Année 2018

Université Lille

THÈSE

pour obtenir le grade de
Docteur,
spécialité Informatique et Applications

présentée et soutenue publiquement par

Anis Kacem

le 12 Décembre 2018

Novel Geometric Tools for Human Behavior Understanding

préparée au sein du laboratoire CRIStAL UMR CNRS 9189 et l'IMT Lille-Douai
sous la direction de
Mr. Mohamed Daoudi

COMPOSITION DU JURY

Mr. Nicu Sebe	Rapporteur	Professor, University of Trento, Italy
Mr. Frederic Jurie	Rapporteur	Professor, Université de Caen, France
Mr. Mohamed Daoudi	Directeur de la thèse	Professor, IMT Lille Douai, France
Mr. Boulbaba Ben Amor	Encadrant de la thèse	Professor, IMT Lille Douai, France
Mr. Juan Carlos Alvarez-Paiva	Président	Professor, Université de Lille, France
Mme. Tinne Tuytelaars	Examinatrice	Professor, KU Leuven, Belgium
Mme. Catherine Achard	Examinatrice	Professor, Sorbonne Université, France

Abstract

Developing intelligent systems dedicated to human behavior understanding has been a very hot research topic in the few recent decades. Indeed, it is crucial to understand the human behavior in order to make machines able to interact with, assist, and help humans in their daily life.. Recent breakthroughs in computer vision and machine learning have made this possible. For instance, human-related computer vision problems can be approached by first detecting and tracking 2D or 3D landmark points from visual data. Two relevant examples of this are given by the facial landmarks detected on the human face and the skeletons tracked along videos of human bodies. These techniques generate temporal sequences of landmark configurations, which exhibit several distortions in their analysis, especially in uncontrolled environments, due to view variations, inaccurate detection and tracking, missing data, etc. In this thesis, we propose two novel space-time representations of human landmark sequences along with suitable computational tools for human behavior understanding. Firstly, we propose a representation based on trajectories of Gram matrices of human landmarks. Gram matrices are positive semi-definite matrices of fixed rank and lie on a nonlinear manifold where standard computational and machine learning techniques could not be applied in a straightforward way. To overcome this issue, we make use of some notions of the Riemannian geometry and derive suitable computational tools for analyzing Gram trajectories. We evaluate the proposed approach in several human related applications involving 2D and 3D landmarks of human faces and bodies such us emotion recognition from facial expression and body movements and also action recognition from skeletons. Secondly, we propose another representation based on the barycentric coordinates of 2D facial landmarks. While being related to the Gram trajectory representation and robust to view variations, the barycentric representation allows to directly work with standard computational tools. The evaluation of this second approach is conducted on two face analysis tasks namely, facial expression recognition and depression severity level assessment.

The obtained results with the two proposed approaches on real benchmarks are competitive with respect to recent state-of-the-art methods.

Résumé : Nouvelles approches géomètriques pour l'analyse du comportement humain

Récemment, le développement de systèmes intelligents dédiés pour la compréhension du comportement humain est devenu un axe de recherche très important. En effet, il est très important de comprendre le comportement humain pour rendre les machines capables d'aider et interagir avec les humains. Pour cela, plusieurs approches de l'état de l'art commencent par détecter automatiquement un ensemble de points 2D ou 3D, appelés marqueurs, sur le corps et/ou le visage humain à partir de données visuelles. L'analyse des séquences temporelles de ces marqueurs pose plusieurs défis dus aux erreurs de suivi et aux variabilités temporelles et de pose. Dans cette thèse, nous proposons deux nouvelles représentations spatio-temporelles avec des outils de calcul appropriés pour la compréhension du comportement humain. La première consiste à représenter une séquence temporelle de marqueurs par une trajectoire de matrices de Gram. Les matrices de Gram sont des matrices semi-définies positives de rang fixe et vivent dans un espace non-linéaire dans lequel les outils d'apprentissage automatique conventionnels ne peuvent pas être appliqués directement. Nous évaluons l'efficacité de notre approche dans plusieurs applications, impliquant des marqueurs 2D et 3D de visages et de corps humain, tels que la reconnaissance des émotions à partir des expressions faciales la reconnaissance d'actions et des émotions à partir des données de profondeur 3D. La deuxième représentation proposée dans cette thèse est basée sur les coordonnées barycentriques des marqueurs de visages 2D. Cette représentation permet d'utiliser les outils de calcul et d'apprentissage automatique tels que les techniques d'apprentissage de métrique. Les résultats obtenus en reconnaissance des expressions faciales et en mesure automatique de la sévérité de la dépression à partir du visage montrent tout l'intérêt de la représentation barycentrique combinée à des techniques d'apprentissage automatique. Les résultats obtenus avec les deux méthodes proposées sur des bases de données réelles montrent la compétitivité

de nos approches avec les méthodes récentes de l'état de l'art.

Dedications and acknowledgments

This thesis would not have been possible without the guidance of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of the present research.

First of all, I would like to express my very great appreciation to **Prof. Mohamed Daoudi**, my PhD supervisor, for his immense knowledge, attention and everyday communication from the very early stage of this research as well as for his extraordinary advises. His perpetual energy and enthusiasm in research extremely motivated me in my studies. I appreciate his contributions of time and ideas to make my work productive and stimulating. I learned from Prof. Daoudi how to dissect the challenges into smaller parts and to look at them from different angles. Thanks once again go to him. Without his immense support, this work would have never seen the light as it is today.

I am particularly grateful to **Prof. Boulbaba Ben Amor**, my PhD co-supervisor, for his insightful comments and efforts, for the hard questions which helped me to widen my research from various perspectives, but also for his valuable assistance, supervision and crucial contribution to this thesis. His professional and academic touches had an important effect on my work that I will benefit from, for a long time to come.

I must express my gratitude to **Prof. Juan Carlos Alvarez-Paiva**, for his continued support and encouragement. I have been extremely lucky to have an amazing and valuable professor of Mathematical Sciences who helped me a lot in formulating the encountered problems and solutions, and who responded to my questions and queries so directly. Under his guidance, I successfully overcame many difficulties and learned a lot too.

I would like to offer my special thanks to **Prof. Stefano Berretti** from University of Florence in Italy for the continuous assistance of my PhD study and related research, for his patience, motivation, and priceless advises. I could not have imagined having a better

mentor for my PhD study. In spite of the distance, I really appreciated his willingness to help whenever I ran into a trouble or had a question about my research.

I would like to thank **Prof. Jeffrey Cohn** and **Dr. Zakia Hammal** from Carnegie Mellon University in USA, for their valuable time, co-operation, and generosity which helped in making this work better. Their deep insights helped me at an important stage of my research. I am also indebted towards them for their encouragement and especially for the nice collaboration that we had in the last year of my research. It was a great pleasure and honor to work with them.

I thank all the persons that I met in the Equipex (plaine images), including Prof. Jean-Louis Nandrino, Prof. Yann Coelho, Prof. Henrique Sequeira, Dr. Mohamed Ladrouz and Dr. Laurence Delbarre who warmly welcomed me and provided essential materials for my research in the Equipex.

Special thanks to my PhD committee members for taking the time to participate in the evaluation of this work: **Prof. Nicu Sebe, Prof. Frederic Jurie, Prof. Tinne Tuytelaars, and Prof. Catherine Achard**. Their presence in this special day despite their highly charged agendas is really an honour for me.

This work has been funded by the region Hauts-de-France and Institut-Mines-Telecom Lille-Douai.

I would like to thank all my colleagues and the research group GT-image for the stimulating discussions, help and suggestions: Dr. Hassen Drira, Dr. Hazem Wannous, Dr. Jose Mennesson, Dr. Pierre Tirilly, Prof. Jean-Philipe Vandeborre, Dr. Taleb Alashkar, Dr. Meng Meng, Dr. Maxime Devannes, Dr. Quentin De Smedt, Dr. Paul Audain Desrosiers, Dr. Vincent Leon, Dr. Vicent Itier, Sarah Ribet, Oussema Bouafif, Benjamin Szczapa. I thank our collaborator Naima Otberdout, from the University of Rabat in Morocco for the fruitful discussions and for her support. My special thanks go to my close friend and colleague Omar Ben Tanfous for his help and support, for the sleepless nights where we were working

together before deadlines, and for all the fun we had in the last three years. I would like to thank my friends for accepting nothing less than excellence from me: Souhail, Alaeddine, David, Ile, Sameh, Aymen, Slim, Mehdi, Hsan, Maha, Faten, Skander, Ilyes, Nedra, Syrine, Imen, Samia, Fares, Malek, Ikbel, Enjie, Ahmed, and Eya.

Last but not least, I must express my very profound gratitude to my parents Jalel and Karima. Their infinite support, valuable advises, and compassion were the secret ingredients to reach the end of my PhD successfully. Thanks go to my grandmothers Dalila and Mansoura for their continuous prayers, to my brother Samih for his permanent support and precious advises, to my sister Syrine and his little family for providing me continuous encouragements along these last three years of study. This accomplishment would not have been possible without them. I also dedicate this work for my deceased aunt Amel who highly contributed in making me a better person.

Table of contents

List of figures

List of tables

Nomenclature

$\mathcal{O}(d)$ Orthogonal group / the set of $d \times d$ orthogonal matrices

$\mathcal{SO}(d)$ Special orthogonal group / the set of $d \times d$ orthonormal matrices

$so(d)$ Lie algebra of the special orthogonal group $\mathcal{SO}(d)$ / the tangent space at the identity matrix in $\mathcal{SO}(d)$

$\mathcal{V}_{d,n}$ Stiefel manifold / the set of $n \times d$ orthonormal matrices

$\mathcal{G}(d,n)$ Grassmann manifold / the set of $d-$subspaces in $\mathbb{R}^n$

$\mathcal{P}_d$ Riemannian manifold of $d \times d$ positive defnite matrices

$\mathcal{S}^+(d,n)$ Riemannian manifold of $n \times n$ positive semidefnite matrices of fixed rank d

Z Landmark configuration

G Gram matrix of a landmark configuration

$\beta_G(t)$ Gram matrix trajectory of a landmark sequence

Λ Barycentric coordinates of a facial landmark configuration

$R_{\alpha,\beta,\gamma}$ Head pose representation in the special orthogonal group $\mathcal{SO}(3)$ formed by the 3 degrees of freedom yaw α, pitch β, and roll γ

H Head pose representation in Lie algebra of a landmark configuration

Chapitre 1

Introduction

1.1 Motivation and challenges

Developing intelligent systems dedicated to human behavior understanding has been a very hot research topic in the few recent decades. Indeed, it is crucial to understand the human behavior in order to make machines able to interact with, assist, and help humans in their daily life. The need for such tools is more acute for health care applications. Recent breakthroughs in Computer Vision and Machine Learning have made this possible. For instance, human-related Computer Vision problems can be approached by first detecting and tracking landmark points from visual data. One relevant illustration is given by the 3D locations of the body joints, termed 3D skeletons, automatically detected in depth streams of human bodies, and their use in action and daily activity recognition. As far as the human face communicates important behavioral and feeling cues, several approaches have addressed the problem of 2D/3D facial landmark points detection and tracking in video flows of human faces. These techniques generate temporal sequences of landmark configurations, which exhibit several distortions in their analysis, especially in uncontrolled environments, due to view variations (*e.g.*, affine or projective transformations), inaccurate

detection and tracking, missing data, etc. This thesis is mainly focused on analyzing these temporal sequences with the aim of proposing novel effective space-time representations along with suitable computational tools for human behavior understanding.

1.2 Thesis contributions

In this thesis, we propose novel geometric tools for human behavior understanding. Specifically, we consider the moving 2D/3D tracked landmark points on the human face or body and propose effective representations along with suitable analyzing tools for human behavior understanding. The main contributions of this thesis can be summarized to:

— A novel space-time representation of human landmark sequences (tracked from faces and bodies) based on Gram matrix trajectories of landmark configurations [P1, P5, P6]. Despite the large use of these matrices in several research fields, to our knowledge, this is the first application in shape analysis. The space of Gram matrices of n landmark points, termed the cone of Positive Semi-Definite $n \times n$ matrices of fixed-rank d ($d = 2$ or $d = 3$ for 2D or 3D landmark configurations, respectively) is a non-linear manifold where standard computational and machine learning tools are not applicable. To overcome this problem, a comprehensive study of the Riemannian geometry of the Positive Semi-Definite cone is conducted to derive suitable analyzing and classification tools for Gram matrix trajectories.

— Evaluation of the proposed framework on Gram matrix trajectories in different human behavior understanding tasks involving 2D facial landmarks and 3D skeletons tracked on the human body. Specifically, we evaluate the effectiveness of the proposed approach in 2D facial expression recognition [P5, P6] and action and emotion recognition from 3D skeletons [P1] on several benchmarks and demonstrate its competitiveness with respect to the state-of-the-art.

— Another affine-invariant representation for the specific case of 2D facial landmarks

based on their barycentric coordinates [P4]. We show that such representation is closely related to the conventional Grassmannian representation which is a part of the Gram matrix representation. In contrast to the non-linear Grassmannian representation, the barycentric representation lie in Euclidean space allowing the use of standard computational and machine learning tools.

— Evaluation of the effectiveness of the barycentric representation in two different facial analysis tasks, namely facial expression recognition [P4] and depression severity level assessment [P3].

1.3 Organization of the manuscript

The manuscript is organized as follows: In chapter 2, we will introduce the task of human behavior understanding and the use of tracked human landmark sequences to tackle this problem, then review the related recent state-of-the-art approaches. In chapter 3, we will present a novel geometric framework on Gram matrix trajectories and its evaluation in facial expression recognition from 2D facial landmarks and action and emotion recognition from 3D skeletons. Chapter 4 introduces another representation for the specific case of 2D facial landmark sequences based on their barycentric coordinates with applications to facial expression recognition and depression severity level assessment. Finally, in chapter 5 we will conclude this thesis, expose its limitations, and present some ongoing and future work.

Publications

[P1] **A. Kacem**, M. Daoudi, B. Ben Amor, S. Berretti, J.C. Alvarez-Paiva, A Novel Geometric Framework on Gram Matrix Trajectories for Human Behavior Understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), accepted in August 2018 for publication in an upcoming issue.

[P2] N. Otberdout, **A. Kacem**, M. Daoudi, L. Ballihi, S. Beretti, Deep Covariance Descriptors for Facial Expression Recognition, September 2018, British Machine Vision Conference (BMVC 2018).

[P3] **A. Kacem**, Z. Hammal, M. Daoudi, J.F. Cohn, Detecting Depression Severity by Interpretable Representations of Motion Dynamics, IEEE International Conference on Automatic Face and Gesture Recognition, Workshop on Analysis for Health Informatics (FG-AHI 2018).

[P4] **A. Kacem**, M. Daoudi, J.C. Alvarez-Paiva, Barycentric Representation and Metric Learning for Facial Expression Recognition, IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018).

[P5] **A. Kacem**, M. Daoudi, B. Ben Amor, J.C. Alvarez-Paiva, A Novel Space-Time Representation on the Positive Semidefinite Cone for Facial Expression Recognition, IEEE International Conference on Computer Vision (ICCV 2017), pp. 3180-3189.

[P6] **A. Kacem**, M. Daoudi, B. Ben Amor, Analyse de Trajectoires sur des Variétś de Matrices pour la Reconnaissance des Expressions Faciales, Journés Francophones des Jeunes Chercheurs en Vision par Ordinateur (ORASIS 2017).

Chapitre 2

State-of-the-art on Analyzing Landmark Sequences for Human Behavior Understanding

2.1 Introduction

Several human behavior understanding methods firstly detect and track a set of landmark points and use them for the analysis of the video. Two relevant examples of these landmark points are given by the tracked skeletons of the human body and the tracked fiducial points on the human face. With this assumption, the problem of analyzing videos is turned to analyzing the motion of the landmark points. In this chapter, we will introduce the task of human behavior understanding and its applications in real world. Then, we will expose the motivations and challenges of using only human landmark sequences for this task and focus on the state-of-the-art on analyzing them.

2.2 Human behavior understanding

2.2.1 Terminology

Human behavior is the responses of individuals or groups of humans to internal and external stimuli. It refers to the array of every physical action and observable emotion associated with individuals [134]. These responses, usually termed behavioral signals, consist of a set of temporal changes in neuromuscular and physiological activity that can last from a few milliseconds (a blink) to minutes (talking) or hours (sitting) [87]. As explained in [41], other types of messages conveyed by behavioral signals include affective states (*e.g.*, fear, joy, stress), manipulators (actions used to act on objects in the environment or self-manipulative actions like lip biting), emblems (culture-specific interactive signals like wink or thumbs up), and so on.

In this thesis, we are interested in endowing machines with intelligent systems that are able to understand some of these human behavioral signals from visual data. That is to say, given a video of a person conveying a behavioral signal (*e.g.*, joy, drinking water, fear, etc.) we would like to make machines able to automatically recognize the nature of this signal.

2.2.2 Applications

Understating the human behavior has a broad range of applications in different fields.

— Human computer interaction: Human computer interaction designs were first dominated by direct manipulation and then delegation. They involved conventional interface devices such as keyboard, mouse and visual displays, and assumed that the human will be explicit and fully attentive while controlling information and command flow [87]. Accurate human behavior understanding tools can highly improve the interfaces between humans (users) and computers (cars, robots, etc.) by providing a more natural, less-restrictive, and effective human-computer interfaces.

— Health care: The need for developing intelligent systems dedicated to human behavior understanding is more acute in health care. Indeed, these intelligent systems can assist clinicians in their diagnosis and help them in effectively applying treatments. Taking this direction, several works tried to automatically measure the intensity of pain level [146] from human faces, other works tried to measure the level of depression severity [35].

— Social psychology: In social psychology, researchers study the psychological processes involved in persuasion, conformity, and other forms of social influence. Human behavior understanding solutions are crucial in order to better understand these processes since they are usually observable. For instance, based on the assumption of the universality of basic facial expressions [40], several works tried to automatically recognize these facial expressions.

— Surveillance and security: Violent extremism and evolving terrorist threat raise a persistent risk of attacks which reinforce the critical requirement for anticipating and responding to evolving threats. Understanding human behaviors can help with this issue by anticipating dangerous human interactions (*e.g.*, punching, kicking, etc.) [138] or analyzing the affective state of suspected persons.

In literature, two basic human behavior understanding tasks were extensively studied. The first task is facial expression recognition which consists of automatically recognizing one of the basic facial expressions conveyed by a person during a time slot (*e.g.*, anger, disgust, fear, joy, neutral, sadness, surprise). The second task consists of recognizing actions performed by humans based on their bodies. In this thesis, we will focus on these two basic tasks and tackle two other emerging tasks, namely emotion recognition from body movements and depression severity level assessment from human faces.

2.3 Human landmarks

Several human-related Computer Vision problems can be approached by first detecting and tracking landmarks from visual data.

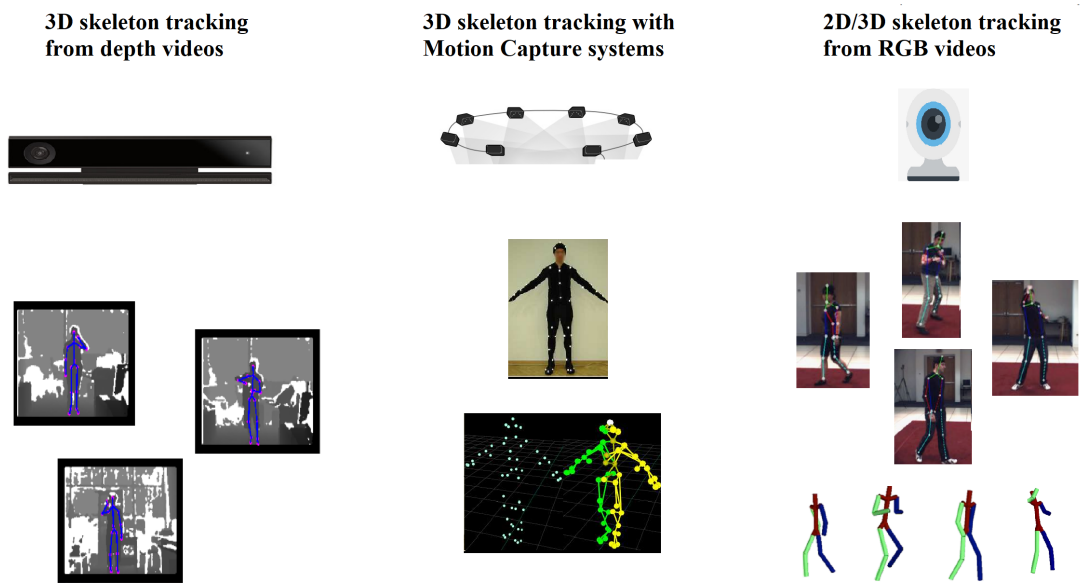

FIGURE 2.1 – Examples of human skeletons detected from different modalities [55, 89, 18]

A relevant example of this is given by the estimated 3D location of the joints of the skeleton in depth streams [106], and their use in action and daily activity recognition [123, 131, 59]. In this case, for each frame of the depth video a set of 3D joints are detected on some articulations of the human body forming a 3D skeleton. In Fig. 2.1, we show an example of a tracked skeleton in a depth video provided by a Kinect V2 sensor. Hence, the problem of analyzing human body motion in a depth video could be efficiently turned to studying the motion of the 3D skeleton along the video. More sophisticated solutions for automatic tracking of the 3D skeleton do exist, as the IR body markers used in MoCap systems, but they are expensive in cost and time. These systems provide a large number of joints with high temporal resolution and accurate estimations (see Fig. 2.1). Recently, advances in human pose estimation methods from RGB videos have also made the tracking of 2D/3D skeletons

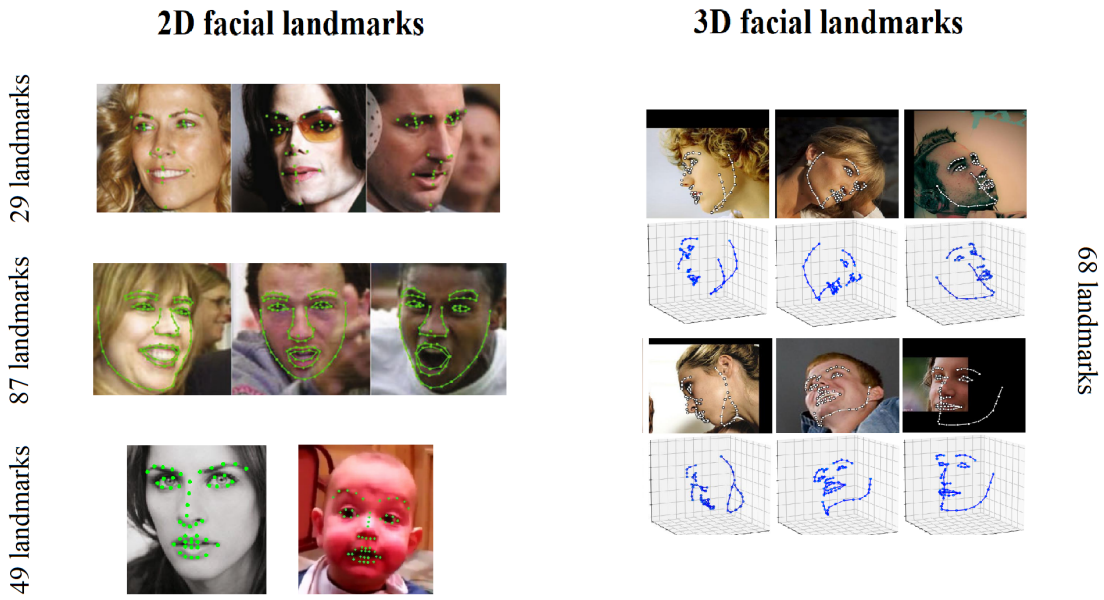

FIGURE 2.2 – Examples of 2D/3D facial landmark detection from RGB videos [21].

in RGB videos possible and have shown an impressive performance [113, 5, 18].

Another relevant example of human landmark tracking is represented by the face, for which several approaches have been proposed for fiducial points detection and tracking in video [8, 137, 23, 21]. These methods detect a set of 2D key points localized at relevant positions of the human face. For instance, several methods opted for detecting landmark points around the eyes, eyebrows, nose, and mouth. Other systems, considered additional landmarks around the chin. In the left panel of Fig. 2.2, we show some examples of 2D facial landmark estimations. One can note that such estimations could lead to distortions in the analysis due to large pose variations. To overcome this problem, some works tried to estimate the 3D locations of these landmark points from only RGB videos [21, 114]. Examples of these 3D estimations are illustrated in the right panel of Fig. 2.2.

It is important to note that, in addition to their impressive performance, most of these methods are real-time solutions for tracking human landmarks.

2.3.1 Why landmark sequences for human behavior understanding?

In this thesis, we will focus on designing effective landmark based solutions for some human behavior understanding tasks. One of our motivations for this choice is driven by the recent impressive advances in human landmark tracking. As mentioned above, recently landmark detection and tracking methods from human faces and bodies became reliable and accurate. They are robust to illumination changes that occur in RGB images, and in some cases robust to occlusions (see the woman wearing sunglasses in the left panel of Fig. 2.2). By considering the tracked landmarks instead of the original images, we take advantage of the robustness of tracking methods to these classical problems in Computer Vision and expect the same robustness for our landmark based solutions.

Furthermore, considering only tracked landmarks reduces the complexity of the visual data. Instead of using a large number of pixels in each frame of the original video, which could make the analysis computationally intense, landmark trackers bring a brief summary of the frame by providing only a set of relevant 2D/3D points (the number of points typically varies from 15 to 90 points). Hence, landmark based solutions are expected to be more efficient and less computational expensive than other solutions, which makes them more suitable for real-time applications.

2.3.2 Challenges

While powerful and robust to many Computer Vision problems, human landmark tracking techniques generate temporal sequences of landmark configurations which exhibit several challenges:

— View variations: The 2D or 3D locations provided by the coordinates of the tracked landmarks are relative to the position of the camera. However, human behavioral signals belonging to the same category (*e.g.*, drinking water), can occur in different positions w.r.t the camera. In Fig. 2.3 we show some examples of static landmark

configurations (skeletal and facial landmarks) conveying similar behavioral signals but in different positions w.r.t the camera. These variations prevent us from directly using the original 2D or 3D locations of the landmark points. Accordingly, one should filter out these view variations from the estimated landmarks in order to effectively analyze the human behavioral signals. From the viewpoint of static landmark configurations, these view variations can be seen as undesirable rigid transformations affecting the landmarks which can be summarized to rotations, translations, and scaling in the 3D case, and to more complex projective transformations in the 2D case of landmarks.

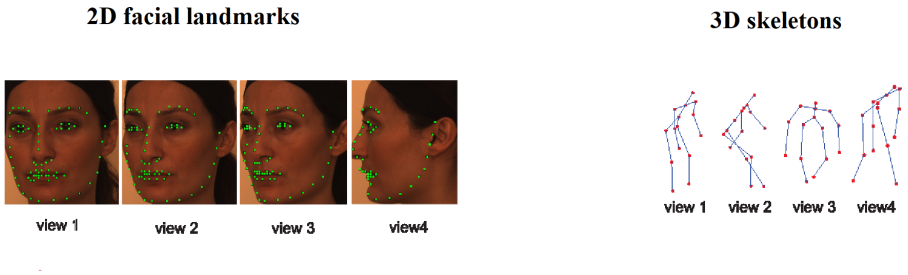

FIGURE 2.3 – Impact of the view variations on the human landmarks. Left: 2D facial landmarks with different views. Right: 3D skeletons with different views.

— Rate variations: The human behavioral signals that we would like to analyze are subject to high temporal variations. For instance, two persons do not perform the same action (*e.g.*, drinking water) at the same time and for the same duration. Consequently, we cannot simply compare the static landmark configurations of the two corresponding landmark sequences in order to know whether they are similar or not. Effective landmark based solutions should take into account these temporal (rate) variations in the analysis of human landmark sequences.

