Year: 2015 Order Number: 41826

3D Dynamic Facial Sequences Analysis for Face Recognition and Emotion Detection

By

Taleb ALASHKAR

2-Nov-2015

Thesis Committee

Mr. Renaud Seguier	Professor, CentraleSupélec Rennes, France	Reviewer
Mr. Stefanos Zafeiriou	Senior Lecturer, Imperial College London, UK	Reviewer
Ms. Bernadette Dorizzi	Professor, Télécom SudParis, France	Examiner
Mr. Stefano Berretti	Associate Professor, University of Fierenze, Italy	Examiner
Mr. Mohamed Daoudi	Professor, Télécom Lille, France	Supervisor
Mr. Boulbaba Ben Amor	Associate Professor (HDR), Télécom Lille, France	Supervisor
Ms. Hyewon Seo	CNRS Researcher (HDR), University of Strasbourg, France	Invited

Abstract

In this thesis, we have investigated the problems of identity recognition and emotion detection from facial 3D shapes animations (called 4D faces). In particular, we have studied the role of facial (shapes) dynamics in revealing the human identity and their exhibited spontaneous emotion. To this end, we have adopted a comprehensive geometric framework for the purpose of analyzing 3D faces and their dynamics across time. That is, a sequence of 3D faces is first split to an indexed collection of short-term sub-sequences that are represented as matrix (subspace) which define a special matrix manifold called, Grassmann manifold (set of k-dimensional linear subspaces). The geometry of the underlying space is used to effectively compare the 3D sub-sequences, compute statistical summaries (e.g. sample mean, etc.) and quantify densely the divergence between subspaces. Two different representations have been proposed to address the problems of face recognition and emotion detection. They are respectively (1) a dictionary (of subspaces) representation associated to Dictionary Learning and Sparse Coding techniques and (2) a time-parameterized curve (trajectory) representation on the underlying space associated with the Structured-Output SVM classifier for early emotion detection. Experimental evaluations conducted on publicly available BU-4DFE, BU4D-Spontaneous and Cam3D Kinect datasets illustrate the effectiveness of these representations and the algorithmic solutions for identity recognition and emotion detection proposed in this thesis.

Keywords: 4D face recognition, Grassmann manifold, Sparse coding, Spontaneous emotion detection, Early detection, depth videos, Grassmann trajectories, pain detection.

PUBLICATIONS LIST

Submitted journal papers

1. T. Alashkar, B. Ben Amor, M. Daoudi and S. Berretti. *"3D Spontaneous Emotion Detection by Analyzing Trajectories on Grassmann Manifolds"*, **Submitted** to IEEE Transaction on Affective Computing, Sep-2015.

2. T. Alashkar, B. Ben Amor, M. Daoudi and S. Berretti. *"Modeling Shape Dynamics on Grassmann Manifolds for 4D Face Recognition"*, **in preparation**.

Papers in international conferences and workshops

1. T. Alashkar, B. Ben Amor, S. Berretti and M. Daoudi, *"Analyzing Trajectories on Grassmann Manifold for Early Emotion Detection from Depth Videos"*, in 11^{th} IEEE International Conference on Automatic Face and Gesture Recognition, Ljubliana, Slovenia, May-2015.

2. T. Alashkar, B. Ben Amor, M. Daoudi and S. Berretti, *"A Grassmannian Framework for Face Recognition of 3D Dynamic Sequences with Challenging Conditions"*, in NORDIA ECCV Workshop 2014, Zurich, Switzerland, Sep-2014.

3. T. Alashkar, B. Ben Amor, M. Daoudi and S. Berretti, *"3D Dynamic Database for Unconstrained Face Recognition"* in 3D Body Scanning Technology, Lugano, Switzerland, Oct-2014.

National conferences

1. T. Alashkar, B. Ben Amor, S. Berretti and M. Daoudi, *"Analyse des trajectoires sur une Grassmannienne pour la detection d'emotions dans des videos de profondeur"*, in journees francophones des jeunes chercheurs en vision par ordinateur (ORASIS), Amiens, France, Jun-2015.

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

DEDICATION AND ACKNOWLEDGEMENTS

My efforts towards this thesis would not have been completed without the precious contribution of lots of people.

First of all, I would like to express my deep gratitude to **Prof.Mohamed Daoudi**, my PhD supervisor, for giving me the chance to work under his guidance for three years. His support, guidance and patience gave me the driving force to go forward through my PhD project. He was always available with his long experience to stand next to me to overcome any obstacle or challenge. I learned from Pro.Daoudi the importance of how to define precisely the scientific challenges in a research project, how to emphasize my effort on the right place, and that new ideas not that valuable without strong and coherent experimental validation.

Also, I would like to thank **Dr.Boulbaba Ben Amor**, my PhD co-supervisor sincerely for all his efforts and attention paid in all levels of my thesis. His advises, suggestions and everyday communication were very supportive and helpful for me and they let my work always have better shape and more valuable meaning. The most important skills, I have learned during my work with Dr.Ben Amor are how to position and link our new research ideas with the state of the art works and the patience to go through every single detail of a problem to figure out the best solution. Also, the art of illustrating complex scientific ideas in elegant and simple way for the community by writing a scientific papers or giving a talk and presentation.

My thanks also go for **Dr.Stefano Berretti** from University of Florence in Italy for his cooperation and involvement in my PhD project. In spite of the distance, the scientific meetings and discussions with Dr.Berretti always end with more clear picture in my mind about the problem and more deep thoughts about the best way to address it. His professional and academic touches had an important effect on my work. I leaned a lot from him during the three months mobility period under his direct supervision in MICC Lab in University of Florence.

A great thank for my all PhD committee members for their time and efforts they made to make my PhD defense, and a special thanks for my PhD thesis reporters for their valuable comments and feedbacks that draw my attention to important and detailed aspects. I would like to thank all our research group **3D SAM** members, Dr.Hazem Wannous, Dr.Hassen Drira, Prof.Jean Phillipe Vandeborre, Dr.Paul Auain Desrosiers and all people in CRTIStAL Lab and Telecom Lille Engineering School.

A special thanks for my graduated PhD colleagues Dr.Rim Slama and Dr. Xia Baiqiang, and current PhD students Maxime Devane, Vinecent Leon, Quentin Poloche and Meng Meng for the great and friendly research atmosphere we had together during three years. My great thanks for all my friends: May Kahoush, Ali Mickael, Ibrahim Hatem, Ragheed Alhydar, Achille Mickael who stood always next to me. Their moral support and compassion helped me a lot reach the end of my PhD successfully.

At last, i would like to thank all of my family members especially my brother Jafar for their priceless support and love.

This thesis has been funded by *Institut Mines-Telecom* under the program Futur et Ruptures

Table of Symbols

Table 0.1: List of symbols used and their definition in the thesis

Symbol	Definition/Explanation
$SO(n)$	Special Orthogonal Group of $\mathbb{R}^n$
$T_\mu(M)$	Tangent space to the manifold M at point μ
$\|\|A\|\|_F$	Frobenius norm of the matrix A
$\mathscr{G}_k(\mathbb{R}^n)$	Grassmann Manifold of k-dimension subspaces of $\mathbb{R}^n$
$\mathscr{L}_k(\mathbb{R}^n)$	Stiefel Manfiold of orthogornal matrices of size $n \times k$
$\mathscr{X}, \mathscr{Y}$	Subspaces on Grassmann manifold
log_μ	Logarithm map projects Grassmann elements to T_μ
exp_μ	Exponential map returns vector on T_μ to Grassmann manifold
$dist(.,.)$	distance on manifold
d_{Geo}	Geodesic distance
d_{proj}	Projection distance
d_{BC}	Binet-Cauchy distance
d_{Max}	Max Correlation distance
d_{Min}	Min Correlation distance
d_{Pro}	Procrustes distance
d_{Geo}	Geodesic distance
$\mathbb{D}$	Dictionary of atoms D_i
ω	window size (number of frames in 3D sequence)

CHAPTER 1

Introduction

The human facial analysis is a major field of research in computer vision and pattern recognition. The high interest in human faces comes not only from its ability to reveal the person's identity [102] or the demographic information (gender, age, ethnicity, etc.) [57], but also because it is considered as an important emotional and awareness communication channel, which reflects some of our cognitive activities and well-being [76] (sickness, stress, fatigue, ...). One of the most important applications of face analysis is identity recognition because it spans several applications, such as law enforcement, surveillance systems, access control, etc. [158]. The non-intrusive nature of human faces is its main advantage against other biometrics, like iris, fingerprint, voice, and hand geometry, which makes it more acceptable from end-users. That is, in face-based recognition (commercial) systems, there is no need to ask the person to make any physical contact with the system, just being constantly in front of the camera for a few seconds is enough. Recently, significant efforts have been paid to recognize people identity from recorded footages without any cooperation from their side by using surveillance cameras as done in the *Multiple Biometric Grand Challenge MBGC*[1] [129], it was also subject of several evaluation contests [21, 102] and recent research studies [41]. All these studies argue that robust face recognition in real-world conditions is still a distant goal.

From another perspective, the human face is considered as the major non-verbal communication channel between human beings that shows a person's emotional states via different facial expressions. The pioneering study conducted by *Paul Ekman* and

[1] http://www.nist.gov/itl/iad/ig/mbgc.cfm

his colleagues [44] approved the universality of six facial expressions (happiness, anger, sadness, fear, disgust, and surprise), where people from different cultures show the same facial expressions for the same feelings [24]. The strong acceptance of this affirmation in psychology opened the door to computer vision researchers to argue the discovery and consider it to design their automated facial expression analysis algorithms. However, since the human emotional states are more complicated than these basic six expressions in real world scenarios, researchers have focused recently on the automatic recognition of complex affects, such as thinking, hesitating, nervousness, etc. A more realistic annotation for human emotional states recognition is proposed, known as arousal-valence continuous human emotions charts [111]. In this annotation, the valence dimension indicates if the emotional state is positive or negative and its degree. The arousal dimension indicates the degree of activation of this state. To have an automatic recognition system, several studies confirmed that incorporating the body, like its posture and movements with the facial information can give a better understanding for human affects [94, 135]. Thus, facial expressions classification and emotional state detection draw increasing attention for several fields [3, 100], like in psychology, healthcare, robotics, and human-machine interaction.

Our faces also can provide a strong evidence about our cognitive state, like the degree of attention and physical state, pain and fatigue. Several applications started to appear in computer vision to improve human-machine interaction, like attention assessment application in online learning environment [58], fatigue detection for drivers from eye movement and head gesture [98], physical pain detection [10], stress detection [81], etc.

1.1 Motivation and challenges

Facial visual data analysis started several decades ago with 2D still color (or grayscale) images and the use of this data permitted to fulfill some applications, such as face recognition under strictly constrained conditions [153]. 2D still images show poor performance in spontaneous facial expression analysis and action units recognition, since they lack the temporal information [11]. Also, performance of 2D face recognition in real world scenarios based on still images, like surveillance system [13], face detection [116] and face recognition in the wild [144, 159] decreases significantly due to several challenges like: illumination variation, pose variation, self-occlusions by hands, hair or the face itself (when changing the head pose), external occlusions by objects, like sunglasses or scarf, scale variation and facial deformations.

All of these challenges motivated researchers to exploit 2D dynamic (video) data to solve such problems because: (i) the additional spatial information available in 2D videos can compensate the low-quality facial images, since we might have the face from different point of views or different distances; (ii) the temporal information resides in the 2D videos more effective in facial expressions and action units classification, since they are by nature dynamic actions. Even in face recognition application, using 2D video data can improve to certain limits the performance against previously mentioned challenges. Evaluations, such as MBGC, investigated unconstrained face recognition from still images and videos (2D), and showed distinctly that face recognition in adverse conditions is still a distant goal [11]. A second alternative is given by the availability of 3D acquisition systems, which opened the way to develop new solutions to face recognition and expression classification from 3D data. Since 3D face recognition approaches use the 3D geometry of the face, they have the advantage of being robust against illumination and pose variations [22]. However, most of the existing solutions are tested on datasets collected under well-controlled settings using either static acquisition systems, like laser scanners [101] or dynamic stereo-vision systems for 3D acquisition [128]. In general, such systems need offline processing to obtain the 3D face model. These limitations made current 3D approaches inconvenient for realistic scenarios [65]. More recent advancements of 3D acquisition technologies, like structured-light and time-of-flight scanners, made 3D dynamic systems available in the market at a lower cost. In spite of all these benefits, the streams of 3D images (depth, meshes, unstructured point clouds, etc.) present serious drawbacks, such as missing data when using a single-view capture system, depth acquisition noise, changes of spatial resolution, size of space-time data (non-availability yet of spatio-temporal compression techniques), which require the use of adapted methodologies and appropriate tools to handle these issues. My thesis is put forward in that context and proposes new compact representations and efficient algorithms for processing and analyzing 4D (i.e., 3D+t) data, for the purpose of face recognition and emotion detection.

After deciding the static data representation, one important choice will be the representation of their temporal evolution to perform efficient processing and address the above-mentioned problems. An emerging solution widely explored in 2D domain is mapping the original videos into a matrix manifold featuring suitable properties for the analysis [88]. Among these matrix domains, the Grassmann (space of k-dimensional linear subspaces of the Euclidean space $\mathbb{R}^n$ (called the ambient space) emerges as an interesting choice. In particular, one can cite: (i) its ability to produce compact low-rank representation for the original video data, which can handle missing and noisy data. Instead of performing feature extraction, as proposed in several works, our aim is to transform the original data and keep the possibility to (faithfully) reconstruct it back from the derived representation; (ii) it simplifies the computational complexity of comparing two dynamic 3D videos by performing it using a small number of inner products; (iii) the advanced statistical inference tools recently developed to fit the nonlinear structure of these Riemannian domains [59, 132]. For these reasons, our modeling of the temporal evolution of human 3D faces is based on mapping the original 4D data to Grassmann manifolds. Based on this idea, we introduce several contributions in this thesis.

1.2 Thesis contributions

In this thesis, we have studied the contribution of 3D facial dynamics (i.e., temporal evolution) for identity recognition and spontaneous emotion detection. Our study leads to several questions of two kinds, methodological and practical. The questions related to the methodology to be adopted are – (1) which representation is the most suitable for analyzing 3D faces and their dynamics? (2) How to compare 3D video clips under pose variations, missing and noisy data in an efficient way? (3) How the problem of dense correspondence over the 3D video can be resolved? (4) Is it possible to produce statistical summaries, like the mean, which allow us to perform data clustering efficiently? The practical questions are as follows – (1) Can the 3D facial deformations exhibited in our daily-life reveal our identity? (2) How to perform sequential (partial) analysis of 3D facial sequences to allow real time emotion detection?

In the following, we summarize our methodological and practical contributions, when considering (separately) the target applications. We recall that, when the same geometrical framework related to the subspace representation is common for the applications, two differences could be highlighted in a higher level. In fact, in 4D face recognition we adopt a **dictionary (of subspace) representation** coupled with sparse coding techniques, where a **trajectory (curve) representation** on Grassmann manifolds associated with an early event detector is proposed for (early) spontaneous emotion detection.

Face recognition from dynamic 3D data

In this part, we investigate the contribution of the temporal evolution of 3D faces (i.e., their shape's dynamic deformation) in identity recognition using 4D data. To this end, we adopt an (optimized) subspace representation of the flows of curvature-maps computed on 3D facial frames, after normalizing their pose. Such representation allows us to embody the shape as well as its temporal evolution within the same subspace representation. Then, we use recently-developed techniques of dictionary learning and sparse coding over the space of fixed-dimensional subspaces, called Grassmann manifolds, to perform face recognition. To show the effectiveness of the proposed method, we have conducted extensive experiments on the BU-4DFE dataset, and we discuss here obtained results with respect to current literature. Besides, two classification methods have been proposed: a Grassmann Nearest-Neighbor classifier (GNNC) involving geometric mean subspaces for subject classes, and a Grassmann Sparse Representation Classifier (GSRC) performed on the sparse representations of the subspaces. While the latter is inspired by an extrinsic solution, the former is an intrinsic solution. The GSRC is computationally cheaper and achieves better accuracy compared to GNNC. It also scores competitive performance with respect to the approaches previously proposed. Our evaluations showed clearly that considering the face shape's behavior over time improves the face recognition accuracy under both expression-specific and non-specific settings. We also investigated the proposed geometric approach on challenging face recognition scenarios under pose variation and other challenges, like facial expressions, talking, walking, internal and external occlusion from our collected database.

Spontaneous emotions detection in 4D data

We propose a unified framework for the purpose of online emotion detection, such as happiness or physical pain, in-depth videos. Our approach consists of mapping the videos onto the Grassmann manifold (i.e., the space of k-dimensional linear subspaces) to build time-parameterized trajectories. To do that, depth videos are decomposed into short-time clips, each approximated by a k-dimensional linear subspace, which is in turn a point on the Grassmann manifold. Considering the temporal evolution of subspaces gives rise to a precise mathematical representation of trajectories on the underlying manifold. Extracted spatio-temporal features based on computing the velocity vectors along the observed trajectories, termed Geometric Motion History (or GMH), are fed into an early event detector based on Structured Output SVM, thus enabling online emotion detection. Experimental results obtained on the publicly available Cam3D Kinect and BP4D-Spontaneous database validate the proposed solution. When the first database has served to exemplify the proposed framework on depth sequences of the upper part of the body (depth-bodies) from depth-consumer cameras, the same framework is also applied to high-resolution and long 4D-faces for physical pain detection, using the second database.

New full 3D/4D face dataset

In addition to the contributions presented above, we have collected a new 3D/4D FR database of 58 subjects, which presents the following features: (1) It includes the most common face recognition challenges in real-world like scenarios, such as pose variation, facial expressions, talking, walking, multiple persons in the scene, internal and external occlusions, which have not been included in any 4D database so far; (2) The low-resolution of the 3D scans is more convenient to simulate 4D face acquisition under less constrained conditions; (3) Free head movement is permitted during recording the 3D videos on the subject due to the wide field-of-view of the used 3D scanner. In addition to the 3D facial sequences (uncontrolled), we have also collected, for each subject, a full 3D static model with high-resolution (up to 50k vertices), with the texture mapped on it. We have conducted preliminary experiments on this dataset, in addition to our evaluation on publicly available datasets – BU-4DFE [148], Cam3D [90], and BP-4D Spontaneous emotion dataset [152].

1.3 Organization of the manuscript

After this general introduction, the rest of the thesis consists of four chapters and a general conclusion, as follows:

Chapter 2 provides a comprehensive state-of-the-art on dynamic face analysis from different imagery channels, with a particular emphasis on approaches which use 4D data. We first motivate the shift from 2D to 3D, then to 4D data, for both target applications face recognition and emotion classification and detection. A particular focus will be given to the recently-developed approaches, which exploit 4D data (meshes, depth images, point clouds, etc.) in a facial analysis.

In **Chapter 3**, we first recall essential background materials of the Grassmann geometry (distances, tangent space, geodesic, velocity vector, Karcher mean computation, etc.), then we derive our representations using (1) dictionary of subspaces and related tools, such that the sparse coding and dictionary learning, and (2) trajectory of subspaces representation and sequential analysis tools. The exploitation of these representations will be investigated in the next two chapters, respectively.

Chapter 4 presents our geometric framework for face recognition from 3D dynamic videos, which is based on sparse coding on Grassmann manifold and its comparison with baseline algorithms and previous studies. Experimental evaluation and discussions on the publicly-available BU-4DFE database are reported. In this chapter, we also describe our new Full 3D/4D face dataset and open the horizon to 4D face recognition in unconstrained conditions, with some preliminary experimental results.

In **Chapter 5**, trajectory analysis on Grassmann and Stiefel manifolds is presented with two applications: First, the early detection of spontaneous emotional states from depth videos of the upper part of the body. The importance of incorporating the upper part of the body with the facial data is exemplified here using the segmented Kinect Cam3D dataset. Second, the application of our framework to early detection of spontaneous physical pain affect from high-resolution 3D facial videos is presented. An experimental illustration and comprehensive discussion of the ability of trajectories on Grassmann manifolds to model 4D facial data is given in this chapter.

Chapter 6 summarizes the main contributions, states the main limitations of the proposed approaches and opens some perspectives and future directions.

CHAPTER 2

STATE-OF-THE-ART ON DYNAMIC FACE ANALYSIS

2.1 Introduction

Face analysis represents a major scope of study in computer vision and pattern recognition fields due to its wide range of applications in biometrics, human machine interaction, affective computing, etc. The design of any proposed solution in this domain related strongly to the availability of the imaging systems in the first place. In the last few years, 3D dynamic acquisition systems with both high- and low-resolution became available at affordable prices on the market. This technological innovation opened a new direction in front of facial analysis to exploit the richness of the new modality (3D+t or 4D). Researchers in computer vision needed to answer the fundamental questions concerning 3D dynamic systems like: what is the main additive values such new imaging systems carry into facial analysis problems? To which extent, using 3D dynamics can solve challenges that 2D (static/dynamic) and 3D static systems can't solve? What are the main limitations and constraints related to the adoption of such systems in automatic facial analysis solutions?

The starting point to find answers was collecting new databases that include the basic challenges and problems needed to be solved in facial analysis domain. Till now, more attention was paid to facial expressions analysis and human affects understanding from 3D dynamics, than face recognition problem. Another important aspect we would like to highlight here is the new trend in facial expression and emotional states recognition approaches to move from acted (or posed) to spontaneous and realistic, which are harder to solve, but more useful and valuable for real-world applications. Also, to make the facial expressions and emotional states and affects, like physical pain detection, much more useful in action, moving into early recognition and detection is very important. Early recognition and detection means that the proposed automatic system should be able to recognize the expression and give a decision as early as possible (i.e., with low-latency) and not to wait until the end of the state. Investigating these challenges and what performance 3D dynamic data have under such conditions is a major interest for our work in this thesis.

In this chapter, we review most significant contributions made in face analysis from dynamic data, in particular using 3D imaging systems. A review of face recognition from 3D dynamic data is presented in Sect. 2.2, and for emotion recognition in Sect. 2.3. A short review for spontaneous emotion detection and classification from 2D videos is presented in Sect. 2.4. Sect. 2.6 reviews the literature on physical pain recognition from facial data. In Sect. 2.7, a review on early event detection from dynamics data is drawn. The most important 3D dynamic facial databases in the community are discussed with a comparison in Sect. 2.8. In Sect. 2.9, we conclude and discuss where our work in this thesis stands according to the literature.

2.2 Face recognition from dynamic data

Dynamic face recognition approaches started with 2D color image modality. The main motivations for using the 2D videos for such problem come from the fact that dynamic faces can overcome real-world challenges. For example, (i) **the pose variation:** the availability of dynamic sequence from different poses for the individual can help to obtain a complete information for the face; (ii) **noise or missing data:** The 2D facial sequences can compensate such problems partially by its information richness; and the (ii) **facial temporal dynamic:** which can improve the identity recognition process.

There are four categories for face recognition from 2D/3D video: 1) image set-based approaches (called also multiple-instance), where the order of the images through the time is ignored (i.e. the motion information is not considered here); 2) motion-based approaches where only the motion information is considered; 3) super-resolution approach which consists to fuse several 2D/3D frames of low resolution to build higher resolution image; and 4) sequence-based approaches, where the image order is considered since they exploit the spatio-temporal information together to make the recognition process. Even though face recognition approaches from 2D videos can give better performance under illumination, pose variation and occlusion than 2D still, they can improve only to a certain limit. A complete survey about face recognition from 2D videos can be found in [11].

From another perspective, the advancement in imaging technologies made the 3D static scanning systems available on the market for research and industrial applications. The availability of such 3D static imaging systems led to a quantum leap in facial analysis applications, because of its efficiency in solving profound challenges in 2D static and dynamic domains, which are illumination change and pose variation. Also, it opens the door in front of merging the 2D texture information and the 3D geometry of human faces for robust face recognition solutions. 3D static solutions show higher performance than 2D static and dynamic solutions under pose variation, illumination changes and in the presence of occlusion (we refer the reader to [2] for a comprehensive discussion). In the last few years, 3D dynamic imaging systems started to appear combining the advantages of dynamic information alongside the 3D information in two main models: 1) high-resolution, but expensive, 3D dynamic acquisition systems, like the Di4D acquisition system. This system gives high- temporal and spatial resolutions and needs to make the acquisition under highly conditioned environment. It also requires an offline reconstruction process; 2) the low-resolution depth-consumer cameras, such as the Microsoft Kinect, which give depth data in low-resolution at 30fps in real-time, and are available at affordable price even for personal use. An overview of 3D dynamic facial sequences analysis is depicted in Fig. 2.1.

In the following sections, we review the state-of-the-art approaches based on this taxonomy, where a first level of categorization is made based on the target applications.

2.2.1 Motion-based approaches

Starting from the fact of human face is a dynamic surface by nature i.e. besides its constant shape feature it has its motion which is an important non-verbal communication channel. The face non-rigid dynamic can be categorized into a) the speech production movement, b) the facial expression and c) the eye gaze changes. Several studies from the psychology field addressed the question of: **How facial motion information affect face recognition process in human perception?** Actually, even some studies in the literature claimed that the motion information has no effect on the recognition process such as [33],[23], several other studies revealed evidence and findings that approve that the recognition could be improved in certain context [125],[108]. From the cognitive

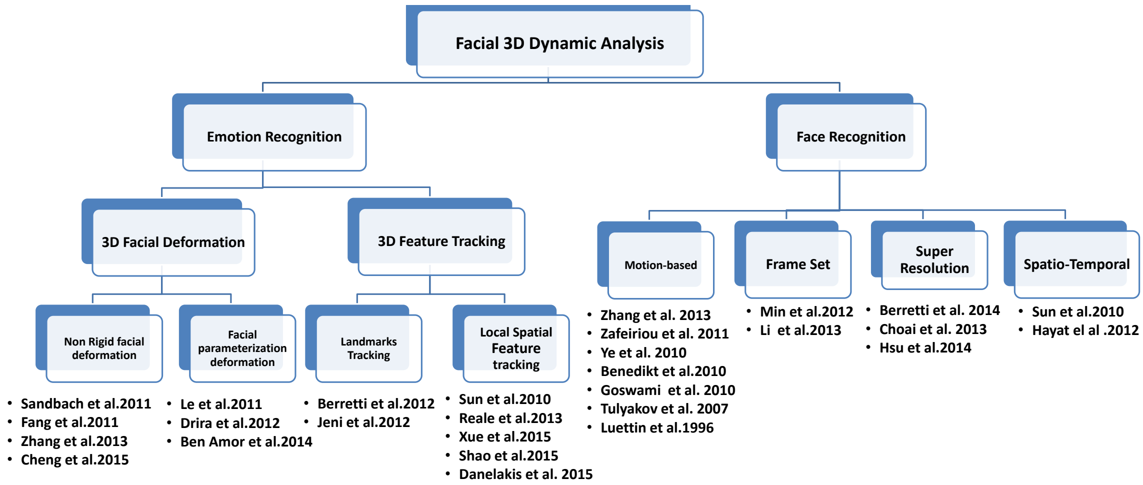

Figure 2.1: Taxonomy of 3D dynamic facial sequences analysis approaches in the two main targeted applications; face recognition and emotion classification.

point-of-view the motion information can support the identity recognition from facial sequences and there are two main directions here:

- The first direction of physiological studies posit that people depend firstly on the face structure static feature since it is consistent during the time and the they dynamic non-rigid facial deformations are not granted to be repeated reliably but the motion information can play a role in recognition when the quality of the face is degraded. Knight et Johnston [75] conducted a study to evaluate the role of the motion information and they found that the dynamic of the face gives better recognition when the quality of the shape is degraded significantly but not when the face image in a good quality. LANDER et al. [77] study showed that motion information improve the recognizing in low quality image and for famous faces more than others since the facial subtle changes need more time to be learned.

- The second direction posits that the additional views available from seeing the human motion information help the observer to infer the 3D structure of the face which is based on structure-from-motion concept. Also, they claim that the non-rigid deformation on the facial image gives cues about the 3D structure of the face. Pike et al.[104] study showed that seeing a human face in motion gives better recognition than in static or by seeing an image set that doesn't preserver the order of the deformation through the time and the motion information is more than a sum of multiple view of one static face image.

In computer vision community it was agreed that one of the challenges of the face recognition in 2D and 3D domain is its sensitivity to facial expression variations and several approaches are proposed to build expression-invariant face recognition systems such as Chang et al.[29]. Recently and inspired by works of physiology that approved that possibility to have idiosyncratic models from facial motions several works start to appear to investigate the efficiency of considering the facial dynamics as a biometric signature. Most of the works in this direction focus on speech production lips movement tracking over the time where few of them started to appear more recently that study the facial deformation which is not related to speech production. One of the first works on speech-related motion investigation as a behaviometrics is proposed by Luettin, et al.[87]. In this work the lips boundary and the intensity of the mouth area is tracked over 2D video to build spatio-temporal descriptor using HMM. The authors approve the possibility of identifying the speaker in both text dependent and text independent scenarios. Goswami et al. [54] proposed a method that models the appearance and the dynamics features of the lips region for speaker verification. The promising results obtained in this study made using the moving lips as a primary biometric modality is acceptable after it was seen as a soft-biometric before. An extension for this work is presented in [28]. Benedikt et al. investigated in [17] the uniqueness and permanence of facial action units that comes from verbal and non-verbal facial actions. Evaluation is conducted on 3D videos and it showed that the speech-related action units gives better performance in identification and verification than the speech-unrelated such as smile and disgust. Zhang et al. [152] proposed to distinguish between twins faces using the facial motion information extracted from their talking profiles. This study shows that the talking profile can be a good biometric for twins identification. Several works appeared to address the person identity recognition out of lips motion such as [107] [54], [47].

For speech-unrelated works that take the whole facial region deformation as a biometric, one of the earliest works that approved the feasibility of using facial motion as a biometric is presented in [34]. In [156], Zhang et al. proposed to capture the an elastic strain pattern which describes the anatomical and bio-mechanical characteristics of the facial tissue . This extracted pattern can serve as a new biometric to identify the person. This elastic strain pattern computed by applying finite elements method and the experimental study is conducted on a small 3D face dataset. Tulyakov et al in [131] modeled the facial motion information by computing the displacement between corresponding facial keypoints in two different images of the same person one in neutral state and the other in the apex of the expression state. The resulted pattern out of this distances showed that it can be used as a biometric for person verification on two datasets. Zafeiriou and Pantic in [149] also conducted a study to evaluate the efficiency of using the motion information out of smile/laughter spontaneous episode on a small dataset for person identification. Authors compute a motion complex vector fields between the neutral frame and the apex frame using the Free Form Deformation (FFD) algorithm and used complex data reduction technique such as complex LDA and PCA. The obtained results give evidence that the spontaneous smile/laughter facial expression is able to verify the identity of the person automatically. Previously mentioned works used facial motion as biometrics are limited to certain type of facial expressions such as smile, Ye et al.[147] proposed more general motion-based face recognition approach. In this method, author extracted identity evidence from various types of facial motions in a local manner and it is called Local Deformation Profile (LDP).

2.2.2 Frame-set approaches

One approach to exploit 3D dynamic data is by applying fusion at the decision level, which gives a more robust recognition process where the order of frames is not taken into account. These approaches that use more than one 3D frame for the person to learn his/her identity can improve the recognition. An example of such methods is proposed in [96], where a real-time 3D face recognition system using multiple RGB-D instances is presented. This approach shows that exploiting majority voting between multiple instances for short time, from 0.5 to 4 seconds, gives 100% recognition rate, while using the same approach on single depth image achieves 97.9% on a real-world small dataset of 20 subjects. Li et al. [79] proposed an algorithm for face recognition under varying poses, expressions, illumination and disguise from depth and color flows. For every subject, there are 89 RGB-D images under different combinations of pose, illumination, facial expressions and occlusion. 18 RGB-D images for every subject under different conditions used for learning two dictionaries one for depth and another for texture information separately, then a fusion is made at the decision level. The testing probe is one of the remaining samples. This work shows that using a set of images that covers different conditions for learning the subject class can give better recognition rate than using only one. Also, fusing the depth and the color channels gives better results of 96.7% compared to the result of the depth channel taken alone of 88.7%.

2.2.3 Super-resolution approaches

Another approach to deal with 3D dynamic data is to register the 3D depth or 3D available meshes to build a **super-resolution** face with higher quality and details. Thus, one can obtain better recognition rate than using single 3D frames to decide. The fusion here happened at the data level to have higher resolution data. Several works adopted this method for face recognition from 3D dynamic data, like in [20] where Berretti et al. investigated the impact of 3D facial scans resolution on the recognition rate by building super-resolution 3D models from consumer depth camera. A sequence of depth frames has been preprocessed, aligned and finally merged to create a super-resolution 3D face. Comparing this synthetic 3D face with 3D high-resolution model captured by $3dMD$ system shows that using the reconstructed (super-resolution) model outperforms single depth or high-resolution models acquired using a high-resolution system. In a similar way, Choi et al. in [32] have proposed a comparison study, in face recognition problem, between three methods – 1) single depth frame vs. set of depth frames, 2) single depth frame vs. another set of depth frames, 3) 3D model vs. 3D model, where this 3D model is constructed by registering a set of depth frames. The experimental results on a small dataset consisted of 20 RGB-D videos of 10 subjects show that 3D vs. 3D model approach gives the higher recognition rate. Hsu et al. [64] showed that super-resolution method can improve the recognition rate across pose variation. The 3D model captured from a depth sequence can help to have different 2D texture images of the probe in different pose settings to match the gallery texture image poses, which leads to better recognition rate. The main limitation of this method is the consuming time of the registration-merge process. Also, it might require annotated landmarks.

2.2.4 Spatio-temporal approaches

Since human face is a 3D surface with high dynamics features by nature, the spatio-temporal representation that can encompass both the 3D shape features and its motion traits through the time will be the most natural modeling and it is believed that it allows more efficient face analysis. This believe is supported by the success achieved in face recognition approaches that incorporated the dynamic traits with the static features but in 2D video such as [43], [83]. Also, several works start to appear recently that succeed to exploit the motion facial information as a biometric for identification and verification tasks out of 2D [147],[149] and 3D videos such as [17].

In this category, the 3D dynamic data should be aligned and tracked precisely through time to build a spatio-temporal descriptor. Here, unlike the frame set approaches, the frame order and alignment is critical to have a robust representation. In [128], Sun et al. proposed a spatio-temporal approach that uses a generic deformable 3D face model to track facial deformations in both space and time. To have an accurate temporal representation of the face deformation over time, a vertex tracking technique is applied to adapt the 3D generic model with each (static) scan. Thus, each 3D scan can be modeled by a spatial-temporal feature vector that describes the shape and the motion information to have an efficient representation. Two types of Hidden Markov Models (HMM) are trained – a Temporal (T-HMM), which models the motion information (inter-frame), and the Spatial (S-HMM) that model the geometrical face information on the same face (intra-frame). The two HMMs are combined to have (ST-HMM) at the decision-level. This approach applied on face recognition on 60 subjects of BU-4DFE dataset in Expression-dependent and -independent settings and it gives 97.89% and 94.14%, respectively. The main limitation of this approach is its high computational (time) cost especially for the vertex tracking step, which also needs a set of 22 landmarks annotation to be done. In their work, Sun et al. have made a comparative study with 2D-video and 3D static face recognition and have shown clearly the usefulness of the dynamic 3D data in face recognition. More recently, in [61] Hayat et al. proposed an automatic face recognition approach from 3D videos on BU-4DFE database. After automatic face detection, cropping and alignment of static frames, 3D scans are converted into depth frames. The face depth videos are divided into 4×4 non-overlapping video cuboids. A dynamic version of Local binary pattern (LBP) descriptor called (TOP-LBP) computed on each dynamic cuboids in three spaces (XY), (XT) and (YT), where X, Y is the depth frame dimension and T is the time dimension. LBPs-TOP computed from all video cuboids are concatenated to form a feature vector for the complete video using multi-class support vector machine (SVM) algorithm. Evaluating this method on all BU-4DFE database using 10-fold cross validation gives 92.68% of recognition rate. The promising results obtained from this spatio-temporal approaches show clearly the importance of including the temporal information beside the spatial in the 3D domain to have higher recognition rates. The main drawbacks reside in the tracking solution, which is very slow and sensitive to noise and missing data.

According to the very few works on 3D dynamic face recognition, this direction in face recognition is still not well explored. The results obtained from current literature state the importance of the temporal information with the spatial information. To address these issues in current spatio-temporal methods, we proposed to use in this thesis an optimized **subspace representation**. The main advantages of this choice are its ability to keep the spatial and temporal information as singular vectors, it provides compact and lower dimension representation of the high dimension original data, which makes 3D video classification and comparison faster, and it is a faithful representation since we can always come back to the original data from the subspace basis. The mathematical notation for this representation is introduced in the next Chapter 3 and our 4D face recognition approach presented in Chapter 4.

2.3 Emotion recognition from dynamic data

The non-verbal channel plays an important role in human-to-human communication, especially in feelings and emotional states recognition. This statement is confirmed in the study proposed by Mehrabian et al. [95], which states that in some context the visual, vocal and verbal elements participate in 55%, 38%, and 7% in feelings and attitude communication, respectively. Such studies motivated researchers in computer vision and affective computing to develop automated systems for emotional states and human affects detection and understanding from facial expressions and body language visual data [151]. The dynamic nature of facial expressions of human face motivated to model and analyze this problem in 2D videos in an early stage. The challenges that affect 2D videos, especially the pose variation and illumination changes, can hinder accurate facial expression analysis. More comprehensive survey of video-based facial expression analysis, challenges and limitations can be found in [97, 114].

