Shape Analysis for Human Behavior Understanding

Hassen Drira

Université de Lille

mémoire préparé par

Hassen Drira

pour obtenir **L'Habilitation à Diriger des Recherches** en Informatique

soutenue publiquement le 02/07/2020

Shape Analysis for Human Behavior Understanding

Analyse de formes pour la compréhension du comportement humain

---

Composition du jury:

M. Olivier COLOT	Président de jury	Pr., University de Lille, France
Mme Bernadette DORIZZI	Rapporteure	Pr., Télécom SudParis, France
M. Edmond BOYER	Rapporteur	D.R., INRIA Grenoble Rhône-Alpes, France
M. Vittorio MURINO	Rapporteur	Pr., University of Verona, Italy
M. Atilla BASKURT	Examinateur	Pr., INSA Lyon, France
M. Christophe ROSENBERGER	Examinateur	Pr., ENSICAEN, France
M. Mohamed DAOUDI	Garant	Pr., IMT Lille Douai, France

Laboratoire CERI SN, IMT Lille Douai
Laboratoire CRIStAL, université de Lille
Ecole Doctorale Sciences Pour l'Ingénieur

Acknowledgements

Thanks for reading...

Contents

Acknowledgements i

1 Activity Report 1
1.1 Curriculum Vitae . . . 2

2 Introduction 38

3 Action Recognition based on Skeleton Data 42
3.1 Introduction . . . 42
3.2 Background and definitions . . . 43
3.2.1 Sparse representations . . . 44
3.2.2 Riemannian geometry . . . 45
3.2.3 Kendall's shape space . . . 46
The case of planar shapes . . . 47
3.2.3.1 Geometric tools on the manifold . . . 47
Exponential map . . . 48
Logarithm map . . . 48
Intrinsic mean . . . 48
3.2.3.2 Hilbert space embedding of the manifold . . . 48
3.3 Related work on representations of trajectories on Riemannian manifolds . . . 49
3.3.1 Riemannian sparse coding and dictionary learning . . . 49
3.3.2 Trajectory representations on Riemannian manifolds . . . 50
3.4 Human motion modeling framework . . . 52
3.4.1 Intrinsic approach . . . 52
3.4.1.1 Intrinsic Sparse Coding . . . 53
3.4.1.2 Intrinsic Dictionary Learning . . . 54
An efficient dictionary initialization for faster learning . . . 55
3.4.2 Extrinsic approach . . . 56
3.4.2.1 Kernel Sparse Coding . . . 56
3.4.2.2 Kernel Dictionary learning . . . 57
3.5 Properties of the latent space . . . 58
3.5.1 Reconstruction of trajectories . . . 58
3.5.2 Efficient tangent space projections . . . 58
3.5.3 Denoising of skeletal shapes . . . 59
3.5.4 On the vector structure of the latent space . . . 60

3.6 Facial Expression and Action Recognition with Sparse Representations . . . 61
3.6.1 Related work . . . 62
3.6.1.1 3D action recognition . . . 65
3.6.1.2 2D facial expression recognition . . . 66
3.6.2 Temporal modeling and classification . . . 67
3.6.2.1 Dynamic time warping, Fourier pyramid and SVM . . . 68
Temporal re-sampling of trajectories . . . 68
3.6.2.2 Long short-term memory network . . . 69
3.6.2.3 Dictionary structure . . . 69
3.7 Experimental evaluation . . . 70
3.7.1 3D action recognition . . . 70
3.7.1.1 Datasets . . . 70
3.7.1.2 Experimental settings . . . 71
3.7.1.3 Results and discussions . . . 71
Comparison to Riemannian methods on MSR-Action, Florence3D and UTKinect datasets . . . 71
Comparison to State-of-the-art . . . 72
3.7.1.4 Comparison to extrinsic SCDL . . . 74
3.7.2 2D Facial Expression Recognition . . . 75
3.7.2.1 Macro-Expression Recognition . . . 76
Results and discussions . . . 76
3.7.2.2 Micro-Expression Recognition . . . 77
3.7.3 Ablation study . . . 79
3.8 Discussions . . . 81
3.9 Online human-object recognition . . . 82
3.9.1 Object Feature . . . 82
3.9.2 Online action recognition . . . 83
3.10 Conclusion . . . 84

4 3D Face Analysis for gender and expression recognition 85
4.1 Introduction . . . 85
4.2 Optimal Deformations (Dense Scalar Field) . . . 86
4.3 Gender classification using 3D face . . . 91
4.3.1 Introduction . . . 91
4.3.1.1 Related work on 3D-based gender classification . . . 91
4.3.1.2 Methodology and contributions . . . 93
4.3.2 Feature Extraction Methodology . . . 95
4.3.2.1 Face symmetry description . . . 95
4.3.2.2 Face averageness description . . . 96
4.3.3 Gender classification . . . 97
4.3.3.1 Feature Selection . . . 98
4.3.3.2 Random Forest . . . 99
4.3.4 Experiments . . . 100
4.3.4.1 Data preprocessing . . . 100
4.3.4.2 Robustness to variations of age and ethnicity . . . 101
4.3.4.3 Robustness to expression variations . . . 103
4.3.4.4 Comparison with state of the art . . . 104

4.3.5 Random Forest Regression . . . 110
4.3.6 Experiments . . . 111
4.3.6.1 Gender-general experiment . . . 112
4.3.6.2 Gender-specific experiment . . . 113
4.4 Expression Recognition from 3D Dynamic Faces . . . 115
4.4.1 Related Work . . . 116
4.4.2 Our Method and Contributions . . . 117
4.4.3 Geometric Facial Deformation . . . 118
4.4.3.1 Effect of the Nose Tip Localisation Inaccuracy on the DSF Computation 119
4.4.3.2 DSF Compared to other Features . . . 120
4.4.4 Expression Recognition using DSFs . . . 120
4.4.4.1 Mean Shape Deformation with Random Forest Classifier . . . 121
4.4.4.2 3D Motion Extraction with HMM Classifier . . . 122
4.4.5 Experimental Results . . . 126
4.4.5.1 BU-4DFE Database: Description and Preprocessing . . . 126
4.4.5.2 Mean deformation-based Expression Classification . . . 129
4.4.5.3 HMM-based Expression Classification . . . 132
4.4.5.4 Discussion and Comparative Evaluation . . . 134
4.5 Subtle Facial Motions for Effective 4D Expression Recognition . . . 136
4.5.1 Experimental Results . . . 138
4.5.1.1 Dataset Description and Experimental Settings . . . 139
4.5.1.2 Cross-dataset Evaluation on BP4D . . . 141
4.6 Conclusions . . . 141

5 Towards Statistical Analysis of Surfaces 143
5.1 Introduction . . . 143
5.2 Statistical Analysis of 3D Faces . . . 145
5.2.1 Mathematical Framework . . . 145
5.2.2 Applications . . . 148
5.3 Gauge Invariant Framework for Shape Analysis of Surfaces . . . 154
5.3.1 Gauge theory in shape analysis . . . 158
5.3.2 Path straightening for geodesic calculation . . . 160
5.3.3 Discretization of infinite dimensional space . . . 161
5.3.4 Examples of geodesics obtained by path-straightening . . . 162
5.3.5 Classification of 3D shapes . . . 163
5.4 Conclusion . . . 164

6 Ongoing Research and Perspectives 165
6.1 Geometry-aware Deep Learning . . . 167
6.1.1 Shape-GAN for motion generation . . . 167
6.1.2 Towards Deep learning on Shape spaces . . . 168
6.2 Facial emotion recognition in Adverse Conditions . . . 168
6.2.1 Body movement for emotion recognition . . . 169
6.2.2 Towards Extrinsic analysis of 3D shape for subtle expression recognition . . . 169
6.3 Functional Analysis on manifold . . . 170

Bibliography **172**

Dedicated to my parents, my wife and my sons.

Chapter 1

Activity Report

1.1 Curriculum Vitae

Hassen DRIRA, 38 years old, married, 2 children

Associate Professor with IMT Lille Douai,

Member of CRIStAL laboratory UMR CNRS 9189,

Cité scientifique - Rue Guglielmo Marconi

59653 Villeneuve d'Ascq Cedex BP 20145 - FRANCE

Tel: + (33) 3 20 33 64 29

E-mail: hassen.drira@imt-lille-douai.fr

Web: https://sites.google.com/site/hdrirahomepage/

My Google Scholar Profile

Positions

September 2012 - Present — Associate Professor, **IMT Lille Douai**,
Numeric teaching and research department, IMT Lille Douai,
Member of CRIStAL laboratory (UMR CNRS/Lille 9189) – FRANCE.

- **May 2019** — Visiting Professor, Digital Research Center of Sfax http://www.crns.rnrt.tn – Tunisia.
- **April 2013 - May 2013** — Visiting Professor, Statistical Shape Analysis and Modeling Group (SSAMG), Department of Statistics, Florida State University – USA.

September 2011 - August 2012 — Postdoctoral researcher, CRIStAL laboratoy (ex LIFL). University of Lille.
January 2008 - August 2011 — PhD student, university of Lille , CRIStAL laboratory (ex LIFL).

- **January 2010 - August 2010** — Assistant Professor (A.T.E.R.) IMT Lille Douai (ex-Télécom Lille), university of Lille.
- **September 2008 - August 2010** — Assistant Professor (*Moniteur d'initiation à l'enseignement supérieur*), IMT Lille Douai (ex-Télécom Lille).
- **July 2009 - August 2009** — Invited researcher, Statistical Shape Analysis and Modeling Group (SSAMG), Department of Statistics, Florida State University – USA.

September 2007 - January 2008 — Assistant Professor, *Institut Supérieur en Informatique* ISI, university Tunis el Manar – Tunisia.

Education

2008 - 2011 — PhD in Computer Science, with highest honor, university of Lille, France.
Title: Statistical Analysis on manifolds for 3D Face recognition,
Associated project: ANR project FAR3D (Partners: *Ecole Centrale de Lyon*, Eurécom, Thalès, university of Lille).
Committee:
Liming Chen, professor, Ecole Centrale de Lyon, France (president).
Pietro Pala, professor, university of Firenze, Italy (reviewer).
Remco Veltkamp, professor, Utrecht university, Holland (reviewer).
Anuj Srivastava, professor, Florida State University, USA (examiner).
Mohamed Daoudi, professor, IMT Lille Douai, France (Thesis director).
Boulbaba Ben Amor, IMT Lille Douai, France (Co-supervisor).

2005 - 2007 — Master in Computer Science [1], *Ecole Nationale des Sciences de l'Informatique* ENSI, Tunisia.
2003 - 2006 — Engineer in Computer Science, ENSI, Tunisia.

Research Grants (as PI)

International projects

Human behavior in an INTelligent environment: COMINT
Period: 2019-2021.
Type of project: PHC Toubkal; bilateral project with Morocco.
Partners: university of Mohamed V in Rabat Moroco, start up AQUILAE France and University of Technology of Troyes UTT France.
Budget: 40 000 euros.
Role in the project : Principal Investigator and scientific responsible.

Emotionally Intelligent Tutorial System based on facial expression recognition
Period: 2018.
Type of project: franco-tunisian Alliance for Research in Digital Education-PReNum: *Projets de recherche en numérique éducatif.*
Partners: National Institute of Applied Science and Technology INSAT, Tunisie.
Budget: 10 000 euros.
Role in the project: co-PI, scientific responsible, french side.
Principal results of the project: in addition to scientific results, this project funded the conference AICO: Artificial Intelligence with applications on E-learning and digital education COnference http:

[1] This master is proposed for the 40 best engineer students among 220 students.

//www.arts-pi.org.tn/AICO2018/index.php.

National projects

Geometric Deep learning: GeoDeep

Period: 2018.

Type of project: PEPS 2018 *Intelligence Artificielle et Apprentissage Automatique* CNRS INSMI, INS2I -AMIES.

Partners: start up AQUILAE, University of Technology of Troyes UTT, France.

Budget: 7 000 euros.

Role in the project: Principal Investigator (PI)

Principal results of the project: in addition to scientific results, a collaboration with AQUILAE start-up and UTT has been set up thanks to this project (UTT and AQUILAE are partners in PHC Toubkal project 2019-2021). **The COMINT project has been chosen as a success story for the 80th of CNRS, organized by AMIES (*Agence pour les Mathématiques en Intéraction avec l'Entreprise et la société*) https://amies-stories.sciencesconf.org/program.**

Project funded by IMT Lille Douai

PrintHead

Period : 2018.

Type of project: "Performance reserve" IMT Lille Douai.

Partners: Automatic department, computer science department and civil engineering department, IMT Lille Douai.

Budget: 67 000 euros.

Role in the project: responsible, computer science department partner.

Participation to projects

DEFI- 2016-2018

Type of project: PHC Utique; bilateral project with Tunisia.

PI: Boulbaba Ben Amor.

Budget: 32 850 euros.

Role in the project: co-supervision of one PhD student (Nadia Hosni).

3D Face Analyzer- 2011-2014

Type of project: ANR blanc - NSFC, project in collaboration with China and *Ecole centrale de Lyon.*

PI: Mohamed Daoudi.

Workpackage leader: Boulbaba Ben Amor.
Role in the project: participation to project meetings, participation to deliverables writing, collaboration with PhD students within the project and co-signing 3 international journals (J4, J8 and J9) and 6 international conferences (C16, C17, C18, C19, C20 and C21) .

Students Supervision

Ph.D. thesis (3)

former students (2)

2014 - 2016 — **Meng Meng**, *Analysis and recognition of human actions and interactions with objects*,

- Institution: university of Lille.
- Defense: January 9, 2017.
- Funding: Futur & Rupture program, Institut Mines-Télécom.
- Supervision rate: 50%.
- Current situation: Postdoctoral researcher, North Carolina Central University, USA.
- Committee:
  - Reviewer: Stefano Berretti (Professor, university of Firenze, Italy).
  - Reviewer: Djamel Merad (Associate Professor, university Aix-Marseille, France).
  - Thesis examiner: Frédéric Lerasle (Professor, university Paul Sabatier, France).
  - Thesis examiner: Catherine Soladie (Associate professor, Centrale Supélec, France).
  - President: Hichem Snoussi (Professor, University of Technology of Troyes UTT, France).
  - Thesis examiner: Christian Wolf (Associate professor, INSA Lyon, France).
  - Director of thesis: Mohamed Daoudi (Professor, IMT Lille Douai, France).
  - Supervisor: Hassen Drira (Associate professor, IMT Lille Douai, France).
  - Supervisor: Jacques Booneart (Associate professor, IMT Lille Douai, France).

- Abstract: in this thesis, we have investigated the human object interaction recognition by using the skeleton data and local depth information provided by RGB-D sensors. There are two main applications we address in this thesis: human object interaction recognition and abnormal activity recognition. First, we propose a spatio-temporal modeling of human-object interaction videos for

on-line and off-line recognition. In the spatial modeling of human object interactions, we propose low-level feature and object related distance feature which adopted on on-line human object interaction recognition and abnormal gait detection. Then, we propose object feature, a rough description of the object shape and size as new features to model human-object interactions. This object feature is fused with the low-level feature for online human object interaction recognition. In the temporal modeling of human object interactions, we proposed a shape analysis framework based on low-level feature and object related distance feature for full sequence-based off-line recognition. Experiments carried out on two representative benchmarks demonstrate the proposed method are effective and discriminative for human object interaction analysis. Second, we extend the study to abnormal gait detection by using the on-line framework of human object interaction classification. The experiments conducted following state-of-the-art settings on the benchmark shows the effectiveness of proposed method. Finally, we collected a multi-view human object interaction dataset involving abnormal and normal human behaviors by RGB-D sensors. We test our model on the new dataset and evaluate the potential of the pro- posed approach.

- Supervision team: I have supervised this thesis with Mohamed Daoudi, professor at IMT Lille Douai and Jacques Boonaert, Associate professor at IMT Lille Douai.
- Publications co-signed with the student:
  - a paper in Image Vision Computing journal J3 (Impact Factor 2.159)
  - 3 international conference papers (C11, C14 and C15)
- Observation: I participated in the supervision of this thesis and with the departure of Mohamed Daoudi (sabbatical) in 2016 I took more autonomy over the supervision and I co-signed a paper in Image Vision Computer journal (IF 2.159) with the student (Meng Meng, Hassen Drira, Jacques Boonaert: Distances evolution analysis for online and off-line human object interaction recognition. Image Vision Comput. 70: 32-45 (2018)).

November 2016-December 2019 — Amor Ben Tanfous, *Geometric analysis of 3D shapes for the analysis of human behavior*,

- Institution: university of Lille.
- Defense date: December 3, 2019.
- Funding: University of Lille.
- Supervision rate: 50%.
- Current situation: Postdoctoral researcher at Artificial and Natural Intelligence Toulouse Institute ANITI, university of Toulouse, France.

- **Thanks to the thesis's work, the former student Amor Ben Tanfous is awarded with a "Special Mention" at the AFRIF (French branch of the IAPR) PhD thesis prize competition (2019)** https://lnkd.in/gHWDCFg.

- Abstract: Designing intelligent systems to understand video content has been a hot research topic in the past few decades since it helps compensate the limited human capabilities of analyzing videos in an efficient way. In particular, human behavior understanding in videos is receiving a huge interest due to its many potential applications. At the same time, the detection and tracking of human landmarks in video streams has gained in reliability partly due to the availability of affordable RGB-D sensors. This infer time-varying geometric data which play an important role in the automatic human motion analysis. However, such analysis remains challenging due to enormous view variations, inaccurate detection of landmarks, large intra- and inter- class variations, and insufficiency of annotated data. In this thesis, we propose novel frameworks to classify and generate 2D/3D sequences of human landmarks. We first represent them as trajectories in the shape manifold which allows for a view-invariant analysis. However, this manifold is nonlinear and thereby standard computational tools and machine learning techniques could not be applied in a straightforward manner. As a solution, we exploit notions of Riemannian geometry to encode these trajectories based on sparse coding and dictionary learning. This not only overcomes the problem of nonlinearity of the manifold but also yields sparse representations that lie in vector space, that are more discriminative and less noisy than the original data. We study intrinsic and extrinsic paradigms of sparse coding and dictionary learning in the shape manifold and provide a comprehensive evaluation on their use according to the nature of the data (*i.e.* face or body in 2D or 3D). Based on these sparse representations, we present two frameworks for 3D human action recognition and 2D micro- and macro- facial expression recognition and show that they achieve competitive performance in comparison to the state-of-the-art. Finally, we design a generative model allowing to synthesize human actions. The main idea is to train a generative adversarial network to generate new sparse representations that are then transformed to pose sequences. This framework is applied to the task of data augmentation allowing to improve the classification performance. In addition, the generated pose sequences are used to guide a second framework to generate human videos by means of pose transfer of each pose to a texture image. We show that the obtained videos are realistic and have better appearance and motion consistency than a recent state-of-the-art baseline.

  - Reviewer: Alice Caplier (Professor, Grenoble INP, University Grenoble Alpes, France).
  - Reviewer: Sylvain Calinon (Senior Researcher, Idiap Research Institute, Switzerland).
  - President of committee: Bernadette Dorizzi (Professor, Télécom SudParis, France).
  - Thesis examiner: Josef Kittler (Professor, university of Surrey, Britain).
  - Director of thesis: Boulbaba Ben Amor (senior Scientist, Inception Institute of Artificial Intelligence (IIAI) since May 2019 and professor at IMT Lille Douai until April 2019).
  - Supervisor: Hassen Drira (Associate professor, IMT Lille Douai).

- Supervision team: I have supervised this thesis with Boulbaba Ben Amor, Senior Scientist, Inception Institute of Artificial Intelligence (IIAI) since May 2019 and professor at IMT Lille Douai until April 2019).
- Principal publications with the student:
  - a paper has been published in the prestigious journal IEEE Transactions on Pattern Analysis and Machine Intelligence TPAMI J2 (Impact Factor 17.73).
  - a paper has been published in the major conference of computer vision CVPR C3 (rank A*).

Ongoing thesis (1)

January 2017-present — **Nadia Hosni**, *Analysis of geometric functional data for 3D approach recognition*,

- Institution: thesis under joint supervision (*cotutelle*) between university of Manouba, Tunisia and university of Lille, France.
- Expected defense date: December 2020.
- Funding: project PHC Utique DEFI.
- Supervision rate: 30%.
- Supervision team: I co-supervise the thesis with Boulbaba Ben Amor, senior Scientist Inception Institute of Artificial Intelligence (IIAI) since May 2019 and professor at IMT Lille Douai until April 2019, Faten Cheieb, professor at the University of Tunis Carthage and Faouzi Ghorbel, professor at the University of Manouba.
- Principal publications with the student: 1 oral paper C4 at ICPR conference (rank B) the selection rate for oral presentation at ICPR 2018 is 10% and a paper is under revision with CVIU journal J1S.

Participation in thesis supervision within international collaborations (2)

2014 - 2018 — **Raouia Mokni**, *Fusion of the shape and appearance of the palm print for identity recognition*,

- Institution: university of Sfax, Tunisia.
- Defense date: September 21, 2018.

- Funding: university of Sfax.
- Current situation: Assistant professor, College of Computer Engineering and Science, Prince Sattam Bin Abdulaziz University, Kharj, Saudi Arabia.
- Abstract: Biometrics is a potentially powerful technology for the identification of a person using his physiological traits (palmprint, iris, face, etc.) and/or behavioral traits (voice, signature, etc.). Biometrics traits aim to guarantee a person's safety by eliminating suspicion about his identity and facilitating this identification. Each biometric trait contains several representations, which have discriminatory characteristics for recognition and may be affected by different changes in an uncontrolled environment. In this context, our thesis focuses on the recognition of persons through their palmprints. Our goal is, on the one hand, to offer a deep analysis for palmprint representations, such as the shape of the principal lines and the texture pattern, using several metrics and tools that are invariant to the different changes, and, on the other hand, to design an Intra-Modal biometric system based on the fusion of these different representations of the palmprint. Our main contributions involve the proposition of the principal approaches: (1) Structural approach based on the analysis of the palmprint representation related to the principal lines shapes, (2) Global approach based on the analysis of the representation related to texture patterns, and (3) Intra-modal approach based on two complementary types of fusion: (i) Uni- Representation (based on the fusion of multiple descriptors to analyze the texture, at the feature level based on the correlation concept), and (ii) Multi-representations (based on the fusion of different representations of palmprint, such as the principal line shapes and the texture pattern, at both the feature and score levels). These different approaches have proven their effectiveness by reaching promising recognition rates which are competitive with other proposed approaches for the recognition of persons through their palmprints.
- Supervision team: I participated in the supervision of the thesis as part of a collaboration with Monji Kherallah, professor at the university of Sfax.
- Principal publications co-signed with the student:
  - a paper in the journal Multimedia Tools and APplications MTAP J5 (Impact Factor : 1.5)
  - a paper in the Journal of Digital Crime and Forensics J1
  - a paper is under revision with the journal Pattern Analysis and Applications J3-S (Impact Factor : 1.28)
  - 3 papers in international conferences (C5, C9 and C10).
- Observation: **I received Raouia Mokni at IMT Lille Douai for 2 months in 2015. The stay was funded by the program (bourse d'alternance).**

January 2016-present — **Malek Boujebli**, *Recognition of human actions in a smart home*,

- Institution: university of Manouba, Tunisia.
- Expected date of defense: December 2020.
- Funding: university of Manouba.
- Supervision team: I participate in the supervision of the thesis as part of a collaboration with Riadh Farah, professor at the university of Manouba and Makram Mestiri, associate professor at the university of Manouba.
- Principal publications with the student :
  - A paper is submitted to the journal Pattern Recognition Letters J2-S
  - A paper is published in an international conference C8
- Observation: **I received Malek Boujebli at IMT Lille Douai for 3 months in 2016 and 3 months in 2018. The stay was funded by the program (bourse d'alternance).**

Masters (3)

2018: Rim Zayani, *Deep learning on manifolds for the recognition of abnormal gait.*

- Defense: October 2018.
- Internship setting: Master2 internship at CRIStAL/IMT Lille Douai.
- Funding: PEPS project GeoDeep (CNRS).
- Supervision rate: 100%

2015: Jihed hadj Ali, *Embedding of a 3D face recognition solution.*

- Defense date: November 2015.
- Internship setting: Master2 at CRIStAL/IMT Lille Douai.
- Funding: Program 'Master for international students' by university of Lille.
- Supervision rate: 100%

2014: Karolina Golec, *Parallel programming of a statistical shape analysis framework in 3D.*

- Defense: July 2014.

- Internship setting: Master2 at CRIStAL/IMT Lille Douai.
- Funding: Program 'Master for international students' by university of Lille.
- Supervision rate: 60% (co-supervision with Pr. Mohamed Daoudi).

Collaborations

I have set up new national and international collaborations to deepen certain areas of research.

- Collaboration on Riemannian geometry and 3D face analysis,
  - **Ohio State University, USA**,
    - Type of collaboration: international collaboration.
    - Partner: Sebastian Kurtek, department of Statistics, Ohio State University USA.
    - Principal results of the collaboration: co-signing a paper in the international journal Computer & Graphics J7 and set up and co-organisation of international workshop Diff-CVML in conjunction with BMVC 2015, CVPR 2016 and CVPR 2017.
- Collaborations on the analysis and recognition of actions,
  - **University of Tunis El Manar, Tunisia**,
    - Type of collaboration: international collaboration.
    - Partner: Pr. Walid Barhoumi, LIMTIC laboratory, university Tunis El Manar, Tunisia.
    - Mobility within the collaboration:
      - I received Amani El Aoud, PhD student (university of El Manar), at IMT Lille Douai for a month in May 2017.
      - I received Prof Walid Barhoumi for 2 weeks at IMT Lille Douai in May 2017, his stay was funded by the program: *Séjour Scientifique Haut Niveau* SSHN (franco-tunisien program).
    - Principal results of the collaboration: co-signing 2 international conference papers (rank B) C2 and C7. Both papers have been accepted for oral presentation and C2 was nominated for best paper award.
  - **University Mohamed V, Morocco**,
    - Type of collaboration: international collaboration.

    - Partner: Dr. Lahoucine Ballihi, associate professor at university Mohamed V, Morocco.
    - Context of collaboration: PHC Toubkal project (2019-2021).
    - Mobility within the collaboration: mobility of seniors (7 days per year) and for PhD students (6 months per year).
  - **University of Manouba, Tunisia**,
    - Type of collaboration: international collaboration.
    - Partners: Pr. Riadh Farah and Dr. Makram Mestiri, Riadi laboratory, university of Manouba, Tunis, Tunisia.
    - Mobility within the collaboration: I received Malek Boujebli, PhD student, twice during 3 months at IMT Lille Douai in 2016 and in 2018 as part of a program exchange of Franco-Tunisian doctoral students (*Bourse d'alternance*).
    - Principal results of the collaboration: co-supervision of one thesis. 1 paper is submitted to Pattern Recognition Letters journal (J2-S) and 1 conference paper C8).

- Collaboration on geometric deep learning,
  - **ACQUILAE (Start'up)**,
    - Type of collaboration: national collaboration.
    - Context of the collaboration: PEPS project (2018) and a PHC Toubkal project (2019-2021).
  - **University of Technology of Troyes UTT, France**,
    - Type of collaboration: national collaboration.
    - Partner: Pr. Hichem Snoussi, UTT.
    - Context of the collaboration: PEPS project (2018) and PHC Toubkal project (2019-2021).

- Collaboration on palmprint biometrics,
  - **University of Sfax, Tunisia**,
  - Type of collaboration: international collaboration.
  - Partner: Pr. Monji Kherallah, faculty of sciences, university of Sfax, Tunisia.
  - Mobility within the collaboration: I received Raouia Mokni, PhD student, for two months in 2015 at IMT Lille Douai as part of a franco-tunisian doctoral student exchange program (*bourse d'alternance*).
  - Principal results of the collaboration: co-supervision of one thesis and co-signing of 2 international journal papers (J1 and J5), a paper submitted to international journal (J3-S) and 3 international conference papers (C5, C9 and C10).

In addition to these new collaborations that I set up, I continued to collaborate with the teams that are already collaborating with 3D-SAM team.

- Collaboration on the recognition of 3D objects and and the comparison of surfaces using geometry differential geometry on infinite dimension manifolds.

  - **Statistical Shape Analysis and Modeling Group (SSAMG), Florida State University, USA**.
    - Partner: Pr. Anuj Srivastava, department of Statistics FSU, USA.
    - I spent 2 months in Florida from April to May 2013 as visiting professor.
    - The collaboration resulted in a publication in the prestigious journal IEEE Transactions (IEEE TPAMI J6 impact factor 17.73).
  - **Painlevé Laboratoiry, university of Lille**,
    - Partner: Dr. Barbara Tumpach, associate professor at university of Lille.
    - Context of the collaboration: CNRS delegation (6 months) of Barbara Tumpach at Laboratory of Fundamental Computer Science of Lille LIFL, university of Lille in collaboration with Pr. Mohamed Daoudi.
    - Project: set up mathematical tools for recognizing three-dimensional objects and develop a method of comparing surfaces using differential geometry on manifolds of infinite dimension.
    - The collaboration resulted in a publication in the prestigious journal IEEE Transactions (IEEE TPAMI J6 impact factor 17.73)

- Collaboration on facial expression recognition from 3D facial sequences.

  - **Media Integration and Communication Center (MICC), université de Florence, Italie**,
    - Partner: Prof. Stefano Berretti, university of Firenze.
    - Results of the collaboration
      - a publication in the prestigious journal IEEE Transactions on Cybernatics (J9 Impact factor : 10.38).
      - a publication in the International Conference on Pattern recognition ICPR (oral paper C21) and 2 publications in HBU and EUVIP workshops (Ch19 and Ch20).
      - 2 book chapters (Ch1 and Ch2).
  - **IRIP Laboratory (Laboratory of Intelligent Recognition and Image Processing), Beihang University, China**,

  - Partners: Prof. Yunhong Wang and Dr. Di Huang, Beihang university, China.
  - Context: international ANR project (3D Face Analyzer).
  - Mobility within the collaboration: participation of the workshop on 3D face analyzer project in Beijing, China (2012) and Chamonix, France (2014).
  - Results of the collaboration
    - a publication in the prestigious journal IEEE Transactions on Affective Computing (J4 Impact factor: 6.28)
    - a publication in the Internnational Conference on Pattern recognition ICPR (oral paper C12.

Scientific animation activities

Set up and organization of international workshops

– Founder and co-organizer of international workshop Differential Geometry in Computer Vision and Machine Learning *Diff-CVML*.

- Objective of the workshop: the objective of the workshop is to bring together researchers dealing with aspects of shape analysis and differential geometry for computer vision in order to promote new interdisciplinary collaborations. The goal is to identify new problems as well as potential solutions. The value of this workshop is to bring together people from communities traditionally considered to work in different fields and rarely meet at the same conferences.
- I was co-chair of the first three editions of the workshop, then other researchers took over; this workshop begins to gain momentum by taking place with the conference CVPR (rank A*).
- Edition 2015
  - The first edition of the workshop http://www-rech.telecom-lille.fr/diff-cv2015/ was organized in conjunction with the conference British Machine Vision Conference BMVC 2015 http://www.bmva.org/bmvc/2015/.
  - Co-chairs: **Hassen Drira**, Sebastian Kurtek (Ohio State University, USA) and Pavan Turaga (Arizona State University, USA).
- Edition 2016
  - the second edition of the workshop http://www-rech.telecom-lille.fr/diff-cv2016/ was organized in conjunction with the conference CVPR 2016 (rank A*, major conference in computer vision) http://cvpr2016.thecvf.com/.

  - Co-chairs: **Hassen Drira**, Sebastian Kurtek (Ohio State University, USA), Pavan Turaga (Arizona State University, USA) and Anuj Srivastava (Florida State University, USA).

- Edition 2017
  - the third edition of the workshop http://www-rech.telecom-lille.fr/diffcvml2017/index.html was organized in conjunction with the conference CVPR 2017 (rank A*, major conference in computer vision) http://cvpr2017.thecvf.com/.
  - Co-chairs : **Hassen Drira**, Mehrtash Harandi (CVRG, Australian National University, Canberra, Australia), Sebastian Kurtek (Ohio State University, USA), Pavan Turaga (Arizona State University, USA), Minh Ha Quang (PAVIS, Italian Institute of Technology, Genoa, Italy), Vittorio Murino (PAVIS, Italian Institute of Technology, Genoa, Italy)

Awards

– The GeoDeep project (PEPS 2018) that I coordinated was chosen as a **successful collaboration story** during the 80 years of the CNRS, the event was organized by the association AMIES (Agency for Mathematics in Interaction with Business and society) https://amies-stories.sciencesconf.org/program.

– Thanks to the thesis's work, the former student Amor Ben Tanfous is awarded with a "Special Mention" at the AFRIF (French branch of the IAPR) PhD thesis prize competition (2019) https://lnkd.in/gHWDCFg.

– Best paper award in the international conference VISAPP 2014, (rank B).
Baiqiang Xia, Boulbaba Ben Amor, Mohamed Daoudi, Hassen Drira: Can 3D Shape of the Face Reveal your Age? VISAPP (2) 2014: 5-13 C16.

– Nominated for the best paper award in the international conference VISAPP 2019 (rank B).
Amani Elaoud, Walid Barhoumi, Hassen Drira, Ezzeddine Zagrouba: Weighted Linear Combination of Distances within Two Manifolds for 3D Human Action Recognition, VISIGRAPP, VISAPP 2019. C2,

Scientific societies

– Member IEEE.

– Member of GDR-ISIS, Image et Vision theme.

– Member of *Association Française de Reconnaissance et d'interprétation des Formes* (AFRIF).

Services to the community

– Review activities in international journals

- IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI
- IEEE Transactions on Multimedia
- Pattern Recognition
- IEEE Transactions on Image Processing
- Journal of Electronic Imaging
- Computer Vision and Image Understanding CVIU
- Journal of Imaging
- IET Image Processing
- Transactions on Multimedia Computing Communications Applications
- Image and Vision Computing journal IVC
- Multimedia Systems
- Signal Processing: Image Communication journal
- Multimedia Systems journal.
- International Journal of Pattern Recognition and Artificial Intelligence
- Advances in Multimedia journal
- Applied Mathematical Modelling journal.

– Program committee member (reviewer) in the main conferences in the area

- International Conference on Pattern Recognition CVPR 2020 (rank A*)
- International Conference on Computer Vision ICCV 2019 (rank A*)
- ACM Multimedia 2019, 2020 (rank A*)
- National Conference of the American Association for Artificial Intelligence AAAI 2020 (rank A*)
- European Conference on Computer Vision, ECCV 2020 (rank A)
- IEEE Winter Conference on Applications of Computer Vision WACV 2017-2018-2019-2020 (rank A)

- IAPR International Conference on Pattern Recognition ICPR (rank B)

Conferences Organizing Committees

– Member of organizing committee (Demo co-char) *CBMI 2020, demo co-chair.* https://cbmi2020.univ-lille.fr/organization#chairs

– *Program co-chair* of the international workshop Representation, analysis and recognition of shape and motion FroM Imaging data RFMI 2019. http://www.arts-pi.org.tn/rfmi2019/Committees.php#Committees

– *Co-chair* of the 1st international COnference on Intelligence Artificial and its applications: E-learning digital education http://www.arts-pi.org.tn/AICO2018/comites.php

– Member of local organizing committee of *Shape Modeling International Shape Modeling International 2015, Lille, 2015.*

– Member of local organizing committee *CORESA 2012 (COmpression et REprésentation des Signaux Audiovisuels), Lille, 24-25 mai 2012.*

Invited talks and Seminars

– Instituto Italiano di Tecnologia - Pattern Analysis and Computer Vision (PAVIS), Genova Italy, September 2019.

– DLSTA, Deep Learning Spring school: Theory, tools and Applications, Sousse Tunisia, March 2018.

– LIMTIC Lab., university Tunis El Manar, Tunisia, November 2017.

– CRISTAL Lab., university Manouba, Tunisia, April 2017.

– Faculty of Sciences of Sfax, Tunisia, April 2014.

– Statistical Shape Analysis and Modeling Group (SSAMG), Florida State University, USA, April 2013.

– SSAMG Group, Florida State University, USA, April 2013.

– IRIP Lab., Baihang University, Chine, January 2012.

PhD defense committee

- Member of the committee of thesis defense of Charbel Chahla (thesis examiner), University of Technology of Troyes (UTT), September 2017.

Publications

International journals (11)

[J1] **Raouia Mokni**, Hassen Drira, Monji Kherallah: Deep-Analysis based on Correlation Concept of Palmprint Representation for Human Biometrics identication, International Journal of Digital Crime and Forensics 40-58 (2020).

[J2] **Amor ben Tanfous**, Hassen Drira, Boulbaba Ben Amor: Sparse Coding of Shape Trajectories for Facial Expression and Action Recognition, IEEE Trans. Pattern Anal. Mach. Intell. 2019 **(Impact Factor: 17.73).**

[J3] **Meng Meng**, Hassen Drira, Jacques Boonaert: Distances evolution analysis for online and off-line human object interaction recognition. Image Vision Comput. 70: 32-45 (2018) **(Impact Factor: 2.747)**

[J4] Qingkai Zhen, Di Huang, Hassen Drira, Boulbaba Ben Amor, Yunhong Wang, Mohammed Daoudi: Magnifying Subtle Facial Motions for Effective 4D Expression Recognition, IEEE Transactions on Affective Computing, 2017.) **(Impact Factor: 6.62)**.

[J5] **Raouia Mokni**, Hassen Drira, Monji Kherallah: Combining shape analysis and texture pattern for palmprint identification. Multimedia Tools Appl. 76(22): 23981-24008 (2017) **(Impact Factor: 2.1)**.

[J6] Alice Barbara Tumpach, Hassen Drira, Mohamed Daoudi, Anuj Srivastava: Gauge Invariant Framework for Shape Analysis of Surfaces. IEEE Trans. Pattern Anal. Mach. Intell. 38(1): 46-59 (2016) **(Impact Factor: 17.73)**.

[J7] Sebastian Kurtek, Hassen Drira: A comprehensive statistical framework for elastic shape analysis of 3D faces. Computers & Graphics 51: 52-59 (2015) **(Impact Factor: 1.3)**.

[J8] Baiqiang Xia, Boulbaba Ben Amor, Hassen Drira, Mohamed Daoudi, Lahoucine Ballihi: Combining face averageness and symmetry for 3D-based gender classification. Pattern Recognition 48(3): 746-758 (2015) **(Impact Factor: 5.89)**.

[J9] Boulbaba Ben Amor*, Hassen Drira*, Stefano Berretti, Mohamed Daoudi, Anuj Srivastava, 4D Facial Expression Recognition by Learning Geometric Deformations , accepted for publication as regular paper in the IEEE Tran. on Cybernetics (12), 2443-2457, 2014. **(Impact Factor: 10.38)**. *: equal contributions

[J1-PhD] Hassen Drira, Boulbaba Ben Amor, Anuj Srivastava, Mohamed Daoudi, Rim Slama, 3D Face Recognition under Expressions, Occlusions, and Pose Variations . IEEE Trans. Pattern Anal. Mach. Intell. 35(9): 2270-2283, 2013. **(Impact Factor: 17.73)**.

[J2-PhD] Boulbaba Ben Amor, Hassen Drira, Lahoucine Ballihi, Anuj Srivastava, Mohamed Daoudi, An experimental illustration of 3D facial shape analysis under facial expressions . Annales des Télécommunications 64(5-6): 369-379, 2009. **(Impact Factor: 1.412)**.

Articles under review, or in progress (3)

[J1-S] Nadia Hosni, Boulbaba Ben Amor, Hassen Drira, Faten Cheieb: A functional PCA Framework on the Shape Space for 3D Gait Analysis and Assessment, submitted to Computer Vision and Image Understanding (CVIU).

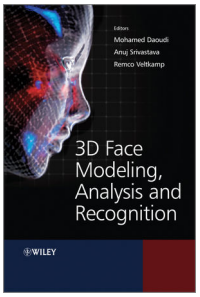
[J2-S] Malek Boujebli, Hassen Drira, Makram Mestiri, Riadh Farah: Rate-invariant Modeling in Lie Algebra for Activity Recognition, submitted to Pattern Recognition Letters journal.

[J3-S] Raouia Mokni, Hassen Drira, Monji Kherallah: A Riemannian Analysis and salient geometrical features of Palmprint Shapes for Human Biometrics, submitted to Pattern Analysis and Applications journal.

Book chapters (2)

Contributor in two chapters of the book 3D Face Modeling, Analysis and Recognition[2], Editeurs: Mohamed Daoudi, Anuj Srivastava, Remco Veltkamp, ISBN: 978-0-470-66641-8, Wiley, Août 2013.

- [Ch1] Chapter III, 3D Face Surface Analysis and Recognition Based on Facial Curves , Hassen Drira, Stefano Berretti, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava, Alberto del Bimbo and Pietro Pala.
- [Ch2] Chapter V, Applications , Stefano Berretti, Boulbaba Ben Amor, Hassen Drira, Mohamed Daoudi, Anuj Srivastava, Alberto del Bimbo and Pietro Pala.