— Intra-class variations: Another challenge of human behavior understanding from landmark sequences consists of the large variations that can be present within the same category of human behavioral signals. Indeed, behavioral signals of the same category could be different from one person to another or even for the same person.

A relevant example of this is given by the facial expressions (*e.g.*, sadness) which can be expressed differently by different persons (see Fig. 2.4).

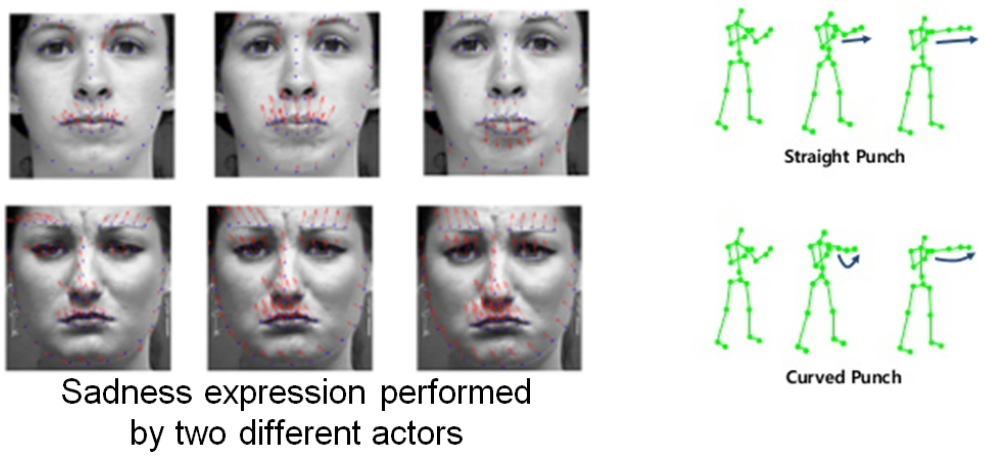

FIGURE 2.4 – Examples of intra-class variations. Left: facial expression (sadness) [93]. Right: human action (punch) [2].

— Inaccurate tracking and missing data: Despite the advances in tracking human landmarks as mentioned in the previous section, inaccurate tracking can occur especially in unconstrained environments and challenging conditions. Fig. 2.5 shows some failure cases of landmark detection from human faces (left) and bodies (right).

While there have been many efforts in the analysis of temporal sequences of landmarks, the problem is far from being solved and the current solutions are facing many technical and practical problems.

2.4 Temporal modeling and classification of landmark sequences

In this section, we review some recent state-of-the-art methods on analyzing human landmark sequences for some human behavior understanding tasks. In particular, we present

2. Example taken from www.slideshare.net/NaverEngineering/human-action-recognition

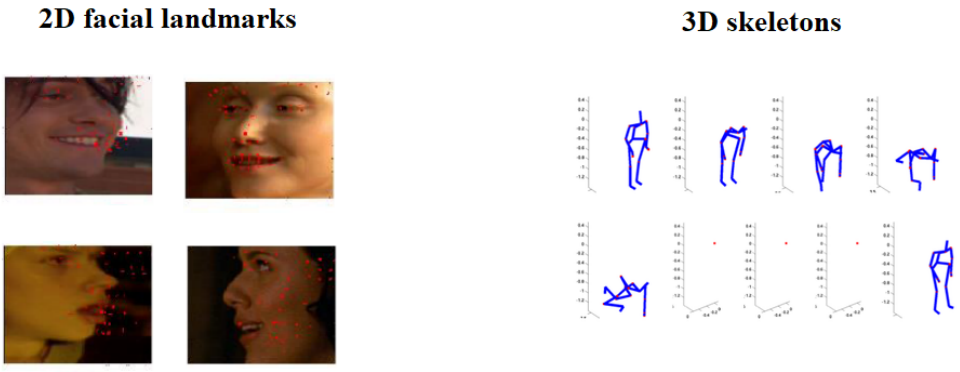

FIGURE 2.5 – Examples of inaccurate tracking of human landmarks. Left: Failure cases of 2D facial landmarks. Right: Failure cases of 3D skeletons.

some recent works that use 2D or 3D landmarks of human faces or bodies (*i.e.*, skeletons) with applications to human behavior understanding. These state-of-the-art approaches are organized into four categories: Riemannian methods, deep learning methods, kernel methods, and probabilistic methods with a focus on Riemannian approaches. An overview of the considered works and their categorizations is sketched in Fig. 2.6.

2.4.1 Probabilistic methods

Several approaches included the use of probabilistic models for different applications of human behavior understanding.

The authors in [83] explored the use of Hidden Markov Models (HMMs) in 3D action recognition. They decomposed the human skeleton into different body parts (*i.e.*, legs+torso, arms, and head) and learned the dynamics of each body part with a single HMM forming a weak classifier. A boosting algorithm is finally used on these weak classifiers to provide a final prediction. HMMs were also adopted by several works after a feature extraction step. For instance, in [136] histograms of 3D joints were computed and encoded into a sequence of visual words which were modeled and classified using HMMs.

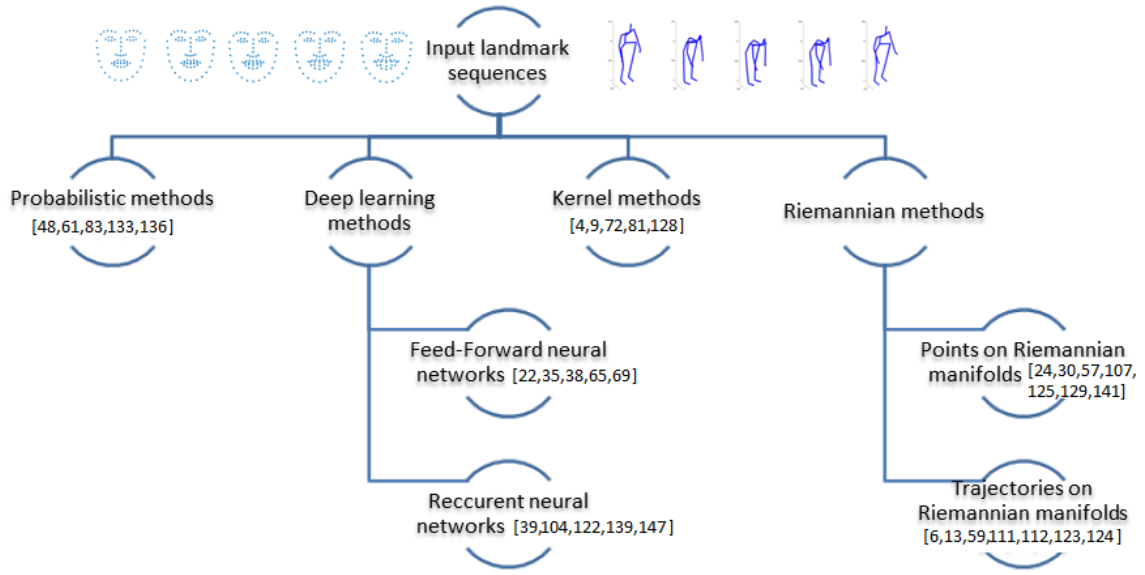

FIGURE 2.6 – Overview of state-of-the-art methods on analyzing landmark sequences for human behavior understanding.

Other probabilistic models such us Conditional Random Fields (CRFs) were also adopted. In the context of 2D facial expression recognition, the authors in [61] proposed a method to capture the subtle motions within expressions using a variant of CRFs called Latent-Dynamic Conditional Random Fields (LDCRFs) on both geometric and appearance features. They illustrate experimentally that variations in shape are much more important than appearance for facial expression recognition.

Time-slice based methods such us HMMs [83] or LDCRFs [61] represent an activity as a sequence of instantaneously occurring events, and as a result they can only capture a small portion of the temporal relations. Starting from this observation, Wang *et al.* [133] introduced a unified probabilistic framework based on an Interval Temporal Bayesian Network (ITBN) built from the movements of 2D facial landmarks. ITBN models a complex activity as sequential or overlapping primitive events (*i.e.*, temporal entity), and each event spans over a time interval. The authors show that the proposed ITBN outperforms other time-slice

based methods such us HMMs in recognizing facial expressions.

Most of the methods listed above focus on modeling the transitions between the frames in order to capture the changes in human landmarks. However, important patterns could be provided by discriminative static observations as well [127]. Aware of this issue, G. Hernando *et al.* [48] proposed a forest-based classifier called transition forests to discriminate both static pose information and temporal transitions between pairs of two independent frames. Applications were shown in 3D action recognition and detection.

2.4.2 Kernel methods

Over the last years, kernel methods have established themselves as powerful tools for many Computer Vision tasks. Based on the fundamental concept of defining similarities between objects they allow, *e.g.*, the prediction of properties of new objects based on the properties of already known ones [74].

Taking this direction, the authors in [81] proposed two time-series kernels computed from 3D facial landmarks for expression recognition. Specifically, they considered the temporal evolution of normalized 3D facial landmarks as a time-series in $\mathbb{R}^{3\times n}$, where n denotes the number of landmark points. A pseudo kernel based on Dynamic Time Warping (DTW) similarities was derived from all the time-series in the dataset. DTW is a dynamic programming based algorithm that allows a temporal alignment of two time-series. Since DTW was adopted, the computed kernel is not positive definite, thus does not satisfy Mercer's theorem. Consequently, an approximated version of this kernel was considered. Another global alignment kernel, which is a smoother version of DTW but results in a positive definite kernel, was also used in this work.

More recently, Bagheri *et al.*, [9] tackled the problem of 3D action recognition by computing time-series kernels. Here also, a DTW kernel was computed but was not approximated with a positive definite kernel. The authors introduced another kernel based

on Longest Common Subsequence (LCSS) similarity measure which consists of counting the number of pairs of points from two sequences that match. In contrast to [81], where approximated versions of the computed kernels were used for SVM classification, the authors in this work opted for another variant of SVM called pairwise proximity function SVM (ppfSVM) [50]. The latter learns a proximity model of the data and only requires the definition of a proximity function which can be the DTW or LCSS similarity measures.

From the other perspective, the authors in [72] proposed two kernel-based tensor representations named sequence compatibility kernel (SCK) and dynamics compatibility kernel (DCK) based on a set of RBF kernels computed over 3D skeletal sequences. These can capture the higher-order relationships between the joints. The first captures the spatio-temporal compatibility of joints between two sequences, while the second kernel uses the intra-sequence joint differences, thus capturing the dynamics as the spatio-temporal co-occurrences of the joints. Tensors are then formed from these kernels to train SVM.

Finally, Multiple Kernel Learning (MKL) was also adopted on different extracted spatio-temporal features from human landmark sequences [128, 4]. In these works, MKL have shown an impressive performance in fusing different features at the kernel level of SVM classifiers.

2.4.3 Deep learning methods

Recently, Deep Learning (DL) became one of the most powerful tools in many Computer Vision tasks. The idea behind DL is to learn the best features to the problem at hand, by defining suitable objectives and network architectures. Many recent approaches for analyzing human landmark sequences used DL in order to jointly model the dynamics (*i.e.*, extract features) and classify the landmark sequences for human behavior understanding. These approaches can be categorized in two groups; the group of methods using feed-forward neural networks (*e.g.*, Convolution Neural Networks, Auto-encoders, etc.) and the group of methods using Recurrent Neural Networks (RNNs). In RNNs, network units have recurrent

connections such that information about previous activations can be propagated over time. In contrast to RNNs, the information in feed-forward networks moves in only one direction, forward, from the input nodes, through the hidden nodes, to the output nodes.

2.4.3.1 Feed-forward neural networks

In [65], the authors proposed a neural network architecture called Deep Temporal Geometry Network (DTGN) for facial expression recognition from 2D facial landmark sequences. The facial landmarks were firstly normalized then concatenated over time to form a single vector representation which is fed to a neural network. The architecture of DTGN consists of Fully Connected (FC) layers and softmax.

In the context of 3D action recognition, the authors in [38] proposed to use Convolutional Neural Networks (CNNs). Specifically, the three coordinates of all skeleton joints in each frame were separately concatenated by their physical connections. A matrix was then generated by arranging the representations of all frames in chronological order, then quantified and normalized into an image. The obtained image represented the skeletal sequences and was finally fed into a hierarchical spatial-temporal adaptive filter banks model for representation learning and recognition. CNNs were also investigated in 3D action recognition in [69], but in a different way. The authors generated three clips corresponding to the three channels of the cylindrical coordinates of a skeleton sequence. A deep CNN model and a temporal mean pooling layer were used to extract a compact representation from each frame of the clips. The output CNN representations of the three clips at the same timestep were concatenated, resulting in different feature vectors. Another neural network (FC layers and Softmax) was used on these feature vectors for action classification.

Dibeklioglu *et al.*, [35] tackled the problem of measuring depression severity level from 2D facial landmark sequences. They used Stacked Denoising Auto-Encoders (SDAE) to encode the static observations of 2D facial landmark sequences. By doing so, the authors

obtained a more discriminative low-dimensional feature representation of the static facial landmarks. They exploited this representation to derive motion features such us velocities and accelerations. Deep auto-encoders were also explored for 3D action recognition. For instance, they were used in [22] to encode the dynamics of the skeletal sequences. In this work, three different temporal encoder structures were proposed (*i.e.*, symmetric, time-scale, and hierarchy encoding) which were designed to capture different spatial-temporal patterns.

2.4.3.2 Recurrent neural networks

Several solutions have experimented the application of Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) networks to the case of 2D/3D human landmarks for human behavior understanding.

This approach was followed by Veeriah *et al.* [122] who presented a family of differential RNNs (dRNNs) that extend LSTM by a new gating mechanism to extract the derivatives of the internal state (DoS). The DoS was fed to the LSTM gates to learn salient dynamic patterns in 3D skeleton data.

Du *et al.* [39] proposed an end-to-end hierarchical RNN for skeleton based action recognition. First, the human skeleton was divided into five parts, which are then feed to five subnets. As the number of layers increases, the representations in the subnets are hierarchically fused to be the inputs of higher layers. The final representations of the skeleton sequences are fed into a single-layer perceptron, and the temporally accumulated output of the perceptron is the final decision.

To ensure effective learning of the deep model, Zhu *et al.* [147] designed an in-depth dropout algorithm for the LSTM neurons in the last layer, which helps the network to learn complex motion dynamics. To further regularize the learning, a co-occurrence inducing norm was added to the network's cost function, which enforced the learning of groups of co-occurring and discriminative joints.

A part aware LSTM model was proposed by Shahroudy *et al.* [104] to utilize the physical structure of the human body to improve the performance of the LSTM learning framework. Instead of keeping a long-term memory of the entire body's motion in the cell, this is split to part-based cells. In this way, the context of each body part is kept independently, and the output of the part based LSTM (P-LSTM) unit is represented as a combination of independent body part context information. Each part cell has therefore its individual input, forget, and modulation gates, but the output gate is shared among the body parts.

LSTMs were also used in combination with RNNs. For example, the authors in [139] decomposed the 2D facial landmark configurations into different facial parts (*e.g.*, eyes, mouth, etc.), then used bi-directional RNNs and LSTMs to learn the dynamics of facial expressions from these facial parts.

While being well-suited for periodic data, RNNs and LSTMs perform less well when confronted with aperiodic time series [22].

2.4.4 Riemannian methods

Most of the approaches listed above, did not take into account the geometric nature of the feature space. Indeed, the extracted features or representations of the landmark sequences may lie on non-linear manifolds where standard computational and machine learning techniques are not applicable in a straightforward manner. A well-know example of this is given by the covariance matrices which are positive definite matrices and lie on a non-linear manifold [117, 7, 24, 84]. To illustrate this issue, let us consider two points that correspond to the feature representations of two landmark sequences. Assume that these points lie on a non-linear space (*e.g.*, a linear combination of them may lie out of the original space). We show an example of this illustration in Fig. 2.7. If we would like to compute the Euclidean distance between them, we would connect them with a straight line, as show in red in Fig. 2.7, and measure its length. This measure would not inform on the

real proximity of these two points on the underlying feature space. In contrast, one should find a geodesic path connecting these two points which is the shortest path connecting them on the non-linear space, as depicted in green in Fig. 2.7, and measure its length to obtain a geodesic distance. By doing so, we are given a more meaningful measure about the proximity of the feature points on the manifold.

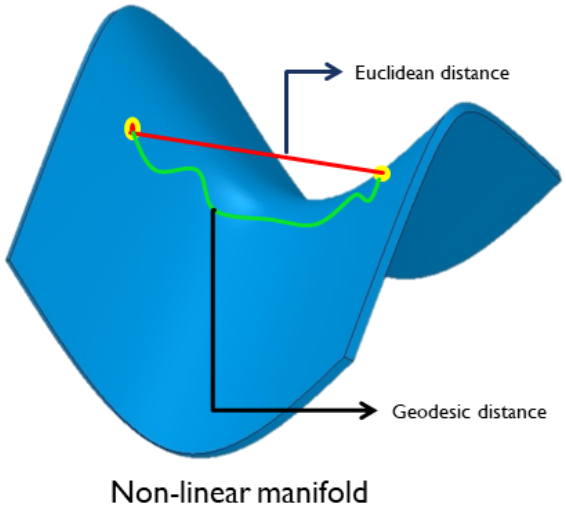

FIGURE 2.7 – Illustration of the non-linearity problem. Best viewed in color.

This issue opened the way to the use of metric and differential-geometric techniques in the study and classification of moving landmarks. Taking this direction, several works opted for the use of Riemannian geometry in order to overcome this problem [123, 13, 30, 67, 6]. The idea here is to define a smoothly varying inner product on each tangent space of the manifold to obtain a Riemannian metric. By defining a Riemannian metric on the manifold, one can locally exploit the vector space structure of the tangent space to define various geometric notions on the manifold including the geodesic distance mentioned above. Other important notions are the logarithm and exponential maps. The former is an operation that maps a point on a Riemannian manifold to a tangent space attached to another point on the manifold. The exponential map is its inverse operation. Further explanations of the notion of

Riemannian manifolds are provided in the next chapter. In what follows, we will present two families of Riemannian methods for analyzing human landmark sequences. Given a sequence of human landmarks, the first family embeds this sequence into one feature representation lying on a Riemannian manifold while the second represents the moving landmarks as a time-parametrized curve (*i.e.*, trajectory) on a Riemannian manifold.

2.4.4.1 Landmark sequences as points on Riemannian manifolds

In the work of Slama *et al.* [107] for 3D action recognition, a temporal sequence was represented as a Linear Dynamical System (LDS). The observability matrix of the LDS was then approximated by a finite matrix [115]. The subspace spanned by the columns of this finite observability matrix corresponds to a point on a Grassmann manifold. Thus, the LDS is represented at each time-instant as a point on the Grassmann manifold. Each video sequence is modeled as an element of the Grassmann manifold, and action learning and recognition is cast to a classification problem on this manifold. Proximity between two spatio-temporal sequences is measured by a distance between two subspaces on the Grassmann manifold. Taking the same direction, Huang *et al.* [57] formulated the LDS as an infinite Grassmann manifold, and proposed a formulation for sparse coding and dictionary learning on this manifold. One drawback of these methods is that LDS can only capture linear relationship between successive frames. Aware of this limitation, Venkataraman *et al.* [125] proposed a shape-theoretic framework for analysis of non-linear dynamical systems. Applications were shown to activity recognition using motion capture and RGB-D sensors, and to activity quality assessment for stroke rehabilitation.

Taking another direction, Devanne *et al.* [30] proposed to formulate the action recognition task as the problem of computing a distance between trajectories generated by the joints moving during the action. An action is then interpreted as a parametrized curve and is seen as a single point on the hyper-sphere by computing its Square Root Velocity Function (SRVF) [109]. However, this approach does not take into account the relationship between

the joints.

The authors of [24] and [129] proposed to map full skeletal sequences into the manifold of Symmetric Positive Definite (SPD) matrices. That is, given an arbitrary sequence, it is summarized by a covariance matrix, which is a SPD matrix, derived from the velocities computed from neighboring frames or from the 3D landmarks themselves, respectively. In both of these works kernelized versions of covariance matrices are considered.

Zhang *et al.* [141] represented temporal landmark sequences using regularized Gram matrices derived from the Hankel matrices of landmark sequences. The authors show that the Hankel matrix of a 3D landmark sequence is related to an Auto-Regressive (AR) model [76], where only the linear relationships between landmark static observations are captured. The Gram matrix of the Hankel matrix is computed to reduce the noise and is seen as a point on the positive semi-definite manifold. To analyze/compare the Gram matrices, they regularized their ranks resulting in positive definite matrices and considered metrics on the positive definite manifold. This approach was evaluated in the 3D action recognition task.

2.4.4.2 Landmark sequences as trajectories on Riemannian manifolds

One promising idea is to formulate the motion features as trajectories on the underlying manifolds. Indeed, features computed from static landmark configurations often lie on non-linear manifolds [123, 124, 111, 13]. Hence, landmark sequences can be seen as trajectories on this manifold. In contrast to the first family of Riemannian methods, the temporal structure of landmark sequences is preserved allowing desirable operations in the manifold such us interpolation.

Taking this direction, Taheri *et al.* [111] proposed to represent 2D facial landmarks in the Grassmann manifold. This representation is invariant to affine transformations allowing a robust analysis under view variations. In order to capture the facial expressions from these landmark representations, the authors computed the velocity vectors between

successive frames using the logarithm map. A parallel transport of these velocity vectors to a fixed tangent space of the manifold was also used in this work in order to have all the velocity vectors in the same tangent space. By mapping all the velocity vectors to a fixed tangent space, this method depends on the chosen fixed tangent space and involves several approximations which can introduce distortions in the analysis.

In [123], Vemulapalli *et al.* proposed a Lie group trajectory representation of the skeletal data on the product space of Special Euclidean (SE) groups for 3D action recognition. For each frame, the latter representation is obtained by computing the Euclidean transformation matrices encoding rotations and translations between different joint pairs. The temporal evolution of these matrices is seen as a trajectory on $SE(3) \times \cdots \times SE(3)$ and mapped to the tangent space of a reference point. A one-versus-all SVM, combined with Dynamic Time Warping (DTW) and Fourier Temporal Pyramid (FTP) is used for classification. One limitation of this method is that mapping trajectories to a common tangent space using the logarithm map could result in significant approximation errors. Aware of this limitation, the same authors proposed in [124] a mapping combining the usual logarithm map with a rolling map that guarantees a better flattening of trajectories on Lie groups. Based on the same lie group representation of human skeletons, the authors in [59] proposed a deep network architecture in lie groups. The proposed network transforms the lie group representations (*i.e.,* rotation matrices) into more desirable ones for action recognition. Several special layers were introduced in this work (*e.g.*, RotMap layer, RotPooling layer, etc.).

Anirudh *et al.* [6] started from the two Riemannian trajectory based representations mentioned above, in Lie Groups [123] and in Grassmann manifold [111]. They proposed a statistical framework for analyzing Riemannian trajectories called Transported Square-Root Velocity Fields (TSRVF), which has desirable properties including a rate-invariant metric and vector space representation. Based on this framework, they proposed to learn an embedding such that each trajectory is mapped to a single point in a low-dimensional Euclidean space, and the trajectories that differ only in temporal rates map to the same

point. The TSRVF representation and accompanying statistical summaries of Riemannian trajectories are used to extend existing coding methods such as PCA, KSVD, and Label Consistent KSVD to Riemannian trajectories. In the experiments, it is shown such coding efficiently captures trajectories in action recognition, stroke rehabilitation, visual speech recognition, clustering, and diverse sequence sampling.

Ben Amor *et al.* [13] represented 3D skeletal shapes on the Kendall's shape space by removing translations, rotations, and scaling information for the purpose of 3D action recognition. A landmark sequence is then seen as a trajectory on the Kendall's shape space. Following [110], they used an elastic metric that considers the time-warping on a Riemannian manifold, thus allowing trajectories registration and the computation of statistics on the trajectories (*e.g.*, resampling, mean trajectory, etc.). To classify trajectories (3D landmark sequences), the authors computed the mean trajectories of each class and extracted for each trajectory a feature vector formed by distances to mean trajectories of each class. However, the mean trajectory of a class is not a significant statistical summary of the trajectories belonging to the same class, especially in cases of high intra-class variations. Hence, the feature vector of distances to mean trajectories could not be robust to intra-class variations. Based on the same Kendall trajectory representation, Ben Tanfous *et al.* [112] used an intrinsic formulation for spare coding and dictionary learning to encode trajectories on Kendall's shape space. By doing so, a trajectory on Kendall's shape space is parsed to a sequence of sparse codes that can be fed to any standard machine learning pipeline. Two classification pipelines were used for the task of 3D action recognition: a pipeline of DTW-FTP-SVM, and a bidirectional LSTM.

2.4.4.3 Classification on Riemannian manifolds

As mentioned above, one problem that arises when considering a representation of landmark sequences in a Riemannian manifold is how to adapt machine learning techniques to effectively work on the manifold-valued data. In current literature, two families of

approaches have been used to handle the non-linearity of Riemannian manifolds:

— The first family maps the points on the manifold to a tangent space where traditional learning techniques can be used for classification [111, 123, 6]. Mapping data to a tangent space only yields a first-order approximation of the data that can be distorted, especially in regions far from the origin of the tangent space. Moreover, iteratively mapping back and forth, *i.e.*, Riemannian Logarithmic and Exponential maps, to the tangent spaces significantly increases the computational cost of the algorithm.

— The second family embeds a manifold in a high dimensional Reproducing Kernel Hilbert Space (RKHS), where Euclidean geometry can be applied [62, 24, 129]. The Riemannian kernels enable the classifiers to operate in an extrinsic feature space without computing tangent space and log and exp maps. Many Euclidean machine learning algorithms can be directly generalized to an RKHS, which is a vector space that possesses an important structure: the inner product. Such an embedding, however, requires a kernel function defined on the manifold which, according to Mercer's theorem, should be positive definite.

2.5 Conclusion

Motivated by the recent advances in human landmarks detection and tracking, we focused on landmark based solutions for human behavior understanding. However, in practice one should take into account several challenges exhibited by human landmark sequences (*e.g.*, view and rate variations, inaccurate tracking, etc.) in order to develop reliable human behavior understanding solutions. In this chapter, we presented a multitude of landmark based state-of-the-art solutions which were categorized into four main groups (*i.e.*, probabilistic, kernel based, deep learning, and Riemannian methods). Most of probabilistic methods focused more on modeling the transitions between static frames and neglected modeling static landmark configurations which could provide important patterns. In kernel

methods, one should define a positive definite kernel in order to satisfy Mercer's theorem. This puts additional constraints in defining suitable similarity measures between landmark sequences. While powerful, Deep Learning methods require a large amount of data to achieve the expected performance. However, collecting large visual datasets for human behavior understanding tasks is not straightforward.

Most of the approaches categorized above, did not take into account the geometric nature of the feature space. Indeed, the extracted features or representations of the landmark sequences may lie on non-linear manifolds where standard computational and machine learning techniques are not applicable in a straightforward manner. Riemannian methods use some basics of the Riemannian geometry to define suitable computational tools on special non-linear manifolds. These methods were categorized in two subgroups. The first group models a landmark sequence as a single point in a Riemannian manifold, while the second models it as a trajectory lying on the manifold. In contrast to the single point representation, trajectory based representation preserves the original temporal structure of the landmark sequences and provides desirable operations in the manifold such us interpolation. In this thesis, we will focus on Riemannian trajectory based representations of the landmark sequences for different human behavior understanding tasks such us action recognition and facial expression recognition.