The problems of pose variations and illumination changes can be solved in 3D modality, which had a great advancement in last few years where several 3D dynamic databases were collected for facial expression and action units recognition as discussed later in detail. In addition to this technological feasibility of studying facial expressions in 3D dynamic space, the human face itself is a 3D dynamic surface by nature. Several approaches appeared in last few years in the literature addressing the problem of automatic facial expressions analysis from 3D dynamic data, either from high-resolution 3D data or low-resolution depth data. The methodologies used in these approaches fall in two main groups – the 3D feature tracking approaches, and the second group including the 3D deformation based approaches, which depend on estimating the non-rigid deformation between static 3D frames themselves or by fitting a generic model.

2.3.1 3D feature tracking approaches

In this category of 3D dynamic facial sequences analysis, there are two methods. The first one is called local feature tracking. In this method, the 3D facial scans are divided into small patches around keypoints or landmarks, a local 3D feature is extracted from each patch and tracked along the video to have a spatio-temporal descriptor. The second method is called landmarks tracking approach. It focuses only on the keypoints or landmarks themselves not on the facial patches around them, where some distances between predefined landmarks on the facial scan are computed and tracked over the time to model the 3D facial dynamics data.

Local feature tracking approaches

Tracking the local spatial information on 3D faces through the video is one of the most common methodologies. Selecting the local descriptor is a critical point, and the 3D scans alignment is very important.

One of the earliest studies that addressed facial expression recognition from 3D dynamic scans is proposed by Sun et al. [128], which was applied for 3D dynamic face recognition and it is discussed in the previous section. The same approach was applied to classify the six facial expressions on the frame level using LDA classifier. The average recognition rate is 83.7%, which is better than the results obtained from 2D videos and 3D static approaches on the same database. More recently, Reale et al. [105] from the same group have proposed a 4D spatio-temporal descriptor called *Nebula* for 4D facial expressions and movement analysis. The starting point to build this descriptor is aligning the 3D frames precisely, and creating a local spherical voxel around the starting frame points, through the time. The curvature is computed for the points of this voxel and they are assigned into different label values to create the feature vector. Besides the curvature value, the least curvature polar angles are computed alongside the value to create the proposed spatio-temporal representation, which can be sensitive to the speed of performing the facial expression or the action unit. A histogram for each facial region is created from the three computed values (curvature value and the two polar angles) concatenating all histograms' regions to have the final feature vector. This new 4D descriptor is evaluated on BU-4DFE database for facial expression recognition and on BP4D-Spontaneous database for action units detection, and the reported results show that it outperforms previous spatio-temporal descriptors. This dynamic feature vector is computed on a subsequence of N frames, unlike the other vectors that are computed between two frames, which can speed-up the performance.

Furthermore, in [143] Xue et al. proposed a descriptor to analyze 3D expression sequences changing over time. The facial area is divided into spherical patches of radius r around 68 annotated landmarks. Applying Discrete Cosine Transform (DCT) on the three dimensions of the data results in a spatio-temporal representation for the 3D facial patch through the time (called 3D-DCT). Concatenating the 3D-DCTs for all facial patches gives the final facial representation for the 3D subsequence of size N. Passing these high-dimensional feature vector into *minimal redundancy maximal relevance (mRMR)* allows to keep the most relevant features. Then, a nearest neighbor classifier is applied to recognize the six facial expressions on BU-4DFE database, which gives 78.8% recognition rate on average. The results obtained from this work state that producing the spatio-temporal features from subsequences is more appropriate than extracting it from frame-to-frame deformation method.

In order to investigate the performance of 3D dynamic facial analysis on low-resolution depth data, Shao et al. [118] produced three different resolution depth videos from 3D videos of BU-4DFE database. In this work, the authors proposed to divide the facial area at both the gray scale (obtained from 2D color images) and the depth video (obtained from 3D scans) into spatio-temporal cuboids, then applying LBP-TOP descriptor on the volume data to have a robust representation for every cuboid. Pooling the feature vectors obtained from the grayscale and the depth cuboids together allows learning codebooks that represent dynamic facial expression videos by few feature vectors using sparse coding. Conditional Random Fields learning algorithm is used to classify the six expressions, and it obtains 83.07%, 79.38%, 69.1% for the six expressions using the three different decreasing resolutions, respectively. Addressing the same problem with 8 landmarks only, Danelakis et al. in [37] proposed a geometrical descriptor called Heart Kernel Signature (HKS). This descriptor is computed around each landmark on the 3D mesh itself and on the normal vectors estimated at each vertex, then concatenating the set to build a spatial feature vector of the scan. Applying a wavelet-based transformation on these spatial features over time gives rise to the spatio-temporal representation. Evaluation results on BU-4DFE dataset show theirsuperiority against many others.

Landmarks tracking approaches

In this method, the 3D face is represented only by the landmarks position themselves and some distances among them. Tracking this simple spatial representation through the video gives the spatio-temporal representation used to classify the expression embedded in the data. Berretti et al. [19] addressed the problem of facial expression recognition by proposing a real-time landmark tracking approach for analyzing 4D data. The method starts by detecting the nose tip first, then automatically detecting other facial landmarks around the mouth and eyes regions. A set of distances between mouth region areas, nose and mouth borders, and eyes area is computed to describe each 3D facial scan. These distances are normalized in two steps to be independent of the person. Finally, a HMM classifier is used for recognition evaluation on three expressions (Happy, Angry and Surprise) out of BU-4DFE database and achieves 76.3% classification rate on average. Another landmark-tracking-based approach is proposed by Jeni et al. [67] that addressed the independent person facial expression problem under pose variation in 2D and 3D dynamic facial data. In this method, the difference between landmarks of the neutral frame and the others through the video is measured and passed to multi-class SVM classifier. Evaluation on CK+ 2D video and BU-4DFE datasets shows interesting results. In this method, selecting stable landmarks tracking algorithm is very important for robust facial expression recognition performance.

2.3.2 3D facial deformation approaches

The main idea behind approaches in this category is the fitting accuracy performed between 3D frames and the reference to be able to measure the temporal evolution through the time. They are divided into two methods: the non-rigid facial deformation and the parametrized facial deformation.

Non-Rigid facial deformation approaches

The principle of these methods is the ability to capture the temporal deformation of the 3D facial scans by fitting a reference model to the 3D frames of the video. For example, in [112] a fully automatic approach for analyzing facial expression is introduced. After preprocessing and alignment of 3D frames of one video, the motion temporal information obtained by computing the Free Deformation Model (FDD) initially presented in [110] between successive frames and a quad-tree decomposition is applied to the resulted FDDs vectors to have more accurate feature description. Feature selection and training step are implemented in the same time using GentleBoost classifiers one for onset and another for offset segments. The temporal modeling is performed using HMMs, where the full expressions is considered as one HMM of 4 steps: the expression starts with *neutral*, then *onset*, *apex* and ends with *offset*. This approach is evaluated on three expressions (Happy, Angry, Surprise) available in BU-4DFE database and a comparison with 2D video data is conducted. Obtained results, 81.93% recognition rate, show that 3D dynamic data gives a higher performance. An extension of this work is presented in [113].

A fully automatic 4D facial expression analysis approach is presented in [46]. In this work, Fang et al. proposed a new 4D data registration approach that preserves temporal coherence between successive scans and robustness against outliers. The LBP-TOP descriptor initially proposed in [157] is implemented on the difference maps between 3D video frames and the first frame. Evaluation on BU-4DFE database gives 74.63% on the six expressions, it gives 96.71% when it is tested on three expressions (Angry, Happy and Surprise) and it gives 95.75% when it is tested on (Happy, Sad and Surprise). A Similar approach was proposed by the same authors in [45]. Different registration algorithms are evaluated including ICP (Iterative Closest Point) and more advanced mesh matching techniques, like MeshHOG and Spin Images with application to facial expression recognition from 3D static and dynamic scans. Since template fitting used in these approaches is important especially under facial expressions, recently, Cheng et al. [31] proposed a new algorithm to adapt a 3D model to a high-resolution depth scan. This fitting algorithm, called Active non-rigid ICP, can handle the highly deformable nature of the face by learning statistical models for local regions. Combining these statistical models with non-rigid Iterative Closet Point (ICP) algorithm, which is used also in [8], is implemented to have robust fitting. Evaluating the performance of the new fitting algorithm is approved by its higher performance on facial expression recognition from BU-4DFE database especially in strongly deformed scans, like in surprise expression.

Facial parameterization-based approaches

In [39] and [15], the authors proposed a Riemannian framework, which allows dealing with 3D face registration and pose normalization. The authors started a parameterization based on radial curves emanating from the nose tip with fixed rotation angle between them. These curves allow to capture the geometry of the 3D face where every curve consists of fixed number of points. To capture the dynamic facial deformation through the video, they used Riemannian method for shape analysis of curves to compute the Dense Scalar Fields (termed DSF). This DSF is the tangent vector field between the corresponding curves that belong to two different 3D faces after considering each curve as an element of a Riemannian manifold. Two classification schemes are proposed (1) using a multi-class Random Forest algorithm applied on the mean deformations and (2) HMM classifier applied on the motion. The authors provided evaluations on BU4DFE database with an average recognition rate of 93.21%. In [78], Le et al. presented a spatio-temporal method, which uses the planar iso-level curves as 3D face parameterization. These level curves give the spatial information of the 3D facial scan, and they used the Chamfer distance between corresponding curves of successive frames to capture the temporal evolution over time. Resulted features represent a spatio-temporal information, and they are passed to a HMM classifier. The evaluation results reported on happy, sad and surprise expressions gives 92.22% in average recognition rate from BU-4DFE dataset. A recent and more comprehensive survey on facial expression recognition from 3D video sequences is published in [38].

2.4 Spontaneous emotion recognition

Within the efforts dedicated to bring spontaneous facial expressions from 2D to 3D, databases started recently to appear considering this aspect. For example, in [7] Sherin et al. presented a Kinect based facial expressions database for recognizing seven acted and spontaneous expressions of 32 subjects. Zhang et al. [152] presented a high-resolution 3D dynamic spontaneous facial expression database with 3D and 2D textured videos for 41 subjects. Mahmoud et al. [90] created a 2D texture and depth video database for complex mental states including, in addition to the face, the upper part of the body. In particular, incorporating these latter data in the dataset can help in understanding the complex emotional and affect states.

Several works appeared in last few years addressing the spontaneous facial expressions classification. In [36], Cruz et al. proposed a bio-inspired approach for spontaneous facial emotion analysis. Authors of this work were motivated by the cognitive principle according to which the human vision system pays more attention to the parts of the scene with the highest dynamics. This approach implemented this principle by unfixed video down-sampling rate. The results confirmed that temporal video down-sampling according to the temporal change is more efficient than uniform rate down-sampling and faster than using the full video frame rate. This method is limited mainly by the influence of the accuracy of the apex labeling on the performance. Abd El Meguid et al. [3] proposed a fully automatic framework for spontaneous facial expressions detection and classification using random forest classifier. This framework works independently of the training dataset, and in unconstrained scenarios with pose and illumination variation, also providing real-time performance. In [103], an aggression detection out of other spontaneous facial expressions framework is presented by PiƒÖtkowska and Martyna. Senechal et al. [117] present an algorithm to detect spontaneous asymmetric facial expressions, like (smark) out of natural symmetric facial expressions from 2D videos. In [150], Zeng et al. proposed a one-class classification problem to distinguish between emotional facial expressions and non-emotional ones. A kernel subspace method is applied to model the facial expressions with support vector data description classifier and validated on Adult Attachment Interview (AAI) database [109]. Kamarol et al. [72] proposed a new spatio-temporal feature extraction method with application to spontaneous facial expressions classification, which outperforms the state of the art feature extraction methods in term of classification rate and computational time. Liu and Yin [82] proposed a new descriptor for spontaneous facial expression analysis, but using thermal video images. More detailed and comprehensive surveys on automatic human affect detection and recognition from facial expressions are available in [25, 114].

From this review, one can note **the increasing interest in this research direction, recently**. This thesis investigated this problem as it will be presented in Chapter 5.

2.5 Subspace representation for face classification

Subspace representation for dynamic facial information either for image sets or for image sequences (videos) showed a great success in this field of computer vision. Shigenaka et al. [119] proposed a Grassmann distance mutual subspace method (GD-MSM) and Grassmann Kernel Support Vector Machine (GK-SVM) comparison study for the face recognition problem from a mobile 2D video database. In [89], Lui et al. proposed a geodesic distance based algorithm for face recognition from 2D image sets. In this work, they exploited the canonical correlation analysis between two subspaces and used geodesic distance to consider the whole geometry of the subspace in the similarity score. Experiments conducted on 2D face image datasets show better recognition for this approach over others. More recently, Wang et al. [66] proposed learning projection distance on Grassmann manifold for face recognition from image sets. Every image set is represented as a Gaussian distribution over the manifold to model the data overall distribution, not only the image sets information, which results in an improved recognition. Turaga et al. [132] presented a statistical method for video based face recognition. These methods use subspace-based models and tools from Riemannian geometry of the Grassmann manifold. Intrinsic and extrinsic statistics are derived for maximum-likelihood classification applications. In [60], Harandi et al. proposed a

Grassmann Discriminant Analysis (GDA) approach, which is an extension of the Linear Discriminant Analysis (LDA) algorithm to work with nonlinear structures. A graph embedding framework is used in this work to build two within-class and between-class similarity graphs, which move the classification problem from non-linear Grassmann manifold into vector linear space. The application of this approach to face recognition and object classification shows good results.

From these presented works, the subspace representation of the 2D facial image sets or sequences showed a high performance and robustness against noise, missing data. Besides, it reduced the computational costs of comparing two image sets in many to many scenario and converted it into two low dimensional linear subspaces comparison. All of that, gives us the motivation to explore the performance of subspace representation for modeling 3D dynamic data for the first time.

2.6 Physical pain detection in videos

Physical pain detection and estimation from facial images attracted more attention recently due to its important applications in healthcare systems, clinical treatment especially for people in a coma, under surgery or suffering from speech organs disorders. Lucey et al. [85] presented a facial video database (known as UNBC-McMaster Shoulder Pain Expression archive) for people suffering from shoulder pain with action unit coding on the frame level of the video. The same authors extended the work in [86], by proposing an Active Appearance Model (AAM) system that can detect the frame with pain expression out of others in 2D texture videos. A full automatic pain intensity estimation approach from 2D image sequences from UNBC-MacMaster database is presented by Kaltwang et al. [71]. In [74], Khan et al. proposed a new facial descriptor called pyramid local binary pattern (PLBP), with application to pain detection on UNBC-Macmaster database. Their approach gives near real-time performance, with high recognition rate. Unlike previously mentioned works, Sikka et al. [122] proposed sequence level spatial-temporal descriptor instead of frame level to exploit the advantage of temporal information in the 2D video in combination with bag-of-words framework. This approach gives better results on MacMaster Shoulder Pain database approving the positive effect of temporal information on recognizing pain.

Since the works listed above are based on 2D images, they are affected by pose and illumination variations, which can be solved by moving to 3D imaging systems. Following other facial computer vision problems, pain recognition may be considered in 3D facial databases. In BP4D-Spontaneous 3D dynamic database [152], there is one task of spontaneous physical pain experience for 41 subjects. Zhang et al. [154] proposed a pain-related action units detection on BP4D database using binary edge feature representation. This approach exploits the available temporal information alongside the 3D facial scans as well as their robustness against pose variation. A more comprehensive survey on pain detection from facial expressions can be found in [10].

From the review above, it emerges the importance of the early detection aspect for several applications, especially computer machine interaction, and **the very limited works that addressed this problem for spontaneous facial expression from 3D dynamic data. This was the main motivation to orient part of the work in this thesis to explore the opportunities and limitations that 3D dynamic data have for a such complex scenario**.

2.7 Early event detection in videos

The majority of video analysis methods propose expression classification based on the observation of the entire 3D dynamic sequence (i.e., a decision is taken once the full sequence is observed). In these works, no emphasis is placed on the responsiveness, i.e., on the capability to produce a correct classification just from a partial observation, as short as possible, of the sequence. This latter capability is indeed expected to be of great relevance to real contexts of application. Studying the trade-off between the accuracy and observation size for rapid recognition is an important topic in a wide spectrum of applications, ranging from video security to clinical treatments. This aspect has been investigated through several studies, in different domains and from different perspectives. Indeed, the trade-off between the accuracy and observation size for rapid recognition is an important topic in a wide spectrum of real applications. Schindler and Van Gool [115], first investigated this aspect by evaluating how many frames were required to enable action classification in RGB-videos. They found that short action snippets with as few as 1–7 frames were almost as informative as the entire video. This aspect has been addressed in few works. Su et al. [126] presented a high-frame-rate 3D facial expressions recognition system, based on an early AdaBoost classifier, but the test dataset was limited to few subjects and the facial expressions were posed, with a very high temporal resolution. The six basic expressions are collected five times for the same person with 100 fps as a temporal resolution. The concatenated animations of facial markers position in the 3D space are used as a feature vector after refining them by wavelet spectral subtraction. In [127], Su and Sato proposed an early recognition framework based on RankBoost with application to facial expression recognition. Starting from the fact that the intensity of the facial expression generally increases from the onset to the apex monotonically, this increase is learned by weak rankers in the same temporal order. Applying the weight propagation on the weak rankers, the early recognition system is built. Results are reported on the Cohn-Kanade (CK) 2D video dataset, and on a small 3D high temporal resolution dataset of six subjects.

More recently, Hoai and De la Torre [63] proposed a learning formulation for early event detection. Their maximum-margin framework is devised for training temporal event detectors capable of recognizing partial events, thus enabling early detection with minimal latency. Their method extends the Structured Output SVM to accommodate sequential data. They showed the effectiveness of the framework for detecting facial expressions, recognizing hand gestures, and classifying human activities from video sequences.

2.8 Dynamic facial datasets

In this section, a comprehensive survey for the spontaneous dynamic 2D and 3D dataset oriented for facial expressions problems will be survey and the most important 3D dynamic (4D) facial analysis datasets.

2.8.1 Spontaneous dynamic facial expression datasets

Facial expressions classification and emotional states detection remained for long-time focusing on acted facial expressions due to the difficulty of collecting and annotating spontaneous and natural facial expression databases. Recently, more attention has been paid to the analysis of spontaneous facial expression and emotion detection. Several databases have been collected for this purpose as reviewed hereafter.

The FeedTUM database [137] proposed by Wallhoffet et al. who tried to solve the problem of deliberated facial expression in dynamic databases. So, it gathered the basic six emotions from 18 different individuals. To achieve spontaneous facial expressions, they played video clips or still images after a short introduction phase instead of telling the person to play a role. This includes that head moves in all directions are also allowed. Videos are captured using Sony XC-999P camera that gives images with size of 640×480 pixels, a color depth of 24 bits and a frame rate of 25 frames per second. Due to capacity reasons, the images where converted into 8 Bit JPEG-compressed images with a size of 320×240. The DaFEx database [12] proposed by Battocchi et al. is a database created with the purpose of providing a benchmark for the evaluation of the facial expressibility of Embodied Conversational Agents (ECAs). DaFEx consists of 1008 short videos containing emotional facial expressions of the 6 Ekman's emotions plus the neutral expression. The facial expressions were recorded by 8 Italian professional actors (4 male and 4 female) in two acting conditions ("utterance" and "no- utterance") and at 3 intensity levels (high, medium, low). For capturing videos, a Canon MV360i was placed on tripod is used. After a post-processing step, final data saved in .avi format yielding a final size on screen of 360×288 pixels. To overcome the challenges of illumination variation in imaging conditions, Wang et al in [139] created NVIE 2D videos, which contain visible and thermal infrared images for natural and posed database for six basic expressions of 100 subjects. Two cameras have been used for this task, a DZ-GX25M 2D visible camera with 30 fps as temporal resolution which gives 704×480 image sizes. A SAT-HY6850 infrared camera with 25 frames per second as temporal resolution, which gives images of size 320×240 and wave band $8-14\ \mu m$. The LIRIS-ACCEDE database proposed by Baveye et al. in [14] is a large 2D videos database collected from public available films and movies with extensive annotation for affective content analysis. It contains 9,800 clips that last between 8 to 12 seconds extracted from 160 different movies. This database is annotated in Arousal-Valence space by experts and available for public use. The MAHNOB-HCI multimodal database is proposed in [123] by Soleymani et al. It contains facial videos, voice data, eye gaze data and peripheral/central nervous system physiological signals for 27 subjects. Spontaneous emotions induced by showing videos to the participants. Facial visual data captured using two imaging systems, one Allied Vision Stingray F-046C, a color camera, and five Allied Vision Stingray F-046B, monochrome cameras. The temporal resolution for all cameras is 60 fps and the spatial resolution is 780×580.

In [92], Mavadati et al. created a spontaneous action units intensity database called DISFA. There are 27 subjects in this database that were collected using a high-resolution ($1,024 \times 768 pixels$) BumbleBee point gray stereo-vision system at 20 frames per second under uniform illumination. Action units' intensity levels are annotated using a scale from 0 (action unit not activated) to 5 (maximum intensity) by two FACS expert coder. 66 landmarks were annotated using an Active Appearance Model (AAM) method. In [93], Mckeown collected an audio-visual database, SEMAINE, for spontaneous effective states by interaction between an operator and the participant consisted of 20 subjects. The operator plays four different roles to evoke four different emotional states for the participants. A high-resolution imaging system is used, which consists of five synchronized cameras that record by 50 fps as temporal resolution and 780×580 as spatial resolution. Annotation is made on five dimensions – Valence, Activation, Power, Anticipation/Expectation – with the addition of Overall Emotional Intensity. SMIC is a spontaneous micro-expression database proposed in [80] by Xiaobai et al. It contains 164 micro-expression 2D video clips that belongs to 16 subjects which can be a benchmark for micro-expressions detection and recognition approaches. 16 movies are used to induce the spontaneous emotions of the participants. A high speed (HS) camera (PixeLINK PL-B774U, 640×480) of 100 fps was used to collect the database in addition to another normal speed 25 fps imaging system, which consists of a normal visual camera and a near infrared camera of spatial resolution 640×480 both.

Within the efforts dedicated to bring spontaneous facial expressions from 2D into 3D, new databases started to appear recently considering this aspect. Mahmoud et al. in [90] collected a set of 108 audio/video segments of natural complex mental states of 7 subjects. Each video is acquired with the Kinect camera, including both the appearance (RGB) and depth information. The data capture natural facial expressions and the accompanied hand gestures. The emotional states are: Agreeing, Bored, Disagreeing, Disgusted, Excite, Happy, Interested, Sad, Surprised, Thinking and Unsure. This database was collected using two cameras: the HD cameras provide 720×576 pixel resolution color images at 25 fps and the Kinect sensor provides a color image and a disparity map, which is the inverse of depth values, at 30 fps. In [7], depth spontaneous facial expressions VT-KFER database is proposed for acted and spontaneous facial expressions. It includes 7 expressions, which are happiness, sadness, surprise, disgust, fear, anger, and neutral for 32 subjects. A set of 121 automatically detected facial landmarks is provided with the depth frames with their correspondence on 2D texture images. The Microsoft Kinect camera was used in the acquisition.

From works summarized in Table 2.1, we can notice the increasing interest is moving from acted facial expressions and action units into the spontaneous ones, which are closer to real world scenarios, but more challenging for automatic recognition and detection. Also, most of the works induced the spontaneous emotions by showing specific videos in front of the participants, other techniques hiring professional actors, taking videos from movies or making interaction with an operator. Most recently, **spontaneous facial emotion analysis brought into 3D domain** by collecting depth and 3D high-resolution spontaneous datasets.

Table 2.1: Overview of spontaneous facial expressions and action units datasets.

Reference	# Subject	Type	Imaging Systems	Purpose
FeedTUM [137]	18	2D video	Sony XC-999P	FER
DaFEx [12]	8 actors	2D video	Canon MV360i	FER
NVIE [139]	100	2D video/ Infrared	DZ-GX25M / HY6850	FER
LIRIS-ACCEDE [14]	NA	2D video	from movies	FER
MAHNOB-HCI [123]	27	2D video	Vision Stingray F-046C	FER
DISFA [92]	27	2D video	BumbleBee stereo-vision	Aus
SEMAINE [93]	20	2D video	Color and Gray cameras	FER
SMIC [80]	16	2D video	HS PixeLINK PL-B774U / Near Infrared camar	Micro FER
Depth Corpus [90]	7	2D video/ Depth	HD cameras / MS Kinect 1.0	ESR
VT-KFER [7]	32	2D video/ Depth	MS Kinect 1.0	FER
BP4D [152]	41	3D video/ 2D video	Di3D system	FER/ AUs

2.8.2 3D dynamic facial databases

In recent years, several facial 3D dynamic databases have been introduced to analyze the dynamic nature of human faces, mainly for expression/emotion and action units recognition. The BU-4DFE dataset, collected by Yin et al. [148] consists of 4D faces (sequences of 3D faces). The database included 101 subjects and was created using the Di4D (Dimensional Imaging) passive stereo-photogrammetry imaging system. It contains sequences of the six prototypical facial expressions with their temporal segments (neutral-onset-apex-offset-neutral) with each sequence lasting approximately 4 seconds. The temporal and spatial resolution is 25 fps and 35,000 vertices, respectively. The main limits of this database are that it contains posed facial expressions, and restricted acquisition environment (well-controlled illumination and frontal view of the subject's face), which makes it far from real scenarios. Cosker et al. [35] presented the first database that contains coded examples of dynamic 3D Action Units (AUs) in D3DFACS. There are 10 subjects in this dataset, including 4 FACS experts, and they were asked to perform 38 AUs in various combinations. Totally, there are 519 AUs sessions at 60 fps as temporal resolution. Each action unit consisting of 90 frames approximately. An FACS expert coded the peak of each sequence. It is more oriented for AUs recognition, captured under highly conditioned framework with posed facial expressions, too. The Hi4D-ADSIP

Table 2.2: Comparison of existing 4D Face databases.

Database	# Subjects	Temporal Resolution	Spatial Resolution	Illumination condition	Pose variation
Bu4DFE: [148]	101	25 fps	35k	Controlled	Limited
BP4D-Spon: [155]	41	25 fps	40k	Controlled	Limited
D3DFACS: [35]	10	60 fps	30k	Controlled	No
Hi4DADSIP:[91]	80	60 fps	20k	Controlled	Limited

database, presented by Matuszewski et al. in [91] is a 3D dynamic facial database, which contains facial articulation. Both, the temporal resolution, 60 fps, and the spatial resolution, 2352×1728 pixels per frame, are highly recorded using the Di4D system. In total, there are 80 subjects in this dataset with 3360 sequences. Subjects have various age, gender and race. The seven basic facial expressions are included with seven facial articulations. The main reason to include these articulations is to support the clinical research on facial dysfunctions. The facial expression recognition algorithm was applied to validate the part of the database containing standard facial expressions. Two different algorithms in static and dynamic mode are applied. In addition, a psycho-physical experiment that was used to formally evaluate the accuracy of recorded expressions is conducted. Where the first dataset is publicly available, the two last ones are private.

Finally, Zhang et al. [155] have created a high-resolution spontaneous 3D dynamic facial expression Database, called BP4D-Spontaneous. Also, for this dataset, the Di4D system was used for the acquisition, but the expressions are not posed; instead they are spontaneously conveyed by the participants. Expressions include happiness or amusement, sadness, surprise, embarrassment, fear or nervous, physical pain, anger or upset and disgust. There are 41 participants in this database. For each subject, 3D and 2D videos lasting about 1 minute for each scenario are captured. Manually annotated action units (FACS AU) by a certified FACs coders, automatically tracked facial landmarks and head pose in 3D/2D videos are provided with the database. Table 2.2 presents a comparison between existing dynamic 3D face datasets and the dynamic part of our 3D-4D database.

From this summary, one can note the following points – the great recent interest of the community in facial analysis from **dynamic data** is motivated by the importance of the new dimension (time) for better understanding of facial expressions, emotions and action units; Most of these datasets are designed for **facial expressions and/or action units** problem and do not address face recognition.

2.9 Conclusion

In this chapter, we have reviewed prior work to facial analysis from dynamic data, in particular for two applications – face recognition and emotion classification. A taxonomy of current literature is first presented, then a set of papers have been discussed in each category. From this review, one can first note the novelty of the topic – exploiting 4D data for face understanding. Only very few research groups have made advanced studies and have confirmed the interest of using sequences of 3D facial shapes instead of video data. However, the proposed approaches are computationally expensive in general and often need 3D landmarks annotation and tracking. The most promising representations and methodologies are derived from 3D (static) approaches, such that template fitting and non-rigid registration, which are time-consuming and sensitive to noisy and missing data. The above-mentioned challenges have motivated us, in this thesis, to focus on representations suitable for dynamic data based on subspace methods as a first modeling level. In a next level, two major representations based on **dictionaries** and **trajectories** are proposed for dictionary learning and sequential analysis, respectively.

In the next chapter, we shall introduce essential mathematical materials of Grassmann manifolds and computational tools needed to introduce our contributions.

CHAPTER 3

Geometric Framework for Modeling 3D Facial Sequences

3.1 Introduction

From the previous chapter, one can note the important aspects and motivations, which lie behind our choice to work on 3D dynamic facial sequences for face recognition and early detection of spontaneous emotional states and affects. Inside, the very first and important question, which needs to be answered is – **Which representations of static and dynamic shapes are more suitable to study such problems?**

In this chapter, we start presenting the dynamic 3D data and the subspace representation adopted in our solutions. In Sect. 3.2, we introduce the subspace representation and why it is selected in this work. The notation of Grassmann manifold is given with the definitions of several distances and metrics in Sect. 3.3. Sect. 3.4 presents statistical learning algorithms that are very important to manipulate and classify the original dynamic data from the subspace representation. The dynamic representing of 3D sequences as trajectories on Grassmann and Stiefel manifolds and how to model spatio-temporal information from these trajectories using distances and velocity vectors are presented in Sect. 3.5. Finally, we conclude the chapter in Sect. 3.6.

3.2 Which data of interest?

In Fig. 3.1, we show an example of 3D sequence acquired by a single-view structured-light 3D scanner with a large field-of-view. One can appreciate the deformations of the 3D scan over time. In addition, the frames present different poses of the body, and include undesirable parts, such as the neck, the shoulders, etc.

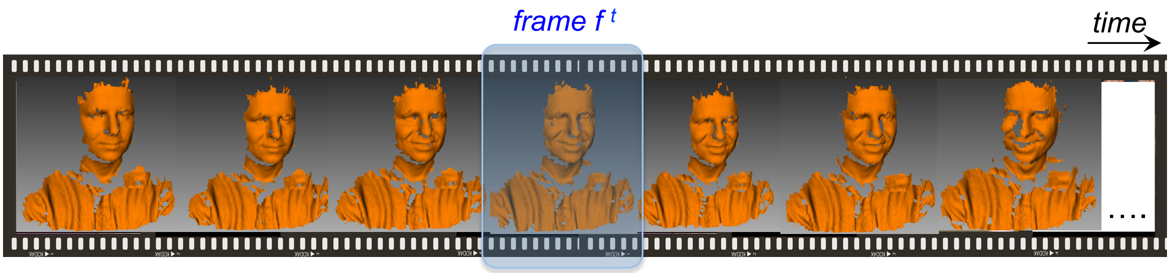

Figure 3.1: Equally-spaced 3D frames of a sample dynamic facial sequence (of the author) conveying a happiness expression. The sequence shows some challenges, such as pose variations, incomplete data and noise.

Actually, from the state-of-the-art, we note the two main categories to model 4D facial scans are: First, 3D feature tracking, which depends basically on the accuracy of detecting landmarks through the video; Second, the 3D deformation of facial scans by comparison with reference model or another scan. These two methodologies can be affected badly when applying them on noisy 3D facial data captured under unconstrained scenarios. Starting from this point and encouraged by the results achieved in 2D dynamic facial analysis, we decided to use the subspace representation to model 3D videos in this work. The compactness of this representation derives from projecting the high dimensional data in low-dimensional representation while keeping the informative part, being able at the same time of discarding noise and compensate missing data. If needed, one can come back from the new compact representation into the original data due to its faithfulness.

Now, let us consider two 3D facial videos V_1, V_2, we want to know for example if they belong to the same person class (for identity recognition) or they convey the same emotion (facial expression recognition). The main question here is: **How can we measure the similarity between these two videos?** This similarity measure is the first step for going further toward classification and statistical learning algorithms. By modeling these two videos as k dimensional linear subspaces $\mathscr{X}, \mathscr{Y}$ on $\mathbb{R}^n$, these subspaces lie naturally in space of linear subspaces, which is a special Riemannian manifold called Grassmann manifold. Over this non-flat manifold, the length of the shortest path between two elements (subspaces) is well defined as a geodesic distance. Several techniques have been developed in the literature in order to find a linear projection of high-dimensional data into a lower finite dimension linear subspace. The main motivation for adopting this representation is its ability to reveal a hidden principle structure of the raw data, compensating for missing parts and discarding noise. Principle Component Analysis (PCA) [68] is one of the most common approaches for dimensionality reduction, and it has been used early for face recognition in the *Eigenfaces* approach [133]. Another data reduction technique related to PCA is the Singular Value Decomposition (SVD). SVD is often used when the informative data are more related to the global structure than the variation, so keeping the mean can be meaningful in these cases whereas it is removed in PCA method.

A great interest has been paid recently to matrix manifolds and their use to solve computer vision problems [88]. Advanced mathematical and statistical learning algorithms have been already defined on these manifolds. Learning approaches solved the problem of non-linearity representation by intrinsic methods that start from the fact these manifolds have a linear structure locally [132] or extrinsic approaches that embed the non-linear manifold into another manifold with a linear structure [59]. The principle of modeling real world data in low-dimensional linear subspaces approved its efficiency in numerous applications, like object recognition from image sets and videos [132], spatio-temporal dynamic system representation [9], image analysis and filtering [134], object tracking [84], etc. More recently, several learning approaches on manifold appeared that address the spatio-dynamic modeling as a trajectory on the manifold, which showed efficient performance on several computer vision applications, like in action classification [9, 16]. The ability to represent a sequence of subspaces as a parameterized trajectory by the time can be an excellent solution for emotional states and complex affects detection from 3D dynamic data.

3.3 Geometry of Grassmann manifolds

The Riemannian manifold by definition is a nonlinear topological structure that has a Euclidean space property locally with a defined metric that can give a similarity measure between two elements on the manifold. Let us have two sets of points A and B in one space, and the relation between their elements is equivalence, i.e., every certain set of points from set A is equivalent to one specific point in set B. This relation defines the group B as a quotient of group A. Following this quotient principle, the geometry of Stiefel $\mathscr{L}_k(\mathbb{R}^n)$ and Grassmann $\mathscr{G}_k(\mathbb{R}^n)$ manifolds will be presented as a quotients of the special orthogonal group $SO(n)$.

3.3.1 Special orthogonal group

The generalized linear group $GL(n)$ of $n \times n$ non-singular matrices forms a differentiable manifold. Even though the differentiable manifold is not a vector space, it can be considered subsets of Euclidean space locally. Later, we will see the importance of this property of local linearity for adapting the Euclidean mathematical and statistical tools to these manifolds. Since the GL(n) is a differentiable manifold and a group at the same time, it forms a Lie Group LG(n). The Special Orthogonal Group $SO(n)$ obtained by considering the subset of orthogonal matrices with determinant $+1$. Thus, $SO(n)$ is a submanifold of $LG(n)$ and keeps Lie Group structure.

The first step towards doing differential calculus on a manifold is to specify the tangent space. For the identity matrix I, which is an element of $SO(n)$, the tangent space $T_I(SO(n))$ is the set of all $n \times n$ skew-symmetric matrices given by:

$$T_I(SO(n)) = \{X \in \mathbb{R}^{n\times n} \mid X + X^T = 0\}. \tag{3.1}$$

Definition 3.3.1. The *Tangent Space* $T_O(SO(n))$ at any point $O \in SO(n)$ is a rotation of the identity matrix tangent space $T_I(SO(n))$, and it is given formally as:

$$T_O(SO(n)) = \{OX | X \in T_I(SO(n))\}. \tag{3.2}$$

After defining the tangent space, let us define an inner product for any $X, Y \in T_O(SO(n))$ where $< X, Y >= tr(XY^T)$ and tr is the sum of the diagonal elements in the matrix, the group $SO(n)$ becomes a Riemannian manifold. Starting from the bi-invariant Riemannian structure obtained, it is possible to measure the length of paths on a manifold.

Definition 3.3.2. Let us have two points $O_1, O_2 \in SO(n)$, the *Riemannian Metric* between these two points can be defined as the infimum of the length of all smooth paths on $SO(n)$, which has O_1 as a beginning and O_2 as an end given by:

$$d(O_1, O_2) = \inf_{\{\alpha:[0,1]\to SO(n) \ |\alpha(0)=O_1, \alpha(1)=O_2\}} \int_0^1 \sqrt{\langle \frac{d\alpha(t)}{dt}, \frac{d\alpha(t)}{dt} \rangle}\, dt. \tag{3.3}$$

The path $\hat{\alpha}$, which achieves the above minimum is a geodesic between O_1 and O_2 on $SO(n)$. This geodesic can be computed from the matrix exponential as well. It is important to highlight that the geodesic here is a constant speed curve defined by its initial velocity and it is different from the geodesic distance, which is a Riemannian distance between two points on the Grassmann manifold.