[2] http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470666412.html

Articles in selective international peer-reviewed conferences and workshops (27)

[C1] Rasha Friji, Hassen Drira, Faten Chaieb: Geometric Deep Learning on Skeleton Sequences for 2D/3D Action Recognition. VISIGRAPP (5: VISAPP) 2020: 196-204. (**rank B, oral paper**)

[C2] Amani Elaoud, Walid Barhoumi, Hassen Drira, Ezzeddine Zagrouba: Weighted Linear Combination of Distances within Two Manifolds for 3D Human Action Recognition. VISIGRAPP (VISAPP) 2019 (**rank B, oral paper nominated for best paper award**)

[C3] Amor ben Tanfous, Hassen Drira, Boulbaba Ben Amor : Coding Kendall's Shape Trajectories for 3D Action Recognition. IEEE Computer Vision and Pattern Recognition CVPR 2018 (**rank A***)

[C4] Nadia Hosni, Hassen Drira, Faten Cheieb, Boulbaba Ben Amor: 3D Gait Recognition based on Functional PCA on Kendall's Shape Space. ICPR 2018. (**rang B, oral paper**)

[C5] Raouia Mokni, Hassen Drira, Monji Kherallah: Efficient Personal Identication Intra-Modal System by fusing Left and Right Palms. In the 2018 IEEE International Conference on Intelligent Systems Design and Applications ISDA 2018. (**rank C**)

[C6] Kaouthar Larbi, Wael Ouarda, Hassen Drira, Boulbaba Ben Amor, Chokri Ben Amar: DeepColorFASD: Face Anti Spoofing Solution Using a Multi Channeled Color Spaces CNN. SMC 2018: 4011-4016 (**rank B**)

[C7] Amani Elaoud, Walid Barhoumi, Hassen Drira, Ezzeddine Zagrouba: Analysis of Skeletal Shape Trajectories for Person Re-Identification. ACIVS 2017: 138-149. (**rank B, oral paper**)

[C8] Malek Boujebli, Hassen Drira, Makram Mestiri, I. R. Farah: Rate invariant action recognition in Lie algebra. ATSIP 2017: 1-7

[C9] Raouia Mokni, Anis Mezghani, Hassen Drira, Monji Kherallah: Multiset Canonical Correlation Analysis: Texture Feature Level Fusion of Multiple Descriptors for Intra-modal Palmprint Biometric Recognition. PSIVT 2017: 3-16. (**rank B**)

[C10] Raouia Mokni, Hassen Drira, Monji Kherallah: Fusing Multi-Techniques based on LDA-CCA and their Application in Palmprint Identification System. AICCSA 2017: 350-357. (**rank C**)

[C11] Meng Meng, Hassen Drira, Mohamed Daoudi, Jacques Boonaert: Human Object Interaction Recognition Using Rate-Invariant Shape Analysis of Inter Joint Distances Trajectories. CVPR Workshops 2016: 999-1004

[C12] Qingkai Zhen, Di Huang, Yunhong Wang, Hassen Drira, Boulbaba Ben Amor, Mohamed Daoudi: Magnifying subtle facial motions for 4D Expression Recognition. ICPR 2016: 2252-2257. (**rank B, oral paper**).

[C13] Imed Yehyaoui, Tarek Frikha, Mohamed Abid, Hassen Drira: Embedded adaptation for 3D face analysis using Elastic Riemannian algorithm. IDT 2016: 65-70

[C14] Meng Meng, Hassen Drira, Mohamed Daoudi, Jacques Boonaert: Detection of Abnormal Gait from Skeleton Data. VISIGRAPP (3: VISAPP) 2016: 133-139. (**rank B, oral paper**)

[C15] Meng Meng, Hassen Drira, Mohamed Daoudi, Jacques Boonaert:Human-object interaction recognition by learning the distances between the object and the skeleton joints. FG 2015: 1-6.

[C16] Baiqiang Xia, Boulbaba Ben Amor, Mohamed Daoudi, Hassen Drira: Can 3D Shape of the Face Reveal your Age ? VISAPP 2014 (**rank B, Best paper award**)

[C17] Baiqiang Xia, Boulbaba Ben Amor, Hassen Drira, Mohamed Daoudi, Lahoucine Ballihi: Gender and 3D facial symmetry: What's the relationship? 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013). (**rank C**)

[C18] Baiqiang Xia, Boulbaba Ben Amor Huang Di, Daoudi Mohamed, Wang Yunhong, Hassen Drira: Enhancing Gender Classification by Combining 3D and 2D Face Modalities, 21th European Signal Processing Conference (EUSIPCO) (2013).

[C19] Mohamed Daoudi, Hassen Drira, Boulbaba Ben Amor, Stefano Berretti: A dynamic geometry-based approach for 4D facial expressions recognition. EUVIP 2013: 280-284.

[C20] Hassen Drira, Boulbaba Ben Amor, Mohamed Daoudi, Stefano Berretti: A Dense Deformation Field for Facial Expression Analysis in Dynamic Sequences of 3D Scans. HBU 2013: 148-159.

[C21] Hassen Drira, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava, Stefano Berretti: 3D dynamic expression recognition based on a novel Deformation Vector Field and Random Forest. ICPR 2012: 1104-1107. (**rank B, oral paper**)

[C22] Wael Ben Soltana, Mohsen Ardabilian, Pierre Lemaire, Di Huang, Przemyslaw Szeptycki, Liming Chen, Nesli Erdogmus, Lionel Daniel, Jean-Luc Dugelay, Boulbaba Ben Amor, Hassen Drira, Mohamed Daoudi, Joseph Colineau: 3D face recognition: A robust multi-matcher approach to data degradations. ICB 2012: 103-110.

[C23] Remco C. Veltkamp, Stefan van Jole, Hassen Drira, Boulbaba Ben Amor, Mohamed Daoudi, Huibin Li, Liming Chen, Peter Claes, Dirk Smeets, Jeroen Hermans, Dirk Vandermeulen, Paul Suetens: SHREC '11 Track: 3D Face Models Retrieval. 3DOR 2011: 89-95.

[C24] Hassen Drira, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava: Pose and Expression-Invariant 3D Face Recognition using Elastic Radial Curves. BMVC 2010: 1-11. (**rank B, oral paper**)

[C25] Hassen Drira, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava: Elastic radial curves to model 3D facial deformations 3DOR@MM 2010: 75-80.

[C26] Hassen Drira, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava: Nasal Region Contribution in 3D Face Biometrics Using Shape Analysis Framework. ICB 2009: 357-366.

[C27] Hassen Drira, Boulbaba Ben Amor, Anuj Srivastava, Mohamed Daoudi: A Riemannian analysis of 3D nose shapes for partial human biometrics. ICCV 2009: 2050-2057. (**rank A***)

National reviewed conferences (4)

[C27] Nadia Hosni, Hassen Drira, Boulbaba Ben Amor: ACP fonctionnelle sur l'espace de formes de Kendall pour l'analyse de la démarche en D3, Traitement et Analyse de l'Information Méthodes et Applications (TAIMA), Hammamet, 2018.

[C28] Hassen Drira, Rim Slama, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava: Une nouvelle approche de reconnaissance de visages 3D partiellement occultés„ 18e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle (RFIA), Lyon, 2012.

[C29] Boulbaba Ben Amor, Hassen Drira Daoudi Mohamed, Ardabilian Mohsen, Ben Soltana Wael, Chen Liming, Lemaire Pierre, Erdogmus Nesli, Dugelay Jean Luc, Colineau Joseph: Fusion d'Experts pour une Biométrie Faciale 3D Robuste aux Déformations, 18e congrès

francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle (RFIA), Lyon, 2012.

[C30] Hassen Drira, Boulbaba Ben Amor, Mohamed Douai, Anuj Srivastava: Analyse riemannienne de la forme du nez pour la reconnaissance de visage 3D, COmpression et REprésentation des Signaux Audiovisuels CORESA 2009, Toulouse, 2009.

Ph.D. Thesis

[T1] Hassen Drira, *Statistical computing on manifolds for 3D face analysis and recognition*, Ph.D. Dissertation, defense date: July, 4, 2011, university of Lille1, N°40556.

Pedagogical responsibilities and teaching activities

Created by the merger of Mines Douai and Télécom Lille on January 1st, 2017, IMT Lille Douai is the largest graduate school of engineering north of Paris, training the general engineers and digital experts of the future. Each year, IMT Lille Douai, an IMT school in partnership with the University of Lille, awards degrees to over 500 talented engineers, trained to anticipate economic and social changes. At IMT Lille Douai, we offer different training courses for different audiences.

- IMT Lille Douai offers Specialist Master's degrees to students with a High School Diploma + 5 in a scientific field or a High School Diploma + 4 with proof of 3 years' professional experience connected with the subject of the intended Specialist Master's degree. These Specialized Masters are accredited by the Conférence des Grandes Ecoles and allow them to exercise the functions of technical director or project manager in their respective areas of expertise.
- IMT Lille Douai offers apprenticeship courses that are accessible by application to students at L2, BTS or DUT level. This training takes place alternately: school - company. Apprentices benefit from both consistent professional experience and solid training in the scientific, technological and human fields.
- IMT Lille Douai offers an engineering cycle (high school diploma + 3) after a preparatory class. The Mines-Télécom entrance exam is the main pathway to the IMT Lille Douai engineering program. This entrance exam is arranged for 12 schools based on a separate ranking for each of the MP, PC, PES, PT, TIS, ATS (Higher technician) and BCPES courses of study, for recruiting students from the second year of preparatory classes (Maths specialism) or who have an equivalent education.
- IMT Lille Douai offers an engineering program that is accessible after the french scientific baccalauréat and the GEIPI Polytech entrance exam. This excellent recruitment process allows the

best scientific high school students to join the school for a 5-year training program to get the engineering degree.

I assume different educational responsibilities in each of the training courses offered by IMT Lille Douai.

1. **Pedagogic responsible of the specialist master "Cybersecurity Engineering"**

   Since March 2019, I have taken the responsibility of the specialist Master Cybersecurity Engineering. The specialist Master Cybersecurity Engineering is accredited by the CGE (Conference des Grandes Ecoles) and labeled SecNumEdu by the ANSSI (Agence Nationale de la Sécurité des Systèmes d'Information).

   Content of the four course units and scientific and technological project:

   - UV FUNS (FUndamental Networks and Security) – ECTS, will be added (in September 2020) to the Mastère's program, 7 ECTS, this UV is also proposed to engineer students (M1 and M2 level).
   - UV TOP (TheOry & Practice of information security), 8 ECTS, this UV is also proposed to engineer students (M1 and M2 level).
   - UV SRS (System and Network Security), 9 ECTS, this UV is also proposed to engineer students (M1 and M2 level).
   - UV CIS (Cybersecurity of Industrial Systems and Services), 12 ECTS, this UV is exclusively proposed to Master students.
   - Scientific and technological project – Duration: 130 working hours in groups of two (9 ECTS)

   Several cybersecurity companies support and provide input to the course through lectures from their experts, providing platforms, as well as proposing and supervising projects, etc. The coordination of the Master includes the coordination of the intervention of the various industrial partners.

   - Orange Cyberdefense and its Cybersecurity Centre of Excellence (PEC), in which IMT Lille Douai is a stakeholder
   - Stormshield, a subsidiary of Airbus Defence and Space CyberSecurity, which has RD facilities near to the school
   - Advens, an information security management specialist
   - Wavestone, a leading consulting firm
   - The Innovation Centre of Contactless Technologies (CITC), a key player in the Internet of Objects and ambient intelligence

A 5-month professional dissertation to be completed in the workplace.

The coordination of the Masters includes the organization of the admission juries, **the coordination of the course units offered in the training**, the follow-up of internships in companies, the organization of the defense of professional theses, the organization of the final jury. Since taking responsibility, I have had to perform the following tasks:

- **Mounting a dossier to register the Master to the National Directory of Professional Certification or RNCP** https://certificationprofessionnelle.fr/
- **Renew the SecNumdu-certification by the ANSSI** https://www.ssi.gouv.fr/entreprise/formations/secnumedu/**.**
- **Renew the CGE accrediattion** https://www.cge.asso.fr/presentation-de-la-formation-lab
- **Adaptation of the content of UV TOP and SRS which are offered to engineering students at M2 level with the new courses at IMT Lille Douai.**
- **Adding the new UV FUNS to the study program in order to reduce the number of hours in TOP and SRS and make it possible to offer them in the new courses at IMT Lille Douai. Moreover, the FUNS UV includes some fundamental courses on Networks and security allowing the increase the targeted students for the Master.**
- **Follow-up of internships in companies, organizing the defense of professional theses and organizing the final jury.**
- **Renewal of interventions of companies (Orange, Advens, Stormshield, etc) and planning of new collaborations with other companies (ATOS, OVH, MANIKA).**
- **Selection of candidates, interviews and organization of admission juries.**
- **2019-present: coordination of te UV CIS (Cybersecurity of Industrial Systems and Services)**
  - Number of hours seen by the student : 120 hours
  - Targeted audience: specialist Master 'Cybersecurity Engineering' students.
  - Number of students in 2019-2020: 5.
  - Number of speakers to manage: 6
  - Content : Industrial control systems (ICS) security, Business continuity planning and risk management, Security management and SIMEs (Security Information Event Management), Dependability: cybersecurity, a new challenge?, Advanced network security (II) – Security of Software-Defined Networking (SDN), Network Virtualisation Functions (NFV), Assessment methodologies for authentification approaches and attack and counter-attack methods (spoofing and anti-spoofing), Cloud computing and cloud security, The internet of objects (IoT) and security, Research project, Conferences

- **2019-present: coordination of the UV TOP (TheOry & Practice of information security)** This UV is offered to engineering students and master students (more informations are given in the following section on responsibilities in the engineering courses).
- **2019-present: coordination of the UV SRS (System and Network Security)** This UV is offered to engineering students and master students (more informations are given in the following section on responsibilities in the engineering courses)

2. **Pedagogical responsibilities and coordination in the engineers studies**

At IMT Lille Douai, the coordination of a *Unité de Valeur* (UV) or a course is a complete task which involves the creation of the course in terms of lectures, exercises, and laboratory courses, the design of the teaching materials, including room reservations/management. The coordination of teaching staff, the design and grading of exams, the participation in committees, etc.

- **2019-present: coordinator of the UV TOP (TheOry & Practice of information security)**
  - Number of hours seen by the student: 120 hours
  - Targeted audience: engineer students M2 level (36) and specialist Master 'Cybersecurity Engineering' students (5).
  - Number of students in 2019-2020: 41.
  - Number of speakers to manage: 9
  - Content : Legal and statutory cybersecurity requirements, Networks in a nutshell – models, architecture and protocols, Cryptography – basics and applications, Biometric authentification – systems and uses, Trust and reputation management (in English), Organisational audits – basics, risk analysis and the ISO 2700x benchmark standards, Risk analysis project

- **2019-present: coordinator of the UV SRS (System and Network Security)**
  - Number of hours seen by the student: 120 hours
  - Targeted audience: engineer students M2 level (36) and specialist Master 'Cybersecurity Engineering' students (5).
  - Number of students in 2019-2020: 41.
  - Number of speakers to manage: 9.
  - Content: Recent cyber attacks and cyber defence systems, Hacking and technical audits, Network security (I), Intrusion detection systems, or IDS (in English), Digital forensics, Controlled access models and Kerberos authentification protocol, Image processing and biometric pattern recognition, WiFi network security, CNSA Stormshield certification, Conferences

- **2018-present: co-designer and coordinator of the UV FUndamentals of Networks and Security (FUNS) in the new courses of IMT Lille Douai**
  - Number of hours seen by the student: 90 hours
  - Targeted audience: engineer student level M1, M2 and the specialist master 'Cybersecurity Engineering' from September 2020.
  - Number of students in 2019: 20 students M1 + 81 students L3 (transition year).
  - Number of speakers to manage: 8.
  - Accomplished tasks: co-designing of the UV in the new courses of IMT Lille Douai, coordination, and adding to the 'Cybersecurity Engineering' specialist master.
  - Content: Quality of networks (Stochastic process and queue theory), security of networks, Voice IP.

- **2012-2019: coordinator of the UV Stochastic process and queue theory in Télécom Lille courses**
  - Number of hours seen by the student: 30 hours.
  - Targeted audience: engineer students level L3 (Télécom Lille courses).
  - Number of students: 120 students from 2012 to 2017, (60 students in 2018 and 81 in 2019).
  - Number of speakers to manage: 5.
  - Accomplished tasks: Updating the course, setting up a new lab and coordination.
  - Content: Stochastic process and queue theory, graph modeling, ...
- **2012-2014: coordinator of the UV Data structure and C programming**
  - Number of hours seen by the student: 64 hours.
  - Targeted audience: engineer students level L2.
  - Number of students: 120.
  - Number of speakers to manage: 5.
  - Content: C Programming, Data structure (lists, tables), C programming, pointers (dynamic memory allocation) ...

3. **Pedagogic responsibilities and coordination within the apprenticeship study program**

For this apprenticeship training, in addition to coordinating the module and face-to-face teaching, I provide distance learning (15h E-learning).

2012-present: Coordinator of the UV Stochastic process and queue theory within the apprenticeship.

- Number of hours seen by the student: 30 hours
- Targeted audience: engineer students level M1 (apprentices).
- Number of students: 60 students.
- Number of speakers to manage: 2.
- Accomplished tasks: updating the course, setting up a new lab and coordination.
- Content: Stochastic process and queue theory, graph modeling

4. **Internship tutoring**

- Since the 2012, I have been the tutor of several engineering students in studies project (final internship in fifth year) every year, as needed (2 PFE per year on average). Since 2019, I have been following the internships of students of the Specialized Master 'CyberSecurity Engineering'.
- During their course, to supplement the courses (lecture in amphitheater, labs) and internships in companies, students also follow lessons in the form of projects, called Scientific and Technological Projects. Each year, I propose and follow several projects of this kind. The subjects are always oriented towards my research activities. This allows the student to feel involved in an educational project with a direct utility for the teacher-researcher and not a simple pedagogical duty with no other utility than to give a mark.
- IMT Lille Douai gives students the opportunity to do their final year on a professionalization contract. This possibility allows students who wish to spend their year in alternation at a weekly rate of 3 days at school and 2 days in company until the end of studies project of 24 consecutive weeks in company. I follow up, as needed, student engineers on professionalization contracts.

5. **Teaching activities**

The teaching times given in this paragraph are understood as in-class hours teaching (face-to-face), that is to say without coordination, corrections, etc. In Table 1.1, details of my teaching activities are reported in terms of lectures/exercise courses/laboratory courses.

TABLE 1.1: Teaching activities in hours (lectures/exercise courses/lab. courses) per field/specialty and year. (*: E-learning, **: coaching)

	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020
GAN	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	4,5/-/6
Face modeling and analysis	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	6/-/6	6/-/6	6/-/6	6/-/6	6/-/6	6/-/6	6/-/6	6/-/6
Biometrics	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	4,5/-/15
JAVA programming	-/-/-	-/-/-	-/-/-	-/-/-	-/-/15	-/-/-	-/-/-	-/-/-	-/-/-	-/-/15	-/-/15	-/-/15	-/-/-
Stochastic process and queue theory (apprenticeship)	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	9/26*/6	9/26*/6	9/26*/6	9/21*/6	9/21*/6	9/21*/6	9/21*/6	9/21*/6
Stochastic process and queue theory	-/-/-	-/15/-	-/15/-	-/15/-	-/-/-	12/15/6	12/15/6	12/15/6	12/15/6	12/15/6	12/15/6	12/15/6	12/15/6
Data Structure	-/-/-	-/-/-	-/-/-	-/12/18	-/-/-	-/12/78	-/12/78	-/12/78	-/12/18	-/12/18	-/12/18	-/12/18	-/12/18
Language theory and compilation	-/-/-	-/-/-	-/-/-	-/12/18	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
Introduction to algorithmic	-/-/-	-/12/24	-/12/24	-/12/24	-/12/24	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
3D Face Biometric	-/-/-	-/-/-	-/-/-	1,5/-/3	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
Unix and system programming	-/-/-	-/-/-	-/-/-	-/-/6	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
Entrepreneurship Challenge	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/8**/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
Advanced Data Structure	-/-/-	-/12/-	-/12/-	-/12/-	-/12/-	-/12/-	-/12/-	-/12/-	-/12/-	-/12/-	-/-/-	-/-/-	-/-/-
Video processing	-/-/-	-/-/6	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
O.S. UNIX	15/-/15	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
Web design	18/-/30	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
Propositional logic and predicate calculus	-/30/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
Internship tutor	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/15/-	-/15/-	-/15/-	-/15/-	-/15/-	-/15/-	-/25/-	-/25/-
Project tutor	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-	-/5/-	-/5/-	-/5/-	-/5/-	-/5/-	-/5/-	-/5/-	-/5/-
Total number of hours (ETD)	124,5	69	63	134,25	78	229,5	221,5	221,5	161,5	170,5	168,5	168,5	194

The teaching times given in this paragraph are understood as in-class hours teaching (face-to-face), that is to say without coordination, corrections, etc.

(a) Associate professor at IMT Lille Douai (September 2012 - present)

- Generative Adversial Networks (GAN)
  - Type of intervention: lectures and laboratory courses
  - Number of hours: 4,5 hours lectures and 6h laboratory courses
  - Targeted audience: engineer students level M2 (Number of students in 2020: 28)
  - Intervention years: 2019-2020 to present
  - Course objectives: the aim is to explain the principle of Generative Adviersial Networks (GAN). A python project (laboratory courses) on skeletal data is proposed to develop a skeleton sequence generator with predefined actions.
- Face modeling and analysis
  - Type of intervention: lectures and laboratory courses
  - Number of hours: 6 hours lectures and 6 hours laboratory courses
  - Targeted audience: engineer students level M2 (Number of students in 2019-2020: 20)
  - Intervention years: 2012-2013 to present
  - Course objectives: the objective is to give an overview of the taxonomy on face analysis approaches with different applications. A Python project is proposed to develop an approach for recognizing facial expressions.
- Biometrics
  - Type of intervention: lectures and laboratory courses
  - Number of hours: 4,5 hours lectures and 15h laboratory courses
  - Targeted audience: engineer students level M2 and Cybersecurity Specialist Master students (Number of students in 2019-2020: 41)
  - Intervention years: 2019-2020 to present
  - Course objectives: the aim is to define the notion of biometric and give an overview of biometric systems and their performance measurement. A python project (laboratory courses) on face recognition using eigenfaces is proposed (implement eigenfaces algorithm, implement ROC curves, CMC curves, etc).
- JAVA programming
  - Type of intervention: laboratory courses, project tutoring and evaluation
  - Number of hours: 15 hours
  - Targeted audience: Engineer students level L3
  - Years of intervention : 2016-2017, 2017-2018, 2018-2019
  - Course objective: oriented-object programming

- Stochastic process and queue theory
  - Type of intervention: lectures, tutorials (exercises), laboratory courses
  - Number of hours: 12 hours lectures, 15 hours Tutorials (exercises) and 6 hours laboratory courses
  - Targeted audience: Engineer students level L3
  - Year of intervention: 2012-present
  - Course objective: Network modeling by stochastic process and queue theory, graph modeling
  - Note: since 2019, this course is part of UV FUNs (proposed in M1 and M2 level) in the new course program of IMT Lille Douai
- Stochastic process and queue theory (apprenticeship)
  - Type of intervention: E-learning, lectures, tutorials (exercises), laboratory courses
  - Number of hours: 26 hours E-learning, 9 hours lectures, 6 hours Tutorials (exercises) and 6 hours laboratory courses
  - Targeted audience: Engineer students (apprentices) level M1
  - Year of intervention: 2012-present
  - Course objective: Network modeling by stochastic process and queue theory, graph modeling
- Data Structure
  - Type of intervention: tutorials (exercises), laboratory courses, project tutoring and evaluation
  - Number of hours: 12 hours tutorials (exercises) and 18 hours laboratory courses, project (30 hours per group, project tutoring until 2014-2015)
  - Targeted audience: Engineer students level L2
  - Year of intervention: 2012-present
  - Course objective : C programming, Data structure (lists, trees), dynamic allocation of memory
- Advanced Data Structure
  - Type of intervention: tutorials (exercises)
  - Number of hours: 12 hours tutorials (exercises)
  - Targeted audience: engineer students Télécom Lille L3, (30 students)
  - Years of intervention: 2012-2013, 2013-2014, 2014-2015, 2015-2016, 2016-2017
  - Course objective: this module introduces from a theoretical and practical point of view some advanced data structures commonly used in IT: linked lists, hash tables, balanced trees, etc.
- Entrepreneurship Challenge

- Type of intervention: Coaching
- Number of hours: 8 hours
- Targeted audience: Engineer students level M1 and Economy-gestion students M1 level, 4 groups of 5 students each.
- Year of intervention: 2012-2013
- Objective:
  i. Promote an innovative idea for company creation integrating,
  ii. Encourage entrepreneurship,
  iii. Implement a pedagogy by project,
  iv. Make the students work in full-scale teams, from different horizons and disciplines
  v. Mobilize individual and collective creative energies around an innovative project,
  vi. Implement a transversal educational project,
- I had the opportunity to assist students with different company projects and take advantage of my technical expertise to advise them on the best technological choices while validating the feasibility of ideas.

(b) Temporary lecturer and researcher (ATER) at university of Lille/Télécom Lille (January 2011 - August 2011)

In order to consolidate my teaching skills at the end of the thesis, I applied for a position of Temporary lecturer and researcher (ATER). This position allowed me to carry out the following interventions at Télécom Lille.

- Language theory and compilation
  - Type of intervention: tutorials (exercises), laboratory courses, project tutoring and evaluation
  - Number of hours: 12 hours tutorials (exercises), 18 hours laboratory courses
  - Targeted audience: Engineer students level L2, (20 students)
  - Year of intervention: 2010-2011
  - Course objective: the aim of this course is to address the main theories necessary for understanding, compiling and interpreting programming languages. Using Flex and Bison software, students are assisted in a project that involves designing simple compilers.
- Introduction to algorithmic
  - Type of intervention: tutorials (exercises), laboratory courses, project tutoring and evaluation
  - Number of hours: 12 hours tutorials (exercises), 24 hours laboratory courses
  - Targeted audience: Engineer students level L1, (20 students)
  - Year of intervention: 2010-2011

  - Course objective: the aim of this module is to introduce to students the notion of algorithms complexity. To illustrate the course, the tutorials present various sorting algorithms and a study of their complexity is proposed. From a practical point of view, this module is validated by a C implementation project aiming to measure in practice the time complexity of these different sorting algorithms in a lexicographic research context on a large corpus.

- Introduction to Unix and system programming in C language
  - Type of intervention: tutorials (exercises)
  - Number of hours: 6 hours
  - Targeted audience: Engineer students level L3, (20 students)
  - Year of intervention: 2010-2011
  - Course objective: this course gives an introduction to the Unix operating systems as well as basic system programming. After a few lessons (introduction to Unix, C language reminders and input / output system calls), the course is mainly made up of laboratory courses with the final aim of programming a project in C: generally the programming of a system command simplified. This course is mainly made up of laboratory courses which allow students to remember the principles of programming in C but also to learn Unix.
- Data Structure
  - Type of intervention: tutorials (exercises), laboratory courses, project tutoring and evaluation
  - Number of hours: 12 hours tutorials (exercises) and 18 hours laboratory courses
  - Targeted audience: Engineer students level L2
  - Year of intervention: 2010-2011
- Advanced Data Structure
  - Type of intervention: tutorials (exercises)
  - Number of hours: 12 hours tutorials (exercises)
  - Targeted audience: engineer students Télécom Lille L3, (30 students)
  - Years of intervention: 2010-2011
- Stochastic process and queue theory
  - Type of intervention: tutorials (exercises)
  - Number of hours: 15 hours Tutorials (exercises)
  - Targeted audience: Engineer students level L3
  - Year of intervention: 2010-2011
- 3D Face Biometric
  - Type of intervention: lectures and laboratory courses

  - Number of hours: 1,5 hours lectures and 3 hours laboratory courses
  - Targeted audience: Engineer student level M2 (Multimedia option), 15 students
  - Year of intervention: 2010-2011
  - Course objective: The objective of this course is to present a method of 3D facial recognition (which I developed in my thesis) and initiate students to measure the performance of biometric systems (ROC and CMC curves).

(c) Vacations at Télécom Lille, university of Lille. I assured vacations in order to diversify my teaching skills.

- JAVA programming
  - Type of intervention: laboratory courses, project tutoring and evaluation
  - Number of hours: 15 hours
  - Targeted audience: Engineer students level L3
  - Years of intervention: 2011-2012
  - Course objective: oriented-object programming
- Advanced Data Structure
  - Type of intervention: tutorials (exercises), laboratory courses
  - Number of hours: 12 hours tutorials (exercises)
  - Targeted audience: engineer students Télécom Lille L3, (30 students)
  - Years of intervention: 2011-2012
- Introduction to algorithmic
  - Type of intervention: tutorials (exercises), laboratory courses, project tutoring and evaluation
  - Number of hours: 12 hours tutorials (exercises), 24 hours laboratory courses
  - Years of intervention: 2011-2012
  - Targeted audience: Engineer students level L1, (20 students)

(d) Higher Education Monitor at Télécom Lille (September 2008 - August 2010)

During my thesis, I applied to the Higher Education Monitor. This position allowed me to carry out the following interventions at Télécom Lille.

- Introduction to algorithmic
  - Type of intervention: tutorials (exercises), laboratory courses, project tutoring and evaluation
  - Number of hours: 12 hours tutorials (exercises), 24 hours laboratory courses
  - Years of intervention: 2008-2009, 2009-2010;
  - Targeted audience: Engineer students level L1, (20 students)
- Advanced Data Structure

  - Type of intervention: tutorials (exercises), laboratory courses, project tutoring and evaluation
  - Number of hours: 12 hours tutorials (exercises)
  - Targeted audience: engineer students Télécom Lille L3, (30 students)
  - Years of intervention: 2008-2009, 2009-2010
- Stochastic process and queue theory
  - Type of intervention: tutorials (exercises)
  - Number of hours: 15 hours Tutorials (exercises)
  - Targeted audience: Engineer students level L3
  - Year of intervention : 2008-2009, 2009-2010
- Video processing
  - Type of intervention: laboratory courses
  - Number of hours: 6 hours laboratory courses
  - Targeted audience: engineer students Télécom Lille M2, (35 students)
  - Year of intervention: 2008-2009;
  - Course objective: these laboratory courses are intended to introduce students to video signal processing techniques. The goal of the project is to create blue screen processing software, capable of automatically modifying the background image of a blue screen sequence according to the sound context of the sequence. The implementation is based on the free Transcode platform.

(e) Assistant professor at Higher Institute of Computer Science (ISI), Tunisia (September 2007 - January 2008)
After my master, I had the opportunity to have a contract as assistant professor at the Higher Institute of Computer Science (ISI) in Tunis, Tunisia, school of engineers in 6 years (3 years License then 3 years engineering cycle) - El Manar university.

- Operating Systems UNIX
  - Type of intervention: lectures and laboratory courses
  - Number of hours: 15 hours lectures and 15 hours laboratory courses
  - Targeted audience: L1 students, (30 students)
  - Intervention year: 2007-2008
  - Course Objective: This course aims to familiarize students with UNIX operating systems as well as to introduce them to the principles of shell programming.
  - Note: It was the first year of Tunisia's experience with the LMD system and I had to prepare the course. It was the opportunity for me to initiate myself to lectures. I wrote most of the exam validating the course and participated to its correction.
- Web design

  - Type of intervention: lectures and laboratory courses
  - Number of hours: 18 hours lectures, 30 hours laboratory courses
  - Targeted audience: L3 students, (30 students)
  - Intervention year: 2007-2008
  - Course Objective: this course aims to enrich the training of students with a touch of Graphic Computing by initiation to programming techniques of web pages, mainly using Adobe Photoshop and Adobe Flash.
  - I participated in writing the exam validating the course and participated to its correction.
- Propositional logic and predicate calculus
  - Type of intervention: tutorials, exercises
  - Number of hours: 30 hours
  - Targeted audience: engineer students level L3, (3 groups, 25 students each)
  - Intervention year: 2007-2008

The quality and variety of the courses given during my contract as an assistant professor in Tunisia, then as an lecturer then ATER and finally Associate professor, allowed me to acquire a good educational experience. In addition, the tutoring of students during their internships or end-of-study project, allowed me to share my scientific knowledge. During my lessons, I also had the opportunity to teach to an audience of different levels: from L1 where the pedagogy and supervision of students must be more important, to M2 where we can leave more freedom to students but which requires greater personal preparation to be able to easily answer their questions.

Chapter 2

Introduction

As one of the most active research areas in computer vision, visual analysis of human motion attempts to detect, track and identify people, and more generally, to interpret human behaviors, from image sequences involving humans. Human motion analysis has attracted great interests from computer vision researchers due to its promising applications in many areas: (1).

- **Smart Surveillance**: In today's surveillance systems, video contents are viewed continuously by human operators. Smart surveillance systems can analyse an event online and provide appropriate intimation using computer based human motion and behavioural analysis. Smart surveillance is required for access control in special areas like military territory, distant human identification, counting the persons and congestion analysis, detection of abnormal behaviour at shopping malls, railway stations, hospitals, government buildings, commercial premises, and schools (2).

- **Behavioural Biometrics**: Nowadays, the use of the gait pattern as a biometric has become popular. The main reason is that the recognition of the gait pattern does not require subject cooperation as compared to the other biometrics (3).

- **Gesture and Posture Recognition and Analysis**: For a more advanced natural interface with computers and computerized systems, human gesture and posture recognition is an important key. It has promising applications such as gaming, sign language recognition, controlling devices, and others (4, 5).

- **Robotics**: Human motion analysis plays an important role in robotics for humanoid robot control, to imitate human motions in a robot in virtual and augmented environments (6).

- **Medical**: The medical field uses human motion recognition for the study and analysis of Orthopaedics, Neurology, Musculoskeletal disorders, body posture, and fitness. It is also useful to design intelligent systems to assist elderly people and physically / mentally disabled ones (7, 8).

- **Sports and Exercise**: In sports, motion recognition is useful to analyze athletic movements and to design affordable and efficient frameworks for training. An environment for rehabilitation exercise with a feedback system at remote places or in the presence of an expert is designed (9). (10) proposed a monitoring system for the exercises of elderly people. These kinds of systems will definitely be useful for patients and old age people.

- **Art and Entertainment**: Motion recognition is useful in analyzing, learning, and an emotional understanding of artistic dance movements as in dances like Bharatnatyam, and Salsa. Kale and Patil (11) have recognized Bharatnatyam dance sequence from depth data. This also helps to increase the effectiveness of a scene, and the alteration of movements required for quality and the impact of acting.

Depending on complexity, human motion is conceptually categorized into gestures, actions, activity, interactions, and group activities. Representation and recognition methodologies are decided from tracking and initialization of human body in video. Broad approaches for representation are 2-D Kinematic or stick figure, 3-D kinematic or shape model and image model.

Over the last decade, I have directed my research towards the topic of shape analysis of imaging data with application to human behavior analysis. I have investigated the analysis of the shape of sparse and dense, 2D and especially 3D, static and dynamic representations of human in order to recognize his characteristics (soft-biometrics, like gender and age), understand his emotion (facial expression recognition) and recognize his activity (human action recognition).

With the emergence of modern shape theory and related approaches based on differential geometry, I have been attracted by the elegant theory and relevant geometric and statistical tools that it offers, in particular viewing shapes as elements of finite- or infinite-dimensional manifolds, the definition of Riemannian structures (or metrics) on these manifolds, and computing statistics on them (sample mean of shapes, sample covariance, etc. These tools are suitable and computationally efficient to be applied to pattern recognition problems. Throughout this *habilitation*, my goal was to develop shape analysis frameworks for 2D and 3D landmarks (facial landmarks, skeleton), 3D faces, dynamic faces and spherical surfaces with specific goal of defining a shape space for each representation of human imaging, the shape space is defined as the invariant under the action of groups modeling the undesirable transformations. The shape variations are modeled by action of Lie groups on shape representations, as previously formulated by D.G. Kendall and Ulf Grenander which leads to the **Group theory**.

Pioneering work related to modern shape theory have investigated different representation of data ranging from landmarks (a dataset of registered anchor points), in particular the Kendall's shape analysis methodology, to landmark-free geometrical representations, including 3D continuous curves and 3D continuous surfaces. Shapes of natural or man-made objects extracted from imaged scenes are important cues used in the detection, recognition and classification. It points out the external form of someone or

something as produced by their outline and is usually viewed as a set of landmarks or continuous boundary. A formal and intuitive definition of shape has been introduced in (12) then in (13) and considered later by many researchers in this field:

Shape *is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object (taken from (13)).*

The seminal work of Kendall (12) (and later of (13)) and Bookstein (14) resulted in elegant and comprehensive statistical shape analysis theory, that influenced the modern theory of shapes and inspired many researchers with the introduction of methods and techniques derived from differential geometry. Herein, shapes are represented by **sets of ordering landmark points** and their statistical variability imposes to deal with a set of Euclidean shape-preserving transformations such us **scaling**, **translation**, and **rotation**. Starting with this space, Kendall methodically and very rigorously proceed to remove variability due to translation, rotation, and (optionally) scaling to arrive at their space of shape representations, which is the space of orbits under the action of the rotation group. He also equipped this orbit space with a Riemannian metric, making possible to quantify shape divergences (geodesic distance) and to provide geodesics between shapes. The notion of *geodesic distance* is a basic tool for statistical shape analysis. Later, in (13) the authors have discussed planar Procrustes analysis to highlight the main components of shape analysis. They have provided an environment for the development of a statistical theory of shapes via mean shapes and tangent space probability models to Kendall's shape manifolds.

The first contribution of this *habilitation* is to propose an intrinsic sparse coding and dictionary learning SCDL on the Kendall Shape Space with application to action recognition (3D landmarks) and expression recognition (2D landmarks). A comparative study to an extrinsic sparse coding is also presented to understand the benefit of each methodology. Intrinsic methods perform calculus on shape spaces by projecting the manifold valued data to the tangent spaces. Whereas extrinsic methods project the data, via a kernel mapping, to an euclidean space (bigger dimension in general) to allow euclidean calculation. This contribution is presented in chapter 3

From a similar point-of-view, Grenander's shape theory formulation (15) viewed **continuous shapes** as points on an infinite-dimensional, differentiable manifold. The variations between shapes are modeled by actions of *Lie groups* (deformations) on this manifold. Low-dimensional groups, such as rotation, translation and scaling, change the object instances keeping the shape unchanged, while high-dimensional groups (diffeomorphisms) smoothly change the object shapes (16). This theory proposed to view the set of shape representations (the shape space) as quotient of the pre-shape space, obtained by modding out shape-preserving transformations. (17) introduced a representation for planar curves parametrized by arc-length in which each curve is represented by its angle function $\theta(t)$, defined as the elevation angle of the tangent vector of the curve at t (values are chosen so that θ is continuous). This representation is invariant to translation and reparametrization, and can be made rotation-invariant by vertically shifting each angle function so that it has an average value of π. However, since all curves are required to be

parametrized by arc length, it is not possible to reparametrize curves to improve the registration between them. (18, 19) presented a special representation of curves, called the Square-Root Velocity Function (SRVF), under which a specific elastic metric becomes an $\mathbb{L}^2$ metric and simplifies the shape analysis.

In order to capture and model the deformations of the face, we propose to use the same facial representation by radial curve (20, 21). We have exploited the notion of **shooting vector** along a geodesic to represent the facial deformations between the faces, end points of the geodesic, and derive our Dense Scalar Fields. This novel geometric feature has been used to define efficient high level features for soft-biometrics recognition (gender and age) and facial expression recognition. This represents **the second contribution of this *habilitation*** and is presented in chapter 4.

The third contribution of this *habilitation* is presented in chapter 5. The key idea lies in defining manifolds of 3D surfaces, there is no need to parameterize the surface by a reference point and collection of curves as proposed in chapter 4. The first step is to adapt a recent elastic shape analysis framework (22, 23) to the case of hemispherical surfaces, and explore its use in a number of 3D face processing applications including face deformation, template computation, summarization of variability in different expression classes, random generation of 3D faces from a Gaussian-type generative model, and symmetry analysis. Second, we present a novel framework for shape analysis of spherical surfaces. The novelty lies in defining the Riemannian metric directly on the quotient (shape) space, rather than inheriting it from pre-shape space, and using it to formulate a path energy that measures only the normal components of velocities along the path. In other words, we define and solve for geodesics directly on the shape space and avoid complications resulting from the quotient operation. This comprehensive framework is invariant to arbitrary parameterizations of surfaces along paths, a phenomenon termed as gauge invariance. Finally Chapter 6 presents my current research and some perspectives for future directions.

Chapter 3

Action Recognition based on Skeleton Data

The main results presented in this chapter have been published in the following international journal: IEEE PAMI (2019) (24), IVC (2017) (25), in addition to the main conference CVPR 2018 (26).

3.1 Introduction

The availability of real-time skeletal data estimation solutions (27, 28) and reliable facial landmarks detectors (29–31) has pushed researchers to study shapes of landmark configurations and their dynamics. For instance, 3D skeletons have been widely used to represent human actions due to their ability in summarizing the human motion. Another example is given by the 2D facial landmarks and their tremendous use in facial expression analysis. However, human actions and facial expressions observed from visual sensors are often subject to view variations which makes their analysis complex. Considering this non-trivial problem, an efficient way to analyze these data takes into account view-invariance properties, giving rise to shape representations often lying to nonlinear shape spaces (12, 32, 33). David G. Kendall (12) defines the shape as the geometric information that remains when location, scale, and rotational effects are filtered out from an object. Accordingly, we represent 2D landmark faces and 3D skeletons as points in the 2D and 3D Kendall's spaces, respectively. Further, when considering the dynamics of these points, the corresponding representations become trajectories in these spaces (32). However, inferencing such a representation remains challenging due to the *nonlinearity* of the underlying manifolds. In the literature, two alternatives have been proposed to overcome this problem for different Riemannian manifolds – they are either *Intrinsic* (34–37) or *Extrinsic (kernel-based)* (38–41). On one hand, intrinsic solutions tend to project the manifold-valued data to a tangent space at a reference point (32, 37, 42). While it solves the problem of nonlinearity of the manifold of interest, this solution could introduce distortions, especially when the projected points are far from the reference point. On the other hand, extrinsic solutions are

based on embeddings to higher dimensional Reproducing Kernel Hilbert Spaces (RKHS), which are vector spaces where Euclidean geometry applies. These methods bring the advantage that, as evidenced by kernel methods on $\mathbb{R}^n$, embedding a lower dimensional space in a higher dimensional one gives a richer representation of the data and helps capturing complex patterns. Nevertheless, to define a valid RKHS, the kernel function must be positive definite according to Mercer's theorem (43). Several works in the literature have studied kernels on the 2D Kendall's space. For instance, the Procrustes Gaussian kernel is proposed in (40) as positive definite. In contrast, to our knowledge, such a kernel has not been explored for the 3D Kendall's space.

Motivated by the success of sparse representations in several recognition tasks (36, 38, 44), we propose to code shape trajectories using Riemannian sparse coding and dictionary learning (SCDL) in the shape manifold. We will explore both intrinsic and extrinsic paradigms of this technique and provide the main benefits of the resulting representations to model human actions and facial expressions.