Chapitre 3

Novel Geometric Framework on Gram Matrix Trajectories for Emotion and Activity Recognition

3.1 Introduction

In this chapter, we propose a novel space-time geometric representation of human landmark configurations and derive tools for comparison and classification. We model the temporal evolution of landmarks as parametrized trajectories of Gram matrices on the Riemannian manifold of positive semidefinite matrices of fixed-rank. Our representation has the benefit to bring naturally a second desirable quantity when comparing shapes – the spatial covariance – in addition to the conventional affine-shape representation. We derived then geometric and computational tools for rate-invariant analysis and adaptive re-sampling of trajectories, grounding on the Riemannian geometry of the underlying manifold. Specifically, our approach involves three steps: (1) landmarks are first mapped into the Riemannian manifold of positive semidefinite matrices of fixed-rank to build

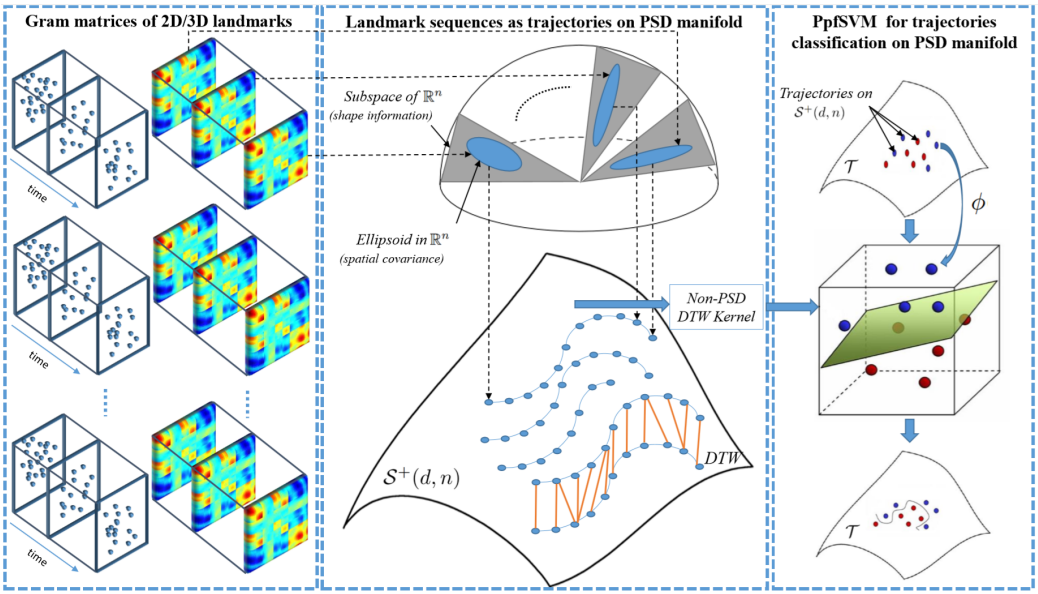

FIGURE 3.1 – Overview of the proposed approach. Given a landmark sequence, the Gram matrices are computed for each landmark configuration to build trajectories on $\mathcal{S}^+(d,n)$. A moving shape is hence assimilated to an ellipsoid traveling along d-dimensional subspaces of $\mathbb{R}^n$, with $d_{\mathcal{S}^+}$ used to compare static ellipsoids. Dynamic Time Warping (DTW) is then used to align and compare trajectories in a rate-invariant manner. Finally, the ppfSVM is used on these trajectories for classification.

time-parameterized trajectories; (2) a temporal warping is performed on the trajectories, providing a geometry-aware (dis-)similarity measure between them; (3) finally, a pairwise proximity function SVM is used to classify them, incorporating the (dis-)similarity measure into the kernel function. An overview of the proposed framework is shown in Fig. 3.1. We show that such representation and metric achieve competitive results in applications as action recognition and emotion recognition from 3D skeletal data, and facial expression recognition from 2D facial landmarks. Experiments have been conducted on several publicly available up-to-date benchmarks.

3.2 Gram matrix for shape representation

Let us consider an arbitrary sequence of landmark configurations $\{Z_0, \ldots, Z_\tau\}$. Each configuration Z_i $(0 \leq i \leq \tau)$ is an $n \times d$ matrix of rank d encoding the positions of n distinct landmark points in d dimensions. In our applications, we only consider the configurations of landmark points in two- or three-dimensional space (*i.e.*, d=2 or d=3) given by, respectively, $p_1 = (x_1, y_1), \ldots, p_n = (x_n, y_n)$ or $p_1 = (x_1, y_1, z_1), \ldots, p_n = (x_n, y_n, z_n)$. We are interested in studying such sequences or curves of landmark configurations up to Euclidean motions. In the following, we will first propose a representation for static observations, then adopt a time-parametrized representation for temporal analysis.

As a first step, we seek a shape representation that is invariant up to Euclidean transformations (rotation and translation). Arguably, the most natural choice is the matrix of pairwise distances between the landmarks of the same shape augmented by the distances between all the landmarks and their center of mass p_0. Since we are dealing with Euclidean distances, it will turn out to be more convenient to consider the matrix of the squares of these distances. Also note that by subtracting the center of mass from the coordinates of the landmarks, these can be considered as *centered*: the center of mass is always at the origin. From now on, we will assume $p_0 = (0, 0)$ for $d = 2$ (or $p_0 = (0, 0, 0)$ for $d = 3$). With this provision, the augmented pairwise square-distance matrix $\mathcal{D}$ takes the form,

$$\mathcal{D} := \begin{pmatrix} 0 & \|p_1\|^2 & \cdots & \|p_n\|^2 \\ \|p_1\|^2 & 0 & \cdots & \|p_1 - p_n\|^2 \\ \vdots & \vdots & \vdots & \vdots \\ \|p_n\|^2 & \|p_n - p_1\|^2 & \cdots & 0 \end{pmatrix},$$

where $\|\cdot\|$ denotes the norm associated to the l^2-inner product $\langle \cdot, \cdot \rangle$. A key observation is that the matrix $\mathcal{D}$ can be easily obtained from the $n \times n$ Gram matrix $G := ZZ^T$. Indeed,

the entries of G are the pairwise inner products of the points $p_1, \ldots, p_n$,

$$G = ZZ^T = \langle p_i, p_j \rangle, \quad 1 \leq i, j \leq n \ , \tag{3.2.1}$$

and the equality

$$\mathcal{D}_{ij} = \langle p_i, p_i \rangle - 2\langle p_i, p_j \rangle + \langle p_j, p_j \rangle, \quad 0 \leq i, j \leq n \ , \tag{3.2.2}$$

establishes a linear equivalence between the set of $n \times n$ Gram matrices and the augmented square-distance $(n+1) \times (n+1)$ matrices of distinct landmark points. On the other hand, Gram matrices of the form ZZ^T, where Z is an $n \times d$ matrix of rank d are characterized as $n \times n$ positive semidefinite matrices of rank d. For a detailed discussion of the relation between positive semidefinite matrices, Gram matrices, and square-distance matrices, we refer the reader to Section 6.2.1 of [31]. The space of these matrices, called the positive semidefinite cone $\mathcal{S}^+(d, n)$, is a not a vector space and is mostly studied when endowed with a Riemannian metric. In the next section, we will briefly review some basics of the Riemannian geometry of the manifolds of interest, then express the Riemannian geometry of the space of Gram matrices (*i.e.*, positive semi-definite matrices of fixed rank).

3.3 Riemannian geometry of the space of Gram matrices

3.3.1 Mathematical preliminaries

A manifold is a topological space that is locally homeomorphic to the *dim*-dimensional Euclidean space $\mathbb{R}^{dim}$, where *dim* is the dimensionality of the manifold. A differentiable manifold is a topological manifold equipped with a differential structure that allows differential calculus on the manifold. The tangent space at a given point on a differentiable manifold is a vector space that consists of the tangent vectors of all possible curves passing through the point. A Riemannian manifold is a differentiable manifold equipped with a smoothly varying inner product on each tangent space. The family of inner products on

all tangent spaces is known as the Riemannian metric of the manifold [62]. By definng a Riemannian metric on the manifold, one can exploit the vector space structure of the tangent space to define various geometric notions on the manifold. As mentioned in Section 2.4.4 of the previous chapter, one can compute the *geodesic distance* between two points on the manifold which is the length of the shortest curve (*i.e.*, *geodesic*) connecting these two points. Two other important operations in Riemannian manifolds are the logarithm (log) and exponential (exp) maps. To illustrate these two operations, let us consider two points X and Y lying on a Riemannian manifold $\mathcal{M}$. Let $T_X\mathcal{M}$ be the tangent space attached to the point X as depicted in Fig. 3.2. The logarithm map $\log_X(Y)$ of the point Y to the tangent space $T_X(\mathcal{M})$ attached to X results in a vector V in $T_X(\mathcal{M})$. This vector summarizes the path that should be taken in $\mathcal{M}$ to connect X and Y. In contrast, the exponential map $\exp_X(V)$ maps back the vector V to the manifold $\mathcal{M}$ resulting in a curve $\gamma(t)$ in $\mathcal{M}$ connecting X and Y. It is important to note that the computation of these operations depends on the nature of the manifold and the defined Riemannian metric.

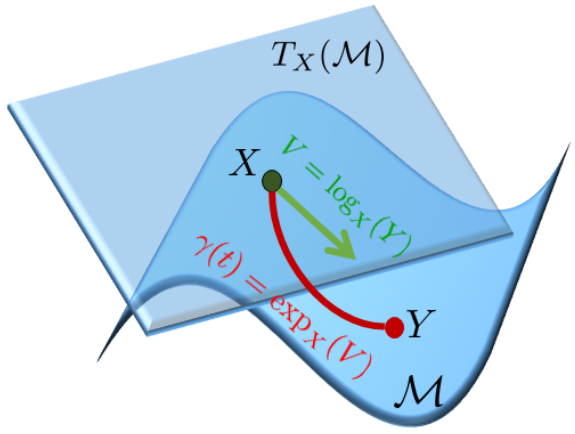

FIGURE 3.2 – Logarithm and exponential maps on Riemannian manifolds

Conveniently for us, the Riemannian geometry of the space of positive semidefinite matrices of fixed rank (*i.e.*, Gram matrices) was studied in [19, 43, 85, 120]. To have a better understanding of the geometry of this space, we first define two manifolds that are

extensively used in Computer Vision namely, the Grassmann manifold and the Riemannian manifold of positive definite matrices.

3.3.1.1 Grassmann manifold

A Grassmann manifold $\mathcal{G}(d,n)$ is the set of the d-dimensional subspaces of $\mathbb{R}^n$, where $n > d$. A subspace $\mathcal{U}$ of $\mathcal{G}(d,n)$ is represented by an $n \times d$ matrix U, whose columns store an orthonormal basis of this subspace. Thus, U is said to span $\mathcal{U}$, and $\mathcal{U}$ is said to be the column space (or span) of U, and we write $\mathcal{U} = span(U)$. Indeed, the set of $n \times d$ matrices with orthonormal columns forms a manifold known as the Stiefel manifold $\mathcal{V}_{d,n}$. Points on $\mathcal{G}(d,n)$ are equivalence classes of $n \times d$ matrices with orthonormal columns (*i.e.*, points on $\mathcal{V}_{d,n}$), where two matrices are equivalent if their columns span the same d-dimensional subspace. The geometry of the Grassmannian $\mathcal{G}(d,n)$ is then easily described by the map

$$span : \mathcal{V}_{d,n} \rightarrow \mathcal{G}(d,n) \ , \tag{3.3.1}$$

that sends an $n \times d$ matrix with orthonormal columns U to their span $span(U)$. Given two subspaces $\mathcal{U}_1 = span(U_1)$ and $\mathcal{U}_2 = span(U_2) \in \mathcal{G}(d,n)$, the geodesic curve connecting them is

$$span(U(t)) = span(U_1 \cos(\Theta t) + M \sin(\Theta t)) \ , \tag{3.3.2}$$

where Θ is a $d \times d$ diagonal matrix formed by the d *principal angles* between $\mathcal{U}_1$ and $\mathcal{U}_2$, while the matrix M is given by $M = (I_n - U_1 U_1^T) U_2 F$, with F being the pseudo-inverse of Θ. The Riemannian geodesic distance between $\mathcal{U}_1$ and $\mathcal{U}_2$ is given by

$$d^2_{\mathcal{G}}(\mathcal{U}_1, \mathcal{U}_2) = \|\Theta\|^2_F \ . \tag{3.3.3}$$

3.3.1.2 Riemannian manifold of positive definite matrices

It is known to be the positive cone in $\mathbb{R}^d$, and has been extensively used to study covariance matrices [116, 97, 16]. A symmetric $d \times d$ matrix R is said to be positive definite

if and only if $v^{\mathrm{T}}Rv > 0$ for every non-zero vector $v \in \mathbb{R}^d$. $\mathcal{P}_d$ is mostly studied when endowed with a Riemannian metric, thus forming a Riemannian manifold. A number of metrics have been proposed for $\mathcal{P}_d$, the most popular ones being the Affine-Invariant Riemannian Metric (AIRM) and the log-Euclidean Riemannian metric (LERM) [7]. In this study, we only consider the AIRM for its robustness [117].

With this metric, the geodesic curve connecting two SPD matrices R_1 and R_2 in $\mathcal{P}_d$ is

$$R(t) = R_1^{1/2} \exp(t \log(R_1^{-1/2} R_2 R_1^{-1/2})) R_1^{1/2} \ , \tag{3.3.4}$$

where $log(.)$ and $exp(.)$ are the matrix logarithm and exponential, respectively. The Riemannian distance between R_1 and R_2 is given by

$$d^2_{P_d}(R_1, R_2) = \| \log (R_1^{-1/2} R_2 R_1^{-1/2}) \|^2_{\mathrm{F}} \ , \tag{3.3.5}$$

where $\|.\|_F$ denotes the Frobenius matrix norm.

For more details about the geometry of the Grassmannian $\mathcal{G}(d, n)$ and the positive definite cone $\mathcal{P}_d$, readers are referred to [2, 12, 19, 91].

3.3.2 Riemannian manifold of positive semi-definite matrices of fixed rank

Given an $n \times d$ matrix Z of rank d, its polar decomposition $Z = UR$ with $R = (Z^T Z)^{1/2}$ allows us to write the Gram matrix ZZ^T as UR^2U^T. Since the columns of the matrix U are orthonormal, this decomposition defines a map

$$\begin{aligned} \Pi : & \mathcal{V}_{d,n} \times \mathcal{P}_d \to \mathcal{S}^+(d, n) \\ & (U, R^2) \mapsto U R^2 U^T \ , \end{aligned}$$

from the product of the Stiefel manifold $\mathcal{V}_{d,n}$ and the cone of $d \times d$ positive definite matrices $\mathcal{P}_d$ to the manifold $\mathcal{S}^+(d, n)$ of $n \times n$ positive semidefinite matrices of rank d. The map Π

defines a principal fiber bundle over $\mathcal{S}^+(d,n)$ with fibers

$$\Pi^{-1}(UR^2U^T) = \{(UO, O^TR^2O) : O \in \mathcal{O}(d)\} \ ,$$

where $\mathcal{O}(d)$ is the group of $d \times d$ orthogonal matrices. Bonnabel and Sepulchre [19] used this map and the geometry of the *structure space* $\mathcal{V}_{d,n} \times \mathcal{P}_d$ to introduce a Riemannian metric on $\mathcal{S}^+(d,n)$ and study its geometry.

3.3.2.1 Tangent space and Riemannian metric

The tangent space $T_{(U,R^2)}(\mathcal{V}_{d,n} \times \mathcal{P}_d)$ consists of pairs (M,N), where M is a $n \times d$ matrix satisfying $M^TU + U^TM = 0$ and N is any $d \times d$ symmetric matrix. Bonnabel and Sepulchre defined a *connection* (see [71, p. 63]) on the principal bundle $\Pi : \mathcal{V}_{d,n} \times \mathcal{P}_d \to \mathcal{S}^+(d,n)$ by setting the horizontal subspace $\mathcal{H}_{(U,R^2)}$ at the point (U,R^2) to be the space of tangent vectors (M,N) such that $M^TU = 0$ and N is an arbitrary $d \times d$ symmetric matrix. They also defined an inner product on $\mathcal{H}_{(U,R^2)}$: given two tangent vectors $A = (M_1,N_1)$ and $B = (M_2,N_2)$ on $\mathcal{H}_{(U,R^2)}$, set

$$\langle (A,B) \rangle_{\mathcal{H}_{U,R^2}} = \mathrm{tr}(M_1^TM_2) + k \ \mathrm{tr}(N_1R^{-2}N_2R^{-2}) \ , \tag{3.3.6}$$

where $k > 0$ is a real parameter.

It is easily checked that the action of the group of $d \times d$ orthogonal matrices on the fiber $\Pi^{-1}(UR^2U^T)$ sends horizontals to horizontals isometrically. It follows that the inner product on $T_{UR^2U^T}\mathcal{S}^+(d,n)$ induced from that of $\mathcal{H}_{(U,R^2)}$ via the linear isomorphism $D\Pi$ is independent of the choice of point (U,R^2) projecting onto UR^2U^T. This procedure defines a Riemannian metric on $\mathcal{S}^+(d,n)$ for which the natural projection

$$\rho : \mathcal{S}^+(d,n) \to \mathcal{G}(d,n)$$
$$G \mapsto range(G) \ ,$$

is a Riemannian submersion. This allows us to relate the geometry of $\mathcal{S}^+(d,n)$ with that of the Grassmannian $\mathcal{G}(d,n)$.

3.3.2.2 Pseudo-geodesics and closeness in $\mathcal{S}^+(d,n)$

Bonnabel and Sepulchre [19] defined the *pseudo-geodesic* connecting two matrices $G_1 = U_1 R_1^2 U_1^T$ and $G_2 = U_2 R_2^2 U_2^T$ in $\mathcal{S}^+(d,n)$ as the curve

$$\mathcal{C}_{G_1 \to G_2}(t) = U(t) R^2(t) U^T(t), \forall t \in [0,1] \,, \tag{3.3.7}$$

where $R^2(t) = R_1 \exp(t \log R_1^{-1} R_2^2 R_1^{-1}) R_1$ is a geodesic in $\mathcal{P}_d$ connecting R_1^2 and R_2^2, and $U(t)$ is the geodesic in $\mathcal{G}(d,n)$ given by Eq. (3.3.2). They also defined the *closeness* between G_1 and G_2, $d_{\mathcal{S}^+}(G_1, G_2)$, as the square of the length of this curve:

$$d_{\mathcal{S}^+}(G_1, G_2) = d^2_{\mathcal{G}}(\mathcal{U}_1, \mathcal{U}_2) + k d^2_{\mathcal{P}_d}(R_1^2, R_2^2) = \|\Theta\|_F^2 + k \| \log R_1^{-1} R_2^2 R_1^{-1} \|_F^2 \,, \tag{3.3.8}$$

where $\mathcal{U}_i$ $(i = 1, 2)$ is the *span* of U_i and Θ is a $d \times d$ diagonal matrix formed by the principal angles between $\mathcal{U}_1$ and $\mathcal{U}_2$.

The closeness $d_{\mathcal{S}^+}$ consists of two independent contributions: the square of the distance $d_{\mathcal{G}}(span(U_1), span(U_2))$ between the two associated subspaces, and the square of the distance $d_{\mathcal{P}_d}(R_1^2, R_2^2)$ on the positive cone $\mathcal{P}_d$ (Fig. 3.3). Note that $\mathcal{C}_{G_1 \to G_2}$ is not necessarily a geodesic and therefore, the closeness $d_{\mathcal{S}^+}$ is not a true Riemannian distance.

3.3.3 Affine-invariant and spatial covariance information of Gram matrices

An alternative affine shape representation, considered in [12] and [111], associates to each configuration Z the d-dimensional subspace $span(Z)$ spanned by its columns. This representation, which exploits the geometry of the Grassmann manifold $\mathcal{G}(d,n)$ of d-dimensional subspaces in $\mathbb{R}^n$ is invariant under *all* invertible linear transformations. By fully encoding the set of all mutual distances between landmark points, the proposed Euclidean shape representation supplements the affine shape representation with the knowledge of the $d \times d$ positive definite matrix R^2 that lie on $\mathcal{P}_d$.

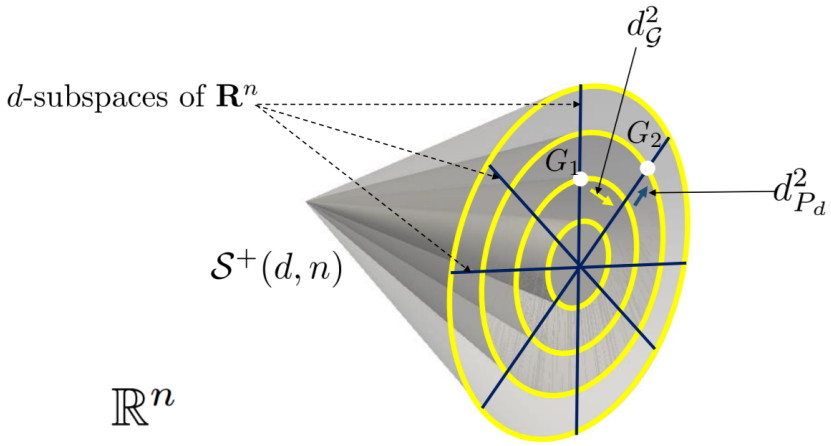

FIGURE 3.3 – A pictorial representation of the positive semidefinite cone $\mathcal{S}^+(d,n)$. Viewing matrices G_1 and G_2 as ellipsoids in $\mathbb{R}^n$; their closeness consists of two contributions: $d^2_{\mathcal{G}}$ (squared Grassmann distance) and $d^2_{\mathcal{P}_d}$ (squared Riemannian distance in $\mathcal{P}_d$).

From the viewpoint of the landmark configurations Z_1 and Z_2, with $G_1 = Z_1 Z_1^T$ and $G_2 = Z_2 Z_2^T$, the closeness $d_{\mathcal{S}^+}$ encodes the distances measured between the affine shapes $span(Z_1)$ and $span(Z_2)$ in $\mathcal{G}(d,n)$ and between their spatial covariances in $\mathcal{P}_d$. Indeed, the spatial covariance of Z_i $(i = 1, 2)$ is the $d \times d$ symmetric positive definite matrix

$$C = \frac{Z_i^T Z_i}{n-1} = \frac{(U_i R_i)^T (U_i R_i)}{n-1} = \frac{R_i^2}{n-1} \,. \tag{3.3.9}$$

The weight parameter k controls the relative weight of these two contributions. Note that for $k = 0$ the distance on $\mathcal{S}^+(d,n)$ collapses to the distance on $\mathcal{G}(d,n)$. Nevertheless, the authors in [19] recommended choosing small values for this parameter. The experiments performed and reported in Section 3.6 are in general accordance with this recommendation.

3.4 Gram matrix trajectories for temporal modeling of landmark sequences

We are able to compare static landmark configurations based on their Gramian representation G, the induced space, and closeness introduced in the previous Section. We need a natural and effective extension to study their temporal evolution. Following [13, 111, 123], we defined curves $\beta_G : I \rightarrow \mathcal{S}^+(d,n)$ (I denotes the time domain, *e.g.,* $[0,1]$) to model the spatio-temporal evolution of elements on $\mathcal{S}^+(d,n)$. Given a sequence of landmark configurations $\{Z_0,\ldots,Z_\tau\}$ represented by their corresponding Gram matrices $\{G_0,\ldots,G_\tau\}$ in $\mathcal{S}^+(d,n)$, the corresponding curve is the trajectory of the point $\beta_G(t)$ on $\mathcal{S}^+(d,n)$, when t ranges in $[0,1]$. These curves are obtained by connecting all successive Gramian representations of shapes G_i and G_{i+1}, $0 \leq i \leq \tau-1$, by pseudo-geodesics in $\mathcal{S}^+(d,n)$. Algorithm 1 summarizes the steps to build trajectories in $\mathcal{S}^+(d,n)$ for temporal modeling of landmark sequences.

Algorithm 1: Computing trajectory $\beta_G(t)$ in $\mathcal{S}^+(d,n)$ of a sequence of landmarks

input : A sequence of centered landmark configurations $\{Z_0,\cdots,Z_\tau\}$, where $Z_{0\leq i\leq\tau}$ is an $(n\times d)$ matrix ($d=2$ or $d=3$) formed by the coordinates $p_1=(x_1,y_1),\cdots,p_n=(x_n,y_n)$ or $p_1=(x_1,y_1,z_1),\cdots,p_n=(x_n,y_n,z_n)$.

output: Trajectory $\beta_G(t)_{0\leq t\leq\tau}$ and pseudo-geodesics $\mathcal{C}_{\beta_G(t)\rightarrow\beta_G(t+1)}$ in $\mathcal{S}^+(d,n)$

/ Compute the Gram matrices of centered landmarks */*

for $i \leftarrow 0$ **to** τ **do**

  $G_i \longleftarrow Z_iZ_i^T = \langle p_l, p_k\rangle,\ \ 1\leq l,k\leq n$

  / Compute the Polar decomposition*[1] *of $Z_i = U_iR_i$ */*

  $G_i \longleftarrow U_iR_i^2U_i^T$

/ Compute the pseudo-geodesic paths between successive Gram matrices */*

$\beta_G(0) \longleftarrow G_0$

for $t \leftarrow 0$ **to** $\tau-1$ **do**

  $\mathcal{C}_{\beta_G(t)\rightarrow\beta_G(t+1)} \longleftarrow \mathcal{C}_{G_t\rightarrow G_{t+1}}$ given by Eq. (3.3.7) connecting G_t and G_{t+1} in $\mathcal{S}^+(d,n)$

  $\beta_G(t+1) \longleftarrow G_{t+1}$

return trajectory $\beta_G(t)_{0\leq t\leq\tau}$ and pseudo-geodesics $\mathcal{C}_{\beta_G(t)\rightarrow\beta_G(t+1)}$ in $\mathcal{S}^+(d,n)$

1. To compute the polar decomposition, we used the SVD based implementation proposed in [54].

3.4.1 Rate-invariant comparison of Gram matrix trajectories

A relevant issue to our classification problems is – how to compare trajectories while being invariant to rates of execution? One can formulate the problem of temporal misalignment as comparing trajectories when parameterized differently. The parameterization variability makes the distance between trajectories distorted. This issue was first highlighted by Veeraraghavan *et al.* [121] who showed that different rates of execution of the same activity can greatly decrease recognition performance if ignored. Veeraraghan *et al.* [121] and Abdelkader *et al.* [1] used the Dynamic Time Warping (DTW) for temporal alignment before comparing trajectories of shapes of planar curves that represent silhouettes in videos. Following the above-mentioned state-of-the-art solutions, we adopt here a DTW solution to temporally align our trajectories. More formally, given m trajectories $\{\beta_G^1, \beta_G^2, \ldots, \beta_G^m\}$ on $\mathcal{S}^+(d,n)$, we are interested in finding functions γ_i such that the $\beta_G^i(\gamma_i(t))$ are matched optimally for all $t \in [0,1]$. In other words, two curves $\beta_G^1(t)$ and $\beta_G^2(t)$ represent the same trajectory if their images are the same. This happens if, and only if, $\beta_G^2 = \beta_G^1 \circ \gamma$, where γ is a re-parameterization of the interval $[0,1]$. The problem of temporal alignment is turned to find an optimal warping function $\gamma^\star$ according to,

$$\gamma^\star = \arg\min_{\gamma \in \Gamma} \int_0^1 d_{\mathcal{S}^+}(\beta_G^1(t), \beta_G^2(\gamma(t)))\, \mathrm{d}t \;, \tag{3.4.1}$$

where Γ denotes the set of all monotonically-increasing functions $\gamma : [0,1] \to [0,1]$. The most commonly used method to solve such optimization problem is DTW. Note that accommodation of the DTW algorithm to the manifold-value sequences can be achieved with respect to an appropriate metric defined on the underlying manifold $\mathcal{S}^+(d,n)$. Having the optimal re-parametrization function $\gamma^\star$, one can define a (dis-)similarity measure between two trajectories allowing a rate-invariant comparison:

$$d_{DTW}(\beta_G^1, \beta_G^2) = \int_0^1 d_{\mathcal{S}^+}(\beta_G^1(t), \beta_G^2(\gamma^\star(t)))\, \mathrm{d}t \;. \tag{3.4.2}$$

From now, we shall use $d_{DTW}(.,.)$ to compare trajectories in our manifold of interest $\mathcal{S}^+(d,n)$.