Definition 3.3.3. Let us have a matrix A of size $n \times n$, the *Matrix Exponential* of A exp(A) can be computed as follows:

$$exp(A) = I + \frac{A}{1!} + \frac{A^2}{2!} + \frac{A^3}{3!} + \dots \tag{3.4}$$

Starting from this equation, it is possible to define geodesics on $SO(n)$ as follows: Let us have an orthonormal matrix $O \in SO(n)$ and any skew-symmetric matrix X, $\alpha(t) = Oexp(tX)$ is the unique geodesic in $SO(n)$ passing through O with velocity vector OX at $t = 0$.

The exponential map is very important for statistics on the manifold, because it allows moving a point from the tangent space to the manifold.

Definition 3.3.4. If M is a Riemannian manifold and $p \in M$, the *Exponential Map* $exp_p : T_p(M) \to M$, is defined by $exp_p(v) = \alpha_v(1)$, where α_v is a geodesic starting at p. In the case of $SO(n)$, the exponential map $exp_O :\ T_O(SO(n)) \to SO(n)$ is given by:

$$exp_O(X) = Oexp(X), \tag{3.5}$$

where the exponential map of O is the multiplication between O and its matrix exponential.

3.3.2 Stiefel manifold

Definition 3.3.5. *Stiefel manifold* is a set of k-dimensional orthonormal bases in $\mathbb{R}^n$ where $1 \le k \le n$.

Since every basis is represented by a matrix of size $n \times k$ with orthonormal columns, this set can be seen as a quotient space of $SO(n)$ as follows: We can consider $SO(n-k)$ as a subgroup with smaller rotations on $SO(n)$ by defining an embedding function $\phi_1 : SO(n-k) \to SO(n)$ as:

$$\phi_1(W) = \begin{bmatrix} I_k & 0 \\ 0 & W \end{bmatrix} \in SO(n). \tag{3.6}$$

Now, we consider $O_1, O_2 \in SO(n)$ to be equivalent, i.e., $O_1 \sim O_2$, *if* $O_1 = O_2\phi_1(W)$ for some $W \in SO(n-k)$, where $\phi_1(SO(n-k))$ represents the rotations of $SO(n)$, which rotates only the last $(n-k)$ components in $\mathbb{R}^n$ and keeping the first (k) without any rotation. Thus, we defined a new equivalence relation between orthogonal matrices of size $n \times n$, where they are identical if the first k columns are identical regardless of the rest $(n-k)$ columns, and this class is given by:

$$[O]_\alpha = \{O\phi_1(W) \mid W \in SO(n-k)\}. \tag{3.7}$$

Since all $[O]_\alpha$ have the same k first columns, we represent all elements of $[O]_\alpha$ by one submatrix $U \in \mathbb{R}^{n\times k}$. So, **Stiefel manifold** of dimension k is the set of these equivalence elements, i.e., a quotient space of the Special Orthogonal group $SO(n)$ and it is given simply by:

$$\mathscr{L}_k(\mathbb{R}^n) = SO(n)/SO(n-k). \tag{3.8}$$

Definition 3.3.6. One possibility to define a *Stiefel metric* between two elements of this manifold is given by the Frobenius norm. Consider two elements of Stiefel manifold $X, Y \in \mathscr{L}_k(\mathbb{R}^n)$. The *Frobenius metric* is defined by:

$$d_{stiefel}(X,Y) = \|X - Y\|_F \tag{3.9}$$

where $\| \,.\, \|_F$ is the standard Frobenius norm, where $\|A\|_F = \sqrt{tr(AA^t)}$.

3.3.3 Grassmann manifolds

Definition 3.3.7. The *Grassmann manifold* is the set of all k-dimensional subspaces of $\mathbb{R}^n$. Since that every $n \times k$ orthonormal matrix and all its rotations on $SO(n)$, that make different element of Stiefel manifold, represent the same subspace on Grassmann manifold.

To define a structure of quotient space for Stiefel manifold $\mathscr{L}_k(\mathbb{R}^n)$, let us consider $S(k) \times S(n-k)$ as a subgroup of $SO(n)$ defined by the function $\phi_2 : (SO(k) \times SO(n-k) \to SO(n)$ as:

$$\phi_2(W_1, W_2) = \begin{bmatrix} W_1 & 0 \\ 0 & W_2 \end{bmatrix} \in SO(n). \tag{3.10}$$

$O_1 \sim O_2$ if $O_1 = O_2\phi_2(W_1, W_2)$ for some $W_1 \in SO(k)$ and $W_2 \in SO(n-k)$ O_1 and O_2 are equivalent if the first k columns of O_1 are rotations of the first k columns of O_2 and the same for the rest $(n-k)$ columns. An equivalence class is given by:

$$[O]_\beta = \{O\phi_2(W_1, W_2) \mid W_1 \in SO(k),\ W_2 \in SO(n-k)\}. \tag{3.11}$$

Then, the set of all these equivalence classes form the Grassmann manifold $\mathscr{G}_k(\mathbb{R}^n)$ and it can be given formally as a quotient space of Special Orthogonal Group SO(n):

$$\mathscr{G}_k(\mathbb{R}^n) = SO(n)/(SO(k) \times SO(n-k)). \tag{3.12}$$

Consequently, it is a quotient space of Stiefel manifold $\mathscr{L}_k(\mathbb{R}^n)$:

$$\mathscr{G}_k(\mathbb{R}^n) = \mathscr{L}_k(\mathbb{R}^n)/SO(k). \tag{3.13}$$

From this definition for the Grassmann manifold, our adopted representation of the 3D dynamic facial sequence of m frames lies naturally on these two manifolds. This achieved after applying dimension-reduction technique on the original data, like k-singular value decomposition $(k-SVD)$.

The main motivation for dealing with Grassmann manifold as a quotient space of the special orthogonal group $SO(n)$ is that it allows us to inherit systematically the well defined geodesics and tangent planes of the $SO(n)$.

Definition 3.3.8. The *Tangent Space* of a Grassmann manifold $\mathscr{G}_k(\mathbb{R}^n)$ can be induced directly from the tangent space of the $SO(n)$ since it is a quotient space of it as follows:

Let us have M/L as a quotient space of M under the action of a group $L \subset M$. Now, for any point $p \in M$, a vector $v \in T_p(M)$ can be considered as tangent to M/L, since it is perpendicular to the tangent space $T_p(pL))$ where $T_p(pL)$ is a subspace of $T_p(M)$. Following the same principle, we define the tangent space of $\mathscr{G}_k(\mathbb{R}^n)$, while $M = SO(n)$ and $L = \phi_2(SO(k) \times SO(n-k))$ with ϕ_2 given in Eq. 3.10.

The tangent space $T_I(L)$ is considered as a subspace of $T_I(SO(n))$ by defining the embedding function ϕ_T:

$$\phi_T(A_1, A_2) = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \in T_I(SO(n)). \tag{3.14}$$

The tangent vectors to $SO(n)$ and perpendicular to the space $(T_{I_k}(SO(k)) \times T_{I(n-k)(SO(n-k))})$ can be considered the tangent of $\mathscr{G}_k(\mathbb{R}^n)$ after multiplication on right by matrix $J \in \mathbb{R}^{n\times k}$, which includes the first k columns of $I_n \in \mathbb{R}^{n\times n}$. The tangent space at $[J]$ is given by:

$$T_{[J]} = \{ \begin{bmatrix} 0 \\ B^T \end{bmatrix} \setminus B \in \mathbb{R}^{k\times(n-k)}\} \tag{3.15}$$

If we have $[U] \in \mathscr{G}_k(\mathbb{R}^n)$, and $O \in SO(n)$, then $U = O^T J$. The tangent space at $[U]$ is given by:

$$T_{[U]}(\mathscr{G}_d(\mathbb{R}^n)) = \{O^T G \setminus G \in T_{[J]}(\mathscr{G}_k(\mathbb{R}^n))\} \tag{3.16}$$

3.3.4 Exponential and logarithm map on Grassmann manifolds

Since the Grassmann manifold is a quotient space of special orthogonal group $SO(n)$, it inherits the definition of exponential map that projects a point from the manifold into the tangent vector space and its inverse, the logarithm map, that returns the point from the tangent space to the manifold. These two algorithms are essentials to solve statistical learning and optimization problems on Grassmann manifold by intrinsic manner. In [50], Gallivan et al. presented efficient computational methods to implement these two algorithms.

Definition 3.3.9. Let us have two subspaces $\mathscr{X}_1, \mathscr{X}_2 \in \mathscr{G}_k(\mathbb{R}^n)$ represented by two matrices X_1, X_2 of size $n \times k$. We need a method to calculate the velocity parameter V that travels from $\mathscr{X}_1$ to $\mathscr{X}_2$ in the unit time called the *velocity matrix*.

The algorithm proposed by Gallivan et al. in [50] to compute this structure is given by:

1. Compute the $n \times n$ orthogonal completion Q of X_1.
2. Compute the thin decomposition of $Q^T X_2$ given by:
$Q^T X_1 = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix} = \begin{bmatrix} \Gamma(1) \\ \Sigma(1) \end{bmatrix} V_1^T.$
3. Compute the angles given by the *arcsin* and *arcos* of the diagonal elements of Γ and Σ respectively. Form the diagonal matrix Θ containing θs on its diagonal,
4. Compute $V = M_2 \Theta M_1$.

Definition 3.3.10. Let us have a $\mathscr{X} \in \mathscr{G}_k(\mathbb{R}^n)$, which is represented by an orthogonal matrix X of size $n \times k$ with a direction matrix $A \in \mathbb{R}^{(n-k)\times k}$ that gives the direction of the geodesic flow. The geodesic path $\beta(t)$ of X at each time instance (t) is given by: $\beta(t) = Q exp(tA)J$ Where $Q \in SO(n)$ and $Q^T X = J$ and $J = [I_k; O_{n-k,k}]$ called a *moving geodesic.*

The main steps to sample the geodesic path $\beta(t)$ presented in [50] are:

1. computing the completion matrix of X, Q of size $n \times n$ by QR decomposition of X.
2. Apply SVD to decompose the direction matrix $A = USV^T$.
3. Compute the diagonal matrices $\Gamma(t)$ and $\Sigma(t)$ of size $k \times k$ from diagonal elements of S, such that $\gamma_i(t) = cos(t\theta_i)$ and $\sigma_i(t) = sin(t\theta_i)$, where Θ is the diagonal elements of S (the principle angles).
4. $\beta(t) = \begin{bmatrix} U\Gamma(t) \\ -V^T\Sigma(t) \end{bmatrix}$ for various values of $t \in [0,1]$.

To illustrate these algorithms on Grassmann, let us have μ as an element of $\mathcal{G}_k(\mathbb{R}^n)$, the tangent space defined on the manifold at this point is T_μ. Using the logarithm map, we can project point $X_1 \in \mathcal{G}_k(\mathbb{R}^n)$ to the vector space T_μ to have V_1 tangent vector. This operation can be defined as:

$$log_\mu : \mathcal{G}_k(\mathbb{R}^n) \to T_\mu(\mathcal{G}_k(\mathbb{R}^n)) \tag{3.17}$$

Also, we can project V_2 from the vector space to the Grassmann manifold to have X_2 element using the inverse operation (exponential map), which is given by:

$$exp_\mu : T_\mu(\mathcal{G}_k(\mathbb{R}^n)) \to \mathcal{G}_k(\mathbb{R}^n) \tag{3.18}$$

Fig. 3.2 depicts these ideas on the Grassmann manifold.

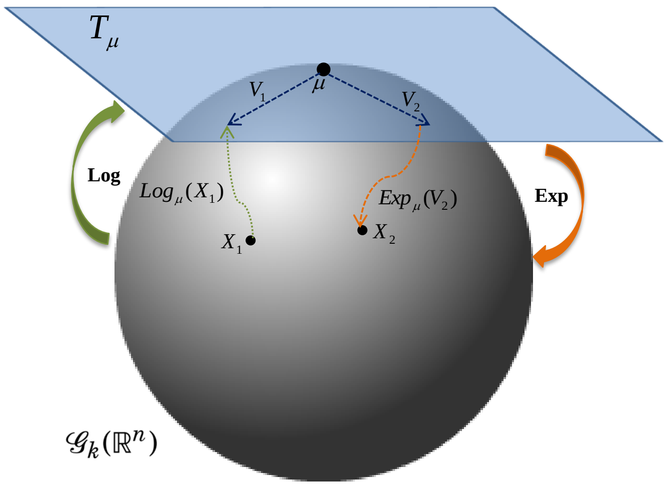

Figure 3.2: Illustration of a tangent plane at point μ and tangent vectors with their map to the Grassmann manifold with Exponential and Logarithm map functions.

3.3.5 Distances on Grassmann manifolds

The idea of using the Grassmann manifold representation is that a subsequence of 3D or depth scans can be cast to a matrix representation, and thus mapped to a unique point on the manifold. In this way, computing the similarity between two subsequences is transformed to the problem of computing a Riemannian distance between two points on the manifold.

It is important to differentiate between *distance* and *metric* terms on Grassmann. The term *distance* is used to refer to similarity measure between two subspaces, which has a non-negative value and invariant to any rotation of the subspace basis.

Definition 3.3.11. Let us have a function $d : \mathcal{G}_k(\mathbb{R}^n) \times \mathcal{G}_k(\mathbb{R}^n) \to \mathbb{R}$, d is a *Grassmann Distance* if $d(\mathcal{X},\mathcal{Y}) = d(\mathcal{X}R_1,\mathcal{Y}R_2), \quad \forall R_1,R_2 \in SO(k)$.

The *metric* is a distance, but it should satisfy the following conditions for any $\mathcal{X}_1,\mathcal{X}_2,\mathcal{X}_3 \in \mathcal{G}_k(\mathbb{R}^n)$

1. $d(\mathcal{X}_1,\mathcal{X}_2) \geq 0$,
2. $d(\mathcal{X}_1,\mathcal{X}_2) = 0$ if and only if $\mathcal{X}_1 = \mathcal{X}_2$,
3. $d(\mathcal{X}_1,\mathcal{X}_2) = d(\mathcal{X}_2,\mathcal{X}_1)$,
4. $d(\mathcal{X}_1,\mathcal{X}_2) \leq d(\mathcal{X}_2,\mathcal{X}_3) + d(\mathcal{X}_1,\mathcal{X}_3)$.

More specifically, let $\mathcal{X}$, $\mathcal{Y}$ denote a pair of subspaces of dimension k on $\mathcal{G}_k(\mathbb{R}^n)$. The Riemannian distance between $\mathcal{X}$ and $\mathcal{Y}$ is the length of the shortest path connecting the two points on the manifold (i.e., the geodesic distance). The problem of computing this distance can be solved using the notion of *Principle Angles or Canonical Correlation*, introduced by Golub and Loan [52] as an intuitive and computationally efficient way for defining the distance between two linear subspaces.

In fact, there is a set of principal angles $\Theta = [\theta_1,\ldots,\theta_k]$ $(0 \leq \theta_1,\ldots,\theta_k \leq \pi/2)$, between the subspaces $\mathcal{X}$ and $\mathcal{Y}$ (see Fig. 3.3), recursively defined as follows:

$$\theta_k = \cos^{-1}\left(\max_{u_k \in \mathcal{X}} \max_{v_k \in \mathcal{Y}} \langle u_k^T, v_k \rangle\right), \tag{3.19}$$

where u_k and v_k are the vectors of the basis spanning, respectively, the subspaces $\mathcal{X}$ and $\mathcal{Y}$, subject to the additional constraints: $\langle u_k^T, u_k \rangle = \langle v_k^T, v_k \rangle = 1$, being $\langle .,. \rangle$ the inner product in $\mathbb{R}^n$; and $\langle u_k^T, u_i \rangle = \langle v_k^T, v_i \rangle = 0$ $(\forall\, k,i : k \neq i)$.

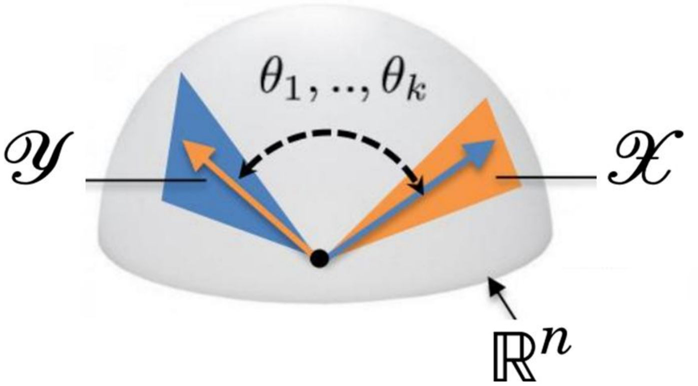

Figure 3.3: Principal angles $\Theta = [\theta_1, .., \theta_k]$ computed between two linear subspaces $\mathscr{X}$ and $\mathscr{Y}$ of the Grassmannian $\mathscr{G}_k(\mathbb{R}^n)$.

In other words, the first principal angle θ_1 is the smallest angle between all pairs of unit basis vectors in the two subspaces and the *cosine* of the first principle angles is the first canonical correlation. The k^{th} principal angle and canonical correlation are defined in a similar manner. Based on the definition of the principal angles, the geodesic distance between $\mathscr{X}$ and $\mathscr{Y}$ can be defined as [42]:

$$d^2_{Geo}(\mathscr{X}, \mathscr{Y}) = \|\Theta\|_2 = \sum_i^k \theta_i^2 \,. \tag{3.20}$$

Accordingly, the geodesic distance could be interpreted as the magnitude of the smallest rotation that takes $\mathscr{X}$ to $\mathscr{Y}$. Given the matrices X, Y, where $\mathscr{X} = Span(X)$ and $\mathscr{Y} = Span(Y)$, the principle angles can be computed by applying SVD on the matrix $X^T Y$ as follows:

$$X^T Y = \mathscr{U}(\cos\Theta)\mathscr{V}^T, \tag{3.21}$$

where $\mathscr{U} = [u_1, \dots, u_k]$, $\mathscr{V} = [v_1, \dots, v_k]$, and $\cos\Theta = diag(\cos\theta_1, \dots, \cos\theta_k)$. The principle angles are ordered in non-decreasing form as follows:

$$0 \leq \theta_1 \leq ...\theta_k \leq \pi/2.$$

consequently, the canonical correlation is in non-increasing order:

$$1 \geq \cos\theta_1 \geq ... \cos\theta_k \geq 0.$$

This distance is used to measure the similarity between two linear subspaces, even though with two different dimensions, permitting to smooth the effect of noisy data, at the same time showing robustness with respect to acquisition variations.

Based on the notion of principle angles, several other distances and metrics on the Grassmann manifold were proposed in the literature. The most used distances and metrics are given below with a discussion about their different geometrical meaning.

Projection metric: It is defined as the l_2 norm of sin of the principle angles between two subspaces:

$$d^2_{proj}(\mathscr{X},\mathscr{Y}) = \sum_{i=1}^{k} \sin(\theta_i)^2 = k - \sum_{i=1}^{k} \cos^2(\theta_i)\,. \tag{3.22}$$

This distance can be computed easily from the product of $X^T Y$. From equation.3.21, the relation between SVD and $X^T Y$ we can get:

$$d^2_{proj}(\mathscr{X},\mathscr{Y}) = k - \sum_{i=1}^{k} \cos^2(\theta_i) = k - ||X^T X - Y^T Y||^2_F\,, \tag{3.23}$$

where $||.||^2_F$ is the Frobenius norm on the matrix.

This Projection distance is a Grassmann distance because it is invariant to different representations and it is a metric as well.

Binet-Cauchy distance: It is defined as a function of the product of canonical correlations:

$$d_{BC}(\mathscr{X},\mathscr{Y}) = (1 - \prod_{i}^{k} \cos^2\theta_i)^{1/2}\,. \tag{3.24}$$

It is computed from from the SVD of $X^T Y$ as:

$$d^2_{BC}(\mathscr{X},\mathscr{Y}) = 1 - \prod_{i}^{k} \cos^2\theta_i = 1 - det(X^T Y)^2\,, \tag{3.25}$$

This distance is a Grassmann distance and a metric as well.

Max Correlation: It is based on using only the smallest principle angle θ_1, which gives the largest canonical correlation as:

$$d_{Max}(\mathscr{X},\mathscr{Y}) = (1 - \cos^2\theta_1)^{1/2} = \sin\theta_1\,. \tag{3.26}$$

It is a Grassmann distance but not a metric since it can be 0 even though the two subspaces are not the same, so this can be a limitation for its use.

Min Correlation: It is the opposite of *Max Correlation*, where it is based on only the largest principle angle θ_k, which gives the lowest canonical correlation.

$$d_{Min}(\mathscr{X},\mathscr{Y}) = (1 - \cos^2\theta_k)^{1/2} = \sin\theta_k\,. \tag{3.27}$$

It can also be rewritten as:

$$d_{Min}(\mathscr{X},\mathscr{Y}) = ||X^T X - Y^T Y||_2, \tag{3.28}$$

where $||.||_2$ is the matrix l_2 norm given by:

$$||A||_2 = max_{x\neq 0}\frac{||Ax||_2}{||x||_2}, \quad A \in \mathbb{R}^{m\times n}.$$

This distance is a Grassmann distance and satisfies the metric conditions.

Procrustes distance: It is defined as the minimum distance between all possible subspaces spanned by two bases as:

$$d_{Proc}(\mathscr{X},\mathscr{Y}) = 2\left(\sum_{i=1}^{k} \sin(\theta_i/2)\right)^{1/2}. \tag{3.29}$$

It can also defined as:

$$d_{Proc}(\mathscr{X},\mathscr{Y}) = \min_{R_1,R_2\in\mathbb{O}()} ||XR_1 - YR_2||_F. \tag{3.30}$$

By definition, the Procrustes distance is invariant under different representations and furthermore is a valid metric.

The selection of the best distance for an application depends mainly on the data nature. For example, *Max Correlation* can be a good choice when the subspaces are scattered, and the data is noisy, and then we can depend on the largest canonical correlation only. The *Min Correlation* gives an opposite performance, since it uses the smallest canonical correlation. Thus, it can be a good choice when the subspaces are very close to each other and there is a slight difference among them. *Binet-Cauchy distance* performance is close to *Min Correlation* since it seeks for the smallest possible distance even though it uses all principle angles. Distances like Geodesic, Projection and Procrustes, give intermediate performance between the Max and the Min correlation distances. An experimental analysis for all of these metrics on 4D face recognition problems will be presented for 4D face recognition problem in the next Chapter 4.

These measures capture different aspects of the distance on the manifold and can help to explore the data distribution in the subspace represented by the singular vectors for recognition and classification tasks.

3.4 Statistical learning on Grassmann manifolds

The subspace representation of 3D dynamic sequences as elements on Grassmann manifold and how to measure similarity using different distances and metrics have been introduced in the previous section. Now, the most important concept is how statistical learning approaches can be adapted to work properly on such non-linear structure in order to combine advantages of subspace modeling with the statistical learning tools. There are two main directions for statistical learning on Grassmann manifold in the literature:

Intrinsic Method – This method relies on the basic idea of mapping the points of the Grassmann manifold into a fixed tangent space using the logarithm map function (i.e., a vector space) [27, 141]. The main constraint of this method is the computation of logarithm map function, which does not have an explicit formula in the case of Grassmann manifolds. This makes its estimation numerically not too accurate, especially for the points far from the tangent space position and also it is time consuming. We will discuss basic intrinsic methods like Karcher mean and k-means learning on Grassmann later on.

Extrinsic Method – To avoid intrinsic method limitations, this method consists to embed the Grassmann manifold into a larger Euclidean space by predefined projection mapping function, like in [124] and [136]. Here the computation is relatively simple by comparison to intrinsic but the non-uniqueness embedding solution can lead to non-uniqueness of statistics. The adaptation of the well-known dictionary learning and sparse coding on Euclidean to work properly on non-flat Grassmann manifold will be presented.

The implementations of these two types of learning on Grassmann with experimental analysis on face recognition from 4D data are presented in the next Chapter.

3.4.1 Sample (Karcher) mean computation

As mentioned above, an important tool in shape (and its temporal evolution) analysis is given by the computation of statistical summaries. For a set of given subspaces $\mathbb{P} = \{\mathscr{P}_i\}_{i=1}^{m}$, where $\mathscr{P}_i \in \mathscr{G}_k(\mathbb{R}^n)$ (i.e., points on the underlying manifold), a sample mean μ is a point on the Grassmannian, which minimizes the mean squared error [73] with respect to the canonical metric d_{Geo} previously defined in Eq. 3.20.

This algorithm starts by initializing the mean to the first subspace in the set initially, then it uses the *Log Map* algorithm to project all $\mathbb{P}$ elements on the tangent space of the current mean point as depicted in Fig. 3.4. Then, computing the average vector from all tangent vectors of the data points. The current mean moved in the direction of the average vector by a certain step to have the new mean after projecting it back on the manifold by using *Exp Map* algorithm. This loop is repeated till the convergence of the

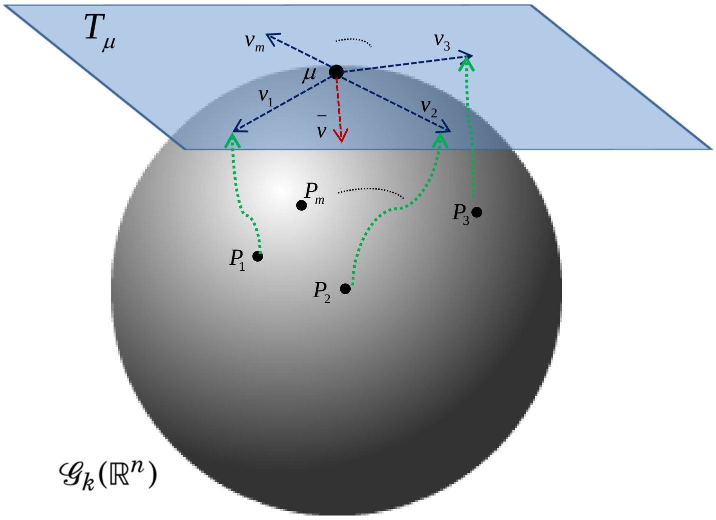

Figure 3.4: Estimation of a Karcher mean of a set of Grassmann elements.

average vector norm to a predefined value. The steps to compute μ are summarized in Algorithm 1.

Algorithm 1 – Mean Sample Estimation over $\mathscr{G}_k(\mathbb{R}^n)$

Require: $\mathbb{P} = \{\mathscr{P}_i\}_{i=1}^m, \; where \;\; \mathscr{P}_i \in \mathscr{G}_k(\mathbb{R}^n), \;\; \epsilon > 0$ typically $\epsilon = 0.5$; τ : Threshold value

Initialize $\mu_0 \leftarrow \mathscr{P}_0$, $i \leftarrow 0$

repeat

Compute $v_i \leftarrow exp_{\mu_i}^{-1}(\mathscr{P}_j)$ for $j = 0, \ldots, m$

Compute the average tangent vector $\bar{v} \leftarrow \frac{1}{m} \sum v_i$

Move μ_i according to $\mu_{i+1} \leftarrow exp_{\mu_i}(\epsilon \bar{v})$

$i \leftarrow i + 1$

until $(||\bar{v}|| \leq \tau)$

Ensure: μ the estimated mean of $\mathbb{P}$ set

3.4.2 Grassmann k-means algorithm

Karcher mean is an efficient statistical tool on Grassmann manifold, where more important learning algorithm can be based on it. The K-means unsupervised learning algorithm defined on Euclidean vector space can be extended to address the non-linear structure of Grassmann manifold depending on Karcher mean. Let us have a set of m subspaces $\mathbb{P} = \{\mathscr{P}_i\}_{i=1}^m$ on Grassmann manifold. It is required to group these subspaces in N classes according to their similarity measure by finding the mean of them $(\mu_1, \mu_2, ..., \mu_N)$. The same expectation Minimization EM-algorithm used in Euclidean k-means is used here on minimizing the geodesic distances squares. First, an assignment of classes means is done randomly from the subspaces set. Every subspace will be assigned to the nearest class center in Expectation step, and the Karcher mean is computed for every class members in Minimization step. These two steps are repeated a certain number of times, which should be predefined according to the nature of the data. These steps are summarized in Algorithm 2.

Algorithm 2 – K-means clustering on $\mathscr{G}_k(\mathbb{R}^n)$

Require: $\mathbb{P} = \{\mathscr{P}_i\}_{i=1}^m$, where $\mathscr{P}_i \in \mathscr{G}_k(\mathbb{R}^n)$, , N: Number of classes, M: Max number of iterations

Initialize the classes center randomly $(\mu_1^0, \mu_2^0, ..., \mu_N^0)$, $i \leftarrow 0$

repeat

Compute the distance between $\mathbb{P}$ members and cluster centers

Assign every $\mathscr{P}_i$ the closest cluster

Re-computer clusters centers $(\mu_1^i, \mu_2^i, ..., \mu_N^i)$ using Algorithm 1

$i \leftarrow i+1$

until $(j = M)$

Ensure: The N cluster centers: $(\mu_1^M, \mu_2^M, ..., \mu_N^M)$

3.4.3 Sparse coding and dictionary learning

Recently, the sparse coding and dictionary learning showed a great success in several related topics like signal processing [142], image classification [51, 160] and face recognition [140, 145], where a given signal or image can be approximated effectively as a combination of few members (atoms) of a learned dictionary. The success of sparse coding in several computer vision problems motivated to extend this learning approach from vector space to nonlinear manifolds, like Grassmann [50, 132], in order to represent a subspace as the combination of few subspaces of a dictionary. However, in so doing, the main issue is the non-linearity of the Grassmann manifold, which implies using tools from differential geometry. Since this often requires intensive computation, these solutions are less attractive for 2D and 3D video modeling and analysis.

The problem of *sparse coding* has been solved in $\mathbb{R}^n$ Euclidean space by minimizing the following quantity, which includes a coding cost function with a penalty term related to the sparsity of the result:

$$l(x,\mathcal{D}) = \min_y \|x - \mathcal{D}y\|_2^2 + \lambda\|y\|_1 \tag{3.31}$$

where $x \in \mathbb{R}^n$ is the sample signal to be coded, $\mathcal{D}$ is a dictionary (a $n \times N$ matrix being N the number of training samples) with atoms $D_i \in \mathbb{R}^n$ in its columns, and λ the sparse regularization parameter. The vector $y \in \mathbb{R}^N$ is the new latent sparse representation of the original data, which contains many zeros. The problem of *dictionary learning* consists of minimizing the total coding cost for all the samples $\{x^t \in \mathbb{R}^n\}_{1 \le t \le N}$ of the training set, over all choices of codes and dictionaries as follows:

$$h(\mathcal{D}) = \min_{\{x^t\},\mathcal{D}} \frac{1}{N} \sum_{t=1}^{N} l(x^t,\mathcal{D}). \tag{3.32}$$

In order to combine advantages of subspace modeling with the powerful sparse coding representation, it is essential to handle the non-linearity of the Grassmann manifold. An *extrinsic method* consists to embed the Grassmann manifolds into a smooth sub-manifold of the space of symmetric matrices [146], as will be adopted in this work. This embedding is performed by a projection mapping function already used in [124] and [136].

Formally, let's have a set of points, for example subspaces that represent 3D dynamic facial sequences in this work, $\mathbb{X} = \{X_i\}_{i=1}^m$, where $Span\{X_i\} \in \mathcal{G}_k(\mathbb{R}^n)$. We need to be able of representing each point (subspace) as a linear combination of a few atoms of a dictionary of subspaces $\mathbb{D} = \{D_1, D_2, \ldots, D_j\}$ using the sparse coding technique.

For any $\mathcal{X} = Span(X) \in \mathcal{G}_k(\mathbb{R}^n)$ the mapping $\partial : \mathcal{G}_k(\mathbb{R}^n) \to Sym(\mathbf{n})$, such that $\partial(\mathcal{X}) = XX^T = \hat{X}$ is computed.

The mapping function ∂ is isometric, as it preserves the curve length between the Grassmann manifold and the manifold of Symmetric matrices $Sym(\mathbf{n})$ [62]. A natural choice of metric on the manifold of symmetric matrices $Sym(\mathbf{n})$ is the Frobenius inner product. For any $Span(X)$, $Span(Y) \in \mathcal{G}_k(\mathbb{R}^n)$, $Frobenius(X,Y) = Tr(\hat{X},\hat{Y}) = \|X^TY\|_F^2$. With this embedding, Eq. (3.31) can be rewritten by considering the embedding $\hat{X}$ of a given query subspace $\mathcal{X}$:

$$l(\mathcal{X},\mathcal{D}) = \min_y \|\hat{X} - \hat{\mathcal{D}}y\|_F^2 + \lambda\|y\|_1, \tag{3.33}$$

where $\hat{\mathcal{D}}$ denotes the dictionary with atoms elements of $Sym(\mathbf{n})$ and y the sparse representation. This convex optimization problem is solvable as a vectorized sparse coding problem, as depicted in Algorithm 3.

Algorithm 3 – Sparse Coding on $\mathcal{G}_k(\mathbb{R}^n)$

Require: A given dictionary $\mathbb{D} = \{\mathcal{D}_i\}_{i=1}^N \in \mathcal{G}_k(\mathbb{R}^n)$ where $\mathcal{D}_i = Span(D_i)$ of size N. Query subspace $\mathcal{X} \in \mathcal{G}_k(\mathbb{R}^n) = Span(X)$

for $i, j \leftarrow 1$ to N **do**

$\mathbb{K}(\mathbb{D})_{i,j} \leftarrow \|D_i^T D_j\|_F^2$

end for

$\mathbb{K}(\mathbb{D})_{N\times N} = U\Sigma U^T$

$A = \Sigma^{1/2} U^T$

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

$\mathcal{K}(X, \mathbb{D})_i \leftarrow \|X^T D_i\|_F^2$

end for

$x^* \leftarrow \Sigma^{-1/2} U^T \mathcal{K}(X, D)$

Ensure: $y^* \leftarrow arg\ \min_y \|x^* - Ay\|^2 + \lambda \|y\|_1$

In Algorithm 3, the training set of (labeled) subspaces is considered as the dictionary $\mathcal{D}$ of size N (i.e., the training set size); (i) A similarity matrix between dictionary elements $\mathbb{K}(\mathcal{D})$ is computed based on the Frobenius inner product; (ii) Singular Value Decomposition (SVD) is applied to $\mathbb{K}$ (i.e., $\mathbb{K} = U\Sigma V^T$) to compute the A matrix, which is the weighted singular vectors of $\mathbb{K}$; (iii) The similarity matrix $\mathcal{K}(X, \mathcal{D})$ between testing and training samples is computed on the induced space. The decomposition of Eq. (3.33) shows that the sparse coding problem can be formulated as:

$$l(\mathcal{X}, \mathcal{D}) = \min_y \|x^* - Ay\|^2 + \lambda \|y\|_1 , \tag{3.34}$$

where $x^* = \Sigma^{-1/2} U^T \mathcal{K}(X, \mathcal{D})$.

We can see that this algorithm ends up by representing every subspace by a linear feature vector called a sparse code. This sparse code allows us to reconstruct the related subspace from a dictionary of subspaces. Thus, we are in a Euclidean space and several learning and classification algorithms will be available to classify this new linear representation of the subspace as will be discussed in the next Chapter 4.

3.5 Trajectories on Riemannian manifolds

In the previous section, we discussed the subspace representation of 3D dynamic facial sequences and its ability to capture the global structure and the variation over time of the dynamic face. In some cases, the 3D dynamic video is divided into shorter subsequences and every one is modeled as a separate subspace to overcome some problems, like pose variation or high variability in facial surface. The statistical tools could be applied to this multiple-instances representation, like Karcher mean, k-means clustering and sparse coding are useful if the order of the subsequences is not important, like in face recognition problem. In other cases, the important information is not only in the subsequences, but also in the temporal evolution of the facial data over time from on subspace to another. This temporal information can be captured from the difference between successive subspaces that belong to the same video. For example, in the case of emotional state that is conveyed through a complete video. Here, keeping the order of subsequences and the ability to extract difference between ordered subspaces is very important to obtain the spatio-temporal description of this emotional state. Now, it is important to define: **How can we capture the spatio-temporal information conveyed through the complete 3D video that is represented as a set of subspaces?** The proposed solution in this work is by considering the set of subspaces as a parametrized trajectory on Riemannian manifold by time. Here, every subspace represents an instance (t). Considering such trajectory of subspaces gives us the ability to measure and capture the temporal evolution through time between neighboring subspaces of the trajectory or according to a reference subspace. The concept of time-parametrized curves (trajectories) analysis on Riemannian manifold introduced in [69] and applied to several computer vision problems, like action recognition [9] and 3D action recognition [16]. In the latter paper, Ben Amor et al. have addressed the problem of action/activity recognition from skeletal data (acquired using Kinect-like cameras). They have proposed a suite of geometric tools for processing static and dynamic shapes as elements and trajectories in the well-known Kendall shape space (which provides invariance to scale, translation and rotation), respectively. The main ingredient introduced in [16] is an elastic metric for aligning pairwise (or multiple) trajectories.