We start by presenting some preliminaries and mathematical definitions in section 3.2. Later on, we review some existing representations of trajectories on Riemannian manifolds along with scientific challenges on the matter in section 3.3. In Section 3.4, we describe the solutions that we adopt to code trajectories in the shape manifold before presenting their properties in the following section. Section 3.6 presents the application of the proposed methodology in the context of expression recognition and human action recognition and sections 3.7 and 3.8 present the experimental results and discussions. Finally, we propose in section 3.9 an online human action recognition including interaction with objects and we conclude the chapter with a conclusion section.

3.2 Background and definitions

Data encoding techniques have been broadly studied in literature as they play a crucial role in data analysis. In this dissertation, we aim to represent human skeletons and facial landmark configurations with a coding technique that yields sparse representations. This classical approach assumes that the data are defined in Euclidean space. However in our study, we are applying important shape transformations on the data which turn to be elements of a nonlinear space. As a consequence, classical coding approaches could not directly apply to these data and their nonlinear extension is required. As a solution, the Riemannian geometry of the well-known shape manifold can be used, offering several geometric tools which enables the extension of coding to nonlinear spaces. In the following, we start by describing the aforementioned coding technique. Then, we present basic notions of Riemannian geometry with a particular focus on the Kendall's shape space which is our manifold of interest in this dissertation.

3.2.1 Sparse representations

The notion of sparsity as a way to model signals has generated significant interest in the past decade (36, 38, 44). In particular, sparse representations have proved to be successful in various computer vision and machine learning problems including classification (45), segmentation (46), image denoising and inpainting (47), visual tracking (48) and face recognition (49), to cite a few. The basic assumption of this model is that, given a dictionary $\mathcal{D} = \{d_1, d_2, ..., d_N\}$ with N elements called *atoms*, a query x has a sparse vector representation $w \in \mathbb{R}^N$ with k non-zero elements under the dictionary $\mathcal{D}$:

$$x = \mathcal{D}w, \tag{3.1}$$

with $\|w\|_0 \leq k$, where $\|.\|_0$ represents the ℓ_0-pseudonorm which counts the number of non-zero elements in a vector. In Equation 3.1, the query x is represented as a linear combination of k atoms of the dictionary $\mathcal{D}$. The set of indices of these k active atoms is called the support Su. As such, this model assumes that the signal x resides in a low dimensional subspace, which is spanned by the k atoms of $\mathcal{D}$ that correspond to the support of w. An illustration of this model is given in Figure 3.1.

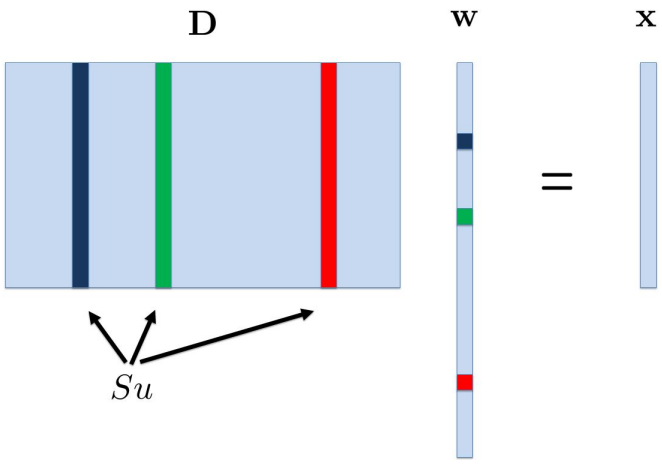


FIGURE 3.1: Schematic sparse model.

A common approach to estimate the optimal representation w is known as *sparse coding* which includes a penalty function $f(w)$ in the following optimization problem:

$$l_E(x, \mathcal{D}) = \min_w \|x - \sum_{i=1}^{N} [w]_i \, d_i\|_2^2 + \lambda f(w), \tag{3.2}$$

where $w \in \mathbb{R}^N$ denotes the vector of codes comprised of $\{[w]_i\}_{i=1}^N$, $f : \mathbb{R}^N \to \mathbb{R}$ is the sparsity inducing function defined as the ℓ_1 norm, and λ is the sparsity regularization parameter. Eq. (3.2) seeks to optimally approximate x (by $\hat{x}$) as a linear combination of atoms, *i.e.*, $\hat{x} = \sum_{i=1}^N [w]_i \, d_i$, while tacking into account a particular sparsity constraint on the codes, $f(w) = \|w\|_1$. This sparsity function has the role of forcing x to be represented as only a small number of atoms.

The problem of sparse coding assumes that the dictionary is known. In practice, when the dictionary is learned from the data, significant improvements can be made on the reconstruction of x. This problem is of wide interest and it is known as the dictionary learning problem. Several techniques have been developed to solve this problem and train dictionaries from data. Popular examples are the Method of Optimal Directions (MOD) (50) and the K-SVD method (51), which generalizes the K-means clustering algorithm to learn overcomplete dictionaries. Besides, using adaptive dictionaries (52) is usually the best choice in terms of the reconstruction performance that can be achieved. In addition, it enables to control the amount of induced sparsity over the reconstruction performance.

Formally, given a finite set of t training observations $\{x_1, x_2, ..., x_t\}$ in $\mathbb{R}^k$, learning adaptive dictionaries (52) is defined as to jointly minimize the coding cost over all choices of atoms and codes according to:

$$l_E(\mathcal{D}) = \min_{\mathcal{D},w} \sum_{i=1}^{t} \left\| x_i - \sum_{j=1}^{N} [w_i]_j d_j \right\|_2^2 + \lambda f(w_i). \tag{3.3}$$

To solve this non-convex problem, a common approach alternates between the two sets of variables, $\mathcal{D}$ and w, such that: (1) Minimizing over w while $\mathcal{D}$ is fixed is a convex problem (*i.e.*, sparse coding). (2) Minimizing Eq. (3.3) over $\mathcal{D}$ while w is fixed is similarly a convex problem.

In our work, we are interested in the extension of SCDL to the Kendall's shape space. Before introducing this manifold, in the following section, we present basic notions on Riemannian geometry.

3.2.2 Riemannian geometry

A manifold $\mathcal{M}$ is a Hausdorff topological space which locally resembles a Euclidean space $\mathbb{R}^n$, where n is the dimension of the manifold. The tangent space at a point on the manifold provides us with a vector space of tangent vectors that give an idea of direction on the manifold. A Riemannian manifold is differential and equipped with a metric on the tangent spaces. A Riemannian metric allows us to compute angle and length of directions (tangent vectors). A Riemannian metric is a continuous collection of inner products on the tangent space at each point of the manifold. It is usually chosen to provide robustness to some geometrical transformations. Furthermore, it makes it possible to define several geometric notions on a Riemannian manifold such as the geodesic distance between points on the manifold. In fact, points on a Riemannian manifold are connected with smooth curves. Assuming the Riemannian metric, one can compute the length of a given curve. The curves yielding the minimum distance for any two points of the manifold are called geodesics which are analogous to straight lines in $\mathbb{R}^n$. The length of a geodesic defines the *geodesic distance*. The geodesic distance induced by the Riemannian metric is the most natural measure of dissimilarity between two points lying on a Riemannian manifold. However, in practice, many other nonlinear distances which do not necessarily arise from Riemannian metrics can also be useful for measuring dissimilarity on manifolds (53). Two other essential operations on Riemannian manifolds are

the *logarithm map* (log) and *exponential map* (exp). As illustrated in Figure 3.2, the exponential map $\exp_Z(.) : T_Z\mathcal{M} \to \mathcal{M}$ projects a tangent vector from the tangent space at a point Z to the manifold. It guarantees that the length of the tangent vector is equal to the geodesic distance. The logarithm map $\log_Z(.) = \exp^{-1}(.) : M \to T_Z\mathcal{M}$ projects a point on the manifold to the tangent space $T_Z\mathcal{M}$ at another point.

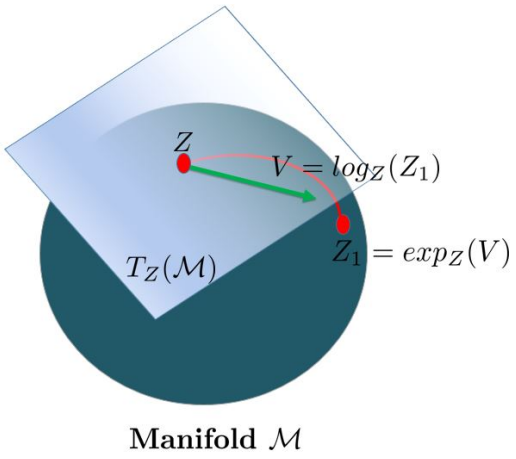


FIGURE 3.2: Exponential map and logarithm map operations on a Riemannian manifold. The red curve represents the geodesic connecting the points Z and Z_1. The green arrow represents the tangent vector V on the tangent space $T_Z(\mathcal{M})$.

Several Riemannian manifolds have been widely studied in the computer vision literature. Examples include the Lie group SO(3) formed by 3D rotation matrices, the unit n-sphere S^n formed by normalized histograms, the manifold of symmetric positive definite (SPD) matrices, the Grassmann manifold defined as linear subspaces of $\mathbb{R}^n$, and the Kendall's shape space (also known as the shape manifold). For an overview on Riemannian geometry and manifolds, we refer the reader to useful resources on the topic (54, 55). In the following section, we will outline the geometric properties of the manifold considered in this work, the Kendall's shape space.

3.2.3 Kendall's shape space

Let us consider a set of n landmarks in $\mathbb{R}^m$ $(m = 2, 3)$. To represent its shape, Kendall (12) proposed to establish equivalences with respect to shape-preserving transformations that are translations, rotations, and global scaling. Let $Z \in \mathbb{R}^{n\times m}$ represent a configuration of landmarks. To remove the translation variability, we follow (56) and introduce the notion of Helmert sub-matrix, a $(n-1) \times n$ sub-matrix of a commonly used Helmert matrix, to perform centering of configurations. For any $Z \in \mathbb{R}^{n\times m}$, the product $HZ \in \mathbb{R}^{(n-1)\times m}$ represents the Euclidean coordinates of the centered configuration. Let $\mathcal{C}_0$ be the set of all such centered configurations of n landmarks in $\mathbb{R}^m$, *i.e.*, $\mathcal{C}_0 = \{HZ \in \mathbb{R}^{(n-1)\times m} | Z \in \mathbb{R}^{n\times m}\}$. $\mathcal{C}_0$ is a $m(n-1)$ dimensional vector space and can be identified with $\mathbb{R}^{m(n-1)}$. To remove the scale variability, we define the pre-shape space to be: $\mathcal{C} = \{Z \in \mathcal{C}_0 | \|Z\|_F = 1\}$; $\mathcal{C}$ is a unit sphere in $\mathbb{R}^{m(n-1)}$

and, thus, is $m(n-1)-1$ dimensional. The tangent space at any pre-shape Z is given by: $T_Z(\mathcal{C}) = \{V \in \mathcal{C}_0 | \text{trace}(V^T Z) = 0\}$. To remove the rotation variability, for any $Z \in \mathcal{C}$, we define an equivalence class: $\bar{Z} = \{ZO | O \in SO(m)\}$ that represents all rotations of a configuration Z. The set of all such equivalence classes, $\mathcal{S} = \{\bar{Z} | Z \in \mathcal{C}\} = \mathcal{C}/SO(m)$ is called the *shape space* of configurations. The tangent space at any shape $\bar{Z}$ is $T_{\bar{Z}}(\mathcal{S}) = \{V \in \mathcal{C}_0 | \text{trace}(V^T Z) = 0, \ \ \text{trace}(V^T Z U) = 0\}$, where U is any $m \times m$ skew-symmetric matrix. The first condition makes V tangent to $\mathcal{C}$ and the second makes V perpendicular to the rotation orbit. Together, they force V to be tangent to the shape space $\mathcal{S}$. Assuming standard Riemannian metric on $\mathcal{S}$, the geodesic between two points $\bar{Z}_1, \bar{Z}_2 \in \mathcal{S}$ is defined as:

$$\alpha(t) = \frac{1}{\sin(\theta)}(\sin((1-t)\theta)Z_1 + \sin(t\theta)Z_2 O^*), \tag{3.4}$$

where $\theta = \cos^{-1}(\langle Z_1, Z_2 O^* \rangle)$, $\langle ., . \rangle$ is the inner product on $\mathcal{S}$, and O^* is the optimal rotation that aligns Z_2 with Z_1: $O^* = argmin_{O \in SO(m)} \|Z_1 - Z_2 O\|_F^2$. θ is also the geodesic distance between $\bar{Z}_1$ and $\bar{Z}_2$ in the shape space $\mathcal{S}$, representing the optimal deformation to connect $\bar{Z}_1$ to $\bar{Z}_2$ in $\mathcal{S}$. For $t = 0$, $\alpha(0) = \bar{Z}_1$ and for $t = 1$ we have $\alpha(1) = \bar{Z}_2$. Note that Kendall's shape space is a complete Riemannian manifold such that the logarithm map operator $\log_{\bar{Z}}$ is defined for all $\bar{Z} \in \mathcal{S}$ (see Section 3.2.3.1 for its definition). As a consequence, the geodesic distance between two configurations $\bar{Z}_1$ and $\bar{Z}_2$ can be computed as $\mathbf{d}_{\mathcal{S}}(\bar{Z}_1, \bar{Z}_2) = \| \log_{\bar{Z}_1}(\bar{Z}_2) \|_{\bar{Z}_1}$, where $\|.\|_{\bar{Z}_1}$ denotes the norm induced by the Riemannian metric at $T_{\bar{Z}_1}(\mathcal{S})$.

The case of planar shapes For $m = 2$, a 2D landmark configuration can be initially represented as a n-dimensional complex vector whose real and imaginary parts respectively encode the x and y coordinates of the landmarks. In this case, the pre-shape space is defined, after removing the translation and scale effects, as: $\mathcal{C} = \{z \in \mathbb{C}^{n-1} | \|z\| = 1\}$; $\mathcal{C}$ is a complex unit sphere of dimension $2(n-1)-1$. The rotation removal consists of defining, for any $z \in \mathbb{C}^{n-1}$, an equivalence class $\bar{z} = \{zO | O \in SO(2)\}$ that represents all rotations of a configuration z. The final shape space $\mathcal{S}$ is the set of all such equivalence classes $\mathcal{S} = \{\bar{z} | z \in \mathcal{C}\} = \mathcal{C}/SO(2)$. To measure the distance between two shapes $\bar{z_1}$ and $\bar{z_2}$, we define the most popular distance on the 2D Kendall's shape space, named the full Procrustes Distance (12), as

$$d_{FP}(\bar{z_1}, \bar{z_2}) = (1 - |\langle z_1, z_2 \rangle|^2)^{1/2}, \tag{3.5}$$

where $\langle \cdot, \cdot \rangle$ and $|.|$ denote the inner product in $\mathcal{S}$ and the absolute value of a complex number, respectively.

3.2.3.1 Geometric tools on the manifold

Considering the spherical structure of $\mathcal{C}$, analytic expressions of the logarithm and exponential maps are well defined (12, 56) and can be easily adapted to $\mathcal{S}$. These operations allow to compensate the lack of vector structure in the shape manifold by working on tangent spaces. In this section, we define

these operators, then use them to define a useful algorithm allowing to compute the mean shape of manifold-valued points, namely intrinsic mean.

Exponential map allows projection from a tangent space to the manifold. It applies the shooting vector to a source shape and provides the deformed (target) shape. It is defined, for any $V \in T_{\bar{Z}}(\mathcal{S})$, by,

$$\exp_{\bar{Z}}(V) = \left[\cos(\theta)Z + \frac{\sin(\theta)}{\theta}V\right]. \tag{3.6}$$

Logarithm map allows to map a point on the manifold to the tangent space at another point. It represents the *shooting vector* at the first shape (source) to the second shape (target). Its mathematical expression is given explicitly by,

$$\log_{\bar{Z}_1}(\bar{Z}_2) = \frac{\theta}{\sin(\theta)}(Z_2O^* - \cos(\theta)Z_1), \tag{3.7}$$

for source shape $\bar{Z}_1$ and target $\bar{Z}_2$, with θ as above.

Intrinsic mean An important advantage of the Riemannian approach is the ability to compute statistics on a set of manifold-valued points. One can use the notion of Karcher mean (57) to define an average shape. This represents an intrinsic mean and can be used as representative of a group of points on the manifold. Let $\{\bar{Z}_1, ..., \bar{Z}_k\}$ be a set of points on $\mathcal{S}$. We define an objective function $\Psi : \mathcal{S} \to \mathbb{R}$, $\Psi(\bar{Z}) = \sum_{i=1}^{k} d_{\mathcal{S}}(\bar{Z}_i, \bar{Z})^2$. The intrinsic mean is obtained by minimizing this objective function, which is commonly solved using a standard algorithm that we describe in Algorithm 1.

Algorithm 1 Computing intrinsic mean on $\mathcal{S}$

InputinputOutputoutput A set of shapes $\mathcal{Z} = \{\bar{Z}_j\}_{j=1}^m$ and ϵ_1, ϵ_2 small Initialize $\hat{\mu} \leftarrow Z_1, i \leftarrow 0$

repeat $|\bar{v}| < \epsilon_1$ Compute $v_j \leftarrow \log_{\hat{\mu}_i}(\bar{Z}_j)$ Compute average tangent vector $\bar{v} \leftarrow \frac{1}{k}\sum_{j=1}^m v_j$ Update $\hat{\mu}_i$ according to $\hat{\mu}_{i+1} \leftarrow \exp_{\hat{\mu}_i}(\epsilon_2\bar{v})$ $i \leftarrow i+1$ **return** $\hat{\mu}$, a sample Mean of $\mathcal{Z} = \{\bar{Z}_j\}_{j=1}^m$

As discussed previously, mappings to a tangent space allow to compensate the lack of vector structure on the shape manifold. We refer to this solution as an *intrinsic* approach. Another solution to seek a vector representation of the data is known as an *extrinsic* approach which embeds the manifold-valued data to a higher-dimensional vector space namely the Hilbert space.

3.2.3.2 Hilbert space embedding of the manifold

A Hilbert space $\mathcal{H}$ is an (often infinite-dimensional) inner product space which is complete with respect to the norm induced by the inner product. A Reproducing Kernel Hilbert Space (RKHS) is a special kind

of Hilbert space of functions on some nonempty set $\mathcal{S}$ in which all evaluation functionals are bounded and hence continuous. The inner product of an RKHS of functions on $\mathcal{S}$ can be defined by a bivariate function on $\mathcal{S} \times \mathcal{S}$, known as the reproducing kernel of the RKHS (53).

The SCDL algorithms depend only on the notion of inner product, which allows us to measure distances. Therefore, they can be easily extended to Hilbert spaces. The embedding of the shape manifold to RKHS brings the main advantage of transforming the nonlinear manifold into a vector space where one can directly apply standard (Euclidean) algorithms. In addition, it gives a richer representation of the data in a higher-dimensional space which helps identifying complex patterns. To define an inner product in $\mathcal{H}$, one can use a kernel function $f : (\mathcal{S} \times \mathcal{S}) \to \mathbb{R}$ which makes the resulting space a RKHS without the need of computing the actual vectors. This procedure, known as the *kernel trick*, is commonly used in machine learning when the designed algorithm only relies on measures of similarities, i.e., on inner products. However, to define a valid RKHS, the kernel function must be positive definite according to Mercer's theorem (43). In particular, for the 2D shape manifold, the authors in (40) proved the positive definiteness of the Procrustes Gaussian kernel $k_P : (\mathcal{S} \times \mathcal{S}) \to \mathbb{R}$ which is defined as

$$k_P(\bar{z_1}, \bar{z_2}) := exp(-{d_{FP}}^2(\bar{z_1}, \bar{z_2})/2\sigma^2), \tag{3.8}$$

where d_{FP} is the full Procrustes Distance defined in Eq.(3.5) and σ is the kernel parameter. This kernel is positive definite for all $\sigma \in \mathbb{R}$. In Section 3.4.2, it will be used for the extension of SCDL to Hilbert space.

3.3 Related work on representations of trajectories on Riemannian manifolds

In this section, we provide a brief literature overview of Riemannian SCDL approaches. Then, we discuss different representations of trajectories on Riemannian manifolds and highlight their limitations.

3.3.1 Riemannian sparse coding and dictionary learning

The basic definition of SCDL as defined in Section 3.2.1 assumes that the query points as well as the dictionary atoms are defined in vector space. However, most suitable image descriptors often lie to nonlinear manifolds (55). Thus, to perform SCDL on these data while respecting the geometric structure of Riemannian manifolds, the classical problem of SCDL needs to be extended to its nonlinear counterpart. The main issue here arises from the fact that the linear combination of atoms is not possible on a nonlinear manifold since the resulting point may not even be in the manifold, as illustrated in Figure 3.3. Previous works addressed this problem (36, 38, 41, 44, 58–60). For instance, a straightforward solution was proposed in (58, 61) by embedding the manifolds of interest into vector space, the tangent

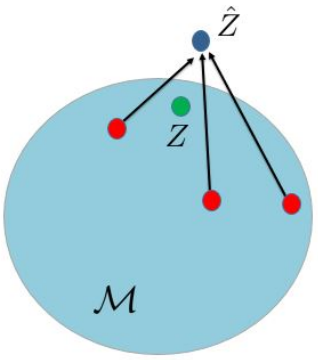


FIGURE 3.3: Z is a data point on the manifold $\mathcal{M}$ and the red circles represent the dictionary atoms. The linear approximation $\hat{Z}$ with the atoms may be out of the manifold $\mathcal{M}$.

space at a reference point. However, this solution does not take advantage of the entire Riemannian structure of the manifold since on a tangent space, only distances to the reference point are equal to true geodesic distances. To overcome this problem, Ho *et al.* (36) proposed a general framework for SCDL in Riemannian manifolds by working on the tangent bundle. Here, each point is coded on its attached tangent space into which the atoms are mapped. By doing so, only distances to the tangent point are needed. Their proposed dictionary learning method includes an iterative update of the atoms using a gradient descent approach along geodesics. This general solution essentially relies on mappings to tangent spaces using the logarithm map operator. Although it is well defined for several manifolds, analytic formulation of the logarithm map is not available or difficult to compute for others. Therefore, some studies (38, 39, 41, 59) proposed to embed the Riemannian manifold into RKHSs which are vector spaces where linear SCDL becomes possible. Recently, Harandi *et al.* (38) proposed to map the Grassmann manifolds into the space of symmetric matrices to overcome the latter problem and preserve several properties of the Grassmann structure. They also proposed kernelized versions of the SCDL algorithms to handle the nonlinearity of the data, similarly proposed in (62) for symmetric positive definite matrices. Throughout this dissertation, we will investigate two paradigms of Riemannian SCDL in the shape manifold with the aim of coding trajectories while tackling different challenges that we will discuss in the following section.

3.3.2 Trajectory representations on Riemannian manifolds

A sequence of points that evolve over time on a manifold can be seen as a time-series. In the case of a Riemannian manifold, a time-series is usually denoted by *trajectory*. Analysis of these trajectories is challenging due to the nonlinearity of underlying spaces. Several representations of landmark sequences lie to nonlinear manifolds. In many approaches, the Riemannian geometry of these manifolds is exploited to analyze the corresponding representations and a common solution to solve for their nonlinearity consists on mapping the manifold-valued data to a common tangent space. A popular example is given by the Lie group and its use for skeletal trajectory analysis. For instance, Vemulapalli *et al.* (37) proposed to represent 3D skeletal sequences in the product space of Special Euclidean (Lie) groups

$SE(3)^n$. To this end, for each frame of a sequence, the Euclidean transformation matrices encoding rotations and translations between different joint pairs are computed. Hence, the dynamics of these matrices is seen as a trajectory on $SE(3) \times ... \times SE(3)$. To overcome the nonlinear nature of this manifold, this representation is mapped to the Lie algebra $\mathfrak{se}(3)^n$ which is a vector space, the tangent space at the identity element. However, mapping points to a common tangent space may introduce undesirable distortions, especially when the mapped points are far from the tangent point. Aware of this limitation, the authors in (63) proposed a mapping of trajectories on Lie groups combining the usual logarithm map with a rolling map that guarantees a better flattening of trajectories on Lie groups. Taking another direction, Anirudh *et al.* (42) extended the framework of Transported Square-Root Velocity Fields (TSRVF) (64) by modeling trajectories of human actions on the Grassmann manifold and the product space of Lie groups $SE(3) \times ... \times SE(3)$. They tackled the problems of high-dimensionality of the feature space and its nonlinearity and proposed to learn a low-dimensional embedding using a manifold functional variant of principal component analysis. Hence, each trajectory is mapped to a single point in a low-dimensional Euclidean space. Another approach (32) proposed a different solution by extending the Kendall's shape theory to trajectories. Accordingly, translation, rotation, and global scaling are filtered out from each skeleton to quantify the shape. Then based on the TSRVF, they defined an elastic metric to jointly align and compare trajectories.

Riemannian trajectories were also considered in the analysis of 2D facial landmark sequences. Taheri *et al.* (65) proposed to represent 2D facial landmarks in the Grassmann manifold which makes the resulting representation invariant to affine transformations and hence robust to view variations. To capture facial expressions from these representations, the authors computed the velocity vectors between successive frames using the logarithm map. To obtain velocities in the same tangent space, they applied a parallel transport of these velocity vectors to a common tangent space. However, their method depends on the choice of this tangent space. In another work (66), 2D facial landmark sequences were first represented as trajectories of Gram matrices in the manifold of positive semidefinite matrices of rank 2. A similarity measure is then provided by temporally aligning trajectories while taking into account the geometry of the manifold.

Most of the methods described above share a common drawback which consists on mapping all the manifold-valued data to a reference tangent space. The major problem of this strategy is that distortions could be introduced especially when the mapped points are far from the tangent point. Moreover, comparing the resulting tangent vectors by computing distances between them is not accurate since only distances to the tangent point are equal to true geodesics. Therefore, our goal is to go beyond this drawback in the proposed intrinsic representation. On the other hand, to our knowledge, extrinsic approaches were not studied in the literature of human modeling with Riemannian trajectories. Hence, we also aim to explore this direction.

3.4 Human motion modeling framework

In this section, we present our human motion modeling, *i.e.* representation of actions and facial expressions, which is used throughout this thesis work for the tasks of human motion classification and generation. Given a set of sequences of 2D/3D skeletons or facial landmarks, our approach consists in three main steps:

- Embedding each frame (*i.e.* landmark configuration) of a sequence to the Kendall's pre-shape space $\mathcal{C}$ by filtering out translation and scale.
- Learning a dictionary from all training samples (*i.e.* static pre-shapes on $\mathcal{C}$). In this step, when an operation involves two configurations on $\mathcal{C}$ such as the logarithm map or the geodesic distance computation, rotation is filtered out by aligning one pre-shape into the second by applying the Procrustes algorithm (12). By doing so, we consider dictionary learning on the Kendall's shape space $\mathcal{S}$.
- Using the learned dictionary, each frame (*i.e.* pre-shapes on $\mathcal{C}$) of a sequence is coded using Riemannian sparse coding in $\mathcal{S}$ after filtering out rotation as done in the previous step.

As a main result of applying the proposed framework, each frame of an input sequence gives rise to a sparse code vector and frames that have similar shapes are expected to have similar code vectors. As a consequence, the input sequence will turn to a smoothly-evolving time-series (see Figure 3.4).

(A) **(B)**

FIGURE 3.4: Example results of the proposed framework. An input sequence is transformed to a smoothly-evolving sparse time-series. The X-axis represents the time frame. (A) 3D skeletal sequence. (B) 2D facial landmark sequence.

In this work, we investigate two approaches of Riemannian SCDL. The first is intrinsic and is based on projections to tangent spaces, while the second is extrinsic and relies on embeddings to RKHS. We illustrate these two approaches in Figure 3.5. In the following, we describe each of them.

3.4.1 Intrinsic approach

We propose to adapt a general intrinsic formulation of Riemannian SCDL (36) to the case of Kendall's shape space. This allows to transform a 2D/3D landmark configuration lying on the shape manifold to a sparse vector.

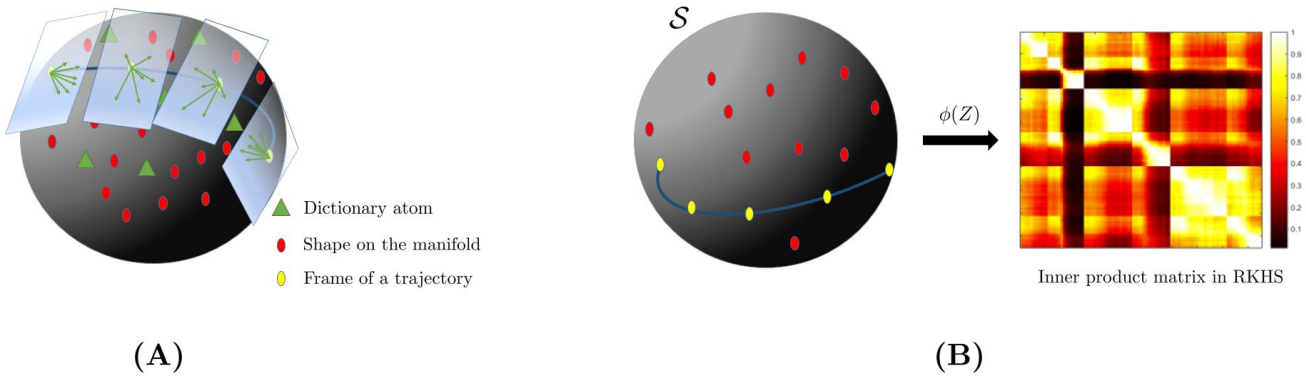


FIGURE 3.5: Illustration of the adopted solutions to overcome the nonlinearity of the shape manifold. (A) The intrinsic approach maps the data on the manifold to tangent spaces using the logarithm map opertor. (B) The extrinsic approach embeds the manifold-valued data to RKHS by computing the inner product matrix using a positive definite kernel function.

3.4.1.1 Intrinsic Sparse Coding

Let $\mathcal{D} = \{\bar{d}_1, \bar{d}_2, ..., \bar{d}_N\}$ be a dictionary on $\mathcal{S}$, and similarly the query $\bar{Z}$ is a point on $\mathcal{S}$. Accordingly, the problem of sparse coding involves the geodesic distance defined on $\mathcal{S}$ and thus the Euclidean formulation in Equation (3.2) (in Section 3.2.1) becomes

$$l_{\mathcal{S}}(\bar{Z}, \mathcal{D}) = \min_{w}(d_{\mathcal{S}}(\bar{Z}, F(\mathcal{D}, w))^2 + \lambda f(w)). \tag{3.9}$$

Here, $F : \mathcal{S}^N \times \mathbb{R}^N \to \mathcal{S}$ denotes an encoding function that generates the approximated point $\hat{\bar{Z}}$ on $\mathcal{S}$ by combining atoms with codes. Note that in the special case of Euclidean space, $F(\mathcal{D}, w)$ would be a linear combination of atoms. However, in the Riemannian manifold $\mathcal{S}$, we have forsaken the structure of vector space which makes the linear combination of atoms lying on $\mathcal{S}$ no longer applicable, since the approximated $\hat{\bar{Z}}$ may lie out of the manifold. An interesting alternative is the intrinsic formulation of Eq. (3.9), when considering that $\mathcal{S}$ is a complete Riemannian manifold, thus, the geodesic distance $d_{\mathcal{S}}(\bar{Z}, \bar{d}) = \| \log_{\bar{Z}}(\bar{d}) \|_{\bar{Z}}$ (as explained in Section 3.2.3). As a consequence, the cost function in (3.9) can be written as

$$l_{\mathcal{S}}(\bar{Z}, \mathcal{D}) = \min_{w} \| \sum_{i=1}^{N} [w]_i \log_{\bar{Z}}(\bar{d}_i) \|_{\bar{Z}}^2 + \lambda f(w), \tag{3.10}$$

where $\log_{\bar{Z}}$ denotes the logarithm map operator that maps each atom $\bar{d} \in \mathcal{S}$ to the tangent space $T_{\bar{Z}}(\mathcal{S})$ at the point $\bar{Z}$ being coded, and $\|.\|_{\bar{Z}}$ is the norm induced by the Riemannian metric at $T_{\bar{Z}}(\mathcal{S})$. Mathematically, this allows to partially compensate the lack of vector space structure on $\mathcal{S}$, as illustrated in Figure 3.6. To avoid the solution $w = 0$, we imposed in Eq. (3.10) an important additional affine constraint defined as $\sum_{i=1}^{N} [w]_i = 1$. By this formulation of sparse coding, we only compute distances to the tangent point, hence we avoid the commonly induced distortions when working in a reference tangent

space. By substituting the logarithm map by its explicit formulation in Eq. (3.10), we have

$$l_{\mathcal{S}}(\bar{Z}, \mathcal{D}) = \min_{w} \| \sum_{i=1}^{N} [w]_i \frac{\theta}{\sin(\theta)} (d_i O^* - \cos(\theta) Z) \|_{\bar{Z}}^2 + \lambda f(w). \tag{3.11}$$

In practice, Eq. (3.11) is computed by first finding the optimal rotation O^* between Z and each atom d_i via the Procrustes algorithm (12). Then, we solve for w using the state-of-the-art CVXPY optimizer (67). In Algorithm 2, we provide a summary of the sparse coding approach on the shape manifold.

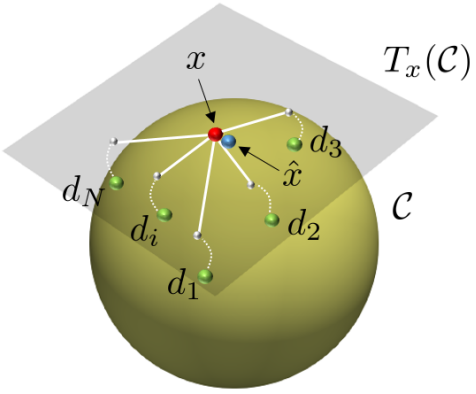


FIGURE 3.6: Pictorial of the sparse coding approach on the pre-shape space $\mathcal{C}$. The approximation of $x \in \mathcal{C}$ could be viewed as a weighted intrinsic mean of the atoms of a dictionary $\mathcal{D} = \{d_i\}_{i=1}^N$.

Algorithm 2 Riemannian sparse coding algorithm

InputinputOutputoutput Dictionary $\mathcal{D} = \{\bar{d}_i\}_{i=1}^N$, $\bar{d}_i \in \mathcal{S}$; $\bar{Z} \in \mathcal{S}$ (query) Sparse codes vector $w^*$of the query $\bar{Z}$ */* Projection of $\mathcal{D}$ into $T_{\bar{Z}}(\mathcal{S})$ */*

for $i = 1$ to N **do** $\mathcal{V}_i \leftarrow \log_{\bar{Z}}(\bar{d}_i)$

$w^* = \text{argmin}_w \| \sum_{i=1}^N [w]_i \mathcal{V}_i \|_2^2 + \lambda f(w)$
return Sparse codes vector $w^*$of the query $\bar{Z}$

3.4.1.2 Intrinsic Dictionary Learning

Learning a discriminative dictionary $\mathcal{D}$ typically yields accurate reconstruction of training samples and produces discriminative sparse codes. We propose a dictionary learning algorithm based on the sparse coding framework described above. Let $\mathcal{D} = \{\bar{d}_1, \bar{d}_2, ..., \bar{d}_N\}$ be a dictionary on $\mathcal{S}$, and similarly $\{\bar{Z}_1, \bar{Z}_2, ..., \bar{Z}_t\}$ is a set of t training samples on $\mathcal{S}$. Similarly to the sparse coding problem, we introduce in Eq. (3.3) the geodesic distance defined on $\mathcal{S}$ computed as $d_{\mathcal{S}}(\bar{Z}, \bar{d}) = \| \log_{\bar{Z}}(\bar{d}) \|_{\bar{Z}}$. As a consequence, the problem of dictionary learning on Kendall's shape space is written as

$$\min_{\mathcal{D}, w} \sum_{i=1}^{t} \left\| \sum_{j=1}^{N} [w_i]_j \log_{\bar{Z}_i} \bar{d}_j \right\|_{\bar{Z}_i}^2 + \lambda f(w_i), \tag{3.12}$$

with the important affine constraint $\sum_{j=1}^{N} [w]_j = 1$. Similarly to the Euclidean case, the optimization problem can be solved by iteratively performing sparse coding while fixing $\mathcal{D}$, and optimizing $\mathcal{D}$ while fixing the sparse codes. Algorithm 3 summarizes the different steps of dictionary learning.

Algorithm 3 Riemannian dictionary learning algorithm

InputinputOutputoutput Training set $\mathcal{Z} = \{\bar{Z}_j\}_{j=1}^{m}$, where $\bar{Z}_j \in \mathcal{S}$; Dictionary $\mathcal{D} = \{\bar{d_i}\}_{i=1}^{N}, \bar{d_i} \in \mathcal{S}$ Dictionary initialization using Bayesian clustering and PGA (see Section 3.4.1.2))
/ Processing */*

for $k = 1$ to *nIter* **do** Sparse Coding using Algorithm 2; w_j^* are the output sparse codes.
*/*Dictionary update Step */*

for $a = 1$ to N **do** Updating atom i using line-search algorithm
$w^* = \text{argmin}_w \| \sum_{i=1}^{N} [w]_i \mathcal{V}_i \|_2^2 + \lambda f(w)$
$\bar{d_a}^* = \text{argmin}_{\bar{d_a}} \sum_{j=1}^{m} \| [w_a] \log_{\bar{Z}_j}(\bar{d_a}) \|_{\bar{Z}_j}^2 + \| \sum_{i=1; i \neq a}^{N} [w_i] \log_{\bar{Z}_j}(\bar{d_i}) \|_{\bar{Z}_j}^2 + \lambda f(w_j)$ **return** Dictionary $\mathcal{D} = \{\bar{d_i}\}_{i=1}^{N}, \bar{d_i} \in \mathcal{S}$

An efficient dictionary initialization for faster learning The performance of sparse coding depends on the number of the dictionary elements N, and an empiric choice of N can be time consuming, especially when it comes to large datasets. As a solution, we propose an initialization step that enables an automatic inference on N and accelerates the convergence of the dictionary learning algorithm. To this end, we propose to cluster the training shapes by adapting the Bayesian clustering of shapes of curves method proposed in (68). The latter brings the advantage of automatically inferring the number of clusters from a set of data. From each cluster, main representatives will then be selected to constitute the initial dictionary.

Kernel-based clustering of shapes for dictionary learning – In Figure 3.7, we show a qualitative result of clustering 3D skeletal shapes and 2D facial shapes. To achieve it, an inner product matrix is first computed from the training data based on the kernel function defined in Section 3.2.3.2. Note that in the 3D case, this kernel is positive definite for only certain values of the kernel parameter σ. Thus, its empiric choice is required to seek positive definiteness. The inner product matrix is then modeled using a Wishart distribution. To allow for an automatic inference on the number of clusters, prior distributions are carefully assigned to the parameters of the Wishart distribution. Then, posterior is sampled using a Markov chain Monte Carlo procedure based on the Chinese restaurant process for final clustering. For details, we refer the reader to (68), where the authors presented the Bayesian clustering method to segment shapes of curves. In our work, we propose to adapt their approach to cluster shapes of 2D/3D landmark configurations where the only difference resides on the computation of the distance matrix.

Atoms inference from clusters – To select the best representatives of a cluster, the mean shape is a suitable candidate but it is not sufficient to summarize the intra-cluster variability. For that, we propose to apply principal geodesic analysis (PGA), first proposed by (69). Specifically, all elements of a culster

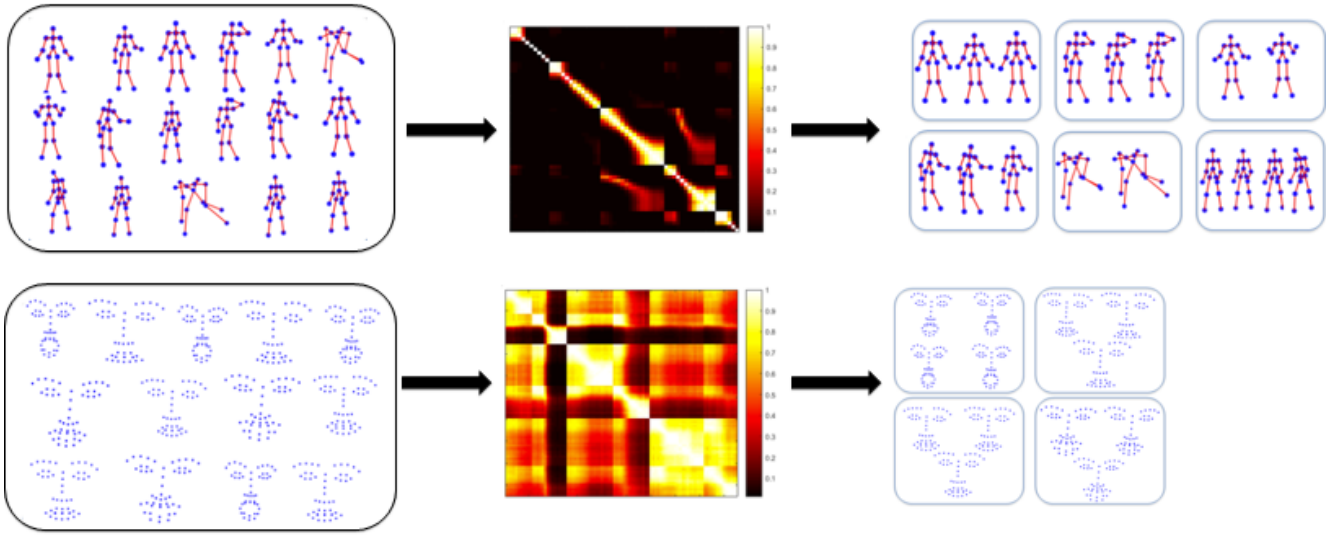


FIGURE 3.7: Pictorial of the proposed clustering approach. Landmark configurations (left) are mapped from the the shape manifold to RKHS by computing the inner product matrix from the data (Middle). Bayesian clustering is then applied on this matrix to construct the final clusters (right) whose number is automatically inferred.

are mapped to the tangent space at the mean shape $T_{\bar{\mu}}(\mathcal{S})$. Then, principal component analysis (PCA) is applied in this vector space to induce the main components (tangent vectors). Finally, the induced vectors are mapped back to the manifold $\mathcal{S}$ using the exponential map operator to represent inital atoms. This procedure is applied to all the clusters. Note that an important advantage of performing PGA in each cluster rather than in the whole training set is to avoid the problematic case of having points in the manifold that are far from the tangent point.