3.4.2 Adaptive re-sampling

One difficulty in video analysis is to capture the most relevant frames and focus on them. In fact, it is relevant to reduce the number of frames when no motion happened, and "introduce" new frames, otherwise. Our geometric framework provides tools to do so. In fact, interpolation between successive frames could be achieved using the pseudo-geodesics defined in Eq. (3.3.7), while their length (closeness defined in Eq. (3.3.8)) expresses the magnitude of the motion. Accordingly, we have designed an adaptive re-sampling tool that is able to increase/decrease the number of samples in a fixed time interval according to their relevance with respect to the geometry of the underlying manifold $\mathcal{S}^+(d,n)$. Relevant samples are identified by a relatively low closeness $d_{\mathcal{S}^+}$ to the previous frame, while irrelevant ones correspond to a higher closeness level. Here, the down-sampling is performed by removing irrelevant shapes. In turn, the up-sampling is possible by interpolating between successive shape representations in $\mathcal{S}^+(d,n)$, using pseudo-geodesics.

More formally, given a trajectory $\beta_G(t)_{t=0,1,\ldots,\tau}$ on $\mathcal{S}^+(d,n)$ for each sample $\beta_G(t)$, we compute the closeness to the previous sample, *i.e.,* $d_{\mathcal{S}^+}(\beta_G(t),\beta_G(t-1))$: if the value is below a defined threshold ζ_1, the current sample is simply removed from the trajectory. In contrast, if the distance exceeds a second threshold ζ_2, equally spaced shape representations from the pseudo-geodesic curve connecting $\beta_G(t)$ to $\beta_G(t-1)$ are inserted in the trajectory.

3.5 Classification of Gram matrix trajectories

Our trajectory representation reduces the problem of landmark sequence classification to that of trajectory classification in $\mathcal{S}^+(d,n)$. That is, let us consider $\mathcal{T} = \{\beta_G \ : [0,1] \rightarrow \mathcal{S}^+(d,n)\}$, the set of time-parameterized trajectories of the underlying manifold. Let $\mathcal{L} =$

$\{(\beta_G^1, y^1), \ldots, (\beta_G^m, y^m)\}$ be the training set with class labels, where $\beta_G^i \in \mathcal{T}$ and $y^i \in \mathcal{Y}$, such that $y^i = f(\beta_G^i)$. The goal here is to find an approximation h to f such that $h : \mathcal{T} \to \mathcal{L}$. In Euclidean spaces, any standard classifier (*e.g.,* standard SVM) may be a natural and appropriate choice to classify the trajectories. Unfortunately, this is no more suitable in our modeling, as the space $\mathcal{T}$ built from $\mathcal{S}^+(d, n)$ is non-linear. As mentioned and discussed in the previous chapter, a function that divides the manifold is rather a complicated notion compared with the Euclidean space. To overcome this issue, we adopt two classification schemes based on the (dis-)similarity measure d_{DTW} that uses the geometry-aware closeness $d_{\mathcal{S}^+}$ namely, k-Nearest Neighbor and Pairwise proximity function SVM classifiers.

3.5.1 Pairwise proximity function SVM

Inspired by a recent work of [9] for action recognition, we adopted the *pairwise proximity function SVM* (ppfSVM) [50, 51]. The ppfSVM requires the definition of a (dis-)similarity measure to compare samples. In our case, it is natural to consider the d_{DTW} defined in Eq. (3.4.2) for such a comparison. This strategy involves the construction of inputs such that each trajectory is represented by its (dis-)similarity to all the trajectories, with respect to d_{DTW}, in the dataset and then apply a conventional SVM to this transformed data [51]. The ppfSVM is related to the arbitrary kernel-SVM without restrictions on the kernel function [50].

Given m trajectories $\{\beta_G^1, \beta_G^2, \ldots, \beta_G^m\}$ in $\mathcal{T}$, following [9], a proximity function $\mathcal{P}_{\mathcal{T}} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_+$ between two trajectories $\beta_G^1, \beta_G^2 \in \mathcal{T}$ is defined as,

$$\mathcal{P}_{\mathcal{T}}(\beta_G^1, \beta_G^2) = d_{DTW}(\beta_G^1, \beta_G^2) \; . \tag{3.5.1}$$

According to [50], there are no restrictions on the function $\mathcal{P}_{\mathcal{T}}$. For an input trajectory $\beta_G \in \mathcal{T}$, the mapping $\phi(\beta_G)$ is given by,

$$\phi(\beta_G) = [\mathcal{P}_{\mathcal{T}}(\beta_G, \beta_G^1), \ldots, \mathcal{P}_{\mathcal{T}}(\beta_G, \beta_G^m)]^T \; . \tag{3.5.2}$$

The obtained vector $\phi(\beta_G) \in \mathbb{R}^m$ is used to represent a sample trajectory $\beta_G \in \mathcal{T}$. Hence, the set of trajectories can be represented by a $m \times m$ matrix P, where $P(i,j) = \mathcal{P}_{\mathcal{T}}(\beta_G^i, \beta_G^j)$, $i,j \in \{1,\ldots,m\}$. Finally, a linear SVM is applied to this data representation. Further details on ppfSVM can be found in [9, 50, 51]. In Algorithm 2, we provide a pseudo-code for the proposed trajectory classification in $\mathcal{S}^+(d,n)$.

Algorithm 2: Classification of trajectories in $\mathcal{S}^+(d,n)$

input : m training trajectories in $\mathcal{S}^+(d,n)$ with their corresponding labels $\{(\beta_G^1, y^1), \ldots, (\beta_G^m, y^m)\}$
One testing trajectory β_G^{test} in $\mathcal{S}^+(d,n)$
output: Predicted class y^{test} of β_G^{test}
/ Model training */*
for $i \leftarrow 1$ **to** m **do**
  for $j \leftarrow 1$ **to** m **do**
    $P(i,j) = \mathcal{P}_{\mathcal{T}}(\beta_G^i, \beta_G^j)$ w.r.t Eq. (3.5.1)
Training a linear SVM on the data representation P
/ Testing phase */*
$\phi(\beta_G^{test}) = [\mathcal{P}_{\mathcal{T}}(\beta_G^{test}, \beta_G^1), \ldots, \mathcal{P}_{\mathcal{T}}(\beta_G^{test}, \beta_G^m)]^T$
$y^{test} \longleftarrow$ Linear SVM using the feature vector $\phi(\beta_G^{test})$
return Predicted class y^{test}

The proposed ppfSVM classification of trajectories on $\mathcal{S}^+(d,n)$ aims to learn a proximity model of the data, which makes the computation of a pairwise distance function using the DTW (dis-)similarity measure on all the trajectories of the dataset quite necessary. For more efficiency, one can consider faster algorithms for trajectories alignment such us [96, 28].

3.5.2 K-Nearest neighbor

As a baseline classifier, we use used k-nearest neighbor solution, where for each test trajectory (sequence), we computed the k-nearest trajectories (sequences) from the training set using the same (dis-)similarity measure d_{DTW} defined in Eq. (3.4.2). The test sequence is then classified according to a majority voting of its neighbors, (*i.e.*, it is assigned to the class that is most common among its k-nearest neighbors).

3.6 Experimental evaluation

To validate the proposed framework, we conducted extensive experiments on three human behavior understanding applications. These scenarios show the potential of the proposed solution when landmarks capture different information on different data. First, we addressed the problem of activity recognition from depth sensors such as the Microsoft Kinect. In this case, 3D landmarks correspond to the joints of the body skeleton, as extracted from RGB-Depth frames. The number of joints per skeleton varies between 15 and 20, and their position is generally noisy. Next, we addressed the new emerging problem of finding relationships between body movement and emotions using 3D skeletal data. Here, landmarks correspond to physical markers placed on the body and tracked with high temporal rate and good estimation of the 3D position by a Motion Capture (MoCap) system. Finally, we evaluated our framework on the problem of facial expression recognition using landmarks of the face. In this case, 49 face landmarks are extracted in 2D with high accuracy using a state-of-the-art face landmark detector.

3.6.1 3D action recognition

Action recognition has been performed on 3D skeleton data as provided by a Kinect camera in different datasets. In this case, landmarks correspond to the estimated position of 3D joints of the skeleton (d=3). With this assumption, skeletons are represented by $n \times n$ Gram matrices of rank 3 lying on $\mathcal{S}^+(3, n)$, and skeletal sequences are seen as trajectories on this manifold.

As discussed in Section 3.2, the information given by the Gram matrix of the skeleton is linearly equivalent to that of the pairwise distances between different joints. Thus, considering only some specific subparts of the skeletons can be more accurate for some actions. For instance, it is more discriminative to consider only the pairwise distances between the joints of left and right arms for actions that involve principally the motion

of arms, (*e.g.*, *wave hands*, *throw*). Accordingly, we divided the skeletons into three body parts, *i.e.*, left/right arms, left/right legs and torso, while keeping a coarse information given by all the joints of the skeleton. In Fig. 3.4, we show an example of the proposed Kinect skeleton decomposition into three body parts. For an efficient use of the information given by the different body parts, we propose a late fusion of four ppf-SVM classifiers that consists of: (1) training all the body part classifiers separately; (2) merging the contributions of the four body part classifiers. This is done by multiplying the probabilities $s_{i,j}$, output of the SVM for each class j, where $i \in \{1, 2, 3, 4\}$ denotes the body part. The class $\mathcal{C}$ of each test sample is determined by

$$\mathcal{C} = \arg\max_{j} \prod_{i=1}^{4} s_{i,j}, \ j = 1, \ldots, n_{\mathcal{C}} \ , \tag{3.6.1}$$

where $n_{\mathcal{C}}$ is the number of classes.

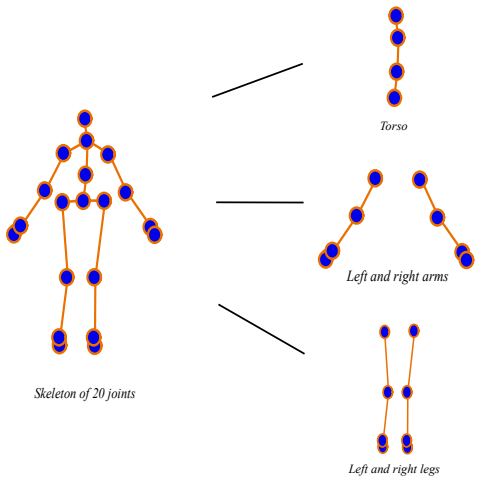

FIGURE 3.4 – Decomposition of the Kinect skeleton into three body parts.

3.6.1.1 Datasets

We performed experiments on four publicly available datasets showing different challenges. All these datasets have been collected with a Microsoft Kinect sensor.

UT-Kinect dataset [136] – It contains 10 actions performed by 10 different subjects. Each subject performed each action twice resulting in 199 valid action sequences. The 3D locations of 20 joints are provided with the dataset.

Florence3D dataset [103] – It contains 9 actions performed two or three times by 10 different subjects. Skeleton comprises 15 joints. This is a challenging dataset due to variations in the view-point and large intra-class variations.

SYSU-3D dataset [55] – It contains 480 sequences. In this dataset, 12 different activities focusing on interactions with objects were performed by 40 persons. The 3D coordinates of 20 joints are provided in this dataset. The SYSU-3D dataset is very challenging since the motion patterns are highly similar among different activities.

SBU Interaction dataset [138] – This dataset includes 282 skeleton sequences of eight types of two-persons interacting with each other, including *approaching*, *departing*, *pushing*, *kicking*, *punching*, *exchanging objects*, *hugging*, and *shaking hands*. In most interactions, one subject is acting, while the other subject is reacting.

3.6.1.2 Experimental settings and parameters

For all the datasets, we used only the provided skeletons. The adaptive re-sampling of trajectories discussed in Section 3.4.2 has been not applied on these data. The motivation is that this operation tries to capture small shape deformations of the landmarks and this can amplify the noise of skeleton joints. For the SBU dataset, where two skeletons of two interacting persons are given in each frame, we considered all the joints of the two skeletons. In this case, a unique Gram matrix is computed for the two skeletons modeling the interaction between them. In this dataset, the decomposition into body parts is performed only for the acting person since the other person is reacting in a coarse manner.

As discussed in Section 3.3.3, our body movement representation involves a parameter k that controls the contribution of two information: the affine shape of the skeleton at time t,

and its spatial covariance. The affine shape information is given by the Grassmann manifold $\mathcal{G}(3, n)$, while the spatial covariance is given by the SPD manifold $\mathcal{P}_3$. We recall that for $k = 0$, the skeletons are considered as trajectories on the Grassmann manifold $\mathcal{G}(3, n)$. For each dataset, we performed a cross-validation grid search, $k \in [0, 3]$ with a step of 0.1, to find an optimal value k^*. In the case of skeleton decomposition into body parts, a different parameter k is used for computing the distance of each body part, (*i.e.*, one parameter each for arms, legs, and torso, and one parameter for the whole skeleton). Each parameter k is evaluated separately by a cross-validation grid search in the classifier of the relative body part.

To allow a fair comparison, we adopted the most common experimental settings in literature. For the UT-Kinect dataset, we used the *leave-one-out cross-validation* (LOOCV) protocol [136], where one sequence is used for testing and the remaining sequences are used for training. For the Florence3D dataset, a *leave-one-subject-out* (LOSO) schema is adopted following [30, 127, 141]. For the SYSU3D dataset, we followed [55] and performed a *Half-Half* cross-subject test setting, in which half of the subjects were used for training and the remaining half were used for testing. Finally, a 5-*fold* cross-validation was used for the SBU dataset. Note that the subjects considered in each split are those given by the datasets (SYSU3D and SBU). All our programs were implemented in Matlab and run on a 2.8 GHZ CPU. We used the multi-class SVM implementation of the LibSVM library [25].

3.6.1.3 Results and discussion

In Table 3.1- 3.2, we compare our approach with existing methods dealing with skeletons and/or RGB-D data. Overall, our approach achieved competitive results compared to recent state-of-the-art approaches.

On the UT-Kinect dataset, we obtained an average accuracy of 96.48%, when considering the full skeletal shape. Using a late fusion of classifiers based on the body parts, as

TABLE 3.1 – Overall accuracy (%) on the UT-Kinect and Florence3D datasets. Here, $^{(D)}$: depth; $^{(C)}$: color (or RGB); $^{(G)}$: geometry (or skeleton); *: Deep Learning based approach; last row: ours

	UT-Kinect		**Florence3D**	
Method	**Prot.**	**Acc (%)**	**Prot.**	**Acc (%)**
$^{(G+D)}$ 3D^{2}CNN [80]*	LOSO	95.5	–	–
$^{(G)}$ LARP [123]	5-fold	97.08	5-fold	90.88
$^{(G)}$ Gram Hankel [141]	LOOCV	**100**	–	–
$^{(G)}$ Motion trajectories [30]	LOOCV	91.5	LOSO	87.04
$^{(G)}$ Elastic func. coding [6]	5-fold	94.87	5-fold	89.67
$^{(G)}$ Mining key poses [127]	LOOCV	93.47	LOSO	**92.25**
$^{(G)}$ NBNN+parts+time [103]	–	–	LOSO	82
$^{(G)}$ LSTM-trust gate [77]*	LOOCV	97.0	–	–
$^{(G)}$ JL-distance LSTM[140]*	5-fold	95.96	–	–
Traj. on $\mathcal{G}(3,n)$ (full body)	LOOCV	92.46	LOSO	75 ± 5.22
Traj. on $\mathcal{G}(3,n)$ - BP Fusion	LOOCV	96.48	LOSO	76.4 ± 5.37
Traj. on $\mathcal{S}^+(3,n)$ (full body)	LOOCV	**96.48**	LOSO	**88.07± 4.8**
Traj. on $\mathcal{S}^+(3,n)$ - BP Fusion	LOOCV	**98.49**	LOSO	**88.85 ± 4.6**

described in Section 3.6.1, the performance increased to 98.49% outperforming [77, 30, 127]. The highest average accuracy for this dataset was reported in [141] (100%), where Gram matrices were used for skeletal sequence representation, but in a completely different context. Specifically, the authors of [141] built a Gram matrix from the Hankel matrix of an Auto-Regressive (AR) model that represented the dynamics of the skeletal sequences. The used metric for the comparison of Gram matrices is also different than ours as they used metrics in the positive definite cone by regularizing their ranks, *i.e.*, making them full-rank.

On the SBU dataset, the fusion of body parts achieved the highest accuracy reaching 93.7%. We observed that all the interactions present in this dataset are well recognized, *e.g.*, *hugging* (100%), *approaching* (97.5%), etc., except *pushing* (74.7%), which has been mainly confused with a very similar interaction, *i.e.*, *punching*. Here, our approach is ranked second after [140], where an average accuracy of 99.02% is reported. In that work, the authors compute a large number of joint-line distances per frame making their approach time consuming.

On the SYSU3D dataset, our approach achieved the best result compared to skeleton based approaches. We report an average accuracy of 80.22% with a standard deviation of

TABLE 3.2 – Overall accuracy (%) on the SBU interaction, and SYSU-3D datasets. Here, $^{(D)}$: depth; $^{(C)}$: color (or RGB); $^{(G)}$: geometry (or skeleton); *: Deep Learning based approach; last row: ours

	SBU Interaction		**SYSU-3D**	
Method	**Prot.**	**Acc (%)**	**Prot.**	**Acc (%)**
$^{(G+D+C)}$ Dynamic features [55]	–	–	Half-Half	**84.9 ± 2.29**
$^{(G+D+C)}$ LAFF [56]	–	–	Half-Half	80
$^{(G)}$ LAFF (SKL) [56]	–	–	Half-Half	54.2
$^{(G)}$ Dynamic skeletons [55]	–	–	Half-Half	75.5 ± 3.08
$^{(G)}$ LSTM-trust gate [77]*	5-fold	93.3	Half-Half	**76.5**
$^{(G)}$ JL-distance LSTM[140]*	5-fold	**99.02**	–	–
$^{(G)}$ Co-occurence LSTM[147]*	5-fold	90.41	–	–
$^{(G)}$ Hierarchical RNN[39]*	5-fold	80.35	–	–
$^{(G)}$ SkeletonNet[68]*	5-fold	93.47	–	–
$^{(G)}$ STA-LSTM[108]*	5-fold	91.51	–	–
Traj. on $\mathcal{G}(3,n)$ (full body)	5-fold	76.3 ± 3.26	Half-Half	73.26 ± 2.27
Traj. on $\mathcal{G}(3,n)$ - BP Fusion	5-fold	83.56 ± 4.72	Half-Half	76.61 ± 2.86
Traj. on $\mathcal{S}^{+}(3,n)$ (full body)	5-fold	**88.45 ± 2.88**	Half-Half	**76.01 ± 2.09**
Traj. on $\mathcal{S}^{+}(3,n)$ - BP Fusion	5-fold	**93.7 ± 1.59**	Half-Half	**80.22± 2.09**

2.09%, when the late fusion of body parts is used. Our approach, applied to the full skeleton, still achieved very competitive results and reached 76.01% with a standard deviation of 2.09%. Combining the skeletons with depth and color information, including the object, Hu *et al.* [55] obtained the highest performance with an average accuracy of 84.9% and a standard deviation of 2.29%.

On the Florence3D dataset, we obtained an average accuracy of 88.07%, improved by around 0.8% when involving body parts fusion. While high accuracies are reported for coarse actions, *e.g.*, *sitting down* (95%), *standing up* (100%), and *lacing* (96.2%), finer actions, *e.g.*, *reading watch* (73.9%) and *answering phone* (68.2%) are still challenging. Our results are outperformed by [127, 123], where the average accuracies are greater than 90%.

From the reported results on the four different datasets, we can observe the large superiority of the Gramian representation over the Grassmann representation. For the Florence3D and SBU datasets, we report an improvement of about 12%. For UT-Kinect and SYSU3D, the performance increased by about 3%. Note that these improvements over the Grassmannian representation are due to the additional information of the spatial covariance

given by the SPD manifold in the metric. The contribution of the spatial covariance is weighted with a parameter k. As discussed in Section 3.6.1.2, we performed a grid search cross-validation to find the optimal value k^* of this parameter. In Fig. 3.5, we report the accuracies obtained when considering the whole skeletons for different values of k. The optimal values are $k^* = 0.05$, $k^* = 0.81$, $k^* = 0.25$, and $k^* = 0.09$ for the the UT-Kinect, SBU, Florence3D, and SYSU3D datasets, respectively. These results are in concordance with the recommendation of Bonnabel and Sepulchre [19] to use relative small values of k.

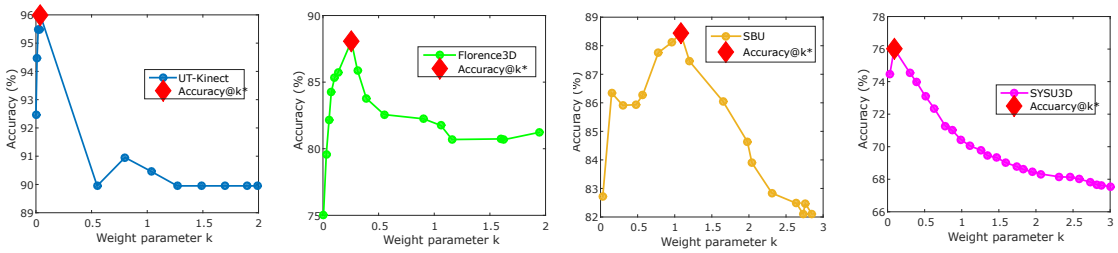

FIGURE 3.5 – Accuracy of the proposed approach when varying the weight parameter k: results for the UT-Kinect, Florence3D, SBU, and SYSU-3D datasets are reported from left to right.

Confusion matrices. In order to evaluate the effectiveness of our approach on recognizing the different actions, we report the obtained confusion matrices on the four datasets used in the experiments.

In Fig. 3.6, we show the confusion matrix for the UT-Kinect dataset. We can observe that all the actions were well recognized. The few confusions happened between *"pick up"* with *"walk"*, *"carry"* with *"walk"*, and *"clap hands"* with *"wave hands"*.

On the human interaction SBU dataset, as shown in Fig. 3.7, the highest performance was achieved for *"departing"* and *"hugging"* interactions (100%), while *"pushing"* interaction was the least recognized (74.7%). The latter was mainly confused by our approach with a similar interaction (*i.e.*, *"punching"*).

Fig. 3.8 depicts the confusions of our approach on the human-object interaction dataset SYSU3D. Unsurprisingly, *"sit chair"* and *"move chair"* were the most recognized interactions

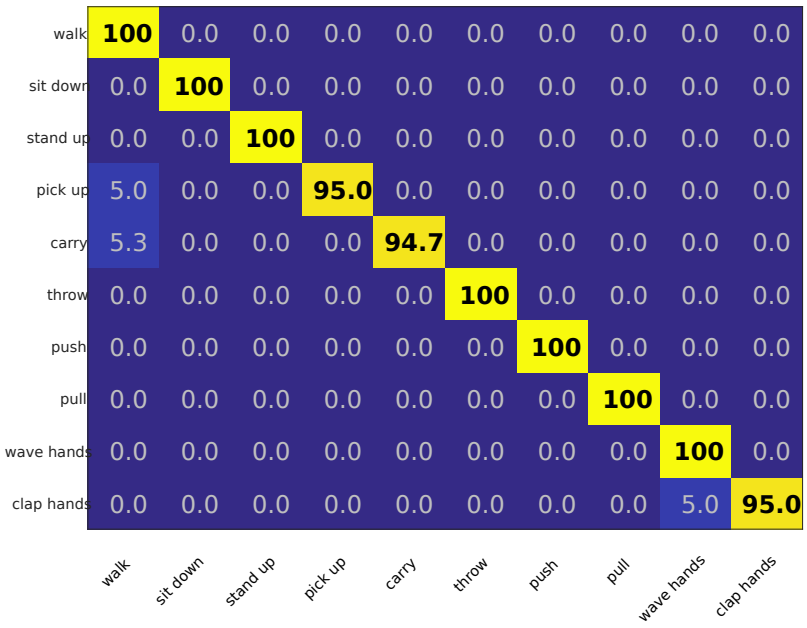

FIGURE 3.6 – Confusion matrix for the UT-Kinect dataset.

($>$ 95%). In accordance with [55], the lowest performance was achieved for *"call phone"* interaction (65.8%), which was mutually confused with *"drinking"*. These two interactions involve similar patterns (raising one arm to the head) that could be more similar with the inaccurate tracking of the skeletons. Other examples of such mutual confusions include the interactions *"take from wallet"* (70.5%) with *"play phone"* (72.8%) and *"mopping"* (74.5%) with *"sweeping"* (73.2%).

Finally, we report in Fig. 3.9 the confusion matrix for the Florence 3D dataset. Similarly to the reported results on the UT-Kinect dataset, the best performance was recorded for the *"stand up"* (100%) and *"sit down"* (95%) actions. Correspondingly to the obtained results on the SYSU3D dataset, the main confusions concerned *"drink"* (76.2%) with *"answer phone"* (68.2%). Furthermore, it is worth noting that, in this dataset, several actions are performed with the right arm by some participants, while others acted it with the left arm. This could explain the low performance achieved by our approach on distinguishing *"read watch"*, where only one arm (left or right) is raised to the chest, from *"clap hands"*, where the two arms

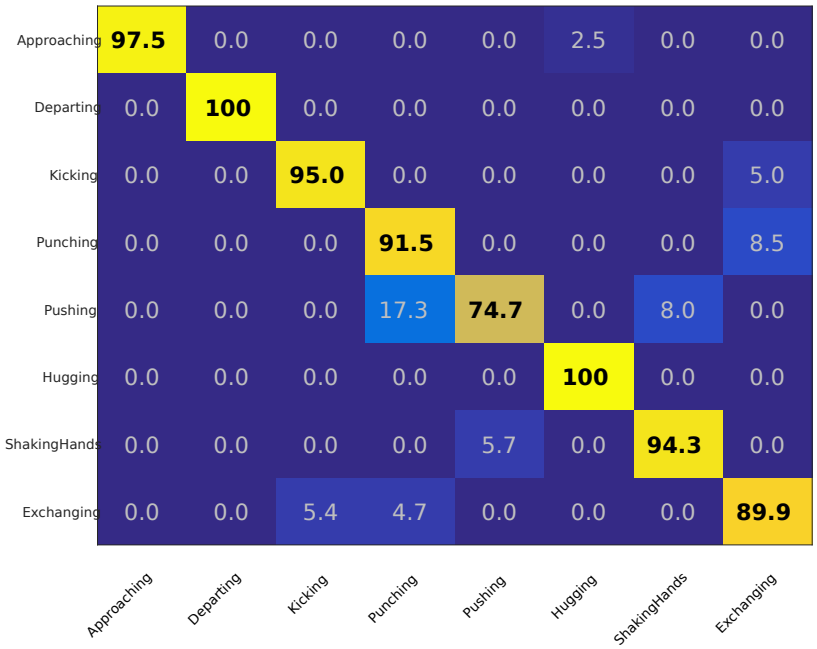

FIGURE 3.7 – Confusion matrix for the SBU dataset.

are raised to merely the same position.

Baseline experiments. In this paragraph, we discuss the effect of using the different steps in our framework and their computational complexity compared to baselines. Results of this evaluation are reported in Table 3.3. Firstly, in the top part of Table 3.3, we studied the computational cost of the proposed pipeline in the task of 3D action recognition and report running time statistics for the different steps of our approach on UT-Kinect dataset. Specifically, we provide the necessary execution time for: (1) an arbitrary trajectory construction in $\mathcal{S}^+(3,n)$ as described in Algorithm 1; (2) comparison of two arbitrary trajectories with the proposed version of DTW; (3) testing phase of an arbitrary trajectory classification with ppfSVM in $\mathcal{S}^+(3,n)$ as described in Algorithm 2.

