of radial curves representing the facial surface. In practice, the size of V_α^i is 1×5000 considering 5000 points on the face. The second step consists of applying a common technique of dimensionality reduction, which is Linear Discriminant Analysis (LDA), in order to reduce the dimension of our feature space. We use LDA to project the data onto a lower dimensional feature space. This step allows to enhance the discriminative power of our feature relatively to the six prototypic expressions, and transforms the n-dimensional feature space to an optimal d-dimensional space, in our approach we deal with six expressions so n is reduced to $d = 5$. The individual 5-dimensional feature vector extracted for the 3D frame at instant t of a sequence is indicate as f^t in the following. Once extracted, the feature vectors are used to train HMMs and to learn deformations due to expression along a temporal sequence of frames.

4.3 Expression Classification based on HMMs

Let $\lambda = \{A, B, \pi\}$ denote an HMM to be trained and N be the number of hidden states in the model. We indicate the states as $S = \{S_1, S_2, ..., S_N\}$, and the state at instant time t is q_t. The state transition probability distribution is indicated as $A = \{a_{ij}\}$, where $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$, with $1 \leq i, j \leq N$. In a discrete domain, states of the model can emit symbols from a discrete alphabet derived from the physical output of the system being modeled. In the case the inputs of the model are multidimensional feature vectors with values in a continuous domain, the alphabet of symbols can be derived using clustering techniques. In this way, feature vectors are clustered around a number of cluster centers that are used as *codebook* of the input vectors. The individual symbols are indicated as $V = \{v_1, v_2, \ldots, v_M\}$, being M the number of distinct cluster centers. Given an observation v_k, $B = \{b_j(k)\} = P(v_k \ at \ t | q_t = S_j)$ is the observation probability distribution in state j, that is the probability that the

observation k being produced from state j, independent of t. Finally, with $\pi = \{\pi_i\}$ is denoted the initial probability array, being $\pi_i = P(q_1 = S_i)$.

In our case, sequences of 3D frames constitute the temporal dynamics to be classified, and each prototypical expression is modeled by an HMM (a total of 6 HMMs λ^{expr} is required, with $expr \in \{an, di, fe, ha, sa, su\}$). Four states per HMM (N=4) are used to represent the temporal behavior of each expression. This corresponds to the idea that each sequence starts and ends with a neutral expression (state S_1); The frames that belong to the temporal intervals where the face changes from neutral to expressive and back from expressive to neutral are modeled by the *onset* (S_2) and *offset* (S_4) states, respectively; Finally, the frames corresponding to the highest intensity of the expression are captured by the apex state (S_3). Fig. 4.6 exemplifies the structure of the HMMs in our framework.

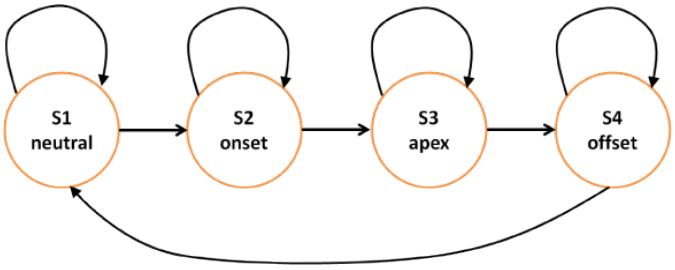

Figure 4.6 – *Structure of the HMMs modeling a 3D facial sequence. The four states model, respectively, the neutral, onset, apex and offset frames of the sequence. As shown, from each state it is possible to remain in the state itself or move to the next one (Bakis or left-right HMM).*

The training procedure of each HMM is summarized as follows:

- The *codebook* is first constructed by clustering the feature vectors into 32 symbols, using the standard LBG algorithm;
- Observation sequences $O = \{O_1, O_2, \cdots, O_T\}$, are derived from the 3D expression sequences, where each O_t denotes an observation at time t expressed by the feature vector f^t;
- The HMM λ^{expr} is initialized with random values. The *Baum-Welch* algorithm [Rab89] is used to perform learning from a set of training sequences, thus estimating the model parameter $\lambda^{expr} = \{A, B, \pi\}$ when $P(O|\lambda^{expr})$ is maximized.

Given a 3D sequence to be classified, it is processed as in Sect. 4.2, so that each feature vectors f^t corresponds to a *query* observation $O = \{O^1 \equiv f^1, \ldots, O^T \equiv f^T\}$. Then, the query observation O is presented to the six HMMs λ^{expr} that model different expressions, and the *Viterbi* algorithm is used to determine the best *path* $\bar{Q} = \{\bar{q}^1, \ldots, \bar{q}^T\}$, which corresponds to the state sequence giving a maximum of likelihood to the observation sequence O. The likelihood along the best path, $p^{expr}(O, \bar{Q}|\lambda^{expr}) = \bar{p}^{expr}(O|\lambda^{expr})$, is considered as a good approximation of the true likelihood given by the more expensive *forward* procedure [Rab89], where all the possible paths are considered instead of the best one. Finally, the sequence is classified as belonging to the class corresponding to the HMM whose log-likelihood along the best paths is the greatest one.

4.4 Experimental Results

The proposed framework for facial expression recognition from dynamic sequences of 3D face scans has been experimented using the BU-4DFE database. Main characteristics of the database and results are reported in the following.

4.4.1 The BU-4DFE Database

To investigate the dynamics of 3-D model sequences with varied facial expressions, we employed a dynamic 3D facial expression database that has been created at *Binghamton University* [YCS*08]. The 3D scans have been constructed by capturing a sequence of stereo images of subjects exhibiting facial expressions and producing a depth map for each frame according to a passive stereo-photogrammetry approach. The range maps are then combined to produce a temporally varying sequence of 3D scan. Subjects were requested to perform the six prototypic expressions separately, in such a way that each expression sequence contains neutral expressions in the beginning and the end. In particular, each expression was performed gradually from neutral appearance, low intensity, high intensity, and back to low intensity and neutral. Each 3D sequence captures one expression at a rate of 25 frames per second and each 3D sequence lasts approximately

4 seconds with about 35,000 vertices per scan (i.e., 3D *frame*). The database consists of 101 subjects (58 female and 43 male) including 606 3D model sequences with 6 prototypic expressions and a variety of ethnic/racial ancestries (i.e., 28 Asian, 8 African-American, 3 Hispanic/Latino, and 62 Caucasian). Examples of 3D frames sampled from the *happy* and *surprise* 4D sequences of subject F045 are given in Fig. 4.5.