3.4.2 Extrinsic approach

The SCDL algorithms only depend on the notion of inner product. In the following, we will discuss how they can be easily extended to RKHS using the Procrustes Gaussian Kernel.

3.4.2.1 Kernel Sparse Coding

A closed-form solution of kernel sparse coding is proposed in (38). To derive it, let us first define $\phi : \mathcal{S} \to \mathcal{H}$ a mapping to RKHS induced by the kernel $k(\bar{z_1}, \bar{z_2}) = \phi(\bar{z_1})^T \phi(\bar{z_2})$, where $\bar{z_1}, \bar{z_2} \in \mathcal{S}$. For a query shape $\bar{z} \in \mathcal{S}$, extending Eq. 3.2 to RKHS yields

$$l_{\mathcal{H}}(\bar{z}, \mathcal{D}) = \min_{w} \|\phi(\bar{z}) - \sum_{i=1}^{N} [w]_i \, \phi(\bar{d_i})\|_2^2 + \lambda f(w), \tag{3.13}$$

with $\sum_{i=1}^{N} [w]_i = 1$. In Eq. 3.13, since the sparsity term depends entirely on w, only the reconstruction term needs to be kernelized. Expanding the latter gives

$$\|\phi(\bar{z}) - \sum_{i=1}^{N} [w]_i \, \phi(\bar{d}_i)\|_2^2 = \phi(\bar{z})^T \phi(\bar{z})$$

$$= -2 \sum_{i=1}^{N} [w]_i \, \phi(\bar{d}_i)^T \phi(\bar{z}) + \sum_{i,j=1}^{N} [w]_i \, [w]_j \, \phi(\bar{d}_i)^T \phi(\bar{d}_j)$$

$$= k(\bar{z}, \bar{z}) - 2w^T k(\bar{z}, D) + w^T K(D, D) w, \tag{3.14}$$

where $k(\bar{z}, \mathcal{D})$ is the N-dimensional kernel vector computed between the query $\bar{z}$ and the dictionary atoms, and $K(\mathcal{D}, \mathcal{D})$ is the $N \times N$ kernel matrix computed between the atoms. An efficient solution of kernel sparse coding can be obtained by considering $U\Sigma U^T$ as the SVD of the symmetric positive definite kernel $K(D, D)$, and $k(\bar{z}, \bar{z})$ as a constant term (independent on w). Thus, Eq. 3.14 can be written as the least-squares problem in $\mathbb{R}^N$: $\min_w \|\tilde{z} - \tilde{D}w\|_2^2$, where $\tilde{D} = \Sigma^{1/2} U^T$ and $\tilde{z} = \Sigma^{-1/2} U^T k(\bar{z}, D)$ (we refer to (38) for the proof). In this work, this approach is applied in the Kendall's shape space by using the Procrustes Gaussian Kernel defined in Section 3.2.3.2.

3.4.2.2 Kernel Dictionary learning

Similarly to Euclidean dictionary learning, the extrinsic Riemannian formulation is based on an alternating optimization strategy to update weights and atoms. While the first step is obtained with extrinsic sparse coding presented above, the second is presented in what follows. Given the codes from the first step, the problem of dictionary learning can be viewed as optimizing Eq. (3.13) over $\mathcal{D}$. The main idea here is to represent $\mathcal{D}$ as a linear combination of the training samples Y in RKHS according to the Representer theorem (70). The resulting weights for the M training samples are stacked in a $M \times N$ matrix V, which gives $\phi(D) = \phi(Y)V$. Since only the first term in Eq. 3.13 depends on $\mathcal{D}$, the problem of dictionary update can be written as $U(V) = \|\phi(Y) - \phi(Y)VW\|_2^2$, where W is the $N \times M$ matrix of sparse codes obtained from the first step. The latter can be expanded to:

$$U(V) = Tr(\phi(Y)(I_M - VW)(I_M - VA)^T \phi(Y)^T)$$

$$= Tr(K(Y, Y)(I_M - VW - W^T V^T + VWW^T V^T)).$$

To obtain the updated dictionary that is now defined by V, the gradient of $U(V)$ is zeroed out w.r.t V. This gives $V = (WW^T)^{-1} W = W^{\dagger}$, where † is the pseudo-inverse operator.

3.5 Properties of the latent space

Sparse coding of trajectories give rise to time-series defined in vector space which we refer to as the latent space. In this section, we describe the main properties of the obtained representations in this space.

3.5.1 Reconstruction of trajectories

As illustrated in Figure 3.8, an important advantage of the intrinsic approach is that it enables to recover a sparse code back to the original manifold. This can be extremely useful for visualization purposes and for certain tasks such as motion generation. Shape reconstruction is achieved with respect to the pre-learned dictionary by applying the weighted intrinsic mean algorithm described in Algorithm 4. The idea here is based on the intrinsic mean, *i.e.* Algorithm 1, where we now: 1) Initialize the approximated shape $\hat{\mu}$ as the linear combination of codes and the corresponding atoms from the dictionary. By doing so, we can obtain a good approximation of the original shape, knowing that the code vector is sparse and supposing that atoms with non-zero coefficients (elements of code vector) are the closest to the tangent point; 2) Compute tangent vectors as $\log_{\hat{\mu}_i}(w_j \bar{D}_j)$ and the mean tangent vector $\bar{v}$. Then update $\hat{\mu}$ by moving $\bar{v}$ to the average direction. This is done iteratively until the norm of $\bar{v}$ is sufficiently close to zero.

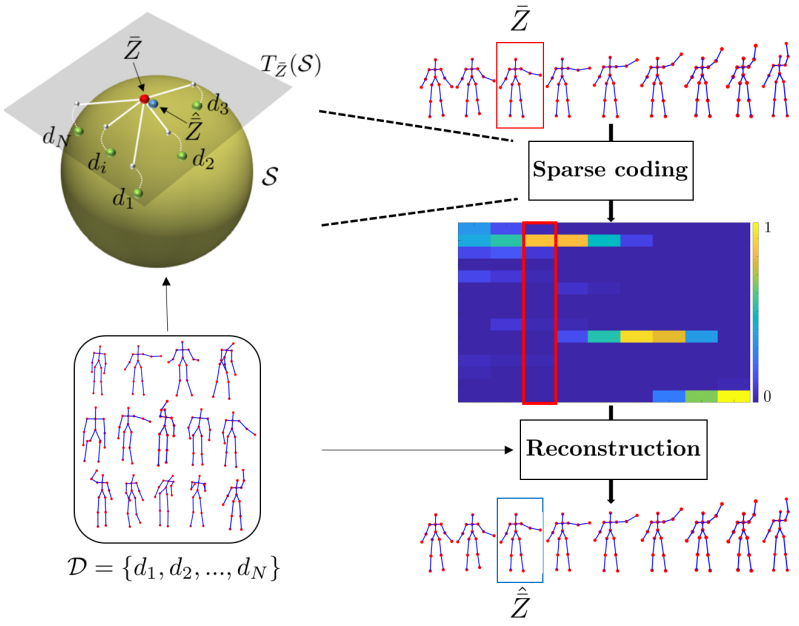


FIGURE 3.8: Given the pre-trained dictionary $\mathcal{D}$, skeletal trajectories in the shape manifold can be reconstructed from the space of sparse codes using the weighted intrinsic mean algorithm.

3.5.2 Efficient tangent space projections

As explained in Section 3.3.2, many approaches that model sequences of human landmarks as trajectories on Riemannian manifolds (32, 37, 65) share a major drawback, that of mapping manifold-valued points

Algorithm 4 Weighted intrinsic mean for shape reconstruction from code

InputinputOutputoutput A vector of codes shapes $\{w_j\}_{j=1}^m$, a dictionary $\mathcal{D} = \{\bar{D}_j\}_{j=1}^m$ and ϵ_1, ϵ_2 small Initialize $\hat{\mu} \leftarrow \frac{1}{m}\sum_{j=1}^m w_j D_j, i \leftarrow 0$

repeat $|\bar{v}| < \epsilon_1$ Compute $v_j \leftarrow \log_{\hat{\mu}_i}(w_j \bar{D}_j)$ Compute average tangent vector $\bar{v} \leftarrow \frac{1}{k}\sum_{j=1}^m v_j$ Update $\hat{\mu}_i$ according to $\hat{\mu}_{i+1} \leftarrow \exp_{\hat{\mu}_i}(\epsilon_2 \bar{v})$ $i \leftarrow i+1$ **return** $\hat{\mu}$, the reconstructed shape

to a reference tangent space which may introduce distortions especially when points are far from the tangent point. In contrast, in our coding approach, each point is coded on its attached tangent space, where the dictionary atoms are mapped (see Figure 3.9 for illustrations of tangent space approximation strategies). By doing so, we only compute distances to the tangent point which are equal to true geodesics. Furthermore, even though we map all the atoms to a tangent space where some of them may be far from the tangent point, in practice, distortions are usually avoided since our sparse coding scheme tends to code a point using the closest atoms to it and attributes zeros to the rest, distant atoms. As a consequence, assuming that the dictionary is well learned (*i.e.* the atoms cover all the space of training shapes), our approach considerably alleviates the non-trivial problem of trajectory distortions caused by tangent space approximations.

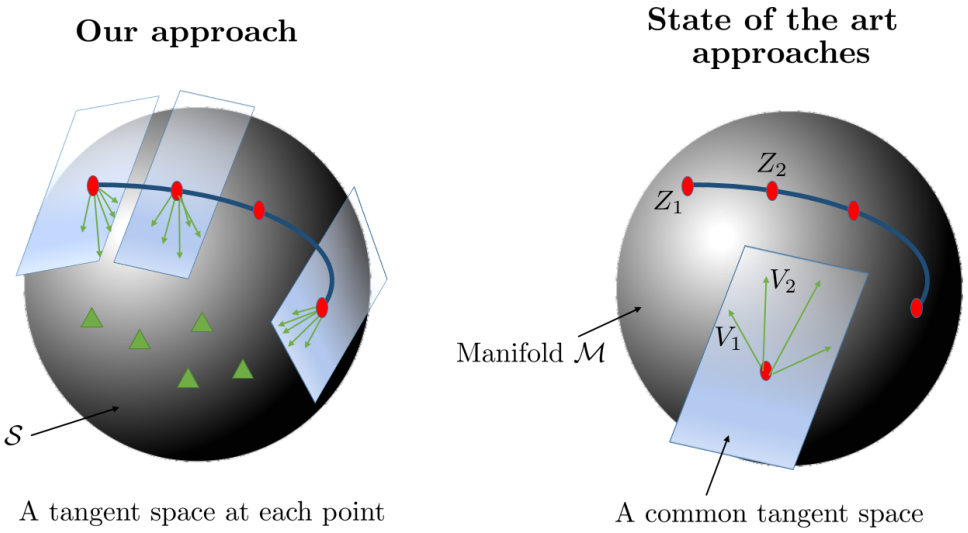


FIGURE 3.9: Schematic tangent space projections using our method compared to state-of-the-art.

3.5.3 Denoising of skeletal shapes

Skeletal joints obtained from low cost sensors are often noisy leading to abnormal skeletal poses. This is a non-trivial issue for applications as action recognition since the performance of a landmark-based recognition approach relies essentially on the accuracy of the extracted landmarks. One advantage of the proposed sparse coding approach is that it naturally enables denoising of skeletons when assuming a clean

dictionary. Figure 3.10 presents an illustration of this property using data collected from the state-of-the-art MSR-Action 3D dataset (71). Here, only certain joints are poorly estimated in a body pose, *e.g.,* right and left knees, hence the global shape is preserved. Sparse coding attempts to approximate this shape using the closest atoms. Assuming that the dictionary does not contain abnormal shapes, the resulting approximation is expected to recover the input abnormal shape. The question now is how to obtain a clean dictionary? Recall that to train a dictionary, skeletons are collected from all training sequences. We point out that in general, noise appears in only certain frames of a sequence. In addition, actions usually evolves smoothly over time. Considering these two information, one can compute distances between successive frames and discard skeletons with a relatively great distance to the previous.

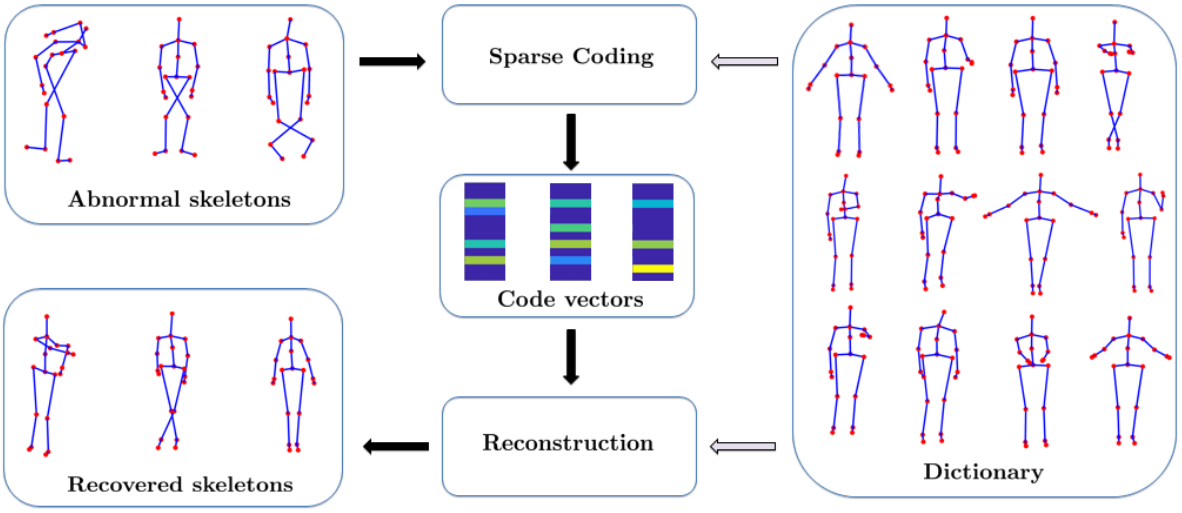


FIGURE 3.10: Denoising of skeletal shapes using sparse coding. Abnormal skeletons are presented in the top left. Code vectors (middle) are obtained after sparse coding. They are then reconstructed with respect to the dictionary (right) to recover the abnormal skeletons.

3.5.4 On the vector structure of the latent space

Standard notions of statistics (*e.g.,* mean computation, interpolation, etc.) and analysis of time-series (*e.g.*, temporal alignment, temporal modeling, etc.) need significant modification to account for the nonlinearity of the shape manifold. In most cases, these operations become highly involved in terms of computational complexity, and often result in iterative procedures further increasing the computational load. The proposed sparse coding approach allows to exploit the linear nature of the latent variables as well as their low-dimensionality to compute standard statistics on the data and apply standard techniques to process time-series, rather than performing them directly on the manifold. For instance, one can compute the mean code linearly and recover it back to the manifold and still obtain a point on the manifold that is very close to the intrinsic mean shape of the same data. An illustration of this is given in Figure 3.11. One can also interpolate linearly between latent variables and obtain meaningful interpolates when mapped back to the shape manifold, see Figure 3.12. Now, we turn our attention to our main concern which is to consider time-evolving shapes that represent actions or facial expressions.

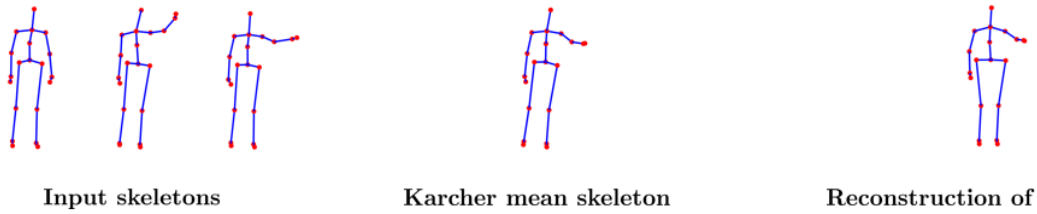


FIGURE 3.11: Mean shape computation on a set of skeletons. Left: Input 3D skeletons in the Kendall's shape space. Middle: Mean shape computed using the intrinsic mean algorithm. Right: Mean shape computed by first sparse coding the input skeletons, then computing the mean code and reconstruct it with the dictionary.

As explained previously, sparse coding of each shape of a trajectory gives rise to a smoothly-varying time-series that is naturally defined in vector space. Thereby, assuming the linearity of the latent space, one can apply machine learning techniques as well as post-processing methods dedicated to Euclidean time-series (up or down temporal resampling, denoising of time-series, temporal alignment of different time-series, etc.), without any manifold assumption.

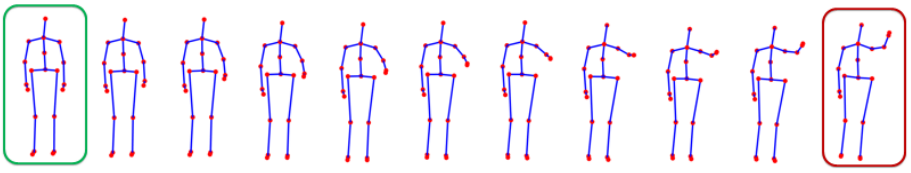


FIGURE 3.12: An example of linear interpolation between two latent variables: source (left) and target (right). Shapes in this figure are the result of mapping the interpolates from the latent space to the manifold.

3.6 Facial Expression and Action Recognition with Sparse Representations

In the previous section, we proposed a novel framework to encode trajectories in the shape manifold. We explored two alternatives of Riemannian SCDL allowing to represent human actions and facial expressions as sparse times-series in vector space. The first is an intrinsic solution where SCDL is performed on the manifold tangent spaces while the second is based on embedding the manifold-valued data to Hilbert spaces. We demonstrate the effectiveness of these sparse representations in two recognition tasks: the 3D action recognition and 2D facial expression recognition (both micro and macro expressions). Specifically, we will show that the intrinsic coding approach is efficient to code 3D shape trajectories while the extrinsic method is suitable for 2D trajectories. In the context of the addressed classification problems, these coding techniques bring two main advantages: (1) Sparse coding of shapes is performed with respect to a Riemannian dictionary. Hence, the resulting sparse times-series are expected to be more discriminative than the data themselves. In addition, they are robust to noise, knowing that SCDL is a

powerful denoising tool as demonstrated in Section 3.5.3; (2) Using sparse time-series as discriminative features allows us to perform both temporal modeling and classification in vector space, avoiding the more difficult task of classification on the manifold. To this end, we will study and compare two different pipelines for temporal modeling and classification. The first is based on a standard machine learning technique while the second is a deep learning approach based on RNNs. An overview of the proposed approaches is given in Figure 3.13.

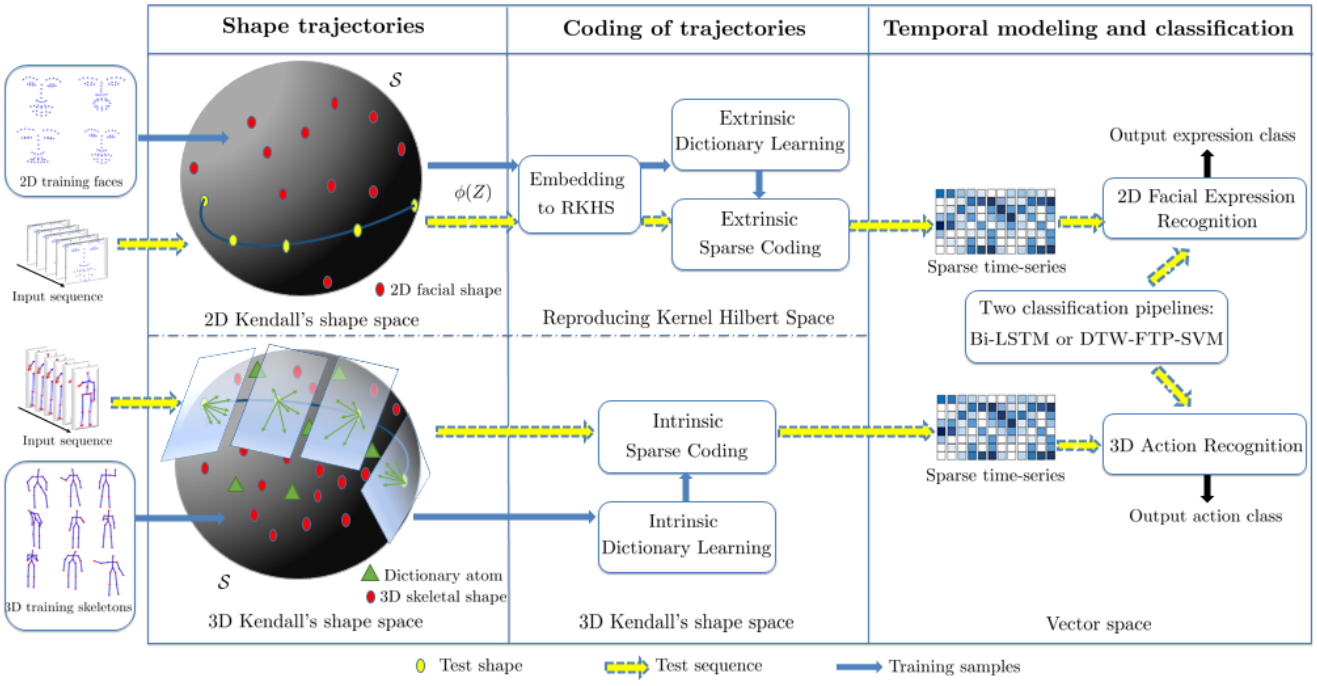


FIGURE 3.13: Overview of the proposed frameworks. Trajectories of 2D facial expressions (respectively 3D actions) are encoded using extrinsic (respectively intrinsic) SCDL in the Kendall's shape space. Temporal modeling and classification are then performed on the obtained time-series in vector space.

Our main contributions in this section are:

- Application of proposed the framework to: 3D action recognition, 2D micro- and 2D macro- facial expression recognition. Extensive experiments are conducted on seven commonly-used datasets to show the competitiveness of the proposed approach to state-of-the-art.
- A comparative study on the intrinsic and extrinsic paradigms of Riemannian SCDL in the 2D and 3D shape manifolds. To the best of our knowledge, this work is the first to apply and compare both approaches to dynamic 2D and 3D shapes.

3.6.1 Related work

The typical framework of human motion recognition using skeletons or faces comprises the following phases, see Figure 3.14. A feature extraction step to capture lower-level information from the data.

This step can account to the spatial information only or also considers the temporal dimension. We consider two main categories of features: hand-crafted features and learned features by means of deep learning. Such representations can be used as input to a temporal modeling stage to capture the temporal dependencies in time-series. The output is then fed to a classification stage which can consist on a standard classifier (e.g. k-nearest neighbors, SVM, etc.) or a deep learning framework.

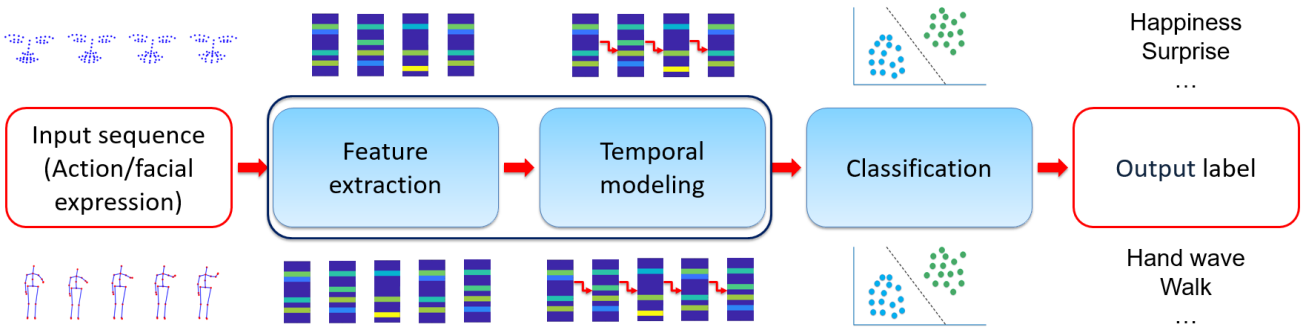


FIGURE 3.14: Overview of a typical landmark-based action/facial expression recognition approach.

Most of the landmark-based approaches in the literature follow the above pipeline to represent and classify 3D actions or 2D facial expressions. At large, we can regroup them into two main categories: classical methods which are based on hand-crafted feature extraction and deep learning methods which automatically learn features by designing suitable network architectures and objectives for the task at hand, see Figure 3.15.

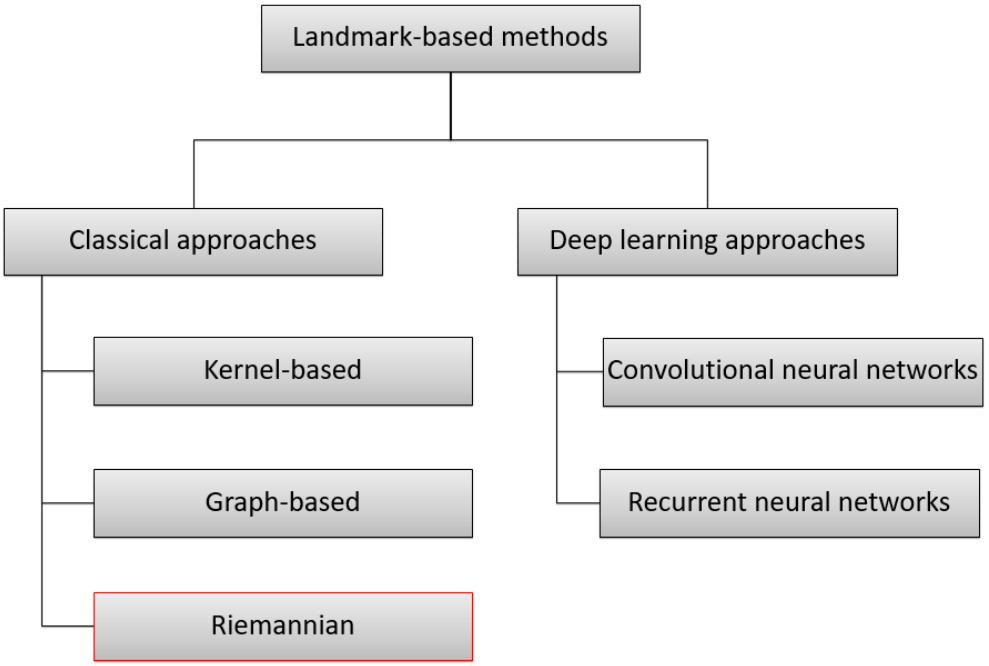


FIGURE 3.15: Our categorisation of state-of-the-art approaches. Representations of 2D/3D sequences of landmarks (skeletons or faces) can be either hand-crafted or learned. The former representations can be categorised into: kernel-based, graphical models or Riemannian. The latter can be achieved using deep learning techniques.

In one hand, we categorise the classical methods into three groups:

Graph-based methods Graph is a powerful tool for modeling structured objects. Since the representation of a human skeleton or face is in essence composed of points that are connected to each other, it is natural to perceive it as a graph. Therefore, several approaches opted for graphs to model the spatial dependencies between human joints as well as their temporal evolution. Probabilistic models such as Hidden Markov models (HMMs), which also falls in this category, have been widely used to model the temporal evolution of human pose sequences.

Kernel methods attempt to compute similarities between data-points (*e.g.* landmark configurations, features extracted from them, etc.) using kernel functions. This enable them to operate in a high-dimensional, implicit feature space. The latter is also called the inner product space since only inner products between data-points are computed, without ever computing the coordinates of the data in the new space. These approaches are known to bring a richer representation of the original data since the inner product space is usually higher-dimensional which helps classification methods to identify complex patterns.

Riemannian methods Several representations of landmark sequences may lie to nonlinear manifolds where traditional computational tools and machine learning techniques cannot be directly applied. In fact, in contrast to vector spaces, these manifolds are characterized with nonlinear topology that algorithms have to take into account. As an example, to compute the similarity between two points on a nonlinear manifold, the Euclidean distance is no longer suitable since it does not represent the real proximity between them. As a solution, several approaches studied the Riemannian geometry of these manifolds to define a metric which is obtained by defining a smoothly varying inner product on each tangent space of the manifold. Hence, the vector space structure of these tangent spaces can be exploited to overcome the nonlinearity of Riemannian manifolds.

On the other hand, feature learning using deep learning techniques has been receiving increasing attention in recent years due to their ability of designing powerful features without the need of heavy human labor and domain expert knowledge to develop effective feature extraction methods. Their success is also attributed to the access to Graphics Processing Units (GPUs) as well as to large scale labeled datasets which allow to design networks with millions of parameters. These deep learning techniques consist of multi-layer neural networks with a number of connected units (neurons). These layers represent three types: input layer with input units receiving information to be processed, output layer with output units giving the result of the network, and hidden layers with hidden units which process the data. The training objective of these neural networks consists of learning the weights of the connections between neurons in order to determine the mapping between the input and the output. The most popular deep learning methods that are commonly used in action and facial expression recognition are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The former are based on the powerful convolution operation and belong to the category of feed-forward neural networks where the information moves only in one direction, that of the output layer. The latter are a kind of neural networks that takes

sequential input and infer sequential output by sharing recurrent connections (parameters) between time steps, in contrast to CNNs.

In the following, we review some examples of methods from the categories discussed above in the context of 3D action recognition and 2D facial expression recognition, with a focus on Riemannian approaches since our proposed framework is also Riemannian.

3.6.1.1 3D action recognition

Over the last decade, many techniques have been proposed for action recognition from 3D skeletal data. As shown in Figure 3.15, several methods are based on the extraction of hand-crafted features. Examples include the use of spatio-temporal graphical models to represent and classify action sequences. Considering a human action as transitions between body poses over time, G. Hernando *et al.* (72) proposed a forest-based classifier called transition forests to discriminate both static pose information and temporal transitions between pairs of two independent frames. Another work (73) modeled a human action as a set of semantic parts called *motionlets* obtained by tracking then segmenting the trajectory of each joint. By combining the motionlets and their spatio-temporal correlations, they proposed an undirected complete labeled graph to represent a video, and a subgraph-pattern graph kernel to measure the similarity between graphs, then to classify videos.

In the category of kernel-based approaches, two kernel-based tensor representations named sequence compatibility kernel (SCK) and dynamics compatibility kernel (DCK) were introduced in (74). 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 models a sequence dynamics as the spatio-temporal co-occurrences of the joints. Tensors are then formed from these kernels to train SVM.

The above-mentioned approaches did not make any manifold assumptions on the data representation. However, several shape representations and their dynamics often lie to nonlinear manifolds. As discussed in the category of Riemannian approaches, many approaches exploited the Riemannian geometry of nonlinear manifolds to analyze skeletal sequences. The latter are considered in different Riemannian manifolds such as the Lie group, the Grassmann manifold, the shape manifold, etc. Since the work presented in this dissertation is based on a representation in the shape manifold, we consider it as part of this category. In Section 3.3.2, we described some representations of 3D skeletal sequences on Riemannian manifolds. Here, we discuss their corresponding temporal modeling and classification steps. In (37), sequences are represented as trajectories in the product space of Lie groups $SE(3) \times ... \times SE(3)$ then mapped to the Lie algebra $\mathfrak{se}(3)^n$, the tangent space at the identity element. To handle the rate variability in human actions, the obtained time-series in $\mathfrak{se}(3)^n$ are aligned by means of the popular dynamic time warping (DTW) algorithm. To handle temporal misalignment as well as the noise present in the data, the aligned sequences are processed using Fourier temporal pyramid (FTP). Finally, the obtained features are classified using a one-vs-all SVM. In (63), the authors also applied DTW on the Lie

group representation of actions then classified final curves using a one-vs-all linear SVM after mapping the curves to the Lie algebra using rolling maps which ensures a better flattening of the Lie group. DTW and SVM were also used in (42) to classify their low-dimensional representation of actions which were initially represented as trajectories in the Lie group. Based on the same trajectory representation on $SE(3)$, the authors in (75) proposed a deep learning framework in Lie groups to recognize actions. The proposed architecture includes several special layers (*e.g.* RotMap layer, RotPooling layer, etc.) which accounts to the geometry of the manifold. Taking another direction, the authors in (32) represented skeletal sequences as trajectories in the Kendall's shape space then modeled them using TSRVF. For classification, they computed the mean trajectories of each class and for each trajectory they extracted a feature vector formed by distances to mean trajectories of each class. Finally, they used SVM to classify these feature vectors. Recall that a common drawback of many of these approaches is the mapping of all manifold-valued data to a reference tangent space which may introduce distortions. In contrast, our coding approach in the shape manifold avoids such a problem.

On the other hand, RNNs, which belong to the category of deep learning approaches according to our categorisation in Figure 3.15, have showed promising performance when applied to 3D action recognition. For instance, HBRNN-L (76) applied bidirectional RNNs hierarchically by dividing a skeleton into five parts of neighboring joints. Then, each is separately fed into a bidirectional RNN before fusing their outputs to form the upper-body and the lower-body. Similarly, these latter were fed into different RNNs and their outputs fusion form the global body representation. More recently, the spatio-temporal LSTM (ST-LSTM) (77) extended LSTM to spatio-temporal domains. To this end, the analysis of a 3D skeleton joint considers spatial information from neighboring joints and temporal information from previous frames. In addition, a tree-structure based method allows to better describe the adjacency properties among the joints. This method is further improved by a gating mechanism to handle noise and occlusion.

3.6.1.2 2D facial expression recognition

The task of facial expression recognition (FER) consists in recognizing the basic emotions, *e.g.*, fear, surprise, happiness, etc. Recall that facial landmarks are located in certain regions of the face such as the mouth, the eyes, eyebrows, etc. As stated in (78), the motion of these landmarks defining facial expressions allows to characterize the emotional state of humans. Therefore, representing and classifying sequences of facial landmarks has been widely used in the literature of FER.

Falling in the category of graph-based approaches, a geometric approach was proposed in (79) which introduced a unified probabilistic framework based on an interval temporal Bayesian network (ITBN) built from the movements of landmark points. Aware of the small variations along a facial expression, the authors in (80) proposed a method to capture the subtle motions within facial expressions using

a variant of Conditional Random Fields (CRFs) called Latent-Dynamic CRFs (LDCRFs) on geometric features.

Taking the direction of Riemannian methods, Taheri *et al.* (65) proposed to represent 2D facial sequences as parameterized trajectories on the Grassmann manifold of 2-dimensional subspaces in $\mathbb{R}^n$ (n is the number of landmarks) which is an affine-invariant shape representation. To capture the facial deformations, they used geodesic velocities between facial shapes and finally, classification was performed by applying LDA then SVM. In the work of Kacem *et al.* (66), 2D facial landmark sequences were first represented as trajectories of Gram matrices in the manifold of positive semidefinite matrices of rank 2. A similarity measure is then provided by temporally aligning trajectories while taking into account the geometry of the manifold. This measure is finally used to train a pairwise proximity function SVM.

More recent approaches exploited deep neural networks. In (81), two neural network architectures were proposed for image videos (DTAN) and 2D facial landmark sequences (DTGN) which are combined (forming DTAGN) to predict final emotions. In particular, DTGN showed to be efficient by using only 2D landmark sequences, when applied separately. Another approach is proposed in (82) where a Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN) is responsible for analyzing the temporal information of facial landmark sequences and a Multi-Signal Convolutional Neural Network (MSCNN) is designed to extract spatial features from still frames. These two networks are combined to boost the performance of facial expression recognition.

Although the macro facial expression recognition problem has seen considerable advances, micro-expression recognition is still a relatively challenging task (83). Micro-expressions are brief facial movements characterized by short duration, involuntariness and subtle intensity. In the literature, previous methods opted for extracting hand-crafted features from texture videos such as LBP-TOP and HOOF (84).

More recently, deep learning methods were proposed to tackle the problem by applying CNNs (85, 86) and RNNs (86). To our knowledge, only the method of (87) is entirely based on analyzing 2D facial landmark sequences. Their work is based on computing the point-wise distances between adjacent landmark configurations along a sequence which is stacked in a matrix. The latter was seen as an input image to a CNN-LSTM-based classifier. However, their approach was only evaluated on a synthesized dataset produced from a macro-expression dataset. In our work, we will show that we achieve state-of-the-art results on a commonly-used micro-expression dataset using only 2D landmark data.

3.6.2 Temporal modeling and classification

Let $\{\bar{Z}_1, \bar{Z}_2, ..., \bar{Z}_L\}$ be a sequence of skeletons representing a trajectory on $\mathcal{S}$. As described previously, we code each skeleton $\bar{Z}_i$ into a sparse vector $w_i \in \mathbb{R}^N$ with respect to a dictionary $\mathcal{D}$. As a consequence, each trajectory is mapped to an N-dimensional function of sparse codes and the problem of classifying trajectories on $\mathcal{S}$ is turned to classifying N-dimensional sparse codes functions in Euclidean space, where

standard tools designed for Euclidean time-series (*e.g.*, temporal modeling, machine learning techniques, etc.) could be directly applied. We adopt and evaluate two temporal modeling and classification schemes to recognize actions and facial expressions.

3.6.2.1 Dynamic time warping, Fourier pyramid and SVM

A non-trivial challenge in recognizing actions and facial expressions resides in their rate variability since they can be executed at different speed. Thus, a typical landmark sequence representation has to be invariant to the execution rate. In other words, sequences belonging to the same class should typically have similar parametrization. A commonly used technique, namely Dynamic Time Warping (DTW) can temporally align one sequence into another by finding the optimal re-parameterization that minimizes a similarity measure between the two. In our work, we exploit the vector structure of our latent space to directly apply DTW on the sequential data, avoiding the more difficult task of temporal alignment of trajectories on the manifold. Another important post-processing is to further filter out noise present in the data. For instance, Fourier Temporal Pyramid (FTP) (88) have shown to be very effective for recognition tasks involving noisy data as it maps a time-series to the Fourier domain and eliminates the high frequency elements. Moreover, FTP is known to be robust to temporal mis-alignment. In our work, we apply a pipeline of DTW and FTP, then classify final features using a one-vs-all linear SVM. By doing so, we handle rate variability, temporal misalignment and noise, and classify final features, respectively.

It is important to note that to be able to apply the FTP approach, the pose sequences should have the same temporal length. For that, we need to apply a temporal re-sampling to all sequences which can be achieved in two alternatives:

- After projecting the input sequences to the Kendall's shape space, we can apply a Riemannian re-sampling algorithm as described in Algorithm 5. This alternative is better applied when the dataset contains long sequences. Thus, one can perform down-sampling of trajectories which will reduce the computational cost of the sparse coding step.
- As explained in Section 3.5.4, one can exploit the vector structure of the latent space to apply algorithms designed for Euclidean time-series which are in most cases faster than their nonlinear equivalent. Thus, one can apply SCDL on the input trajectories, then perform a Euclidean temporal re-sampling on the obtained sequences of sparse codes.

Temporal re-sampling of trajectories Temporal up-sampling allows to increase time-steps in a sequence for more accuracy, while temporal down-sampling decreases them to reduce the computation complexity of algorithms as an example. Besides, one may need to have a set of sequences with the same length by applying up or down sampling to each sequence. This can be achieved in the shape manifold by applying the algorithm described in Algorithm. 5.

Algorithm 5 Re-sampling of trajectories on $\mathcal{S}$

InputinputOutputoutput Given a trajectory $\alpha(t)_{t=t_1,...,t_n}$, we seek $\alpha(s)_{s=s_1,...,s_l}$ where $l < n$ for down-sampling and $l > n$ for up-sampling.

for $i = 1$ to l **do** Find t_{i_1}, t_{i_2} such that $t_{i_1} \leq s_i \leq t_{i_2}$
Compute $w_1 = \frac{s_i - t_{i_1}}{t_{i_2} - t_{i_1}}$ and $w_2 = \frac{t_{i_2} - s_i}{t_{i_2} - t_{i_1}}$
$x = \alpha(i_1), y = \alpha(i_2), \theta = d_{\mathcal{S}}(x, y)$ then
$\alpha(s_i) = \frac{1}{sin(\theta)}(sin(w_2\theta)x + sin(w_1\theta)y)$ **return** Re-sampled trajectory

3.6.2.2 Long short-term memory network

Modeling sequential data using recurrent neural networks (RNNs) has been widely used in different computer vision tasks and has led to breakthrough results in natural language processing (89), speech recognition (90), etc. RNNs are a kind of neural networks that take sequential input and infer sequential output by sharing parameters between time steps. They are trained using back-propagation over time. However, standard RNNs lack the ability of learning long-term dependencies as they suffer from the problem of vanishing gradient (91). Tackling this problem, Long short-term memory (LSTM) network (92) is equipped with a gating mechanism that learns which information is relevant to keep or forget during training. Thereby they better handle the problem of learning long-term dependencies in sequential data. In the context of action recognition, many works in the literature opted for LSTMs to model and classify actions (77, 93, 94). In (93), the authors propose a part-aware LSTM where they divide a skeleton configuration into body parts, hence instead of keeping a long-term memory of the entire body's motion in the cell, they split it to part-based cells. Another work (94) tackles the problem of view variations in the captured actions, similarly to our work but instead of seeking a view-invariant representation then use it for classification, they proposed a view-adaptive LSTM-based network that automatically regulates observation viewpoints during the occurrence of an action. In our work, we aim to take advantage of the view-invariant nature of our representation as well as its compactness (sparsity) and vector space structure to apply an LSTM directly on the SCDL time-series. Further, we explore the use of Bi-directional LSTM (Bi-LSTM). Bi-LSTM is an extension of LSTM that presents each sequence backwards and forwards to two separate recurrent networks, providing context both from the future and past, respectively (95). We will experimentally show that Bi-LSTM can achieve slight improvements over the traditional LSTM in recognizing human motion.

3.6.2.3 Dictionary structure

In the context of classification, one may exploit the important information of data labels to construct more discriminative feature vectors. To this end, we propose to build *class-specific* dictionaries, similarly to (96). Formally, let S be a set of labeled sequences on $\mathcal{S}$ belonging to q different classes $\{c_1, c_2, ..., c_q\}$, we aim to build q class-specific dictionaries $\{D_1, D_2, ..., D_q\}$ in $\mathcal{S}$ such that each D_j is learned using skeletons belonging to training sequences from the corresponding class c_j. In this scenario, coding a

query shape $\bar{Z} \in \mathcal{S}$ is done with respect to each $D_{j,1\leq j\leq q}$, independently. As a result, q vectors of codes are obtained. These vectors are then concatenated to form a global feature vector W. As will be discussed in Section 3.7.3, this yields more discriminative feature vectors for classification.