Then, we evaluated the proposed metric with respect to other metrics used in state of the art solutions. Specifically, given two matrices G_1 and G_2 in $\mathcal{S}^+(3,n)$, we compared our results with two other possible metrics: (1) as proposed in [132, 141], we used $d_{\mathcal{P}_n}$

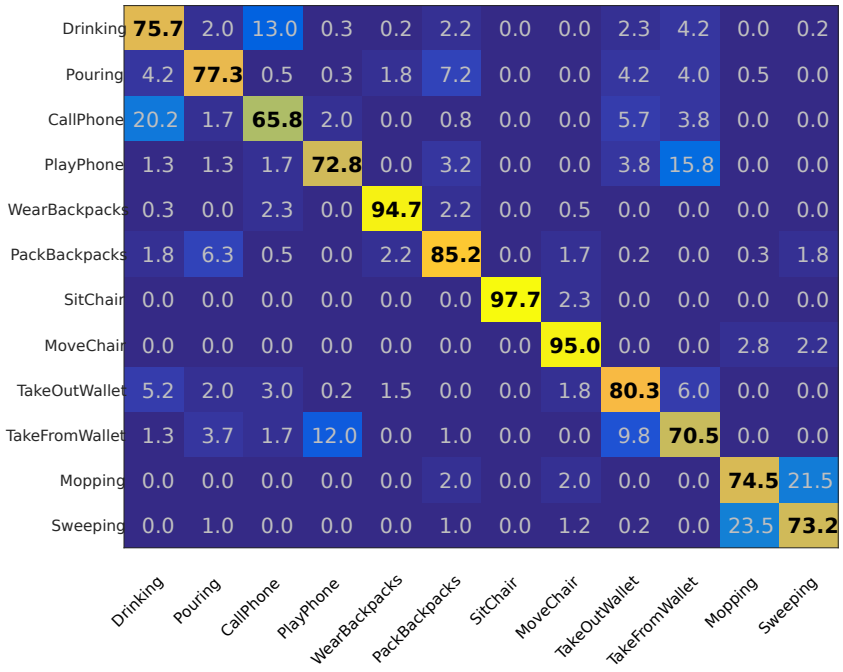

FIGURE 3.8 – Confusion matrix for the SYSU3D dataset.

that was defined in Eq. (3.3.8) to compare G_1 and G_2 by regularizing their ranks, *i.e.*, making them n full-rank, and considering them in $\mathcal{P}_n$ (the space of n-by-n positive definite matrices), $d_{\mathcal{P}_n}(G_1, G_2) = d_{\mathcal{P}_n}(G_1 + \epsilon I_n, G_2 + \epsilon I_n)$; (2) we used the Euclidean flat distance $d_{\mathcal{F}^+}(G_1, G_2) = \|G_1 - G_2\|_F$, where $\|.\|_F$ denotes the Frobenius-norm. Note that the provided execution times are relative to the comparison of two arbitrary sequences. We can observe that in Table 3.3, the closeness $d_{\mathcal{S}^+}$ between two elements of $\mathcal{S}^+(3, n)$ defined in Eq. (3.3.8) is more suitable compared to the distance $d_{\mathcal{P}_n}$ and the flat distance $d_{\mathcal{F}^+}$ defined in literature. This demonstrates the importance of considering the geometry of the manifold of interest. Another advantage of using $d_{\mathcal{S}^+}$ over $d_{\mathcal{P}_n}$ is the computational time as it involves n-by-3 and 3-by-3 matrices instead of n-by-n matrices.

To show the relevance of aligning the skeleton sequences in time before comparing them, we conducted the same experiments without using Dynamic Time Warping (DTW). In this case, the performance decreased by around 5% and 7% on UT-Kinect and SBU datasets,

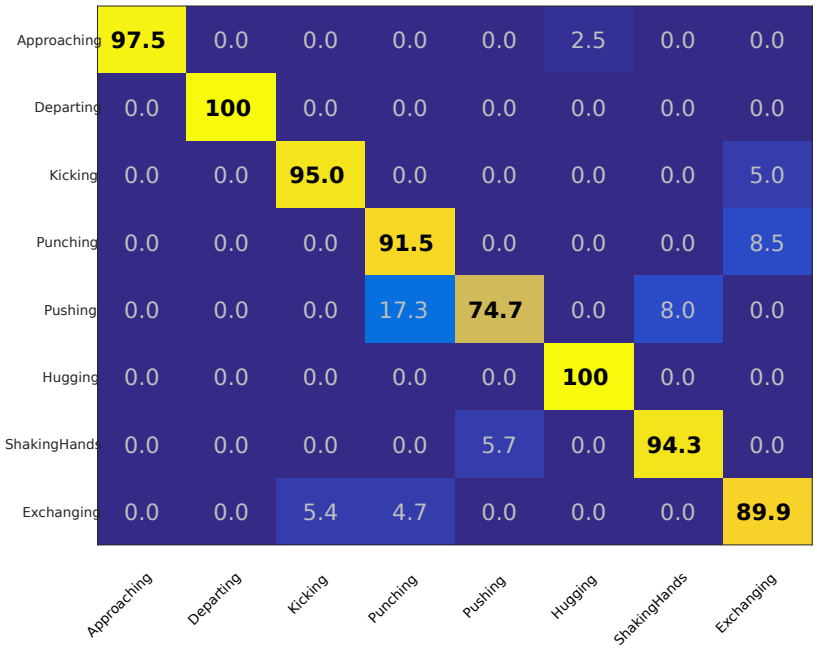

FIGURE 3.9 – Confusion matrix for the Florence dataset.

respectively. Here, the provided execution times are relative to the comparison of two arbitrary sequences on UT-Kinect dataset. Furthermore, we also compared the proposed ppfSVM classifier with a k-nearest neighbor classifier. The number of nearest neighbors k to consider for each dataset is chosen by cross-validation. Using the k-NN classifier, we obtained an average accuracy of 91.96% with $k = 5$ neighbors on UT-Kinect and 61.06% with $k = 4$ on the SBU dataset. These results are outperformed by the ppfSVM classifier.

Finally, in Table 3.3 we provide the obtained accuracies when considering the different body parts separately on all the datasets. Unsurprisingly, the highest accuracy is achieved by left and right arms in all the datasets compared to the torso and the legs, since the majority of the actions are acted using arms. One can note the considerable improvements realized by the late fusion compared to the whole skeleton in all the datasets, especially in the SBU and SYSU3D datasets, where we report improvements of about 5% and 4%, respectively.

TABLE 3.3 – Baseline experiments on the UT-Kinect, SBU, SYSU3D, and Florence3D datasets

Pipeline component	**Time** (s)
Trajectory construction in $\mathcal{S}^+(3,n)$	0.007
Comparison of trajectories in $\mathcal{S}^+(3,n)$	0.93
Classification of a trajectory in $\mathcal{S}^+(3,n)$	147.71

Distance	**UT-Kinect** (%)	**Time** (s)
Flat distance $d_{\mathcal{F}^+}$	92.96	0.06
Distance $d_{\mathcal{P}_n}$ in $\mathcal{P}_n$	94.98	1.66
Closeness $d_{\mathcal{S}^+}$	**96.48**	0.93

Temp. alignment	**UT-Kinect (%)**	**SBU (%)**	**Time** (s)
No DTW	91.46	81.36± 2.78	0.02
DTW	**96.48**	**88.45± 2.88**	0.93

Classifier	**UT-Kinect (%)**	**SBU (%)**
K-NN – $\mathcal{G}(3,n)$	86.93	42.72 ± 5.68
Ppf-SVM – $\mathcal{G}(3,n)$	92.46	76.3 ± 3.26
K-NN – $\mathcal{S}^+(3,n)$	91.96	61.06 ± 2.3
Ppf-SVM – $\mathcal{S}^+(3,n)$	**96.48**	**88.45± 2.88**

Body parts	**UT-Kinect (%)**	**SBU (%)**
Arms only	87.94	80.96 ± 5.53
Legs only	35.68	83.36 ± 2.41
Torso only	72.36	80.58 ± 2.16
Whole body	96.48	88.45 ± 2.88
Late BP Fusion	**98.49**	**93.7 ± 1.59**

Body parts	**Florence3D (%)**	**SYSU3D (%)**
Arms only	75.72 ± 8.45	73.88 ± 2.64
Legs only	42.44 ± 7.69	37.6 ± 2.10
Torso only	54.33 ± 10.62	49.36 ± 3.94
Whole body	88.07 ± 4.8	76.01± 2.09
Late BP Fusion	**88.85 ± 4.6**	**80.22 ± 2.09**

3.6.2 3D emotion recognition from body movements

Recently, the study of computational models for human emotion recognition has gained increasing attention not only for commercial applications (to get feedback on the effectiveness of advertising material), but also for gaming and monitoring of the emotional state of operators that act in risky contexts such as aviation. Most of these studies have focused on the analysis of facial expressions, but important clues can be derived by the analysis of the dynamics of body parts as well [53]. Using the same geometric framework that was

proposed for action recognition, we evaluated our approach in the task of emotion recognition from human body movement. Here, the used landmarks are in 3D coordinate space, but with better accuracy and higher temporal resolution, with respect to the case of action recognition.

3.6.2.1 Dataset

Experiments have been performed on the Body Motion-Emotion dataset (P-BME), acquired at the Cognitive Neuroscience Laboratory (INSERM U960 - Ecole Normale Supérieure) in Paris [53]. It includes Motion Capture (MoCap) 3D data sequences recorded at high frame rate (120 frames per second) by an Opto-electronic Vicon V8 MoCap system wired to 24 cameras. The body movement is captured by using 43 landmarks that are positioned at joints.

To create the dataset, 8 subjects (professional actors) were instructed to walk following a predefined "U" shaped path that includes forward-walking, turn, and coming back. For each acquisition, actors moved along the path performing one emotion out of five different emotions, namely, *anger*, *fear*, *joy*, *neutral*, and *sadness*. So, each sequence is associated with one emotion label. Each actor performed at maximum five repetitions of a same emotional sequence for a total of 156 instances. Though there is some variation from subject to subject, the number of examples is well distributed across the different emotions: 29 *anger*, 31 *fear*, 33 *joy*, 28 *neutral*, 35 *sadness*.

3.6.2.2 Experimental settings and parameters

Since MoCap skeletons are in 3D coordinate space, we followed the same steps that have been proposed for action recognition, including the decomposition into body parts. An example of this decomposition on MoCap skeletons is shown in Fig. 3.10. Note that the same late fusion of body part classifiers, as mentioned in the previous Section, is adopted.

A cross-validation grid search has been performed to find an optimal value for the weight parameter k.

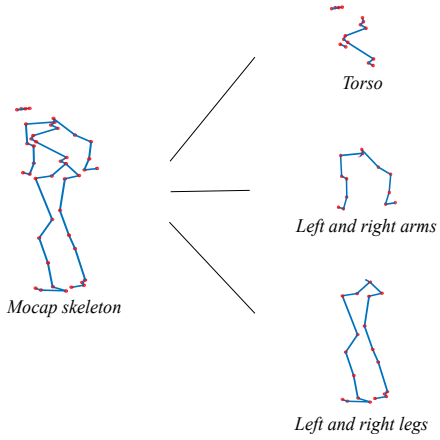

FIGURE 3.10 – Decomposition of the MoCap skeleton into three body parts.

Experiments on the P-BME dataset were performed by using a *leave-one-subject-out* cross validation protocol. With this solution, iteratively, all the emotion sequences of a subject are used for test, while all the sequences of the remaining subjects are used for training.

3.6.2.3 Results and discussion

In Table 3.4, we provide the obtained results as well as a comparative study with baseline experiments on the P-BME dataset.

Similarly to the reported results for action recognition, the proposed fusion of body part classifiers achieved the highest performance with an average accuracy of 81.99% and standard deviation of 4.36%. Considering only the skeletons (without body parts) in the classification, the performance decreased to an average accuracy of 78.15%.

In Fig. 3.11 (left), we report the confusion matrix of different emotions. The diagonal dominance of the matrix can be observed with the best results scored by *neutral* and *anger* (more than 80%), followed by *fear* (71%), *joy* (about 67%), with the lowest accuracy for

TABLE 3.4 – Comparative study of the proposed approach with baseline experiments on the P-BME dataset. First rows: state-of-the-art action and emotion recognition methods and human evaluator; second rows: baseline experiments; last row: ours

Method	**Accuracy (%)**
Human evaluator	74.20
COV3D [29]	71.14 ± 6.77
LARP [123]	74.8 ± 3.17
Traj. on $\mathcal{S}^+(3,n)$ - Flat metric	57.41 ± 8.43
Traj. on $\mathcal{S}^+(3,n)$ - No DTW	63.23 ± 8.62
Traj. on $\mathcal{S}^+(3,n)$ - kNN	68.9 ± 7.63
Traj. on $\mathcal{G}(3,n)$	66.35± 6.43
Traj. on $\mathcal{G}(3,n)$ - BP Fusion	67.09 ± 6.82
Traj. on $\mathcal{S}^+(3,n)$	**78.15 ± 5.79**
Traj. on $\mathcal{S}^+(3,n)$ - BP Fusion	**81.99 ± 4.36**

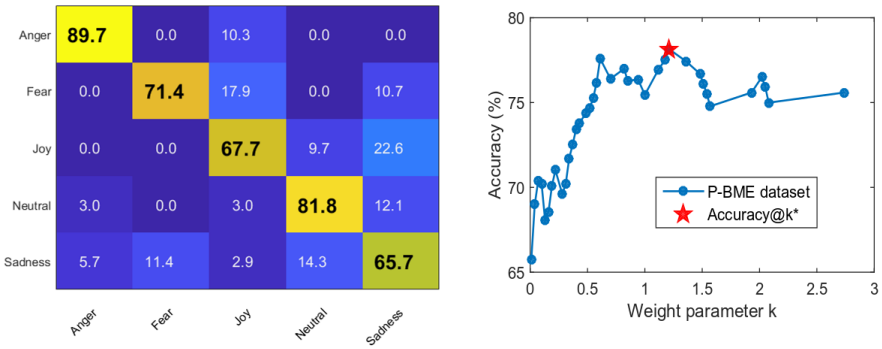

FIGURE 3.11 – P-BME dataset: Confusion matrix (left). Impact of the parameter k on emotion recognition accuracy (right).

sadness (about 65%). In Fig. 3.11 (right), we report the obtained results for $k \in [0, 3]$ with a step of 0.1.

As mentioned in Section 3.4.1, an important step in our approach is the temporal alignment. Avoiding this step and following the same protocol, we found that the performance decreased to 63.23%.

Recently, Daoudi *et al.* [29] proposed a method for emotion recognition from body movement based on covariance matrices and SPD manifold. They used the 3D covariance descriptor (COV3D) of skeleton joints across time to represent sequences without a special handling of the dynamics. They reported and average accuracy of 71.4%. They

also performed a user based test in order to evaluate the performance of the proposed classification method in comparison with a human-based judgment. In this test, thirty-two naive individuals were asked to perform a force-choice task in which they had to choose between one of the five emotions. This resulted in an average value of about 74%. It is relevant to note that the user based test being based on RGB videos provides to the users much more information for evaluation, including the actor's face. Notably, our method is capable to score better results based on the skeleton joints only.

We also compared our results with the Lie algebra relative pairs (LARP) method proposed by Vemulapalli *et al.* [123] for skeleton action recognition. In that work, each skeleton is mapped to a point on the product space of $SE(3) \times SE(3) \cdots \times SE(3)$, where it is modeled using transformations between joint pairs. The temporal evolution of these features is seen as a trajectory on $SE(3) \times SE(3) \times \cdots \times SE(3)$ and mapped to the tangent space of a reference point. A one-versus-all SVM combined with Dynamic Time Warping and Fourier temporal pyramid (FTP) is used for classification. Using this method, an average accuracy of 74.8% was obtained, which is about 8% lower than ours.

The highest accuracy (78.15%) is obtained for $k^* = 1.2$. For $k = 0$, the skeletons are considered as trajectories on the Grassmann manifold $\mathcal{G}(3, n)$, and the obtained accuracy is around 66%, which is 12% lower than the retained result. In order to show the importance of choosing a well defined Riemannian metric in the space of interest, we conducted the same experiments by changing the metric $d_{\mathcal{S}^+}$ defined in Eq. (3.3.8) with a flat metric, defined as the Frobenius norm of the difference between two Gram matrices (skeletons). For this experiment, we report an average accuracy of 57.41% being lower of about 21% than using $d_{\mathcal{S}^+}$.

In Table 3.5, we report the obtained accuracies per emotion for each body part. With this evaluation, we are able to identify body parts that are more informative to a specific emotional state. We can observe that *Anger*, *Fear*, and *Joy* are better recognized with the whole body, while *Neutral* and *Sadness* are better recognized with arms. One can note that

the performance for these two emotions increases after body part fusion compared to the whole body only, notably through the contribution of arms.

TABLE 3.5 – Comparative study of emotion recognition (%) on the P-BME dataset using different parts of the body and our proposed method. Anger (An), Fear (Fe), Joy (Jo), Neutral (Ne), Sadness (Sa), Accuracy (Acc)

Method	**An**	**Fe**	**Jo**	**Ne**	**Sa**	**Acc**
Legs only	55.1	**64.3**	35.5	57.6	60	59.17
Arms only	55.2	57.1	45.2	**84.8**	**71.4**	69.42
Torso only	**82.76**	50	48.4	**75.7**	54.3	67.23
Full body	**89.6**	**78.5**	**58.0**	**72.7**	**65.7**	**78.15**
Late BP Fusion	**89.7**	**71.4**	**67.7**	**81.8**	**65.7**	**81.99**

Finally, we evaluated our approach when considering subsequences of the original sequence. In Table 3.6, we provide the obtained results and the execution time of the testing phase, when considering only 25%, 50%, 75%, and 100% of the sequence. The execution time is recorded for a test sequence of $1,118$ frames (about 8 seconds) when considering separately the four temporal subsequences. The highest execution time is about 2 seconds, which is satisfactory considering the high frame-rate of the data. Unsurprisingly, the best accuracy is obtained when considering the whole sequence. The performance decreases when shorter subsequences are used to perform emotion recognition.

TABLE 3.6 – Emotion recognition accuracy using different sequence lengths on the P-BME dataset

Sequence length	**Accuracy (%)**	**Exec. time (s)**
25% of the sequence	61.20 ± 7.52	1.90
50% of the sequence	67.27 ± 6.36	1.93
75% of the sequence	70.88 ± 6.81	1.95
100% of the sequence	$\mathbf{78.15 \pm 5.79}$	**1.99**

3.6.3 2D facial expression recognition

We evaluated our approach also in the task of facial expression recognition from 2D landmarks. In this case, the landmarks are in a 2D coordinate space, resulting in a Gram matrix of size $n \times n$ of rank 2 for each configuration of n landmarks. The facial sequences are then seen as time-parameterized trajectories on $\mathcal{S}^+(2, n)$.

3.6.3.1 Datasets

We conducted experiments on four publicly available datasets – CK+, MMI, Oulu-CASIA, and AFEW datasets.

Cohn-Kanade Extended (CK+) dataset [82] – It contains 123 subjects and 593 frontal image sequences of posed expressions. Among them, 118 subjects are annotated with the seven labels – *anger* (An), *contempt* (Co), *disgust* (Di), *fear* (Fe), *happy* (Ha), *sad* (Sa) and *surprise* (Su). Note that only the two first temporal phases of the expression, *i.e.*, neutral and onset (with apex frames), are present.

MMI dataset [118] – It consists of 205 image sequences with frontal faces of 30 subjects labeled with the six basic emotion labels. In this dataset each sequence begins with a neutral facial expression, and has a posed facial expression in the middle; the sequence ends up with the neutral facial expression. The location of the peak frame is not provided as a prior information.

Oulu-CASIA dataset [143] – It includes 480 image sequences of 80 subjects, taken under normal illumination conditions. They are labeled with one of the six basic emotion labels. Each sequence begins with a neutral facial expression and ends with the apex of the expression.

AFEW dataset [33] – Collected from movies showing close-to-real-world conditions, which depict or simulate the spontaneous expressions in uncontrolled environment. The task is to classify each video clip into one of the seven expression categories (the six basic emotions plus the neutral).

3.6.3.2 Experimental settings and parameters

All our experiments were performed once facial landmarks were extracted using the method proposed in [8] on the CK+, MMI, and Oulu-CASIA datasets. On the challenging

AFEW dataset, we have considered the corrections provided in [2] after applying the same detector. The number of landmarks is $n = 49$ for each face. In this case, we applied the adaptive re-sampling of trajectories proposed in Section 3.4.2 that enhances small facial deformations and disregards redundant frames. This step involves two parameters ζ_1 and ζ_2 for up-sampling and down-sampling, respectively. These two parameters are chosen so that all the trajectories in the dataset have the same length, equal to the median length. For the parameter k, the same procedure as for action and emotion recognition from body movement is applied.

To evaluate our approach, we followed the experimental settings commonly used in recent works. Following [42, 65, 79], we have performed 10-fold cross validation experiments for the CK+, MMI, and Oulu-CASIA datasets. In contrast, the AFEW dataset was divided into three sets: training, validation and test, according to the protocols defined in EmotiW'2013 [32]. Here, we only report our results on the validation set for comparison with [32, 42, 79].

3.6.3.3 Results and discussion

On CK+, the average accuracy is 96.87%. Note that the accuracy of the trajectory representation on $\mathcal{G}(2, n)$, following the same pipeline is 2% lower, which confirms the contribution of the covariance embedded in our representation.

An average classification accuracy of 79.19% is reported for the MMI dataset. Note that based on geometric features only, our approach grounding on both representations on $\mathcal{S}^+(2, n)$ and $\mathcal{G}(2, n)$ achieved competitive results with respect to the literature (see Table 3.7). On the Oulu-CASIA dataset, the average accuracy is 83.13%, hence 3% higher than the Grassmann trajectory representation. This is the highest accuracy reported in literature (refer to Table 3.8). Finally, we reported an average accuracy of 39.94% on the AFEW dataset. Despite being competitive with respect to recent literature (see Table 3.8),

2. http://sites.google.com/site/chehrahome

these results evidence that AFER "in-the-wild" is still challenging.

We highlight the superiority of the trajectory representation on $\mathcal{S}^+(2,n)$ over the Grassmannian (refer to Table 3.7 and Table 3.8). This is due to the contribution of the covariance part further to the conventional affine-shape analysis over the Grassmannian. Recall that k serves to balance the contribution of the distance between covariance matrices living in $\mathcal{P}_2$ with respect to the Grassmann contribution $\mathcal{G}(2,n)$. The optimal performance are achieved for the following values – $k^*_{CK+} = 0.081$, $k^*_{MMI} = 0.012$, $k^*_{Oulu-CASIA} = 0.014$ and $k^*_{AFEW} = 0.001$. In Fig. 3.12, we study the method when varying the parameter k (closeness). The graphs report the method accuracy on CK+, MMI, Oulu-CASIA, and AFEW, respectively.

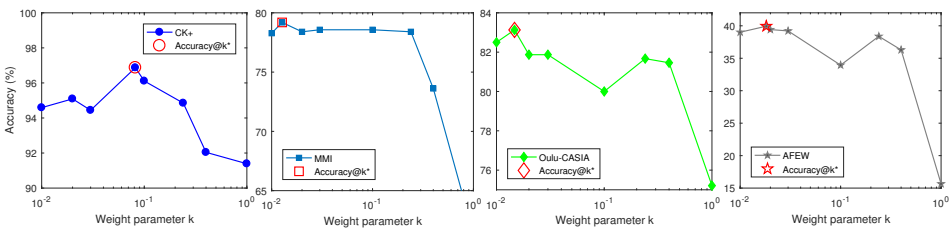

FIGURE 3.12 – Accuracy of the proposed approach when varying the weight parameter k on, from left to right, CK+, MMI, Oulu-CASIA and AFEW.

In the left panel of Fig. 3.13, we show the confusion matrix on the CK+ dataset. While individual accuracies of *"anger"*, *"disgust"*, *"happiness"*, and *"surprise"* are high (more than 96%), recognizing *"contempt"* and *"fear"* is still challenging (less than 92%). In the right panel of the same figure, we can observe that the best accuracy on the MMI dataset was also achieved for *"happiness"* followed by *"surprise"*. Also in this case, the lowest performance was recorded for *"fear"* expression.

As shown in Fig. 3.13, on the Oulu-CASIA dataset the highest performance was reached for *"happiness"* (91.3%) and *"surprise"* (93.8%) expressions; *"Disgust"*, *"fear"*, and *"sadness"* were the most challenging expressions in this dataset ($< 79\%$). Unsurprisingly for the AFEW dataset, the *"neutral"* (63.5%), *"anger"* (56.3%), and *"happiness"* (66.7%) expressions are

TABLE 3.7 – Overall accuracy (%) on CK+ and MMI datasets. Here, $^{(A)}$: appearance (or color); $^{(G)}$: geometry (or shape); *: Deep Learning based approach; last row: ours

Method	**CK+**	**MMI**
$^{(A)}$ 3D HOG (from [65])	91.44	60.89
$^{(A)}$ 3D SIFT (from [65])	-	64.39
$^{(A)}$ Cov3D (from [65])	92.3	-
$^{(A)}$ STM-ExpLet [79] (10-fold)	**94.19**	**75.12**
$^{(A)}$ CSPL [145] (10-fold)	89.89	73.53
$^{(A)}$ F-Bases [99] (LOSO)	**96.02**	**75.12**
$^{(A)}$ ST-RBM [42] (10-fold)	**95.66**	**81.63**
$^{(A)}$ 3DCNN-DAP [78] * (15-fold)	87.9	62.2
$^{(A)}$ DTAN [65] * (10-fold)	91.44	62.45
$^{(A+G)}$ DTAGN [65] * (10-fold)	**97.25**	**70.24**
$^{(G)}$ DTGN [65] * (10-fold)	92.35	59.02
$^{(G)}$ TMS [61] (4-fold)	85.84	-
$^{(G)}$ HMM [133] (15-fold)	83.5	51.5
$^{(G)}$ ITBN [133] (15-fold)	86.3	59.7
$^{(G)}$ Velocity on $\mathcal{G}(n,2)$[111]	82.8	-
$^{(G)}$ traj. on $\mathcal{G}(2,n)$ (10-fold)	94.25 ± 3.71	78.18 ± 4.87
$^{(G)}$ **traj. on $\mathcal{S}^+(2,n)$ (10-fold)**	**96.87 ± 2.46**	**79.19 ± 4.62**

better recognized over the rest (see the right confusion matrix in Fig. 3.14).

It is important to note that the *"fear"* expression was the most challenging expression in all the datasets. In fact, this expression involves several action unit activations (*i.e.*, AU1+AU2+AU4+AU5+AU7+AU20+AU26) [47] that are quite difficult to detect by using only geometric features.