3.5.1 Trajectories on Grassmann manifolds

As far as Grassmann trajectories are concerned in the present study, let $t \longmapsto \mathscr{T}(t)$ be a parameterized curve on $\mathscr{G}_k(\mathbb{R}^n)$, and $V(t)$ the velocity (tangent) vector following the geodesic path between $\mathscr{X}(t)$ and $\mathscr{X}(t+\delta)$. The tangent vector is an element of $T_{\mathscr{X}(t)}(\mathscr{G}_k(\mathbb{R}^n))$. Note that the parameter t denotes the time in our target application as follows. If $[f^0,\dots,f^s]$ denotes a 3D sequence acquired in the time interval $[0,s]$, consequently, the underlying trajectory represents the full (or partial) available time-space observations in the same time-interval. This provides a precise mathematical representation of trajectories on the Grassmannian, and allows deriving interesting quantities to analyze flows of 3D or depth sequences for human emotion detection as it will be investigated in Chapter 5 for early detection of spontaneous emotional states. If needed, one can define the space of trajectories easily by $\mathscr{G}_k(\mathbb{R}^n)^{[0,s]}$, and extend the distance definition of the Grassmannian to this space by integrating d_{Geo} (Eq. (3.20)) over the parameter interval $[0,s]$. This is actually a proper distance between trajectories defined on the Grassmann manifold.

After solving the problem of representing, we need to define a mapping function ζ as follows:

Definition 3.5.1. For any $\mathscr{X}_1, \mathscr{X}_2 \in \mathscr{G}_k(\mathbb{R}^n)$, the mapping $\zeta : \mathscr{G}_k(\mathbb{R}^n) \times \mathscr{G}_k(\mathbb{R}^n) \rightarrow \mathbb{R}^m$ such that $\zeta(\mathscr{X}_1, \mathscr{X}_2) = Z$ where $Z \in \mathbb{R}^m$ and $m << n$.

Scanning the trajectory $\mathscr{T}(t)$ through time t using this $\zeta(t)$ function and concatenating the feature over time results in the final spatio-temporal feature vector of the 3D dynamic video in Euclidean space of $\mathbb{R}^{m\times s}$, where s is the size of $\mathscr{T}(t)$. Thus, we can implement Euclidean classification methods, like the Structured Output Support Vector Machine (SO-SVM) for the sequential analysis and classification of such features by the time as will be addressed in Chapter 5. Figure 3.5 illustrates this mapping function of time parametrized trajectories on Riemannian manifold into Euclidean space.

In this work, two methods to define ζ will be presented: the first depending on the instantaneous speed between trajectory elements, and the second depending on computing the velocity vector between the trajectory elements to capture more information than the speed.

3.5.2 Instantaneous speed along trajectories

One intuitive alternative to analyze trajectories on Stiefel or Grassmann manifolds is to consider the evolution of their instantaneous speed. In particular, given an observed portion of the trajectory in the time interval $[0,t]$, the instantaneous speed can be computed as the distance between neighboring points $\mathscr{X}(t)$ and $\mathscr{X}(t+\delta)$ along the trajectory. In this case, the function ζ substituted by Geodesic distance (d_{Geo}) on Grassmann manifold and the Stiefel distance (the Frobenius norm) on Stiefel manifold with parameter δ as a constant integer, $\delta = \{1,2,3\ldots\}$. These distances can be concatenated in a one-dimensional vector characterizing the temporal evolution along the trajectories.

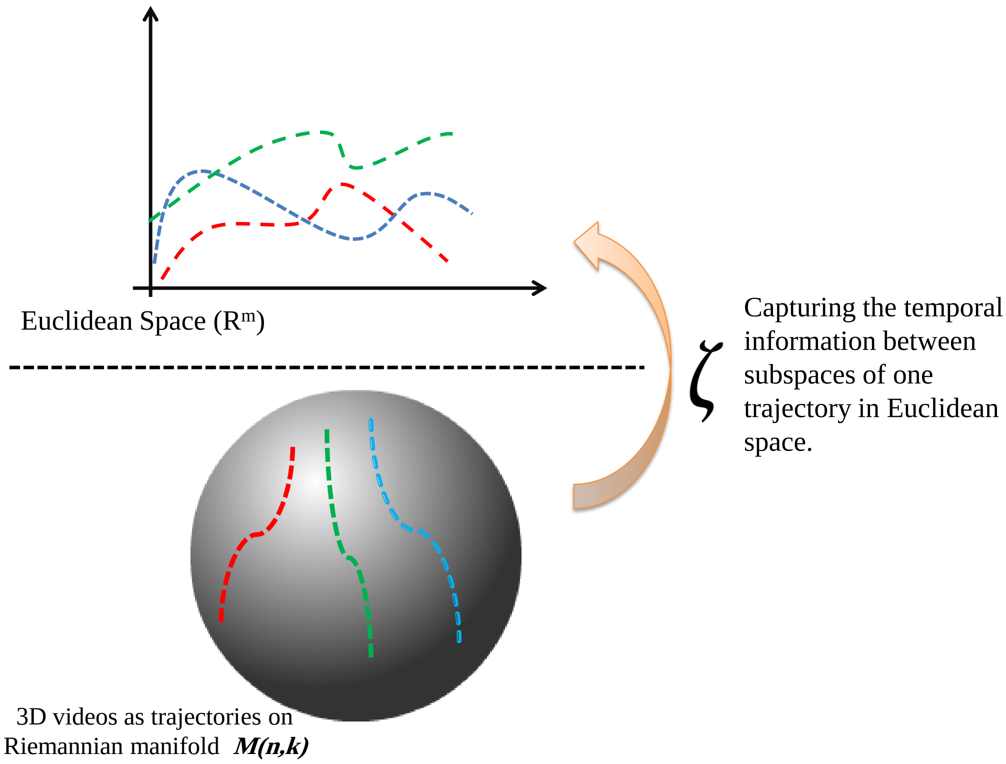

Figure 3.5: Illustration of ζ function and how to capture the spatio-temporal Euclidean feature vector from parametrized trajectory on Riemannian manifold.

One can view this quantity (geodesic distance between subspaces of the same trajectory) as the norm of the shooting (initial velocity) vector between subspaces. Thus, the feature vector of instantaneous speed along the trajectories captures the rhythm (temporal) and amplitude (spatial) of the facial deformations, which could be of great interest for emotion detection. However, this quantity is limited to study the amplitude of the deformation (as a single scalar) for each frame. A natural way to get more complete idea about the (spatial) deformations is to use the velocity vector itself (instead of its norm). Next section provides a detailed description of the velocity vector and its use in physical pain detection will be presented in Chapter 5.

3.5.3 Transported velocity vector fields of trajectories

The quantities presented in the previous approach allow us to quantify the motion's amplitude and the temporal rhythm along the trajectories defined on Riemannian manifolds like Grassmann and Stiefel. To show how the full motion information (face deformation/body and head gestures) one should look at the fields of velocity vectors instead of their norms, along the trajectories on Grassmann manifolds. However, these velocity vectors belong to different tangent spaces ($V(t) \in T_{\mathscr{X}}(\mathscr{G}_k(\mathbb{R}^n))$). One possible solution to this issue is to translate the velocity vector fields to the same and fixed tangent space (e.g., the identity tangent space $\mathscr{I} = span(\begin{bmatrix} I_k \\ 0 \end{bmatrix})$ which is given: $\mathscr{T}_i(t)$

Definition 3.5.2. Let us have a trajectory of subspaces $t \leftarrow \mathscr{T}(t)$ on Grassmann manifold $\mathscr{G}_k(\mathbb{R}^n)$, and let V be a tangent vector defined along the geodesic path $\mathscr{T}(.)$. Then, V said to be *Parallel transported:* along $\mathscr{T}(.)$ if:

$$\nabla_{\dot{\mathscr{T}}(.)} V = 0 \tag{3.35}$$

for all t, where $\dot{\mathscr{T}}(.)$ refers to the tangent vector to $\mathscr{T}(.)$ at t [4].

Overall, after computing the velocity vectors $\dot{\mathscr{X}}(t)$ between neighboring points on the trajectory, $\mathscr{X}(t)$ and $\mathscr{X}(t+\delta)$, we use the parallel transport on Grassmann manifold to translate it to the fixed tangent space. Repeating this operation for all velocity vectors along the trajectory results in an equivalent representation in a vector space (the tangent space attached to the identity element) to compute $V(t)_{X \to \mathscr{I}}$. Hence, the obtained transported velocity vector field reflects the way the motions are exhibited by the face or the body.

One can view the field of (transported) velocity vectors as a basic dynamic model to characterize the motion along Grassmann trajectories. That is, each velocity vector is by definition the first derivative of the geodesic path between subspaces, taken at the initial point of the geodesic. It is important to note that one can recover the initial trajectory knowing the velocity vector field and the initial point of the trajectory. Finally, a more complex dynamic model could be derived by including, in addition, the velocity vector fields the acceleration vector fields, and so on.

3.6 Conclusion

In this chapter, we have introduced a compact subspace representation of 3D videos and the motivation behind adopting it in our work. The technique of computing subspace from original data is discussed as well as the new nonlinear domain obtained from the linear subspaces of our data, called Grassmann manifold. The mathematical background and the geometrical properties of the underlying manifold such the the definition of metrics metrics on it to compare subspaces, the local linearity of this manifold, which induces the intrinsic learning approaches using tangent spaces. Also, the extrinsic learning method by embedding the non-flat manifold into another smooth manifold with a linear structure are discussed. Performing advanced learning, like sparse coding and dictionary learning, which can present several benefits (efficiency, ...) in classification and recognition are presented.

Also, how this Riemannian structure can support the sequential (partial) modeling/analysis of 3D dynamic data as time-parametrized curves of subspaces, and how we are able to capture the temporal information resides through these trajectories on the manifold to get relevant spatio-temporal representations. In the next chapter, we will introduce our approach to study the contribution of facial dynamics to the face recognition problem. Application and experimental illustrations of the mathematical tools introduced in this chapter will be used in the next one.

CHAPTER 4

Face Recognition from 4D Data

4.1 Introduction

As a first targeted application of the Grassmann representations, in particular the dictionary representation (presented in Section 3.4.3 of the previous Chapter), the present chapter introduces our 4D face recognition approach. The main task addressed here is to study the contribution of facial 3D shape's evolution over time in identity recognition. This topic is new and a few studies exist [128] until now, where the majority of current approaches exploit the 3D static shape of the face with a lack of investigation of its behavioral biometric. Moving from shape analysis of 3D static faces to dynamic faces (4D faces) gives rise to several new challenges related to the nature of the data and the processing algorithms. **1) Which static and dynamic shape representation is the most suitable for 4D face analysis? 2) How can the temporal dimension contribute in face recognition? 3) How efficient is it to compute statistical summaries on dynamic 3D faces? 4) From a perspective of face classification, which relevant features and classification algorithms can be used? 5) What are the challenges that unconstrained face recognition meets when working on 3D dynamic data?**

In this chapter, we aim to answer the above questions, by proposing a comprehensive framework for modeling and analyzing 3D facial sequences (4D faces), with an experimental illustration in face recognition from 4D sequences. The rest of the chapter is organized as follows – after an overview of the proposed solution presented in Sect. 4.2, in Sect. 4.3 the methodology of modeling 4D faces on Grassmann manifold is introduced; Our 3D dynamic face recognition framework is presented in Sect. 4.4; Experimental results and their discussions are given in Sect. 4.5. A new dataset for 4D face recognition in adverse conditions with preliminary evaluation experiment is presented in Sect. 4.6. Our conclusions and main findings out of the proposed approach are drawn in Sect. 4.7.

4.2 Overview of the proposed solution

Most of the recent face recognition approaches use sets of 2D still images (with different illumination or pose) or 2D videos as a data source. Besides, the subspace representation showed promising results with possible methodological and application extensions related to the geometry of the underlying manifolds (i.e., Grassmann manifolds), such as domain adaptation [53], multiple motion segmentation [26], video clustering [121], filtering [106] and others [88]. Also, advanced classification techniques, which lie on the non-linear nature of the data have been proposed, such as the extrinsic solutions to the problem of sparse coding and dictionary learning on Grassmann manifold [59]. On the other side, the use of the 3D facial shape for recognition purposes has been well explored [18, 40], in particular with the availability of the FRGC dataset and related experiments [101]. In contrast, little attention has been paid to the role of the shape dynamics (behavior) in identity recognition. In particular, Sun et al. [128] developed a vertex-flow tracking method to enable face recognition from temporal sequences of 3D face scans. In this method, they have showed the usefulness of 4D faces in the recognition process, instead of 2D videos and static 3D shapes. However, this approach is computationally expensive.

Following this new and promising line of research, we conducted a comprehensive study to investigate the role of 3D face dynamics in face recognition. To this end, as illustrated by the pipeline in Fig. 4.1, after a preprocessing step, we compute 3D surface curvature from each 3D static mesh of a sequence, and project it to a 2D map (called curvature-map). A sequence of curvature-maps is then shaped in a matrix form by reshaping the 2D maps to column vectors. A (compact) k-Singular Value Decomposition (k-SVD) is used to produce the subspace basis from the first k singular vectors, that is regarded as a point on a Grassmann manifold. These vectors build our spatio-temporal signature, which will be used in the recognition process in combination with both intrinsic and extrinsic classification methods on the underlying manifold. In particular, extrinsic methods based on sparse coding and dictionary learning achieved the best performances.

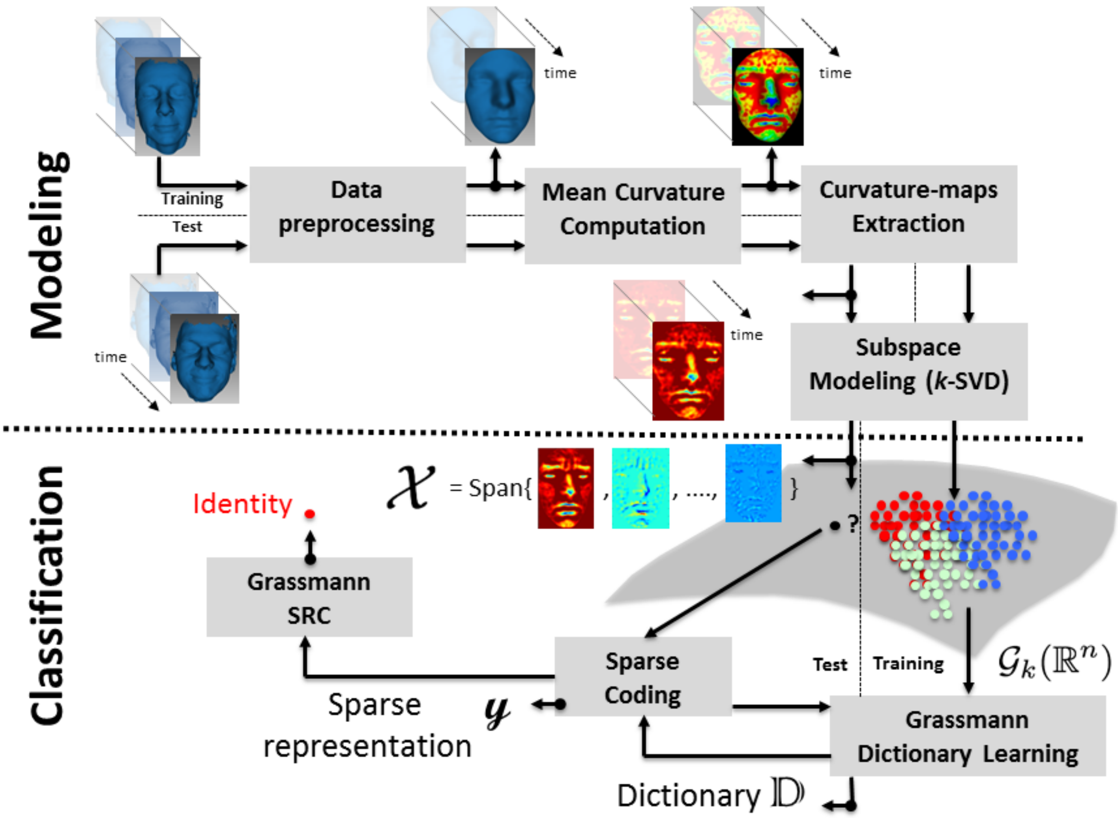

Figure 4.1: Overview of the proposed approach: top – modeling the shape and its dynamics using a subspace representation; bottom – classification of space representations using the SRC algorithm.

Figure 4.1 shows the above-mentioned method based on sparse coding and dictionary learning. The main contributions in this part of the thesis are:

- A fully automatic and computationally cheap face recognition approach using 4D data. To the best of our knowledge, this is the first study in the literature, which explores the subspace modeling methodology with advanced geometric and learning tools for 4D facial domain. Thus, a comprehensive framework is proposed and validated, which spans from the description of the 3D static shape and the modeling of its dynamics to an adequate classification schema;

- An in-depth investigation of the 3D shape dynamics contribution to face recognition is conducted, either in the case the facial expression is controlled or not.

- Instead of using the conventional autoregressive and moving average (ARMA) model for spatio-temporal analysis, which separates the appearance of visual data and their temporal evolution, our goal is to keep the shape and its motion in the same representation for identity recognition. The latter data is then represented by an optimized subspace, using the k-SVD orthogonalization procedure. The (optimized) subspace representation is suitable to process 3D data, which usually present missing parts (holes) and noise due to the acquisition process;

- An extensive experimental analysis, involving the BU-4DFE dataset and three classification schemes based on intrinsic and extrinsic methods: (1) A nearest-neighbor (NN) algorithm performed on the Grassmann manifold with respect to the (subjects) classes mean; (2) a variant of Grassmann Discriminant Analysis (GDA), called Graph-embedding GDA [60]; (3) A Sparse Representation-based Classification (SRC) derived from the Grassmann Dictionary Learning (GDL) approach [140][59].

- A new 3D/4D dynamic database of 58 subjects is collected in our laboratory to explore 4D face recognition problem in diverse conditions such as pose variation, expressions, talking, walking, internal and external occlusions and several persons in the scene. A preliminary evaluation on this new database has been conducted.

4.3 Modeling 4D-faces on Grassmann manifold

The idea of modeling multiple instances of visual data, like set of images or video sequences, as linear subspaces for classification and recognition tasks has revealed its efficiency in many computer vision problems [56, 132, 133]. The advantages of using this compact low-dimensional for 3d dynamic data representation can be summarized in its robustness against noise or missing parts in the original data; The ease of comparing two subspaces instead of two sets of 3D scans in Euclidean space; and the availability of computational tools from differential geometry makes working on non-linear data structure (e.g., the space of k-dimensional subspaces) possible and allows managing the non-Euclidean nature of these subspaces. Accordingly, in this work, we adopt the subspace representation solution for analyzing 4D facial sequences. To our knowledge, this is one of the earliest investigations on modeling the temporal evolution of 3D facial shapes with application to face recognition. Studying the effects of these two aspects together is still an open problem in computer vision domain.

In the proposed solution, we consider 3D scans of the face acquired continuously via a dynamic 3D scanner, thus producing a temporal 3D sequence with the dynamic evolution of the 3D face. Using these data, the proposed approach is designed to exploit the spatio-temporal information. To achieve this goal, a subspace modeling technique is applied as follows: (*i*) The 3D scans are preprocessed by cropping the facial region from the rest of the scan, then pose normalization, denoising via smoothing, and holes filling are performed; (*ii*) The mean curvature on 3D surfaces is computed, so that a flow of curvature-maps is produced by projection; (*iii*) The k-SVD orthogonalization procedure is applied to subsequences of the curvature-maps to obtain an orthonormal basis spanning an optimized subspace. This subspace represents an element of a Grassmannian manifold.

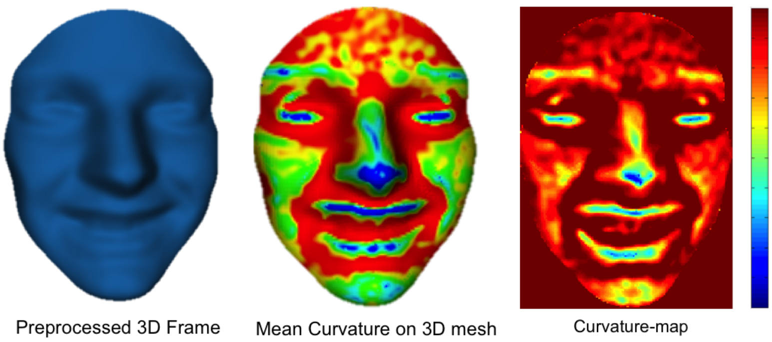

Figure 4.2: 3D static facial shape representation using the mean curvature. From left to right, the pre-processed 3D face, the mean curvature computed on the 3D mesh, and the normalized curvature-map are reported.

The first step of this framework is illustrated in Fig. 4.2. On the left, the preprocessed face scan is reported; The mean curvature computed on the mesh is reported in the middle. The curvature map projected on a 2D image of size $\hat{n} \times \hat{m}$ is shown on the right. This latter map extracted for each frame of a sequence constitutes the data source for our spatio-temporal analysis. More formally, let S_m be a 3D dynamic face sequence with m frames. A subsequence of $\omega < m$ frames is indicated with $S_\omega = \{f_1, f_2, \dots, f_\omega\}$, where each f_i is a curvature-map of linearized size $n = \hat{n} \times \hat{m}$, that is $S_\omega \in \mathbb{R}^{n \times \omega}$, and ω is regarded as the *window size*. Applying the k-SVD orthogonalization procedure where $S_\omega = U \Sigma V^T$, and the k first columns of U matrix provide the dominant k-left singular vectors of S_ω. The subspace spanned by these vectors is an element of the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$.

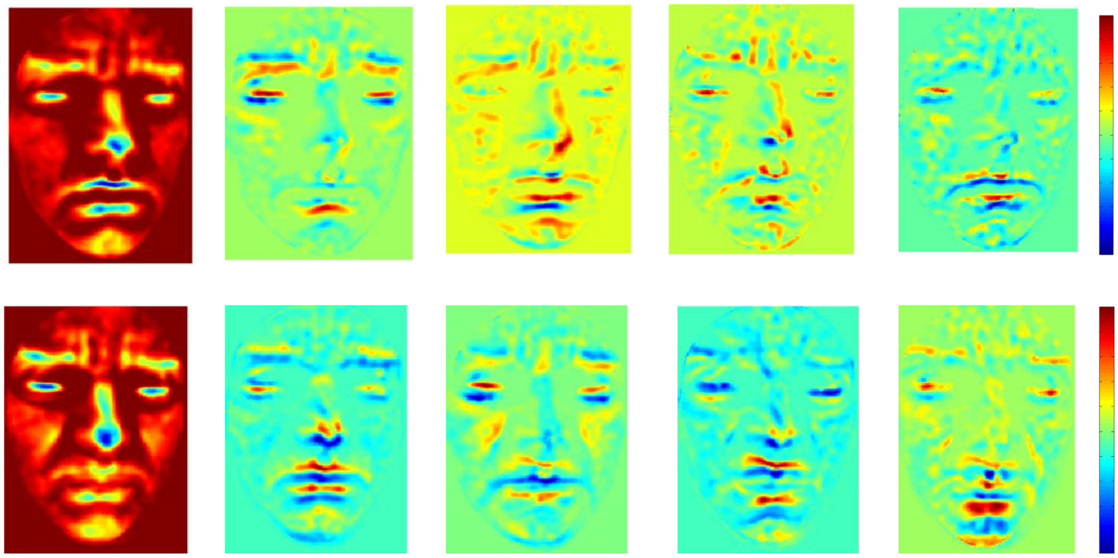

Figure 4.3: Visual illustration of two subspaces (i.e., points on the Grassmann manifold) using their singular vectors derived from SVD *orthogonalization* on sequences of $\omega = 50$ frames (*angry*, top row – *disgust*, bottom row). From left to right, the 5-dominant left singular-vectors (subspace of order 5) of the original data are shown. The first column represents the common shape description over the sequence. While the remaining columns capture the dominant facial motions of the face.

Figure 4.3 shows, as color maps, the matrices representing the subspaces computed from two different 3D facial sequences. It can be appreciated that a subspace (k first dominant left singular vectors of the original matrix of data) can be viewed as the mean shape computed over the subsequence (leftmost images), followed by the dominant deformations (remaining images on the right). These deformation images are different from each other, and change in respect to the expression exhibited by the face (*angry* in the first row, and *surprise* in the second). The histogram equalization is used here (except for the images in the left column) to highlight the location of the deformation areas, using cold to warm colors. Colors in between reflect the most stable areas of the curvature-maps over the 3D video. The singular value decomposition technique provides us a measure to evaluate the importance of the information that every singular vector carries in relative to the original data. This evaluation can be obtained from the singular values which are the diagonal elements of the matrix Σ. Equation 4.1 gives the percentage of the information resides in every first k vectors, thus we can decide the threshold to stop considering the left ones, which is 90% in our case:

$$Y_k = \frac{\Sigma_{i=1}^{k}\lambda_i}{\Sigma_{i=1}^{\omega}\lambda_i}\,, \tag{4.1}$$

where λ_i is the singular value corresponding to singular vector U_i.

In Fig. 4.4, we report the percentage of the information kept (after the matrix factorization) as a function of the number of singular vectors for different window size $\omega \in \{6, 10, 15, 20, 25\}$ given in Eq. 4.1.

From Fig. 4.4, the amount of information increases by considering more singular vectors, till arriving to 100% by using all of them. Interestingly, in all the cases, about 90% of the information of a sequence is captured by considering less than half of the singular-vectors. This observation suggests us the identity information mainly resides in the few first dominant singular vectors. While the remaining components contain the noise and redundant information. From this illustration and discussion, the concept of compact and low dimensional representation appears clearly.

4.4 Identity recognition algorithms

To perform face recognition from the 3D facial shapes and their temporal evolution, the flow of curvature-maps is first divided into clips (subsequences) of size ω. Then, each clip is modeled as an element of Grassmann manifold via k-SVD orthogonalization. More formally, given a sequence of curvature-maps $\{m_0, \cdots, m_t\}$, a predefined size of a sliding window ω, and a fixed order of subspaces k, the idea is to consider the maps under the temporal interval $[t - w + 1, t]$ and to compute the corresponding subspace $\mathscr{X}_t$. This results in a collection of subspaces, elements of Grassmann manifold, which represent the 3D video sequence (after curvature computation).

The main goal of such representation is to capture the 3D shape of the face as well as its dynamics (spatio-temporal description) to perform face recognition. In the following, we present two classification methodologies are used in this work: The Grassmann Nearest Neighbor classifier (GNNC) based on one of the distances defined on Grassmann manifold have been defined in Sect. 3.3.5. This classification method involves the Karcher mean subspace estimation (Algorithm 1) to compute a representative target subspace for each subject class; The second method uses Grassmann Sparse Representation (GSR) classifier adapted to classify the testing subspaces depending on their sparse codes implemented as detailed in Sect. 3.4, Algorithm 3.

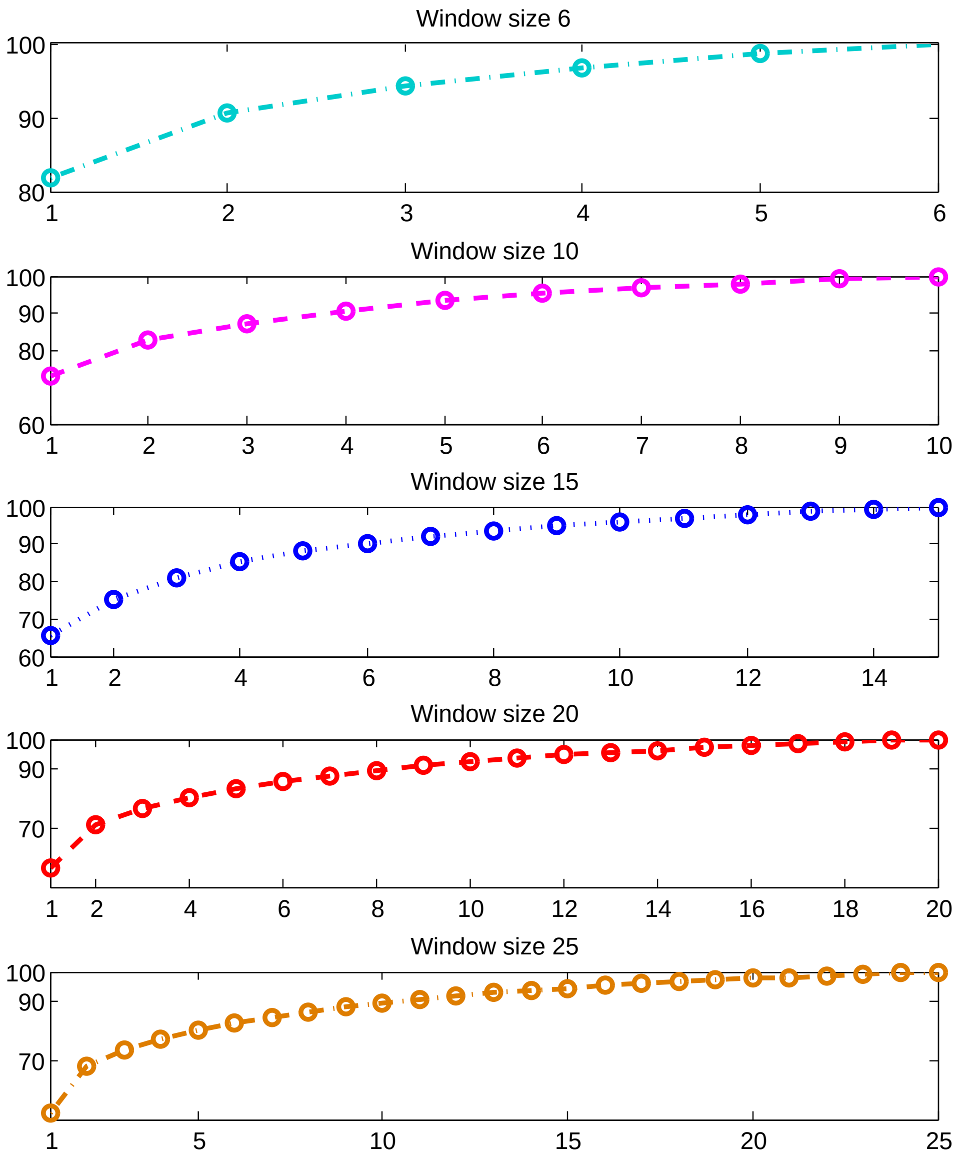

Figure 4.4: Information Y_k captured by the first k singular vectors returned by SVD as a function of λ. Results for different window size are reported.

4.4.1 Grassmann Nearest-Neighbor Classifier (GNNC)

In this approach, for each subject a mean (representative) subspace is computed out of the subspaces that belong to the same subject in the training set (i.e., more than one subsequence is used in the training for each individual) by applying Karcher mean Algorithm 1. These means constitute the gallery subspaces used for recognition. According to this, given a probe subspace $\mathscr{X} = Span(X)$, it is compared against the gallery mean subspaces using one of the distances defined on the Grassmann manifold (see Sect. 3.3.5). Finally, the probe subspace is assigned to one class using the Grassmann Nearest-Neighbor classifier. Figure 4.5 illustrates the idea of computing the principal angles between subspaces, which are served to compute the distances.

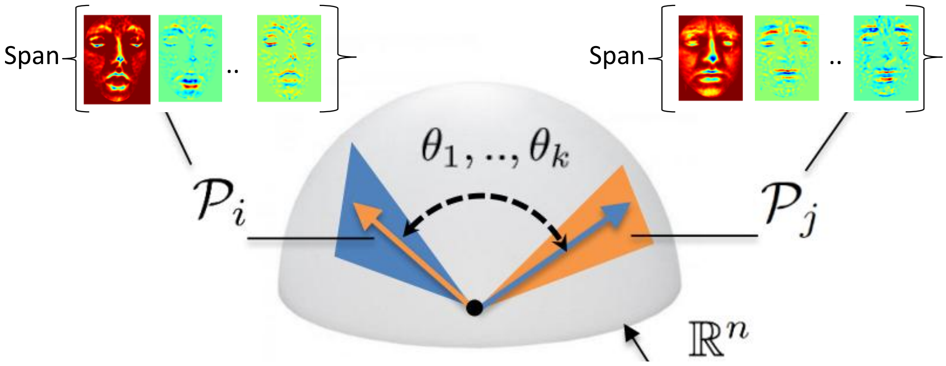

Figure 4.5: Comparing the similarity of two 3D dynamic subsequences after presenting them as two subspaces P_i, P_j of dimension k on $\mathbb{R}^n$.

Learning one mean subspace for the subject (class) as a representative makes the recognition much faster than using an exhaustive search that compares the probe against all the subspaces in the training. Algorithm 4 summarizes the classification steps, where the mean subspace estimation is performed off-line. While the comparison of the probe subspace to the gallery subspaces is performed online.

Algorithm 4 – Grassmann Nearest-Neighbor Classification

Require: Set of training subspaces $\mathbb{X} = \{\mathscr{X}_i\}_{i=1}^{m} \in \mathscr{G}_k(\mathbb{R}^n)$ where $\mathscr{X}_i = Span(X_i)$, belong to C classes, the query sample $\mathscr{Y} = Span(Y) \in \mathscr{G}_k(\mathbb{R}^n)$

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

  Compute the Karcher mean μ_i using Algorithm 1

end for

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

  $d_i(\mathscr{Y}) = \text{dist}(\mathscr{Y}, \mu_i)$ // *one of the distances of Sect. 3.3.5*

end for

Ensure: Identity$(\mathscr{Y}) \leftarrow arg\ \min_i(d_i(\mathscr{Y}))$

In this algorithm, dist(.,.) denotes one of the Grassmann distances defined in Sect. 3.3.5. A comparison study of these distances performance is presented in the experimental evaluation in Sect. 4.5

(a)

(b)

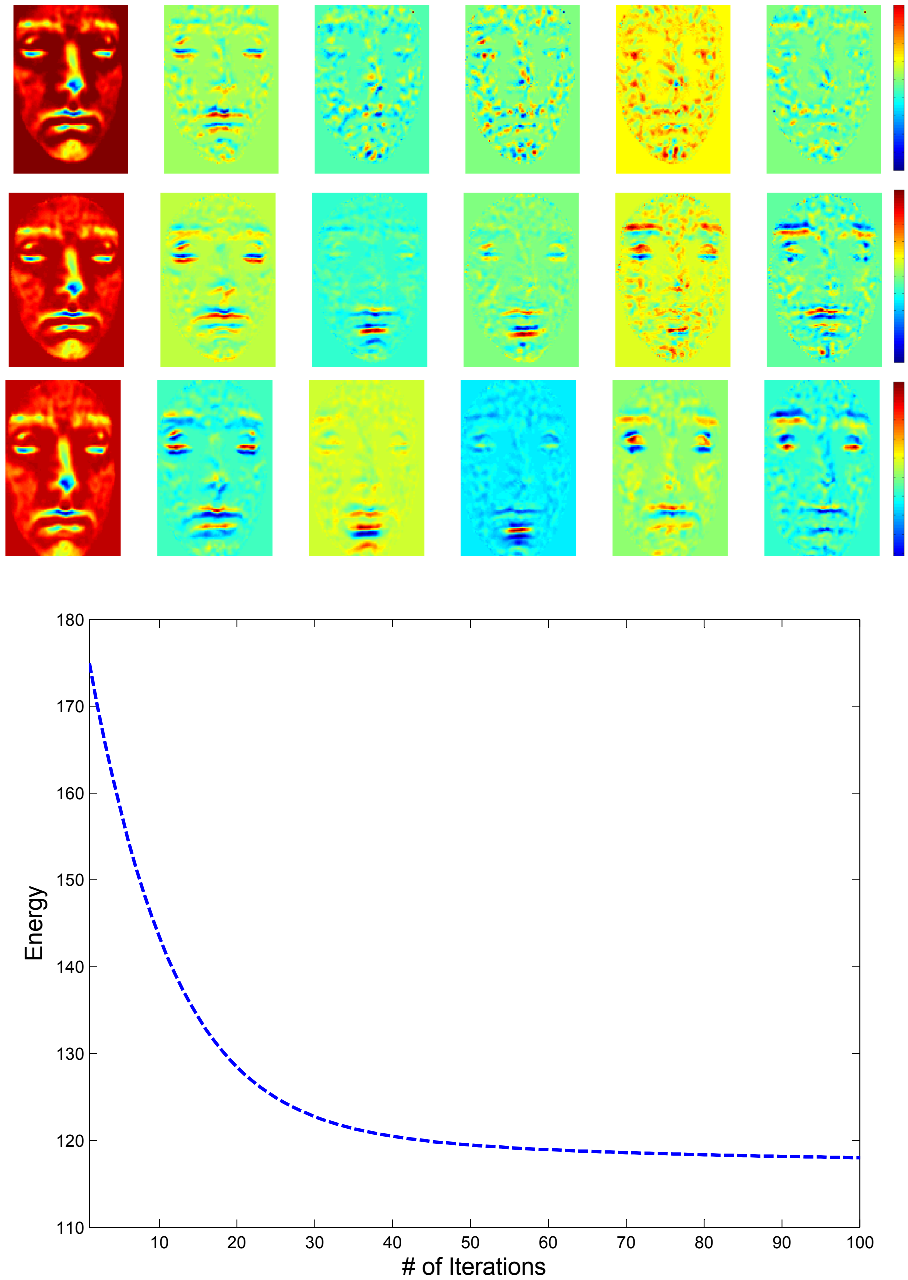

Figure 4.6: (a) Each row represents a sample mean subspace dimension computed on the subsequences of the same person with different expressions. The first 6 dominant singular-vectors are used to represent the sequences in each case. Three different window size are instead considered passing from the top to the bottom row ($\omega = 6, 25, 50$, respectively) where ω refers to number of 3D frames in the original 3D sequence; (b) The energy (i.e., $||\bar{v}||$) minimized in Algorithm 1 for estimating the mean subspace.