3.7 Experimental evaluation

We perform extensive experiments to evaluate the effectiveness of the proposed frameworks in the tasks of: 3D action recognition, 2D macro- and micro- facial expression recognition. We provide comparisons to some state-of-the-art approaches on several publicly available datasets in addition to comparisons between the intrinsic and extrinsic paradigms of the proposed SCDL approach. Moreover, we perform baseline experiments to evaluate some properties of our recognition frameworks.

3.7.1 3D action recognition

3.7.1.1 Datasets

We evaluate the proposed skeletal representation using four benchmark datasets presenting different challenges: Florence3D-Action (97), UTKinect-Action (98), MSR-Action 3D (71), and the large-scale NTU-RGBD dataset (99).

Florence3D-Action dataset consists of 9 actions performed by 10 subjects. Each subject performed every action two or three times for a total of 215 action sequences. The 3D locations of 15 joints collected using the Kinect sensor are provided. The challenges of this dataset consist of the similarity between some actions and also the high intra-class variations as same action can be performed using left or right hand.

UTKinect-Action dataset consists of 10 actions performed twice by 10 different subjects for a total of 199 action sequences. The 3D locations of 20 different joints captured with a stationary Kinect sensor are provided. The main challenge of this dataset is the variations in the view point.

MSR-Action 3D dataset consists of 20 actions performed by 10 different subjects. Each subject performed every action two or three times for a total of 557 sequences. The 3D locations of 20 different joints captured with a depth sensor similar to Kinect are provided with the dataset. This is a challenging dataset because of the high similarity between many actions (*e.g.*, *hammer* and *hand catch*).

NTU-RGB+D is one of the largest 3D human action recognition datasets. It consists of $56,000$ action clips of 60 classes. 40 participants have been asked to perform these actions in a constrained lab environment, with three camera views recorded simultaneously. Each Kinect sensor estimates and records 25 joints coordinates reported in the 3D camera's coordinate system.

3.7.1.2 Experimental settings

For the first three datasets, we followed the cross-subject test setting of (100), in which half of the subjects was used for training and the remaining half was used for testing. Reported results were averaged over ten different combinations of training and test data. For Florence3D-Action and UTKinect-Action datasets, we followed an additional setting for each: Leave-one-actor-out (LOAO) (97, 101) and Leave-one-sequence-out (LOSO) (98), respectively. For MSR-Action3D dataset, we also followed (71) and divided the dataset into three subsets AS1, AS2, and AS3, each consisting of 8 actions, and performed recognition on each subset separately, following the cross-subject test setting of (100). The subsets AS1 and AS2 were intended to group actions with similar movements, while AS3 was intended to group complex actions together. In all experiments, we performed recognition based on the two classification schemes: Bi-LSTM and DTW-FTP-SVM.

For NTU-RGB+D, the authors of this dataset recommended two experimental settings that we follow: 1) Cross-subject (X-Sub) benchmark with 39,889 clips from 20 subjects for training and 16,390 from the remaining subjects for testing; 2) Cross-view (X-View) benchmark with 37,462 and 18,817 clips for training and testing. Training clips in this setting come from the camera views 2 and 3 while the testing clips are all from the camera view 1. Due to the huge amount of data in this dataset, we construct dictionaries using the kernel clustering approach presented in Section 3.4.1.2 since it is less time consuming than the dictionary learning optimization problem. Note that NTU-RGBD dataset contains two types of actions: daily activities where performed by one actor and interactions between two actors. For the latter, we perform sparse coding of each actor's skeleton separately. Further, since the closeness between the two actors is a relevant information, we compute the Euclidean distance between their center of mass and concatenate it to the feature vector obtained after sparse coding. Moreover, we compute displacement vectors, as described in Section 3.7.2.2 for micro-expressions, and fuse them with the rest of the features. Finally, we perform temporal modeling and classification using Bi-LSTM. For this dataset, we do not suggest using the first classification pipeline as it would be highly consuming in terms of computation of DTW and FTP due to the huge amount of data that NTU-RGBD contains.

3.7.1.3 Results and discussions

Comparison to Riemannian methods on MSR-Action, Florence3D and UTKinect datasets
The first row of methods in Table 3.1 reports the recognition results of different Riemannian approaches. Since in (32) human actions are also represented as trajectories in the Kendall's shape space, we report additional results of (32) on Florence3D and UTKinect datasets to give more insights about the strength of our coding approach compared to the method of (32). In Table 3.1, it can be seen that we obtain better results than all Riemannian approaches on the three datasets. We recall that one common drawback of these methods is to map trajectories on manifolds to a reference tangent space, where they compute distances between different points (other than the tangent point). This may introduce distortions, especially

TABLE 3.1: Overall recognition accuracy (%) on MSR-Action 3D, Florence Action 3D, and UTKinect 3D datasets. In the first column: $^{(R)}$: Riemannian approaches; $^{(N)}$: other recent approaches; Last row: our approach.

Dataset	**MSR-Action 3D**		**Florence 3D**		**UTKinect 3D**	
Protocol	**H-H**	**3 Subsets**	**H-H**	**LOAO**	**H-H**	**LOSO**
$^{(R)}$ T-SRVF on Lie group (42)	85.16	–	89.67	–	94.87	–
$^{(R)}$ T-SRVF on $\mathcal{S}$ (32)	89.9	–	–	–	–	–
$^{(R)}$ Lie Group (37)	89.48	92.46	90.8	–	97.08	–
$^{(R)}$ Rolling rotations (63)	–	–	91.4	–	–	–
$^{(R)}$ Gram matrix (102)	–	–	–	88.85	–	**98.49**
$^{(N)}$ Graph-based (73)	–	–	–	91.63	97.44	–
$^{(N)}$ ST-LSTM (77)	–	–	–	–	95.0	97.0
$^{(N)}$ JLd+RNN (103)	–	–	–	–	95.96	–
$^{(N)}$ SCK+DCK (74)	**91.45**	93.96	**95.23**	–	**98.2**	–
$^{(N)}$ Transition-Forest (72)	–	**94.57**	–	94.16	–	–
$^{(R)}$ **Ours** (SVM)	**90.01**	**94.19**	92.85	92.27	**97.39**	97.50
$^{(R)}$ **Ours** (Bi-LSTM)	86.18	86.18	**93.04**	**94.48**	96.89	**98.49**

when points are not close to the reference point. However, our method avoids such a non-trivial problem as coding of each shape is performed on its attached tangent space and the only measures that we compute are with respect to the tangent point. Now, we discuss our results obtained with the first classification scheme, *i.e.*, DTW-FTP-SVM, similarly used in (37, 42, 63). In the three datasets, it is clearly seen that our approach outperforms existing approaches when using the same classification pipeline, which shows the effectiveness of our skeletal representation. For instance, we highlight an improvement of 1.73% on MSR-Action 3D (following protocol (71)) and 1.45% on Florence3D-Action. Now, we discuss the results we obtained using Bi-LSTM. Note that although we do not perform any preprocessing on the sequences of codes before applying Bi-LSTM, our approach still outperforms existing approaches on Florence3D, with 1.64% higher accuracy. However, it performs less well on UTKinect yielding an average accuracy of 96.89% against 97.08% obtained in (37). In MSR-Action 3D, our approach performs better than the method of (42) using the first protocol. Note that in (42), results were averaged over all 242 possible combinations. However, our average accuracy is lower than other approaches following both protocols on this dataset (around 3.5% in the first and 0.62% in the second). Here, it is important to mention that data provided in MSR-Action 3D are noisy (104). As a consequence, using Bi-LSTM without any additional processing step to handle the noise (*e.g.*, FTP) could not achieve state-of-the-art results on this dataset.

Comparison to State-of-the-art We discuss our results with respect to recent non-Riemannian approaches. In all datasets, our approach achieved competitive results.

Florence3D-Action – On this dataset, our method outperforms other methods using Bi-LSTM in the case of LOAO protocol, as shown in Table 3.1. However, using the second protocol, it is 2.19% lower than

(74). The authors of (74) combine two kernel representations: sequence compatibility kernel (SCK) and dynamics compatibility kernel (DCK) which separately achieved 92.98% and 92.77%, respectively. The proposed approach achieves good performance for most of the actions. However, the main confusions concern very similar actions, *e.g.*, *Drink from a bottle* and *answer phone*, as demonstrated by the confusion matrix in Figure 3.16.

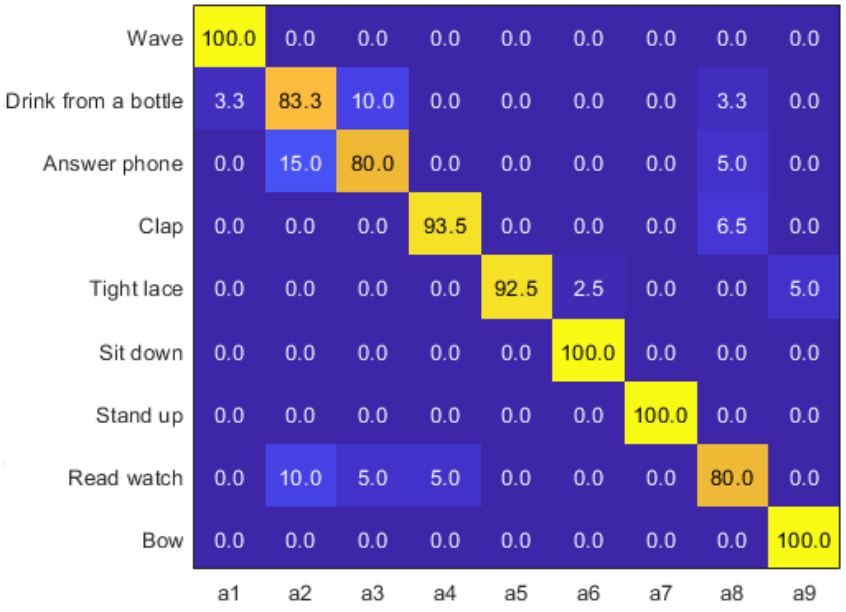


FIGURE 3.16: Confusion matrix on the Florence 3D dataset.

UTKinect – Following the LOSO setting, our approach achieves the best recognition rate, yielding an improvement of 2.49% compared to the method of (77), which is based on an extended version of LSTM. For the second protocol, our best result is competitive to the accuracy of 98.2% obtained in (74). Considering the main challenge of this dataset, *i.e.*, variations in the view point, our approach confirms the importance of the invariance properties gained by adopting the Kendall's representation of shape, hence the relevance of the resulting functions of codes generated using the geometry of the manifold.

MSR-Action 3D – For the experimental setting of (71), our best result is competitive to recent approaches. In particular, on AS3, we report the highest accuracy of 100%. This result shows the efficiency of our approach in recognizing complex actions, as AS3 was intended to group complex actions together. On AS1, we achieved one of the highest accuracies (95.87%). However, our result on AS2 is about 8.9% lower than state-of-the-art best result. This shows that our approach performs less well when recognizing similar actions, as AS2 was intended to group similar actions together. Although our best result is slightly higher than (74), it is lower than the same method when following the experimental setting of (88). This shows that our approach performs better in recognition problems with less classes.

TABLE 3.2: Overall recognition accuracy (%) on NTU-RGB+D following the X-sub and X-view protocols. In the first column: $^{(R)}$: Riemannian approaches; $^{(RN)}$: RNN-based approaches; $^{(CN)}$: CNN-based approaches.

Protocol	X-sub	X-view
$^{(R)}$ Lie Group (105)	50.1	52.8
HB-RNN-L (106)	59.1	64.0
$^{(R)}$ Deep learing on $SO(3)^n$ (75)	61.3	66.9
$^{(RN)}$Deep LSTM (93)	60.7	67.3
$^{(RN)}$Part aware-LSTM (93)	62.9	70.3
$^{(RN)}$ST-LSTM+Trust Gate (77)	69.2	77.7
$^{(RN)}$View Adaptive LSTM (94)	79.4	87.6
$^{(CN)}$Temporal Conv (107)	74.3	83.1
$^{(CN)}$C-CNN+MTLN (108)	79.6	84.8
$^{(CN)}$ST-GCN (109)	**81.5**	**88.3**
$^{(R)}$ **Intrinsic SCDL**	**73.89**	**82.95**

NTU-RGB+D – We report the obtained results for this dataset in Table. 3.2. For both benchmarks, X-view and X-sub, our approach remarkably outperforms other Riemannian representations. For instance, it outperforms the Lie group representation by 23% and 30% on X-sub and X-view protocols. It also surpasses the deep learning on Lie groups method by 12% and 16%. This could demonstrate the ability of our approach to deal with large scale datasets compared to conventional Riemannian approaches. Besides, our method outperforms RNN-based models, HB-RNN-L, Deep LSTM, PA LSTM and ST-LSTM+TG, with the exception of (94). Knowing that we also used an RNN-based model (Bi-LSTM) for temporal modeling and classification, this shows the efficiency of our action modeling. In fact, sparse features obtained after SCDL in Kendall's shape space are remarkably more discriminative than the original data. In order to have a better insight into their corresponding data distributions, we used the t-distributed stochastic neighbor embedding (t-SNE)[1] to visualize original data and SCDL features. From Fig. 3.17, we can observe that the SCDL features are better clustered than the original data in terms of class labels (colors in the figure). Besides, it is worth noting that SCDL is an efficient denoising tool, which is an important advantage when dealing with the often-noisy skeletons extracted with the Kinect sensor.

3.7.1.4 Comparison to extrinsic SCDL

To further evaluate the strength of the proposed intrinsic approach in the context of 3D action recognition, we compare it to extrinsic SCDL. Recall that instead of coding on tangent spaces, the extrinsic approach tends to embed the manifold-valued data into Hilbert spaces which are higher dimensional vector spaces where linear coding becomes possible. The main difficulty here arises from the fact that this embedding relies on a kernel function which, according to Mercer's theorem, should be positive definite. For 2D

[1]t-SNE is a nonlinear dimensionality reduction technique that allows for embedding high-dimensional data into two or three dimensional space, which can then be visualized in a scatter plot.

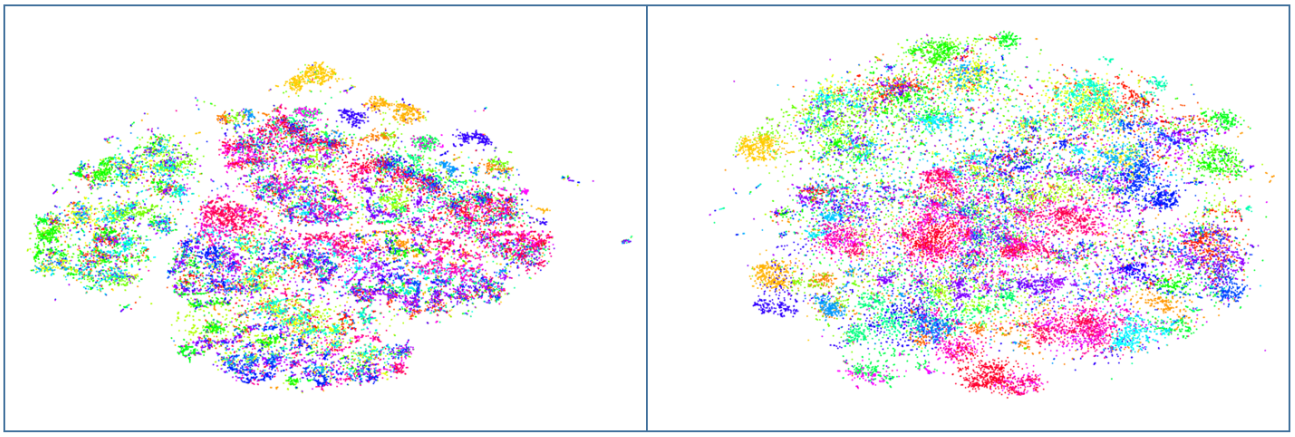


FIGURE 3.17: Visualization of 2-dimensional features of the NTU-RGB+D dataset. Left: original data. Right: the corresponding SCDL features. Each class is represented by a different color.

shapes, we presented in Section 3.2.3.2 a positive definite kernel. However, to the best of our knowledge, the existence of such a kernel has not been proved in the literature for 3D shapes. As a remedy, we adapted the extrinsic SCDL formulation by applying the Procrustes Gaussian kernel defined in Section 3.2.3.2, in which we also adapted the full Procrustes distance to 3D shapes as $d_{FP}(\bar{Z}_1, \bar{Z}_2) = sin(\theta)$ (see Section 4.2.1 of (110)) (θ is the geodesic distance defined in Section 3.2.3). Note that the kernel function relies on a parameter σ. Experimentally, we checked the positive definiteness of the adapted kernel and found out that it is only positive definite for some values of σ. We empirically chose 0.1 for Florence3D, 0.2 for UTKinect, and 0.5 for MSR-Action 3D, as to have valid positive definite kernels. Results reported in Table 3.3 show superiority of the intrinsic method. We argue that this difference comes from the fact that for the 3D case, the positive definiteness constraint on the kernel function reduced the valid space of the kernel parameter σ. Hence, intrinsic SCDL is in this case a better coding solution. In contrast, for the 2D case where we possess a PD kernel, we will show in the next section that extrinsic SCDL is more efficient in 2D recognition tasks.

TABLE 3.3: Comparative evaluation of intrinsic and extrinsic SCDL in recognizing 3D actions.

Dataset	**MSR-Action 3D**		**Florence 3D**		**UTKinect 3D**	
Protocol	H-H	3 Subsets	H-H	LOAO	H-H	LOSO
Extrinsic SCDL	82.52	88.53	85.76	89.03	93.97	94.97
Intrinsic SCDL	**90.01**	**94.19**	**92.85**	**92.27**	**97.39**	**97.50**

3.7.2 2D Facial Expression Recognition

In this application, we extract 49 facial landmarks from human faces in 2D and with high accuracy using a state-of-the-art facial landmark detector (29). We first represent the sequences of landmarks as trajectories in the Kendall's shape space. Extrinsic SCDL is then applied to produce sparse time-series that are finally classified in vector space. We evaluate this approach on two different 2D facial expression recognition tasks, the macro and micro.

3.7.2.1 Macro-Expression Recognition

The task here is to recognize the basic macro emotions, *e.g.*, fear, surprise, happiness, etc. To this end, we applied our approach on two commonly-used datasets namely the Cohn-Kanade Extended dataset and the Oulu-CASIA dataset. Our obtained results are then discussed with respect to state-of-the-art approaches as well as to intrinsic SCDL. For both datasets, we followed the commonly-used experimental setting in (81, 111–113) consisting on a ten-fold cross validation.

Cohn-Kanade Extended (CK+) dataset (114) consists of 327 image sequences performed by 118 subjects with seven emotion labels: *anger*, *contempt*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. Each sequence contains the two first temporal phases of the expression, *i.e.*, neutral and onset (with apex frames).

Oulu-CASIA dataset (115) includes 480 image sequences performed by 80 subjects. They are labeled with one of the six basic emotions (those in CK+, except the *contempt*). Each sequence begins with a neutral facial expression and ends with the expression apex.

Results and discussions Table 3.4 gives an overview of the obtained results on both datasets. Overall, our approach achieved competitive results compared to the literature. For instance, our best result on CK+ (obtained with Bi-LSTM) is by 1.52% lower than the best state-of-the-art result obtained by the method of (81). The latter is based on two neural network architectures trained on image videos and facial landmark sequences. However, when using only the landmark architecture (DTGN), our approach obtained a higher accuracy. Similarly, on Oulu-CASIA, our best result is lower than DTAGN and higher than DTGN. On the other hand, the method of (66) achieved a better performance on both datasets compared to our method. Comparing the confusion matrices (see Table 3.5), the same method seems to better recognize the *sadness* expression while our method is clearly more efficient in recognizing the *contempt* expression. This will be further discussed later on. From Fig. 3.18 and the confusion matrices in Tables 3.6 and 3.5, we can observe that the two expressions: *happiness* and *surprise* are well recognized in the two datasets while the main confusions happened in the two expressions: *fear* and *sadness*, conforming to state-of-the-art results (66, 81). Besides, we highlight the superiority of extrinsic SCDL compared to intrinsic SCDL. The first is performed in RKHS which is a higher dimensional vector space. This helps capturing complex patterns in facial expressions and identifying subtle differences between similar expressions. For instance, an interesting observation could be seen for the *contempt* expression. As stated in (114), the latter is quite subtle and it gets easily confused with other, strong emotions. For this expression, the recognition accuracy obtained with intrinsic SCDL is 55%, compared to 90% obtained with extrinsic SCDL, as shown in Figure 3.18. We argue that this remarkable improvement comes from the mapping to RKHS for the same reasons mentioned above. This observation has pushed us to further evaluate the performance of our approach in the task of micro-expression recognition.

TABLE 3.4: Comparison with state-of-the-art on CK+ and Oulu-CASIA datasets. $^{(A)}$: Appearance-based approaches; $^{(G)}$: Geometric approaches; $^{(R)}$: Riemannian approaches; Last row: our approach.

Method	CK+	Oulu-CASIA
$^{(A)}$ CSPL (113)	89.89	–
$^{(A)}$ ST-RBM (111)	95.66	–
$^{(A)}$ STM-ExpLet (112)	94.19	74.59
$^{(G)}$ ITBN (79)	86.30	–
$^{(G)}$ DTGN (81)	92.35	74.17
$^{(A+G)}$ DTAGN (81)	**97.25**	81.46
$^{(R)}$Shape velocity on Grassmannian (65)	82.80	–
$^{(R)}$Shape traj. on Grassmannian (66)	94.25	**80.0**
$^{(R)}$Gram matrix trajectories (66)	**96.87**	**83.13**
$^{(R)}$**Intrinsic SCDL** (SVM)	91.26	70.37
$^{(R)}$**Intrinsic SCDL** (Bi-LSTM)	89.43	70.24
$^{(R)}$**Extrinsic SCDL** (SVM)	95.62	**77.06**
$^{(R)}$**Extrinsic SCDL** (Bi-LSTM)	**95.73**	73.09

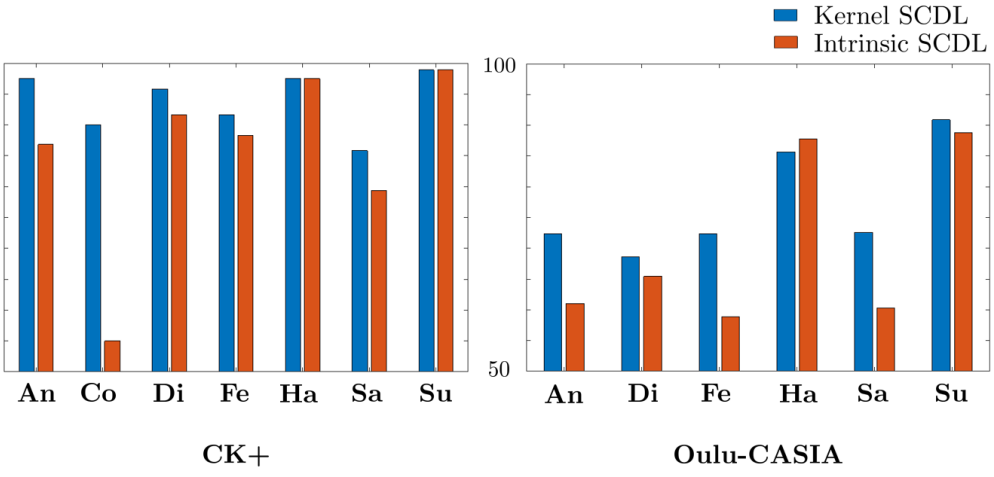


FIGURE 3.18: Recognition accuracy achieved for each emotion class in the CK+ (left) and the CASIA (rigth) datasets, and comparison between extrinsic and intrinsic SCDL approaches.

TABLE 3.5: Confusion matrix on the CK+ dataset.

	An	Co	Di	Fe	Ha	Sa	Su
An	**97.50**	0	2.5	0	0	0	0
Co	10	**90.0**	0	0	0	0	0
Di	2.5	0	**95.83**	0	0	1.67	0
Fe	0	0	0	**91.67**	8.33	0	0
Ha	0	0	0	0	**97.5**	2.5	0
Sa	5.0	0	5.0	2.5	1.67	**85.83**	0
Su	0	1.11	0	0	0	0	**98.89**

3.7.2.2 Micro-Expression Recognition

Micro expressions are brief facial movements characterized by short duration, involuntariness and subtle intensity. We argue that to recognize them, in contrast to macro-expressions, we are more interested in detecting subtle shape changes along a sequence. To this end, we applied the extrinsic SCDL framework as

TABLE 3.6: Confusion matrix on the Oulu-Casia dataset.

	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**
An	**72.33**	12.33	2.11	1.0	12.22	0
Di	14.22	**68.56**	6.22	3.0	8.0	0
Fe	5.22	2.0	**72.33**	5.11	9.33	6.0
Ha	4.0	0	9.33	**85.67**	1.0	0
Sa	15.22	4.11	6.11	2.0	**72.56**	0
Su	0	2.1	5.0	0	2.0	**90.89**

in macro-expression recognition, and to further detect the subtle deformations, we computed displacement vectors as the difference between successive sparse codes of L-dimensional time-series. Then, the resulting sequences of length $L-1$ are finally used for classification. We evaluate our approach on the most commonly-used dataset, namely CASME II.

CASME II dataset (116) contains 246 spontaneous micro-expression video clips recorded from 26 subjects and regrouped into five classes: happiness, surprise, disgust, repression and others. We performed classification based on the commonly used Leave-one-subject-out protocol.

Recall that previous methods that tackled the problem of micro-expression recognition are appearance-based (*i.e.*, using texture images) and to our knowledge, only (87) has studied the problem using 2D facial landmarks. However, their approach was only evaluated on a synthesized dataset produced from CK+ (macro) videos, by selecting the three first frames of an expression, then interpolating between them. For this reason, we compare our results with respect to appearance-based methods, as shown in Table 3.7.

TABLE 3.7: Recognition accuracy on CASME II dataset and comparison with state-of-the-art methods. In the first column: $^{(A)}$: Appearance-based approaches. $^{(R)}$: Riemannian approaches. Last row: our approach.

Method	**Accuracy (%)**
$^{(A)}$ STCLQP(117)	58.39
$^{(A)}$ CNN (85)	59.47
$^{(A)}$ CNN (LSTM) (86)	60.98
$^{(A)}$ LBP-TOP, HOOF (84)	63.25
$^{(A)}$ Optical Strain (118)	63.41
$^{(A)}$ DiSTLBP-IIP (117)	**64.78**
$^{(R)}$**Intrinsic SCDL** (SVM)	43.65
$^{(R)}$**Extrinsic SCDL** (SVM)	**64.62**

We point out the recognition accuracy of 64.62% achieved by our method outperforming state-of-the-art approaches, with the exception of (117). This shows the effectiveness of the adopted extrinsic SCDL in detecting subtle deformations from 2D landmarks, without any apperance-based information as other approaches in the literature.

Compared to the intrinsic approach, it is clear from Table 3.7 that the extrinsic SCDL method is better in recognizing micro-expressions. Recall that the use of extrinsic SCDL to tackle the problem of micro-expression recognition was driven by its good performance in recognizing the contempt emotion, in the CK+ dataset which is characterized by subtle changes along the expression. The obtained results on CASME II hence supports our previous claims.

3.7.3 Ablation study

We examine the effectiveness of the proposed Kendall SCDL schemes by performing several baseline experiments on different datasets.

A. Kendall's shape representation – We evaluate the necessity of the Kendall's shape projection. To this end, we perform temporal modeling and classification on raw data, after a scale and translation normalization, against their application on Kendall SCDL features. On NTU-RGB+D, we applied Bi-LSTM while on Florence 3D, MSR Action 3D and UTKinect, we applied the pipeline DTW-FTP-SVM. Performances are reported in the second and fourth rows of Table 3.8. In all datasets, improvements are remarkably gained with the Kendall's space projection. This is clearly seen in particular on the large scale NTU-RBD+D dataset which presents different view-variations and where the improvement is more than 26%.

TABLE 3.8: Evaluation of the Kendall's shape space representation.

Approach	NTU-RGB+D	Florence	MSR 3D	UTKinect
Raw data	56.5	84.29	87.36	92.67
Linear SCDL	79.20	87.94	89.23	93.58
Kendall SCDL	**82.95**	**92.85**	**90.01**	**97.5**

B. Nonlinear SCDL – In this experiment, we evaluate the importance of the nonlinear formulation of SCDL that we applied on Kendall's space. For that, we compare it to the use of linear SCDL, i.e., by solving for Eq.3.2. Obtained results on the four action recognition datasets, reported in the third row of Table 3.8, clearly show the interest of accounting for the nonlinearity of the manifold when applying SCDL.

C. Sparsity regularization – In this experiment, we evaluate the effect of the sparsity regularization parameter λ (in Eq. (3.9) and Eq. (3.12)) on recognition accuracies obtained using both of the adopted classifiers. To do so, we used half of a training set for learning the dictionary and training the classifiers and the other half for validation. The first graph of Fig. 3.19 shows the impact of increasing λ from 10^{-4} to 1 at steps of 10^{-2}. Further, we report the average sparsity percentage (*i.e.*, number of non-zero codes divided by the total number of codes) for some values of λ to show the coherence of the obtained codes with the proposed theory. As expected, the sparsity percentage increases when increasing λ. We remark that the accuracy reached a maximum value at $\lambda = 0.01$ (37% of sparsity) and $\lambda = 0.02$ (49%

of sparsity) for SVM and Bi-LSTM, respectively. Note that in all previous experiments, λ was chosen empirically so to correspond to these latter percentages of sparsity.

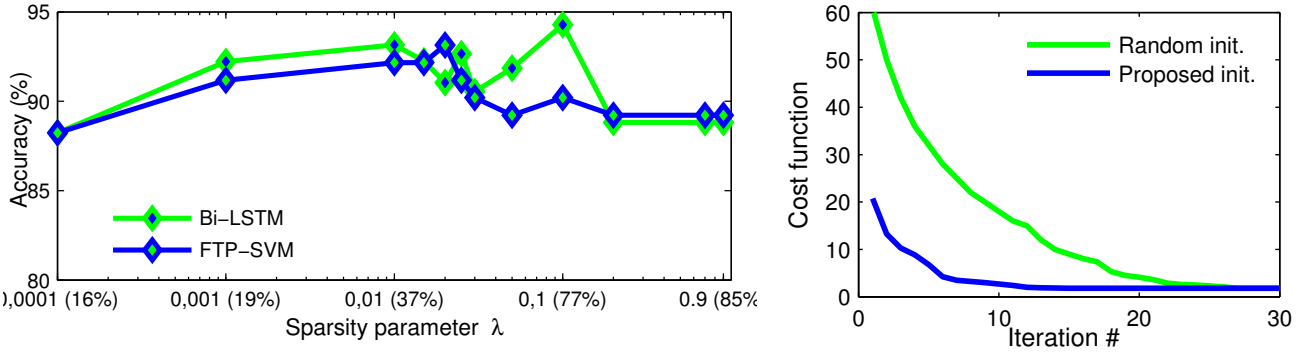


FIGURE 3.19: Left: Accuracy when varying the sparsity regularization parameter λ (% values in the x-axis represent the average sparsity). Right: Dictionary learning objective over iterations for: (1) Random initialization; (2) Our proposed initialization based on Bayesian clustering and PGA.

.

D. Dictionary structure – As described in Section 3.6.2.3, we build class-specific dictionaries. To show the relevance of this structure in the context of classification, we compare it to the case of using a global dictionary, *e.g.*, when label(s are not taken into account. The obtained recognition accuracies using Bi-LSTM and following the LOAO setting are 94.48% and 91.53% for class-specific and global dictionary, respectively. These results clearly prove that the adopted structure is better in classifying actions.

E. Dictionary initialization – In this experiment, we evaluate the performance of our proposed initialization step based on Bayesian clustering of shapes and PGA. To this end, we compare it to the case of random initialization, where atoms are randomly selected from the training set. We train a class-specific dictionary (for class *tight lace* in Florence3D dataset) with the same training data in both cases. For the case of random initialization, we set the number of atoms N to 41 to be equal to that of our proposed initialization. Recall that in our approach, N is automatically inferred to avoid its empiric choice, especially as we build class-specific dictionaries. In Fig. 3.19, on the right graph, we plot the two corresponding dictionary learning objectives over iterations. As it is expected, the proposed initialization shows faster convergence, dividing the overall dictionary learning processing time by approximately two times, when taking into account the execution time of our initialization step.

F. Performance of Bi-LSTM – We compared average accuracies yielded by Bidirectional LSTM and a traditional LSTM. Following LOAO experimental setting, using Bi-LSTM shows an improvement of around 0.7% on Florence Action 3D and 1.2% on NTU-RGBD dataset, indicating the positive effect of learning both future and past contexts to recognize actions.

G. Evaluation on the facial landmark detector The task of facial expression recognition from landmark data relies essentially on the accuracy of the landmark detector. In this experiment, we

evaluate the performance of the landmark detector that we used in our experiments (*i.e.*, Chehra (29)) by comparing it to the newly-released Openface2.0 (30), which gives the option of extracting either 49 or 68 landmarks. In Table 3.9, we report the classification accuracy obtained by applying the pipeline DTW+FTP+SVM on raw landmark data (after a simple scale and translation normalization). Results obtained on CK+ and Oulu-CASIA datasets clearly show a better performance using landmarks extracted with the Chehra detector.

TABLE 3.9: Classification performances when using different landmark detectors.

Landmark detector	Oulu-CASIA dataset	CK+ dataset
Cherha (29) - **49 landmarks**	**76.41**	**93.68**
Openface (30) - **49 landmarks**	70.85	83.73
Openface (30) - 68 landmarks	71.26	82.92

3.8 Discussions

The Kendall's shape representation has proven the efficiency of adopting a view-invariant analysis of the given data. Because of the nonlinearity of the Kendall's manifold, intrinsic and extrinsic solutions of SCDL were comprehensively studied and compared. Regarding the extrinsic solution, the advantage of embedding data from the Kendall's shape space to RKHS is twofold. First, the latter is vector space, thus it enables the extension of linear SCDL to the nonlinear Kendall's space. Second, embedding a lower dimensional space in a higher dimensional one gives a richer representation of the data and helps extracting complex patterns. However, to define a valid RKHS, the kernel function must be positive definite according to Mercer's theorem. On one hand, for the 2D Kendall's space, we have used the Procrustes Gaussian Kernel which is positive definite and shown that for the task of 2D macro facial expression recognition, extrinsic SCDL performs better than intrinsic SCDL. We argue that this is due to the kernel embedding. For instance, we highlight the clear improvement in recognizing the *contempt* emotion in the CK+ dataset. The latter is characterized with subtle deformations that are well captured using the extrinsic approach. This has drove us to evaluate it on the task of 2D micro-expression recognition where the shape deformations along expressions are known to be subtle as well. As expected, the performance of extrinsic SCDL was promising. On the other hand, for the 3D Kendall's space, a positive definite kernel function has not been proposed in the literature. Nevertheless, adapting the PGk to 3D shapes prevented us from exploring the whole space of σ as in this case, this kernel is positive definite for only certain value of this parameter. As a consequence, the performance of extrinsic SCDL in the 3D Kendall's space can be hindered since the quality of the produced codes depends on the value of σ. We argue that this is the main reason behind the better performance obtained using intrinsic SCDL for the task of 3D action recognition. Besides, intrinsic sparse coding of a shape is performed on its attached tangent space, by mapping atoms into it. Compared to Riemannian approaches of the literature, this avoids the common drawback of mapping points to a common tangent space at a reference point which may introduce distortions.

FIGURE 3.20: The illustration of object cube size. In the first row, the green rectangular and red rectangular represent small and large cube respectively. The yellow rectangular in the second row represent the appropriate size of object cube.

3.9 Online human-object recognition

Based on the inter-joints and object-joints distances presented previously, we propose in this section an algorithm for on-line human-object recognition. The classification is based on one frame with N previous ones in the memory (N can be zero). When N is not null, the N-frames sliding window is considered for the on-line classification. The first step in the on-line recognition system we propose is the object feature. This object feature will be fused later with the low-level extracted features to built the final features vector which will be classified on-line using random forest classifier.

3.9.1 Object Feature

A specific object description can be helpful to characterize the human object interaction. But this is a difficult and time consuming way to realize online classification. As we discussed in the previous part, it is insufficient to only use the 3D joint positions to fully model an action, especially when the action includes the interactions between the subject and other objects such as *drinking* and *picking phone*. The extra input like depth information need be adopted in order to have more precise classification.

Motivated by properties of objects, we try to utilize the size and shape information of objects which is more efficient and convenient way for online human-object interaction recognition. When performing an interaction, human usually hold objects by two hands. Moreover, the depth points located around the skeleton joints of two hands contain a lot of messages about the size and shape of objects.

The object is assumed to be present around one hand, thus similarly to the LOP algorithm (119) that counts the number of points inside a given cube around given point (hand for example) and decides the

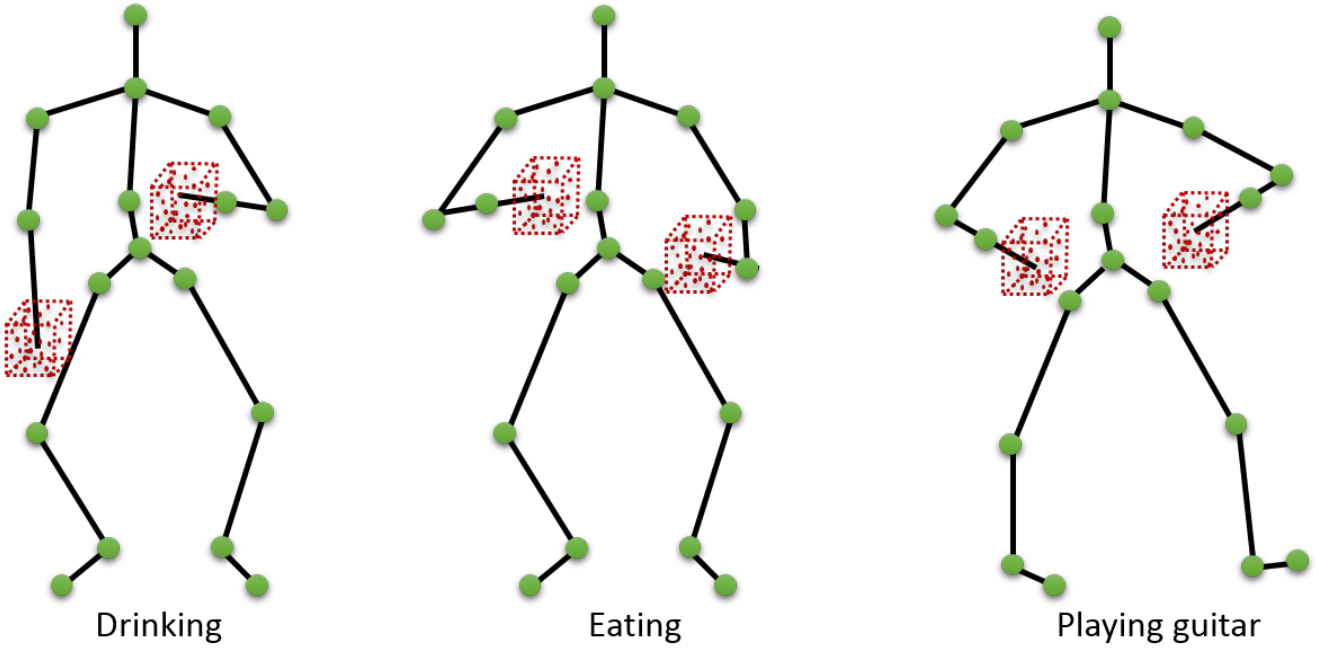


FIGURE 3.21: Examples of our object features on Online RGBD Action Dataset (ORGBD) dataset. The red cube refers to object cube for each action.

presence of an object given a threshold, we extend this algorithm to exploit the number of the points inside the cube and the 3D coordinates of these points to built the object feature. The number of depth points refers to the rough size of objects and the coordinates of these points refer to the rough shape of objects. The PCA algorithm is applied on the coordinates in order to determine the principal directions of the object inside the cube. These directions are concatenated with the number of the points to build the object feature. The feature vector calculation depends on the size of chosen cubes to detect these points. If the cube size is too small like the situation shown at the top left in Fig. 3.20, the green rectangular is too small to show the features of different object. So the resulting feature will not be discriminative for interaction classification. If the cube size is too big like the situation shown at the top right in Fig. 3.20, the red rectangular is so big that contains a lot of context from background and other parts of body. So we have to detect objects in a appropriate size as shown in the second row in Fig. 3.20. In the experiment, the retained size of cubes is 50. A trade of the size of this cube and the results will be discussed later in experimental section.

3.9.2 Online action recognition

The feature vector is the concatenation of the pairwise distances between the joints and the object feature that contains the number of points inside the cube around the hand holding the object and the main directions describing the rough shape of the object given by PCA algorithm. The proposed approach handles also the human action with no object interaction. In this case, the LOP algorithm detects the absence of objects around the hands and an imputation technique is used to fill the missing informations.

In our experiments we employed the mean imputation method, which consists of replacing the missing values by the means of values already calculated in presence of the object from the training set. For the classification task we used the multi-class version of Random Forest algorithm. The Random Forest algorithm was proposed by Leo Breiman in (120) and defined as a meta-learner comprised of many individual trees. It was designed to operate quickly over large datasets and more importantly to be diverse by using random samples to build each tree in the forest. Diversity is obtained by randomly choosing attributes at each node of the tree and then using the attribute that provides the highest level of learning. Once trained, Random Forest classify a new action from an input feature vector by putting it down each of the trees in the forest. Each tree gives a classification decision by voting for that class. Then, the forest chooses the classification having the most votes (over all the trees in the forest). In our experiments we used Weka multi-class implementation of Random Forest algorithm by considering 100 trees. A study of the effect of the number of the trees is reported later in the experimental part.