Comparative Study with the State-of-the-Art. In Table 3.7 and Table 3.8, we

TABLE 3.8 – Overall accuracy on Oulu-CASIA and AFEW dataset (following the EmotiW'13 protocol [32])

Method	**Oulu-CASIA**	**AFEW**
$^{(A)}$ HOG 3D [70]	70.63	26.90
$^{(A)}$ 3D SIFT [101]	55.83	24.87
$^{(A)}$ LBP-TOP [144]	68.13	25.13
$^{(A)}$ EmotiW [32]	-	27.27
$^{(A)}$ STM [79]	-	29.19
$^{(A)}$ STM-ExpLet [79]	74.59	31.73
$^{(A+G)}$ DTAGN [65] * (10-fold)	**81.46**	-
$^{(A)}$ ST-RBM [42]	-	**46.36**
$^{(G)}$ **traj. on $\mathcal{G}(2,n)$**	80.0 ± 5.22	39.1
$^{(G)}$ **traj. on $\mathcal{S}^+(2,n)$**	**83.13 ± 3.86**	**39.94**

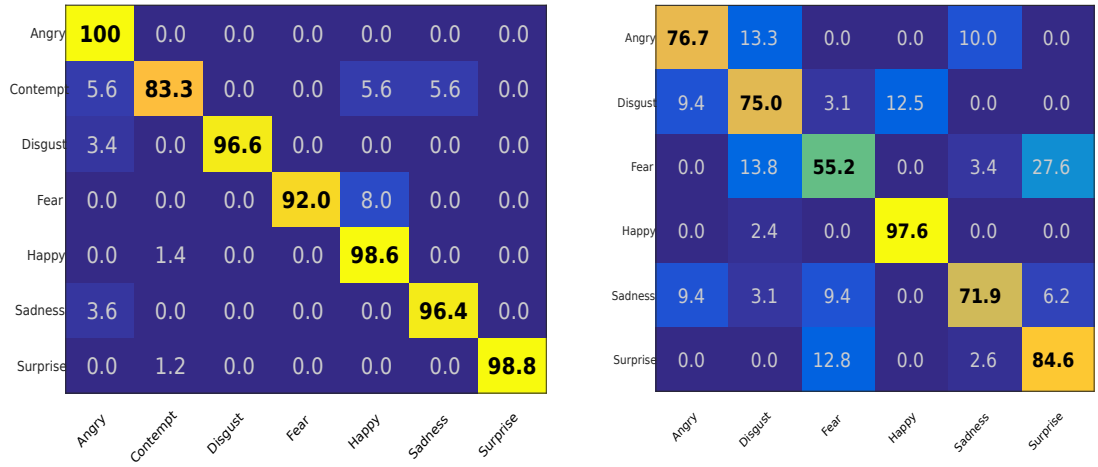

FIGURE 3.13 – Confusion matrices on the CK+ (left) and MMI (right) datasets.

compare our approach over the recent literature. Overall, our approach achieved competitive performance with respect to the most recent approaches. On CK+, we obtained the second highest accuracy. The ranked-first approach is DTAGN [65], in which two deep networks are trained on shape and appearance channels, then fused. Note that the geometry deep network (DTGN) achieved 92.35%, which is much lower than ours. Furthermore, our approach outperforms the ST-RBM [42] and the STM-ExpLet [79]. On the MMI dataset, our approach outperforms the DTAGN [65] and the STM-ExpLet [79]. However, it is behind ST-RBM [42].

On the Oulu-CASIA dataset, our approach shows a clear superiority to existing methods, in particular STM-ExpLet [79] and DTGN [65]. Elaiwat *et al.* [42] do not report any results on this dataset, however, their approach achieved the highest accuracy on AFEW. Our approach is ranked second showing a superiority to remaining approaches on AFEW.

Baseline experiments. Based on the results reported in Table 3.9, we discuss in this paragraph algorithms and their computational complexity with respect to baselines. Firstly, we studied the computational cost of the proposed framework in the task of 2D facial expression recognition on the CK+ dataset. Correspondingly to 3D action recognition settings, we report in the top of Table 3.9 the running time statistics for trajectory

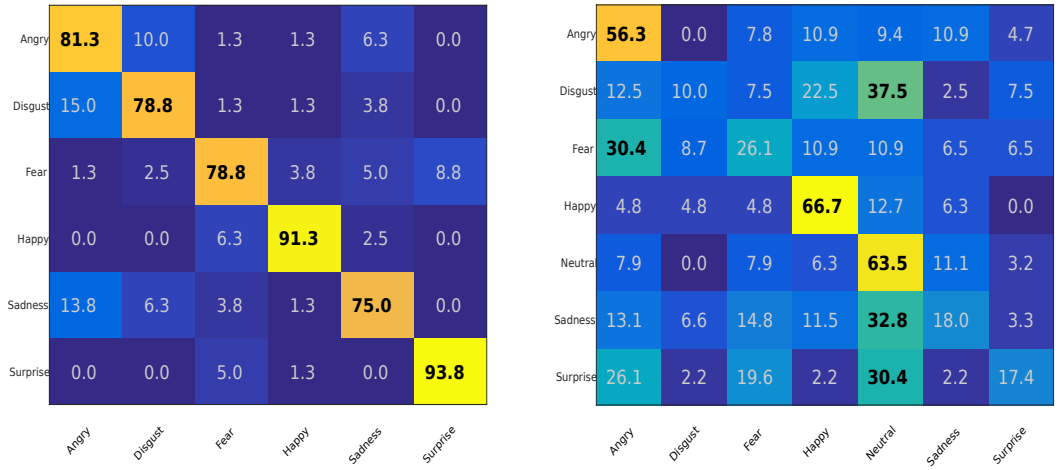

FIGURE 3.14 – Confusion matrices on the Oulu-CASIA (left) and AFEW (right) datasets.

construction, comparison of trajectories, and the testing phase of trajectory classification in $\mathcal{S}^+(2,n)$.

Then, we have used different distances defined on $\mathcal{S}^+(2,n)$. Specifically, given two matrices G_1 and G_2 in $\mathcal{S}^+(2,n)$: (1) we used $d_{\mathcal{P}_n}$ to compare them by regularizing their ranks, *i.e.*, making them n full-rank, and considering them in $\mathcal{P}_n$ (the space of n-by-n positive definite matrices), $d_{\mathcal{P}_n}(G_1,G_2) = d_{\mathcal{P}_n}(G_1+\epsilon I_n, G_2+\epsilon I_n)$; (2) we used the Euclidean flat distance $d_{\mathcal{F}^+}(G_1,G_2) = \|G_1 - G_2\|_F$, where $\|.\|_F$ denotes the Frobenius-norm. The closeness $d_{\mathcal{S}^+}$ between two elements of $\mathcal{S}^+(2,n)$ defined in Eq. (7) is more suitable, compared to the distance $d_{\mathcal{P}_n}$ and the flat distance $d_{\mathcal{F}^+}$ defined in literature. This demonstrates the importance of being faithful to the geometry of the manifold of interest. Another advantage of using $d_{\mathcal{S}^+}$ over $d_{\mathcal{P}_n}$ is the computational time, as it involves n-by-2 and 2-by-2 matrices instead of n-by-n matrices. Note that the provided execution times are relative to the comparison of two arbitrary sequences.

Table 3.9 reports the average accuracy when DTW in used or not in our pipeline, on both the CK+ and MMI datasets. It is clear from these experiments that a temporal alignment of the trajectories is a crucial step, as an improvement of about 12% is obtained on MMI

TABLE 3.9 – Baseline experiments and computational complexity on the CK+, MMI and AFEW datasets

Pipeline component	**Time** (s)
Trajectory construction in $\mathcal{S}^+(2,n)$	0.007
Comparison of trajectories in $\mathcal{S}^+(2,n)$	0.055
Classification of a trajectory in $\mathcal{S}^+(2,n)$	6.28

Distance	**CK+** (%)	**Time** (s)
Flat distance $d_{\mathcal{F}^+}$	93.78 ± 2.92	0.020
Distance $d_{\mathcal{P}_n}$ in $\mathcal{P}_n$	92.92 ± 2.45	0.816
Closeness $d_{\mathcal{S}^+}$	**96.87± 2.46**	0.055

Temporal alignment	**CK+ (%)**	**MMI (%)**	**Time** (s)
without DTW	90.94 ± 4.23	66.93 ± 5.79	0.018
with DTW	**96.87 ± 2.46**	**79.19 ± 4.62**	0.055

Adaptive re-sampling	**MMI (%)**	**AFEW (%)**
without resampling	74.72 ± 5.34	36.81
with resampling	**79.19 ± 4.62**	**39.94**

Classifier	**CK+ (%)**	**AFEW (%)**
K-NN	88.97 ± 6.14	29.77
ppf-SVM	**96.87 ± 2.46**	**39.94**

and of approximately 6% on CK+.

The adaptive re-sampling tool is also analyzed. When it is included in the pipeline, an improvement of about 5% is achieved on MMI and 3% on AFEW.

In the last Table, we compare the results of ppfSVM with a k-Nearest Neighbor classifier for both the CK+ and AFEW datasets. The number of nearest neighbors k to consider for each dataset is chosen by cross-validation. On CK+, we obtained an average accuracy of 88.97% for $k = 11$. On AFEW, we obtained an average accuracy of 29.77% for $k = 7$. These results are outperformed by the ppfSVM classifier.

3.7 Conclusion

In this chapter, we have proposed a geometric approach for effectively modeling and classifying dynamic 2D and 3D landmark sequences for human behavior understanding. Based on Gramian matrices derived from the static landmarks, our representation consists

of an affine-invariant shape representation and a spatial covariance of the landmarks. We have exploited the Riemannian geometry of the space of Gram matrices to define a closeness between static shape representations. Then, we have derived computational tools to align, re-sample and compare these trajectories giving rise to a rate-invariant analysis. Finally, landmark sequences are learned from these trajectories using a variant of SVM, called ppfSVM, which allows us to deal with the nonlinearity of the space of representation. We evaluated our approach in three different applications, namely, 3D human action recognition, 3D emotion recognition from body movement, and 2D facial expression recognition. Extensive experiments on nine publicly available datasets showed that the proposed approach achieves competitive or better results than state-of-art solutions.

Chapitre 4

Barycentric Representation of Facial Landmarks for Expression Recognition and Depression Severity Level Assessment

4.1 Introduction

In the previous chapter, we have introduced a novel shape representation based on the Gram matrix. After matrix decomposition, we have showed that this representation brings two different information; the first one was the spatial covariance given by the positive definite matrix; and the second and most important one was the affine-invariant shape information given by the orthogonal matrix. The latter lies on the Grassmann manifold which is a non-linear space where inference algorithms are not applicable in a straightforward manner. In this chapter we propose an affine-invariant shape representation using barycentric coordinates of 2D facial landmarks. While being closely related to the

conventional Grassmann representation, the barycentric one has the advantage to lie on an Euclidean space. Thanks to the Euclidean nature of the barycentric representation, one can safely use standard computational and machine learning tools. We evaluate the proposed representation in two different face analysis tasks namely, facial expression recognition in unconstrained environments, and automatic assessment of depression severity level.

4.2 Affine-invariant shape representation using barycentric coordinates

As stated in Section 2.3.2 of chapter 2, the analysis of moving landmarks may be distorted by view variations. The problem is more acute when it comes to dealing with 2D landmarks. Indeed, in the 2D case these distortions are due to undesirable projective transformations which should be filtered out to have a robust representation of 2D landmarks to view variations. These projective transformations are difficult to be filtered out, but they can be approximated by affine transformations, especially when the face is far from the camera [111]. In this section we briefly review the main definitions of the affine-invariance with *barycentric coordinates* and their use in 2D facial shape analysis [66].

In order to study the motion of an ordered list of n landmarks $Z_1(t), Z_2(t), \ldots, Z_n(t)$, where t represents the time parametrization and $Z_i(t) = (x_i(t), y_i(t)), 1 \leq i \leq n$, in the plane up to the action of an arbitrary affine transformation, a standard technique is to consider the span of the columns of the $n \times 3$ time-dependent matrix

$$M(t) := \begin{pmatrix} x_1(t) & y_1(t) & 1 \\ \vdots & \vdots & \vdots \\ x_n(t) & y_n(t) & 1 \end{pmatrix} .$$

If at any time t there exists a fixed triplet of landmarks forming a non-degenerate triangle, the rank of the matrix $M(t)$ is constantly equal to 3 and the span of its columns is a curve of three-dimensional subspaces in $\mathbb{R}^n$. In other words, a curve in the Grassmannian $\mathcal{G}(3, n)$,

which is well known [12] to be an affine-invariant of the motion. This convenient way of filtering out the affine transformations opens the way to the use of metric and differential-geometric techniques in the study and classification of moving landmarks [123, 13, 30, 67, 6].

It is worth noting that this representation in $\mathcal{G}(3,n)$ is equivalent to the Grassmann representation in $\mathcal{G}(2,n)$ which was studied and described in the previous chapter [111, 67]. The latter was obtained by centering the 2D landmarks and considering the span of the columns of the $n \times 2$ matrix as an affine-invariant representation in $\mathcal{G}(2,n)$ without adding a column of ones to the matrix formed by the 2D coordinates.

Another convenient and more classic way to filter out affine transformations is through the use of *barycentric coordinates*. This method can be applied given three of the landmarks which form a non-degenerate triangle throughout all their motion. Indeed, assume, without loss of generality, that $Z_1(t)$, $Z_2(t)$, and $Z_3(t)$ are the vertices of a non-degenerate triangle *for every value of* t. In the case of facial shapes, the right and left corners of the eyes and the tip of the nose are chosen to form a non-degenerate triangle (see the red triangle in Fig. 4.1). For $i = 4, .., n$ and at any time t, we can write

$$Z_i(t) = \lambda_{i1}(t)Z_1(t) + \lambda_{i2}(t)Z_2(t) + \lambda_{i3}(t)Z_3(t) \ ,$$

where the numbers $\lambda_{i1}(t)$, $\lambda_{i2}(t)$, and $\lambda_{i3}(t)$ satisfy

$$\lambda_{i1}(t) + \lambda_{i2}(t) + \lambda_{i3}(t) = 1.$$

The last condition renders the triplet of barycentric coordinates $(\lambda_{i1}(t), \lambda_{i2}(t), \lambda_{i3}(t))$ unique. In fact, it is equal to

$$(x_i(t), y_i(t), 1) \begin{pmatrix} x_1(t) & y_1(t) & 1 \\ x_2(t) & y_2(t) & 1 \\ x_3(t) & y_3(t) & 1 \end{pmatrix}^{-1} .$$

If T is an affine transformation of the plane, the barycentric representation of $TZ_i(t)$ in terms of the frame given by $TZ_1(t)$, $TZ_2(t)$, and $TZ_3(t)$ is still $(\lambda_{i1}(t), \lambda_{i2}(t), \lambda_{i3}(t))$. This

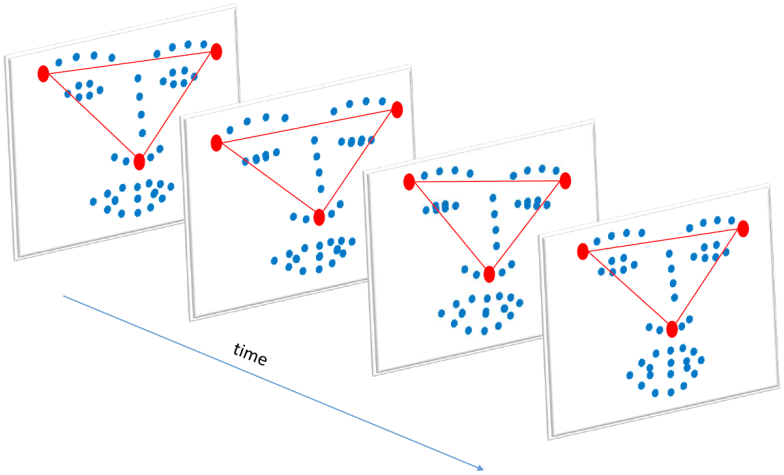

FIGURE 4.1 – Example of the automatically tracked 49 facial landmarks. The three red points denote the facial landmarks used to form the non-degenerate triangle required to compute the barycentric coordinates.

allows us to derive the $(n-3)\times 3$ matrix

$$\Lambda(t) := \begin{pmatrix} \lambda_{41}(t) & \lambda_{42}(t) & \lambda_{43}(t) \\ \vdots & \vdots & \vdots \\ \lambda_{n1}(t) & \lambda_{n2}(t) & \lambda_{n3}(t) \end{pmatrix} .$$

as the affine shape representation of the moving landmarks.

4.2.1 Relationship with the conventional Grassmannian representation

A topological space $\mathcal{M}$ is a topological manifold of dimension *dim* if it is locally Euclidean. That means that every point $X \in \mathcal{M}$ has a neighborhood that is homeomorphic to an open subset of $\mathbb{R}^{dim}$. A coordinate chart (or just a chart on $\mathcal{M}$) is a pair (Σ, Φ), where Σ is an open subset of $\mathcal{M}$ and $\Phi : \Sigma \to \tilde{\Sigma}$ is homeomorphism from Σ to the open set $\tilde{\Sigma} \in \mathbb{R}^{dim}$. The definition of topological manifold implies that each point $X \in \mathcal{M}$ is contained in the domain of some coordinate chart [10]. In the case of the affine-invariant Grassmannian representation in $\mathcal{G}(3,n)$, the points on the Grassmannian corresponding to the facial landmarks are naturally contained in one of the standard charts. It turns out that

passing to this chart is nothing more than taking the barycentric coordinates with respect to a specific triplet of landmark points.

In order to expose the basic relationship between the Grassmannian representation and the barycentric one, let us recall, in a particular case, the usual way to construct charts in the Grassmannian. If $\zeta \in \mathcal{G}(3,n)$ is a subspace that intersects the $(n-3)$-dimensional subspace

$$W = \{(0,0,0,x_4,\ldots,x_n) : x_i \in \mathbb{R}^n \text{ for } i \text{ between 4 and } n\}$$

only at the origin, and $\mathbf{x} = (x_1,\ldots,x_n)$, $\mathbf{y} = (y_1,\ldots,y_n)$, and $\mathbf{z} = (z_1,\ldots,z_n)$ is a basis for ζ, then the 3×3 matrix

$$\begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{pmatrix}$$

is invertible and the $(n-3)\times 3$ matrix

$$\begin{pmatrix} x_4 & y_4 & z_4 \\ \vdots & \vdots & \vdots \\ x_n & y_n & z_n \end{pmatrix} \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{pmatrix}^{-1}$$

is independent of the chosen basis. In this way, the open and dense set of 3-dimensional subspaces transverse to W are put in a bijective correspondence with $\mathbb{R}^{(n-3)\times 3}$.

If we consider the curve in $\mathcal{G}(3,n)$ given by the span of the columns of the matrix

$$M(t) := \begin{pmatrix} x_1(t) & y_1(t) & 1 \\ \vdots & \vdots & \vdots \\ x_n(t) & y_n(t) & 1 \end{pmatrix}$$

and if the landmarks $Z_1(t) = (x_1(t), y_1(t))$, $Z_2(t) = (x_2(t), y_2(t))$, and $Z_3(t) = (x_3(t), y_3(t))$ form a non-degenerate triangle throughout all their motion, then composing this curve with

a chart in the Grassmannian yields the curve of matrices

$$\begin{pmatrix} x_4(t) & y_4(t) & 1 \\ \vdots & \vdots & \vdots \\ x_n(t) & y_n(t) & 1 \end{pmatrix} \begin{pmatrix} x_1(t) & y_1(t) & 1 \\ x_2(t) & y_2(t) & 1 \\ x_3(t) & y_3(t) & 1 \end{pmatrix}^{-1} ,$$

which is just the curve $\Lambda(t)$ encoding the barycentric representation of the landmarks. For more details about the affine-invariance with barycentric coordinates, please refer to the page 81 of the book [14]. In what follows, we will consider the introduced affine-invariant vector Λ, with dimension $m = (n-3) \times 3$, to represent a static facial shape and the curve $\Lambda(t)$ to denote a facial shape sequence.

4.3 Metric learning on barycentric representation for expression recognition in unconstrained environments

Given the facial shape represented by the affine-invariant vector Λ, with dimension $m = (n-3) \times 3$, we seek a suitable metric that is discriminative enough in terms of expression to compare them. The Euclidean distance, defined as the squared l_2-norm of the difference of the vectors, could be a reasonable choice since the defined shapes lie in Euclidean space. However, such distance disregards the specific nature of the considered facial shapes. To overcome this issue, we propose to learn a Mahalanobis distance instead of using the standard Euclidean distance [73]. Given two facial shapes represented by the affine-invariant vectors Λ_i and Λ_j in $\mathbb{R}^m$, the Mahalanobis distance is defined by

$$d^2_{l_{ij}}(\Lambda_i, \Lambda_j) = (\Lambda_i - \Lambda_j)^T A (\Lambda_i - \Lambda_j) \ , \tag{4.3.1}$$

where A is a positive semi-definite (p.s.d) matrix of size $m \times m$. The problem of metric learning is then to find the best p.s.d matrix A that best discriminates the facial expressions, *i.e.,* results in small distances when the facial shapes represent similar expressions and large distances when they represent different expressions.

Let $\mathcal{D} = \{(\Lambda_1, c_1), \ldots, (\Lambda_N, c_N)\}$ represent a set of affine-invariant shapes in $\mathbb{R}^m$ annotated with the corresponding expressions (*e.g.*, $c =$'happy', 'angry', etc.). Let $\{\Lambda_i, \Lambda_j, \Lambda_k\}$ be a triplet of affine-invariant shapes from $\mathcal{D}$ such that (Λ_i, Λ_j) have the same label ($c_i = c_j$), and (Λ_i, Λ_k) with different labels ($c_i \neq c_k$). We aim to find an optimal p.s.d matrix A such that $d^2_{l_{ij}}(\Lambda_i, \Lambda_j) < d^2_{l_{ik}}(\Lambda_i, \Lambda_k)$. That is, we wish to find a p.s.d matrix A that minimizes $d^2_{l_{ij}} - d^2_{l_{ik}} = (\Lambda_i - \Lambda_j)^T A(\Lambda_i - \Lambda_j) - (\Lambda_i - \Lambda_k)^T A(\Lambda_i - \Lambda_k)$. In order to solve this optimization problem, we follow the convenient method described by Shen *et al.* [105], where a boosting is used. This method is based on the observation that any positive semidefinite matrix can be decomposed into a linear combination of trace-one rank-one matrices. It uses rank-one positive semidefinite matrices as weak learners within an efficient and scalable boosting-based learning process.

4.3.1 Facial expression classification

The learned distance does, indeed, assign small distances to similar static facial shapes and large distances to dissimilar shapes. However, as conveying an expression is a temporal process, we are more interested in comparing facial shape sequences. Accordingly, we exploit the learned distance to build a rate-invariant similarity measure between facial shape sequences. Specifically, the Dynamic Time Warping (DTW) algorithm [15], employing the learned distance instead of the standard Euclidean distance, is used to compare two facial sequences.

Following [9, 67], we adopt the *pairwise proximity function SVM* (ppfSVM) [50, 51] to classify the facial sequences. PpfSVM requires the definition of a similarity measure to compare samples. In our case, it is natural to consider the similarity measure given by our version of DTW for such a comparison. An overview of the proposed method is shown in Fig. 4.2.

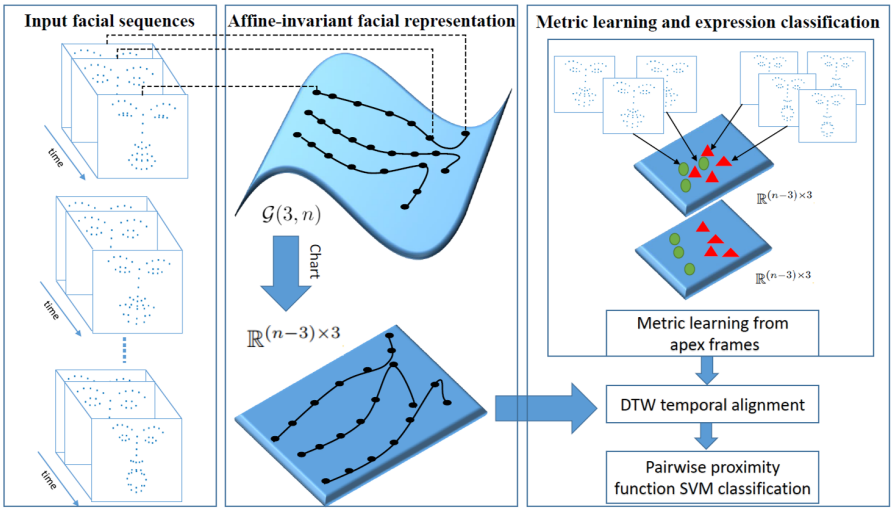

FIGURE 4.2 – Overview of the proposed approach (arycentric representation and metric learning) – After automatic landmark detection for each frame of the video, we represent the resulting shapes through their barycentric coordinates. While being closely related to the affine-invariant Grassmann representation, this representation allows us to work directly on Euclidean space where a metric learning algorithm is applied. Dynamic Time Warping (DTW) using the learned metric is then performed to align the facial sequences. Finally, the ppfSVM exploiting the DTW similarity measure is used as expression classifier.

4.3.2 Experimental results

In order to learn the metric, we use only peak frames from each facial sequence, where the expression reaches its peak. Since peak frames are difficult to detect in uncontrolled facial expressions, we performed the metric learning using extracted landmarks from CK+ dataset [82] which is captured in strict controlled conditions. In this dataset, 309 facial sequences of 118 subjects are annotated with the six labels (the six basic emotions). In all the sequences, the actors start by being neutral then perform the expression until reaching a peak. In our experiments, we only used the five last frames and the first frame from all the sequences. The labels of the five last frames are assigned according to the label of the sequence, while the label of the first frame is always considered as 'neutral'. A total number of 16686 facial shapes are used for the training phase to learn the Mahalanobis distance.

To evaluate the proposed approach, we conducted experiments on the well-known AFEW dataset [33] which was described in the previous chapter. Note that our experiments are made once the facial landmarks are extracted using the method proposed in [8]. The three points used to form the non-degenerate triangle, essential to build the affine-invariant shapes from the landmarks, are the points positioned at the left and right corners of the eye and the nose tip.

All our programs were implemented in Matlab and run on a 2.8 GHZ CPU. We used the multi-class SVM implementation of the LibSVM library [25], and the codes given by [105] for the metric learning.

4.3.2.1 Results and discussions

Following the experimental settings mentioned in the previous Section, we report an accuracy of 38.38%. From the corresponding confusion matrix shown in Fig. 4.3, we can observe that the highest performances are obtained for 'Anger' (51.6%), 'Happiness' (58.7%), and 'Neutral' (55.6%). Since AFEW is a very challenging dataset, the obtained results

are competitive with state-of-art approaches as shown in Table 4.1. We recorded better performance than many appearance based approaches such as SPDNet [58] and STM-ExpLet [79].

Our results are outperformed by the Gram trajectory representation proposed in the previous chapter [67]. However, the execution time of comparing two arbitrary sequences on AFEW dataset is 0.064 seconds with the barycentric approach against 0.84 seconds with the Gram approach. In Table 4.1, we can observe that our results compared to the Gram approach are outperformed by only 1% while being 10 times faster.

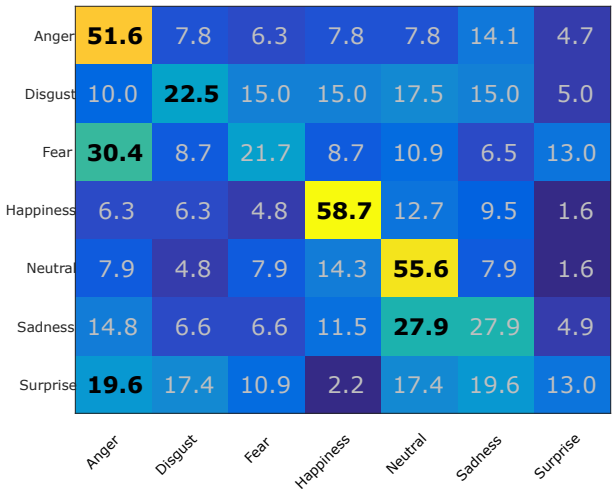

FIGURE 4.3 – Confusion matrix on AFEW dataset

TABLE 4.1 – Overall accuracy AFEW dataset (FER with Barycentric representation)

Method	**Accuracy** (%)
$^{(A)}$ HOG 3D [70]	26.90
$^{(A)}$ HOE [130]	19.54
$^{(A)}$ 3D SIFT [101]	24.87
$^{(A)}$ LBP-TOP [144]	25.13
$^{(A)}$ EmotiW [32]	27.27
$^{(A)}$ STM [79]	29.19
$^{(A)}$ STM-ExpLet [79]	31.73
$^{(A)}$ SPDNet [58]	34.23
$^{(G)}$ Gram Trajectories [67]	**39.94**
$^{(G)}$ **Ours**	**38.38**

To evaluate the different steps of the proposed pipeline, we performed baseline experiments. Firstly, we conducted the same experiments while using alternative representations and metrics. We compared our results with a conventional Grassmann affine-invariant representation coupled with a Riemannian metric given by the subspace angles [12, 111]. The achieved accuracy is around 2.5% lower than ours. We also replaced the learned Mahalanobis distance with a standard Euclidean distance. Here also, the performance decreases by about 3%. In Table 4.2, we show the achieved accuracies by the described alternative representations and metrics and the necessary execution time to compare two arbitrary facial shapes. One can observe that the proposed representation achieves better performance than the Grassmannian while being less time consuming. These results show the effectiveness of the proposed representation and the importance of the metric learning step in our pipeline. As mentioned in the previous Section, we used the five last (peak) frames from the sequences of CK+ dataset to learn the Mahalanobis distance. In Table 4.2, we provide the obtained accuracies when using one, two, five and seven last peak frames from each sequence. The highest accuracy is obtained with the last five frames. Besides, we report in Table 4.2 the average accuracy when DTW is used or not in our pipeline. It is clear from these experiments that a temporal alignment is an important step as an improvement of around 7% is obtained. In the last Table, we compare the results of ppfSVM to a K-NN classifier coupled with the introduced DTW similarity measure. The number of nearest neighbors K is chosen by cross-validation. We obtained an average accuracy of 31.33% for $K = 5$. These results are outperformed by ppfSVM classifier.