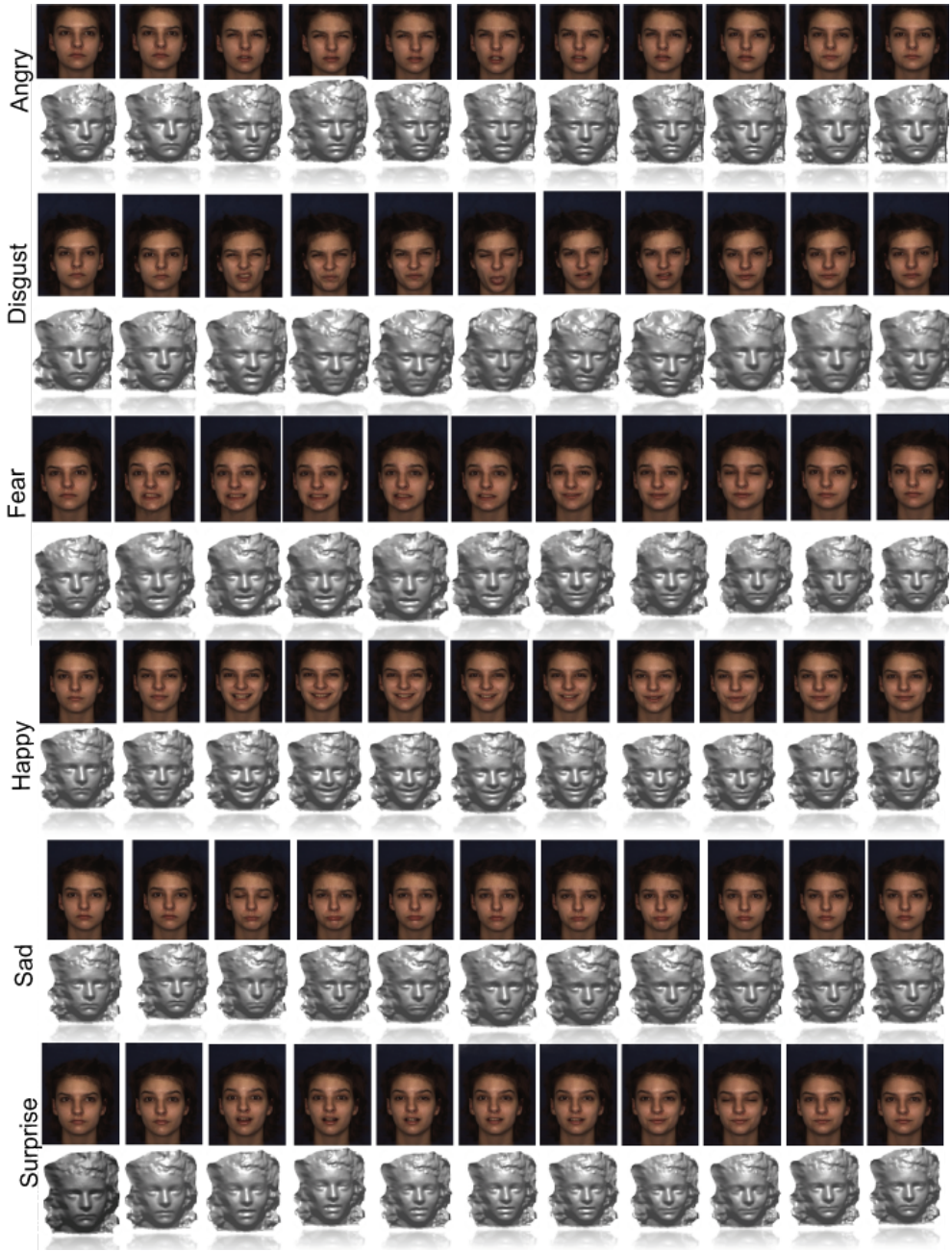

Figure 4.7 – *Video samples displayed in the format of 2D textured images and 3D shape models. A female subject from top to bottom exhibits the expressions of angry , disgust, fear, happy, sad, and surprise, respectively.*

4.4.2 Data Preprocessing

The raw data obtained from even the most accurate 3D scanners is far from being perfect and clean, as it contains spikes, holes and significant noise. A preprocessing step must be applied to remove these anomalies before any further operations can be performed. Thus the preprocessing is an important stage of the recognition systems, especially when knowing that all the features will be extracted from the output of this step. An automatic preprocessing pipeline is developed and is set to apply different tools and follow multiple steps shown in Fig. 4.8 and enumerated below:

1. filling holes: a number of the BU-4DFE scans are affected with holes that often lie in the mouth region and that take part in an acquisition session that meets an open mouth expression. In this case the mouth area is not visible and cannot be acquired by the stereo sensors which causes missing data. A linear interpolation technique is used to fill the missing regions of a given raw 3D face image,

2. nose detection: the nose tip is a key point that is needed for both preprocessing and facial surface representation stages. Several approaches have been proposed to detect this key point [DGA09, YLW09, PB11]. Knowing that in most 3D images, the nose is the closest part of the face to the 3D acquisition system, in this step the nose tip is detected using horizontal and vertical slicing of the facial scan in order to search for the maximum value of the z-coordinate along these curved profiles. Once this is done for the first frame, for the remaining frames of the sequence this detection technique is refined and the search area in a current frame is reduced to a small sphere centered on the nose tip detected in the previous frame.

3. cropping: face boundaries, hair and shoulders are irrelevant parts for our study, and they are usually affected with outliers and spikes. We use the nose tip detected in the previous step to crop out the required facial area from the raw face image. Using a sphere, centered on the nose tip and of a radius determined empirically, we cut the

3D range model and we retain the mesh structure kept inside the sphere.

4. Pose correction: in this step we apply a global registration technique (i.e., ICP) to align meshes of the current frame and the first frame of the sequence. After this rigid alignment enable to adjust the pose of the 3D face and make it similar enough to the first frame pose.

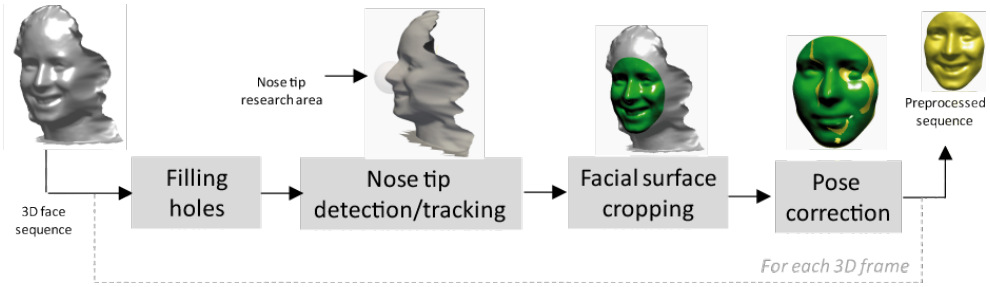

Figure 4.8 – *Preprocessing pipeline.*

4.4.3 Expression Classification Performance

The above preprocessing step must be applied to 3D face data before applying a radial curve-based representation and expression deformation capture operations can be reliably performed. After these preprocessing operations, data of 60 subjects have been randomly selected to perform recognition experiments, whereas the remaining 41 subjects have been used for a preliminary tuning of the proposed algorithms. The 60 subjects are randomly partitioned into 10 sets, each containing 6 subjects, and 10-fold cross validation has been used for test, where at each round 9 of the 10 folds (54 subjects) are used for training while the remaining (6 subjects) are used for test. The recognition results of 10 rounds are then averaged to give a statistically significant performance measure of the proposed solution.

The proposed approach is able to correctly classify all the sequences with an accuracy of 100%. Indeed, the classification model is capable to correctly identify 3D dynamic expression sequences. This provides a measure of the overall capability to classify 3D frames sequences composed of around hundred frames and with a typical behavioral pattern.

In some contexts the classification of individual 3D frames is also relevant in that can permit an online analysis of a 3D video. Following the experimental protocol proposed in [SY08], this is obtained by the definition of a large set of very short subsequences extracted using a sliding window on the original expression sequences. The subsequences have been defined in [SY08] with a length of 6 frames with a sliding step of one frame from one subsequence to the following one. For example, with this approach, a sequence of 100 frames originates a set of $6 \times 95 = 570$ subsequences, each subsequences differing from one frame from the previous one. This accounts for the fact that, in general, the subjects can come into the system not necessarily starting with a neutral expression, but with a generic expression. Classification of these very short sequences is regarded as an indication of the capability of the expression recognition framework to identify individual expressions. According to this, for this experiment we retrained the HMMs on 6 frame subsequences constructed as discussed above. The 4-state structure of the HMMs still resulted adequate to model subsequences. Also in this experiment, we performed 10-folds cross validation, on the overall number of subsequences derived from the 60×6 sequences (31970 in total).

The results obtained by classifying individual 6-frames subsequences of the expression sequences (*frame-by-frame* experiment) are reported in the confusion matrix of Tab. 4.1. Values in the table have been obtained by using 6-frames subsequences as input to the 6 HMMs and using the maximum emission probability criterion as decision rule. It is evident that the proposed approach is capable to accurately classify very short sequences containing very different 3D frames, with an average accuracy of 93.83%. It can be noted that the lower recognition is obtained for the *disgust* expression (91.54%) which is mainly confused with the *angry* and *fear* expression. Interestingly, these three expressions capture negative emotive states of the subjects, so that similar facial muscles can be activated.

	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	**93.95**	1.44	1.79	0.28	2.0	0.54
Disgust	3.10	**91.54**	3.40	0.54	1.27	0.15
Fear	1.05	1.42	**94.55**	0.69	1.67	0.62
Happy	0.51	0.93	1.65	**94.58**	1.93	0.40
Sad	1.77	0.48	1.99	0.32	**94.84**	0.60
Surprise	0.57	0.38	3.25	0.38	1.85	**93.57**

Table 4.1 – *Average confusion matrix for 6-frames subsequences (percentage values).*

4.4.4 Discussion and Comparative Evaluation

To the best of our knowledge, the only three works reporting results on expression recognition from dynamic sequences of 3D scans are [SY08], [SZPR12] and [LTH11]. These works have been verified on the BU-4DFE dataset, but the testing protocols used in the experiments are quite different, so that a direct comparison of the results reported in these works is not possible.