3.10 Conclusion

In this chapter, we investigated human behavior analysis using skeleton and landmark data. Thus, we proposed a novel action and facial expression modeling based on Riemannian sparse coding and dictionary learning in the Kendall shape space manifold. This solution allows to overcome the nonlinear structure of the manifold by mapping a Riemannian trajectory to an Euclidean time-series. In addition to its sparsity and vector structure, this representation allows to reconstruct original trajectory from its latent representation thanks to the dictionary. In addition, it presents a natural denoising tool allowing to alleviate the noise present in the data. We explored both intrinsic and extrinsic solutions of SCDL. The first extends sparse coding to tangent spaces while avoiding a commonly encountered problem which is to map all manifold-valued data to a common tangent space. The second is based on embedding the manifold-valued data to Hilbert space via a positive definite kernel function. These two approaches were evaluated in the context of two recognition tasks: 3D action recognition and 2D facial expression recognition. We have used two temporal modeling and classification schemes on top of the obtained sparse time-series: a deep learning framework based on Bi-LSTM and a pipeline of DTW-FTP-SVM. We proposed also an online approach using a simple spatial modeling to recognize human-object interaction. We have conducted extensive experiments on seven commonly-used datasets and showed that our obtained results are competitive to state-of-the-art. We have compared the two adopted temporal modeling and classification pipelines and discussed the obtained results for different datasets. Further, we presented a comprehensive comparative study on the use of intrinsic and extrinsic SCDL approaches by providing an answer to the question: “Depending on the nature of the data (*i.e* body or face) and its dimension (*i.e* 2D or 3D), when and which technique should we apply?”.

Chapter 4

3D Face Analysis for gender and expression recognition

The main results presented in this chapter have been published in the following international journal: IEEE Cybernatics (2014) (121), IEEE Affective Computing (2017) (122), Pattern Recognition (2015) (123).

4.1 Introduction

Due to the natural, non-intrusive, and high throughput nature of face data acquisition, automatic face recognition has many benefits when compared to other biometrics. Moreover, additional facial attributes in human faces have interested the computer vision community over the last decades. Actually, the soft-biometrics and the facial expressions recognition have been investigated within the the computer vision community with application in several different areas, such as HMI (Human-Machine Interaction), psychology, computer graphics and so on. Soft-biometrics are natural recognizable attribute in human faces. In our daily life, human beings are performing their estimation naturally and effectively, from the face. In sexual dimorphism studies [1] (124), researchers have found that male faces usually possess more prominent features than female faces. Male's face usually has a more protuberant nose, eyebrows, more prominent chin and jaws. The forehead is more backward sloping, and the distance between top-lip and nose-base is longer. Research presented in (125) have also demonstrated that females are smaller in all the concerned anthropometric measurements. The gender classification can help in solving more complicate problems such as age estimation. Age reflects the continuous accumulation of durable effects from the past since birth. Human faces deform with time non-inversely and thus contains their aging information. Among different modalities available for face imaging, 3D scanning has a major advantage over 2D color imaging in that nuisance variables, such as illumination and small pose changes, have a

[1] http://www.virtualffs.co.uk/

relatively smaller influence on the observations. However, the 3D static face is not really sufficient to recognize other facial attributes such as the facial expression as facial expressions are naturally expressed over time. We emphasize that it is more natural to analyze expressions as spatio-temporal deformations of 3D faces, caused by the actions of facial muscles. The importance of facial expressions was first realized and investigated by psychologists, among others. In a seminal work by Mehrabian et al. (126) the relative importance of verbal and nonverbal messages in communicating feelings and attitude is described. In particular, they provided evidence that face-to-face communication is governed by the 7%-38%-55% rule, that balances the relevance of verbal, vocal and visual elements, respectively, in communications. Despite this rigid quantification has since been refuted in later studies, it still provides an indication that the words and tone of the voice form only a part of human communication. The non-verbal elements related to the body language (e.g., gestures, posture, facial expressions) also play an important role. Starting from a different point of view, Ekman (127) conducted the first systematic studies on facial expressions in the late 70s. Ekman also showed that facial expressions can be coded through the movement of face points as described by a set of *action units* (128). Through his experiments, it is demonstrated that there are six *prototypical* facial expressions, representing *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*, plus the *neutral* one that are universally recognized and remain consistent across different ethnicities and cultures. The presence of these prototypical facial expressions is now widely accepted for scientific analysis. More recently, computer vision community has been interested on spontaneous facial expressions and dedicated database have been collected, such as BP4D database. In this chapter, we propose a novel 3D face-based shape analysis framework (called DSF (129)) for capturing the differences between two given 3D faces. This framework is used to classify gender based on static 3D faces and facial expressions based on dynamic 3D faces (4D faces).

4.2 Optimal Deformations (Dense Scalar Field)

In order to capture and model deformations of the face, we propose to represent the facial surface through a set of parameterized radial curves that originate from the tip of the nose. Approximating the facial surface by an ordered set of radial curves, which locally captures its shape can be seen as a parameterization of the facial surface. Indeed, similar parameterizations of the face have shown their effectiveness in facial biometrics (130). The mathematical setup for the shape theory offered here comes from Hilbert space analysis. A facial surface is represented by a collection of radial curves and a Riemannian framework is used to study shapes of these curves. We start by representing facial curves as absolutely continuous maps from $\beta : [0, 1] \rightarrow \mathbb{R}^3$ and our goal is to analyze shapes represented by these maps. The problem in studying shapes using these maps directly is that they change with re-parameterizations of curves. If γ is a re-parameterization function (typically a diffeomorphism from $[0, 1]$ to itself), then under the standard $\mathbb{L}^2$ norm, the quantity $\|\beta_1 - \beta_2\| \neq \|\beta_1 \circ \gamma - \beta_2 \circ \gamma\|$, which is problematic. The solution comes from choosing a Riemannian metric under which this inequality becomes equality and the ensuing analysis simplifies. As described in (131), we represent the facial curves using a new function q, called the

square-root velocity function (SRVF) (see Eq. (4.1)). With the proposed representation, a facial surface is approximated by an indexed collection of radial curves β_α, where the index α denotes the angle formed by the curve with respect to a *reference* radial curve. In particular, the reference radial curve (i.e., the curve with $\alpha = 0$) is chosen as oriented along the vertical axis, while the other radial curves are separated each other by a fixed angle and are ordered in a clockwise manner. As an example, Fig. 4.1(a) shows the radial curves extracted for a sample face with happy expression. To extract the radial curves, the nose tip is accurately detected and each face scan is rotated to the upright position so as to establish a direct correspondence between radial curves having the same index in different face scans. In Fig. 4.1(b)-(c), two radial curves at $\alpha = 90°$ in the neutral and happy scans of the same subject are shown. As emerged in the plot (d) of the same figure, facial expressions can induce consistent variations in the shape of corresponding curves. These variations change in strength from expression to expression and for different parts of the face. In order to effectively capture these variations a Dense Scalar Field is proposed, which relies on a Riemannian analysis of facial shapes.

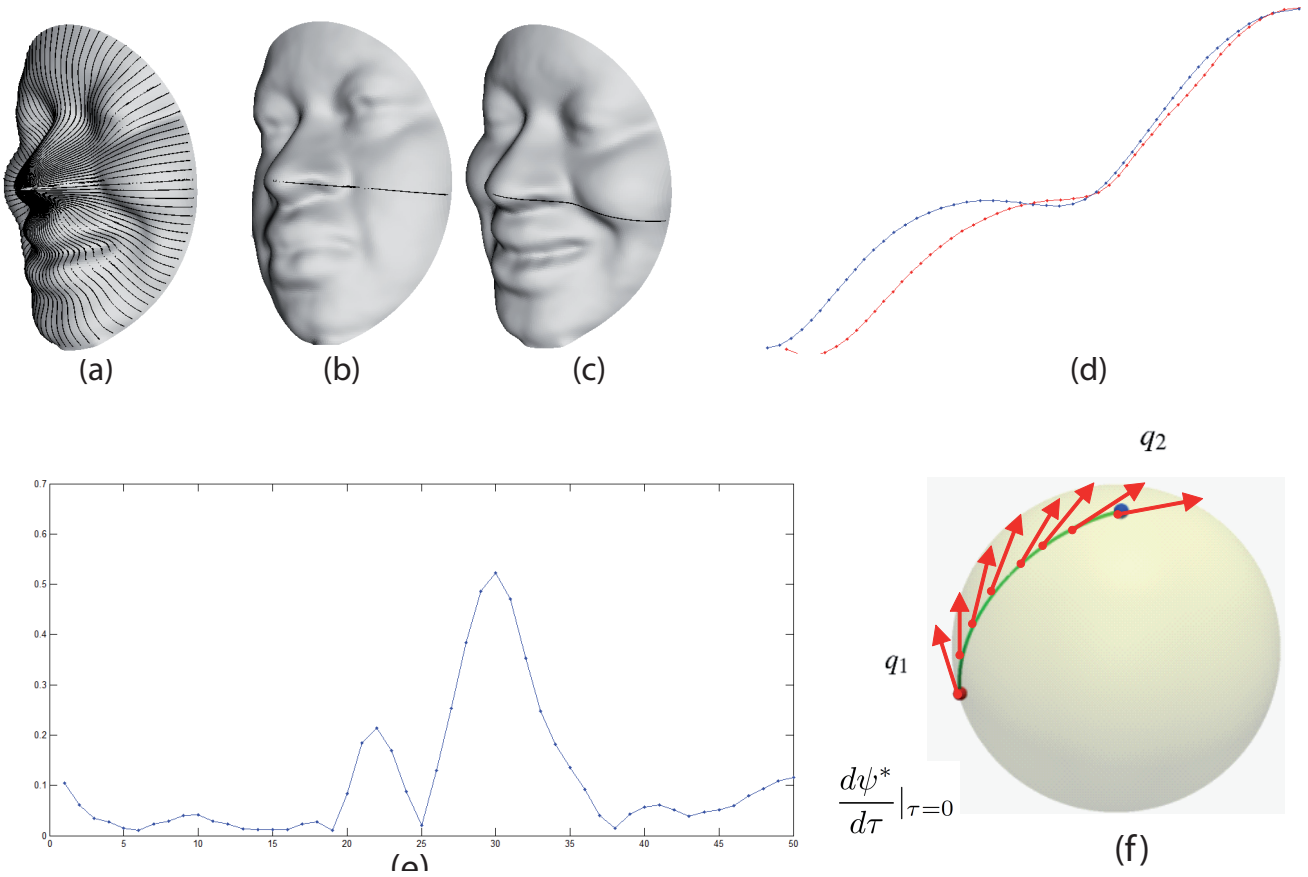


FIGURE 4.1: The figure illustrates: (a) The extracted radial curves; (b)-(c) A radial curve on a neutral face, and the correspondent curve on the same face with happy expression, respectively; (d) The two radial curves are plotted together; (e) The values of the magnitude of $\frac{d\psi^*}{d\tau}|_{\tau=0}(k)$ computed between the curves in (d) are reported for each point k of the curves; (f) The parallel vector field across the geodesic between q_1 and q_2 in the space of curves $\mathcal{C}$.

Considering a generic radial curve β of the face, it can be parameterized as β: $I \to \mathbb{R}^3$, with $I = [0, 1]$,

and mathematically represented through the *square-root velocity function* (SRVF) (131, 132), denoted by $q(t)$, according to:

$$q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}, \quad t \in [0,1]. \tag{4.1}$$

This specific representation has the advantage of capturing the shape of the curve and makes the calculus simpler. Let us define the space of the SRVFs as $\mathcal{C} = \{q : I \to \mathbb{R}^3, \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^3)$, with $\|\cdot\|$ indicating the $\mathbb{L}^2$ norm. With the $\mathbb{L}^2$ metric on its tangent space, $\mathcal{C}$ becomes a Riemannian manifold. Basically, with this parametrization each radial curve is represented on the manifold $\mathcal{C}$ by its SRVF. According to this, given the SRVFs q_1 and q_2 of two radial curves, the shortest path ψ^* on the manifold $\mathcal{C}$ between q_1 and q_2 (called *geodesic path*) is a critical point of the following energy function:

$$E(\psi) = \frac{1}{2} \int ||\dot{\psi}(\tau)||^2 d\tau, \tag{4.2}$$

where ψ denotes a path on the manifold $\mathcal{C}$ between q_1 and q_2, τ is the parameter for traveling along the path ψ, $\dot{\psi} \in T_\psi(\mathcal{C})$ is the tangent vector field on the curve $\psi \in \mathcal{C}$, and $||.||$ denotes the $\mathbb{L}^2$ norm on the tangent space.

Since elements of $\mathcal{C}$ have a unit $\mathbb{L}^2$ norm, $\mathcal{C}$ is a hypersphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^3)$. As a consequence, the geodesic path between any two points $q_1, q_2 \in \mathcal{C}$ is simply given by the minor arc of the great circle connecting them on this hypersphere, $\psi^* : [0,1] \to \mathcal{C}$. This is given by:

$$\psi^*(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta) q_1 + \sin(\theta\tau) q_2\right), \tag{4.3}$$

where $\theta = d_{\mathcal{C}}(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle)$. We point out that $sin(\theta) = 0$, if the distance between the two curves is zero, in other words $q_1 = q_2$. In this case, for each τ, $\psi^*(\tau) = q_1 = q_2$.

The tangent vector field on this geodesic is then written as $\frac{d\psi^*}{d\tau} : [0,1] \to T_\psi(\mathcal{C})$, and is obtained by the following equation:

$$\frac{d\psi^*}{d\tau} = \frac{-\theta}{\sin(\theta)} \left(\cos((1-\tau)\theta) q_1 - \cos(\theta\tau) q_2\right). \tag{4.4}$$

Knowing that on geodesic path, the covariant derivative of its tangent vector field is equal to 0, $\frac{d\psi^*}{d\tau}$ is parallel along the geodesic ψ^* and one can represent it with $\frac{d\psi^*}{d\tau}|_{\tau=0}$ without any loss of information. Accordingly, Eq. (4.4) becomes:

$$\frac{d\psi^*}{d\tau}|_{\tau=0} = \frac{\theta}{\sin(\theta)} \left(q_2 - \cos(\theta) q_1\right) \ (\theta \neq 0). \tag{4.5}$$

A graphical interpretation of this mathematical representation is given in Fig. 4.1. In Fig. 4.1(a), we show a sample face with happy expression and all the extracted radial curves. In Fig. 4.1(b) and Fig. 4.1(c)

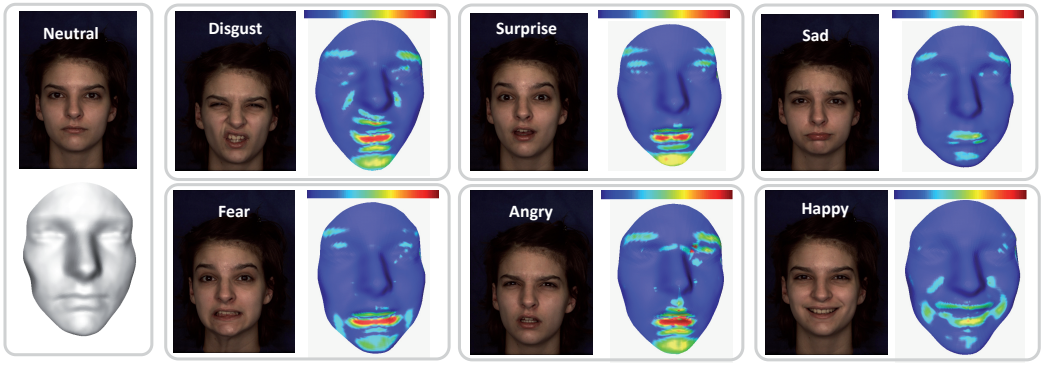


FIGURE 4.2: Deformation Scalar Fields computed between a neutral face of a given subject and the apex frames of the sequences of the six prototypical expressions of the same subject. The neutral scan is shown on the left. Corresponding texture images are also illustrated with each DSFs colormap.

two corresponding radial curves (i.e., radial curves at the same angle α), respectively, on a neutral and a happy face of the same person are highlighted. These curves are reported together in Fig. 4.1(d), where the amount of deformation between them can be appreciated, although the two curves lie at the same angle α and belong to the same person. The amount of deformation between the two curves is calculated using Eq. (4.5), and the plot of the magnitude of this vector at each point of the curve is reported in Fig. 4.1(e) (i.e., 50 points are used to sample each of the two radial curves as reported on the x axis, while the magnitude of the vector field is reported on the y axis). Finally, Fig. 4.1(f) illustrates the idea to map the two radial curves on the hypersphere $\mathcal{C}$ in the Hilbert space through their SRVFs q_1 and q_2, and shows the geodesic path connecting these two points on the hypersphere. The tangent vectors of this geodesic path represent a vector field whose covariant derivative is zero. According to this, $\frac{d\psi^*}{d\tau}|_{\tau=0}$ becomes sufficient to represent this vector field, with the remaining vectors obtained by parallel transport of $\frac{d\psi^*}{d\tau}|_{\tau=0}$ along the geodesic ψ^*.

Based on the above representation, we define a *Dense Scalar Field* capable to capture deformations between two corresponding radial curves β^1_α and β^2_α of two faces approximated by a collection of radial curves.

Dense Scalar Field (DSF)

Let $x_\alpha(t) = ||\frac{d\psi^*_\alpha}{d\tau}|_{\tau=0}(t)||$ be the values of the magnitude computed for each point t of the curves q^1_α and q^2_α; let T be the number of sampled points per curve, and $|\Lambda|$ be the number of curves used per face. According to this, we define the function f by:

$$f : \mathcal{C} \times \mathcal{C} \longrightarrow (\mathbb{R}^+)^{\mathcal{T}},$$

$$f(q^1_\alpha, q^2_\alpha) = (x^1_\alpha, \ldots, x^k_\alpha, \ldots, x^T_\alpha).$$

Assuming that $\{\beta_\alpha^1 | \alpha \in \Lambda\}$ *and* $\{\beta_\alpha^2 | \alpha \in \Lambda\}$ *be the collections of radial curves associated with the two faces* F^1 *and* F^2 *and let* q_α^1 *and* q_α^2 *be their SRVFs, the Dense Scalar Fields (DSF) vector is defined by:*

$$DSF(F^1, F^2) = (f(q_0^1, q_0^2), \ldots, f(q_\alpha^1, q_\alpha^2), \ldots, f(q_{|\Lambda|}^1, q_{|\Lambda|}^2)).$$

The dimension of the DSF vector is $|\Lambda| \times T$.

The steps to compute the proposed DSF are summarized in Algorithm 7.

Algorithm 6 – Computation of the Dense Scalar Field

Require: Facial surfaces F^1 and F^2; T, number of sample points on a curve; $\Delta\alpha$, angle between successive radial curves; $|\Lambda|$, number of curves per face
Ensure: $DSF(F^1, F^2)$, the DSF between the two faces
procedure ComputeDSF($F^1, F^2, T, \Delta\alpha, |\Lambda|$)
$n \leftarrow 0$
while $n < |\Lambda|$ **do**
$\alpha = n \cdot \Delta\alpha$
for $i \leftarrow 1, 2$ **do**
extract the curve β_α^i
compute the SRVF of β_α^i:
$q_\alpha^i(t) \doteq \frac{\dot{\beta}_\alpha^i(t)}{\sqrt{\|\dot{\beta}_\alpha^i(t)\|}} \in \mathcal{C},\ t = 1, \ldots, T$
end for
compute the distance between q_α^1 and q_α^2:
$\theta = d_{\mathcal{C}}(q_\alpha^1, q_\alpha^2) = cos^{-1}(\langle q_\alpha^1, q_\alpha^2 \rangle)$
compute the deformation vector $\frac{d\psi^*}{d\tau}|_{\tau=0}$ using Eq. (4.5) as:
$f(q_\alpha^1, q_\alpha^2) = (x_\alpha(1), x_\alpha(2), \ldots, x_\alpha(T)) \in \mathbb{R}_+^T$
$x_\alpha(t) = |\frac{\theta}{\sin(\theta)} (q_\alpha^2 - \cos(\theta) q_\alpha^1)|,\ t = 1, \ldots, T$
end while
compute $DSF(F^1, F^2)$ as the magnitude of $\frac{d\psi^*}{d\tau}|_{\tau=0}(k)$:
$DSF(F^1, F^2) = (f(q_0^1, q_0^2), \ldots, f(q_{|\Lambda|}^1, q_{|\Lambda|}^2))$ **return** DSF
end procedure=0

The first step to capture the deformation between two given 3D faces F^1 and F^2 is to extract the radial curves originating from the nose tip. Let β_α^1 and β_α^2 denote the radial curves that make an angle α with a reference radial curve on faces F^1 and F^2, respectively. The initial tangent vector to ψ^*, called also the shooting direction, is computed using Eq. (4.5). Then, we consider the magnitude of this vector at each point t of the curve in order to construct the DSFs of the facial surface. In this way, the DSF quantifies the local deformation between points of radial curves β_α^1 and β_α^2, respectively, of the faces F^1 and F^2. In the practice, we represent each face with 100 radial curves, and T=50 sampled points on each curve, so that the DSFs between two 3D faces is expressed by a 5000-dimensional vector.

In Fig. 4.2 examples of the deformation fields computed between a neutral face of a given subject and the apex frames of the sequences of the six prototypical expressions of the same subject are shown. The values of the scalar field to be applied on the neutral face to convey the six different prototypical

expressions are reported using a color scale. In particular, colors from green to red represent the highest deformations, whereas the lower values of the dense scalar field are represented in cyan/blue. As it can be observed, for different expressions, the high deformations are located in different regions of the face. For example, as intuitively expected, the corners of the mouth and the cheeks are mainly deformed for happiness expression, whereas the eyebrows are also strongly deformed for the angry and disgust expressions.

4.3 Gender classification using 3D face

4.3.1 Introduction

Human gender perception is an extremely reliable and fast cognitive process since the face presents a clear sexual dimorphism (133). Humans are remarkably accurate at deciding whether faces of their peers are male or female, even when cues from hair style, makeup, and facial hair are minimized (134). In human face analysis using machines (135), automatic gender classification is an active research area. Developed solutions could be used in human computer interaction (intelligent user interface, video games, etc.), visual surveillance, collecting demographic statistics for marketing (audience or consumer proportion analysis, etc.), and security industry (access control, etc.) as a soft biometrics trait useful to develop efficient face recognition algorithms. Research on automatic gender classification using images goes back to the beginning of the 1990s. Since then, significant progress has been reported in the literature (136–140). Fundamentally, proposed techniques differ in (i) face images (2D or 3D) ; (ii) choice of facial representation, ranging from simple raw 2D pixels or 3D cloud of points to more complex features such as Haar-like, LBP and AAM in 2D and shape index, wavelets and facial curves, in 3D ; and (iii) design of classifiers, for instance Neural Networks, SVM and Boosting methods (136).

4.3.1.1 Related work on 3D-based gender classification

In (141), *Liu et al.* look into the relationship between facial asymmetry and gender. They impose a 2D grid on each 3D face mesh to represent the face with 3D grid points. With the selected symmetry plane which equally separates the face into right and left halves, the distance difference (Euclidean distances to the origin of the cylindrical coordinate system) between each point and its corresponding reflected point is calculated as height differences (HD), and the angle difference between their normal vectors is calculated as orientation differences (OD). Results on 111 full 3D neutral face models of 111 subjects show that statistically significant difference are observed between genders with the overall OD facial asymmetry measurement. This result confirms early claims in anthropomorphic studies, which claim that male faces generally possess a larger amount of asymmetry than female ones (142), (143). They also define a local symmetry measurement named Variance Ratio (VR). They achieve 91.16% and 96.22%

gender recognition rate in testing, respectively. These performances are reported on a private dataset of full 3D faces (instead of 2.5D scans commonly used) including only 111 neutral face scans.

The range and intensity modalities of the face provide different cues of demographic information. In (144), *Lu et al.* provide an integration scheme for range and intensity modalities to classify ethnicity (Asian and Non-Asian) and gender (Male and Female). The best gender classification result using 10-fold cross-validation reported is 91%. Note that for males the result was 95.6% and for females the result was 83%. That is probably due to there are unequal numbers of scans in the dataset between genders. Here also, only neutral scans are considered for the experiments and no study was conducted when varying the facial expressions.

Statistically there are differences in geometry facial features between different genders, such as in the hairline, forehead, eyebrows, eyes, cheeks, nose, mouth, chin, jaw, neck, skin and beard regions (145). In (146), *Han et al.* present a geometry feature based approach for 3D-face gender classification. The volume and area of the forehead, and their corresponding ratio to nose, eyebrows, cheeks and lips are defined to generate feature vectors. RBF-SVM is then applied to classify gender. They select 61 frontal 3D face meshes from the GavabDB database, and carry out 5 experiments, with each experiment containing 48 faces for training and 13 for testing. The average correctness reported is 82.56%. As mentioned, only 61 neutral scans are considered for the experiment which leaves serious questions on the statistical significance of the results. Once more, landmarks are not automatically detectable and the precision required for their positioning requires for manual annotation in both training and testing stages.

In (147), *Wu et al.* use 2.5D facial surface normals (needle-maps) recovered with Shape From Shading (SFS) from intensity images for gender classification. The recovered needle-maps presented in PGA (Principle Geodesic Analysis) parameters not only contain facial shape information, but also the illumination intensity implicitly. They select 260 2D frontal face images from the UND Database. Experiments are done 10 times with 200 faces randomly selected for training and the remaining 60 faces for testing. The best average gender recognition rate reported is 93.6% with both shape and texture considered. To construct the statistical model and apply the SFS method, seven keypoints on the face, manually located, are needed for scans normalization. Also, a small dataset of neutral scans is used to conduct the experiments. As the approaches described above, this method suffer from its dependence to the landmark detection accuracy.

In (148), *Hu et al.* propose a fusion-based gender classification method for 3D frontal faces. Each 3D face shape is separated into four face regions using face landmarks. Fusion is applied to the results of four face regions and the best result reported is 94.3%. Overall experiments are conducted on neutral faces, therefore no attention was given to the robustness to facial expressions. Recall that, deformations caused by the expressions is the most challenging problem in 3D face analysis and recognition.

Recently, in (135), *Toderici et al.* employ MDS (Multi-Dimensional Scaling) and wavelets on 3D face meshes for gender classification. They take 1121 scans of Asian subjects and 2554 scans of White subjects

in FRGCv2 for ethnicity and gender classification. Experiments are carried out subject-independently with no common subject used in the testing stage of 10-fold cross validation. With polynomial kernel SVM, they achieve about 93% correct gender classification rate with an unsupervised MDS approach, and about 94% correctness with the wavelets-based approach. Both approaches significantly outperform the kNN and kernel-kNN approaches. In their experiment, the authors consider only *Asian* and *White* ethnicity classes and leave out 332 scans of 48 subjects of FRGCv2 dataset. Thus, their classifier is trained and tested using only these classes and don't consider more complex ethnicity variations.

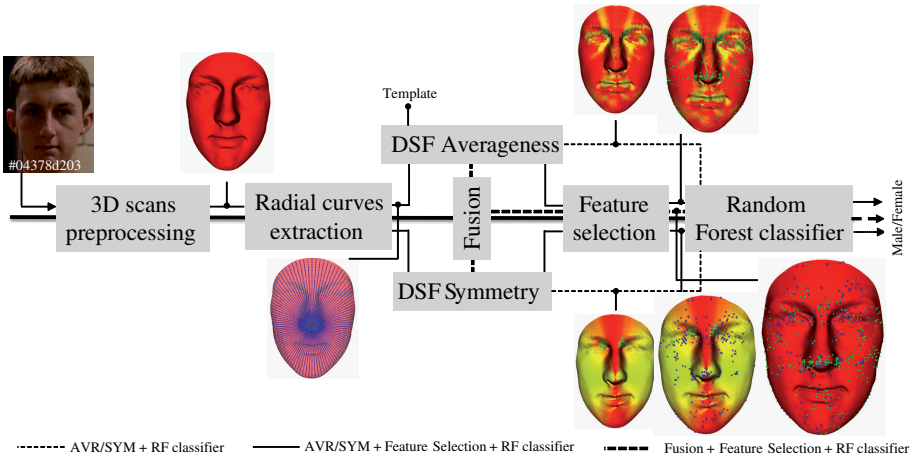


FIGURE 4.3: Outline of the proposed gender classification approach.

More recently, in (149), *Ballihi et al.* extract geometrical features (26 level curves and 40 radial curves) from 3D faces for gender classification. To form a high performance classifier with a minimal set of features, the Adaboost algorithm is used to select salient geometrical facial features. With the salient curves trained by 20 previous 3D faces in the FRGC-1.0 dataset, they obtain a correctness rate of 84.12% with the nearest neighbor algorithm when using the 466 earliest scans of the FRGCv2 dataset as the testing set. Note that, here, training and testing are done separately on different datasets. They also perform a standard 10-fold cross-validation for the 466 earliest scans of FRGCv2, and obtain 86.05% with Adaboost. The approach is performed on the earliest 3D scans of FRGCv2 which consist mainly on neutral faces.

4.3.1.2 Methodology and contributions

From the analysis above, it emerges that a large part of existing works on 2D- and 3D-based gender classification are based on local or global *low-level* descriptors extraction (see table 4.2 for a complete summary) followed by popular classification methods. First of all, we decided to conduct our study using 3D images of the face because of its richness with shape information. Definitely, this allows to capture anatomic differences between male and female, more easily than using texture information. For example, the female brow tends to be more arched than that of the male (which is more horizontal). Other obvious differences appear in the nose and chin which are, usually, more prominent in male compared to female

(150). When analyzing their profiles, men have a more acute nasolabial angle than comparative female (151), and other anatomic differences in other specif area of the face as the forehead, the cheeks and the lips. To the best of our knowledge, no work has been done considering *high-level* cues as face averageness and its bilateral asymmetry except the study in (141) which investigates the relationship between face symmetry and gender. Using sparse measures of height differences (HD), and orientation differences (OD) on a defined grid imposed on full 3D face models, the process requires manual landmarking of seven keypoints on the face model. The main contributions of this work are,

- First, we introduce together face averageness (AVR) and bilateral symmetry (SYM) and provides mathematical tools to densely quantify them on a given 3D shape of face. These primary facial perception cues are rarely considered in the literature of face-based attributes recognition.
- Secondly, we use the proposed framework (Dense Scalar Field DSF) for capturing the averageness/symmetry differences on the face surface. The DSFs grounding on Riemannian shape analysis are capable to densely capture the shape differences in 3D faces (such as averageness/symmetry differences) through face representation of radial curves.
- Thirdly, we investigate the relationship between facial averageness and facial symmetry, through fusion and feature selection methods, to see whether they are complimentary or not in gender classification.
- Last but not the least, our approach for gender classification is fully-automatic, we achieve competitive results compared to the approaches in state-of-the-art on a challenging dataset, FRGCv2, and demonstrate a strong robustness against age, ethnicity and expression variations.

An overview of the proposed approach is depicted in Fig. 4.3. Firstly, in preprocessing, hole-filling, cropping and smoothing are applied to each scan together with nose tip and middle plane detection. We denote the preprocessed face as $\boldsymbol{S}$. The plane which equally separates the preprocessed face $\boldsymbol{S}$ into right and left halves is picked up as the middle plane. This plane $P(t, \overrightarrow{n_h})$ passes through the detected nose tip t and has a horizontal normal $\overrightarrow{n_h}$. Secondly, a DSF extraction step goes after preprocessing. The preprocessed face $\boldsymbol{S}$ is approximated by collection of radial curves emanating from the nose tip after pose normalization. Then, a Dense Scalar Field (DSF) is computed, pair-wisely, to capture the shape differences (averagenesss/symmetry differences) between corresponding radial curves on each indexed point. Thus, we obtain two DSFs for each scan, an averageness DSF and a symmetry DSF. A fusion is then obtained for each scan by concatenating its averageness DSF and symmetry DSF. Thirdly, after DSF extraction, we branch our work into two pipelines. In one pipeline, averageness DSFs, symmetry DSFs and fusion DSFs, together with the gender labels, are directly fed into Random Forest to learn and classify gender. For the second pipeline, we first apply supervised feature selection (FS) algorithm on averageness, symmetry and their fusion DSFs with gender labels and then explore them with Random Forest (RF) to build an automatic classifier.

4.3.2 Feature Extraction Methodology

As mentioned earlier, the second main step of our approach, after scans preprocessing, is to extract densely averageness and symmetry features. Both of them are based on shape analysis of 3D faces using the Dense Scalar Fields that we describe in the section below.

4.3.2.1 Face symmetry description

The idea of face symmetry descriptor is to capture the bilateral symmetry difference in face by DSF. symmetry difference is defined as the deformation from a face point to its corresponding symmetrical point on the other side of face. In practice, symmetry DSF is calculated on each indexed point of the corresponding symmetrical curves in the preprocessed face $\boldsymbol{S}$. Let β_α denote the radial curve that makes an angle α with the middle plane $P_{\boldsymbol{S}}(t, \overrightarrow{n_h})$ from the frontal view of $\boldsymbol{S}$, and $\beta_{2\pi-\alpha}$ denotes the corresponding symmetrical curve that makes an angle $2\pi - \alpha$ with $P_{\boldsymbol{S}}(t, \overrightarrow{n_h})$. The tangent vector field $\dot{\psi_\alpha}^*$ that represents the energy needed to deform β_α to $\beta_{2\pi-\alpha}$ is then calculated. With the magnitude of $\dot{\psi_\alpha}^*$ at each point, located in curve β_α with index k, we build a *symmetry Dense Scalar Field* (symmetry DSF) on the facial surface, $V_\alpha^k = ||\dot{\psi}_\alpha^*|_{(\tau=0)}(k)||$.

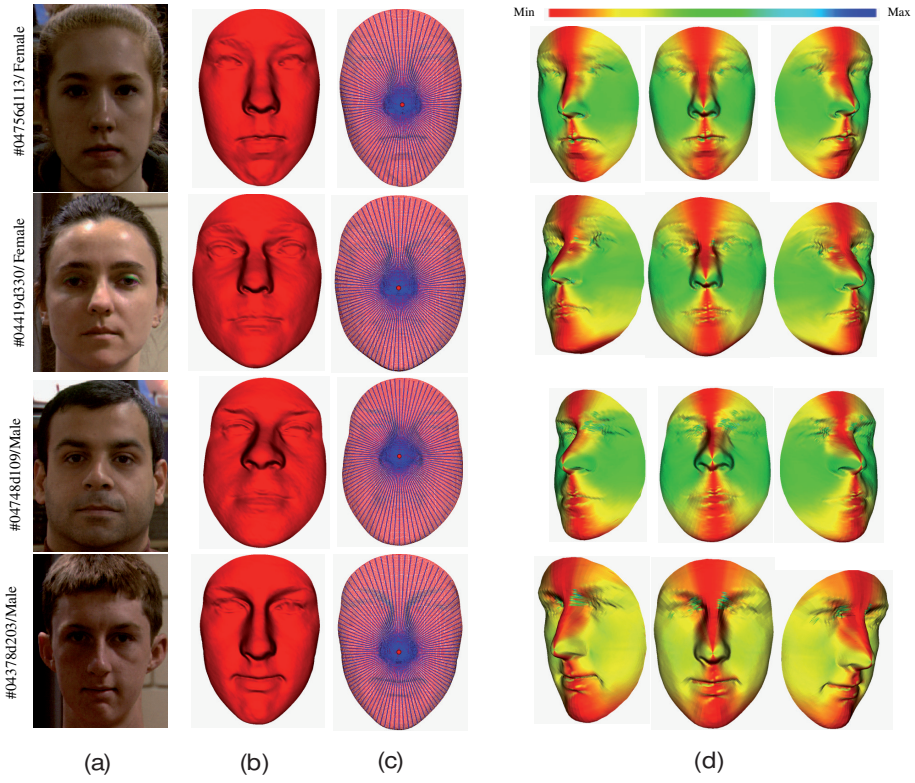


FIGURE 4.4: Illustrations of symmetry DSF on face. (a) 2D intensity image; (b) preprocessed 3D face surface $\boldsymbol{S}$; (c) 3D face $\boldsymbol{S}$ with extracted curves; (d) color-map of symmetry DSF mapped on $\boldsymbol{S}$ with three poses.

This Dense Scalar Field quantifies the shape difference between corresponding symmetrical curves on each point of the preprocessed face $\boldsymbol{S}$. Some examples illustrating this symmetry descriptor are shown in Fig. 4.4. For each subject, face in column (a) shows the 2D intensity image; column (b) illustrates the preprocessed 3D face surface $\boldsymbol{S}$; column (c) illustrates the the 3D face $\boldsymbol{S}$ with extracted curves; column (d) shows the symmetry degree as a color-map of the DSF mapped on $\boldsymbol{S}$. The color bar is shown in the up-right corner. The warm color means the minimum deformation and cold color signifies the maximum deformation. The hotter the color, the lower magnitude of the bilateral asymmetry. In this work, the symmetry DSFs are generated with 200 radial curves extracted from each face and 100 indexed points on each curve. Thus the total volume of each DSF is 20000. Compared with the work of *Liu* in (141), where totally less than 50 VR values were computed on sub-regions of HD or OD face, our experiment with DSF descriptor exceeds significantly in density. The average time consumed for extracting all 200 curves for each face is around 1 seconds, and for generating the bilateral symmetry descriptor (DSF) on all the 200×100 points of each face is 0.058 seconds. The average preprocessing time consumed for each scan is 0.116 seconds. Thus the total computation time (including preprocessing) for each scan is less than 1.2 seconds.

4.3.2.2 Face averageness description

Here, our aim is to capture the differences in face morphology between male and female by comparing their shapes to a defined average face. We claim that such differences change with the face gender. In fact, masculine faces have more prominent features (nose, eyebrows, forehead, mouth, etc.) in comparison to feminine faces. Thanks to DSFs, presented in section 4.2, we are able to capture densely such shape differences as long as a template face is defined. In this section, we answer to the following questions : (a) How do we define the template face to measure morphology differences, and (a) How do we compute the averageness DSFs by analogy with that of symmetry?

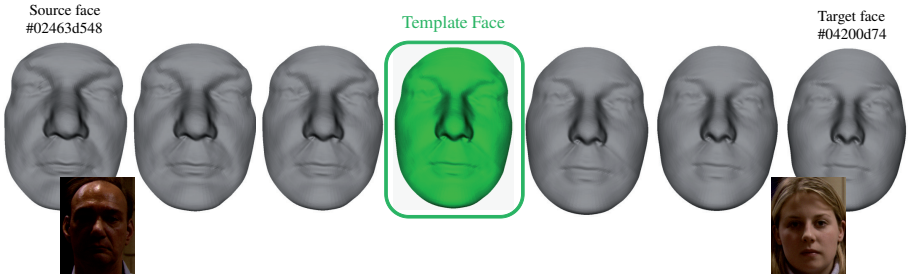


FIGURE 4.5: averageness face template

As shown in Fig. 4.5, the face template is defined as the middle point of a geodesic path which joint a male face *(ID: 02463d548; Age: 48; White)* to a female face *(ID: 04200d74; Age: 21; White)*. We first consider these faces as representative elements of the female and male classes, then, represent them as collections of radial curves. Finally, we compute the geodesic path between pair-wisely curves using eq.

4.3. By interpolation, one can generate the middle point of the geodesic between the faces. Thus, we obtain the averageness face template $\boldsymbol{T}$.

For a preprocessed face $\boldsymbol{S}$, let $\beta_\alpha^{\boldsymbol{S}}$ denote the radial curve that makes an angle α with the middle plane $P_{\boldsymbol{S}}(t, \overrightarrow{n_h})$ from the frontal view of $\boldsymbol{S}$, and β_α^T denotes the curve that makes the same angle α with $P_T(t, \overrightarrow{n_h})$ in the averageness face template $\boldsymbol{T}$. The tangent vector field $\dot{\psi_\alpha}^*$ that represents the energy needed to deform β_α^S to β_α^T is then calculated for each index α. Similar to the symmetry descriptor, with the magnitude of $\dot{\psi_\alpha}^*$ at each point, located in curve β_α^S with index k, we build an *averageness Dense Scalar Field* (averageness DSF) on the facial surface, $V_\alpha^k = ||\dot{\psi}_\alpha^*|_{(\tau=0)}(k)||$. This Dense Scalar Field quantifies the shape difference between corresponding curves of $\boldsymbol{S}$ and $\boldsymbol{T}$ on each indexed point.

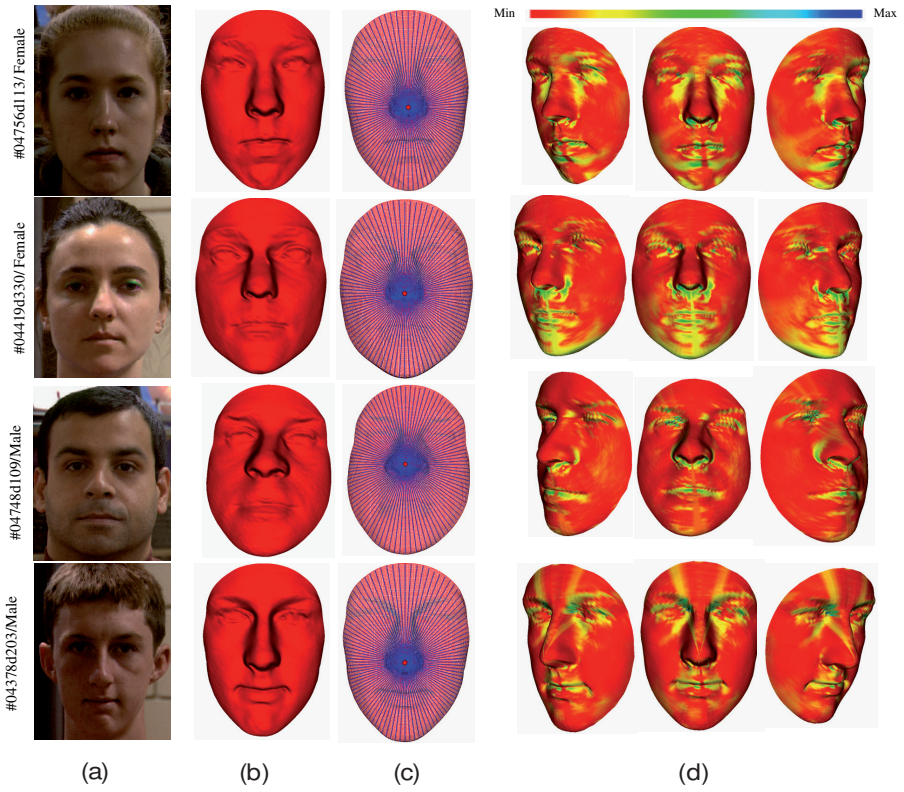


FIGURE 4.6: Illustrations of Average DSF on face. (a) 2D intensity image; (b) preprocessed 3D face surface $\boldsymbol{S}$; (c) the 3D face $\boldsymbol{S}$ with extracted curves; (d) color-map of the Average DSF mapped on $\boldsymbol{S}$ with three poses.