Illustration of Karcher mean computation

In Fig. 4.6(a) three *mean subspaces* obtained with the Karcher mean Algorithm presented in Algorithm 1 are shown, each of them computes the mean facial dynamics for subsequences of the same person under different expressions. In each row, the first six dominant singular vectors are reported, while the window size changes from $\omega = 6$, to 25 and 50, from top to bottom, respectively. Considering the top row, we can notice the first singular vector (first column) captures the main shape information of the face across the sequence and expressions, which is the mean of the global 3D dynamic data, and the second singular-vector captures the main deformation. Less relevant data are included in the remaining singular vectors. This illustration shows clearly the main motivation to choose the subspace representation. This observation (i.e., $k = 2$ is sufficient for $\omega = 6$) is in agreement with the results reported in Fig. 4.4, where the information Y_k captured by considering the first k singular-values reaches 90% considering just the first two singular-values (in the case of window size $\omega = 6$). In contrast, in the second and the third example of Fig. 4.6(a), the first column captures the mean shape, while the remaining singular vectors are required to model the principal deformations of the face. In Fig. 4.6(b), an example of the values of the energy (i.e., $||\bar{v}||$) minimized over the iterations of Algorithm 1 is also plotted.

Figure 4.6 points out the relevance of using the subspace representation for modeling the spatio-temporal behavior of the dynamic face. Besides, discarding less dominant singular vectors allows us to remove the noise and redundancy in the original data. The 3D shape representation obtained here by computing the mean curvature-maps of the 3D face, which is relevant to such analysis. The mean subspace reflects the shape information as well as the dominant deformations of the face in the subsequence (window).

These observations are confirmed in Fig. 4.7, where the visual illustration of the mean computed on sets of subspaces corresponding to the same expressions of different subjects are reported. In this Figure, each row corresponds to the mean of one of the six universal facial expressions – *Angry*, *Disgust*, *Fear*, *Happy*, *Sad* and *Surprise* – as conveyed by 3D dynamic sequences belong to 10 different people.

The window size is set to $\omega = 50$ in each row (the first six dominant singular vectors are shown). It is evident that each expression produces quite different dominant deformations.

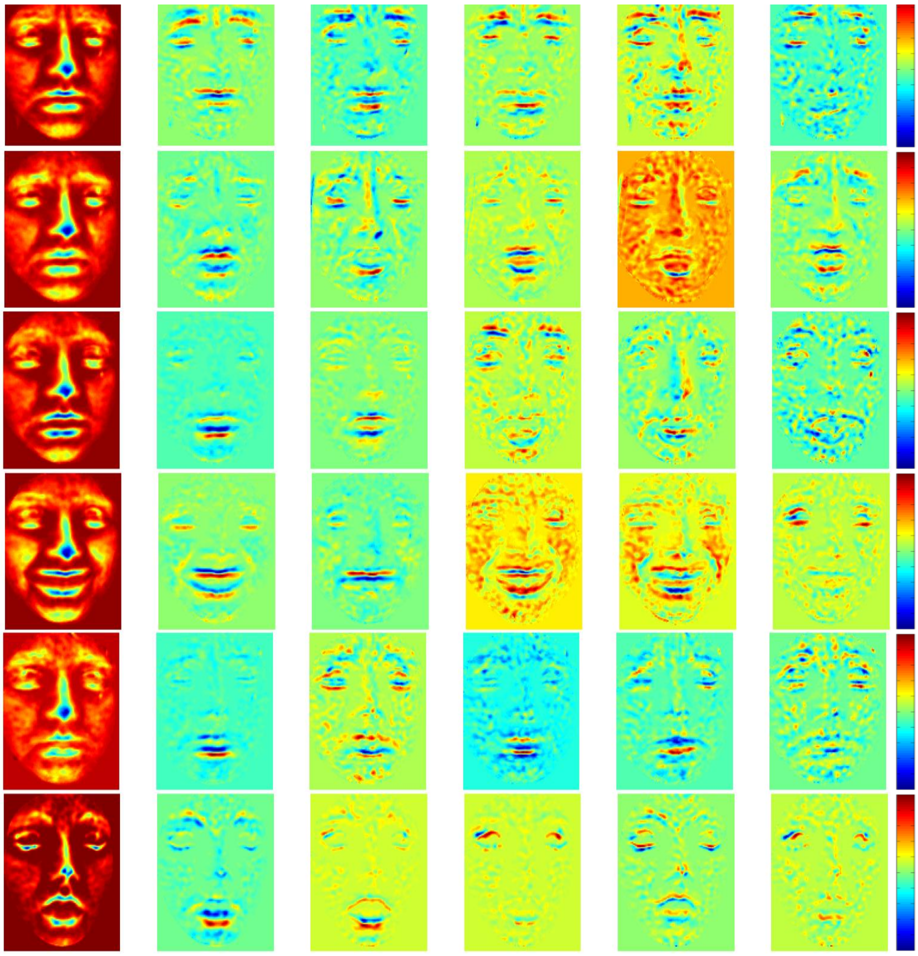

Figure 4.7: Visual illustration of mean subspaces. In every row of the six, we have one subsapce computed from subspaces belonging to 10 different person but they were acting the same expression. Each row represents one of the six universal facial expressions, namely, from top to bottom: *Angry*, *Disgust*, *Fear*, *Happy*, *Sad* and *Surprise*.

4.4.2 Grassmann Sparse Representation Classifier (GSRC)

In this case, the classification is performed on the sparse representation computed according to sparse coding Algorithm 3 presented in Sect. 3.4.3.

In fact, given a test sample, its sparse representation is first computed using the dictionary on the training samples. Consequently, conventional classification methods, like SVM or Nearest-Neighbor can be applied. An alternative solution is to use the Sparse Representation Classifier (SRC) proposed in [140].

Algorithm 5 summarizes the main steps of the classification procedure. The main concept behind this classifier is to reproduce the testing query subspace from non-zero sparse codes that belong to every class in the dictionary separately. Repeating this class-specific estimation and computing the residual error between them and the original query subspace gives a similarity measure. The estimation from the correct class should give the minimum residual error for correct recognition.

The Dirac function has been used in Algorithm 5 allows the selection of the coefficients associated to the i^{th} class. That is, all the elements of this vector are set to be 0 except those which correspond to the i^{th} class.

Algorithm 5 – Grassmann Sparse Representation Classifier

Require: Grassmann Dictionary $\mathbb{D} = \{\mathscr{D}_i\}_{i=1}^{N} \in \mathscr{G}_k(\mathbb{R}^n)$ where $\mathscr{D}_i = Span(D_i)$ with C classes, the test query $\mathscr{X} \in \mathscr{G}_k(\mathbb{R}^n)$ where $\mathscr{X} = Span(X)$ and $XX^T = \hat{X}$

Sparse code estimation of the query as in Algorithm 3

$y^* \leftarrow arg\ \min_y \|x^* - Ay\|^2 + \lambda\|y\|_1$

for i ← to C **do**

$\varepsilon_i(\mathscr{X}) = \|\hat{X} - \Sigma_{i=1}^{N} y_i \hat{D}_i dirac_i(l_j - i)\|_F^2$,

where l_j is the atom label

end for

Ensure: Identity($\mathscr{X}$) ← $arg\ \min_i(\varepsilon_i(\mathscr{X}))$

In summary, face recognition is performed according to the following steps: (1) *Dictionary learning* on the Grassmann manifold - given a training subset of observations, a set of atoms (dictionary) is determined to describe the observations sparsely; (2) *Sparse representation* - given a dictionary and a probe on the underlying manifold, the probe is approximated using a sparse linear combination of atoms that belong to every class from the dictionary separately; (3) *GSR-based classification* - once the training and testing observations are expressed linearly using a sparse representation, it is possible to perform the Grassmann Sparse Representation Classification.

4.5 Experiments and results

To investigate the contribution of facial dynamics in identity recognition using 4D data, we conducted extensive experiments involving the BU-4DFE dataset. This dataset has been collected at the Binghamton University [148] and used in several studies on 4D facial expression recognition. To our knowledge, only two works, Sun et al. [128] and Hayat et al. [61] have reported identification performance on this dataset. To allow a fair comparison with their study, we will consider in the following the same experimental setting.

Before to present experiments and results, a summary of the main characteristics of the BU-4DFE dataset and its pre-processing is presented.

4.5.1 BU-4DFE dataset description and pre-processing

The BU-4DFE database consists of 101 subjects (58 female and 43 male, with an age range of 18 – 45 years old). It includes 606 3D model sequences with 6 universal expressions and a variety of ethnic/racial ancestries. Each participant (subject) was requested to perform the six prototypical expressions – *angry*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise* – separately. The acquisition protocol requires each expression sequence to start and end with neutral facial states. Each expression was performed gradually passing from neutral, low intensity, high intensity, and back to low intensity and neutral (i.e., following the subsequent states *neutral-onset-apex-offset-neutral*).

Actually, as a matter of fact, at a visual inspection some sequences evidence a wrong acquisition, starting with a non-neutral expression. In any case, each 3D sequence captures one expression at a rate of 25 frames per second, lasting approximately 4 seconds, with about 35k vertices per 3D frame (or 3D mesh). As acquisition technology, the Di4D capturing system was used [148], which produces sequences of stereo images and computes 3D meshes of the face based on a passive stereo-photogrammetry approach. The resulting 3D frames of a sequence show a near-frontal pose, with some slight changes occurring mainly in the azimuthal plane. The scans are affected by large outliers, mainly located in the hair, neck and shoulders regions.

In order to remove these imperfections from each 3D frame, an efficient pre-processing pipeline similar to [15] has been performed. The main steps of this pipeline are summarized as follows: (1) For each 3D frame, the holes are filled in; (2) The tip of the nose is detected, then the facial area is cropped using a sphere with radius of $90mm$ centered at the detected nose tip; (3) The pose of each 3D frame is normalized by registering it to the previous one using the Iterative Closest Point (ICP) algorithm. Once the pre-processing is performed, the mean curvature is computed from each 3D frame (Fig. 4.2). Then, the curvature-maps (images) are produced by projection, as described in Sect. 4.3. All these steps are implemented using the Visualization Toolkit (VTK) library[1].

[1] http://www.vtk.org

In the following, we report experimental evaluation and comparative analysis of the proposed approaches using Grassmann Nearest-Neighbor (GNNC) classification on the mean subspaces of each subject class, and Grassmann Sparse-Representation (GSR) based classification computed on the sparse codes, with respect to the current literature.

4.5.2 Experimental setting

Following the protocol proposed in [128], 60 subjects have been considered out of the BU-4DFE, and their sequences are partitioned into subsequences using a window size $\omega = 6$ (with a shifting step of 3 frames). This results into 30 sub-sequences extracted out of every facial expression sequence of the 60 subjects (i.e., each sequence has approximately 90 frames). On these subsequences, experiments have been conducted following two different settings:

- *Expression Independent* (EI) – One expression per subject is used for training, and this expression does not appear in the testing. All the other five expression sequences are used for testing. Since 30 sub-sequences represent each expression sequence, for the 60 subjects a total of $30 \times 60 = 1800$ subsequences is used for training. Five expressions per subject are used for testing, i.e., for each subject we have $5 \times 30 = 150$ test subsequences, with a total for all the 60 subjects of $150 \times 60 = 9000$ subsequences;

- *Expression Dependent* (ED) – For each sequence, the first half (from neutral to nearby the apex of the expression) is used for training, while the remaining half (from the apex of the expression to neutral) is used for testing. As a consequence, the gallery and the probe samples convey similar dynamic behavior, tough with inverse temporal evolution. The number of training subsequences for every subject is $15 \times 6 = 90$, with a total for the 60 subjects of $90 \times 60 = 5400$ subsequences. The same number of subsequences is used for testing.

4.5.3 4D face recognition using GNNC

In this experiment, a window of six frames $\omega = 6$ and shifting step equals to 3 is used (the same as in [128]), with only the first two dominant components kept for representing the subspace ($k = 2$). The GNN-classification method is based on a gallery of subspaces, one per subject, each computed as the mean of the training subsequences for the subject. With the setting above, in the EI scenario, one complete expression is used to compute the mean for each subject, i.e. 30 subsequences; In the ED scenario, the mean is computed on $15 \times 6 = 90$ subsequences with different expressions.

Using the GNN-classifier, a comparison is performed between the ED and EI experiments. Different distances are also considered, which involve the principal angles between subspaces (see Sect. 3.3.5). The average recognition rates are reported in Table 4.1.

Table 4.1: Recognition rates (RR%) for GNN-classification using different distances

Subspace Distance	**ED – RR (%)**	**EI – RR (%)**
Min Correlation	44.75	28.72
Binet-Cauchy	52.83	51,99
Geodesic	73.00	65.00
Procrustes	78.11	66.55
Max Correlation	92.61	67.12
Projection	**93.69**	**68.88**

The observations can be derived from this Table are: (*i*) ED results outperform EI results for each distance measure. This is expected, since in the ED setting there are sequences of the same subject conveying the same expression both in the gallery and probe sets (though with inverse temporal evolution); (*ii*) The different recognition rates scored by the distances provide experimental evidence of the discriminative information distribution across the principle angles. In particular, the highest recognition rate obtained by the *Projection* distance shows that all the singular vectors, and consequently the dynamic information of subsequences, helps in the recognition task by improving the result obtained using just one principle angles (i.e., *Max Correlation* distance). The lowest recognition rate is scored by the *Min Correlation* distance, suggesting us that the subspaces on the manifold are sufficiently separated from each other, thus making them well suited for the identity recognition task.

Results reported in Table 4.1 have been obtained by comparing single instances (subspaces) in the video. Since subsequences are part of a continuous video, it is possible to fuse the decisions of successive subsequence instances to perform recognition. This allows us to design an incremental recognition system over time, where multiple instances are used to decide instead of only one. This idea has been implemented using a majority voting fusion rule, at each time, using all available instances. The experimental results are reported in Fig. 4.8 to show the performance at increasing size of the data have been seen and analyzed along a sequence. From these plots, it is clear that the performance

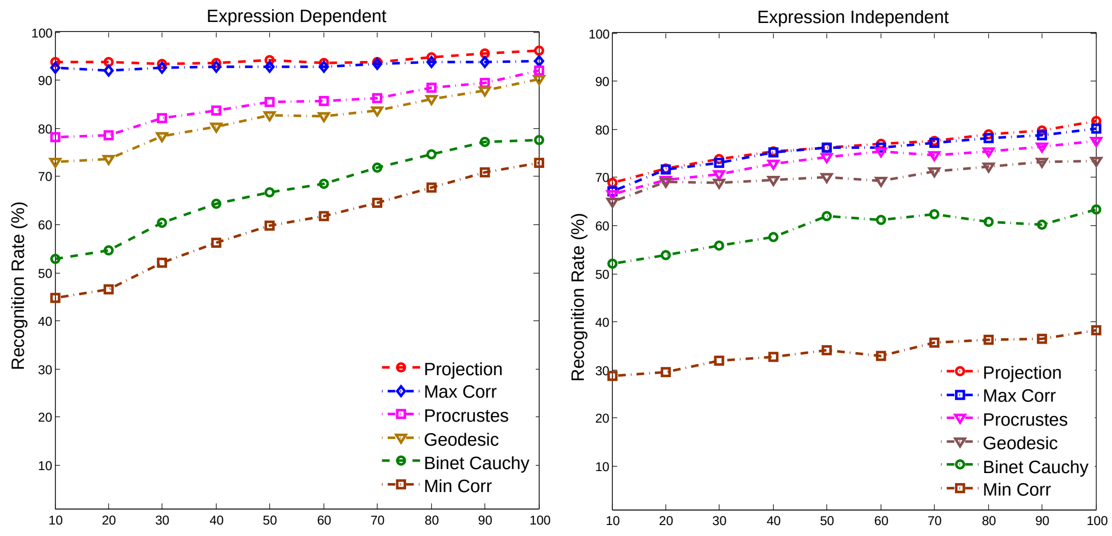

Figure 4.8: Trade-off between accuracy and latency (fraction of the video seen) for different Grassmann metrics/distances in the ED and EI settings.

increase by having longer fraction of the 3D video. This observation is the same, under ED and EI settings.

4.5.4 4D face recognition using GSRC

In these experiments, we use the proposed solution based on Grassmann Sparse Representation algorithm presented before (GSR). A variant of the GDA Grassmann Discriminant Analysis algorithm [56], called GGDA (Graph-embedding GDA) [60] is also used as a baseline to evaluate the effectiveness of the GSR algorithm. In practice, the flow of curvature-maps, for the window of size ω is first mapped to the Grassmann manifold using SVD. Then, the steps described in Sect. 4.4.2 are performed for training and testing. Results under the ED and EI settings are reported. A comprehensive discussion of the experimental results, when varying the window size ω, and the subspace order k is also reported.

Expression Independent (EI) experiment

As a preliminary experiment, we investigated the effect of the subspace order k on the performance. To this end, we apply the GSR algorithm with a varying $k \in \{1,2,3,5,6\}$, while keeping a fixed window size $\omega = 6$ and shifting step equals to 3. So, in this case we have 30 training subspaces for subject, for a total of 1800 subspaces in the training set (dictionary).

The subspace order k is also related to the information carried by the respective eigenvalues through the measure Y_k (see Eq. (4.1)). As shown in Table 4.2, the highest average recognition rate is 84.13%, obtained for $k = 2$. This rate is 3% higher than the average recognition rate obtained for $k = 1$ (using only the first dominant left-singular vector, which corresponds to the common data over the window).

Table 4.2: EI experiment: Effect of the subspace order k on the recognition rate for the GSR algorithm. Subsequences with window size $\omega = 6$ have been used in all the cases

Subspace order k	1	**2**	3	4	5	6
Y_k (%)	81	**90**	94	96	98	100
RR (%)	81.03	**84.13**	81.76	81.22	80.94	80.02

This allows us to make two main conclusions: (*i*) The importance of the facial dynamics in improving the recognition performance. In fact, the optimal parameter $k = 2$ implies that the mean and the first dominant deformations are important in the recognition process. They are given by the first and the second singular-vectors of the orthogonal matrix, respectively; (*ii*) The remaining left-singular vectors are less relevant in the recognition process, including the noise which is present in the 4D acquisition. We note that $k = 2$ allows capturing in average about 90% of the data available in the 4D sub-sequence. Based on these empirical observations, in our next experiments, we will consider 90% of the information for different window size (ω).

We are interested now in studying the effect of varying the size of the window on the performance. In the following experiment, we have varied this parameter in the set $\omega \in \{6, 10, 15, 20, 25\}$. The subspace order k is defined as the number of left singular vectors, which retains 90% of the original data. The corresponding recognition accuracy are reported in Table 4.3. It can be seen that the optimal window size is $\omega = 6$ for both the GSR and the GGDA algorithms.

The reason behind the decreasing accuracy at increasing size of the window is the lack of temporal registration of the curvature-maps. In fact, a large difference between the frames across the window affects negatively the orthogonalization procedure, which assumes dense correspondence between the frames. Interestingly, the accuracy obtained using the GSR (84.13%) substantially improves the accuracy achieved using the GGDA (64.24%), and the GNN-classification (68.88%). This result also evidences the efficiency of sparse coding of subspaces in comparison to the discriminant analysis, which can be affected by the points distribution over the Grassmannian manifold.

Table 4.3: EI experiment: Effect of the window size ω on the recognition accuracy for the GGDA and GSR algorithms. The subspace order k is set to keep 90% of the information

ω, k	**Algorithm**	
	GGDA – RR(%)	**GSR – RR(%)**
6, 2	**64.24**	**84.13**
10, 4	61.15	79.89
15, 6	56.61	76.55
20, 9	50.50	76.59
25, 11	50.60	75.80

Table 4.4 provides additional details by reporting the recognition rates obtained separately for each test expression, by the GGDA and GSR algorithms, and the approach proposed in [128]. The average recognition rate achieved by GSR is 84%, which is about 10% lower than the accuracy reported in [128]. However, differently from the approach proposed by Sun et al., the proposed solution does not require any manual or automatic landmarking of the face, and it is computationally more efficient. In addition, the dense (vertex-level) registration of the 3D frames, which is computationally complex and time-consuming and is not performed in our method. On an opposite side, this operation permits the approach presented in [128] to achieve comparable results throughout all the expressions. In our case instead, we observe the RR decreases by 4% in the case of posed surprise expression, which includes topological variations of the face (i.e., mouth open). Another methodological difference between the two approaches is that Sun et al. designed and trained two separate HMMs called spatial and temporal. In our approach, only 2 singular vectors are used to encode the spatio-temporal information of a 3D facial sequence and can be used to perform GSR classification.

The recognition performance of our solution can be improved by using an increasing fraction of the video. This implies that more than one instance (subsequence) is used to recognize a subject. With this approach, the overall performance of GSRC increases from 84.13% (using only one instance, which represents about 5% of the video) to 95.11% using the whole video (about $4s$). This is illustrated in Fig. 4.9, separately for each expression. This Figure also confirms the difficulty in recognizing subjects which convey the *Surprise* expression.

Table 4.4: EI experiment: Recognition rate obtained using different training expressions compared to the approach in [128]

Training Expression	Method		
	Sun et al. [128]	**GGDA**	**GSR**
Angry	94.12%	61.26%	85.20%
Disgust	94.09%	68.54%	87.70%
Fear	94.45%	69.02%	83.49%
Happy	94.52%	68.56%	83.36%
Sad	93.87%	63.05%	84.86%
Surprise	95.02%	56.07%	80.49%
Overall	94.37%	64.42%	84.13%

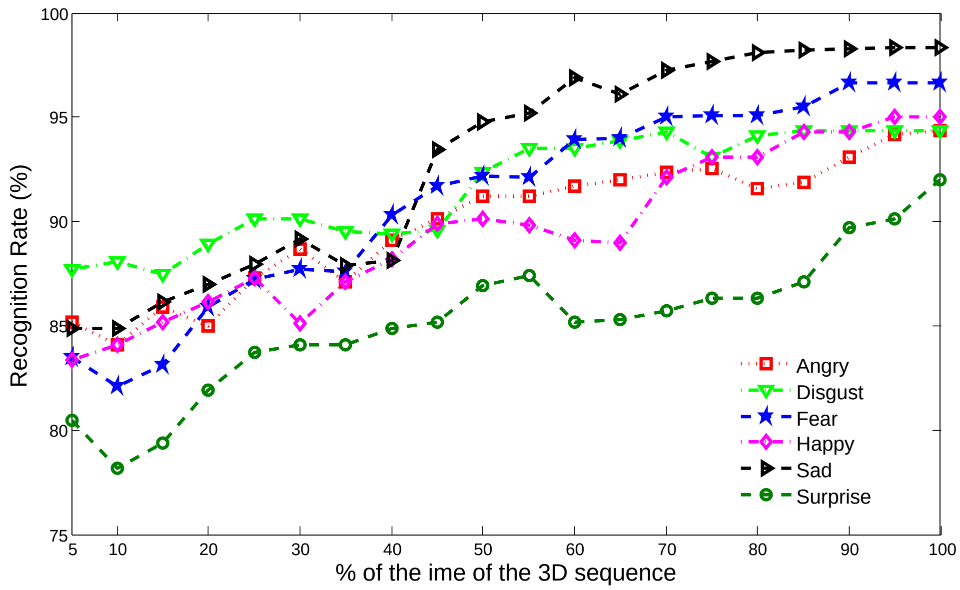

Figure 4.9: EI Experiment: Trade-off between accuracy (RR%) and latency.

In the experiments presented above, only one expression is considered for (identity) training. We have also analyzed the results in the case the training is performed with five expressions, i.e., 9000 for training (150 for each subject), and 1800 for testing (30 per subject), while the test is performed on subsequences from the remaining expression. Results are reported in Table 4.5, which provides a comparison when training with one expression versus training with five. Comparison of these results (using GSRC) shows that increasing the number of samples and their dynamics (even thought they come from different expressions) can significantly increase the recognition rate from 84.13% to 93.37%.

We can also observe that recognizing the subject identity under *Surprise* expression is the most difficult case among the six expressions, due to the large shape changes, where identity recognition under *Sad* expression is the easiest across the time.

Table 4.5: Impact of the training set on the performance: training based on only one expression vs. training based on five expressions

Testing Expression	**Training by one**	**Training by five**
Angry	83.27%	94.50%
Disgust	78.42%	96.30%
Fear	92.21%	98.13%
Happy	86.23%	93.20%
Sad	94.32%	97.73%
Surprise	69.75%	80.40%
Overall	**84.13%**	**93.37%**

Expression Dependent (ED) experiments

In this experiment, the window size is $\omega = 6$, with shifting step, equals to 3, and 30 sub-sequences are obtained from each facial expression sequence, half of which is used for training and a half for testing. Thus, we have 90 training subspaces per subject, and a dictionary of 5400 subspaces.

The GSR-based classifier is used in this experiment. Table 4.6 reports the results obtained using the GSR and the GGDA algorithms on 3D dynamic sequences (4D). In addition, for comparison purposes, we also reported in the Table several results from [128], including Gabor wavelets on 2D videos, LLE, PCA and LDA on 3D static data, and the ST-HMM on 4D data.

It can be seen that both GGDA and GSR outperform state of the art approach. In particular, their accuracy is close or equal to 100% under the ED-setting. Our explanation of the higher accuracy achieved by the GGDA and GSR compared to existing methods is that the optimized SVD-based orthogonalization produces a matrix independent of the time-order of the 3D video clips. That is, comparing two video clips taken from the Onset-Apex and the Apex-Offset gives small distance as the temporal order of the curvature-maps is ignored. This demonstrates the efficiency of using the Grassmann

Table 4.6: ED-experiment: Comparison between the recognition accuracy obtained for the methods proposed in this works, and for the 2D video, 3D static, and 3D dynamic (4D) approaches reported in [128]

Method	**RR (%)**
Gabor-wavelet on 2D videos (from [128])	85.09
LLE on static 3D (from [128])	82.34
PCA on static 3D (from [128])	80.78
LDA on static 3D (from [128])	91.37
ST-HMM on 4D [128]	97.47
GNN on 4D	**93.69**
GGDA on 4D	**98.08**
GSR on 4D	**100**

representation and learning methods defined on in for solving 4D face recognition problem.

4.5.5 Comparative study and discussions

From the experimental results reported above, it emerges the proposed approach, which combines Grassmann representation with an extrinsic learning method achieved promising results in 4D face recognition. We have demonstrated, through extensive experiments, the contribution of the facial dynamics in the recognition process. In Table 4.7, we summarize the obtained results under the ED and EI settings. We also studied the advantage of using the dynamic of shape (3D videos) compared to the dynamic of appearance (2D videos), as reported in the Table with comparison with Sun et al. [128].

Table 4.7: ED and EI results for 2D and 3D videos

Method	**EI – RR (%)**	**ED – RR (%)**
2D video A-HMM [83] (from [128])	67.05	93.97
4D ST-HMM [128]	94.37	97.47
4D GSR	**84.13**	**100**

It is clear from these results that the 3D video modality outperforms the 2D video modality. That is, the dynamics in 3D facial shapes has more discriminating power compared to the dynamics of 2D facial images. When the proposed approach is compared with [128], it is evident that the latter performs better in the ED case, where sequences with different expressions are compared. This indicates the effect of using registration and tracking technique for the robustness against expression differences.

This is mainly due to the dense temporal vertex-tracking approach required before training the HMMs. However, this comes at the cost of an increased computational complexity of the tracking, in addition to the required accurate manual/automatic landmarks detection in the first 3D frame of a sequence.

Table 4.8: Processing time of the proposed pipeline compared to [128]. A 3.2GHz CPU was used in [128], compared to the 2.6GHz CPU used in our work

	Processing time (s)	
Processing Step	**Sun et al. [128]**	**This work**
One 3D frame processing	15	1
One probe recognition	5	0.73
Full video processing - 100 frames	1500	90

The computational aspect is evaluated in Table 4.8, which reports the processing time of the proposed pipeline compared to [128]. From the Table, it emerges the proposed approach is less demanding in processing time. While the method presented in [128] includes time-consuming mesh processing steps, such as conformal mapping, generic model adaptation and vertex-level tracking across the video, our approach benefits from the subspace modeling methodology and sparse coding techniques over the underlying manifold to keep the approach computationally cheap. In addition, it does not use manual or automatic landmark detection and tracking of the face.

4.6 Towards 4D face recognition in adverse conditions

Since most of current 3D dynamic datasets and 4D face recognition works are limited to one face recognition problem that is the facial expressions, several other important problems still not explored like pose variation, occlusion talking, walking, etc. In this section, we present a new 3D/4D dynamic facial database collected basically to address 4D face recognition challenges in real world scenarios. Also, the subspace metric-based approach is implemented to evaluate the performance of this approach under such difficult scenarios in 3D unconstrained videos.

4.6.1 The full 3D/4D face recognition database

All the available 3D dynamic databases are created to address the problem of facial expressions and action units recognition as it can be seen from the literature review in Section 2.8.2. This new 3D dynamic face recognition database implies several common challenges which have not been considered in 3D dynamic before. It is collected using single-view 3D *Artec* scanners with temporal resolution around 15 frames per second. This database can make a contribution in 4D FR research, especially for non-constraint scenarios.

There are 58 subjects in this database, 23 females and 35 males. The age average is 23 years old from different ethnics groups. For each subject, we collected first a full 3D static high-resolution model using the *Artec MHT* 3D scanner with the texture information. the average vertices in every 3D model about 50 k vertices. Figure 4.10 shows an example of a 3D static model from this database with and without texture information.

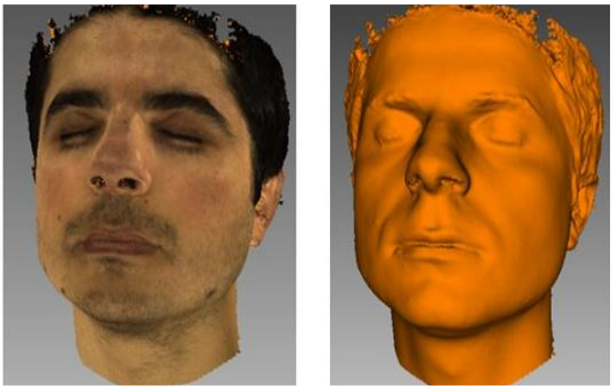

Figure 4.10: Full 3D static model from the database with and without texture information

Second, for every subject, eight 3D videos of 20 seconds using the *Artec L* 3D scanner are recorded. These eight videos are: one segment for the following challenges: facial expression, talking, walking, internal occlusion by hand or hair, external occlusion by scarf or sunglasses, and multiple persons (two or three) and two neutral videos. We refer by neutral that there is no facial expressions or occlusion. All of these videos were recorded under free pose variation in front of the scanner. Since the scanner has been used is a single view scanner, we have only a part of the face in many phases of the video as it is depicted in Fig. 4.11. This database is available for public research use.

More details about the acquisition protocol, settings and database properties, are published in [5].

Figure 4.11: The 3D dynamic sequences acquired under different conditions: a) neutral; b) expressive; c) talking; d) internal occlusion by hand; e) external occlusion by sunglasses; f) walking and g) multiple persons.

4.6.2 Preliminary experiments and results

To validate this new database, we applied a metric-based subspace learning approach to recognize the identity similarly to the framework proposed in Sect. 4.3. Here, different techniques are used to address the new challenges of this new dataset. The problem of pose variation is solved by dividing the long video into short videos of size 15 frames (i.e. about 1 second) where they have one nearby pose. The normal vector is estimated at each vertex for two reasons: First, to make a dense correspondence between successive frames using normal shooting technique presented in [30]. This normal shooting was used to track roughly the vertices through one subsequence to have registration. Second, the map of z component values of the estimated normals are used as a spatial feature vector for every frame. Before that, a down-sampling step is applied to each frame, to produce a constant number of n vertices per frame. These feature vectors are vectorized as columns of one matrix and $k-SVD$ orthogonalization is applied to find the subspace representation of the original data. These steps are illustrated in Fig. 4.12.

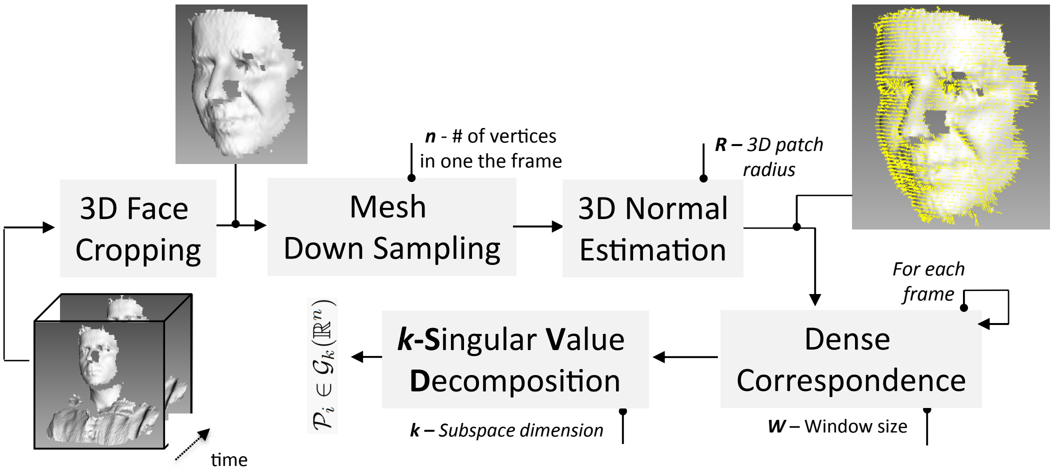

Figure 4.12: Overview of 4D to 4D FR approach under adverse conditions.

As a result of this pipeline, each 4D fragment is viewed as an element of the Grassmannian, and the original problem of 4D-to-4D matching in turned into a distance measurement on Grassmann which can be formulated as follows:

$$g^* = \arg\min_i \, d_{Geo}(\mathscr{X}_{probe}, \mathbb{X}_{g_i}), \tag{4.2}$$

where $d_{Geo}(.,.)$ denotes the geodesic distance between two linear subspaces, and g^* is the closer fragment in the gallery set $\mathbb{X}_{g_i}$ to the probe fragment $\mathscr{X}_{probe}$ according to the used distance. Furthermore, using the Riemannian geometry on Grassmann manifold makes it possible to use other mathematical computations, such as mean computation and k-means clustering explained in Sect. 3.4. As it is explained above, to solve the problem of pose variations the sequence of each subject in the gallery is divided into multiple instances over time. The same procedure is applied to probe sequences. Thus, each 3D temporal fragment of a probe will be compared with all 3D temporal fragments in the gallery. This exhaustive search can be avoided by applying k-means clustering algorithm on the gallery instances to cluster them according to the main pose of the 3D frames. After applying this unsupervised clustering, each cluster uses the *Karcher* mean [73] algorithm on all elements included in the cluster to have a representative mean subspace. In this way, each probe sequence is compared just with the clusters' representative in order to recognize the probe pose first, and then it will be compared only with gallery fragments that have the same pose only. Figure 4.13 illustrates instances from the same subject or from different subjects that have similar poses grouped in the same class.

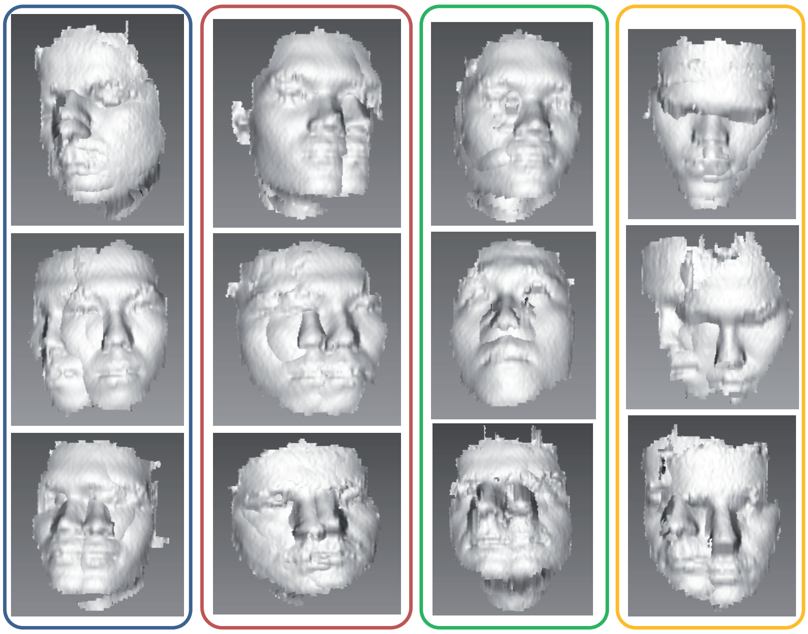

Figure 4.13: Each column shows instances belong to different subjects clustered together due to their nearby poses.

For evaluation, we considered the 58 subject of the database. One of the neutral 3D dynamic sessions is taken as a gallery. After dividing the 20 seconds videos into short subsequences, there are 20 subsequences resulted for every subject, and they are clustered into 5 classes according to their poses. The mean of each class is computed offline as well. For testing, 4 scenarios are considered: Neutral (Ne), Facial expression (Fe), talking (Tk), and external occlusion (EO) for every subject. The subject 3D video that contains these scenarios are tested separately after dividing every video into 20 subsequences as has been done for the gallery. Recognition process includes comparing the probe subsequence with the mean of gallery clusters to estimate its pose first, then comparing it with the instances that belong to this pose to find the identity. Applying majority voting concept to have more robust decision using more than one instances is implemented. The obtained recognition rates for these four scenarios (Ne, Fe, Tk, and Eo) are equal to $72\%, 62\%, 65\%$, and 36%, respectively.