The approach in [SY08] is not completely automatic and also presents high computational cost. In fact, a generic model (i.e., tracking model) is adapted to each depth model of a 3D sequence. The adaptation is controlled by a set of 83 pre-defined keypoints that are manually identified and tracked in 2D. The person-independent expression recognition experiments were performed on 60 selected subjects out of the 101 subjects of the BU-4DFE database, by generating a set of 6-frame subsequences from each expression sequence to construct the training and testing sets. The process were repeated by shifting the starting index of the subsequence every one frame till the end of the sequence. The rationale used by the authors for this shifting was that a subject could come to the recognition system at any time, thus requiring the recognition process could start from any frame. Following a 10-fold cross-validation, an average recognition rates of 90.44% was reported. So, it results that expression recognition results are actually provided not on variable length sequences of 3D depth

frames, but just on very short sequences with a predefined length of 6 frames.

The method proposed in [SZPR12] is fully automatic with respect to the processing of facial frames in the temporal sequences, but uses *supervised* learning to train a set of HMMs. Though performed offline, supervised learning requires manual annotation and counting on a consistent number of training sequences that can be a time consuming operation. In addition, a drawback of this solution is the computational cost due to Free-Form Deformations based on B-spline interpolation between a lattice of control points for nonrigid registration and motion capturing between frames. This hinders the possibility of the method to adhere to a real time protocol of use. Preliminary tests were reported on three expressions: *angry*, *happiness* and *surprise*. Authors motivated the choice of the happiness and anger expressions with the fact that they are at either ends of the valence expression spectrum, whereas surprise was also chosen as it is at one extreme of the arousal expression spectrum. However, these experiments were carried out on a subset of subjects accurately selected as acting out the required expression. Verification of the classification system was performed using a 10-fold cross-validation testing. On this subset of expressions and subjects, an average expression recognition rate of 81.93% is reported.

In [LTH11] an automatic method is also proposed, that uses an *unsupervised* learning solution to train a set of HMMs. In this solution, preprocessing is very important in that an accurate alignment of the 3D mesh of each frame is required in order to extract the level set curves that are used for face representation. This increases the computational cost of the approach making questionable its use where a real time constraint is required. Expression recognition is performed on 60 subjects from the BU-4DFE database for the expressions of *happiness*, *sadness* and *surprise*. Results of 10-fold cross-validation show an overall recognition accuracy of 92.22%, with the highest performance of 95% obtained for the happiness expression.

Tab. 4.2 summarizes the results scored by the above methods com-

pared to those presented in our work. Considering the classification of the entire 4D sequences, our solution clearly outperforms those in [SZPR12] and [LTH11], even working on six expressions instead of just three, evidencing the capability of the proposed face representation to capture salient features to discriminate between different expressions. With respect to the *frame-by-frame* classification experiment, our results are more than 3% better than those in [SY08], with the advantage of using a completely automatic approach and a simpler classification model using temporal HMMs with fewer states.

Average RR	[SY08]*(T-HMM)*	[SY08]*(R2D-HMM)*	[SZPR12][1]	[LTH11][2]	**This work**
Frame-by-frame	80.04 %	90.44 %	73.61 %	-	**93.83 %**
Sequence	-	-	81.93 %	92.22 %	**100 %**

1 Performances provided for *happiness*, *anger* and *surprise* expressions.

2 Performances provides for *happiness*, *sadness* and *surprise* expressions.

Table 4.2 – *Comparison to earlier work.*

4.5 Conclusion

In this chapter, we presented an automatic approach for identity-independent facial expression recognition from 3D video sequences. Through a facial shapes representation by collections of radial curves, a Riemannian shape analysis framework was applied to quantify dense deformations and extract motion from successive 3D frames. Then such dynamic description was trained using HMM after LDA-based feature space transformation. Experiments conducted on the BU-4DFE dataset following state-of-the-art settings show the effectiveness of the proposed approach which outperforms earlier work. A limitation of the approach is the nose tip detection in case of non frontal views and/or occlusion. The BU-4D contains frontal 3D faces without occlusion, however, in realistic scenario, more elaborated techniques should be applied to detect the nose tip.

Conclusion

5

5.1 Summary

In this thesis, we focused on developing a new approach for 3D facial expression recognition. For that matter, we have employed a Riemannian framework to analyze shapes of facial surfaces. This framework has allowed the construction of geodesic paths between arbitrary two surfaces taken under specific representation. The length of the geodesic path between any two surfaces quantifies the difference between them in term of shape, and permits to measure a score of similarity and/or dissimilarity. The optimal deformation from one surface to another through illustration of the geodesic path that can be reconstructed after manipulating calculus and statistics on the shape manifold. There are multiple applications of this geodesic construction. Using this construction we have been able to address the problem of facial expression recognition in two different ways. The first application addresses static 3D facial expression recognition (FER) problem, while the second tackles dynamic 3D (or 4D) FER. For facial expression recognition from 3D static images, we proposed to conduct shape analysis of local regions of 3D facial surface, namely patches. These patches were extracted using level curve-based representation centered on fiducial landmarks. The shape information of each patch is acquired, and shape analysis is derived resulting in quantization of this information using the Reimannian framework. A pairwise comparison were conducted to compute similarity measures between corresponding patches lying on facial surfaces regardless to the identity or the expression of the 3D face models. These measures, obtained after calculation

of geodesic distances , were then employed to build feature vector for describing expressions. These features were considered as entry parameters to several classification algorithms in order to derive facial expression classification. The static BU-3DFE database were used to conduct exhaustive experiments in order validate our approach and to test its robustness. The experimental results demonstrate the effectiveness of the proposed method, and applying machine learning techniques for expression classification, we obtained an average recognition rate that reached 98.81%.
In this approach we presented a novel representation of the 3D face using patches as curve-based local surfaces of the face, and we applied a Reimannian framework for the shape analysis of the extracted patches. Our features for describing the shape of the considered patches are then obtained and are used to feed a set of classifiers to recognize expressions. So in this approach we proposed to couple both shape analysis and machine learning disciplines for addressing 3D facial expression recognition. However, there are several limitations related to the proposed approach and that can be enumerated as follows:

1. this approach is not a single based 3D facial model approach. To conduct our study of shape analysis, we needed for some referential model in order to quantify the shape information. Thus our approach relies on a reference model. In our study, we employed six different subjects, we picked up the six expressive scans corresponding to the six expressions, and each of the expressive model played the role of referential models of that expression class and has been used to compute the geodesic distances between corresponding patches of a gallery face model and the reference model.

2. in this approach we conducted experiments on a set of the BU-3DFE database composed of 60 subjects chosen arbitrarily. We did not take into consideration the whole database, the reason for this is that we wanted to compare our result with the the state-of-the-art one, where this set has been considered for the experiments related to these approaches.

3. in this approach, only two out of four intensity levels of the six ex-

pressions were explored to conduct experiments, the high and the highest intensity levels. These las two levels refer to exaggerated and well-acted expressions conveyed by the subjects. So in our experiments we did not evaluate the performance of our approach in the case of real life facial expressions, and to what extent our approach is able to recognize subtle expressions.

4. this approach is a semi-automatic approach, because it relies on a large set of well defined landmarks. These landmarks were manually annotated by the creators of the BU-3DFE database and were defined based on the facial anatomy structure of the face and by suggestion from a psychologist.

5. the computational cost of our approach is expensive, this is mainly due to both major steps: patches extraction (curve-based surface representation) and the geodesic distances computation (feature calculation). The patches extraction step consists of the representation of local facial surfaces by of a set of level curves, theses curves need to be preprocessed to obtain smooth curves with a fixed number of points for each curve in order to conduct our experiments. As for the computation of the geodesic distances, it is based on an energy minimization problem using gradient descent algorithm (path straightening method) which is a time consuming iterative method.

In a second stage we proposed an approach for recognizing the six prototypic expressions, but this time, from sequence of 3D frames. The goal was to learn facial deformations due to expressions across a 3D video sequence. Such deformations are known to be characterized by subtle variations induced mainly by facial feature points motions. Unlike the proposed approach for static 3D FER, where we defined multiple patches through a closed curves representation for 3D surface representation, we proposed for 4D FER approach to capture the shape information by extracting a set of parametrized radial curves for facial surface representation. Only one point of interest were needed to define this representation, It is the nose tip which we detected automatically through the frame sequence. This nose tip defines the starting point of each curve and is

also one among the control points used faces alignment from video sequences. Since facial surfaces were represented by parametrized curves, we were able to compute the magnitude of deformations resulting from the geodesic path between two faces in consecutive frames of a sequence. Illustration of scalar fields, associated with the magnitude of deformations, provided important visual information about variability, due to expressions, of facial shapes. These scalar fields are used to play the role of features and to be considered as entry parameters to machine learning techniques in order to learn the deformations resulting from the six prototypic expressions. For this end Hidden markov models (HMMs) are applied train the dynamics of the extracted features, six HMMs corresponding to the studied expressions were required. The states of each HMM are modeled regarding the expression behavior along the sequence, that starts from neutral expression, evolves to go through the beginning of the expression (onset) and to reach the maximum (apex), then it turns back to go through the end of it (offset), and terminates with the neutral expression at the end of the sequence. Experiments of our approach are conducted on a set of the BU-4DFE database and the average recognition rate attained 93.83%.