Fig. 4.6 exemplifies this averageness descriptor. For each subject, the face in column (a) shows the 2D intensity image; column (b) illustrates the preprocessed 3D face surface $\boldsymbol{S}$; column (c) shows the 3D face $\boldsymbol{S}$ with extracted curves; column (d) shows color-map of the Average DSF mapped on $\boldsymbol{S}$ with three poses.

4.3.3 Gender classification

Face averageness and symmetry are different structural properties in face perception. Each of these properties has a signaling social role. In this work, we first study individually their relationship with

gender, then we combine them to find out if it enhances gender classification result, which means that they contribute to gender classification in different ways. In practice, we use an *early fusion method* which consists on concatenating the *averageness DSF* and *symmetry DSF* of each scan, to achieve the *fusion DSF*. Then, we explore the performance of the fusion-based classifier, in different scenarios. For that purpose, we use methods from Feature Selection literature and Random Forest classification algorithm, which are detailed below.

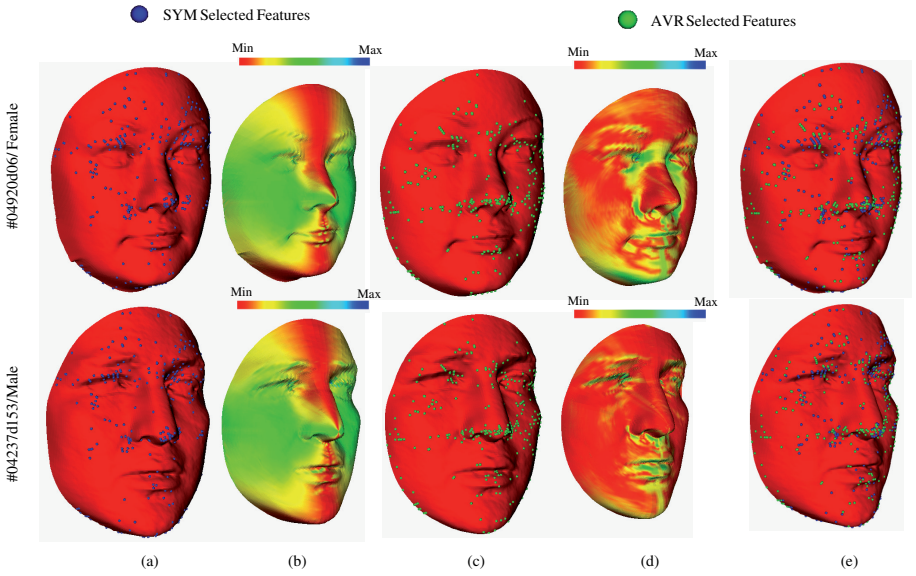


FIGURE 4.7: Feature selection. (a) selected points of symmetry DSF in face; (b) color-map of original symmetry DSF; (c) selected points of averageness DSF in face; (d) color-map of original averageness DSF; (e) selected points of both averageness DSF and symmetry DSF in face.

4.3.3.1 Feature Selection

Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible (152). It is a central problem in machine learning. The earliest approaches for feature selection were ***the filter*** methods. These algorithms use heuristics based on general characteristics of the data to evaluate the merit of feature subsets. Another school of approaches argues that the bias of a particular induction algorithm should be taken into account when selecting features. This method, called ***the wrapper*** (153), uses an induction algorithm along with a statistical re-sampling technique such as cross-validation to estimate the final accuracy of feature subsets. The filter methods operate independently of any learning algorithm-undesirable features are filtered out of the data before learning begins. They are generally much faster than wrapper methods, especially on data of high dimensionality. Since the averageness, symmetry and fusion DSFs are really dense and possibly redundant after DSF extraction, we design a feature selection procedure on the DSFs to get rid of the irrelevant and redundant features. For the merits of filter methods, we chose a filter, named Correlation-based-Feature-Selection (CFS) (152). It is an algorithm that couples the evaluation formula based on an

appropriate correlation measure and a heuristic search strategy. The central hypothesis of CFS is that good feature sets should contain features that are highly correlated with the class, yet uncorrelated with each other. The feature evaluation formula (Pearson's correlation coefficient), based on ideas from test theory, provides an operational definition of this hypothesis. Within CFS, we try two heuristic search strategies, the Best-First search strategy and the Greedy-Step-Wise search strategy. The Best-First search strategy (154) is an AI search strategy that allows back-tracking along the search path. It moves through the search space by greedy hill-climbing augmented with a back-tracking facility. When the path being explored becomes non-improving, the Best-First search will back-track to a more promising previous subset and continue the search from there. The stopping criterion is the number of consecutive non-improving nodes (5 in our experiments) that result in no improvement. For Greedy-Step-Wise, it performs a greedy forward or backward search through the space of attribute subsets. It stops when the addition/deletion of any remaining attributes results in a decrease in evaluation.

After Feature selection, we retain 301 salient points for averageness DSF, 271 salient points for symmetry DSF, and 365 salient points for the fusion. The feature selection procedure significantly reduces the volume and complexity of original DSF description. Fig. 4.7 shows the selected features of averageness DSF and symmetry DSF in faces. Column (a) maps the selected features of symmetry DSF in face; Column (b) shows the color-map of original symmetry DSF on the face ; Column (c) maps the selected points of averageness DSF in face ; Column (d) shows the original averageness DSF on the face; Column (e) maps the selected points of both averageness DSF and symmetry DSF in face. For both averageness DSF and symmetry DSF, we observe dense distribution of salient points around the nose and eyes regions. More salient points exist in forehead regions in averageness DSF, and more salient points exist in cheek regions in symmetry DSF. These observations hint that averageness DSF and symmetry DSF share both commonness and differences. In other words, they are complimentary in face description.

4.3.3.2 Random Forest

Face-based gender classification is a binary classification problem which estimates the gender c of a given test face into Male or Female $c \in \{Male, Female\}$. We carry out gender classification experiments with the well-known machine learning algorithm, Random Forest. Random Forest is an ensemble learning method that grows many classification trees $t \in \{t_1, .., t_T\}$. To classify a new face from an input vector (DSF-based features vectors $V = V_\alpha^k$), each tree gives a classification result and the forest chooses the classification having the most votes. In the growing of each tree (155), firstly, $\boldsymbol{N}$ instances are sampled randomly with replacement from the original data, to make the training set. Then, if each instance comprises of $\boldsymbol{M}$ input variables, a constant number $\boldsymbol{m}$ $\boldsymbol{(m<<M)}$ is specified. At each node of the tree, $\boldsymbol{m}$ variables are randomly selected out of the $\boldsymbol{M}$ and the best split on these $\boldsymbol{m}$ variables is used to split the node. The process goes on until the tree grows to the largest possible extent, without pruning.

The performance of the forest depends on the correlation between any two trees, and the strength of each individual tree. The forest error rate increases when the correlation decreases, or the strength increases. Reducing $\boldsymbol{m}$ reduces both the correlation and the strength. Increasing it increases both. Thus, an optimal $\boldsymbol{m}$ is needed for the trade-off between the correlation and the strength. In Random Forest, the optimal value of $\boldsymbol{m}$ is found by using the oob-error rate (out-of-bag-error rate). It is reported that face classification by Random Forest achieves a lower error rate than some popular classifiers, including SVM (156). As far as we know, there is no reported work in the literature of face-based gender classification using Random Forest.

4.3.4 Experiments

The FRGCv2 database was collected by researchers from the University of Notre Dame and contains 4007 3D face scans of 466 subjects with differences in gender, ethnicity, age and expression (157). For gender, there are 1848 scans of 203 female subjects and 2159 scans of 265 male subjects. The ages of subjects range from 18 to 70, with 92.5% in the 18-30 age group. When considering ethnicity, there are 2554 scans of 319 White subjects, 1121 scans of 99 Asian subjects, 78 scans of 12 Asian-southern subjects, 16 scans of 1 Asian and Middle-east subject, 28 scans of 6 Black-or-African American subjects, 113 scans of 13 Hispanic subjects, and 97 scans of 16 subjects subjects whose ethnicity are unknown. About 60% of the faces have a neutral expression, and the others show expressions of disgust, happiness, sadness and surprise. All the scans in FRGCv2 are near-frontal. With FRGCv2, we perform two types of experiments. The first type is to examine the robustness of our approach to age and ethnicity variations. It uses the 466 earliest scans of each subject in FRGCv2, of which more than 93% are neutral-frontal. The second type extends to examine the robustness of our approach to variations of expressions. It enrolls all the 4007 scans in FRGCv2, about 40% of which are expressive faces. For both types of experiments, results are generated in a subject-independent fashion, using the 10-fold cross-validation approach.

4.3.4.1 Data preprocessing

Since there are holes (caused by absorption of laser in dark areas like eyebrows and eyes), hair and spikes (acquisition noise) in the raw face images, preprocessing is needed to limit their influences. Firstly, through boundary detection, link-up and triangulation, holes are filled in each scan. Secondly, since the scans in FRGCv2 are all near frontal, the nose tip is detected with a simple algorithm; then the mesh is cropped with a sphere centered at nose tip to discard the hair region. Finally a smoothing filter is used to distribute evenly the 3D vertices which capture the original 3D shape. We then apply the ICP algorithm on each scan to rotate the face to the upright-frontal position.

4.3.4.2 Robustness to variations of age and ethnicity

Among the 466 earliest scans, 431 scans are neutral-frontal and 35 are expressional-frontal. In 10-fold cross validation, the 466 scans are randomly partitioned into 10 folds with each fold containing 46-47 scans. In each round, 9 of the 10 folds are used for training while the remaining fold is used for testing. The average recognition rate and standard deviation for 10 rounds then give a statistically significant performance measure of proposed methodology. The relationship between gender classification result and the number of trees used in Random Forest is depicted in Fig. 4.8(a). It evidently demonstrates that a significant relationship exists between gender and facial averageness. Facial symmetry is also closely related with gender, which echoes previous findings in anthropometrical study (142). We perceive also that both the fusion and the feature selection improve the gender classification results. The fusion descriptor outperforms individual averageness and symmetry descriptor. It means that facial averageness and symmetry relate with gender in different ways. They are complimentary in face gender perception. At the same time, results after the feature selection almost always override the results without feature selection. It means that the original averageness DSF and symmetry DSF contain redundant information. Gender-related features are distributed unequally in face regions. The best gender classification result is 93.78%, achieved by 80-Tree Random Forest with the fusion descriptor after feature selection. The result is detailed in the confusion matrix in Table 4.1. The recognition rate for females (92.02%) is slightly lower than for male ones (95.44%). It is probably due to the fact that more male faces were used for training. We also performed a 10-fold 100-repetition experiment with Random Forest under the same setting, which resulted at an average correctness of 92.84% with a standard deviation of 3.58%.

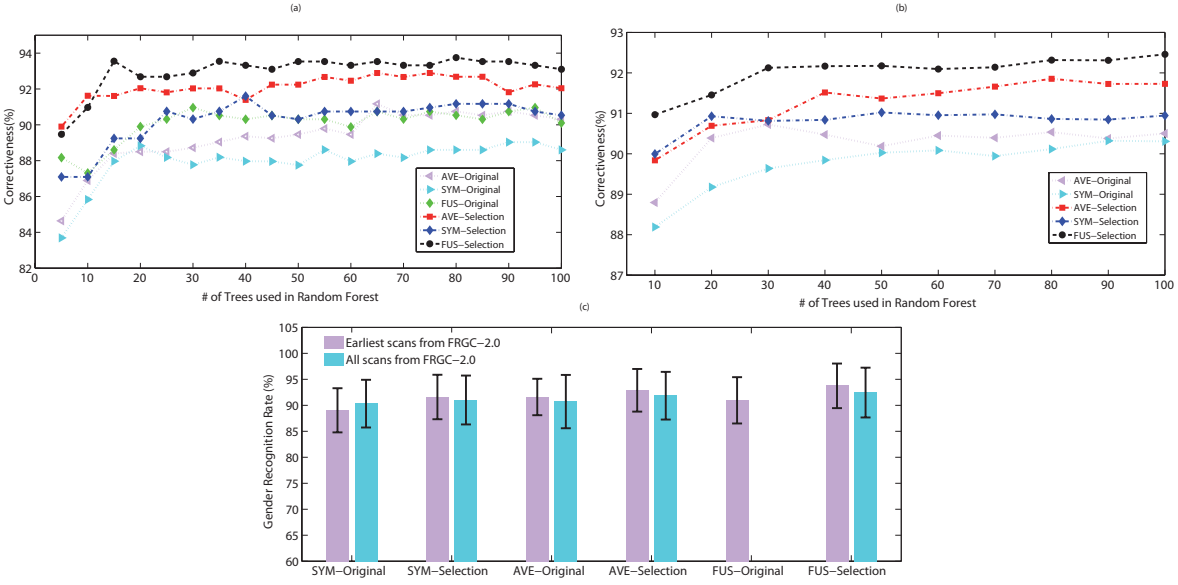


FIGURE 4.8: Results of our approach using Random Forest with different number of trees (a) on 466 earliest scans (mainly neutral), and (b) on whole FRGCv2 dataset (with facial expression variations). (c) Best average recognition rates with standard deviations.

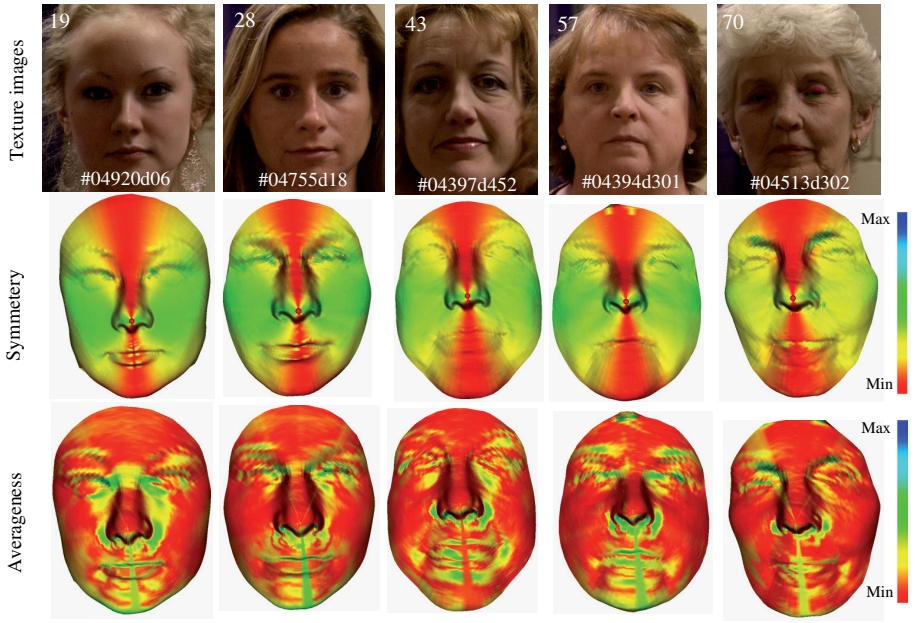


FIGURE 4.9: DSFs on face with different Age.

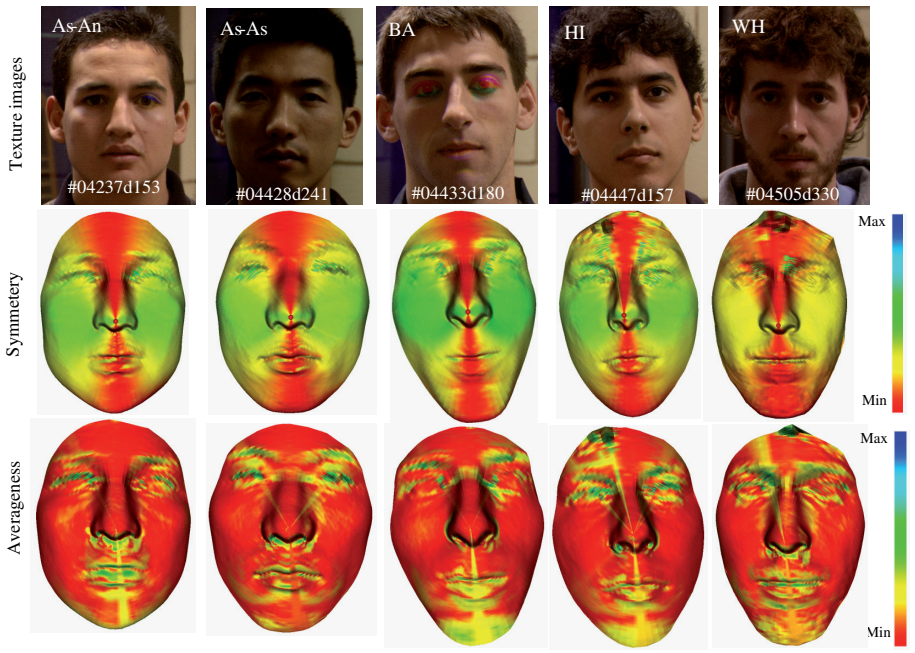


FIGURE 4.10: DSFs on face with different Ethnicity.

Fig. 4.9 illustrates the color-maps of symmetry DSF and averageness DSF on female faces with age differences and Fig. 4.10 illustrates the color-maps of symmetry DSF and averageness DSF on male faces with differences in ethnicity. The information related to age, ethnicity and id of scans is presented in the 2D images in the upper row of each figure. With the middle rows of Fig. 4.9 and Fig. 4.10, we observe that the symmetry deformations of both gender convey a visually symmetrical pattern, where the color-map of left-face is globally in symmetry with the right-face, although sutble local asymmetry exists.

Low-level deformations (red color) are usually located near the middle plane and high-level deformations (yellow and green colors) happen more frequently in farther areas. Deformations in female face changes obviously more smoothly than male. With the lower rows of Fig. 4.9 and Fig. 4.10, we observe that female faces require more deformation in mouth, nose and eye regions to deform from the averageness face template. More subtly, in cheek and forehead regions, the color is more consistent in male faces. All of these observations above stay relatively consistent with changes of age and ethnicity. We believe that these common patterns contribute to the robustness of our approach to variations of age and ethnicity to some extent.

TABLE 4.1: Confusion matrix of RF-based classification.

%	Female	Male
Female	91.63	8.37
Male	4.56	95.44
Recognition Rate =93.78 ± 4.29 %		

4.3.4.3 Robustness to expression variations

For the whole FRGCv2 dataset, we obtained 4005 well preprocessed scans after preprocessing. The failed two scans (with scan id 04629d148 and 04815d208) were resulted from wrong nose tip detection. Considering the ratio of failure is rather tiny ($2/4007<0.0005$), we omit the influence of the two failed scans for the result generation. With the 4005 well preprocessed scans, we first performed the DSF extraction for averageness, symmetry and fusion descriptors, and then did the 10-fold a subject-independent cross-validation with Random Forest. For each round, scans of 46 subjects are randomly selected for testing, and the scans of the remaining subjects dedicated for training. For all the 10 rounds of experiments, no common subjects are used in testing. The relationship between the classification result and the number of trees used in Random Forest is shown in Fig. 4.8(b). We perceive again that both fusion and feature selection improve the results. The best result achieved with fusion descriptor and feature selection is 92.46% ± 4.79 with 100-Tree Random Forest. Considering the FRGCv2 dataset is a really big and challenging dataset which contains as many as 4007 scans with various changes in age, ethnicity and expression, we claim even more confidently that a significant relationship exists between gender and 3D facial averageness/symmetry, and our method is effective and strongly robust to age, ethnicity and even expressions in gender classification.

Fig. 4.11 shows color-maps of DSFs generated for a subject with different expressions. Similar to the perceptions in Fig. 4.9 and Fig. 4.10, we perceive again in the middle row of Fig. 4.11 that the symmetry deformations on both sides of face are globally in symmetry, although tiny local asymmetry exists in areas like eye corners and lips. Low-level deformations (red) always locate near the middle plane and high-level deformations (yellow and green) occur more frequently in farther areas. With the lower rows of Fig. 4.9 and Fig. 4.11, we observe again that female faces require more deformation in mouth, nose and eye

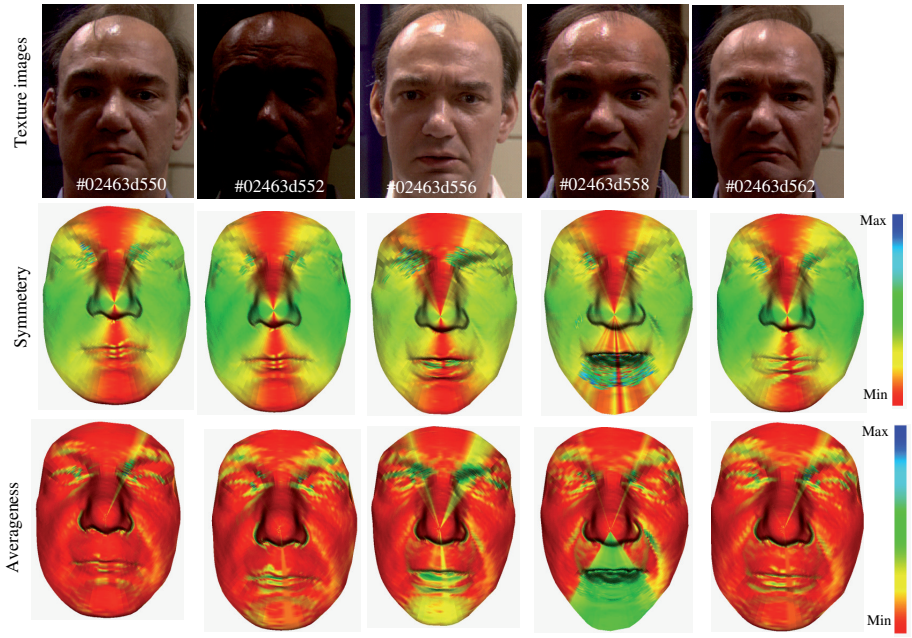


FIGURE 4.11: DSFs on face with different expressions.

regions to deform from the averageness face template. In cheek and forehead regions, the color is more consistent in male faces. All these visible patterns do not change significantly with expression variation. We assume that these patterns contribute to the robustness of our approach to expression changes. Fig. 4.8(c) shows the best gender recognition results (shown as bars) and their standard deviation (shown as black lines) in our experiment. It shows that the gender recognition rate increases with both fusion and feature selection, and the performance of all approach change little between the 466 earliest scans protocol and the whole FRGCv2 dataset protocol. It means our approach is even relatively robust to the volume of the training set.

4.3.4.4 Comparison with state of the art

Table 4.2 gives a comparison of works in previous works. With huge differences in dataset, landmarking, experiment settings and so on, it is difficult to compare and rank these works simply according to the result values. Compared with our work, works in (141), (146), (147) are based on relatively smaller dataset which leave doubts of the statistical significance of their approaches on larger and more challenging datasets, works in (141), (144), (146), (147) require manual landmarking thus their approaches are not fully-automatic, works in (141), (146), (147), (148) use different experiment settings other than the most prevailing 10-fold cross-validation. Our work addressed gender classification in a fully automatic way without manual landmarking, experimented on a large dataset FRGCv2 which contains challenging variations in expression, age and ethnicity, and reached comparable results with literature. The nearest works to ours are done by *Ballihi et al.* in (149) and *George et al.* in (135). With the 466 Earliest scans of FRGCv2 and standard 10-fold cross-validation, *Ballihi et al.* achieved 86.05% correctness in

(149), while we achieved a much higher result of 93.78% with Random Forest when combining facial shape averageness and bilateral asymmetry. In (135), *George et al.* also performed automatic 10-fold cross-validation on FRGCv2 dataset in a subject-independent fashion. Their result (Male: 94% Female: 93%) based on 3676 scans of White and Asian subjects is slightly higher in value comparing with ours (92.46%). However, our experiment on the whole FRGCv2 dataset has covered all the 4007 scans in FRGCv2, thus encountered more challenges from data amount and ethnicity variation. Additionally, during the experiments we found an error in the meta-data of FRGCv2, which mislabeled the gender of one of the subjects (with id 04662, female indeed) as male and resulted in 8 mislabeled scans. We corrected it and carried out our work with the correct meta-data.

TABLE 4.2: Comparison of our approach to earlier studies.

Reference	*Dataset*	*Manual landmarks*	*Features*	*Classifiers*	*Experiment settings*	*Results*	*Shape/Texture*
Ballihi et al. (149)	466 *earliest scans of FRGCv2*	*No*	20 *salient curves selected by Adaboost with faces from FRGC-1.0*	*Adaboost (Classification done only with scans of FRGCv2)*	*10-fold cross-validation*	86.05%	*Shape*
Toderici et al. (135)	3676 *scans from FRGCv2*	*No*	*Wavelets*	*Polynomial kernel-SVM*	10-*fold cross-validation*	*Male:* 94 $\pm$ 5% - *Female:* 93 $\pm$ 4%	*Shape*
Hu et al. (148)	*729 frontal 3D scans from UND and 216 private scans*	*No*	*Curvature-based shape index for 5 face regions*	*RBF-SVM*	5-*fold cross-validation*	94.03%	*Shape*
Han et al. (146)	*61 capture of 61 sujects in GavabDB Dataset*	*Yes*	*Geometry Features*	*RBF-SVM*	5-*fold cross-validation*	82.56 $\pm$ 0.92%	*Shape*
Wu et al. (147)	*Needle maps of* 260 *subjects from UND*	*Yes*	*PGA features*	*Posteriori Probabilities*	*6 experiments with each contains 200 scans for training and 60 for testing*	93.6 $\pm$ 0.04%	*Shape+Texture*
Lu et al. (144)	1240 *scans of* 376 *subjects from UND and MSU datasets*	*Yes*	*Concatenated Grid element values*	*SVM & Posteriori Probabilities*	10-*fold cross-validation*	91 $\pm$ 0.03%	*Shape+Texture*
Liu et al. (141)	*111 scans of 111 subjects from University of South Florida*	*Yes*	*Variance Ratio (Vr) of Features on HD and OD face*	*A linear classifier develop themselves*	100 *repetition with half scans for training and half for testing*	*HD:*91.16 $\pm$ 3.15% *OD:*96.22 $\pm$ 2.30%	*Shape*
Our work[1]	*466 Earliest scans of FRGCv2*	*No*	*AVR+SYM DSFs*	*Random Forest*	10-*fold cross-validation*	93.78 $\pm$ 4.29%	*Shape*
Our work[2]	***All scans of FRGCv2***	*No*	*AVR+SYM DSFs*	*Random Forest*	*10-fold cross-validation in a subject-independent fashion*	92.46 $\pm$ 3.58%	*Shape*

Face age estimation performs important social roles in human-to-human communication. Studies in cognitive psychology, presented as a review by (136), have discovered that human beings develop the

ability of face age estimation naturally in early life, and can be fairly accurate in deciding the age or age group with a given face. These studies, based on subjective age estimation given to face image from human participants, have also found that multiple cues contribute to age estimation, including the holistic face features (like the outline of the face, face shape and texture, etc.), local face features (like the eyes, nose, the forehead, etc.) and their configuration (like the bilateral symmetry of the face (151)). Whereas, claims has also been given that individuals are not sufficiently reliable to make fine-grained age distinctions, and individuals age estimation suffers from the subjective individual factors and contextual social factors.

The aging process is a cumulative, uncontrollable and personalized slow process, influenced by intrinsic factors like the gene and gender, and extrinsic factors like lifestyle, expression, environment and sociality (135, 138). The appearance and anatomy of human faces changes remarkably with the progress of aging (147). The general pattern of the aging process differs in faces of different person (personalized or identity-specific), in faces of different age (age-specific), in faces of different gender (gender-specific), and in different facial components (135–137, 156, 157). Typically, the craniofacial growth (bone movement and growth) takes place during childhood, and stops around the age of 20, which leads to the re-sizing and re-distribution of facial regions, such as the forehead, eyes, nose, cheeks, lips, and the chin. From adulthood to old age, face changes mainly in the skin, such as the color changes (usually darker and with more color changes) and the texture changes (appearance of wrinkles). The shape changes of faces continues from adulthood to old age. With the droops and sags of facial muscle and skin, the faces are tend to be more a shape of trapezoid or rectangle in old faces, while the typical adult faces are more of a U-shaped or upside-down-triangle (136).

Automatic face age estimation is to label a face image with the exact age or age group objectively by machine. With the rapid advances in computer vision and machine learning, recently, automatic face age estimation have become particularly prevalent because of its explosive emerging and promising real-world applications, such as electronic customer relationship management, age-specific human-computer-interaction, age-specific access control and surveillance, law enforcement (e.g., detecting child-pornography, forensic), biometrics (e.g., age-invariant person identification (156)), entertainment (e.g., cartoon film production, automatic album management), and cosmetology. Compared with human age estimation, automatic age estimation yields better performance as demonstrated in (138). The performance of age estimation is typically measured by the mean absolute error (MAE) and the cumulative score (CS). The MAE is defined as the average of the absolute errors between the estimated age and the ground truth age, while the CS, proposed firstly by (149) in age estimation, shows the percentage of cases among the test set where the absolute age estimation error is less than a threshold. The CS measure is regarded as a more representative measure in relation with the performance of an age estimator (145).

As pointed in (136, 158), the earliest age estimation works used the mathematical cardioidal strain model, derived from face anthropometry that measures directly the sizes and proportions in human face, to describe the craniofacial growth. These approaches are useful for young ages, but not appropriate for

adults. After this, abundant works exploiting 2D images have been published in the literature with more complex approaches. Different with the comprehensive surveys given by (136, 158), which categorized the literature concerning different aging modeling techniques, we represented the literature with the different ideas underlying these technical solutions. Based on the previous statements, we describe the face appearance as a function of multiple factors, including the age, the intrinsic factors (permanent factors like gene, gender, ethnicity, identity, etc.), and the extrinsic factors (temporary factors like lifestyle, health, sociality, expression, pose, illumination, etc.).

A. General aging patterns in face appearance. Essentially, face age estimation is to estimate the age of a subject by the aging patterns shown visually in the appearance. To analyze the appearance given in the face image is the basic ways to estimate the age. In the literature of age estimation, works were carried out with several different perceptions of the general aging patterns in face appearance. As aging exhibits similar patterns among different person, several approaches have been designed to learn the general public-level aging patterns in face appearance for age estimation. The most representative ones are the Active-Appearance-Model (AAM) based approaches, the manifold embedding approaches, and the Biologically-Inspired-Feature (BIF) based approaches. The common idea underlying these approaches is to project a face (linearly or non-linearly) into a subspace, to have a low dimensional representation. Respectively, ***(i)*** (144, 147) use an Active Appearance Model (AAM) based scheme for projecting face images linearly into a low dimensional space. The AAM was initially proposed by (159), in which each face is represented by its shape and texture deviations to the mean face with a set of model parameters. Age estimation results with a quadratic regressor showed that the generic aging patterns work well for age estimation. Moreover,(144) illustrated that different face parameters obtained from training are responsible for different changes in lighting, pose, expression, and individual appearance. Considering that these parameters work well for age estimation, we can conclude that these face co-variants are influential in age estimation. ***(ii)*** The goal of manifold embedding approaches is to embed the original high dimensional face data in a lower-dimensional subspace by linear or non-linear projection, and take the embedding parameters as face representation. In the work of (152, 157), the authors extracted age related features from 2D images with a linear manifold embedding method, named Orthogonal Locality Preserving Projections (OLPP). (160) learned age manifold with both local preserving requirements and ordinal requirements to enhance age estimation performance (161) projected each face as a point on the Grassmann Manifold with the standard SVD method, then the tangent vector on these points of the manifold were taken as features for age estimation. ***(iii)*** Inspired by a feed-forward path theory in cortex for visual processing, (137) introduced the biologically inspired features (BIF) for face age estimation. After filtering a image with a Gabor filter and a standard deviation based filter consecutively, the obtained features are processed with PCA to generate lower-dimension BIF features. The results demonstrated the effectiveness and robustness of bio-inspired features in encoding the generic aging patterns. Beyond the public-level aging patterns, there could be some less generic aging patterns when dealing with a subset of faces, such as a group of faces with high similarity, or a temporal sequence of face images for the same person. Based on the observation that similar faces tend to age similarly, (144, 147) presented an

appearance-specific strategy for age estimation. Faces are firstly clustered into groups considering their inter similarity, then training is performed on each group separately to learn a set of appearance-specific age estimators. Given a previously unseen face, the first step is to assign it to the most appropriate group, then the corresponding age estimator makes the age estimation. Experimental results showed that the group-level aging patterns are more accurate in age estimation compared with the generic-aging patterns. In case there is no similar enough face image for a testing face image in the database, (147) presented a weighted-appearance-specific which also yield fine performance. As different individual ages differently, (149, 155) proposed the Aging-Pattern-Subspace (AGES), which studies the individual-level aging patterns from a temporal sequence of images of an individual ordered by time. For a test face, the aging pattern and the age is determined by the projection in the subspace that has the least reconstruction error. Experiments confirm that individual aging patterns contributes to age estimation. As different face components age differently, the component-level aging patterns are studied for age estimation. (162) represented faces with a hierarchical And-Or Graph. Face aging is then modeled as a Markov process on the graphs and the learned parameters of the model are used for age estimation. They found that the forehead and eye regions are the most informative for age estimation, which is also supported by discoveries of (138) using the BIF features.

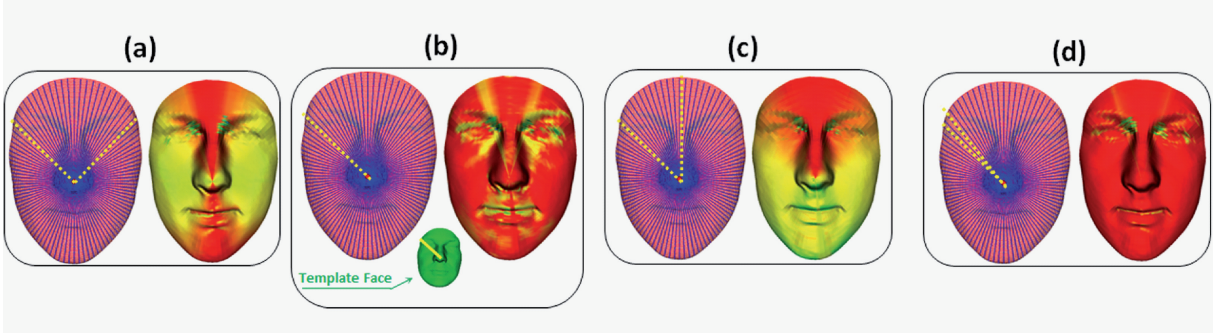


FIGURE 4.12: Illustrations of different DSFs on preprocessed face $\boldsymbol{S}$. (a) Symmetry DSF: the DSF from radial curve $\beta_\alpha^{\boldsymbol{S}}$ to its bilateral symmetrical curve $\beta_{2\pi-\alpha}^{\boldsymbol{S}}$; (b)Averageness DSF: DSF from radial curve $\beta_\alpha^{\boldsymbol{S}}$ in a preprocessed face to radial curve $\beta_\alpha^{\boldsymbol{T}}$ in an average face template (with the same angle index α); (c) Spatial DSF: DSF from radial curve $\beta_\alpha^{\boldsymbol{S}}$ to the middle radial curve $\beta_0^{\boldsymbol{S}}$ in the forehead; (d) Gradient DSF: DSF from radial curve $\beta_\alpha^{\boldsymbol{S}}$ to its neighbor curve $\beta_{\alpha+\Delta\alpha}^{\boldsymbol{S}}$

B. Considering the intrinsic/extrinsic factors in facial aging. As stated at the beginning of this introduction, the appearance of face is influenced by intrinsic factors like the gene, gender, and extrinsic factors like lifestyle, expressions, environment and sociality (135, 138). Several studies have given consideration of the influences of these factors in age estimation with enhanced age estimation performance reported. Specifically, thinking that faces age differently in different age, age-specific approaches are adopted by (144), where age estimation is obtained by using a global age classifier first, then adjusted the estimated age by a local classifier which operates within a specific age range. Similarly, (152, 157) proposed a Locally Adjusted Robust Regressor (LARR) for age estimation, which begins with a SVR-based global age regression, then followed by a local SVM-based classification that adjusts the age estimation in a local age range. All of these age-specific approaches have achieved better performance compared with their

corresponding approaches without local adjustment. Considering that different gender ages differently with age (152, 158), (146, 152, 158, 163) carried out age estimation on male and female groups separately. Considering the individual lifestyle, (147) encoded this information together with facial appearance in age estimation, and demonstrated that the importance of lifestyle in determining the most appropriate aging function of a new individual. (146) gave weights to different lighting conditions for illumination-robust face age estimation. (160) gave consideration of the feature redundancy and used feature selection to enhance age estimation.

As stated before, in the childhood, face deformation mainly takes the form of craniofacial growth with facial features re-sized and re-distributed. From adulthood to old age, with the droops and sags of facial muscle and skin, the old faces usually deform to a trapezoid or rectangle shape from a typically U-shaped or upside-down-triangle in adult face (136). Another significant shape deformation is the introduction of facial wrinkles with aging. While, given the fact that face shape deforms significantly with age in three dimensions, and given the robustness of 3D face scans to illumination and poses compared with 2D face images, all the previous works in the literature used 2D face datasets for age estimation, no work has been done concerning the 3D face. Thus, in this work, we introduce the investigation of age estimation with 3D face scans. The rest of the paper is organized as follows: in section **2**, we present an overview of our methodology and summarize the main contributions; in section **3**, we explain our methodology of features extraction from the 3D faces based on an Riemann framework; in section **4**, we detail the regression strategy for age estimation using Random Forest; experimental results and their discussion are presented in section **5** while section **6** comes to the conclusion of this work.

Methodology and contribution

From the analysis above, it emerges that most of the existing works study age estimation with aging patterns chosen at a specified level and some aging factors enrolled for enhancement. As far as we concern, all these works are based on 2D images, no work concerning 3D face scans has been attached to age estimation. Thus, we introduce in the present work a new study of 3D-base face age estimation to the domain. In our approach, we consider the public-level aging patterns and gender factor for age estimation. First, we extract four types of Dense Scalar Field (DSF) features from each pre-processed face, namely the Average DSF, the Symmetry DSF, the Spatial DSF and the Gradient DSF. These DSFs are derived from different face perception ideas and their computation is grounding on Riemannian shape analysis of facial curves. Then we perform age estimation using Random Forest Regression on each type of DSFs with two protocols: one experiment on DSFs of the whole dataset directly and the other experiment on male and female DSFs separately. We have also designed a simple result-level fusion with different type of the DSFs, to see if the performance improves with all these face perception ideas combined. In summary, the main contributions of this work are as follows. First, as far as we know, this is the first work in 3D-based age estimation. Although 3D face growth has been notice for a long time (124, 164), no work has been reported to 3D face age estimation. Secondly, in this work, we introduce four different perspectives of faces perception for face representation. With the Dense Scalar Field features,

we have obtained significant accuracy with each of the perspectives, compared with typical 2D-based age estimation performance. Last but not the least, we have enhanced the age estimation performance by experimenting on the scans of each gender separately, which confirms that the sexual dimorphism exists in terms of face aging patterns. We have also enhanced the performance by a simple late fusion rule of the four descriptors.

In our approach, the raw 3D face scans are first pre-processed for hole-filling, cropping, smoothing and pose normalization, and then represented by a set of parameterized radial curves emanating from the nose tip of the preprocessed face denoted with $\boldsymbol{S}$. The radial curve that makes an clockwise angle of α with the radial curve which passes through the forehead (β_0) is denoted as β_α, and the neighbor curve of β_α that has an angle increase of $\Delta\alpha$ is denoted as $\beta_{\alpha+\Delta\alpha}$. Such representation can be seen as a approximation of the preprocessed face $\boldsymbol{S}$. To extract the DSF features, one need to first define the correspondence of curves in pair-wise shape comparison. With four different perspectives from face perception, we define four different types of correspondence in pair-wise shape comparison, which results into four different types of DSF features with all the radial curves considered in a face, namely the Symmetry DSF, the Averageness DSF, the Spatial DSF and the Gradient DSF. Figure 4.12 gives an illustration of these DSF features. The Symmetry DSF shown in sub-figure (a) captures the deformation between a pair of bilateral symmetrical radial curves ($\beta_\alpha^{\boldsymbol{S}}$ and $\beta_{2\pi-\alpha}^{\boldsymbol{S}}$) in a preprocessed face $\boldsymbol{S}$. The Symmetry DSF conveys the idea that the bilateral facial symmetry loses with age. The Averageness DSF shown in sub-figure (b) compares a pair of curves with the same angle index from a preprocessed face $\beta_\alpha^{\boldsymbol{S}}$ and an average face template $\beta_\alpha^{\boldsymbol{T}}$. The average face template $\boldsymbol{T}$ (as presented in sub-figure (b)) is defined as the middle point of geodesic deformation path from a representative male scan to a representative female scan. The Averageness DSF represents the idea that faces become more personalized and thus deviates more from the average face shape with age. The Spatial DSF shown in sub-figure (c) captures the deformation of a curve β_α to one reference radial curve β_0 in the forehead in a preprocessed face $\boldsymbol{S}$. As β_0 is the most rigid curve in the face, the Spatial DSF can be perceived as the cumulative deformation from the most rigid part of the face. The Gradient DSF shown in sub-figure (d) captures the deformation between a pair of neighbor curves ($\beta_\alpha^{\boldsymbol{S}}$ and $\beta_{\alpha+\Delta\alpha}^{\boldsymbol{S}}$) in a preprocessed face $\boldsymbol{S}$. In contrast with the Spatial DSF, the Gradient DSF can be viewed as a representation of local deformation on the face. In each sub-figure of Figure 4.12, the left part shows the extracted radial curves in the face and correspondence for curve comparison, the right part shows the corresponding DSF features as color-map on the face, where on each face point, the hotter the color, the lower of the DSF magnitude.