4.4 Facial and head movements analysis for depression severity level assessment

Many of the symptoms of depression are observable. In depression facial expressiveness [94, 100] and head movement [45, 64, 49] are reduced.

TABLE 4.2 – Baseline experiments (FER with barycentric representation)

Distance	**Accuracy** (%)	**Time** (μs)
Subspace angles in $\mathcal{G}(3,n)$	36.81	2967
Euclidean distance	36.55	530
Mahalanobis distance d_l	**38.38**	**568**

Number of peak frames	**Accuracy** (%)
1 peak frame	37.07
2 peak frames	37.59
5 peak frames	**38.38**
7 peak frames	36.29

Temporal alignment	**Accuracy** (%)	**Time** (s)
without DTW	30.8	0.008
with DTW	**38.38**	**0.064**

Classifier	**Accuracy** (%)
K-NN	31.33
ppf-SVM	**38.38**

Yet, systematic means of using observable behavior to inform screening and diagnosis of the occurrence and severity of depression are lacking. Recent advances in computer vision and machine learning have explored the validity of automatic measurement of depression severity from video sequences [3, 119, 135, 35].

Hdibeklioglu and colleagues [35] proposed a multimodal deep learning based approach to detect depression severity in participants undergoing treatment for depression. Deep learning based per-frame coding and per-video Fisher-vector based coding were used to characterize the dynamics of facial and head movement. For each modality, selection among features was performed using combined mutual information, which improved accuracy relative to blanket selection of all features regardless of their merit. For individual modalities, facial and head movement dynamics outperformed vocal prosody. For combinations, fusing the dynamics of facial and head movement was more discriminative than head movement dynamics and more discriminative than facial movement dynamics plus vocal prosody and head movement dynamics plus vocal prosody. The proposed deep learning based method outperformed the state of the art counterparts for each modality.

A limitation of the deep learning approach is its lack of interpretability. The dynamics of

facial, head, and vocal prosody were important, but the nature of those changes during course of depression were occult. From their findings, one could not say whether dynamics were increasing, decreasing, or varying in some non-linear way. For clinical scientists and clinicians interested in the mechanisms and course of depression, interpretable features matter. They want to know not only presence or severity of depression but how dynamics vary with occurrence and severity of depression.

Two previous shallow-learning approaches to depression detection were interpretable but less sensitive to depression severity. In Alghowinem and colleagues [3], head movements were tracked by AAMs [98] and modeled by Gaussian mixture models with seven components. Mean, variance, and component weights of the learned GMMs were used as features. And a set of interpretable head pose functionals was proposed. These included the statistics of head movements and duration of looking in different directions.

Williamson and his colleagues [135] investigated the specific changes in coordination, movement, and timing of facial and vocal signals as potential symptoms for self-reported BDI (Beck Depression Inventory) scores [11]. They proposed a multi-scale correlation structure and timing feature sets from video-based facial action units (AUs [40]) and audio-based vocal features. The features were combined using a Gaussian mixture model and extreme learning machine classifiers to predict BDI scores.

Reduced facial expression is commonly observed in depression and relates to deficits in experiencing positive as well as negative emotion [95]. Less often, greatly increased expression occurs. There are referred to as psychomotor retardation and psychomotor agitation, respectively.

In our study, we propose to capture aspects of psychomotor retardation and agitation using the dynamics of facial and head movement. Participants were from a clinical trial for treatment of moderate to severe depression and had history of multiple depressive episodes. Compared to state-of-the-art deep learning approch for depression severity assessment, we

propose a reliable and clinically interpretable method of automatically measuring depression severity from the dynamics of face and head motion.

After extraction of facial landmarks with a state-of-the-art solution [63], we encode them using the barycentric representation introduced in Section 4.2. Because we are interested in both facial movement dynamics and head movement dynamics, the later is encoded by combining the 3 degrees of freedom of head movement (*i.e.,* yaw, roll, and pitch angles) in a single rotation matrix mapped to Lie algebra to overcome the non-linearity of the space of rotation matrices [123, 124].

To capture changes in the dynamics of head and facial movement that would reflect the psychomotor retardation of depressed participants, relevant kinematic features are extracted (*i.e.,* velocities and accelerations) from each proposed representation. Gaussian Mixture Models (GMM) combined with an improved fisher vector encoding are then used to obtain a single vector representation for each sequence (*i.e.,* interview). Finally, a multi-class SVM with a Gaussian kernel is used to classify the encoded facial and head movement dynamics into three depression severity levels. The overview of the proposed approach is shown in Fig. 4.4.

4.4.1 Facial movements analysis using barycentric coordinates

In order to analyze the facial movements separately from the head movements, we seek for a representation of facial landmarks which is robust to head pose changes. Accordingly, we use the barycentric representation proposed in Section 4.2 to filter out the head pose changes.

Given an ordered list of moving landmarks, $Z_1(t) = (x_1(t), y_1(t)), \ldots, Z_n(t) = (x_n(t), y_n(t))$, we assume that $Z_1(t)$, $Z_2(t)$, and $Z_3(t)$ are the vertices of a non-degenerate triangle *for every value of* t. Here, the right and left corners of the eyes and the tip of the nose are chosen to form a non-degenerate triangle (see the red triangle in Fig. 4.1). As explained

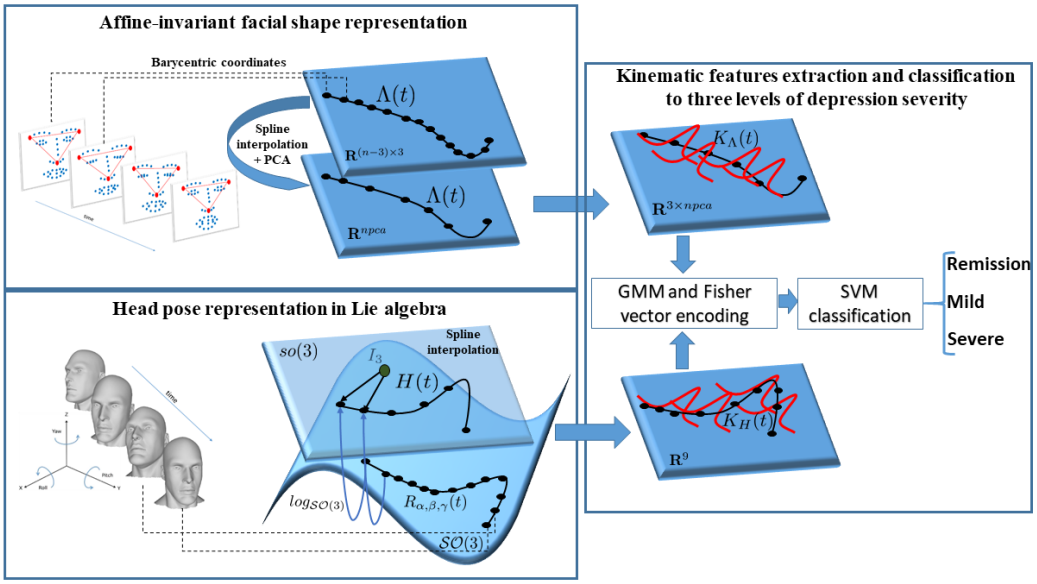

FIGURE 4.4 – Overview of the proposed approach (Depression severity level assessment).

in Section 4.2, the barycentric coordinates encoded in the $(n-3)\times 3$ matrix

$$\Lambda(t) = \begin{pmatrix} x_4(t) & y_4(t) & 1 \\ \vdots & \vdots & \vdots \\ x_n(t) & y_n(t) & 1 \end{pmatrix} \begin{pmatrix} x_1(t) & y_1(t) & 1 \\ x_2(t) & y_2(t) & 1 \\ x_3(t) & y_3(t) & 1 \end{pmatrix}^{-1}$$

form an affine invariant shape representation of the moving landmarks for all $i = 4, .., n$ and at any time t.

4.4.2 Head movements analysis in Lie algebra

Head movements correspond to head nods (*i.e.,* pitch), head turns (*i.e.,* yaw), and lateral head inclinations (*i.e.,* roll) (see Fig. 4.5). Given a time series of the 3 degrees of freedom of out-of-plane rigid head movement, at any time t the yaw is defined as a counterclockwise rotation of $\alpha(t)$ about the z-axis. The corresponding time-dependent rotation matrix is given

by

$$R_\alpha(t) := \begin{pmatrix} \cos(\alpha(t)) & -\sin(\alpha(t)) & 0 \\ \sin(\alpha(t)) & \cos(\alpha(t)) & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

Pitch is a counterclockwise rotation of $\beta(t)$ about the y-axis. The rotation matrix is given by

$$R_\beta(t) := \begin{pmatrix} \cos(\beta(t)) & 0 & \sin(\beta(t)) \\ 0 & 1 & 0 \\ -\sin(\beta(t)) & 0 & \cos(\beta(t)) \end{pmatrix} .$$

Roll is a counterclockwise rotation of $\gamma(t)$ about the x-axis. The rotation matrix is given by

$$R_\gamma(t) := \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\gamma(t)) & -\sin(\gamma(t)) \\ 0 & \sin(\gamma(t)) & \cos(\gamma(t)) \end{pmatrix} .$$

A single rotation matrix can be formed by multiplying the yaw, pitch, and roll rotation matrices to obtain

$$R_{\alpha,\beta,\gamma}(t) = R_\alpha(t) R_\beta(t) R_\gamma(t) . \tag{4.4.1}$$

The obtained time-parametrized curve $R_{\alpha,\beta,\gamma}(t)$ encodes head pose at each time t and lie on a non-linear manifold called the special orthogonal group. The special orthogonal group $\mathcal{SO}(3)$ is a matrix Lie group formed by all rotations about the origin of three-dimensional Euclidean space $\mathbb{R}^3$ under the operation of composition [17]. The tangent space at the identity $I_3 \in \mathcal{SO}(3)$ is a three-dimensional vector space, called the Lie algebra of $\mathcal{SO}(3)$ and is denoted by $\mathsf{so}(3)$. Following [124, 123], we overcome the non-linearity of the space of our representation (*i.e.*, $\mathcal{SO}(3)$), and map the curve $R_{\alpha,\beta,\gamma}(t)$ from $\mathcal{SO}(3)$ to $\mathsf{so}(3)$ using the logarithm map $\log_{\mathcal{SO}(3)}$ to obtain the three-dimensional curve

$$H(t) = \log_{\mathcal{SO}(3)}(I_3, R_{\alpha,\beta,\gamma}(t)) = \log(R_{\alpha,\beta,\gamma}(t)) , \tag{4.4.2}$$

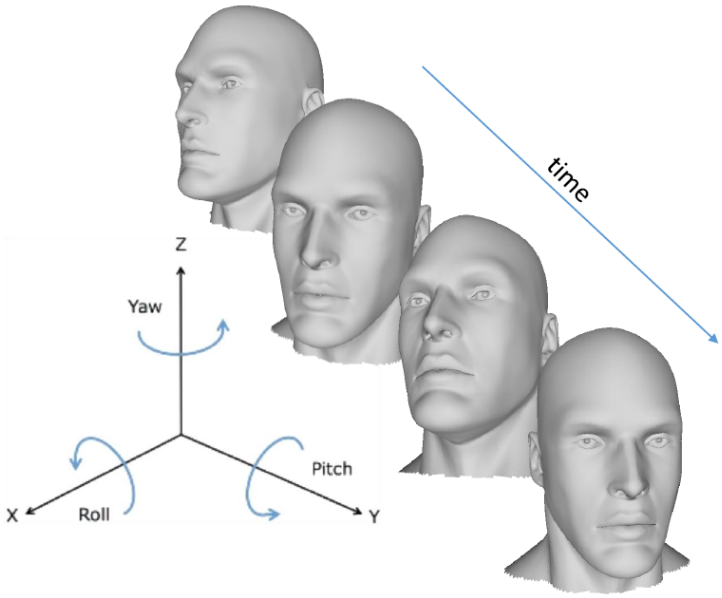

FIGURE 4.5 – Example of the automatically tracked 3 degrees of freedom of head pose

lying on $\mathfrak{so}(3)$. Fore more details about the special orthogonal group, the logarithm map, and the lie algebra, readers are refereed to [124, 123, 17]. In the following, the time series of the 3 degrees of freedom of rigid head movement are represented using the three dimensional curve $H(t)$.

4.4.3 Kinematic features and Fisher vector encoding

To characterize facial and head movement dynamics, we derive appropriate kinematic features based on their proposed representations $\Lambda(t)$ and $H(t)$, respectively.

4.4.3.1 Kinematic features

Because videos of clinical interviews vary in length, the extracted facial and head curves (of different videos) varies in length. The variation in the obtained curves' lengths may introduce distortions in the feature extraction step. To overcome this limitation, we apply

a cubic spline interpolation to the obtained $\Lambda(t)$ and $H(t)$ curves, resulting in smoother, shorter, and fixed length curves. We set empirically the new length of the curve given by spline interpolation to 5000 samples for both facial and head curves.

Usually, the number of landmark points given by recent landmark detectors vary from 40 to 70 points. By building the barycentric coordinates of the facial shape as explained in Section 4.2, this results in high-dimensional facial curves $\Lambda(t)$ with static observations of dimension 120 at least (it can reach 200 if we have 70 landmark points per face). To reduce the dimensionality of the facial curve $\Lambda(t)$, we perform a Principal Component Analysis (PCA) that accounts for 98% of the variance to obtain new facial curves with dimension 20. Then, we compute the velocity $V_\Lambda(t) = \frac{\partial \Lambda(t)}{\partial t}$ and the acceleration $A_\Lambda(t) = \frac{\partial^2 \Lambda(t)}{\partial t^2}$ from the facial sequence $\Lambda(t)$ after reducing its dimension. Finally, facial shapes, velocities, and accelerations are concatenated to form the curve

$$K_\Lambda(t) = [\Lambda(t); V_\Lambda(t); A_\Lambda(t)] \ , \tag{4.4.3}$$

Because head curve $H(t)$ is only three-dimensional no need for data reduction. Velocities and accelerations are directly computed from the head sequence $H(t)$ and concatenated with head pose values to obtain the final nine-dimensional curve

$$K_H(t) = [H(t); V_H(t); A_H(t)] \ . \tag{4.4.4}$$

The curves $K_\Lambda(t)$ and $K_H(t)$ denote the kinematic features over time of the facial and head movements, respectively.

4.4.3.2 Fisher vector encoding

Our goal is to obtain a single vector representation from the kinematic curves $K_\Lambda(t)$ and $K_H(t)$ for depression severity assessment. Following [35], we used the Fisher Vector representation using a Gaussian mixture model (GMM) distributions [148]. Assuming that the observations of a single kinematic curve are statistically independent, a GMM with c

components is computed for each kinematic curve by optimizing the maximum likelihood (ML) criterion of the observations to the c Gaussian distributions. In order to encode the estimated Gaussian distributions in a single vector representation, we use the convenient improved fisher vector encoding which is suitable for large-scale classification problems [92]. This step is performed for kinematic curves $K_\Lambda(t)$ and $K_H(t)$, separately. The number of Gaussian distributions c are chosen by a a leave-one-subject-out cross-validation and are set to 14 for kinematic facial curves and to 31 for kinematic head curves resulting in fisher vectors with dimension $14 \times 20 \times 3 \times 2 = 1680$ for facial movement dynamics and vectors with dimension $31 \times 3 \times 3 \times 2 = 558$ for head movement dynamics.

4.4.4 Assessment of depression severity level

After extracting the fisher vectors from the kinematic curves, the facial and head movements are represented by compact vectors that describe the dynamics of facial and head movements, respectively. To reduce redundancy and select the most discriminative feature set, the Min-Redundancy Max-Relevance (mRMR) algorithm [90] was used for feature selection. The set of selected features are then fed to a multi-class SVM with a Gaussian kernel to classify the extracted facial and head movement dynamics into different depression severity levels. Please note that a leave-one-subject-out cross-validation is performed to choose the number of selected features by mRMR which is set to 726 for facial movement dynamics and to 377 for head movement dynamics.

For an optimal use of the information given by the facial and head movements, depression severity was assessed by late fusion of separate SVM classifiers. This is done by multiplying the probabilities $s_{i,j}$, output of the SVM for each class j, where $i \in \{1, 2\}$ denotes the modality (*i.e.,* facial and head movements). The class $\mathcal{C}$ of each test sample is determined by

$$\mathcal{C} = \arg\max_{j} \prod_{i=1}^{2} s_{i,j},\ j = 1, \ldots, n_{\mathcal{C}}\ \ , \tag{4.4.5}$$

where $n_{\mathcal{C}}$ is the number of classes (*i.e.,* depression severity levels).

4.4.5 Experimental evaluation

4.4.5.1 Dataset

In order to evaluate the proposed approach, we conducted our experiments on the dataset used in [35, 3, 27]. In this dataset, Fifty-seven depressed participants (34 women, 23 men) were recruited from a clinical trial for treatment of depression. At the time of the study, all met DSM-4 criteria [44] for Major Depressive Disorder (MDD). Data from 49 participants was available for analysis. Participant loss was due to change in original diagnosis, severe suicidal ideation, and methodological reasons (*e.g.,* missing audio or video). Symptom severity was evaluated on up to four occasions at 1, 7, 13, and 21 weeks post diagnosis and intake by four clinical interviewers (the number of interviews per interviewer varied). Interviews were conducted using the Hamilton Rating Scale for Depression (HRSD) [52]. HRSD is a clinician-rated multiple item questionnaire to measure depression severity and response to treatment. HRSD scores of 15 or higher are generally considered to indicate moderate to severe depression; scores between 8 and 14 indicate mild depression; and scores of 7 or lower indicate remission [46]. Using these cut-off scores, we defined three ordinal depression severity classes: moderate to severe depression, mild depression, and remission (*i.e.,* recovery from depression). The final sample was 126 sessions from 49 participants: 56 moderate to severely depressed, 35 mildly depressed, and 35 remitted (for a more detailed description of the data please see [35]).

4.4.5.2 Results

We seek to discriminate three levels of depression severity from facial and head movement dynamics separately and in combination. To do so, we used leave-One-Subject-Out cross validation scheme. Performance was evaluated using two criterion. One was the mean

TABLE 4.3 – Classification Accuracy (%) - Comparison with State-of-the-art

Method	Modality	Acc (%)	W. Kappa
J. Cohn *et al.* [27]	Facial movements	59.5	0.43
S. Alghowinem *et al.* [3]	Head movements	53.0	0.42
Dibeklioglu *et al.* [36]	Facial movements	64.98	0.50
Dibeklioglu *et al.* [36]	Head movements	56.06	0.40
Dibeklioglu *et al.* [35]	Facial movements	72.59	0.62
Dibeklioglu *et al.* [35]	Head movements	65.25	0.51
Dibeklioglu *et al.* [35]	Facial/Head movements	**77.77**	**0.71**
Ours	**Facial movements**	**66.19**	**0.60**
Ours	**Head movements**	**61.43**	**0.54**
Ours	**Facial/Head movements**	**70.83**	**0.65**

TABLE 4.4 – Confusion matrix of depression severity level assessment

	Remission	Mild	Severe
Remission	**60.0**	31.42	8.57
Mild	20.0	**68.57**	11.42
Severe	1.78	14.28	**83.92**

accuracy over the three levels of severity. The other was weighted kappa [26]. Weighted kappa is the proportion of ordinal agreement above what would be expected to occur by chance [26].

Consistent with prior work [35], average accuracy was higher for facial movement than for head movement. Facial movement was 66.19%, and head movement was 61.43% (see Table 4.3). When the two modalities were combined, average accuracy increased to 70.83%.

Misclassification was more common between adjacent categories (e.g., Mild and Remitted) than between distant categories (e.g., Remitted and Severe) (Table 4.4). Highest accuracy was found for the difference between severe and mild depression (83.92%).

Evaluation of the system components. To evaluate our approach to encoding

TABLE 4.5 – Evaluation of the Steps to the Proposed Approach - Depression severity level assessment

Facial shapes representation	**Accuracy (%)**
Pose normalization (Procrustes)	63.69
Barycentric coordinates	**66.19**
Head pose representation	**Accuracy (%)**
Angles head pose representation	59.05
Lie algebra head pose representation	**61.43**
Impact of spline interpolation	**Accuracy (%)**
Without spline interpolation	60.36
With spline interpolation	**70.83**
Impact of PCA on facial movements	**Accuracy (%)**
Without PCA	56.19
With PCA	**66.19**
Impact of feature selection (mRMR)	**Accuracy (%)**
Without feature selection	62.50
With feature selection	**70.83**
Classifiers	**Accuracy (%)**
Logistic regression	62.02
Multi-class SVM	**70.83**

movement dynamics of face and head movement with alternative representations. For facial movement dynamics, we compared the barycentric representation with a Procrustes representation. Average accuracy using Procrustes was 3% lower than that for barycentric representation (Table 4.5). For head movements, we compared the Lie algebra representation to a vector representation formed by the yaw, roll, and pitch angles. Accuracy decreased by about 2% in comparison with the proposed approach.

To evaluate whether dimensionality reduction using PCA together with spline interpolation improves accuracy, we compared results with and without PCA and spline interpolation. Omitting PCA and spline interpolation decreased accuracy by about 10%.

To evaluate whether mRMR feature selection and choice of classifier contributed to accuracy, we compared results with and without use of a feature selection step for both Multi-SVM with logistic regression classifiers. When mRMR feature selection was omitted, accuracy decreased by about 8%. Similarly, when logistic regression was used in place of Multi-SVM, accuracy decreased by about 7%. This result was unaffected by choice of kernel.

Thus, use of the any of the proposed alternatives would have decreased accuracy relative to the proposed method.

4.4.6 Interpretation and discussion

In this section, we evaluate the interpretability of the proposed kinematic features (that is, $K_\Lambda(t)$ and $K_H(t)$ defined in Eq. 4.4.3 and Eq. 4.4.4) for depression severity detection. We compute the l2-norm of velocity and acceleration intensities for the face (*i.e.,* $V_\Lambda(t)$ and $A_\Lambda(t)$) and head (*i.e.,* $V_H(t)$ and $A_H(t)$) curves for each video. Since each video is analyzed independently, we compute the histograms of the velocity and acceleration intensities over 10 samples (videos) from each level of depression severity. This results in histograms of 50000 velocity and acceleration intensities for each depression level.

Fig. 4.6 shows the histograms of facial and head velocity (top part) and acceleration

(bottom part) intensities. Results for face are presented in the left panel and those for head in the right panel. For face, the level of depression severity is inversely proportional to the velocity and acceleration intensities. Velocity and acceleration both increased as participants improved from severe to mild and then to remitted. This finding is consistent with data and theory in depression.

Head motion, on the other hand, failed to vary systematically with change in depression severity (Fig. 4.6). This finding was in contrast to previous work. Girard and colleagues [49] found that head movement velocity increased when depression severity decreased. A possible reason for this difference may lie in how head motion was quantified. Girard [49] quantified head movement separately for pitch and yaw; whereas we combined pitch, yaw, and also roll. By combining all three directions of head movement, we may have obscured the relation between head movement and depression severity.

The proposed method detected depression severity with moderate to high accuracy that approaches that of state of the art [35]. Beyond the state of the art, the proposed method yields interpretable findings. The proposed dynamic features strongly mapped onto depression severity. When participants were depressed, their overall facial dynamics were dampened. When depression severity lessened, participants became more expressive. In remission, expressiveness was even higher. These findings are consistent with the observation that psychomotor retardation in depression lessens as severity decreases. Stated otherwise, people more expressive with return to normal mood.

4.5 Conclusion

In this chapter, we proposed a novel affine-invariant representation of 2D facial landmark sequences based on their barycentric coordinates. While being closely related to the conventional Grassmann representation, the latter has the advantage of lying in Euclidean space avoiding the non-linearity problem encountered in Grassmann manifold.

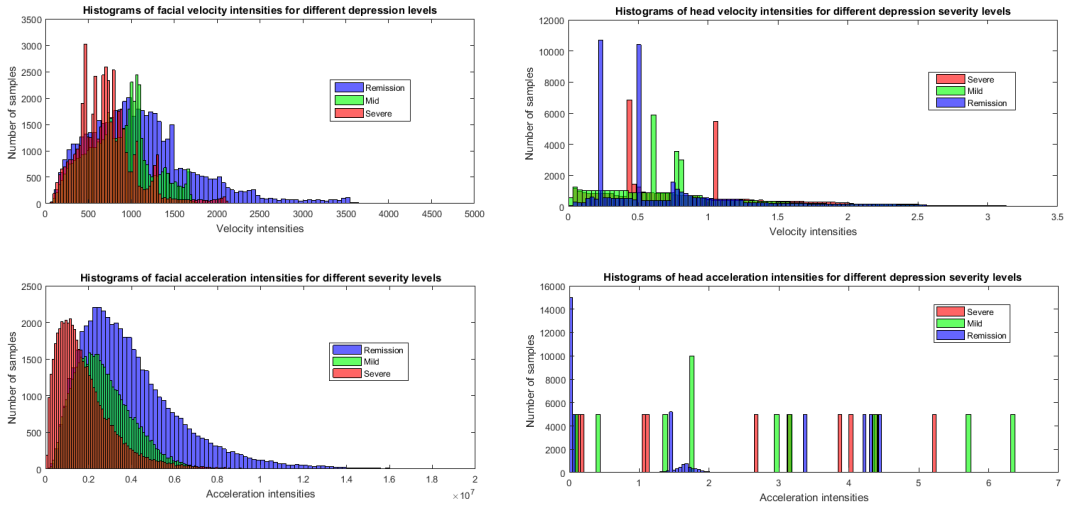

FIGURE 4.6 – Histograms of velocity and acceleration intensities for facial (left) and head (right) movements. Psychomotor retardation symptom is well captured by the introduced kinematic features, especially with those computed from the facial movements.

Applications of the proposed representation have been shown in facial expression recognition in unconstrained environments and depression severity level assessment. In facial expression recognition task, a metric learning was adopted on the barycentric representations to better discriminate between static observations, then a pipeline of DTW and ppfSVM was used with the learned metric for facial sequence classification. For the assessment of depression severity level, kinematic features (*i.e.*, velocities and accelerations) were derived from the barycentric representation and encoded using GMM and Fisher vector encoding. As far as head poses are concerned in depression, we proposed a head pose representation in Lie algebra and applied the same pipeline as for barycentric representation (*i.e.*, kinematic features extraction, GMM and Fisher vector encoding). Finally, SVM is adopted to classify separately and combined the fisher vectors from barycentric and lie algebra representations. The experimental results showed that the proposed approaches achieved comparable performance with sate-of-the-art methods in both facial expression recognition and depression severity level assessment.