The approach we proposed to address 4D FER comes with a novel method of extracting motion features related to dynamic behavior of expressions. This method is based on differential geometry applied on a the manifold of the set of square root velocity representations of curves (radial curves). This motion feature is illustrated by a scalar field (SF) that can be mapped on each frame of an expressive video sequence, resulting in a color map that shows color variations (from blue to red) associated with deformations (from low to high deformations) due to expressions. The advantage of the developed approach, is that under the new characterization of the facial expression dynamics, the deformation feature denoted by SF, we were able to alleviate some of the limitations that affected our approach for 3D FER. In this approach there was no need for a reference model, there was no need, also, for a large set of well defined landmarks, we needed just to detect the nose tip to be the starting point of the radial

curves representing the facial surface. With defining the SF feature, we were able to capture deformation due to expression between two consecutive frames of a video sequences. Since that a video sequence starts from neutral expression and ends with a neutral, with a rate of 25 frames per second, the SF feature permitted to model quite subtle facial deformations.

5.2 Future work

Further work can be done in order to enhance our approaches developed for recognizing facial expressions. For the 3D FER, we can conduct further experiments in order to evaluate the performance of our approach on the whole BU-3DFE database, and on other publicly available and challenging benchmarks such as Bosphorus and Hi4D-ADSIP databases. In our approach we only considered data related to well-acted expressions (high and highest intensity), lower intensity level expressions (low and middle intensity) need to be taken into consideration in order to test to what extent our approach is capable of recognizing more spontaneous facial expressions. Further study can be conducted to define a smaller set of landmark points that could possibly carry relevant information, and that can be considered to be utilized rather than exploring a large number of landmarks and a large set of associated patches that may contain patches with no relevant information.

For the 4D FER, we can further investigate the ability of the SF feature to capture subtle deformations due to expressions with experimenting more dense sequence frames that exceeds the rate of 25 frames per second. Furthermore, in our 4D FER approach we only studied temporal motion of a set of facial points to infer the deformation of a facial surface due to expression variation. Other interesting information can be extracted such as spatial information, curvature and texture information. The spatial information could be the capture of spatial displacement of facial points lying on some iso-curves, and the illustration of smooth changes of their values as their positions change.

Face expression recognition systems have improved a lot over the past decade. The focus has definitely shifted from posed expression recog-

nition to spontaneous expression recognition. Since the face expression research community is paying more interest to the recognition of spontaneous expressions. We can further develop and enhance our 4D facial expression recognition approach in order to construct a system useful for real-world application. Such system can integrate additional capabilities such as face detection, pose-estimation and fiducial points detection. Thus resulting in fully automatic facial expression analyzer. Our approach can be optimized and well designed to be integrated in some low-cost 3D sensors like Kinect for real time FER application. All these issues and progressive enhancement of our facial expression recognition approach can be considered in a future work.

Annexes

A

A.1 Theoretical Background

We present some theoretical Background that enabled us to develop our approaches for facial expressions recognition problem. It is based on a Riemannian framework that treats differential geometry allowing to derive shape analysis. Let M be a manifold and also a submanifold of a Hilbert space, with the Riemannian structure inherited from that larger space. We will denote this large ambient space by V. The formal problem of finding geodesics on M is posed as follows: Say we are given two points p_1 and p_2 in M that we want to join using a geodesic path in M. Let $\mathcal{H}$ be the set of all differentiable paths in M, whose first derivatives are $\mathbb{L}^2$ functions, parameterized by $t \in [0,1]$:

$$\mathcal{H} = \alpha : [0,1] \longrightarrow M,$$

and $\mathcal{H}_0$ be the subset of $\mathcal{H}$ consisting of those paths that start at p_0 and end at p_1:

$$\mathcal{H}_0 = \{\alpha \in \mathcal{H} | \alpha(0) = p_1 and \alpha(1) = p_2\}$$

The desired geodesic is an element of $\mathcal{H}_0$. For elements of $\mathcal{H}$, define an energy function $E : \mathcal{H} \longrightarrow \mathbb{R}_+$ by:

$$E[\alpha] = \frac{1}{2} \int_0^1 \langle \dot{\alpha}(t), \dot{\alpha}(t) \rangle \, dt$$

Some remarks about this definition of E:

- Note that for each t, $\dot{\alpha}(t)$ is an element of $T_{\alpha(t)}(M)$ and, by the assumption on $\mathcal{H}$, $\dot{\alpha} : [0,1] \longrightarrow TM$ is an $\mathbb{L}^2$ function. The inner product appearing inside the integral sign comes, of course, from the Riemannian metric on M.

- E is not the length of the path α although it is closely related. If we use the square root of the integrand (and remove the $\frac{1}{2}$ factor), we obtain the length of the path α. Thus, E is $\frac{1}{2}$ the integral of the square of the instantaneous speed along the curve (the integral is taken with respect to the parameter t.

- the critical points of E on the space $\mathcal{H}_0$ are precisely the geodesic paths on M between p_1 and p_2. Therefore, one way to find a geodesic

is to use the gradients of E to reach its critical points. This is the method we will describe.

We will be using the gradients of E to find its critical points on $\mathcal{H}_0$. It is needed to start with the differential structure of $\mathcal{H}_0$. The tangent spaces of $\mathcal{H}$ and $\mathcal{H}_0$ are:

$$T_\alpha(\mathcal{H}) = \left\{ w : [0,1] \longrightarrow TM | \forall t \in [0,1], w(t) \in T_{\alpha(t)}(M) \right\}$$

where $T_{\alpha(t)}(M)$ is the tangent space of M at the point $\alpha(t) \in M$, and

$$T_\alpha(\mathcal{H}_0) = \{ w \in T_\alpha(\mathcal{H}) | w(0) = w(1) = 0 \}$$

We note that the tangent space element w is a vector field along the path α tangent to M at each point of α.

A.1.1 Definition (Orbit)

Assume that a group G acts on a manifold M. For any $p \in M$, the orbit of p under the action of G is defined as the set $G \cdot p = g \cdot p : g \in G$. We will also denote it by $[p]$. If the orbit of any $p \in M$ is the whole of M, then the group action is said to betransitive. The orbit of a point in M refers to all possible points one can reach in M using the action of G on that point. The orbit of a point can vary in size from a single point to the entire manifold M (which is the case if the action is transitive).

A.1.2 metrics

We start the study of the differential geometry of a space M, M becomes a Riemannian manifold if we impose an inner-product on its tangent spaces. There are several possibilities for the metric:

$\mathbb{L}^2$ **Metric:** One choice is the $\mathbb{L}^2$, for any $v_1, v_2 \in T_m(M)$ tangent space at the element m of M, define the $\mathbb{L}^2$ inner-product:

$$< v_1, v_2 >= \int_{S^1} v_1(x) v_2(x) dx$$

Fisher-Rao Metric: for any $v_1, v_2 \in T_m(M)$ tangent space at the element m of M, the Fisher-Rao (FR) metric is defined by:

$$< v_1, v_2 >_{FR}= \int_{S^1} v_1(x) v_2(x) \frac{1}{\dot{m}(x)} dx$$

Palais Metric: This metric is given by:

$$< v_1, v_2 >_{pal} = v_1(0)v_2(1) + \int_{S^1} \dot{v}_1(x)\dot{v}_2(x)dx$$

A.2 Classifiers

Throughout this thesis, we have been dealing with facial expression classification problem. And we have been led, naturally, to apply several machine learning methods and classifiers. In this appendix we will briefly describe the employed techniques. We begin by the Boosting, which is one type of meta learning techniques that try to build a strong learning algorithm based on a group of weak classifiers,where weak classifier comes from Valiant's framework [Val84]. The boosting method was proposed by Schapire [Sch90] and later improved by Freund [Fre90]. Since then, boosting has been explored by many researchers theoretically and empirically. The most popular algorithm AdaBoost, which was introduced by Freund and Schapire in 1995 [FS95], has successfully solved many practical problems of previous boosting approaches. AdaBoost is also extended to multi-class classification problems and regression problems in [FS95] and [Sch99]. Many experiments have been carried out using AdaBoost and its variants (refer to [Sch02]), including OCR, text filtering, image retrieval, medical diagnosis, etc. In our work, we applied the idea of boosting on 3D Facial Expression Recognition.