Although the results obtained from this dataset is lower than those have been obtained on BU-4DFE database, the considered challenges in each scenario are more difficult, and these primary experience results can be improved by adopting more advanced techniques for faces registration, feature extraction, and learning. More details about this experimental study can be found in [6].

4.7 Conclusion

In this chapter, we have proposed a comprehensive 4D face recognition framework, which adopts a subspace-learning methodology and exploiting the efficiency of low dimensional subspace compact representation of the high dimensional data. While this direction has been widely used in the 2D domain, to our knowledge, this is the first study which brings it to the 3D domain for face recognition problem. As a contribution to our study, we have demonstrated that the shape dynamics (behavior) improves the recognition accuracy. This conclusion is valid even if the training samples (in the gallery) and the probes (to be recognized) present a different facial behavior. Leveraging the geometry of Grassmann manifolds, relevant geometric tools, and advanced Machine Learning tools, i.e., dictionary learning and sparse coding on the underlying manifold and comparing its performance with intrinsic learning methods like the Karcher mean computation. This approach is capable of managing face recognition from dynamic sequences of 3D scans in an effective and efficient way. The main advantages of this framework are: It is completely automatic and computationally less demanding compared to the current literature. Evaluation on BU-4DFE database is conducted, and obtained results outperform previous approaches under the expression-dependent setting and better performance than 2D video and 3D static based approaches. An empirical analysis for proposed approach parameters is reported as well. The importance of exploiting more than one instance to make recognition decision (majority voting through the time) advantage validated on expression independent scenario. A performance comparison of the different defined distances in Grassmann Nearest Neighbor classifier shows the superiority of projection distance over all others.

To bring face recognition from 3D dynamic sequences to more realistic scenarios, new 3D/4D facial database has been collected containing several challenges like pose variation, facial expressions, talking, walking, internal and external occlusion and multiple persons in the scene. Experimental analysis for a primary metric-based subspace learning approach for 4D to 4D face recognition on this new challenging database is reported.

In the following Chapter 5, we will address another main application for 3D dynamic sequences analysis which is spontaneous emotional states and pain affect early detection from depth and 3D high-resolution dynamic data by analyzing trajectories of subspaces on Grassmann manifolds.

CHAPTER 5

Spontaneous Emotion Detection in 4D Data

5.1 Introduction

One of major field of interest in facial sequences analysis is emotions and affects recognition and detection. Most of the current facial expression recognition works in the community consider the six prototypical (basic) expressions derived from psychological study proposed by Ekman [44] and they include *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise* which are collected in acted manner. These posed expressions are different from spontaneous and genuine expressions that are more complex. A more recent alternative to the hard categorical description of human affect is the *dimensional description* [111] in which an affective state is characterized in terms of a small number of latent dimensions, rather than a small number of discrete emotion categories. The dimensional description of emotions is shown in Fig. 5.1 using the *Arousal-Valence* chart. On the horizontal axis, the evaluation dimension is accounted, from displeasure to pleasure; on the vertical axis, the activation dimension is accounted through the arousal state, varying from low-to-high.

In this chapter, we exploit 3D dynamic data representation on Grassmann manifolds as **trajectories**, for the purpose of online spontaneous emotion detection, such as happiness or physical pain from depth or 3D videos. Our approach consists of mapping the video streams onto a Grassmann manifold (i.e., space of k-dimensional linear subspaces) to form time-parameterized trajectories. To this end, depth videos are decomposed into short-time clips, each approximated by a k-dimensional linear subspace, which is in

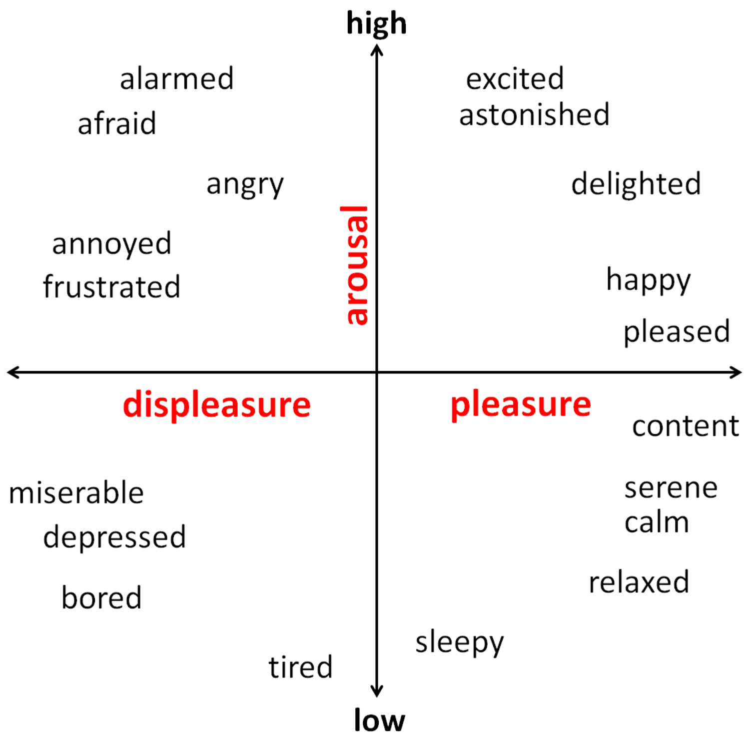

Figure 5.1: Dimensional Arousal-Valence chart of human emotions.

turn a point on the Grassmann manifold that captures the embodied information in the video at that portion. Then, the temporal evolution of subspaces gives rise to a precise mathematical representation of trajectories on the underlying manifold. In the final step, extracted spatio-temporal features based on computing the velocity vectors along the trajectories, termed Geometric Motion History (GMH), are led to an early event detector based on Structured Output SVM. The SO-SVM enables online emotion detection in the 3D video from partial and complete data. Experimental results obtained on the publicly available Cam3D Kinect [90] and BP4D-Spontaneous databases[155] validate the proposed solution. The first database has served to exemplify the proposed framework using depth sequences of the upper part of the body (4D-bodies) collected using depth-consumer cameras, while the second database allowed the application of the same framework to physical pain detection from high-resolution and long 4D-face sequences.

The rest of the chapter is organized as follows: In Sect. 5.2, we outline the main ideas and contributions of the proposed approach; A discussion of the 3D video representation adaptation to an early event-detector framework, which permits emotion detection from a 3D dynamic sequence is presented in Sect. 5.3; The pain detection from 4D data is presented in Sect. 5.4; We showcase the potential of the proposed solution in Sect. 5.5, by reporting results on the Cam3D Kinect database and BP4D-Spontaneous high-resolution database; Finally, our conclusion is in Sect. 5.6.

5.2 Methodology and contributions

In this chapter, we propose an online approach that detects the emotional state from 3D dynamic data as early as possible. The proposed framework is evaluated on two challenging problems: *(a)* Early detection of spontaneous emotional states from depth sequences of the upper part of the body (*depth-bodies*) acquired with a low-resolution sensor. Here, the spontaneous emotions derived from the dynamics of facial expressions and upper body gestures together; *(b)* Early detection of spontaneous physical pain affect from 4D high-resolution facial sequences (*4D-faces*).

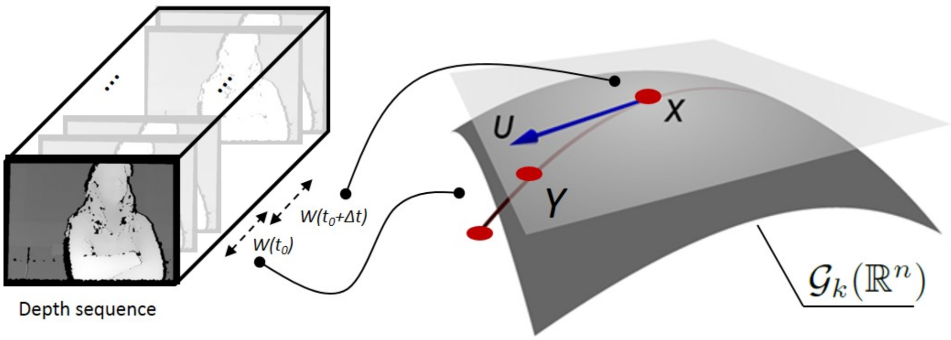

Figure 5.2: Dynamic depth data representation as trajectories on the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$. The streams of depth data at the left, are mapped to associated trajectories on the Grassmannian (right).

The main contribution has been introduced here is a new representation of human space-time 3D/depth data and relevant processing tools. In fact, several inherent challenges arise when analyzing depth videos. The most relevant one derives from the non-linearity of the space-time data. The non-linearity is caused by face deformations or the body gestures. In addition, the rigid transformations, mainly rotations and translations, which span other challenging problems, like missing data. In fact, human body acquisition using depth sensors or single-view 4D scanners includes auto-occlusions (the occlusion of the body by itself). In the literature, solving these issues requires pose normalization as well as temporal registration along the depth-video, which are time-consuming when processing dense data [128]. In our proposed approach, we account for the non-linearity of the data and related transformations as follow: First, we assume linearity in a local (short) time interval, by grouping the depth frames into subsequences of predefined length and regarding each group as a linear subspace (i.e., span of an orthonormal basis, represented by a matrix). This matrix gives rise to an element on a specific well-known Riemannian manifold (Grassmannian manifold); Then, we generalize it to longer videos using curves (i.e., non-linear) on the underlying curved manifold. This manifold-mapping allows faithfully representing the original depth and 3D video data in a cheaper and effective way, and it also shows robustness to noisy and missing data. This latter aspect makes the proposed representation suitable for processing and analyzing videos acquired with depth-consumer cameras, which suffer from low-accuracy and noisy depth measurements as well as incomplete data. Finally, using a Structured Output SVM (SO-SVM) based on sequential analysis of Euclidean spatio-temporal features, our framework is also endowed with online affect state detection capability, thus permitting early event detection.

Figure 5.2 summarizes the idea of mapping short-time depth video clips to a Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$, where k is the dimension of subspaces, and n the ambient space dimension. The positions of points corresponding to successive clips capture the temporal evolution (i.e., dynamics) of the face or the body in 3D videos, shown as a trajectory on the manifold. In particular, the temporal evolution of neighboring points across the trajectory is regarded as a one-dimensional feature vector, called *Geometric Motion History (GMH)* descriptor, which constitutes the input to the SO-SVM early event detector. In summary, the main contributions of this part are:

- A novel representation based on trajectories on Grassmann manifold suitable for modeling 3D/depth sequences and inherent human motions (deformations, gestures, etc.) of non-linear nature;

- A new space-time features vector termed *GMH*, which captures the spatio-temporal information to analyze the dynamic facial or body data 3D data;

- An adaptation of the early event detector developed by Hoai and De la Torre [63] for sequential analysis of Grassmann trajectories. In so doing, we report a clear benefit in early spontaneous emotion detection using the upper part of the body, rather than the face alone, and the efficiency in pain affect detection from 3D high-resolution facial expression sequences.

The proposed framework is also the first one, to our knowledge, capable of addressing early detection of spontaneous emotions in a complex scenario that includes:

- Depth sequences of the upper part of the body acquired with a cost-effective Kinect camera;
- Spontaneous emotions acquired without a rigid protocol (i.e., no assumption on the time when the emotion occurs in the sequence);
- Emotional state detection does not depend only on the temporal dynamics of the 3D face deformations but also on the upper part of the body, including shoulders and arms;
- Early detection of spontaneous physical pain from 4D high-resolution sequences.

5.3 Emotion detection from Kinect depth-bodies

In this scenario, videos of the upper part of the body (face, neck, shoulders and arms/hands) are acquired using a depth-consumer (Kinect) camera.

The first processing step consists in segmenting the upper part of the body from the background in each depth frame of the observed videos. Then, the sequence of the cropped upper body is divided into successive short-time clips, based on a temporal window size ω. For each clip, the cropped depth data (of each frame) of the body are reshaped to a vector of size n, which is then arranged to a matrix $X \in \mathbb{R}^{n\times\omega}$. Applying k-SVD to X, i.e., $X = U\Sigma V^T$, the subspace spanned by the columns of the matrix $U \in \mathbb{R}^{n\times k}$ is retained to represent the original clip. As a result, a video comprising m subsequences of size ω, and each of them is mapped to represent k-dimensional linear subspaces which lies on the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$. These points define a corresponding time parameterized trajectory on the manifold $\mathcal{T}(t)$ as discussed in Section 3.5.1, where every subspace here is a time instance.

This representation by trajectories on Grassmann manifold allows us to reduce the effect of the noise of the acquired depth data, and constitutes an efficient way to sequentially analyze the video streams (when observed) and extract relevant space-time features for online emotion detection. Our idea here is to compute first-order derivatives of the trajectory, and build a history of the motion including both deformations and pose changes. In so doing, the rhythm and the amplitude of the motion can be captured using the norm of the derivation.

5.3.1 Geometric Motion History (GMH)

In this work, we introduce a mono-dimensional feature vector to capture a spatio-temporal description for the 3D dynamic video from its representation as trajectory of subsapces on the Grassmann manifold. From this 4D-depth bodies, this GMH will be built from the instantaneous speed along trajectories as presented in Section 3.5.2. More in details, trajectories on Grassmann manifold can be analyzed by considering the evolution of their instantaneous speed. Given an observed portion of the trajectory $\mathcal{T}(t)$ in the time interval $[0, t]$, the instantaneous speed can be computed as the distance between neighboring points $\mathcal{X}(t)$ and $\mathcal{X}(t+\delta)$ along the trajectory, where δ is an integer constant added to control the evolution step between considered subsapces of the trajectory. The length of the shortest path is computed (Geodesic distance) on Grassmann manifold between the elements of the trajectory with step δ to build the *Geometric Motion History (GMH)* that characterizes the temporal motion of this 3D dynamic video. For an experimental validation of using Grassmann manifold, the same GMH is also built on Stiefel manifold using the Frobenius norm distance, given in Eq. 3.9, as it will be seen in the experimental Section 5.5. Figure 5.3 plots the *GMH* feature vectors obtained for three different depth videos, where the green segment corresponds to the emotion of interest while the black *GMH* segments are obtained for other different emotions. The similar shape exhibited by the *GMH* descriptors in the three cases for the emotion of interest in the middle can be appreciated.

5.3.2 Structured output learning from sequential data

The principle idea behind early detection from sequential data is to find the correct classifier capable of providing a recognition decision from both partial and complete events. This should permit recognition of the emotion of interest while receiving the sequential data and also provide its initial and ending boundary. To this end, in this work, we adopted the Structured-Output SVM (SO-SVM) [63], motivated by some interesting aspects of this classifier: 1) it can be trained on all partial segments and the complete one at the same time; 2) it allows us to model the correlation between the extracted features and duration of the emotion; 3) no previous knowledge is required about the structure of the emotion; 4) it can give better performance than other algorithms in sequence-based applications [99].

Assume a set of *Geometric Motion History (GMH)* feature vectors are computed. Each resulted *GMH* feature vector includes an emotion of interest, which is annotated by a

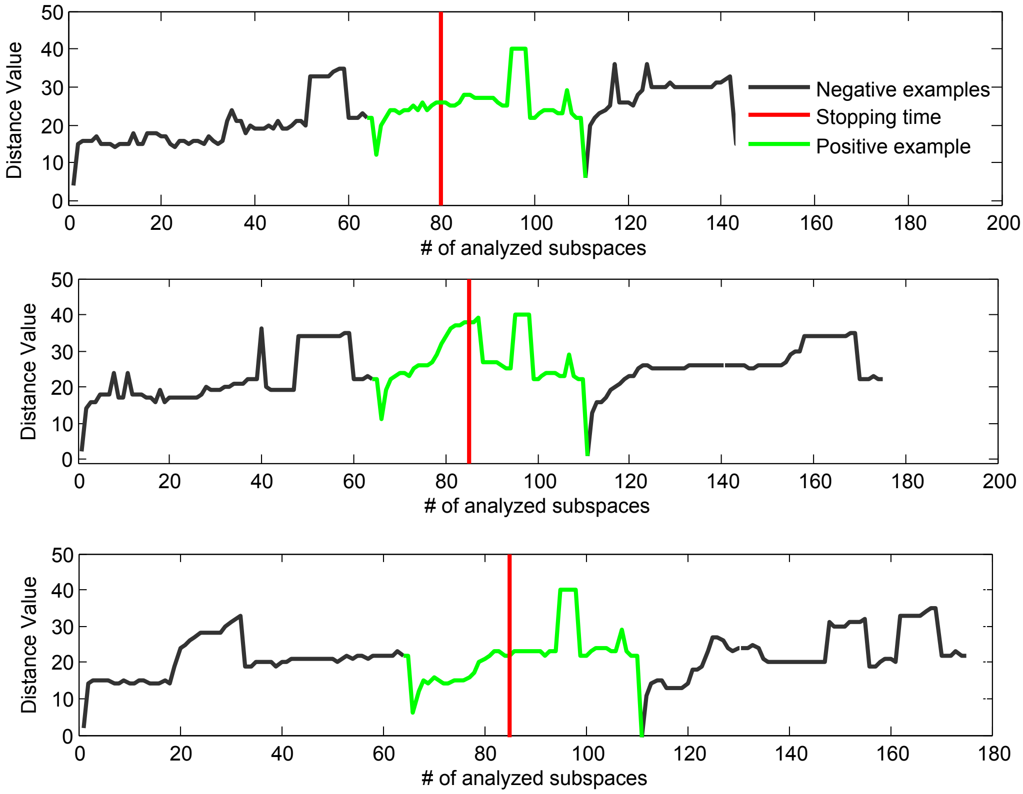

Figure 5.3: Three examples of the *Geometric Motion History* feature vectors extracted using the proposed framework.

pair of values $\langle s^i, e^i \rangle$, representing the start and end time of the emotion, respectively. At any time t^i comprised within the start and end of the emotion $s^i \leq t^i \leq e^i$, all partial emotions sub-segments obtained between $[s^i, t^i]$ will be used to train the structured output early event detector, since these different size sub-segments represent positive samples. All the other parts of the *GMH* are considered, instead, as negative samples. Another important aspect in SO-SVM early detection that always the more complete emotion portion of the video has a higher functional score than the less complete one as depicted in Fig. 5.4

The expected performance from SO-SVM in the testing stage is to fire the detection of the emotion of interest as soon as possible (after it starts and before it ends). As an example, Fig. 5.3 shows (in red) the detection times at which the early detection of the emotion from depth video is performed online. The problem of size variation between the partial segments of the emotion and the complete one is solved by computing the

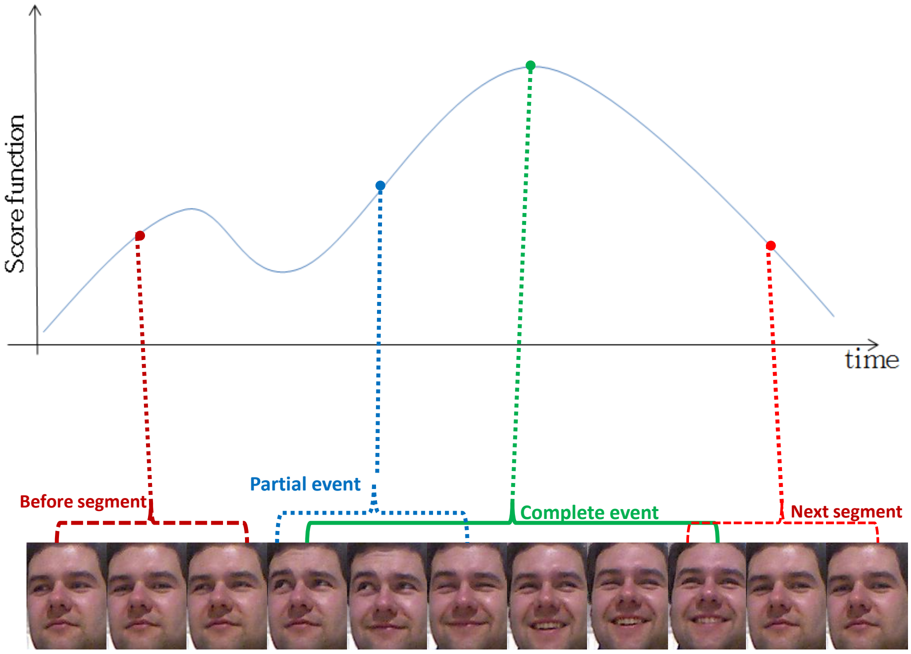

Figure 5.4: Online early detection score for happiness emotion from dynamic data

L_2-normalized histogram for each *GMH* segment to pass it to the SO-SVM. More details about this SO-learning framework can be found in [63]. The task of emotion detection is formulated as an early detection problem, which aims to detect the emotion of interest as quick as possible. This is achieved using SO-SVM, which results in a convex optimization problem [130].

Algorithm 6 summarizes the steps of our proposed method for early emotion detection from depth bodies.

5.4 Physical pain detection from 4D-faces

In this Section, we present a different adaptation of our trajectory based framework to the scenario of spontaneous physical pain detection from high-resolution 4D scans. Two different representations of the facial data are used here. First, the 3D landmarks-based method that uses the 3D facial keypoints available in the video (as a baseline), and the depth frames obtained from the 3D high-resolution scan. Since the detection of physical pain from the face is related to slight and local facial expressions, we proposed to create the Geometric Motion History (GMH) of the 3D video not only by geodesic distance but by using the complete information available in the velocity vector between

Algorithm 6 – Online emotion detection from 4D depth-bodies

Require: Depth bodies videos set $\mathbb{S} = \{S^i_{m_i}\}_{i=1}^{M}$, of size M; every S^i has m_i frames; ω is the window size

Initialization:

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

$\hat{S}^i \leftarrow S^i$ *//Depth preprocessing and taking the region of interest*

$\mathbb{X}^i = \{X^i_1, X^i_2, ..., X^i_N\} \leftarrow \hat{S}^i$ *//Dividing video into N successive subsequences of size ω*

$\mathcal{T}^i(t) \leftarrow kSVD\{X^i_t\}_{t=1}^{N}$ *//Subspace representation of the subsequences as trajectory*

$GMH^i(t) \leftarrow d_{Geo}(\mathcal{T}^i(t), \mathcal{T}^i(t+\delta))$ *//GMH building by computing geodeisc distances between successive subsapces*

end for

Processing:

$D\{i\} = [GMH_L \mid GMH_M \mid GMH_R]$ *//GMH Concatenation with emotion of interest in the middle*

$Label\{i\}] = [s\ ,\ e]$ *//GMH_M start and end points indexes*

Model = SO-SVM(D_{tr},$Label_{tr}$) *//SO-SVM Training*

y* = SO-SVM(D_{tst},Model) *//SO-SVM Testing*

Ensure: y*=[s*,e*] *//Emotion of interest detected boundaries*

two subspaces in the trajectory. To this end, we implemented the transported velocity vector fields formulation presented in Sect. 3.5.3. By this implementation, we intend to illustrate the utility of considering the information carried in velocity vectors to capture densely the deformations. Figure 5.5 shows the landmarks and the depth image with their corresponding 2D texture image that belong to one 3D pain face, taken from the BP4D-Spontaneous expression dataset.

5.4.1 3D landmarks-based Grassmann trajectories

In this solution, we start from a sequence of high-resolution 3D face scans, each of which is labeled with l facial landmarks. The 3D coordinates (x, y, z) of the facial landmarks are considered as a simple baseline descriptor of the face so that each frame is represented by a vector in $\mathbb{R}^{3\times l}$. Starting from this representation, and following the same steps of Sect. 5.3 as dividing the video into subsequences of size ω, applying k-SVD to obtain a trajectory of subspaces for every 3D dynamic pain sequence $\mathcal{T}(t)$ on a Grassmann manifold $\mathcal{G}_k(\mathbb{R}^{3\times l})$. The 3D spatio-temporal information is then captured by computing

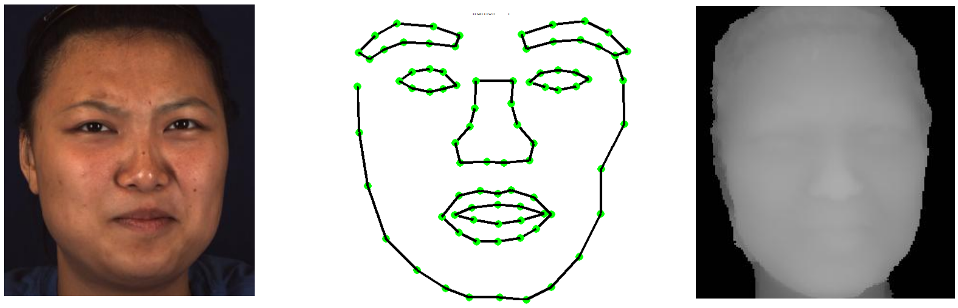

Figure 5.5: From left to right: color image; 3D landmarks; and depth image.

the geodesic distance between successive subspaces by step δ to build the *Geometric Motion History* from dynamic 3D landmarks.

In addition, the change in the instantaneous speed along a trajectory due to both facial deformations between two subspaces of the trajectory with interval δ and the pose variations can be observed. This latter effect is the dominant one in Fig. 5.6, due to a strong pose variation. This represents a problem for emotion recognition from the facial deformation that is addressed by pose normalization as it will be detailed in the experimental part.

From this one-dimensional vector derived from 4D high-resolution facial data using $\delta = 1, 3, 6$ can be observed. The importance of selecting an appropriate value of δ emerges clearly from the Figure. It is evident that the signal resulting for $\delta = 1$ is noisy while the informative change in the subsequence is clearer for $\delta = 3$. Further increasing δ to 6 can cancel information about the emotional evolution through the video.

This solution uses local and sparse information of the 3D shape of the face, and will serve as a baseline to compare with the dense 3D shape representation using depth images.

5.4.2 Depth-based Grassmann trajectories

In this case, we produce a depth image from each 3D model after preprocessing and cropping the facial area. As mentioned earlier, a depth map gives a complete shape description of the face, rather than only the 3D landmarks. Following the same procedure as previously, every subsequence of ω depth frames is modeled as a k-dimensional subspace of $\mathbb{R}^n$, being n the image size after vectorization. This permits us to build a time-parametrized trajectory $\mathcal{T}(t)$ of subspaces on the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$,

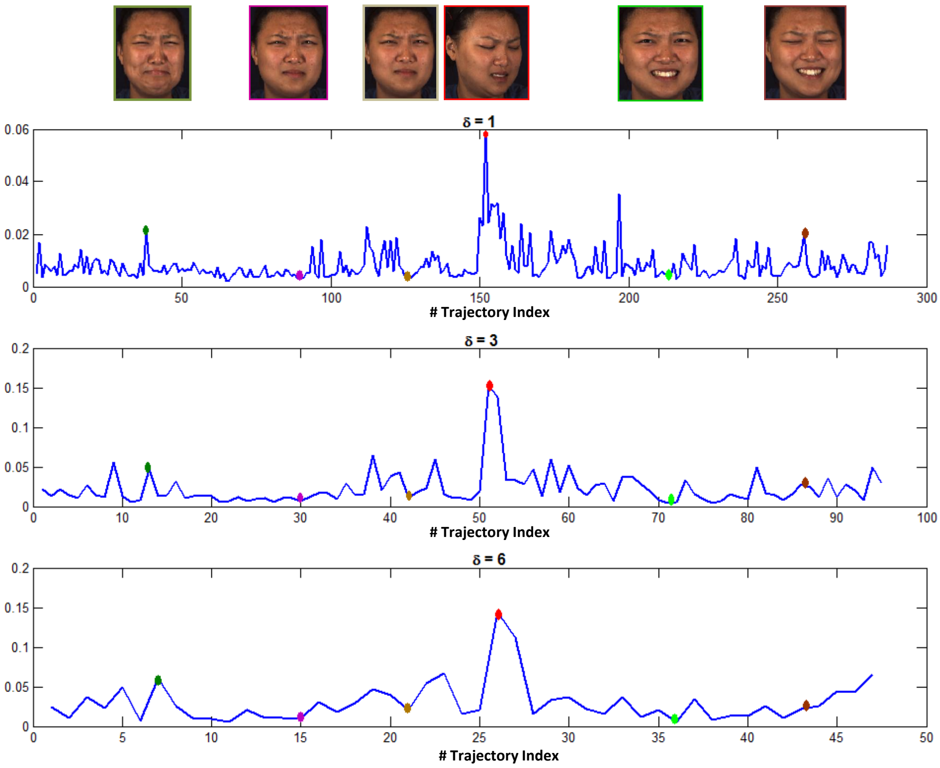

Figure 5.6: The instantaneous speed along a trajectory on Grassmannian manifold computed for a pain depth flow for different values of $\delta = 1, 3, 6$.

similarly to the case studied in Sect. 5.3. In this scenario, in addition to build the *GMH* by computing the geodesic distances between successive subspaces, like in the landmarks representation method, we introduce a more efficient spatio-temporal representation of the facial dynamic data by proposing the *Local Deformation Histogram (LDH)* descriptor that follows the transported vector field formulation.

More in detail, the *LDH* is computed through the following steps. First, the velocity vector V between successive subspaces of a trajectory $\mathscr{T}$ on the Grassmann manifold $\mathscr{G}_k(\mathbb{R}^n)$ is computed and transported to a fixed tangent space $T_I(\mathscr{G}_k(\mathbb{R}^n))$ at the identity element of Grassmann manifold. One possible representation of the parallel transported velocity vector $(V_i) \in T_I(\mathscr{G}_k(\mathbb{R}^n))$ is a matrix of size $n \times k$. Taking the k first columns of this matrix V_i as vectors of size n and reshaping them to the original dimension of the face depth image $\widehat{m} \times \widehat{n}$ gives rise to a k-first components. Visualizing these components as 2D images shows clearly the temporal deformation with respect to its spatial location in the depth image. The first component of the velocity vector contains informative motion data, where the rest contains noise and redundant data.

Then, rather than using the Grassmann distance that quantifies the speed along the trajectory, we propose to exploit the first component of the velocity vector between two subspaces. This new representation for the temporal evolution of the trajectory carries information not only about the speed of the deformation, but also about where and in which direction the deformation occurs as anticipated in Fig. 5.7.

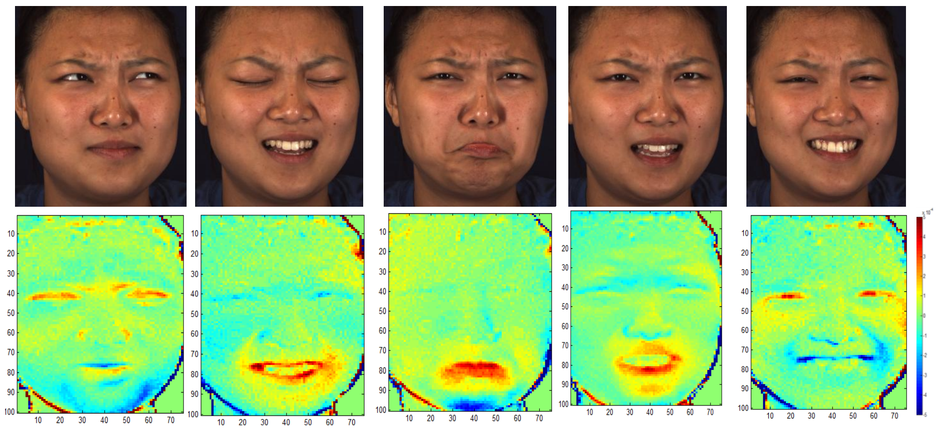

Figure 5.7: The visualization of velocity vectors first components between subspaces of one trajectory with their corresponding 2D texture images. The color maps show where the deformation happens in the face and its direction. Colors around green mean no deformation; from green to red: deformation in the positive z axis direction and from green to blue deformation in the negative z direction. The degree of the color indicates the deformation intensity.

This is illustrated in Fig. 5.7, where positive values indicate a deformation in the forward (positive z axis) direction, while negative values indicate deformation in the backward (negative z axis) direction. The scalar value also indicates the degree of deformation.

In a final step, the matrix is divided into blocks, thus permitting us to localize where the deformation happens in the face, and compute a dual value (positive/negative) histogram for each block. This dual-value histogram gives us an idea about the intensity and the direction of the deformation of the facial area associated with the block. Then, the concatenation of all blocks provides what we call the *Local Deformation Histogram* (LDH) for the velocity vector. The LDH vectors between each two successive subspaces will be concatenated sequentially to build the Geometric Motion History GMH out of the trajectory $\mathscr{T}$ on Grassmann manifold. Fig. 5.8 illustrates these steps.

The beginning and the end of the pain are decided according to certain annotated

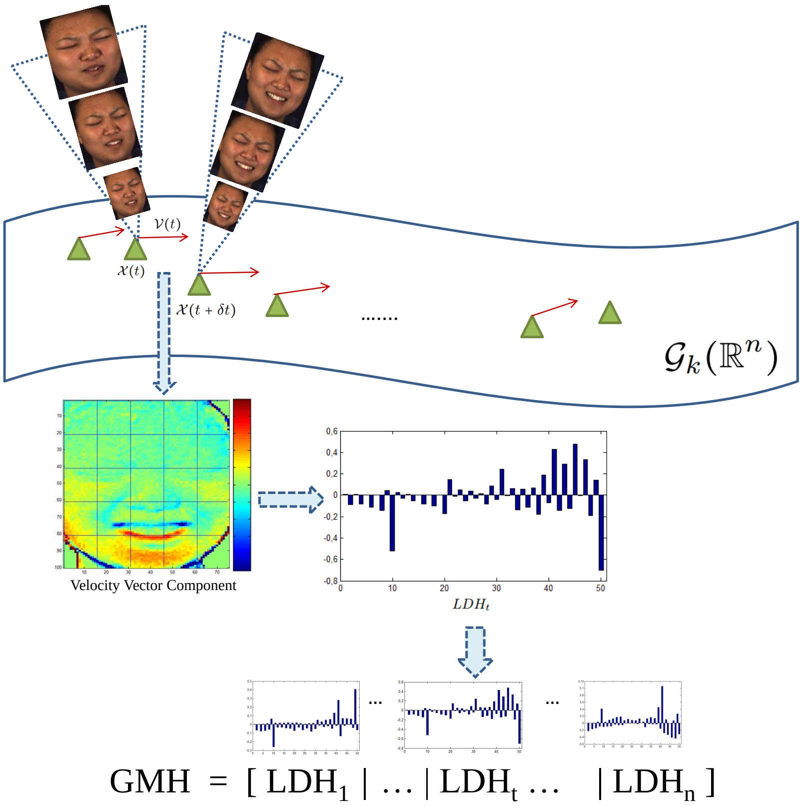

Figure 5.8: Illustration of LDH computation from the velocity vectors (red arrows) between subspaces (green triangles) of the same trajectory. Taking the first component of the velocity vector, and dividing the first component into 5×5 blocks, computing the dual value histogram for every batch and concatenate them together to have the LDH_t. Concatenating LDH for a ω frames gives rise the GMH feature vector, input of the SO-SVM algorithm.

facial action units combination (this aspect will be discussed in more detail in Sect. 5.5). The SO-SVM approach presented above will be used to detect the pain feeling as early as possible from the GMH features extracted from the landmarks and depth representation. Algorithm 7 summarizes the main steps of the pain detection approach from 4D high-resolution data using the local deformation histogram (LDH).

5.5 Experiments and evaluation

To validate the proposed framework, we have conducted several experiments of emotion detection on two different datasets. The first dataset captures depth-videos of the upper

Algorithm 7 – Physical pain detection from 4D-faces

Require: 4D facial scans set $\mathbb{S} = \{S^i_{m_i}\}_{i=1}^{M}$, of size M; every S^i has m_i frames; ω is the window size. $Labels\{L^i\}_{i=1}^{M}$, where $L^i[s,e]$ indicates the start and the end of pain affect in S^i

Initialization:

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

$\hat{S}^i \leftarrow S^i$ *//3D preprocessing and depth generation*

$X_i\{\mathscr{X}^i_1, \mathscr{X}^i_2, ..., \mathscr{X}^i_N\} \leftarrow \hat{S}^i$ *//Dividing video into subsequences*

$\mathscr{T}_i\{1,..,N\} \leftarrow kSVD(X_i\{1,..,N\})$ *//Trajectory building*

$\mathscr{V}_i \leftarrow Velocity(\mathscr{T}_i)$ *//Velocity Vectors between subspaces*

$\mathscr{V}^T_i \leftarrow Transport(\mathscr{V}_i)$ *//Tranportation to one tangent space*

$LDH_i\{1,..,N\} = LDH(\mathscr{V}^T_i)$ *//LDH from velocity vectors*

$GMH_i \leftarrow [LDH_i(1), LDH_i(2)...., LDH_i(N)]$ *//GMH building by concatenation of LDHs*

end for

Processing:

Model = SOSVM($GMH_{tr}, Labels_{tr}$) *//SO-SVM Training*

y* = SO-SVM(GMH_{tst},Model) *//SO-SVM Testing*

Ensure: y*=[s*,e*] *//Pain affect detected boundaries*

part of the body when spontaneous emotions or complex mental states, such as happiness and thinking are exhibited [90]. We will apply our framework on this dataset to obtain early detection of a spontaneous emotional state of interest. The second dataset consists of high-resolution 3D videos of faces also showing spontaneous emotions, like happiness, sadness, physical pain, etc. [152]. On this database, our experiments focus on early detection of spontaneous physical pain using different representations.