Chapitre 5

Conclusion and Future study

5.1 Conclusions and limitations

In this thesis, we proposed novel geometric tools for human behavior understanding based on the analysis of human landmark sequences. Firstly, we proposed a novel geometric framework on Gram matrix trajectories. To overcome the non-linear nature of the space of Gram matrices, its Riemannian geometry was studied to derive suitable analyzing tools for the Gram matrix trajectories. Applications were shown to facial expression recognition from 2D landmarks tracked on the human face in RGB videos, 3D action recognition from 3D skeletons detected on the human body in depth streams, and 3D emotion recognition from body movements captured by motion capture systems. Secondly, we proposed an affine-invariant representation for the specific case of 2D facial landmarks based on their barycentric coordinates. While being related to the Gram matrix representation, the barycentric representation has the advantage of lying in Euclidean space where standard computational and machine learning tools are applicable. The barycentric representation was evaluated in facial expression recognition by applying a standard metric learning algorithm, and in depression severity level assessment by deriving kinematic features along with standard

features encoding techniques.

While powerful, landmark based methods rely on the performance of landmark detectors. If the landmark detector provides inaccurate estimations, this will definitely harm the performance of landmark based solutions for human behavior understanding. Another limitation for using only landmark points is the possible loss of information. Indeed, landmark detectors provide a set of key points on the human face or body which could discard relevant information about the problem at hand. For instance, Fear expression was the most challenging expression in all our experiments since it involves several action unit activations (*i.e.*, AU1+AU2+AU4+AU5+AU7+AU20+AU26) [47] that are quite difficult to detect by using only landmark points.

Moreover, in this thesis we only studied classification tasks (*e.g.*, action or expression classification). That is to say, given a landmark sequence we only focused on classifying into predefined categories (*e.g.*, joy, fear, etc.). However, in some real human related application, one needs to provide a quantity within a fixed interval. For example, for the specific task of pain intensity estimation from faces [146], we should provide a value for each sequence indicating the pain intensity.

5.2 Towards geometry guided deep covariance descriptors for facial expression recognition

Correspondingly to the limitations mentioned in the previous Section, we investigated the use of appearance based methods for static facial expression recognition in collaboration with our colleague Naima Otberdout (PhD student in Mohammed V University of Rabat).

Recently, Deep Convolutional Neural Networks (DCNNs) achieved impressive performance in such tasks. The idea here is to make the network learn the best features from large collections of data during a training phase. However, one drawback of DCNNs is that they

do not take into account the spatial relationships within the face. To overcome this issue, we propose to exploit globally and locally the network features extracted in different regions of the face. This yields a set of DCNN features per region. The question is how to encode them in a compact and discriminative representation for a more efficient classification than the one achieved globally by classical softmax. We propose to encode face DCNN features in a covariance matrix. These matrices have shown to outperform first-order features in many computer vision tasks [116, 117, 84]. In doing this, we exploit the space geometry of the covariance matrices as points on the symmetric positive definite (SPD) manifold. Furthermore, we use a valid positive definite Gaussian RBF kernel on this manifold to train a SVM classifier for expression classification.

Specifically, we start by encoding the facial expression into Feature Maps (FMs) extracted using DCNNs. Here, we use two DCNN models, namely, *VGG-face* [88] and *ExpNet* [37]. A covariance descriptor is then computed over these FMs and is considered for global face representation. We also extract four regions on the input face image around the eyes, mouth, and cheeks (left and right) using a facial landmark detector. By mapping these regions on the extracted deep FMs, we are able to extract local regions in these FMs that bring more accurate information about the facial expression. A local covariance descriptor is also computed for each local region. A RBF kernel endowed with the Log-Euclidean Riemannian metric (LERM) [7] which has been proved to be positive definite [62] is employed for SVM classification. Note that we consider a late fusion of the local and global covariance descriptors by computing a weighted sum of the scores given by the classifier for each region.

Overall, the proposed solution permits us to combine the geometric and appearance features enabling an effective description of facial expressions at different spatial levels, while taking into account the spatial relationships within the face. An overview of the proposed solution is illustrated in Fig. 5.1. The effectiveness of the proposed approach in recognizing basic facial expressions has been evaluated in constrained and unconstrained (*i.e.*, in-the-wild) settings using two publicly available datasets with different challenges:

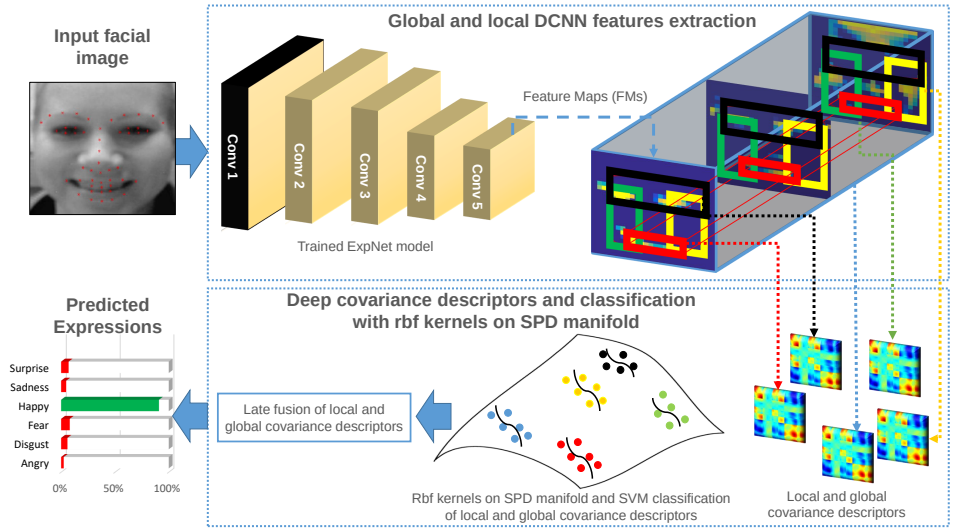

FIGURE 5.1 – Overview of the proposed method.

Oulu-CASIA dataset [143]: Includes 480 image sequences of 80 subjects taken in a constrained environment with normal illumination conditions. For both training and testing, we use the last three peak frames to represent the video resulting in 1440 images. Following the same setting of the state-of-the-art, we conducted a ten-fold cross validation experiment, with subject independent splitting.

Static Facial Expression in the Wild (SFEW) dataset [34]: Consists of $1,322$ static images labeled with seven facial expressions (the six basic plus the neutral one). This dataset has been collected from real movies and targets spontaneous expression recognition in challenging, *i.e.*, in-the-wild, environments. It is divided into training (891 samples), validation (431 samples), and test sets. Since the test labels are not available, here we report results on the validation data.

As initial step, we performed some preprocessing on the images of both datasets. For Oulu-CASIA, we first detected the face using the method proposed in [126]. For SFEW, we used the aligned faces provided by the dataset. Then, in order to detect the facial regions we detected 49 facial landmarks on each face using the Chehra Face Tracker [8]. All frames

were cropped and resized to 224×224, which is the input size of the DCNN models.

In Table 5.1, we compare our proposed global (G-FMs) and local (R-FMs) solutions with the baselines provided by the VGG-face and ExpNet models, without the use of the covariance matrix (*i.e.*, they used the fully connected and softmax layers). On Oulu-CASIA, the G-FMs solution improves by 3.7% and 1.26%, respectively, the VGG-face and ExpNet models. Though less marked, an increment of 0.69% for the VGG-face and of 0.92% for ExpNet has been also obtained on the SFEW dataset. These results prove that the covariance descriptors computed on the convolutional features provide more discriminative representations. Furthermore, the classification of these representations using Gaussian kernel on SPD manifold is more efficient than the standard classification with fully connected layers and softmax, even if these layers were trained in an end-to-end manner. Table 5.1 also shows that the fusion of the local (R-FMs) and global (G-FMs) approaches achieves a clear superiority on the Oulu-CASIA dataset surpassing by 1.25% the global approach, while no improvement is observed on the SFEW dataset. This is due to the failure of landmark detection skewing the extraction of the local deep features.

Dataset	**Model**	**FC-Softmax**	**ours (G-FMs)**	**ours (G-FMs and R-FMs)**
Oulu-CASIA	*VGG Face*	77.8	81.5	–
	ExpNet	82.29	**83.55**	**84.80**
SFEW	*VGG Face*	46.66	47.35	–
	ExpNet	48.26	**49.18**	**49.18**

TABLE 5.1 – Comparison of the proposed classification scheme with respect to the VGG-Face and ExpNet models with fully connected layer and Softmax.

For more details about the method and the conducted experiments, readers are referred to [86].

5.3 Future works

As future works, we aim to investigate the following points:

— In this thesis, we proposed two representations of 2D/3D human landmarks which are robust to view variations. The Gram representation introduced in chapter 3 was invariant to Euclidean transformations, while the barycentric representation presented in chapter 4 was invariant to affine transformations. However, the view variations for 2D landmarks result in projective transformations as stated in Section 4.2 of chapter 4. Future works may include the study of filtering out these complex projective transformations for a more robust representation of 2D landmarks to view variations especially in unconstrained (in-the-wild) environments.

— Recently, Deep Learning (DL) became one of the most successful solutions in many Computer Vision tasks. However, research on DL techniques has mainly focused so far on data defined on Euclidean domains. In this thesis, we were confronted to the problem of non-linearity of data representations (*e.g.*, space-time shape representations on non-linear manifolds). Other examples of non-linear representations include dynamical systems, covariance matrices, and subspace representations. The adoption of conventional DL techniques on these data representations is not straightforward and require adapting optimization techniques to effectively work on the underlying manifold. For instance, in order to conduct an end-to-end classification of the deep covariance descriptors introduced in Section 5.2 instead of using SVM classifier, one should adapt the FC-Softmax to effectively work on the manifold of positive definite matrices. Some recent findings in this direction have show that adapting DL techniques to manifold valued data is possible [60, 59, 58, 20].

— For some human related real applications, we need to anticipate the human behavior rather than understanding it. A relevant example of this is given by autonomous driving systems which should anticipate the behavior of the pedestrians in order to avoid accidents especially when the car is going fast. In this thesis, we only studied classification problems of human behaviors but it would be interesting to investigate the prediction of human behaviors in order to anticipate them [75].

— In the context of facial expression recognition, this thesis mainly focused on recognizing posed basic facial expressions which are not naturally linked to the emotional state of the test subject [102]. Future works may include the study of spontaneous and authentic facial expressions [102, 142].

Bibliographie

[1] Mohamed F. Abdelkader, Wael Abd-Almageed, Anuj Srivastava, and Rama Chellappa. Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. *Computer Vision and Image Understanding*, 115(3):439–455, March 2011.

[2] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Riemannian geometry of grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, 80(2):199–220, 2004.

[3] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parkerx, and Michael Breakspear. Head pose and movement analysis as an indicator of depression. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 283–288, 2013.

[4] Salah Althloothi, Mohammad H Mahoor, Xiao Zhang, and Richard M Voyles. Human activity recognition using multi-features and multiple kernel learning. *Pattern recognition*, 47(5):1800–1812, 2014.

[5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[6] R. Anirudh, P. Turaga, J. Su, and A. Srivastava. Elastic functional coding of riemannian trajectories. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 39(5):922–936, May 2017.

[7] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 56(2):411–421, 2006.

[8] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1859–1866, 2014.

[9] Mohammad Ali Bagheri, Qigang Gao, and Sergio Escalera. Support vector machines with time series distance kernels for action classification. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 1–7, 2016.

[10] Djordje Baralić. How to understand grassmannians? *The Teaching of Mathematics*, pages 147–157, 2011.

[11] AT Beck, CH Ward, M Mendelson, J Mock, and J Erbaugh. An inventory for measuring. *Archives of general psychiatry*, 4:561–571, 1961.

[12] Evgeni Begelfor and Michael Werman. Affine invariance revisited. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2087–2094, 2006.

[13] Boulbaba Ben Amor, Jingyong Su, and Anuj Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(1):1–13, 2016.

[14] Marcel Berger. Geometry, vol. i-ii, 1987.

[15] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[16] S. Bhattacharya, N. Souly, and M. Shah. Covariance of Motion and Appearance Features for Spatio Temporal Recognition Tasks. *ArXiv e-prints*, June 2016.

[17] Mary L Boas. *Mathematical methods in the physical sciences*. Wiley, 2006.

[18] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[19] Silvere Bonnabel and Rodolphe Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1070, 2009.

[20] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[21] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, volume 1, page 4, 2017.

[22] Judith Bütepage, Michael J Black, Danica Kragic, and Hedvig Kjellström. Deep representation learning for human motion prediction and classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2017. IEEE, 2017.

[23] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107:177–190, 2014.

[24] Jacopo Cavazza, Andrea Zunino, Marco San Biagio, and Vittorio Murino. Kernelized covariance for action recognition. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 408–413. IEEE, 2016.

[25] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2(3):27, 2011.

[26] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[27] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction*, pages 1–7, 2009.

[28] Marco Cuturi. Fast global alignment kernels. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 929–936, 2011.

[29] Mohamed Daoudi, Stefano Berretti, Pietro Pala, Yvonne Delevoye, and Alberto Bimbo. Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices. In *Int. Conf. on Image Analysis and Processing*, to appear 2017.

[30] Maxime Devanne, Hazem Wannous, Stefano Berretti, Pietro Pala, Mohamed Daoudi, and Alberto Del Bimbo. 3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold. *IEEE Trans. on Cybernetics*, 45(7):1340–1352, 2015.

[31] Michel Marie Deza and Monique Laurent. *Geometry of cuts and metrics*, volume 15. Springer, 2009.

[32] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge (EmotiW) challenge and workshop summary. In *Int. Conf. on Multimodal Interaction, (ICMI)*, pages 371–372, 2013.

[33] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, 2012.

[34] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In

ACM Int. Conf. on Multimodal Interaction, pages 423–426. ACM, 2015.

[35] Hamdi Dibeklioglu, Zakia Hammal, and Jeffrey F Cohn. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics*, 2017.

[36] Hamdi Dibeklioglu, Zakia Hammal, Ying Yang, and Jeffrey F. Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*, pages 307–310, 2015.

[37] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. In *IEEE Int. Conf. on Automatic Face Gesture Recognition (FG)*, pages 118–126, 2017.

[38] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pages 579–583. IEEE, 2015.

[39] Yong Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, June 2015.

[40] Paul Ekman, Wallace V Freisen, and Sonia Ancoli. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125, 1980.

[41] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98, 1969.

[42] S. Elaiwat, Mohammed Bennamoun, and Farid Boussaïd. A spatio-temporal rbm-based model for facial expression recognition. *Pattern Recognition*, 49:152–161, 2016.

[43] Masoud Faraki, Mehrtash T Harandi, and Fatih Porikli. Image set classification by symmetric positive semi-definite matrices. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 1–8, 2016.

[44] Michael B First, Robert L Spitzer, Miriam Gibbon, and Janet BW Williams. *Structured clinical interview for DSM-IV axis I disorders - Patient edition (SCID-I/P, Version 2.0)*. Biometrics Research Department, New York State Psychiatric Institute, New York, NY, 1995.

[45] Hans-Ulrich Fisch, Siegfried Frey, and Hans-Peter Hirsbrunner. Analyzing nonverbal behavior in depression. *Journal of abnormal psychology*, 92(3):307, 1983.

[46] Jay C Fournier, Robert J DeRubeis, Steven D Hollon, Sona Dimidjian, Jay D Amsterdam, Richard C Shelton, and Jan Fawcett. Antidepressant drug effects and depression severity: A patient-level meta-analysis. *Journal of the American Medial Association*, 303(1):47–53, 2010.

[47] Wallace V Friesen, Paul Ekman, et al. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2(36):1, 1983.

[48] Guillermo Garcia-Hernando and Tae-Kyun Kim. Transition forests: Learning discriminative temporal transitions for action recognition and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 432–440, 2017.

[49] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647, 2014.

[50] Thore Graepel, Ralf Herbrich, Peter Bollmann-Sdorra, and Klaus Obermayer. Classification on pairwise proximity data. *Advances in Neural Information Processing Systems*, pages 438–444, 1999.

[51] Steinn Gudmundsson, Thomas Philip Runarsson, and Sven Sigurdsson. Support vector machines and dynamic time warping for time series. In *IEEE World Congress on*

Computational Intelligence, pages 2772–2776, 2008.

[52] Max Hamilton. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56–61, 1960.

[53] Halim Hicheur, Hideki Kadone, Julie Grèzes, and Alain Berthoz. The combined role of motion-related cues and upper body posture for the expression of emotions during human walking. In *Modeling, Simulation and Optimization of Bipedal Walking*, pages 71–85. Springer Berlin Heidelberg, 2013.

[54] Nicholas J Higham. Computing the polar decomposition with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986.

[55] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5344–5352, 2015.

[56] Jian-Fang Hu, Wei-Shi Zheng, Lianyang Ma, Gang Wang, and Jianhuang Lai. Real-time RGB-D activity prediction by soft regression. In *European Conf. on Computer Vision (ECCV)*, pages 280–296, 2016.

[57] Wenbing Huang, Fuchun Sun, Lele Cao, Deli Zhao, Huaping Liu, and Mehrtash Harandi. Sparse coding and dictionary learning with linear dynamical systems. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3938–3947, June 2016.

[58] Zhiwu Huang and Luc J Van Gool. A Riemannian network for spd matrix learning. In *AAAI*, volume 2, page 6, 2017.

[59] Zhiwu Huang, Chengde Wan, Thomas Probst, and Luc Van Gool. Deep learning on lie groups for skeleton-based action recognition. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1243–1252. IEEE computer Society, 2017.

[60] Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2965–2973, 2015.

[61] Suyog Jain, Changbo Hu, and Jake K. Aggarwal. Facial expression recognition with temporal modeling of shapes. In *IEEE Int. Conf. on Computer Vision Workshops (ICCV)*, pages 1642–1649, 2011.

[62] Sadeep Jayasumana, Richard I. Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Tafazzoli Harandi. Kernel methods on riemannian manifolds with gaussian RBF kernels. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(12):2464–2477, 2015.

[63] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3D face alignment from 2D videos for real-time use. *Image and Vision Computing*, 58:13–24, 2017.

[64] Jyoti Joshi, Roland Goecke, Gordon Parker, and Michael Breakspear. Can body expressions contribute to automatic depression analysis? In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–7, 2013.

[65] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *IEEE International Conference on Computer Vision, ICCV*, pages 2983–2991, 2015.

[66] Anis Kacem, Mohamed Daoudi, and Juan-Carlos Alvarez-Paiva. Barycentric Representation and Metric Learning for Facial Expression Recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition* , Xi'an, China, May 2018.

[67] Anis Kacem, Mohamed Daoudi, Boulbaba Ben Amor, and Juan Carlos Alvarez-Paiva. A novel space-time representation on the positive semidefinite cone for facial expression recognition. In *IEEE Int. Conf. on Computer Vision (ICCV)*, October 2017.

[68] Qiuhong Ke, Senjian An, Mohammed Bennamoun, Ferdous Sohel, and Farid Boussaid. Skeletonnet: Mining deep part features for 3-d action recognition. *IEEE Signal Processing Letters*, 24(6):731–735, 2017.

[69] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4570–4579. IEEE, 2017.

[70] Alexander Kläser, Marcin Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conf. (BMVC)*, pages 1–10, 2008.

[71] S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry*, volume 1. Interscience Publishers, 1963.

[72] Piotr Koniusz, Anoop Cherian, and Fatih Porikli. Tensor representations via kernel linearization for action recognition from 3d skeletons. In *European Conference on Computer Vision*, pages 37–53. Springer, 2016.

[73] Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

[74] Christoph H Lampert et al. Kernel methods in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 4(3):193–285, 2009.

[75] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *European Conference on Computer Vision*, pages 689–704. Springer, 2014.

[76] Binlong Li, Octavia I Camps, and Mario Sznaier. Cross-view activity recognition using hankelets. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1362–1369. IEEE, 2012.

[77] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. *Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition*, pages 816–833. Springer Int. Publishing, Cham, 2016.

[78] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian Conf. on Computer Vision*, pages 143–157, 2014.

[79] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1749–1756, 2014.

[80] Zhi Liu, Chenyang Zhang, and Yingli Tian. 3d-based deep convolutional neural network for action recognition with depth sequences. *Image and Vision Computing*, 55:93–100, 2016.

[81] Andras Lorincz, Laszlo Jeni, Zoltan Szabo, Jeffrey Cohn, and Takeo Kanade. Emotional expression classification using time-series kernels. In *Proceedings of the IEEE Conference on computer vision and pattern recognition workshops*, pages 889–895, 2013.

[82] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara Ambadar, and Iain A. Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 94–101, 2010.

[83] Fengjun Lv and Ramakant Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *European conference on computer vision*, pages 359–372. Springer, 2006.

[84] Bingpeng Ma, Yu Su, and Frédéric Jurie. Bicov: a novel image representation for person re-identification and face verification. In *British Machive Vision Conference*, pages 11–pages, 2012.

[85] Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Regression on fixed-rank positive semidefinite matrices: a riemannian approach. *Journal of Machine Learning Research*, 12(Feb):593–625, 2011.

[86] Naima Otberdout, Anis Kacem, Mohamed Daoudi, Lahoucine Ballihi, and Stefano Berretti. Deep covariance descriptors for facial expression recognition. *arXiv preprint arXiv:1805.03869*, 2018.

[87] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas S Huang. Human computing and machine understanding of human behavior: a survey. In *Artifical Intelligence for Human Computing*, pages 47–71. Springer, 2007.

[88] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conf. (BMVC)*, pages 41.1–41.12. BMVA Press, 2015.

[89] Magdalena Pawlyta and Przemysław Skurowski. A survey of selected machine learning methods for the segmentation of raw motion capture data into functional body mesh. In *Information Technologies in Medicine*, pages 321–336. Springer, 2016.

[90] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[91] Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A riemannian framework for tensor computing. *Int. Journal of Computer Vision*, 66(1):41–66, 2006.

[92] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, pages 143–156. Springer, 2010.

[93] Liliana Lo Presti and Marco La Cascia. A novel time series kernel for sequences generated by lti systems. In *Asian Conference on Computer Vision*, pages 433–451. Springer, 2016.

[94] Babette Renneberg, Katrin Heyn, Rita Gebhard, and Silke Bachmann. Facial expression of emotions in borderline personality disorder and depression. *Journal of behavior therapy and experimental psychiatry*, 36(3):183–196, 2005.

[95] Jonathan Rottenberg, James J Gross, and Ian H Gotlib. Emotion context insensitivity in major depressive disorder. *Journal of abnormal psychology*, 114(4):627, 2005.

[96] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[97] Andres Sanin, Conrad Sanderson, Mehrtash T. Harandi, and Brian C. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 103–110, 2013.

[98] Jason Saragih and Roland Goecke. Iterative error bound minimisation for aam alignment. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 1196–1195. IEEE, 2006.

[99] E. Sariyanidi, H. Gunes, and A. Cavallaro. Learning bases of activity for facial expression recognition. *IEEE Trans. on Image Processing*, PP(99):1–1, 2017.

[100] Gary E Schwartz, Paul L Fair, Patricia Salt, Michel R Mandel, and Gerald L Klerman. Facial expression and imagery in depression: an electromyographic study. *Psychosomatic medicine*, 1976.

[101] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Int. Conf. on Multimedia*, pages 357–360, 2007.

[102] Nicu Sebe, Michael S Lew, Yafei Sun, Ira Cohen, Theo Gevers, and Thomas S Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007.

[103] Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti, Alberto Bimbo, and Pietro Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses.

In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 479–485, 2013.

[104] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, June 2016.

[105] Chunhua Shen, Junae Kim, Lei Wang, and Anton Hengel. Positive semidefinite metric learning with boosting. In *Advances in neural information processing systems*, pages 1651–1659, 2009.

[106] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[107] Rim Slama, Hazem Wannous, Mohamed Daoudi, and Anuj Srivastava. Accurate 3D action recognition using learning on the grassmann manifold. *Pattern Recognition*, 48(2):556–567, 2015.

[108] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, pages 4263–4270, 2017.

[109] Anuj Srivastava, Eric Klassen, Shantanu H Joshi, and Ian H Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1415–1428, 2011.

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

[111] Sima Taheri, Pavan Turaga, and Rama Chellappa. Towards view-invariant expression analysis using analytic shape manifolds. In *IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops (FG)*, pages 306–313, 2011.

[112] Amor Ben Tanfous, Hassen Drira, and Boulbaba Ben Amor. Coding kendall's shape trajectories for 3d action recognition. In *IEEE Computer Vision and Pattern Recognition*, 2018.

[113] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[114] Sergey Tulyakov, László A Jeni, Jeffrey F Cohn, and Nicu Sebe. Consistent 3d face alignment. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):2250–2264, 2018.

[115] Pavan K. Turaga, Ashok Veeraraghavan, Anuj Srivastava, and Rama Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Trans. Pattern Analysis Machine Intelligence*, 33(11):2273–2286, 2011.

[116] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *European Conf. on Computer Vision (ECCV)*, pages 589–600, 2006.

[117] Oncel Tuzel, Fatih Porikli, and Peter Meer. Pedestrian detection via classification on Riemannian manifolds. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, Oct. 2008.

[118] M. F. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Int. Conf. on Language Resources and Evaluation, Workshop on EMOTION*, pages 65–70, Malta, May 2010.

[119] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

[120] Bart Vandereycken, P-A Absil, and Stefan Vandewalle. Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank. In *IEEE/SP Workshop on Statistical Signal Processing (SSP)*, pages 389–392, 2009.

[121] A. Veeraraghavan, R. Chellappa, and A.K. Roy-Chowdhury. The function space of an activity. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 959–968, 2006.

[122] V. Veeriah, N. Zhuang, and G. J. Qi. Differential recurrent neural networks for action recognition. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 4041–4049, Dec 2015.

[123] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3D skeletons as points in a Lie group. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 588–595, 2014.

[124] Raviteja Vemulapalli and Rama Chellapa. Rolling rotations for recognizing human actions from 3d skeletal data. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4471–4479, 2016.

[125] Vinay Venkataraman and Pavan K. Turaga. Shape distributions of nonlinear dynamical systems for video-based inference. *CoRR*, abs/1601.07471, 2016.

[126] Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal on Computer Vision*, 57(2):137–154, 2004.

[127] Chunyu Wang, Yizhou Wang, and Alan L Yuille. Mining 3d key-pose-motifs for action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2639–2647, 2016.

[128] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297. IEEE, 2012.

[129] Lei Wang, Jianjia Zhang, Luping Zhou, Chang Tang, and Wanqing Li. Beyond covariance: Feature representation with nonlinear kernel matrices. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4570–4578, 2015.

[130] Limin Wang, Yu Qiao, and Xiaoou Tang. Motionlets: Mid-level 3d parts for human motion recognition. In *2013 IEEE Conf. on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 2674–2681, 2013.

[131] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. Rgb-d-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding*, 2018.

[132] Ruiping Wang, Huimin Guo, Larry S Davis, and Qionghai Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2496–2503, 2012.

[133] Ziheng Wang, Shangfei Wang, and Qiang Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3422–3429, 2013.

[134] Wikipedia. Human behavior, 2011.

[135] James R Williamson, Thomas F Quatieri, Brian S Helfer, Gregory Ciccarelli, and Daryush D Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. ACM, 2014.

[136] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*, pages 20–27. IEEE, 2012.

[137] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013.

[138] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 28–35, 2012.

[139] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203, 2017.

[140] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 148–157, 2017.

[141] Xikang Zhang, Yin Wang, Mengran Gou, Mario Sznaier, and Octavia Camps. Efficient temporal sequence comparison and classification using Gram matrix embeddings on a riemannian manifold. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[142] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[143] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z. Li, and Matti Pietikäinen. Facial expression recognition from near-infrared videos. *Image Vision Computing*, 29(9):607–619, 2011.

[144] Guoying Zhao and Matti Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Analysis and*

Machine Intelligence, 29(6):915–928, 2007.

[145] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N. Metaxas. Learning active facial patches for expression analysis. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2562–2569, 2012.

[146] Jing Zhou, Xiaopeng Hong, Fei Su, and Guoying Zhao. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 84–92, 2016.

[147] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *AAAI Conf. on Artificial Intelligence (AAAI)*, pages 3697–3703, 2016.

[148] Zoran Zivkovic. Improved adaptive Gaussian mixture model for background subtraction. In *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004.*, pages 28–31, 2004.