A.2.1 AdaBoost

AdaBoost is a very successful machine-learning method that permits to build an accurate prediction rule, its principle is based on finding many rough rules of thumb instead of finding a one highly accurate rule. More simpler, the idea is to build a strong classifier by combining weaker ones. AdaBoost is proven to be an effective and powerful classifier in the category of ensemble techniques. The algorithm takes as input a training examples $(x_1, y_1), .., (x_N, y_N)$ where each x_i $(i = 1, ..., N)$ is an example that belongs to some domain or instance space X, and each label y_i is

a boolean value that belongs to the domain $Y = \{-1, +1\}$, indicating whether x_n is positive or negative example. Along a finite number of iterations $t = 1, 2, ..., T$ the algorithm calls, at each iteration t, the weak classifier (or learner). After T times it generates a set of hypothesis $\{h_t\}_{t=1}^{T}$ such that $h_t \longrightarrow \{-1, 1\}$. The final classifier $H(X)$ is the strongest one, and is given by the combination of these hypothesis, ponderated by their respective weight factors $\{\alpha_t\}_{t=1}^{T}$. The hypothesis h_t and its corresponding weight α_t are determined at each iteration t, the selection of the best hypothesis h_t, at each time t, is done among a set of hypothesis $\{h_j\}_{j=1}^{J}$, where J stands for the number of features considered for the classification task. h_t is equal to h_j that gives the smallest error of classification ϵ_j. The error ϵ_j corresponds to samples that are misclassified, and that will see their associated weight increased in the next iteration $t + 1$. These procedures are presented in Algorithm 1.

Algorithm 2 AdaBoost algorithm.

- Input: set of examples $(x_1, y_1), .., (x_N, y_N)$ where $x_i \in X$ and $y_i = \{-1, +1\}$.
- Let m be the number of negatives examples and l be the number of positive examples. Initialize weights $w_{1,n} = \frac{1}{2m}, \frac{1}{2l}$ depending on the value of y_n.
- For $t = 1, ..., T$x:

  1- Normalize the weights $w_{t,n}$ so that $\sum_{n=1}^{N} w_{t,n} = 1$.

  2- For each feature f_j, train a weak classifier h_j.

  3- The error ϵ_j of a classifier h_j is determined with respect to the weights $w_{t,1}, ..., w_{t,N}$:

  $$\epsilon_j = \sum_{n}^{N} w_{t,n} |h_j(x_n) - y_n|$$

  4- Choose the classifier h_j with the lowest error $\epsilon_j$and set $(h_t, \epsilon_t) = (h_j, \epsilon_j)$.

  5- Update the weights $w_{t+1,n} = w_{t,n}\beta_t^{1-e_n}$, where $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$ and $e_n = 0$, if example x_n is classified correctly by h_t and 1 otherwise

- The final strong classifier is given by:

$$H(x) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} log\frac{1}{\beta_t} h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} log(\frac{1}{\beta_t}); \\ 0 & \text{otherwise.} \end{cases}$$

A.2.2 MultiBoost

Given a set of training data $(x_1, c_1), \ldots, (x_n, c_n)$ where x_i is the input, and each output $c_i \in 1, \ldots, K$ is a class label. We use K to denote the number of possible class labels. Using training data, Multiboost, which is an extension of the original AdaBoost method, permits to find a classification rule so that when given a new input x, we can assign it a class label c from $1, \ldots, K$. Let $T(x)$ denote a weak multi-class classifier that assigns a class label to x. Then the Multiboost algorithm, called also AdaBoostM1, proceeds as the following:

Algorithm 3 AdaBoostM1 algorithm.

- Initialize the observation weights $w_i = \frac{1}{n}, i = 1, 2, \ldots, n$.
- For $m = 1$ to M:
  - Fit a classifier $T_m(x)$ to the training data using weights w_i
  - Compute $err_m = \frac{\sum_{i=1} nw_i \prod(c_i \neq T_m(x_i))}{\sum_{i=1} nw_i}$.
  - Compute $\alpha_m = \log(\frac{1-err_m}{err_m})$
  - Set $w_i \leftarrow w_i \cdot \exp(\alpha_m \cdot \prod(c_i \neq T_m(x_i)))$, $i = 1, \ldots, n$.
  - Re-normalize w_i.
- Output$C(x) = \arg\max_k \sum_{m=1} M\alpha_m \prod(T_m(x) = K)$.

MultiBoost with LDA classifier

MultiBoost with LDA classifier incorporates linear discriminant analysis (LDA) algorithm to implement linear combinations between selected features and generate new combined features. The combined features are used along with the original features in boosting algorithm for improving classification performance. Given a binary classification problem with linear classifiers which are specified by discriminant functions. LDA assumes the covariance matrices of both classes to be equal, $\sum$. We denote

the means by μ_1 and μ_2, and arbitrary feature vector by x define:

$$\begin{aligned} D(x) &= [b; w]^T . [1; x] \\ w &= \sum^{-1} .(\mu_2 - \mu_1) \\ b &= -w^T.\mu \\ \mu &= \tfrac{1}{2}.(\mu_1 + \mu_2) \end{aligned}$$

$D(x)$ is the difference in the distance of the feature vector x to the separating hyperplane described by its normal vector w and the bias b. If $D(x)$ is greater than 0, the observation x is classified as class 2 and otherwise as class 1.

MultiBoost with NB classifier

The Naive Bayes classifier estimates the posterior probability that an instance belongs to a class, given the observed attribute values for the instance. It builds a simple conditional independence classifier. Formally, the probability of a class label value y for an unlabeled instance x containing n attributes $\langle A_1, A_2, \ldots, A_n \rangle$ is given by:

$$\begin{aligned} P(y|x) &= \\ &= P(x|y) . \tfrac{P(y)}{P(x)} \\ &\propto P(A_1, A_2, \ldots, A_n|y) .P(y) \\ &= \textstyle\prod_{j+1} nP(A_j|y).P(y) \end{aligned}$$

The above probability is computed for each class and the prediction is made for the class with the largest posterior probability. The probabilities in the above formulas must be estimated from the training set.

MultiBoost with NN classifier

The Nearest Neighbor pattern classifier has shown to be a powerful tool for multi-class classification. The basic idea of the NN classifier is that whenever we have a new instance to classify, we find its K nearest neighbors from the training data. Given a query instance x_q to be classified:

- Let $x_1, x_2, \ldots, x_k$ denote the k instances from training examples that are nearest to x_q.
- Return the class that represents the maximum of the k instances.

A.3 Support Vector Machine (SVM)

SVM is based on the use of functions that can optimally separate data. When considering the case of two classes for which data are linearly separable, there exists an infinite number of hyperplanes for separating the observations. SVM's goal is to find the optimal hyperplane that separates data with maximizing the distance between the two classes and that goes middle of the two points classes of examples. The nearest points, which are used only for the determination of the hyperplane, are called support vectors. Among the models of SVM, there is linear-SVM and non-linear-SVM. The first are the simplest SVM because they can linearly separate data while for the second ones are used for data that are not linearly separable. In the last case the data are transformed to be represented in a large space where they are linearly separable.