Two evaluation criteria are used to test the performance from the viewpoint of accuracy and timeliness.

- **Area under the ROC curve:** A ROC curve is created by plotting the *True Positive Rate* (TPR) vs. the *False Positive Rate* (FPR) at varying threshold; and the Area Under ROC Curve (AUC) gives the overall performance of the binary classifier to discriminate between positive and negative samples;

- **AMOC curve:** The *Activity Monitoring Operating Characteristic* curve is generally used to evaluate the timeliness of any event surveillance system. It gives an indicator of how fast the detection of the event is, by reporting the *Normalized*

Time to Detection (NTtoD) as a function of False Positive Rate (FPR). In particular, NTtoD is defined as the fraction of the event occurred at one-time instance. For an event starting at s and ending at e in a time series, if the detector fires the event at time t where $s < t < e$, the NTtoD is given by:

$$NTtoD = \frac{t-s+1}{e-s+1}. \tag{5.1}$$

5.5.1 Cam3D Kinect database

In the Cam3D Kinect database [90], Mahmoud et al. collected a set of 108 audio/video segments of natural complex mental states of 7 subjects. Each video is acquired with the Kinect camera, including both the appearance (RGB) and depth (D) information. The data capture natural facial expressions and the accompanying hand gestures. The emotional states are: *Agreeing*, *Bored*, *Disagreeing*, *Disgusted*, *Excite*, *Happy*, *Interested*, *Sad*, *Surprised*, *Thinking* and *Unsure*. These emotional states are more realistic and more complex than the basic well known six facial expressions that are commonly used in the literature. Figure 5.9 shows example frames for four different emotional states. It can be observed the subjects sit at a table in front of the camera showing the upper part of the body, including arms and hands, shoulders and face.

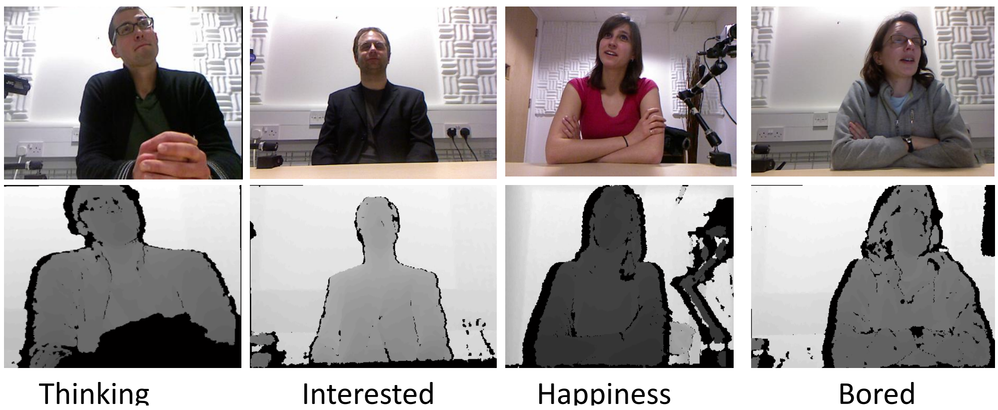

Figure 5.9: Cam3D Kinect database: Example depth frames with their corresponding 2D texture image of different emotional states.

Table 5.1 shows the number of available segments for each emotional state. It can be observed that videos in this dataset provide a sampling of the dimensional description chart of emotions as reported in Fig. 5.1. However, the possibility to use each emotion category in a detection experiment is hindered by the low number of videos comprised by several categories (i.e., less than 8 videos are present in 9 out of the 12 emotion categories, with 5 categories having just 1 or 2 videos). This motivated us to consider the following two experimental scenarios: *Happiness* vs. *others*; and *Thinking* / *Unsure* vs. *others*. Compared to the chart of Fig. 5.1, the first scenario tests the detection of an emotion of interest located in the *high-arousal* / *pleasure* quadrant (positive emotion); the second one refers to an emotion in the *low-arousal* / *displeasure* sector (negative emotion).

Table 5.1: Number of available depth videos for each emotional state

Emotional/Mental State	**# of depth videos**
Agreeing	4
Bored	3
Disagreeing	2
Disgusted	1
Excited	1
Happy	**26**
Interested	7
Neutral	2
Sad	1
Surprised	5
Thinking	**22**
Unsure	**32**

5.5.2 Emotional state detection

We applied the method using speed along trajectories on the manifold (see Algorithm 6) to detect emotional states from two different regions of the dimensional *Arousal-Valence* emotion chart of Fig. 5.1: (1) *Happiness* out of all non-happiness, i.e., *Happiness* vs. *others* (high-arousal/pleasure quadrant); (2) *Thinking* / *Unsure* vs. *others* (*low-arousal* / *displeasure* quadrant). In both the experiments, the videos of the emotion of interest and the videos of the other emotions are divided equally into two halves, one used for training and one for testing. Then, the *Geometric Motion History* feature (GMH) of each video is computed by dividing the video into subsequences of size $\omega = 20$ and subspace dimension $k = 5$. These setting have been chosen empirically. Then, the GMH of the emotion of interest is concatenated with the GMH computed for two videos of different emotional states. Selecting these videos randomly for each concatenation, permitted us to obtain more training and testing data. Some examples of this process are illustrated in Fig. 5.3. Using this setting, we derive a total of 100 GMH for training and the same number for testing. For each generated sequence, the start and the end point of the emotion of interest are known. Experiments in the following explore different aspects of the proposed approach.

In a first experiment, we compare the performance of our trajectory sequential analysis framework using the *Geometric Motion History* feature computed for Grassmann and Stiefel manifold. For the *Happiness* vs. *others* case, Fig. 5.10 shows the ROC and the AMOC curves obtained. From the ROC curves related to the Grassmann, it can be observed that when the FPR is around 20% the TPR reaches 90% for *Happiness* detection. This accuracy decreases significantly (around 50%) at FAR=10%. Comparing the analysis of the trajectories along the Stiefel (dashed curves) and the Grassmann manifold (continuous curves), it clearly emerges the sequential analysis performed on Grassmann manifold outperforms the analysis on Stiefel manifold. The areas under ROC curves are 0.73 and 0.84 on Stiefel and Grassmann, respectively. The same findings can be concluded from comparing Stiefel and Grassmann manifolds for *Thinking/Unsure* emotional state in Fig. 5.10.

This demonstrates the consistency of the subspace based representation $\mathscr{Y} = Span(Y)$ and the associated metric $d_{\mathscr{G}}$ over the matrix representation. This is mainly due to the invariance of the subspace representation to the rotations $O(k)$ as $\mathscr{G}$ is a quotient space of $\mathscr{V}$ under the group action of $O(k)$. The plots on the right of Fig. 5.10 show the evolution of the system latency (the fraction of video needed to make the binary decision) against FPR. For example, the detector achieves 20% of FPR by analyzing 20% of the video segment. Also, in this case, results reported for the Grassmann representation are better than results obtained from the Stiefel representation.

From Fig. 5.10, it is also possible to compare detection accuracy results for *Happiness* and *Thinking/Unsure*. In particular, the *Thinking/Unsure* detection shows a performance decrease with respect to the *Happiness* detection results. The area under the ROC curve is 0.66 and 0.79 on Stiefel and Grassmann manifold, respectively, for *Thinking/Unsure*, while they are 0.73 and 0.84 for *Happiness*. These results confirm the advantage in using the Grassmann rather than the Stiefel representation. From the plot on the right of this Figure, it can be noted that about 20% of the negative samples are recognized to be an element of this class, even if the videos are observed completely. This can be motivated by the "common" neutral behavior exhibited by human when conveying other complex mental states (e.g., agreeing, bored, etc.). This induces a confusion to the

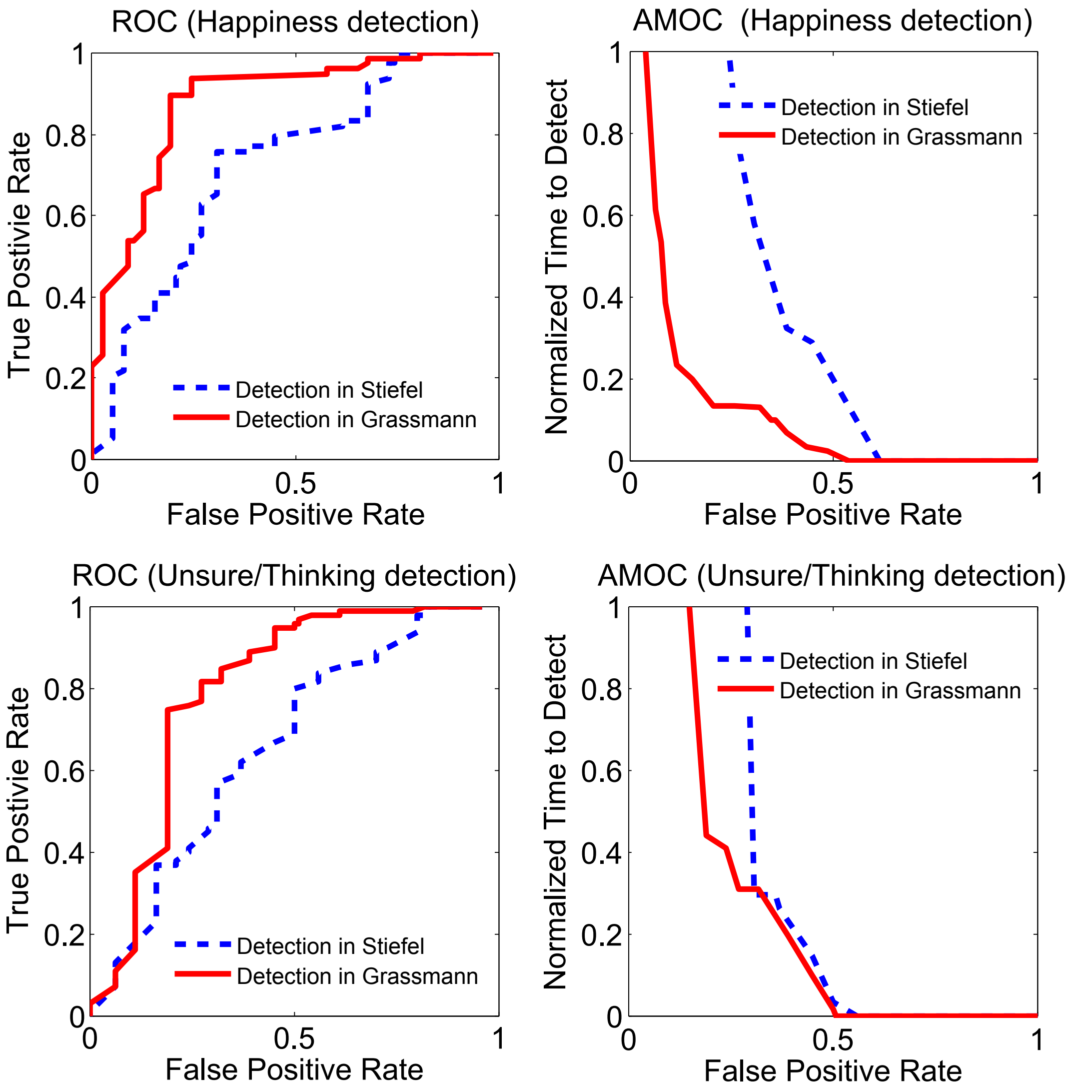

Figure 5.10: ROC and AMOC curves for *Happiness* (top) detection and *Thinking*/*Unsure* detection over Stiefel and Grassmann manifolds.

detector, which was not the case for the *Happiness* detector, as the happiness is often accompanied by body and facial expressions.

To investigate the importance of using the upper part of the body versus using only the face depth spatio-temporal information, we performed experiments with the previous protocol, but considering the upper body in the depth videos to construct the *GMH* on Grassmann manifold, instead of the cropped region of the face. From Fig. 5.11, it is clear that the emotional state exhibited by the upper part of the body is easier to detect than considering the facial region alone when acquired using cost-effective cameras. In the

Happiness experiment, the area under the ROC curve for the upper part of the body and the face only are 0.84 and 0.68, respectively. Performing the same experiment for the *Thinking*/*Unsure* case, the area under the ROC curve is 0.79 and 0.63 for the upper part of the body and the face only, respectively. This result is in agreement with studies like [94, 135], which encourage the use of the upper body with the face to have better performance in automatic emotional state understanding.

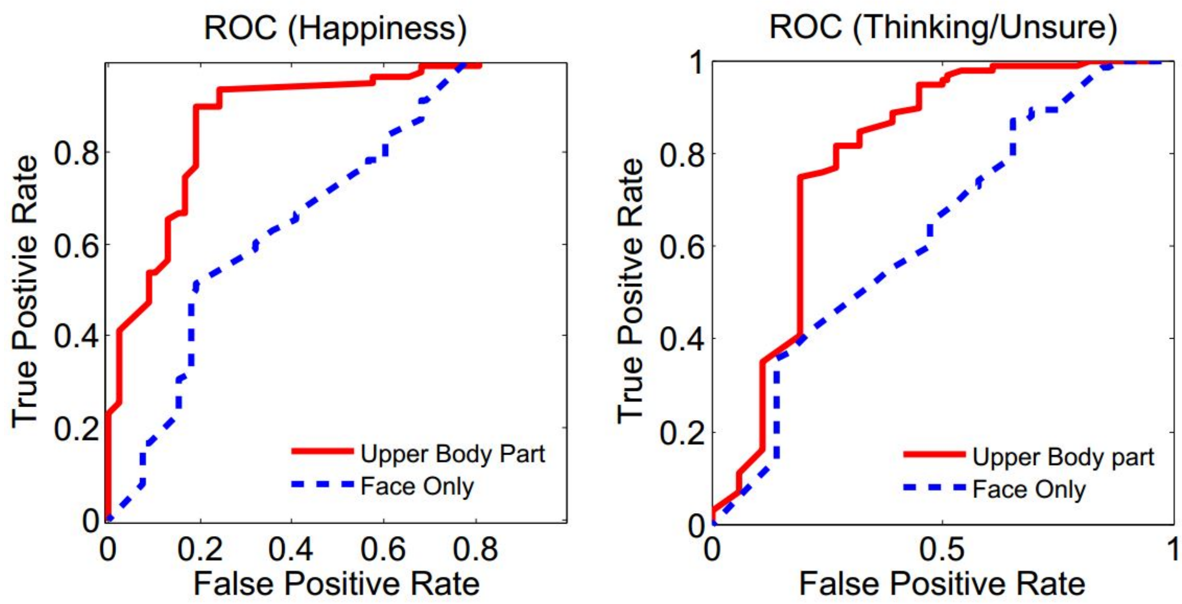

Figure 5.11: ROC curves comparison for *Happiness* and *Thinking*/*Unsure* detection over the Grassmann manifold using the upper body and the face only.

Finally, we also investigated the relevance of the window size (# of frames used to embody the motion in the subspace) and of the subspace dimension. In Fig. 5.12, we consider the Grassmann manifold for *Happiness* detection and compare results for windows of size $\omega = 20$ and $\omega = 5$ (red and blue curves, respectively). The dimension of the subspace is $k = 5$ in both the cases (we remember here, k is the number of singular-values used for the subspace representation). In the first case, with the window size of $\omega = 20$, using five singular values permits us to keep 90% of the original information of the temporal window (we selected this value by empirical experiment); in the second case ($\omega = 5$), we keep 100% of the information as $k = \omega = 5$. So, in this comparison the window size ω is the only changing parameter. The area under the ROC curve for $\omega = 5$ is 0.74, and 0.84 when $\omega = 20$. The observed performance gap between the two cases (a quite marked improvement is noted for $\omega = 20$), clearly evidences the importance of an appropriate setting of these parameters.

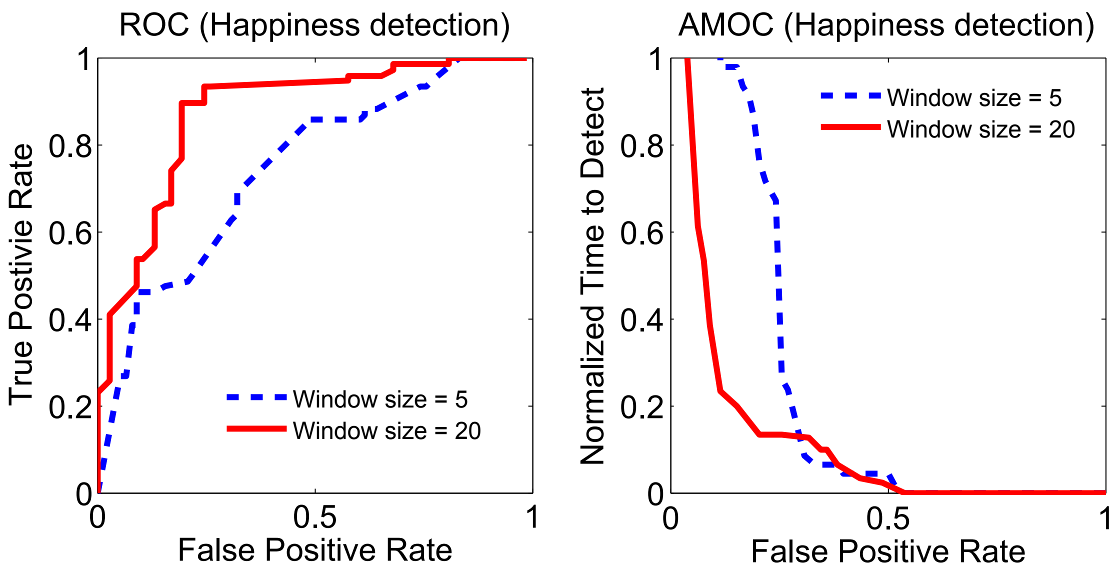

Figure 5.12: ROC and AMOC curves for *Happiness* detection over the Grassmann manifold for two different window size (i.e., $\omega = 5$ and $\omega = 20$).

5.5.3 BP4D-Spontaneous facial expression database

In [152], Zhang et al. proposed Binghamton-Pittsburgh 3D dynamic (4D) spontaneous facial expression database. This database includes 41 subjects acquired using Di3D dynamic face capturing system at 25 fps resolution for 3D videos. There are 8 different tasks for every subject corresponding to the following spontaneous expressions: *Happiness* or *Amusement*, *Sadness*, *Surprise*, *Embarrassment*, *Fear* or *Nervous*, *Physical pain*, *Anger* or *upset* and *Disgust*. This database provides the 3D model and the 2D still images for every video with metadata. The metadata includes for 2D texture images, the 46 landmarks annotation with the pose information, and for 3D models, 83 feature points (landmarks) annotation with the pose information given by the *pitch*, *yaw* and *roll* angles. Facial action units (FAUs) are provided for 20 seconds (about 500 frames) of every task. This AU annotation provides information about the fact a specific AU is activated or not in the frame and its intensity in the case of activation. Figure 5.13 depicts one 3D model with its corresponding 2D texture image for every task.

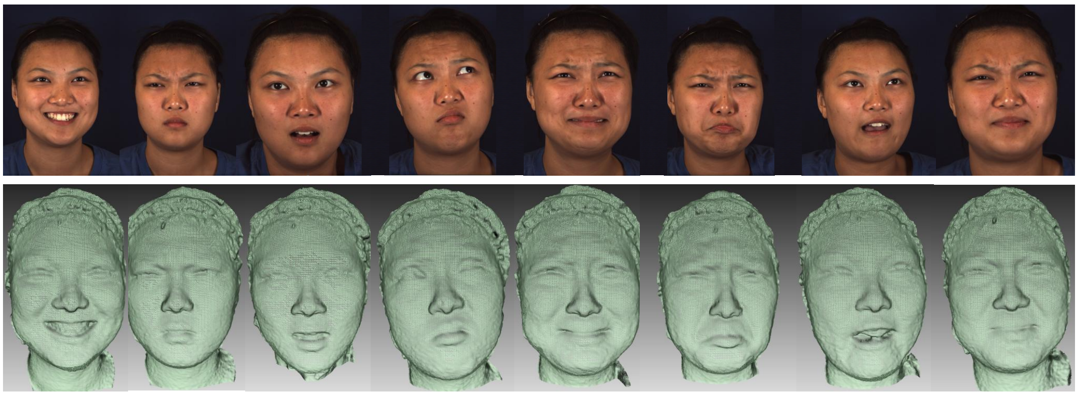

Figure 5.13: BP4D Database: Examples of the eight different spontaneous expressions (tasks) included in the database

2263 ### 5.5.4 Analyzing 4D-faces for physical pain detection

We applied the proposed geometric framework with transported velocity vector fields method as explained in Sect. 5.4 to detect spontaneous physical pain from 3D dynamic facial videos. The spontaneous physical pain is elicited by putting the participant's hand in ice water for each of the 41 subjects. The acquired 3D videos are quite long (their duration is about 20s), and it is known there is a pain emotion through the video, which constitutes our initial ground truth. To have accurate pain affect start and end points during the video as an emotion of interest, we use the FAUs provided annotation. Several studies have been conducted in psychology field to reveal the optimal AUs combination that can define the physical pain emotional state, like [70] where they found the AUs that can be activated in pain affect are those listed in Table 5.2. Parkachin and Solomonin [70] also proposed a pain intensity scale equation (PSPI) considering certain AUs:

$$Pain = AU4 + (AU6||AU7) + (AU9||AU10) + AU43\,. \tag{5.2}$$

Zhang et al. [155] made extensive study to show the mapping between the AUs and the targeted emotion on BP4D database, and they found that AUs {4, 6, 7, 9, 10} are the most common in pain videos. From these results, and the available AUs annotation, we decided the begging and the end of the pain in the videos using the following equation:

$$Pain = AU4 + (AU6||AU7) + (AU9||AU10)\,. \tag{5.3}$$

2264 which states that a *physical pain* is considered as existing if AU4 and (AU6 or AU7) and

Table 5.2: Possible AUs related to pain according to [70].

Action unit	Name	Action unit	Name
4	Brow Lowerer	20	Lip Stretcher
6	Cheek Raising	25	Lip parter
7	Eyelid Tightener	26	Jaw Dropper
9	Nose wrinkler	27	Mouth Stretcher
10	Upper Lip Raiser	43	Eye Closer
12	Lip Corner Puller		

(AU9 or AU10) are activated. Figure 5.14 illustrates the use of AUs combination for pain annotation according to Eq. (5.3).

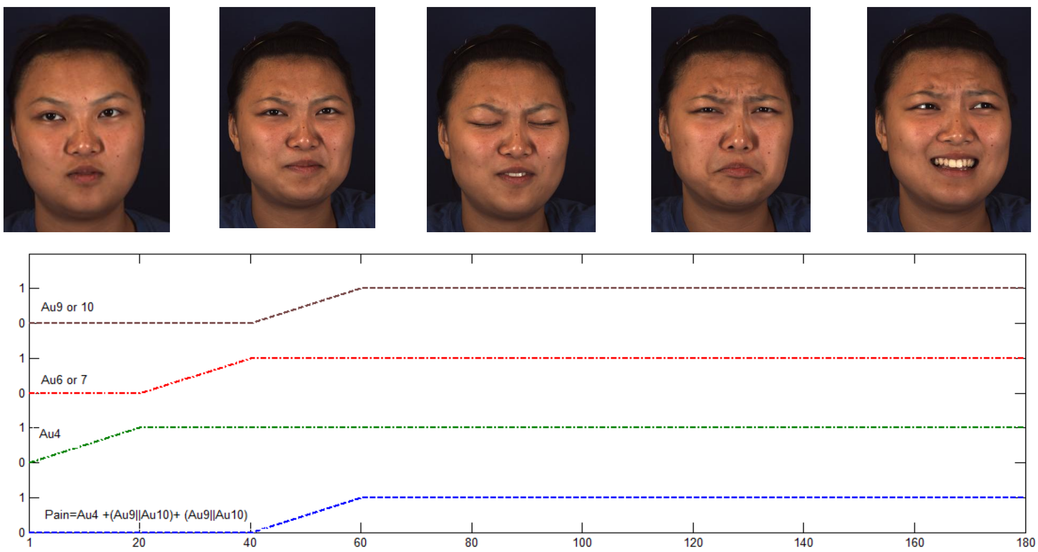

Figure 5.14: Illustration of AU activation during a physical pain video. The horizontal axis gives the frame index in the video, and the vertical axis provides the activation (i.e. of value 1) or non-activation of the AU (i.e. of value 0).

Based on the available AUs annotation in BP4D database, 28 subjects have been selected for the task of physical pain detection (task 6 videos). Half of these subjects (14) are used for training and the second half (14) for testing in the SO-SVM learning framework with the beginning and the end of pain emotion labels. There is no need for concatenation of *GMH* in these experiments since we have long 3D videos and the pain does not start immediately according to the eliciting protocol. Two methods have been investigated in this work to model the 3D video subsequences. Results, for both the cases, are reported in the following, using a window size $\omega = 6$ for deriving the linear subspaces.

3D landmarks-based (baseline) method

In this representation, we use the 3D coordinates (x, y, z) of the 83 landmarks available in BP4D metadata as a representative feature for every 3 frame after vectorizing these values to have a feature vector in $\mathbb{R}^n$ where $n = 83 * 3 = 249$. This approach is regarded as a baseline solution for our work. We model every subsequence of size $\omega = 6$ as one subspace after applying k-SVD, with $k = 2$. These settings are selected empirically. Two experiments are conducted using this representation to study the pose effects and the step δ on the trajectory.

To evaluate the pose normalization effect on the performance, we used the landmarks representation for pain detection from 3D videos with and without the pose normalization in order to investigate how the pose variation affects the pain detection accuracy. The pose is normalized by applying the inverse rotation of the 3D frame pose information given in the metadata. From Fig. 5.15, it is quite clear that the AUC with pose normalization (0.68,0.78,0.76) are higher than without pose normalization (0.63,0.75,0.70) for $\delta = 1, 3, 6$, respectively. These results confirm the negative effect of pose variation in our framework, because the facial deformation resulting from pain affect in correspondence to the landmarks is combined with the changes resulting from the pose variation.

GMH curves on Grassmann manifold can be affected by noisy changes that might occur due to raw data or errors in the registration step. To investigate this aspect, we considered the effect of different smoothing levels applied to the Grassmann trajectory, which corresponds to using different values of δ. This empirical analysis is conducted using the landmarks representation and normalized pose with $\omega = 6$ and $k = 2$. Table 5.3 shows the AUC values for pain detection with this setting for δ from 1 to 5. The best AUC value of 0.78 is obtained for $\delta = 3$. These results show that smoothed trajectories, corresponding to $\delta > 1$, provide better performance up to a certain extent, thanks to the noise removal, but large values of δ (e.g., $\delta = \{4, 5\}$) affect negatively the results, since informative changes along the time can be canceled.

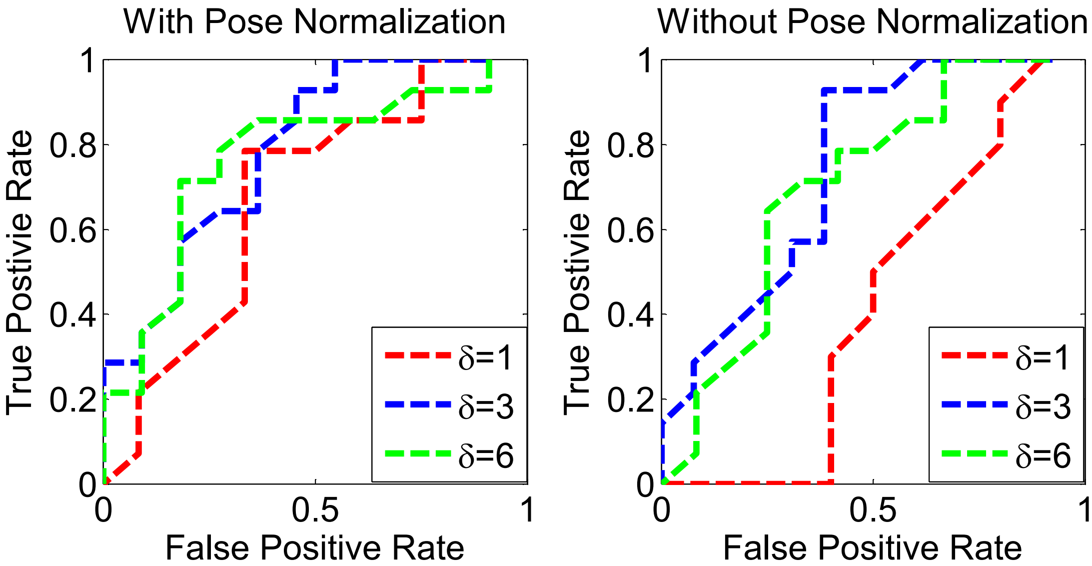

Figure 5.15: ROC curve for the landmarks method. The left plots show the ROC curves after pose normalization for $\delta = \{1,3,6\}$, while the right plots show the performance obtained without pose normalization.

Table 5.3: AUC values for the landmarks method, with and without pose normalization, for $\delta = 1,2,3,4,5$

value of δ	1	2	**3**	4	5
AUC – not normalized pose	0.63	0.69	**0.75**	0.71	0.70
AUC – normalized pose	0.68	0.72	**0.78**	0.75	0.74

Depth representation method

In this approach, the depth images of the face region are used instead of the landmarks. The depth image is obtained by rendering the 3D model after pose normalization, and then the face region is cropped and saved as a depth image of size 100×75. The pain depth video is divided into subsequences of size $\omega = 6$, and every subsequence is modeled as one subspace by applying k-SVD, with $k = 2$ and $\delta = 3$.

Firstly, we compare the performance of the proposed pain detection framework by using two different facial representations: the landmarks, and the depth data of the face region. In both the cases, the geodesic distance is used to create the *GMH* trajectories, with $\omega = 6$ and $k = 2$ under normalized pose. Figure 5.16 shows the ROC and AMOC curves for the two methods. From the ROC curve, we observe the depth representation, whose captures carry more spatio-temporal information, also achieves better performance on pain affect detection. The AUC value obtained using depth flow reached 0.80, compared to the value of 0.78 obtained using the landmarks only. In term of timeliness represented by AMOC curve, we can see that the depth flow scores less false positive rate once the system receives more than 50% of the pain emotion frames.

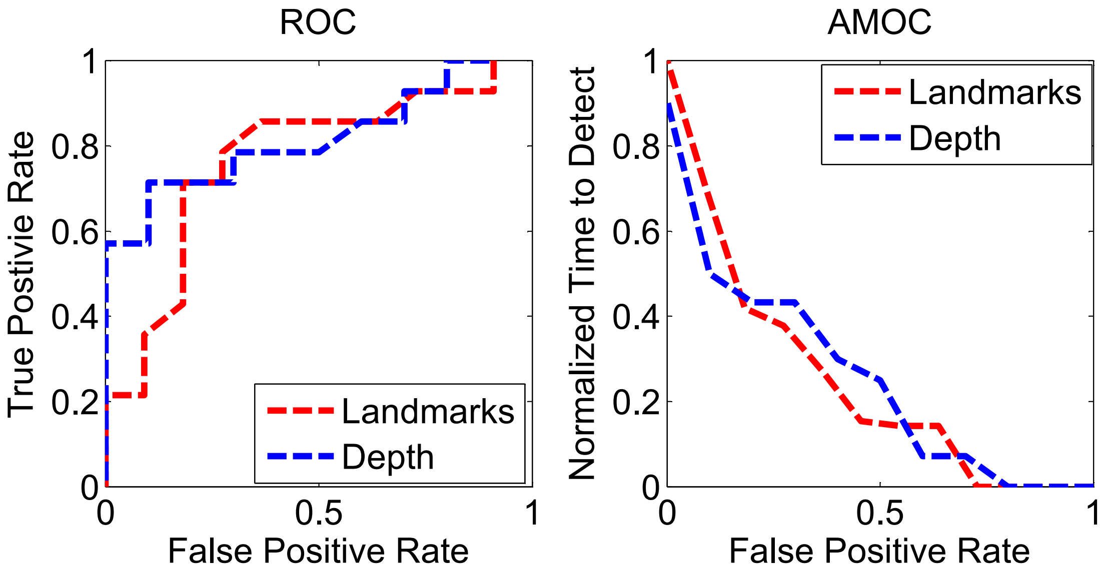

Figure 5.16: ROC and AMOC curves for comparison between pain detection using landmarks and depth representation.

The performance of the GMH computed from the geodesic distances is then evaluated in comparison with our proposed Local Deformation Histogram (LDH) descriptor extracted from the whole velocity vector between two successive subspaces along the trajectory (see Sect. 5.4). In both ehe cases, we used pose normalization with $\omega = 6$, $k = 2$, and $\delta = 3$. Results for this experiment are reported in Fig. 5.17, showing the ROC and AMOC curves for the two methods. The ROC curve on the left shows the superior performance of the LDH representation over the geodesic distance, where the AUC for LDH and geodesic distance is 0.84 and 0.80, respectively. The AMOC curve on the right shows that the two methods are comparable, while the system receives less than 40% of the pain emotion, and the LDH method achieves less false positive rate when more frames are received.

These results confirm the efficiency of using local coding of the temporal facial deformation through the time for pain affect detection from facial expressions. This

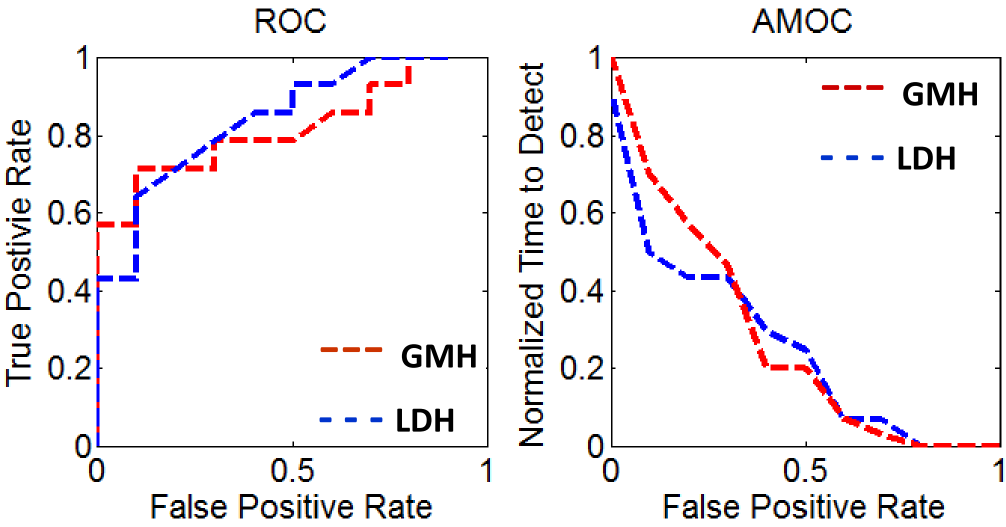

Figure 5.17: ROC and AMOC curves for comparison between pain detection using *Local Deformation Histogram (LDH)* and Grassmann distances-based *GMH*.

representation outperforms the geodesic distance method, which accounts only for the speed of the deformation through the time, thus incurring in potential hiding of important local cues for detection.

5.6 Conclusion

In this chapter, we have introduced a novel geometric framework for early detection of spontaneous emotional states, and experimented its applicability in two different scenarios: *(i)* happiness/thinking-unsure detection in depth videos of the upper part of the body acquired using Kinect-like cameras (depth-bodies); and *(ii)* physical pain detection from 3D high-resolution facial sequences (4D-faces). The key idea of our approach is to represent the stream of depth-images as time-parametrized trajectories of subspaces on a well defined Grassmann manifold. Analyzing the obtained trajectories gives rise to space-time features called *GMH (Geometric Motion History)* computed in two different ways to allow global and local analyze of the deformations and their temporal rhythm along the underlying trajectories. From a perspective of binary classification, we use an adaptation of the SVM algorithm to accommodate sequential (partial) analysis of the features, proposed earlier in [63]. We have experimentally illustrated the effectiveness of the proposed framework using two datasets: the Cam3D contains spontaneous emotions and complex mental states, such as happiness and thinking/unsure, while the BP4D consists of high-resolution 3D facial sequences of a set of eight emotional states, including the physical pain affect. We have performed several experiments including *(i)* global vs. local GMH (using LDH) representations, *(ii)* sparse (3D landmarks) vs. dense (depth) data, *(iii)* Stiefel vs. Grassmann (quotient space of the Stiefel), and *(iv)* the impact of the pose variation on the obtained results. To our knowledge, this is the first work proposing early automated detection of spontaneous emotions and pain acquired from high-resolution and low-resolution depth videos.

We have limited our experiments to an existing early event detector [63] from sequential Euclidean features in order to exemplify the proposed representation. It will be interesting to investigate advanced statistical inference techniques of partial (or full) observations using intrinsic (on the manifold) or extrinsic (e.g., fixed tangent space). In addition, we will apply the same framework to other databases and emotions to make more detailed comparison with other detection approaches.