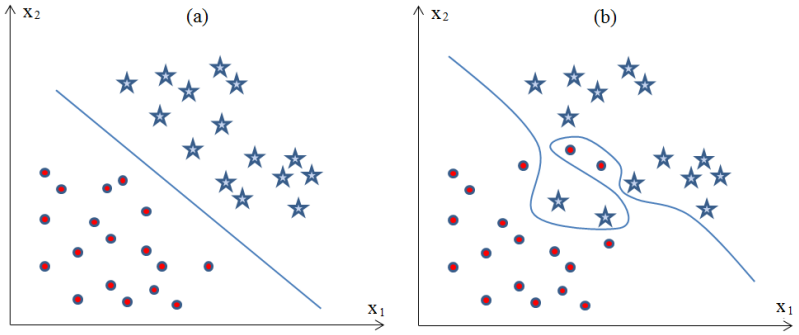

Figure A.1 – *(a)linearly separable data samples represented in a plane and separated by a straight line, (b) non-linearly separable data samples represented in a plane and separated by a curved line.*

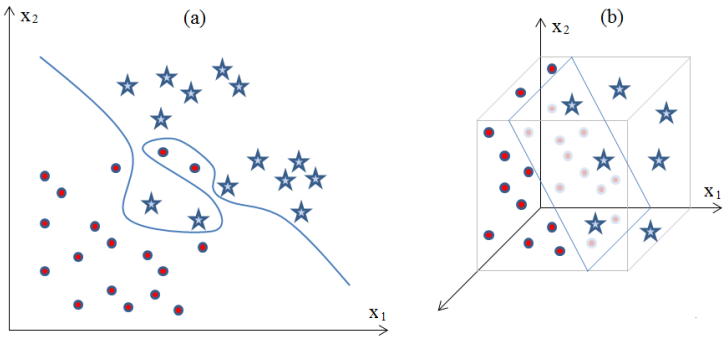

Figure A.2 – *(a)non-linearly separable data samples represented in a plane and separated by a curved line, (b) Plan separation after a transformation of of the same data samples into a 3D space.*

The evaluation of classification techniques is a recurrent problem which often depends on the difficult task of measuring generalization performance, i.e., the performance on new, previously unseen data. For most real-world problems we can only estimate the generalization performance. In order to evaluate a certain learning algorithm, we usually apply a cross-validation scheme.

A.4 Cross-validation

The traditional approach of cross-validation consists in dividing the data to classify into training and testing partitions a number of times. In our studies we applied K-fold cross-validation procedure (with $k = 10$), where the data is first partitioned into k equally (or nearly equally) sized segments or folds. Subsequently k iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining $k - 1$ folds are used for learning. Fellowing this procedure we measure the classification accuracy of the considered classifier.

A.5 Hidden Markov Model

HMM consists of <S, so, O, A, B>

- S : the set of states
- so: the initial state where everything begins
- O : the sequence of observations, each of which is drawn from a vocabulary, $\langle o_1, o_2, \ldots, o_T \rangle$
- A : the transitional probability matrix
- B: the emission probabilities, where $b_i(o_t)$ is the probability of observation o_t generated from state i

A.5.1 Forward-backward probability

- Forward probability: At time t, the probability that

  - we are in state i
  - the observation thus far has been $o1 \dots o_t$
  - $\alpha_t(i) = P(s_t = i, o_1, \dots, o_t|\lambda)$

- Backward probability: At time t and we are in state i, the probability that
  - the observation that follows will be $o_{t+1} \dots o_T$
  - $\beta_t(i) = P(o_{t+a} \dots o_T|s_t = i, \lambda)$

A.5.2 Baum-Welch algorithm

The Baum-Welch algorithm is used to estimate the parameters of a hidden Markov model. It can compute maximum likelihood estimates and posterior mode estimates for the parameters (transition and emission probabilities) of an HMM, when given only emissions as training data.

1. Initialize the parameters to some values
2. Calculate ''forward-backward'' probabilities based on the current parameters
3. Use the forward-backward probabilities to estimate the expected frequencies
   - Expected number of transitions from state i (to state j)
   - Expected number of being in state j (and observing o_t)
   - Expected number of starting in state j
4. Use the expected frequencies to estimate the parameters.
5. Repeat 2 to 4 until the parameters converge.

Bibliography

[AD08] Andrï¿½asson P., Dimberg U.: Emotional empathy and facial feedback. *Journal of Nonverbal Behavior 32* (2008), 215–224. 10.1007/s10919-008-0052-z. (Cited page 6.)

[ADD04] Abboud B., Davoine F., Dang M.: Facial expression recognition and synthesis based on an appearance model. *Sig. Proc.: Image Comm. 19*, 8 (2004), 723–740. (Cited pages 26 and 29.)

[BBADdB11] Berretti S., Ben Amor B., Daoudi M., del Bimbo A.: 3d facial expression recognition using sift descriptors of automatically detected keypoints. *The Visual Computer 27* (2011), 1021–1036. 10.1007/s00371-011-0611-x. (Cited page 73.)

[BBDD10] Berretti S., Ben Amor B., Daoudi M., Del Bimbo A.: Person independent 3D facial expression recognition by a selected ensemble of sift descriptors. In *3rd Eurographics/ACM SIGGRAPH Symposyum on 3D Object Retrieval* (Norrkoping, Sweden, May 2010), pp. 47–54. (Cited pages 72 and 73.)

[BDP*10] Berretti S., Del Bimbo A., Pala P., Ben Amor B., Daoudi M.: Selected sift features for 3D facial expression recognition. In *20th International Conference on Pattern Recognition* (Istanbul, Turkey, Aug. 2010), pp. 4125–4128. (Cited page 39.)

[BM03] Batista G., Monard M. C.: An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence 17* (2003), 519–533. (Cited page 73.)

[BV03] Blanz V., Vetter T.: Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell. 25*, 9 (2003), 1063–1074. (Cited page 24.)

[BY97] Black M. J., Yacoob Y.: Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision 25*, 1 (1997), 23–48. (Cited pages 32 and 34.)

[Can27] Cannon W. B.: The james-lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology 39*, 1 (1927), 106–124. (Cited page 5.)

[cAS10] Çinar Akakin H., Sankur B.: Spatiotemporal features for effective facial expression recognition, September 2010. (Cited pages 32 and 34.)

[ccAS08] Çeliktutan O., Çinar Akakin H., Sankur B.: Multi-attribute robust facial feature localization. In *FG* (2008), pp. 1–6. (Cited page 76.)

[Chr98] Christophe V.: *Les émotions: tour d'horizon des principales théories*. Savoirs mieux. Presses universitaires du Septentrion, 1998. (Cited page 4.)

[CM90] Cottrell G. W., Metcalfe J.: Empath: face, emotion, and gender recognition using holons. In *Proceedings of the 1990 conference on Advances in neural information processing systems 3* (San Francisco, CA, USA, 1990), NIPS-3, Morgan Kaufmann Publishers Inc., pp. 564–571. (Cited pages 27 and 29.)

[CT99] Cootes T., Taylor C.: Statistical models of appearance for computer vision, 1999. (Cited page 23.)

[DBH*99] Donato G., Bartlett M. S., Hager J. C., Ekman P., Sejnowski T. J.: Classifying facial actions. *IEEE Trans. Pattern Anal. Mach. Intell. 21*, 10 (1999), 974–989. (Cited pages 30 and 34.)

[DEP02] Darwin C., Ekman P., Prodger P.: *The Expression of the Emotions in Man and Animals*. Oxford University Press, 2002. (Cited page 4.)

[DGA09] Dibeklioğlu H., Gökberk B., Akarun L.: Nasal region-based 3d face recognition under pose and expression variations. In *Proceedings of the Third International Conference on Advances in Biometrics* (Berlin, Heidelberg, 2009), ICB '09, Springer-Verlag, pp. 309–318. (Cited page 91.)

[Dri11] Drira H.: *Statistical computing on manifolds for 3D face analysis and recognition*. PhD thesis, TELECOM-Lille1, 07 2011. (Cited page 60.)

[ED94] Ekman P., Davidson R.: *The nature of emotion: fundamental questions*. Series in affective science. Oxford University Press, 1994. (Cited page 12.)

[EF71] Ekman P., Friesen W. V.: *Constants Across Cultures in the Face and Emotion*, 1971. (Cited page 52.)

[EF78a] Ekman P., Friesen W.: *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, 1978. Consulting Psychologists Press. (Cited page 53.)

[EF78b] Ekman P., Friesen W.: *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978. (Cited page 78.)

[EFE72] Ekman P., Friesen W., Ellsworth P.: *Emotion in the human face: guide-lines for research and an integration of findings*. Pergamon general psychology series. Pergamon Press, 1972. (Cited page 13.)

[EHSH92] Ekman P., Huang T. S., Sejnowski T. J., Hager J. C.: *Final Report to NSF of the Planning Workshop on Facial Expression Understanding*. Tech. rep., 1992. Available from Human Interaction Lab, LPPI Box 0984, University of California, San Francisco, CA 94143. (Cited page 52.)

[EP97] Essa I. A., Pentland A. P.: Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell. 19* (July 1997), 757–763. (Cited pages 30 and 34.)

[FL03] Fasel B., Luettin J.: Automatic facial expression analysis: a survey. *Pattern Recognition 36*, 1 (2003), 259–275. (Cited page 24.)