CHAPTER 6

Conclusion and Perspectives

Summary and contributions

We have demonstrated, through the study investigated in this thesis, the contribution of 3D facial dynamic behavior in identity recognition and spontaneous emotion early detection. We have proposed a unified framework based on (optimized) subspace representations, which leads to the Grassmann manifolds. When the subspace-based representation is widely used in 2D domain and several computer vision research areas such as face recognition [134], action recognition [138], facial expression recognition [1] and age estimation [48]. To our knowledge, our study is the first one bringing these ideas, with extensions, to 3D dynamic domain. For each targeted application, we have derived a specific representation and efficient classification/detection algorithms. That is, in the context of face recognition, we have used the sparse coding and dictionary learning techniques on Grassmannian to design an efficient solution. We have demonstrated experimentally that considering the temporal evolution (up to certain interval) of the face helps to recognize people both in expression-dependent (same expression) and expression-independent (different expression) scenarios. A comparative study of the proposed solution to the existing method of Sun et al. [128] and two baseline algorithms, GNNC for Grassmann Nearest-Neighbor Classifier and an improved variant of the Grassmann Discriminant Analysis (GGDA) has shown the effectiveness of the proposed solution. In fact, our approach does not need neither landmarks detection nor tracking densely the vertex-flow over the 3D video. Extensive experiments (on the publicly available dataset BU-4DFE) are conducted but remains limited due to the limited number of subjects. Initially, this database is designed to test solutions on 4D (posed) facial expression analysis, where the participants are asked to sit in front of the camera and pose specific expression. Hence, all the 3D frames are near-frontal and of spatial and temporal high-resolutions.

To allow more realistic face recognition tests from 3D video, we have collected a new dataset, which includes several sequences of 58 participants, using a single-view 3D scanner with a large field-of-view to allow people (more than one in some scenarios) moving freely (but up to certain distance) in front of the 3D camera. Preliminary results on this new challenging dataset are reported as well.

As far as the second targeted application of early detection of spontaneous emotion is concerned, a novel (non-linear) representation of long 4D sequences is proposed. It consists to map the original 3D video data to Grassmann manifolds and build time-parametrized curves (or trajectories). Then, a simple dynamic model have been proposed based on the first-order derivation along the curves to capture the facial dynamic spatio-temporal behavior. Finally, we have employed and adapted recently-developed learning techniques for partial Euclidean data analysis. Using this pipeline, we have designed solutions for complex emotional state early detection. The validation has been made in two different scenarios. When the first uses depth-streams acquired via consumer cameras (like the Kinect) and focus on the behavior of the upper-part of the body, the second analyzes high-resolution 4D faces for the purpose of physical pain detection. These test scenarios are context-dependent, i.e., the emotional states and the physical pain are stimulated using the same procedure for all the participants of the databases, Cam3D [90] and BP4D-Spontaneous [152], respectively. Again, we consider these experiments limited due to the limited number of available acquisitions and participants. In contrast, the emotions exhibited in both datasets are spontaneous, which represents a first opportunities to researchers to conduct preliminary studies. Here also, an important set of experiments are conducted to compare the trajectory representation on Grassmann vs. Stiefel manifolds, the depth-based shape representation vs. the landmarks-based representations and to allow studying our approach's behavior when changing some relevant parameters.

As mentioned above, my thesis presents preliminary methodological and practical contributions to the field of face analysis from 4D facial sequences with experimental illustrations in face recognition and emotion detection. However, it opens the door to several perspectives and future work that we summarize in the next section.

Perspectives and future directions

This work is one of the first studies in the field of 4D data analysis for human facial behavior understanding. It is now a shared conviction that the 3D data capture faithfully the facial deformations and allow better understanding of facial behavior, compared to 2D data. Using dynamic 3D data (4D) is suitable as the face is a deformable surface by nature. This work confirms these observations in the context of identity recognition and emotion analysis. However, several issues of two types remain open – practical and methodological/theoretical.

First, the availably of 3D sensors embarked on computers and tablets have pushed, recently, the community to explore the use of depth and color streams together or separately in human behavior analysis. In addition to their attractive cost, RGB-D cameras (and their associated SDKs) present several benefits. That is, the foreground (human body, face, hands, etc.) could be isolated easily from the background in the filmed scene. Second, in spite of its low-resolution and the presence of noise, the depth channel is an additional source of information which reflects a dense (dynamic) shape representation of the face or the body. However, analyzing the dynamic depth channel requires to address several issues such as the noise, incomplete data, occlusions, etc. In this work, we have presented possible solutions to these problems, mainly using the subspace representation of short-time 3D clips. This representation could be improved by introducing some methodological approaches as we will describe next (i.e., smoothing and filtering Grassmann trajectories) and consider recent technical progress which makes available solutions for real-time pose estimation in depth videos. Considering these solutions, one can implement real-time processing algorithms, improve current performances and go to real-world like evaluation of the approaches. In this context and with the help of a master student (Damien Druel), I have started this work with the implementation of first blocks including – depth data acquisition using the Intel RealSense F200 camera. I use available algorithms in the SDK (face detection, landmarks detection and pose estimation), and include our implementation of the subspace-based representation. This gave rise to a preliminary interface to study the proposed methodology, in a realistic way and using depth-consumer cameras.

From a methodological point-of-view, it is now a sharing statement that dense correspondence between 3D frames is required to accurately quantify the facial deformations and the temporal dynamics. Some research groups have tried to tackle this problem by developing vertex flow tracking algorithms and/or model adaptation techniques, under facial deformations. For example, in [15], Ben Amor et al. have proposed a Riemannian approach, which resolves the issues of pose variations and dense correspondence, in the same framework, using elastic radial curves. However, the registration is obtained along the curves, which presents a serious limitation of their approach. In a different way, Sun et al. [128] proposed to use a vertex tracking algorithm, driven by the location of 3D landmarks along the 3D video. This method is time-consuming and unsuitable for real-time processing. Other solutions consist of using or adapting existing algorithms, previously used in static, like the Non-rigid Iterative Closest Point (ICP) [31], the Free Form Deformation (FFD) algorithm [49], or the Thin-plate Splines (TPS) technique [45] to achieve non-rigid registration or template fitting. Their goal is to achieve an accurate frame-to-frame correspondence. In our methodology, we consider short-time clips and assume pixel-to-pixel correspondence, in the same temporal interval (window). Long-term videos are presented by curves (of subspaces) on Grassmann manifolds. Although its capability to face both pose variation and dense correspondence issues, its major limitation is the limited size of the 3D clips. One possible future investigation is associating efficient 3D registration/tracking algorithms to subspace representations to allow increasing the time-interval of the clips (i.e. increase the window size) and study the behavior our the trajectory-based representation.

Another interesting methodological perspective to propose a suite of tools and algorithms for processing trajectories on Grassmann manifolds. The simplicity of their geometry and the availability of geometric formulations and efficient implementations (of geodesics, Karcher mean computation, etc.) make possible to develop the following processing blocks,

- **Smoothing and (median) filtering** of trajectories to allow reducing the effect of the noise, suitable when exploiting depth data. This is possible using algorithms to compute sample (Karcher) means and median samples on a fixed-time window of the trajectories.

- **Resampling (down-sampling or up-sampling)** original trajectories based on the geodesic formulation on Grassmann manifolds. In same cases, processing/analyzing requires increasing their temporal resolution. This is possible by creating new samples between original samples (up-sampling processing). In contrast, the down-sampling step reduces the number of original samples on the trajectory.

- **Novel dynamic models**, which consist in computing n-order derivations of the trajectories (the simplest ones are velocity and acceleration) to characterize their temporal evolution. In the proposed methodology, we have investigated only a first-order dynamic model, which leads to the velocity vector field. This model could be easily extended to a second-order model involving the covariant derivative of velocity vector fields, and so on.

- **Extend existing inference models to analyze curves on Grassmann manifolds** and their use in dynamic 3D data analysis. For example, it will be interesting to adopt techniques designed to analyze time-series to the Grassmann domains (or any other matrix manifold). Some recent works have studied the problem, in the context of object tracking, using particle filtering [55, 120].

All these tools and others could be developed in the continuous domain, which is more suitable for theoreticians. That is, one can start considering continuous and smooth parametrized curves on Grassmann manifolds (i.e., $\mathcal{G}_k(\mathbb{R}^n)$) and to develop proper metrics, statistical summaries and associated algorithms for the space of trajectories (i.e., $\mathcal{G}_k(\mathbb{R}^n)^{[0,t]}$). One difficult problem would be to propose rate-invariant metrics (or dynamic time-warping techniques) for registration and comparison of curves, which basically represent 4D sequences of the same emotion conveyed by different subjects, for example. Based on this methodology, one can push the discretization of the problem to the end step, i.e., when implementing the algorithms. All the ideas presented above, of both methodological and practical order, present the direction of our future investigations.

Bibliography

[1] Improving subspace learning for facial expression recognition using person dependent and geometrically enriched training sets.
Neural Networks, 24(8):814 – 823, 2011.
Artificial Neural Networks: Selected Papers from {ICANN} 2010.

[2] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino.
2D and 3D face recognition: A survey.
Pattern Recognition Letters, 28(14):1885–1906, 2007.

[3] M. Abd El Meguid and M. Levine.
Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers.
Affective Computing, IEEE Transactions on, 5(2):141–154, April 2014.

[4] P.-A. Absil, R. Mahony, and R. Sepulchre.
Optimization algorithms on matrix manifolds.
In *Princeton University Press, Princeton, NJ*, 2008.

[5] T. Alashkar, B. Ben Amor, M. Daoudi, and S. Berretti.
A 3D dynamic database for unconstrained face recognition.
In *5th Int. Conf. of 3D body scanning technology*, Oct 2014.

[6] T. Alashkar, B. Ben Amor, M. Daoudi, and S. Berretti.
A grassmannian framework for face recognition of 3D dynamic sequences with challenging conditions.
In *European Conf. on Computer Vision (ECCV) Workshops*, pages 326–340. Springer International Publishing, 2014.

[7] S. Aly, A. Trubanova, L. Abbott, S. White, and A. Youssef.
Vt-kfer: A kinect-based RGBD+time dataset for spontaneous and non-spontaneous facial expression recognition.

In *Biometrics (ICB), 2015 International Conference on*, pages 90–97, May 2015.

[8] B. Amberg, S. Romdhani, and T. Vetter.
Optimal step nonrigid ICP algorithms for surface registration.
In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[9] R. Anirudh, P. Turaga, J. Su, and A. Srivastava.
Elastic functional coding of human actions: From vector-fields to latent variables.
June 2015.

[10] M. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, A. Elkins, N. Tyler, P. Watson, A. Williams, M. Pantic, and N. Berthouze.
The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset.
Affective Computing, IEEE Transactions on, 2015.

[11] J. Barr, K. Bowyer, P. Flynn, and S. Biswas.
Face recognition from video: a review.
Int. Journal of Pattern Recognition and Artificial Intelligence, 26(5), 2012.

[12] A. Battocchi, F. Pianesi, and D. Goren-Bar.
The properties of dafex, a database of kinetic facial expressions.
In J. Tao, T. Tan, and R. Picard, editors, *Affective Computing and Intelligent Interaction*, volume 3784 of *Lecture Notes in Computer Science*, pages 558–565. Springer Berlin Heidelberg, 2005.

[13] M. Bauml, K. Bernardin, M. Fischer, H. Ekenel, and R. Stiefelhagen.
Multi-pose face recognition for person retrieval in camera networks.
In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 441–447, Aug 2010.

[14] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen.
Liris-accede: A video database for affective content analysis.
Affective Computing, IEEE Transactions on, 6(1):43–55, Jan 2015.

[15] B. Ben Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava.
4-d facial expression recognition by learning geometric deformations.
IEEE T. Cybernetics, 44(12):2443–2457, 2014.

[16] B. Ben Amor, J. Su, and A. Srivastava.
Action recognition using rate-invariant analysis of skeletal shape trajectories.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, PP(99):1–1, 2015.

[17] L. Benedikt, D. Cosker, P. Rosin, and D. Marshall.
Assessing the uniqueness and permanence of facial actions for use in biometric applications.
Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 40(3):449–460, May 2010.

[18] S. Berretti, A. Del Bimbo, and P. Pala.
3D face recognition using iso-geodesic stripes.
IEEE Trans. on Pattern Analysis and Machine Intelligence, 32(12):2162–2177, Dec. 2010.

[19] S. Berretti, A. Del Bimbo, and P. Pala.
Real-time expression recognition from dynamic sequences of 3D facial scans.
In *Proceedings of the 5th Eurographics Conference on 3D Object Retrieval*, EG 3DOR'12, pages 85–92. Eurographics Association, 2012.

[20] S. Berretti, P. Pala, and A. Del Bimbo.
Face recognition by super-resolved 3D models from consumer depth cameras.
Information Forensics and Security, IEEE Transactions on, 9(9):1436–1449, Sept 2014.

[21] J. R. Beveridge, H. Zhang, B. A. Draper, P. J. Flynn, Z. Feng, P. Huber, J. Kittler, Z. Huang, S. Li, Y. Li, M. Kan, R. Wang, S. Shan, X. Chen, H. Li, G. Hua, V. Struc, J. Krizaj, C. Ding, D. Tao, and P. J. Phillips.
Report on the FG 2015 video person recognition evaluation.
In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, May 4-8, 2015*, pages 1–8, 2015.

[22] K. Bowyer, K. Chang, and P. Flynn.
A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition.
Computer Vision and Image Understanding, 101(1):1–15, 2006.

[23] V. Bruce, Z. Henderson, K. Greenwood, P. Hancock, A. Burton, and P. Miller. Verification of face identities from images captured on video. *Journal of Experimental Psychology*, 5:339–360, 1999.

[24] P. Bull. *Communication under the microscope: The theory and practice of microanalysis.* Routledge, 2002.

[25] R. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, 1(1):18–37, Jan 2010.

[26] H. Cetingul and R. Vidal. Intrinsic mean shift for clustering on stiefel and grassmann manifolds. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1896–1902, Miami Beach, FL, USA, June 2009.

[27] H. Cetingul and R. Vidal. Sparse riemannian manifold clustering for hardi segmentation. In *IEEE Int. Symp. on Biomedical Imaging: From Nano to Macro*, pages 1750–1753, March 2011.

[28] C. H. Chan, B. Goswami, J. Kittler, and W. Christmas. Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication. *Information Forensics and Security, IEEE Transactions on*, 7(2):602–612, April 2012.

[29] K. Chang, W. Bowyer, and P. Flynn. Multiple nose region matching for 3d face recognition under varying facial expression. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1695–1700, Oct 2006.

[30] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. In *Robotics and Automation*, volume 3, pages 2724–2729, 1991.

[31] S. Cheng, I. Marras, S. Zafeiriou, and M. Pantic.
Active nonrigid ICP algorithm.
In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'15)*, Ljubljana, Slovenia, May 2015.

[32] J. Choi, A. Sharma, and G. Medioni.
Comparing strategies for 3D face recognition from a 3D sensor.
In *RO-MAN, 2013 IEEE*, pages 19–24, Aug 2013.

[33] F. Christie and V. Bruce.
The role of dynamic information in the recognition of unfamiliar faces.
Memory and Cognition, 26:780–790, 1998.

[34] J. Cohn, K. Schmidt, R. Gross, and P. Ekman.
Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform person identification.
In *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pages 491–496, 2002.

[35] D. Cosker, E. Krumhuber, and A. Hilton.
A facs valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling.
In *Int. Conf. on Computer Vision (ICCV)*, pages 2296–2303, 2011.

[36] A. Cruz, B. Bhanu, and N. Thakoor.
Vision and attention theory based sampling for continuous facial emotion recognition.
Affective Computing, IEEE Transactions on, 5(4):418–431, Oct 2014.

[37] A. Danelakis, T. Theoharis, and I. Pratikakis.
A survey on facial expression recognition in 3D video sequences.
Multimedia Tools and Applications, 74(15):5577–5615, 2015.

[38] A. Danelakis, T. Theoharis, and I. Pratikakis.
A survey on facial expression recognition in 3D video sequences.
Multimedia Tools and Applications, 74(15):5577–5615, 2015.

[39] H. Drira, B. Ben Amor, M. Daoudi, A. Srivastava, and S. Berretti.

3D dynamic expression recognition based on a novel deformation vector field and random forest.
In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1104–1107, 2012.

[40] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama.
3D face recognition under expressions, occlusions, and pose variations.
IEEE Trans. on Pattern Analysis and Machine Intelligence, 35(9):2270–2283, Sept. 2013.

[41] M. Du, A. C. Sankaranarayanan, and R. Chellappa.
Robust face recognition from multi-view videos.
IEEE Transactions on Image Processing, 23(3):1105–1117, 2014.

[42] A. Edelman, T. Arias, and S. Smith.
The geometry of algorithms with orthogonality constraints.
Siam J. Matrix Anal. Appl., 20(2):303–353, 1998.

[43] G. Edwards, C. Taylor, and T. Cootes.
Improving identification performance by integrating evidence from sequences.
In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 1, page 491 Vol. 1, 1999.

[44] P. Ekman.
Universals and cultural differences in facial expressions of emotion.
In *Nebraska Symposium on Motivation*, volume 19, pages 207–283, Lincoln, NE, 1972.

[45] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris.
3D/4D facial expression analysis: An advanced annotated face model approach.
Image and Vision Computing, 30(10):738 – 749, 2012.

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

[47] M.-I. Faraj and J. Bigun.
Audio,Àìvisual person authentication using lip-motion from orientation maps.

Pattern Recognition Letters, 28(11):1368 – 1382, 2007.
Advances on Pattern recognition for speech and audio processing.

[48] Y. Fu and T. Huang.
Human age estimation with regression on discriminative aging manifold.
Multimedia, IEEE Transactions on, 10(4):578–584, June 2008.

[49] M. P. G. Sandbach, S. Zafeiriou and L. Yin.
Static and dynamic 3D facial expression recognition: A comprehensive survey.
Image and Vision Computing, 30(10):683 – 697, 2012.

[50] K. Gallivan, A. Srivastava, X. Liu, and P. Van Dooren.
Efficient algorithms for inferences on grassmann manifolds.
In *Statistical Signal Processing, 2003 IEEE Workshop on*, pages 315–318, 2003.

[51] S. Gao, I.-H. Tsang, and L.-T. Chia.
Laplacian sparse coding, hypergraph laplacian sparse coding, and applications.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(1):92–104, Jan 2013.

[52] G. Golub and C. Van Loan.
Matrix computations (3rd edition).
Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[53] R. Gopalan, R. Li, and R. Chellappa.
Domain adaptation for object recognition: An unsupervised approach.
In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 999–1006, Barcelona, Spain, Nov. 2011.

[54] B. Goswami, C. H. Chan, J. Kittler, and B. Christmas.
Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication.
In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–6, Sept 2010.

[55] I. Gu and Z. Khan.
Grassmann manifold online learning and partial occlusion handling for visual object tracking under bayesian formulation.
In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1463–1466, 2012.

[56] J. Hamm and D. D. Lee.
Grassmann discriminant analysis: A unifying view on subspace-based learning.
In *Int. Conf. on Machine Learning*, ICML '08, pages 376–383, 2008.

[57] H. Han, C. Otto, X. Liu, and A. K. Jain.
Demographic estimation from face images: Human vs. machine performance.
IEEE Trans. Pattern Anal. Mach. Intell., 37(6):1148–1161, 2015.

[58] S. Happy, A. Dasgupta, P. Patnaik, and A. Routray.
Automated alertness and emotion detection for empathic feedback during e-learning.
In *Technology for Education (T4E), 2013 IEEE Fifth International Conference on*, pages 47–50, Dec 2013.

[59] M. Harandi, R. Hartley, C. Shen, B. Lovell, and C. Sanderson.
Extrinsic methods for coding and dictionary learning on grassmann manifolds.
Int. Journal of Computer Vision, 2015, under press.

[60] M. T. Harandi, C. Sanderson, S. A. Shirazi, and B. C. Lovell.
Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching.
In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2705–2712, Colorado Springs, CO, USA, June 2011.

[61] M. Hayat, M. Bennamoun, and A. El-Sallam.
Fully automatic face recognition from 3D videos.
In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1415–1418, Nov 2012.

[62] J. Helmke and K. Huper.
Newton's method on grassmann manifold.
In *Preprint, arXiv:0709.2205*, 2007.

[63] M. Hoai and F. De la Torre.
Max-margin early event detectors.
Int. Journal of Computer Vision, 107(2):191–202, Feb. 2014.

[64] G.-S. Hsu, Y.-L. Liu, H.-C. Peng, and P.-X. Wu.
RGB-D-based face reconstruction and recognition.
IEEE Trans. on Information Forensics and Security, 9(12):2110–2118, Dec. 2014.

[65] G. Hua, M. Yang, E. Learned-Miller, Y. Ma, M. Turk, D. Kriegman, and T. Huang. Introduction to the special section on real-world face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(10):1921–1924, Oct. 2011.

[66] Z. Huang, R. Wang, S. Shan, and X. Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 140–149, 2015.

[67] L. A. Jeni, A. Lorincz, T. Nagy, Z. Palotai, J. Sebok, Z. Szabo, and D. Takacs. 3D shape estimation in video sequences provides high precision evaluation of facial expressions. *Image and Vision Computing*, 30(10):785 – 795, 2012. 3D Facial Behaviour Analysis and Understanding.

[68] I. Jolliffe. *Principal Component Analysis*. John Wiley and Sons, Ltd, 2005.

[69] E. J.Su, S.Kurtek and A.Srivastava. Statistical analysis of trajectories on riemannian manifolds: Bird migration, hurricane tracking and video surveillance. *The Annals of Applied Statistics*, 8(1), April 2014.

[70] P. S. K. Prkachin. The structure, reliability and validity of pain expression: evidence from patients with shoulder pain. *Pain*, 139(2):267 – 274, 2008.

[71] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *Advances in visual computing*, volume 7432 of *Lecture notes in computer science*, pages 368–377, Berlin, Germany, 2012. Springer.

[72] S. K. A. Kamarol, N. S. Meli, M. H. Jaward, and N. Kamrani. Spatio-temporal texture-based feature extraction for spontaneous facial expression recognition.

In *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*, pages 467–470, May 2015.

[73] H. Karcher.
Riemannian center of mass and mollifier smoothing.
Communications on Pure and Applied Mathematics, 30:509–541, 1977.

[74] R. Khan, A. Meyer, H. Konik, and S. Bouakaz.
Pain detection through shape and appearance features.
In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6, July 2013.

[75] B. Knight and A. Johnston.
The role of movement in face recognition.
Visual Cognition, 2:265–273, 1997.

[76] S. Koelstra, M. Pantic, and I. Patras.
A dynamic texture-based approach to recognition of facial actions and their temporal models.
IEEE Trans. Pattern Anal. Mach. Intell., 32(11):1940–1954, 2010.

[77] K. Lander, F. Chrisite, and V. Bruce.
The role of movement in the recognition of famous faces.
Visual Cognition, 6:974–985, 1999.

[78] V. Le, H. Tang, and T. Huang.
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*, pages 414–421, March 2011.

[79] B. Li, A. Mian, W. Liu, and A. Krishna.
Using kinect for face recognition under varying poses, expressions, illumination and disguise.
In *IEEE Work. on Applications of Computer Vision (WACV)*, pages 186–192, Tampa, FL, USA, Jan. 2013.

[80] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen.
A spontaneous micro-expression database: Inducement, collection and baseline.

In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6, April 2013.

[81] Y. M. Lim, A. Ayesh, and M. Stacey.
Detecting emotional stress during typing task with time pressure.
In *Science and Information Conference (SAI), 2014*, pages 329–338, Aug 2014.

[82] P. Liu and L. Yin.
Spontaneous facial expression analysis based on temperature changes and head motions.
In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, pages 1–6, May 2015.

[83] X. Liu and T. Chen.
Video-based face recognition using adaptive hidden markov models.
In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 340–345, Madison, WS, USA, June 2003.

[84] S. Liwicki, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic.
Efficient online subspace learning with an indefinite kernel for visual tracking and recognition.
IEEE Transactions on Neural Networks and Learning Systems, 23:1624–1636, October 2012.

[85] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews.
Painful data: The unbc-mcmaster shoulder pain expression archive database.
In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64, March 2011.

[86] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews.
Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database.
Image and Vision Computing, 30(3):197 – 205, 2012.
Best of Automatic Face and Gesture Recognition 2011.

[87] J. Luettin, N. Thacker, and S. Beet.
Speaker identification by lipreading.
In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 1, pages 62–65 vol.1, Oct 1996.

[88] Y. M. Lui.
Advances in matrix manifolds for computer vision.
Image Vision Computing, 30(6-7):380–388, June 2012.

[89] Y. M. Lui and J. R. Beveridge.
Grassmann registration manifolds for face recognition.
In *European Conf. on Computer Vision*, ECCV'08, pages 44–57, 2008.

[90] M. Mahmoud, T. Baltrŭsaitis, P. Robinson, and L. Riek.
3D corpus of spontaneous complex mental states.
In *Conf. on Affective Computing and Intelligent Interaction*, pages 205–214, Memphis, TN, USA, Oct. 2011.

[91] B. Matuszewski, W. Quan, L.-k. Shark, A. McLoughlin, C. Lightbody, H. Emsley, and C. Watkins.
Hi4d-adsip 3D dynamic facial articulation database.
Image and Vision Computing, 30(10), 2012.

[92] S. Mavadati, M. Mahoor, K. Bartlett, P. Trinh, and J. Cohn.
Disfa: A spontaneous facial action intensity database.
Affective Computing, IEEE Transactions on, 4(2):151–160, April 2013.

[93] G. McKeown, M. Valstar, R. Cowie, and M. Pantic.
The semaine corpus of emotionally coloured character interactions.
In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084, July 2010.

[94] H. Meeren, C. van Heijnsbergen, and B. de Gelder.
Rapid perceptual integration of facial expression and emotional body language.
National Academy of Sciences USA, 102(45):16518–16523, 2005.

[95] A. Mehrabian and M. WIENER.
Decoding of inconsistent communications.
Journal of Personality and Social Psychology, 6(1):109–114, May 1967.

[96] R. Min, J. Choi, G. Medioni, and J.-L. Dugelay.
Real-time 3D face identification from a depth camera.
In *Int. Conf. on Pattern Recognition (ICPR)*, pages 1739–1742, Tsukuba, Japan, Nov. 2012.

[97] B. Mishra, S. Fernandes, K. Abhishek, A. Alva, C. Shetty, C. Ajila, D. Shetty, H. Rao, and P. Shetty.
Facial expression recognition using feature based techniques and model based techniques: A survey.
In *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*, pages 589–594, 2015.

[98] M. S. Mohamad-Hoseyn Sigari, Muhammad-Reza Pourshahabi and M. Fathy.
A review on driver face monitoring systems for fatigue and distraction detection.
International Journal of Advanced Science and Technology, 46(0):73 – 100, 2014.

[99] N. Nguyen and Y. Guo.
Comparisons of sequence labeling algorithms and extensions.
In *Int. Conf. on Machine Learning*, ICML '07, pages 681–688, 2007.

[100] M. Pantic.
Facial Expression Recognition, pages 1–8.
2014.

[101] P. Phillips, P. Flynn, W. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek.
Overview of the face recognition grand challenge.
In *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 947–954, San Diego, CA, USA, June 2005.

[102] P. J. Phillips and A. J. O'Toole.
Comparison of human and computer performance across face recognition experiments.
Image Vision Comput., 32(1):74–85, 2014.

[103] E. Piatkowska and J. Martyna.
Spontaneous facial expression recognition: Automatic aggression detection.
In E. Corchado, V. Snasel, A. Abraham, M. Wozniak, M. Grana, and S.-B. Cho, editors, *Hybrid Artificial Intelligent Systems*, volume 7208 of *Lecture Notes in Computer Science*, pages 147–158. Springer Berlin Heidelberg, 2012.

[104] G. Pike, R. . Kemp, A. Towell, and C. Keith.
Recognizing moving faces: The relative contribution of motion and perspective view information.

Visual Cognition, 4:409–438, 1997.

[105] M. Reale, X. Zhang, and L. Yin.
Nebula feature: A space-time feature for posed and spontaneous 4d facial behavior analysis.
In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8, 2013.

[106] Q. Rentmeesters, P.-A. Absil, P. Van Dooren, K. Gallivan, and A. Srivastava.
An efficient particle filtering technique on the grassmann manifold.
In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 3838–3841, March 2010.

[107] M. Roach, J. Brand, and J. Mason.
Acoustic and facial features for speaker recognition.
In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 3, pages 258–261 vol.3, 2000.

[108] D. Roark, S. Barrett, M. Spence, A. Abdi, and A. O'Toole.
Psychological and neural perspectives on the role of motion in face recognition.
Behavioral and Cognitive Neuroscience Reviews, 2:15–46, 2003.

[109] G. I. Roisman and J. L. Tsai.
The emotional integration of childhood experience: Physiological, facial expressive, and self-reported emotional response during the adult attachment interview.
Developmental Psychology, pages 776–789, 2004.

[110] D. Rueckert, A. Frangi, and J. Schnabel.
Automatic construction of 3D statistical deformation models using non-rigid registration.
In W. Niessen and M. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention ‚Äì MICCAI 2001*, volume 2208 of *Lecture Notes in Computer Science*, pages 77–84. Springer Berlin Heidelberg, 2001.

[111] J. Russell and A. Mehrabian.
Evidence for a three-factor theory of emotions.
Journal of Research in Personality, 11(3):273–294, Sept. 1977.

[112] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert.

A dynamic approach to the recognition of 3D facial expressions and their temporal models.
In *IEEE Conf. on Automatic Face and Gesture Recognition*, pages 406–413, Santa Barbara, CA, Mar. 2011.

[113] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert.
Recognition of 3D facial expression dynamics.
Image and Vision Computing, 30(10):762–773, 2012.
3D Facial Behaviour Analysis and Understanding.

[114] E. Sariyanidi, H. Gunes, and A. Cavallaro.
Automatic analysis of facial affect: A survey of registration, representation, and recognition.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, 37(6):1113–1133, June 2015.

[115] K. Schindler and L. Van Gool.
Action snippets: How many frames does human action recognition require?
In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[116] R. Séguier.
A very fast adaptive face detection system.
International conference on visualization, imaging, and image processing, 2004.

[117] T. Senechal, J. Turcot, and R. El Kaliouby.
Smile or smirk? automatic detection of spontaneous asymmetric smiles to understand viewer experience.
In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8, April 2013.

[118] J. Shao, I. Gori, S. Wan, and J. Aggarwal.
3D dynamic facial expression recognition using low-resolution videos.
Pattern Recognition Letters, 2015.

[119] R. Shigenaka, B. Raytchev, T. Tamaki, and K. Kaneda.
Face sequence recognition using grassmann distances and grassmann kernels.
In *IEEE Int. Joint Conf. on Neural Networks (IJCNN)*, pages 1–7, Brisbane, QLD, Australia, June 2012.

[120] S. A. Shirazi, M. T. Harandi, B. C. Lovell, and C. Sanderson.
Object tracking via non-euclidean geometry: A grassmann approach.
CoRR, abs/1403.0309, 2014.

[121] S. A. Shirazi, M. T. Harandi, C. Sanderson, A. Alavi, and B. C. Lovell.
Clustering on grassmann manifolds via kernel embedding with application to action analysis.
In *IEEE Int. Conf. on Image Processing (ICIP)*, pages 781–784, Orlando, FL, USA, Sept. 2012.

[122] K. Sikka, A. Dhall, and M. Bartlett.
Weakly supervised pain localization using multiple instance learning.
In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8, April 2013.

[123] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic.
A multimodal database for affect recognition and implicit tagging.
Affective Computing, IEEE Transactions on, 3(1):42–55, Jan 2012.

[124] A. Srivastava.
A bayesian approach to geometric subspace estimation.
IEEE Trans. on Signal Processing, 48(5):1390–1400, May 2000.

[125] L. Steede, J. Tree, and H. G.J.
I can't recognize your face but i can recognize its movement.
Cognitive Neuropsychology, 24:451–466, 2007.

[126] L. Su, S. Kumano, K. Otsuka, D. Mikami, J. Yamato, and Y. Sato.
Early facial expression recognition with high-frame rate 3D sensing.
In *IEEE Int. Conf. on Systems, Man, and Cybernetics*, pages 3304–3310, Oct 2011.

[127] L. Su and Y. Sato.
Early facial expression recognition using early rankboost.
In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 1–7, April 2013.

[128] Y. Sun, X. Chen, M. Rosato, and L. Yin.
Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis.

Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 40(3):461–474, May 2010.

[129] M. Tistarelli and M. S. Nixon, editors.
Advances in Biometrics, Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings, volume 5558 of *Lecture Notes in Computer Science*. Springer, 2009.

[130] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun.
Large margin methods for structured and interdependent output variables.
Journal of Machine Learning Research, 6:1453–1484, Sept. 2005.

[131] S. Tulyakov, T. Slowe, Z. Zhang, and V. Govindaraju.
Facial expression biometrics using tracker displacement features.
In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–5, June 2007.

[132] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa.
Statistical computations on grassmann and stiefel manifolds for image and video-based recognition.
IEEE Trans. on Pattern Analysis and Machine Intelligence, 33(11):2273–2286, Nov. 2011.

[133] M. Turk and A. Pentland.
Face recognition using eigenfaces.
In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, pages 586–591, Jun 1991.

[134] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic.
Subspace learning from image gradient orientations.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(12):2454–2466, Dec 2012.

[135] J. Van den Stock, R. Righart, and B. de Gelder.
Body expressions influence recognition of emotions in the face and voice.
Emotion, 7:487–494, August 2007.

[136] R. Vemulapalli, J. Pillai, and R. Chellappa.
Kernel learning for extrinsic classification of manifold features.

In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1782–1789, Portland, OR, USA, June 2013.

[137] F. Wallhoff, B. Schuller, M. Hawellek, and G. Rigoll.
Efficient recognition of authentic dynamic facial expressions on the feedtum database.
In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 493–496, July 2006.

[138] L. Wang and D. Suter.
Learning and matching of dynamic shape manifolds for human action recognition.
Image Processing, IEEE Transactions on, 16(6):1646–1661, June 2007.

[139] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang.
A natural visible and infrared facial expression database for expression recognition and emotion inference.
Multimedia, IEEE Transactions on, 12(7):682–691, Nov 2010.

[140] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma.
Robust face recognition via sparse representation.
IEEE Trans. on Pattern Analysis and Machine Intelligence, 31(2):210–227, Feb. 2009.

[141] Y. Xie, J. Ho, and B. Vemuri.
On a nonlinear generalization of sparse coding and dictionary learning.
In *Int. Conf. of Machine Learning (ICML)*, pages 1480–1488, Atlanta, GE, USA, June 2013.

[142] Y. Xu, Z. Xiao, and X. Tian.
A simulation study on neural ensemble sparse coding.
In *Int. Conf. on Information Engineering and Computer Science (ICIECS)*, pages 1–4, Wuhan, China, Dec. 2009.

[143] M. Xue, A. Mian, W. Liu, and L. Li.
Automatic 4d facial expression recognition using DCT features.
In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 199–206, Jan 2015.

[144] B. Yang, J. Yan, Z. Lei, and S. Li.

Fine-grained evaluation on face detection in the wild.
In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, pages 1–7, May 2015.

[145] M. Yang, D. Zhang, J. Yang, and D. Zhang.
Robust sparse coding for face recognition.
In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 625–632, Colorado Springs, CO, USA, June 2011.

[146] C. Yasuko.
Statistics on special manifolds, lecture notes in statistics.
In *vol. 174. New York: Springer*, 2003.

[147] N. Ye and T. Sim.
Towards general motion-based face recognition.
In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2598–2605, June 2010.

[148] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale.
A high-resolution 3D dynamic facial expression database.
In *IEEE Conf. on Face and Gesture Recognition (FG)*, pages 1–6, Amsterdam, The Netherlands,, Sept. 2008.

[149] S. Zafeiriou and M. Pantic.
Facial behaviometrics: The case of facial deformation in spontaneous smile/laughter.
In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 13–19, June 2011.

[150] Z. Zeng, Y. Fu, G. Roisman, Z. Wen, Y. Hu, and T. Huang.
One-class classification for spontaneous facial expression analysis.
In *Int. Conf. on Automatic Face and Gesture Recognition (FGR)*, pages 281–286, April 2006.

[151] Z. Zeng, M. Pantic, G. Roisman, and T. Huang.
A survey of affect recognition methods: Audio, visual, and spontaneous expressions.
IEEE Trans. on Pattern Analysis and Machine Intelligence, 31(1):39–58, Jan. 2009.

[152] L. Zhang, H. Nejati, L. Foo, K. T. Ma, D. Guo, and T. Sim.

A talking profile to distinguish identical twins.
In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6, April 2013.

[153] X. Zhang and Y. Gao.
Face recognition across pose: A review.
Pattern Recognition, 42(11):2876–2896, 2009.

[154] X. Zhang, L. Yin, and J. F. Cohn.
Three dimensional binary edge feature representation for pain expression analysis.
In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, pages 1–7, May 2015.

[155] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard.
Bp4d-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database.
Image and Vision Computing, 32(10):692 – 706, 2014.

[156] Y. Zhang, S. J. Kundu, D. B. Goldgof, S. Sarkar, and L. V. Tsap.
Elastic face, an anatomy-based biometrics beyond visible cue.
In *In Proceedings of International Conference on Pattern Recognition*, 2004.

[157] G. Zhao and M. Pietikainen.
Dynamic texture recognition using local binary patterns with an application to facial expressions.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29(6):915–928, 2007.

[158] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld.
Face recognition: A literature survey.
ACM Comput. Surv., 35(4):399–458, Dec. 2003.

[159] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li.
High-fidelity pose and expression normalization for face recognition in the wild.
June 2015.

[160] W. Zuo, D. Meng, L. Zhang, X. Feng, and D. Zhang.
A generalized iterated shrinkage algorithm for non-convex sparse coding.

In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 217–224, Sydney, Australia, Dec 2013.