[Fre90] Freund Y.: Boosting a weak learning algorithm by majority. In *Proceedings of the third annual workshop on Computational learning theory* (San Francisco, CA, USA, 1990), COLT '90, Morgan Kaufmann Publishers Inc., pp. 202–216. (Cited page 107.)

[Fri86] Frijda N.: *The emotions*. Studies in emotion and social interaction. Cambridge University Press, 1986. (Cited page 13.)

[FS95] Freund Y., Schapire R. E.: A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory* (London, UK, UK, 1995), EuroCOLT '95, Springer-Verlag, pp. 23–37. (Cited page 107.)

[FZSK11] Fang T., Zhao X., Shah S., Kakadiaris I.: 4d facial expression recognition. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011* (nov. 2011), pp. 1594 –1601. (Cited pages 45 and 48.)

[GWLT09a] Gong B., Wang Y., Liu J., Tang X.: Automatic facial expression recognition on a single 3d face by exploring shape deformation. In *Proceedings of the ACM International Conference on Multimedia* (Beijing, China, Oct 2009), pp. 569–572. (Cited pages 40, 47, and 69.)

[GWLT09b] Gong B., Wang Y., Liu J., Tang X.: Automatic facial expression recognition on a single 3d face by exploring shape

deformation. In *Proceedings of the 17th ACM international conference on Multimedia* (New York, NY, USA, 2009), MM '09, ACM, pp. 569–572. (Cited pages 72 and 73.)

[Hag] Hager J. C.: Dataface. (Cited page 15.)

[HFH*09] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H.: The weka data mining software: An update. *SIGKDD Explor. Newsl 11* (2009), 10–18. (Cited page 70.)

[HH97] Huang C., Huang Y.: Facial expression recognition using model-based feature extraction and action parameters classification. *Journal of Visual Communication and Image Representation 8* (1997), 278–290. (Cited page 34.)

[HL01] Hjelmås E., Low B. K.: Face detection: A survey. *Computer Vision and Image Understanding 83*, 3 (Sept. 2001), 236–274. (Cited page 22.)

[Hua97] Huang C.: Facial expression recognition using model-based feature extraction and action parameters classification. *Journal of Visual Communication and Image Representation 8*, 3 (1997), 278–290. (Cited page 31.)

[HWYW09] He L., Wang X., Yu C., Wu K.: Facial expression recognition using embedded hidden markov model. In *Proceedings of the 2009 IEEE international conference on Systems, Man and Cybernetics* (Piscataway, NJ, USA, 2009), SMC'09, IEEE Press, pp. 1568–1572. (Cited page 29.)

[HZY*08] Hu Y., Zeng Z., Yin L., Wei X., Zhou X., Huang T.: Multi-view facial expression recognition. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on* (sept. 2008), pp. 1 –6. (Cited page 29.)

[Iza77] Izard C.: *Human emotions*. Emotions, personality, and psychotherapy. Plenum Press, 1977. (Cited page 13.)

[Jam07] James W.: *What Is an Emotion?* Wilder Publications, 2007. (Cited pages 5 and 13.)

[JKSJ07] Joshi S., Klassen E., Srivastava A., Jermyn I. H.: A novel representation for riemannian analysis of elastic curves in $\mathbb{R}^n$. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)* (2007). (Cited pages 52 and 58.)

[KB84] Kochanek D. H. U., Bartels R. H.: Interpolating splines with local tension, continuity, and bias control. *SIGGRAPH Comput. Graph. 18*, 3 (Jan. 1984), 33–41. (Cited page 57.)

[K.M91] K.Mase: Recognition of facial expression from optical flow. 3,474–3,483. (Cited pages 30 and 34.)

[KOY*09] Kumano S., Otsuka K., Yamato J., Maeda E., Sato Y.: Pose-invariant facial expression recognition using variable-intensity templates. *International Journal of Computer Vision 83*, 2 (2009), 178–194. (Cited pages 32 and 34.)

[KS06] Klassen E., Srivastava A.: Geodesics between 3d closed curves using path-straightening. In *ECCV (1)* (2006), pp. 95–106. (Cited page 65.)

[Laz91] Lazarus R. S.: Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist 46*, 8 (1991), 819–834. (Cited page 6.)

[LIK09] Li Z., Imai J.-i., Kaneko M.: Facial-component-based bag of words and phog descriptor for facial expression recognition. In *Proceedings of the 2009 IEEE international conference on Systems, Man and Cybernetics* (Piscataway, NJ, USA, 2009), SMC'09, IEEE Press, pp. 1353–1358. (Cited page 29.)

[LJH22] Lange C., James W., Haupt I.: *The Emotions*. Psychology classics. Williams & Wilkins Company, 1922. (Cited page 5.)

[LMC11] Li H., Morvan J.-M., Chen L.: 3d facial expression recognition based on histograms of surface differential quanti-

ties. In *Proceedings of the 13th international conference on Advanced concepts for intelligent vision systems* (Berlin, Heidelberg, 2011), ACIVS'11, Springer-Verlag, pp. 483–494. (Cited pages 72 and 73.)

[LSN08] Lekshmi P., Sasikumar M., Naveen S. S.: Pca based facial expression using wavelets. In *IC-AI* (2008), pp. 286–290. (Cited pages 27 and 29.)

[LTC97] Lanitis A., Taylor C. J., Cootes T. F.: Automatic interpretation and coding of face images using flexible models. *IEEE Trans. Pattern Anal. Mach. Intell. 19* (July 1997), 743–756. (Cited pages 27 and 29.)

[LTH11] Le V., Tang H., Huang T.: Expression recognition from 3d dynamic faces using robust spatio-temporal shape features. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (2011), pp. 414 –421. (Cited pages 45, 48, 94, 95, and 96.)

[lTKC01] li Tian Y., Kanade T., Cohn J. F.: Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* (2001), 97–115. (Cited pages 28 and 29.)

[MMS08] Mpiperis I., Malassiotis S., Strintzis M.: Bilinear models for 3-d face and facial expression recognition. *Information Forensics and Security, IEEE Transactions on 3*, 3 (sept. 2008), 498 –511. (Cited pages 41 and 47.)

[Mow60] Mowrer O.: *Learning theory and behavior*. Wiley, 1960. (Cited page 13.)

[MSF78] M. Suwa N. S., Fujimora K.: A preliminary note on pattern recognition of human emotional expression. Proc. Int'l Joint Conf. Pattern Recognition, pp. 408–410. (Cited page 30.)

[MSJ07] Mio W., Srivastava A., Joshi S.: On shape of plane elastic curves. *International Journal of Computer Vision 73* (2007), 307–324. 10.1007/s11263-006-9968-0. (Cited page 62.)

[NC08] Nair P., Cavallaro A.: Mathcing 3d faces with partial data. In *British Machine Vision Conference* (Leeds, UK, September 2008), pp. 1–10. (Cited page 24.)

[PB11] Peng X., Bennamoun M.: Nose tip detection and tracking in 3d video sequences. In *GRAPP* (2011), pp. 13–22. (Cited page 91.)

[PC96] Padgett C., Cottrell G. W.: Representing face images for emotion classification. In *Neural Information Processing Systems (NIPS)* (1996), MIT Press, pp. 894–900. (Cited pages 27 and 29.)

[PH06] Peter C., Herbon A.: Emotion representation and physiology assignments in digital systems. *Interacting with Computers 18*, 2 (2006), 139–170. (Cited page 12.)

[Plu80] Plutchik R.: *A general psychoevolutionary theory of emotion*, vol. 1. Academic Press, 1980, pp. 3–33. (Cited pages 12 and 13.)

[Plu91] Plutchik R.: *The emotions*. University Press of America, 1991. (Cited page 9.)

[Plu03] Plutchik R.: *Emotions and life: perspectives from psychology, biology, and evolution*. American Psychological Association, 2003. (Cited pages xvi, 9, and 14.)

[PR00] Pantic M., Rothkrantz L.: Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 12 (December 2000), 1424–1445. (Cited page 24.)

[Rab89] Rabiner L.: A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE 77*, 2 (Feb. 1989), 257–286. (Cited pages 88 and 89.)

[RKS10] Rajpathak T., Kumar R., Schwartz E.: Eye detection using morphological and color image processing. *Image Rochester NY*, 1 (2010), 1–6. (Cited page 22.)

[Rus80] Russell J. A.: A circumplex model of affect. *Journal of Personality and Social Psychology 39*, 6 (1980), 1161–1178. (Cited pages xvi, 12, and 13.)

[RYD96] Rosenblum M., Yacoob Y., Davis L. S.: Human expression recognition from motion using a radial basis function network architecture. *IEEE Trans Neural Netw 7*, 5 (1996), 1121–38. (Cited page 31.)

[SAD*08] Savran A., Alyuz N., Dibeklioglu H., Celiktutan O., Gokberk B., Sankur B., Akarun L.: Bosphorus database for 3D face analysis. In *Proceedings of the First COST 2101 Workshop on Biometrics and Identity Management (BIOD)* (Denmark, May 2008). (Cited page 78.)

[Sch90] Schapire R. E.: The strength of weak learnability. *Mach. Learn. 5*, 2 (July 1990), 197–227. (Cited page 107.)

[Sch99] Schapire R. E.: A brief introduction to boosting. In *Proceedings of the 16th international joint conference on Artificial intelligence - Volume 2* (San Francisco, CA, USA, 1999), IJCAI'99, Morgan Kaufmann Publishers Inc., pp. 1401–1406. (Cited page 107.)

[Sch02] Schapire R. E.: The boosting approach to machine learning an overview. *LECTURE NOTES IN STATISTICSNEW YORK-SPRINGER VERLAG 27*, 2 (2002), 1–23. (Cited page 107.)

[SD07] Soyel H., Demirel H.: Facial expression recognition using 3d facial feature distances. *International Conference on Image Analysis and Recognition (ICIAR)* (2007), 831–838. (Cited pages 38, 47, 71, and 72.)

[SGM09] Shan C., Gong S., McOwan P. W.: Facial expression recognition based on local binary patterns: A comprehensive

study. *Image Vision Comput. 27*, 6 (2009), 803–816. (Cited pages 27 and 29.)

[SKJJ10] Srivastava A., Klassen E., Joshi S. H., Jermyn I. H.: Shape analysis of elastic curves in euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence accepted for publication* (2010). (Cited page 62.)

[Slo02] Sloman A.: Architecture-based conceptions of mind. In *In the Scope of Logic, Methodology, and Philosophy of Science (Vol II)* (Dordrecht, 2002), Kluwer, pp. 403–427. (Synthese Library Vol. 316). (Cited page 9.)

[SRY08] Sun Y., Reale M., Yin L.: Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition. *New York* (2008), 1–8. (Cited page 48.)

[SS62] Schachter S., Singer J. E.: Cognitive, social, and physiological determinants of emotional state. *Psychological Review 69*, 5 (1962), 379–399. (Cited page 6.)

[SSB11] Savran A., Sankur B., Bilge M. T.: Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 0 (2011), –. (Cited page 39.)

[SSB12] Savran A., Sankur B., Bilge M. T.: Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units. *Pattern Recognition 45*, 2 (2012), 767–782. (Cited pages 39 and 47.)

[SSDK09] Samir C., Srivastava A., Daoudi M., Klassen E.: An intrinsic framework for analysis of facial surfaces. *International Journal of Computer Vision 82*, 1 (2009), 80–95. (Cited page 55.)

[SY08] Sun Y., Yin L.: Facial expression recognition based on 3d dynamic range model sequences. In *Proceedings of the 10th*

European Conference on Computer Vision: Part II (Berlin, Heidelberg, 2008), ECCV '08, Springer-Verlag, pp. 58–71. (Cited pages 42, 48, 81, 93, 94, and 96.)

[SZPR12] Sandbach G., Zafeiriou S., Pantic M., Rueckert D.: Recognition of 3d facial expression dynamics. *Image and Vision Computing*, 0 (2012), –. (Cited pages 44, 48, 81, 94, 95, and 96.)

[TCJ10] Tong Y., Chen J., Ji Q.: A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE Trans. Pattern Anal. Mach. Intell. 32*, 2 (Feb. 2010), 258–273. (Cited page 34.)

[TH08] Tang H., Huang T.: 3d facial expression recognition based on automatically selected features. *In First IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB)* (2008), 1–8. (Cited pages 38, 47, 71, and 72.)

[TM10] Tsalakanidou F., Malassiotis S.: Real-time 2d+3d facial action and expression recognition. *Pattern Recogn. 43*, 5 (May 2010), 1763–1775. (Cited page 48.)

[Tom62] Tompkins S.: *Affect Imagery Consciousness - Volume I the Positive Affects*. Affect Imagery Consciousness. Springer Publishing Company, 1962. (Cited page 13.)

[Tom63] Tomkins S.: *Affect Imagery Consciousness - Volume II the Negative Affects*. Affect Imagery Consciousness. Springer, 1963. (Cited page 13.)

[TP91] Turk M., Pentland A.: Eigenfaces for recognition. *J. Cognitive Neuroscience 3* (January 1991), 71–86. (Cited page 26.)

[ULK09] Uddin M., Lee J., Kim T.-S.: An enhanced independent component-based human facial expression recognition from video. *Consumer Electronics, IEEE Transactions on 55*, 4 (november 2009), 2216 –2224. (Cited page 34.)

[Val84] Valiant L. G.: A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing* (New York, NY, USA, 1984), STOC '84, ACM, pp. 436–445. (Cited page 107.)

[VNP09] Vretos N., Nikolaidis N., Pitas I.: A model-based facial expression recognition algorithm using principal components analysis. In *Proceedings of the 16th IEEE international conference on Image processing* (Piscataway, NJ, USA, 2009), ICIP'09, IEEE Press, pp. 3265–3268. (Cited page 34.)

[Wat59] Watson J.: *Behaviorism*. University of Chicago Press, 1959. (Cited page 13.)

[WY07] Wang J., Yin L.: Static topographic modeling for facial expression recognition and analysis. *Comput. Vis. Image Underst. 108*, 1-2 (2007), 19–34. (Cited pages 26 and 29.)

[WYWS06] Wang J., Yin L., Wei X., Sun Y.: 3d facial expression recognition based on primitive surface feature distribution. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2006), 1399–1406. (Cited pages 37, 47, 71, 72, and 75.)

[Yac94] Yacoob Y. F.: *Computing spatio-temporal representations of human faces*. PhD thesis, College Park, MD, USA, 1994. AAI9514606. (Cited pages 31 and 34.)

[YCS*08] Yin L., Chen X., Sun Y., Worm T., Reale M.: A high-resolution 3d dynamic facial expression database. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on* (sept. 2008), pp. 1 –6. (Cited pages 78 and 89.)

[YKA02] Yang M.-H., Kriegman D. J., Ahuja N.: Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 1 (Jan. 2002), 34–58. (Cited pages 22 and 23.)

[YLW09] Yang J., Liao Z.-W., Wu Z.-D.: A method for robust nose tip location across pose variety in 3d face data. *2009 Inter-*

national Asia Conference on Informatics in Control Automation and Robotics (2009), 114–117. (Cited page 91.)

[YWS*06] Yin L., Wei X., Sun Y., Wang J., Rosato M. J.: A 3d facial expression database for facial behavior research. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition* (Washington, DC, USA, 2006), FGR '06, IEEE Computer Society, pp. 211–216. (Cited pages 53 and 78.)

[ZBVH09] Zaharescu A., Boyer E., Varanasi K., Horaud R. P.: Surface feature detection and description with applications to mesh matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Miami Beach, Florida, June 2009). (Cited page 45.)

[ZFRK09] Zhi R., Flierl M., Ruan Q., Kleijn B.: Facial expression recognition based on graph-preserving sparse non-negative matrix factorization. In *Proceedings of the 16th IEEE international conference on Image processing* (Piscataway, NJ, USA, 2009), ICIP'09, IEEE Press, pp. 3257–3260. (Cited page 29.)

[ZHDC10] Zhao X., Huang D., Dellandrï¿½a E., Chen L.: Automatic 3D Facial Expression Recognition based on a Bayesian Belief Net and a Statistical Facial Feature Model. In *International Conference on Pattern Recognition (ICPR)* (Aug. 2010), pp. 3724–3727. (Cited pages 40 and 47.)

[ZLSA98] Zhang Z., Lyons M., Schuster M., Akamatsu S.: Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proceedings of the 3rd. International Conference on Face & Gesture Recognition* (Washington, DC, USA, 1998), FG '98, IEEE Computer Society, pp. 454–. (Cited pages 26 and 29.)

[ZPRH09] Zeng Z., Pantic M., Roisman G., Huang T.: A survey of affect recognition methods: Audio, visual, and spontaneous

expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 31*, 1 (jan. 2009), 39 –58. (Cited page 24.)