4.3.5 Random Forest Regression

Age estimation can be considered as a classification problem, when each age is taken as a class label. On the other hand, age estimation can also be considered as a regression problem, since the age could be interpreted as continuous value. Note that there are only 15 subjects of more than 40 years old in FRGCv2, the number of faces is too small to train classifiers for those ages. Thus, in our approach, we

take the age estimation as a regression problem. Similar reason has been used by (137) for choosing the regression strategy for age estimation on the FG-net dataset, where the images from old subjects is also very rare. As summarized by (133), the regression task is, given a labeled set of training data, learning a general mapping which associates previously unseen, independent test data points with their dependent continuous output prediction. In the work of (139), Random Forest regression has demonstrated nice age estimation performance (3.43 MAE) in LOPO experiments for the young age subset of the FG-net dataset. As far as we concern, no studies have investigated the age estimation performance of Random Forest with the overall age distribution. Thus, we adopt the Random Forest in our regression experiments to demonstrate its capability in age estimation. Technically, Random Forest is an ensemble learning method that grows many classification trees $t \in \{t_1, .., t_T\}$. To estimate age from a new face from an input vector (DSF-based feature vector $v = V_\alpha^k$), each tree gives a regression result and the forest take the average of estimated ages as the final result. In the growing of each tree, two types of randomness are introduced consecutively. Firstly, a number of $\boldsymbol{N}$ instances are sampled randomly with replacement from the original data, to make the training set. Then, if each instance comprises of $\boldsymbol{M}$ input variables, a constant number $\boldsymbol{m}$ $\boldsymbol{(m<<M)}$ is specified. At each node of the tree, $\boldsymbol{m}$ variables are randomly selected out of the $\boldsymbol{M}$ and the best split on these $\boldsymbol{m}$ variables is used to split the node. The process goes on until the tree grows to the largest possible extent without pruning, where the resulted subsets of the node are totally purified in label.

4.3.6 Experiments

Our experiments are carried out with Random Forest Regression on FRGCv2 dataset. The FRGCv2 dataset was collected by researchers from the University of Notre Dame and contains 4007 3D near-frontal face scans of 466 subjects, where 203 are female and 263 are male (141). The age of subjects ranges from 18 to 70, with 92.5% in the 18-30 age group. Our experiments are performed with the 466 earliest scans of each subject in FRGCv2. With the 466 earliest scans, we design two experiment protocols. The first protocol, named ***Gender-General-Protocol*** (GGP), experiments on the 466 scans directly with Random Forest Regression. While the second protocol, named ***Gender-Specific-Protocol*** (GSP), separates the 466 scans into male group and female group first, and then performs experiments on each group separately with Random Forest Regression. For all the two protocols, experimental results are generated using the ***Leave-One-Person-Out*** (LOPO) cross-validation strategy, where each time one scan of the concerning data (all 466 scans or scans of each gender) is used as testing face once, with the rest the scans used in training. Thus, there are altogether 466 experiments in the cross-validation in each protocol, and each scan is tested equally only once.

4.3.6.1 Gender-general experiment

As described above, with the *Gender-General-Protocol* (GGP), we perform Leave-One-Person-Out cross-validation experiments directly with the 466 earliest scans of FRGCv2 dataset for each descriptor. Each time one scan is picked out for testing and the rest 465 scans are used for training. Table 4.3 shows the experimental results as the mean of the absolute error between the truth and the estimated age for each tested scan in corresponding age group. By taking the minimum value of the estimated ages given by the four descriptors as the age estimation result, we have also obtained the fusion results, as shown also in Table 4.3. From this table, we observe that we achieve a minimum overall mean absolute error (MAE) about 3.7 years by the Averageness and Spatial DSFs. For the other two descriptors, the overall mean absolute errors are a little higher, while both of them are under 4 years. Thus, from the perception of the overall mean absolute errors, we find that our approach with all of the four descriptors are effective in age estimation. Moreover, when we go inside of the details of these results for each age group, we find that the age estimation performance declines significantly with aging. We assume that the big decrease of the number of scans in aged groups (from about 200 to about 20) accounts largely for this performance decline. From the same table, we also observe that the fusion method, which takes the minimum of the estimated ages concerning each of the four descriptors, yields a better overall mean absolute error of 3.29 years. It means that the age related cues in these descriptors are different and complimentary in age estimation. When going inside of the detail of the fusion result for each age group, we find the enhancement of overall performance is mainly coming from young age groups. It is probably due to the fact that for young age groups, more scans are available in training for each descriptor. Thus the estimation results from each descriptor for young age groups are less biased for making the fusion decision.

TABLE 4.3: Age estimation results for different age groups with the Gender-General-Protocol. (MAE:Mean Absolute Error; AVR: Averageness; SYM: symmetry; GRA: gradient; SPA: spatial; MIN: minimum rule for fusion.)

Age group	MAE AVR	MAE SYM	MAE GRA	MAE SPA	Fuse MIN	♯ of scans
≤ 20	3.48	3.43	3.77	3.30	2.20	185
$(20, 30]$	2.18	2.58	2.32	2.38	1.98	246
$(30, 40]$	9.99	7.60	10.05	8.92	9.18	20
> 40	24.82	23.66	24.56	25.36	25.75	15
Overall	**3.76**	**3.79**	**3.94**	**3.76**	**3.29**	**466**

Figure 4.13 shows the experimental results of Gender-General-Protocol by cumulative scores for the four descriptors. The x-axis is the level of Mean Absolute Error, which represents the mean of the absolute age error (between the truth and estimated age of scan) over the 466 scans. The y-axis show the cumulative score of accuracy by percentage of acceptance. Thus, a point *(a,b)* on the curve shows, with a Mean Absolute Error tolerance of a years, it achieves an acceptance of b percent. We have also captured the fusion result in the same figure by cumulative scores. From Figure 4.13, we observe that with a

Error Level of 5 years, we achieve an acceptance of more than 75% over the 466 scans; when the Error Level is 10 years, the cumulative score of acceptance increases to more than 90%. We also observe that the fusion result is significantly higher compared with the result of each individual descriptor. From these observations, we claim again that our approach concerning all these descriptors are comparably effective in age estimation, and the result-level fusion of these descriptors enhances the age estimation performance.

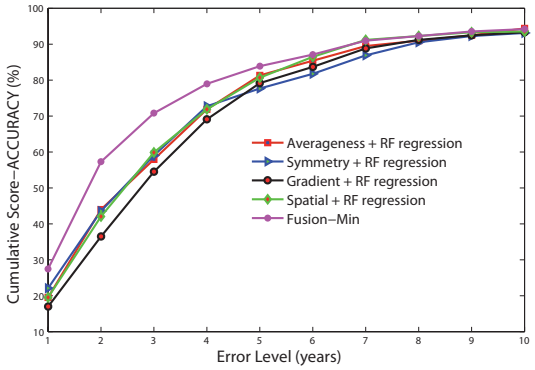


FIGURE 4.13: Age regression results in Leave-One-Person-Out experiment with the Gender-General-Protocol.

4.3.6.2 Gender-specific experiment

Gender and age are natural co-variates in human face. In (134), *Ashok Samal et al.* statistically confirm that sexual dimorphism is strong and widespread among face features, and find out the degree of dimorphism changes as a function age (e.g., the average age at which the sexual dimorphism becomes more significant is around 13). Thus, the face aging effect is considerably different with different gender. In the experiments considering the ***Gender-Specific-Protocol*** (GSP), we first separate the 466 earliest scans of FRGCv2 into male group and female group, then we perform Leave-One-Person-Out cross-validation experiments on male scans and female scans separately for each descriptor. As in the GGP experiments, each time we take one scan in testing and the rest scans in training. The final results for each descriptor are generated by statistically merging the results from each gender.

Table 4.4 shows the experimental results as the mean of the absolute error between the truth and the estimated age for each tested scan in corresponding age group. From Table 4.4, we observe that for all the four descriptors, we always achieve better overall performance with GSP. We also achieve better results in each age group with all these descriptors, except for the symmetry descriptor in the (30,40] age group. With these observations, which indicate that the Gender-Specific-Protocol outperforms the Gender-General-Protocol in age estimation, we confirm the claims in (134), that faces of different gender convey considerably different morphology of aging. Moreover, the overall fusion result outperforms again the result of each descriptor in the GSP experiments, and also the overall fusion result in the GGP experiments. It shows again that the result-level fusion of these descriptors can enhance the age estimation performance.

TABLE 4.4: Results for different age groups with the Gender-Specific-Protocol. (MAE:Mean Absolute Error; AVR: Averageness; SYM: symmetry; GRA: gradient; SPA: spatial.)

Age group	MAE AVR	MAE SYM	MAE GRA	MAE SPA	Fuse MIN	♯ of scans
≤ 20	3.25	3.38	3.46	3.19	2.14	185
(20,30]	2.03	2.16	2.14	2.18	2.04	246
(30,40]	8.97	8.52	9.18	8.81	10.43	20
> 40	20.81	22.59	21.32	22.22	24.05	15
Overall	**3.42**	**3.57**	**3.58**	**3.51**	**3.15**	**466**

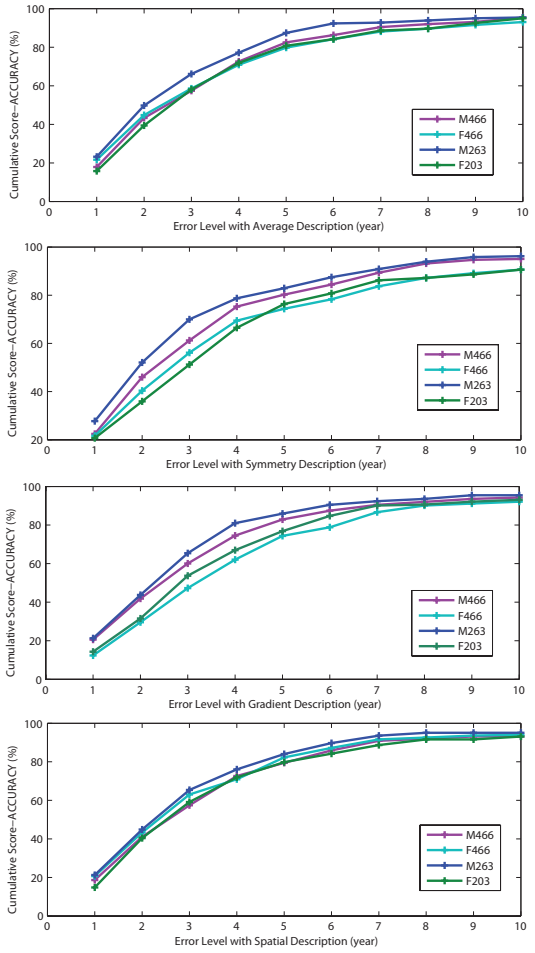


FIGURE 4.14: Comparison of results from the Gender-General-Protocol (GGP) and the Gender-Specific-Protocol (GSP) for each gender. (M-gender-general: male group in GGP experiments; F-gender-general: female group in GGP experiments; M-gender-specific: male group in GSP experiments; F-gender-specific: female group in GSP experiments.)

Figure 4.14 makes a further comparison between the GGP and GSP experiments, with the cumulative scores for each gender and for each descriptor in these two type of experiments. From Figure 4.14, we observe that, only except for the beginning part of result with the female group and symmetry descriptor, the experimental results are always significantly higher for both male and female groups in the GSP experiments for all the descriptors. That is to say, although trained with less data, the GSP

experiments have the advantage of giving better age regression results. One probable explanation for this observation is that, in the GSP experiments, the regression results do not suffer the influence from the scans in the other gender, which conveys a significantly different aging morphology. With Figure 4.14, we further confirm that the aging effect differs with gender.

4.4 Expression Recognition from 3D Dynamic Faces

Automatic recognition of facial expressions emerged as a field of active research, with applications in several different areas, such as human-machine interaction, psychology, computer graphics, transport security (by detecting driver fatigue, for example), and so on.

Ekman (127) showed that facial expressions can be coded through the movement of face points as described by a set of *action units* (128).

These results, in turn, inspired many researchers to analyze facial expressions in video data, by tracking facial features and measuring the amount of facial movements in video frames (165). This body of work demonstrates a collective knowledge that facial expressions are highly dynamical processes, and looking at sequences of face instances can help to improve the recognition performance. We further emphasize that, rather than being just a static or dynamic 2D image analysis, it is more natural to analyze expressions as spatio-temporal deformations of 3D faces, caused by the actions of facial muscles. In this approach, the facial expressions can be studied comprehensively by analyzing temporal dynamics of 3D face scans (3D plus time is often regarded as 4D data). From this perspective the relative immunity of 3D scans to lighting conditions and pose variations give support for the use of 3D and 4D data. Motivated by these considerations, there has been a progressive shift from 2D to 3D in performing facial shape analysis for recognition (166–170), and expression recognition (171, 172). In particular, this latter research subject is gaining momentum thanks to the recent availability of public 3D datasets, like the *Binghamton University* 3D Facial Expression database (BU-3DFE) (173), and the *Bosphorus* 3D Face Database (174). At the same time, advances in 3D imaging technology have permitted collections of large datasets that include temporal sequences of 3D scans (i.e., 4D datasets), such as the *Binghamton University* 4D Facial Expression database (BU-4DFE) (175), the 4D dataset constructed at *University of Central Lancashire* (Hi4D-ADSIP) (176, 177), and the dynamic 3D FACS dataset (D3DFACS) for facial expression research (178), which also includes fully coded FACS. This trend has been strengthened further by the introduction of inexpensive acquisition devices, such as the consumer 3D cameras like Kinect or Asus that provide fast albeit low-resolution streams of 3D data to a large number of users, thus opening new opportunities and challenges in 3D face recognition and facial expression recognition (179, 180).

Motivated by these facts, we focus in this work on the problem of expression recognition from dynamic sequences of 3D facial scans. We propose a new framework for temporal analysis of 3D faces that combines scalar field modeling of face deformations with effective classifiers. To motivate our solution

and to relate it to the state of the art, next we provide an overview of existing methods for 4D facial expression recognition (see also the recent work in (181) for a comprehensive survey on this subject), then we give a general overview of our approach.

4.4.1 Related Work

The use of 4D data for face analysis is still at the beginning, with just a few works performing face recognition from sequences of 3D face scans (180, 182, 183), and some works focussing on facial expression recognition.

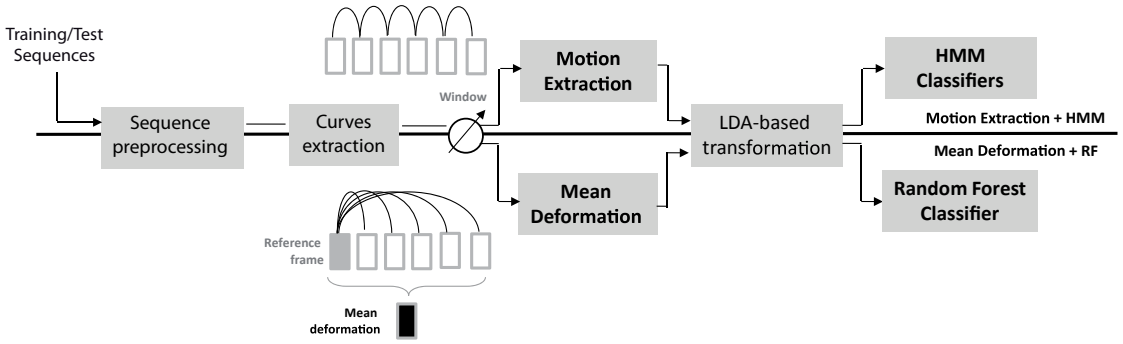


FIGURE 4.15: Overview of the proposed approach. Four main steps are shown: Sequence preprocessing and extraction of the radial curves; Motion extraction and Mean deformation computation; Dimensionality reduction with LDA; HMM- and Random-Forest-based classification. Note that both train and test sequences can go through the upper and lower path in the block-diagram.

In particular, the first approach addressing the problem of facial expression recognition from dynamic sequences of 3D scans was proposed by Sun et al. (184, 185). Their approach basically relies on the use of a generic deformable 3D model whose changes are tracked both in space and time in order to extract a spatio-temporal description of the face. A spatial HMM and a temporal one were used to model the spatial and temporal relationships between the extracted features. Exhaustive analysis was performed on the BU-4DFE database. The main limit of this solution resides in the use of the 83 manually annotated landmarks of the BU-4DFE that are not released for public use.

The approach proposed by Sandbach et al. (186) exploits the dynamics of 3D facial movements to analyze expressions. This is obtained by first capturing motion between frames using Free-Form Deformations and extracting motion features using a quad-tree decomposition of several motion fields. GentleBoost classifiers are used in order to simultaneously select the best features to use and perform the training using two classifiers for each expression. Experiments were reported for three prototypical expressions (i.e., happy, angry and surprise) of the BU-4DFE database. An extension of this work has been presented in (181), where results on the BU-4DFE database using the six universal facial expressions are reported.

In (187) a level curve based approach is proposed by Le et al. to capture the shape of 3D facial models. The level curves are parameterized using the arclength function. The Chamfer distance is applied to

measure the distances between the corresponding normalized segments, partitioned from these level curves of two 3D facial shapes. Results were reported for the happy, sad and surprise sequences of the BU-4DFE database.

Fang et al. (188) propose a fully automatic 4D facial expression recognition approach with a particular emphasis on 4D data registration and dense correspondence between 3D meshes along the temporal line. The variant of the Local Binary Patterns (LBP) descriptor proposed in (189), which computes LBP on three orthogonal planes is used as face descriptor along the sequence. Results are provided on the BU-4DFE database for all expressions and for the subsets of expressions used in (186) and (187), showing improved results with respect to competing solutions. In (190), the same authors propose a similar methodology for facial expression recognition from dynamic sequences of 3D scans, with an extended analysis and comparison of different 4D registration algorithms, including ICP and more sophisticated mesh matching algorithms, as Spin Images and MeshHOG. However, 12 manually annotated landmarks were used in this study.

Recently, Reale et al. (191) have proposed a new 4D spatio-temporal feature named *Nebula* for facial expressions and movement analysis from a volume of 3D data. After fitting the volume data to a cubic polynomial, a histogram is built for different facial regions using geometric features, as curvatures and polar angles. They have conducted several recognition experiments on the BU-4DFE database for posing expressions, and on a new database published in (192) for spontaneous expressions. However, manual intervention is used to detect the onset frame and just 15 frames from the onset one are used for classification, and these frames correspond to the most intense expression.

From the discussion above, it becomes clear that solutions specifically tailored for 4D facial expression recognition from dynamic sequences are still preliminary, being semi-automatic, or are capable of discriminating between only a subset of expressions.

4.4.2 Our Method and Contributions

Due to the increasing importance of shape analysis of objects in different applications, including 3D faces, a variety of mathematical representations and techniques have been suggested, as described above (181, 185, 190). The difficulty in analyzing shapes of objects comes from the fact that: (1) Shape representations, metrics, and models should be invariant to certain transformations that are termed *shape preserving*. For instance, rigid motions and re-parameterizations of facial surfaces do not change their shapes, and any shape analysis of faces should be invariant to these transformations. (2) Registration of points across objects is an important ingredient in shape analysis. Specifically, in comparing shapes of faces, it makes sense that similar biological parts are registered to each other across different faces. Furthermore, it is important to use techniques that allow a joint registration and comparisons of surfaces in a comprehensive framework, rather than in two separate steps. These two issues— invariance and registration—are naturally handled using Riemannian methods where one can choose metrics that are

invariant to certain transformations and form quotient spaces (termed shape spaces) by forming equivalence classes of objects that have the same shape. The elastic Riemannian metric used in this chapter provides a nice physical interpretation of measuring deformations between facial curves using a combination of stretching and bending. These elastic deformations are captured by the Dense Scalar Field features used for classifications. In summary, the main motivation of using a Riemannian approach is to perform registration that matches corresponding anatomical features, and obtain deformation fields that are physically interpretable.

Based on these premises, in this work we extend the ideas presented in (193) to propose an automatic approach for facial expression recognition that exploits the facial deformations extracted from 3D facial videos. An overview of the proposed approach is given in Fig. 4.15. In the preprocessing step, the 3D mesh in each frame is first aligned to the previous one and then cropped. From the obtained subsequence, the 3D deformation is captured based on a Dense Scalar Field (DSF) that represents the 3D deformation between two frames. Linear Discriminant Analysis (LDA) is used to transform derived feature space to an optimal compact space to better separate different expressions. Finally, the expression classification is performed in two ways: (1) using the HMM models for temporal evolution; and (2) using mean deformation along a window with Random Forest classifier. Experimental results show that the proposed approaches are capable of improving the state of art performance on the BU-4DFE database. There are three main contributions in this work,

- Benefit from the Novel Dense Scalar Fields (DSFs) defined on radial curves of 3D faces using Riemannian analysis in shape spaces of curves. These scalar fields accurately capture deformations occurring between 3D faces represented as collections of radial curves;
- A new approach for facial expression recognition from 3D dynamic sequences, that combines the high descriptiveness of DSFs extracted from successive 3D scans of a sequence with the discriminant power of LDA features using HMM and multi-class Random Forest;
- An extensive experimental evaluation that compares the proposed solution with the state of the art methods using a common dataset and testing protocols. Results show that our approach outperforms the published state of the art results.

4.4.3 Geometric Facial Deformation

One basic idea to capture facial deformations across 3D video sequences is to track mesh vertices densely along successive 3D frames. Since, as the resolution of the meshes varies across 3D video frames, establishing a dense matching on consecutive frames is necessary. For this purpose, Sun et al. (184) proposed to adapt a generic model (a tracking model) to each 3D frame using a set of 83 predefined facial landmarks to control the adaptation based on radial basis functions. A second solution is presented by Sandbach et al. (186, 194), where the authors used an existing non-rigid registration algorithm (FFD) (195) based on

B-splines interpolation between a lattice of control points. In this case, dense matching is a preprocessing step used to estimate a motion vector field between 3D frames t and t-1. The problem of quantifying subtle deformations along the sequence still remains a challenging task, and the results presented in (186) are limited to just three facial expressions: *happy*, *angry* and *surprise*.

4.4.3.1 Effect of the Nose Tip Localisation Inaccuracy on the DSF Computation

In the following, we present a study on the effects that possible inaccuracies in the detection of the nose tip can have on the computation of the proposed dense scalar field. In particular, we consider the effects on the shooting directions of the geodesic paths and the radial curves originating from the nose tip.

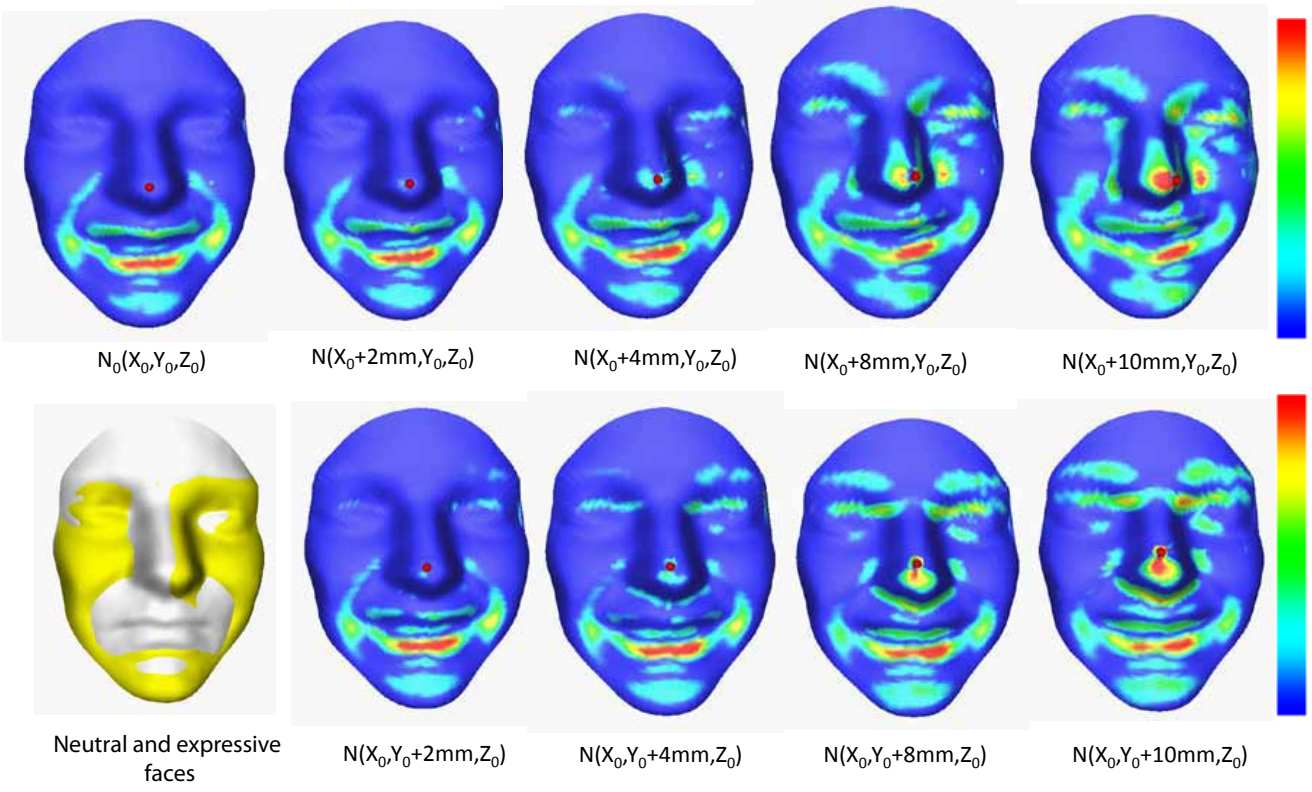


FIGURE 4.16: Effect of the nose tip placement inaccuracy on the shooting directions of the geodesic paths (or the DSFs computation). The first row illustrates DSFs when varying the nose tip position along the X-direction; The second row shows DSFs when the variation is performed along the Y-direction.

We have changed the nose tip coordinates in the X- and Y-directions and have reported the DSFs compuation results (using colormaps on the expressive faces) in Fig. 4.16. As illustrated in this figure, a large localization error ($>$ 4mm) of the nose tip generates false deformations, which could impact negatively the performance of the approach. In fact, our method is based on learning such geometric deformations to build HMMs or Random Forest classifiers. However, the left side of the figure illustrates the fact that the DSFs computation tolerates quite well errors up to 4mm.

4.4.3.2 DSF Compared to other Features

In order to compare the proposed DSF feature against other methods for extracting dense deformation features, we selected the Free-Form Deformation approach, which has been originally defined in Rueckert et al. (196) for medical images, and later on successfully applied to the problem of 3D dynamic facial expression recognition by Sandbach et al. (186, 194). In particular, FFD is a method for non-rigid registration based on B-spline interpolation between a lattice of control points. In addition, we also compared our approach with respect to a baseline solution, which uses the point-to-point Euclidean distance between frames of a sequence. Figure 4.17 reports the results for an example case, where a frame of a happy sequence is deformed with respect to the first frame of the sequence. The figure shows quite clearly as the DSF proposed in this work is capable to capture the face deformations with smooth variations that include, in the example, the mouth, the chin and the cheeck. This result is important to discriminate between different expressions whose effects are not limited to the mouth region. Differently, variations captured by the other two solutions are much more concentrated in the mouth region of the face.

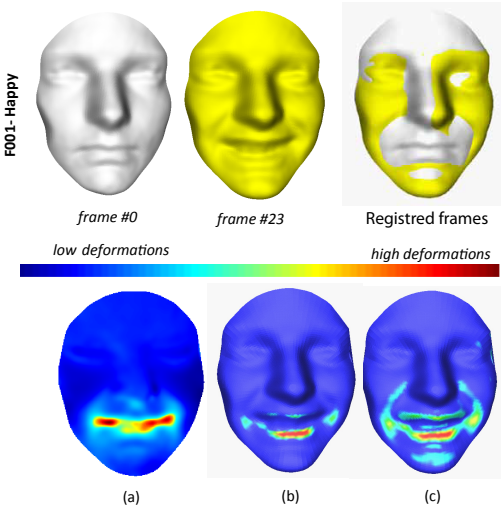


FIGURE 4.17: Comparison of the different features extracted between two frames taken from subject 001 for the happy expression: (a) the Free Form-based Deformations (FFD); (b) the point-to-point Euclidean distances; and (c) the DSFs deformations.

4.4.4 Expression Recognition using DSFs

Deformations due to facial expressions across 3D video sequences are characterized by subtle variations induced mainly by the motion of facial points. These subtle changes are important to perform effective expression recognition, but they are also difficult to be analyzed due to the face movements. To handle this problem, as described in the previous section, we propose a curve-based parametrization of the face that consists in representing the facial surface by a set of radial curves. According to this representation,

the problem of comparing two facial surfaces, a reference facial surface and a target one, is reduced to the computation of the DSF between them.

In order to make possible to enter the expression recognition system at any time and make the recognition process possible from any frame of a given video, we consider subsequences of n frames. Thus, we chose the first n frames as the first subsequence. Then, we chose n-consecutive frames starting from the second frame as the second subsequence. This process is repeated by shifting the starting index of the sequence every one frame till the end of the sequence. In order to classify the resulting subsequences, we propose two different feature extraction and classification framework based on the DSF:

- **Mean Deformation-based features associated to Random Forest classifier**. The first frame of the subsequence is considered as a reference frame and the deformation is calculated from each of the remaining frames to the first one using the DSF. The average deformation of the n-1 resulting DSFs represents the feature vector in this classification scheme and is presented, after dimensionality reduction, to multi-class Random Forest classifiers;
- **3D Motion features combined with HMM classifiers**. The deformation between successive frames in a subsequence are calculated using the DSFs and presented to an HMM classifier preceded by LDA-based dimensionality reduction.

4.4.4.1 Mean Shape Deformation with Random Forest Classifier

The idea here is to capture a mean deformation of the face in the sliding window on the 3D expression sequence. In order to get this feature, the first frame of each subsequence is considered as the reference one, and the dense deformation is computed from this frame to each of the remaining frames of the subsequence. Let F_{ref} denote the reference frame of a subsequence and F_i the i-th successive frame in the subsequence; the successive frame, for example, is denoted by F_1. The DSF is calculated between F_{ref} and F_i, for different values of i ($i = 1, \ldots, n-1$), and the mean deformation is then given by:

$$\overline{DSF} = \frac{1}{n-1} \sum_{i=1}^{n-1} DSF(F_{ref}, F_i). \tag{4.6}$$

Figure 4.18 illustrates one subsequence for each expression with $n = 6$ frames. Each expression is illustrated in two rows: The upper row gives the reference frame of the subsequence and the n-1 successive frames of the subsequences. Below, the corresponding Dense Scalar Fields computed for each frame are shown. The mean deformation field is reported on the right of each plot and represents the feature vector for each subsequence. The feature vector for this subsequence is built based on the mean deformation of the n-1 calculated deformations. Thus, each subsequence is represented by a feature vector of size equal to the number of points on the face (i.e., the number of points used to sample the radial curves of

the face). In order to provide a visual representation of the scalar fields, an automatic labeling scheme is applied: Warm colors (red, yellow) are associated with high $DSF(F_{ref}, F_t)$ values and correspond to facial regions affected by high deformations. Cold colors are associated with regions of the face that remain stable from one frame to another. Thus, this dense deformation field summarizes the temporal changes of the facial surface when a particular facial expression is conveyed.

According to this representation, the deformation of each subsequence is captured by the mean $\overline{DSF}$ defined in Eq. (4.6). The main motivation for using the mean deformation, instead of the maximum deformation for instance, is related to its greater robustness to the noise. In the practice, the mean deformation resulted more resistant to deformations due to, for example, inaccurate nose tip detection or the presence of acquisition noise. In Fig. 4.18, for each subsequence, the mean deformation field illustrates a smoothed pattern better than individual deformation fields in the same subsequence. Since the dimensionality of the feature vector is high, we use LDA-based transformation to map the present feature space to an optimal one that is relatively insensitive to different subjects, while preserving the discriminating expression information. LDA defines the within-class matrix S_w and the between-class matrix S_b. It transforms a n-dimensional feature to an optimized d-dimensional feature, where $d < n$. In our experiments, the discriminating classes are the 6 expressions, thus the reduced dimension d is 5.

For the classification, we used the multi-class Random Forest algorithm. The algorithm was proposed by Leo Breiman in (197) and defined as a meta-learner comprised of many individual trees. It was designed to operate quickly over large datasets and more importantly to be diverse by using random samples to build each tree in the forest. A tree achieves highly non-linear mappings by splitting the original problem into smaller ones, solvable with simple predictors. Each node in the tree consists of a test, whose result directs a data sample towards the left or the right child. During training, the tests are chosen in order to group the training data in clusters where simple models achieve good predictions. Such models are stored at the leaves, computed from the annotated data, which reached each leaf at train time. Once trained, a Random Forest is capable to classify a new expression from an input feature vector by putting it down each of the trees in the forest. Each tree gives a classification decision by voting for that class. Then, the forest chooses the classification having the most votes (over all the trees in the forest).

4.4.4.2 3D Motion Extraction with HMM Classifier

The DSF features described in Sect. 4.4.3, can also be applied for expression recognition according to a different classification scheme. The deformations between successive frames in the subsequence are calculated using the DSF. In particular, the deformation between two successive 3D frames is obtained by computing the pairwise Dense Scalar Field $DSF(F_{t-1}, F_t)$ of correspondent radial curves. Based on this measure, we are able to quantify the motion of face points along radial curves and thus capture the changes in facial surface geometry.

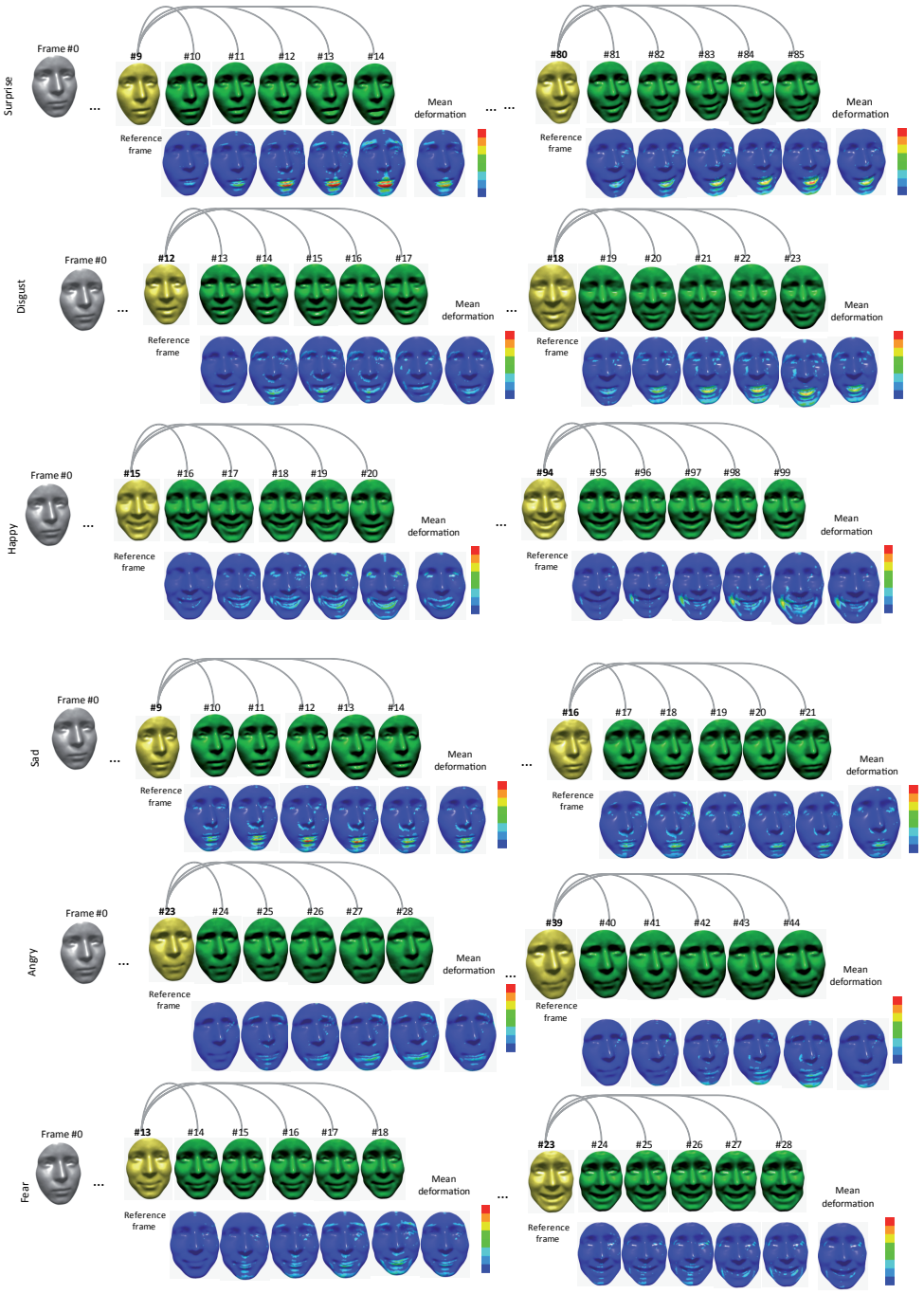


FIGURE 4.18: Computation of dynamic shape deformation on different subsequences taken from the BU-4DFE database. Each expression is illustrated by two rows: the upper one gives the reference frame of the subsequence and the n-1 successive frames. The corresponding deformation fields computed for each frame with respect to the reference one are illustrated in the lower row. The mean deformation field is given on the right of each lower row.

Figure 4.19 illustrates a direct application of the $DSF(F_{t-1}, F_t)$ and its effectiveness in capturing deformation between one facial surface to another belonging to two consecutive frames in a 3D video sequence. This figure shows two subsequences extracted from videos in the BU-4DFE database (happy and surprise expressions are shown on the left and on the right, respectively). For each sequence, the 2D image and

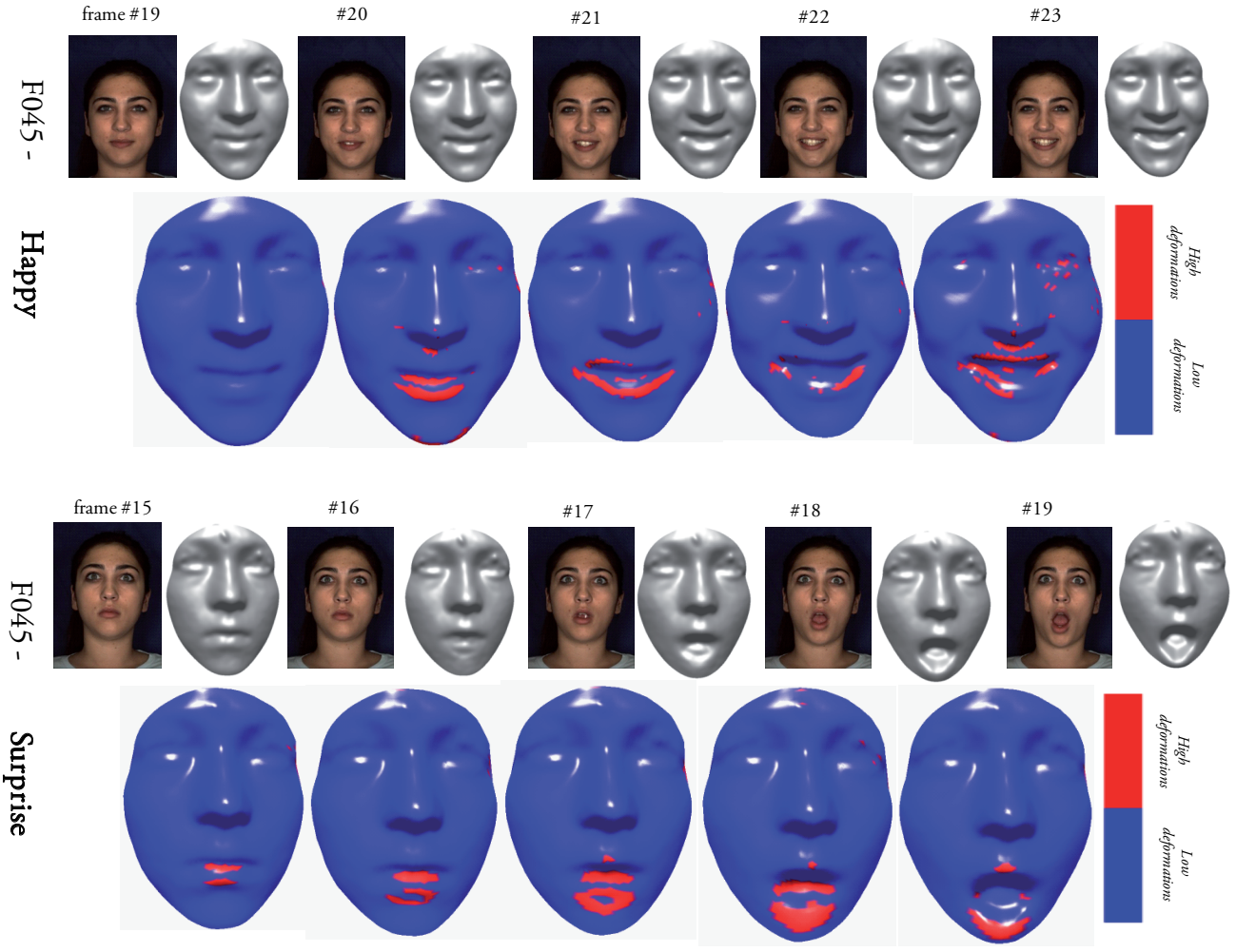


FIGURE 4.19: Examples of DSF deformations between subsequent frames of 3D video sequences: Happy and surprise expressions are shown, respectively, on the left and right.

the 3D scans of some frames are shown in the upper row. In the lower row, the deformation scalar field $DSF(F_{t-1}, F_t)$ computed between consecutive frames (i.e., the current frame and the previous one) in the subsequence is reported. In order to provide a visual representation of the scalar field, an automatic labeling scheme is applied that includes only two colors: The red color is associated with high $DSF(F_{t-1}, F_t)$ values and corresponds to facial regions affected by high deformations. The blue color is associated with regions that remain more stable from one frame to another. As illustrated in Fig. 4.19, for different expressions, different regions are mainly deformed, showing the capability of the deformation fields to capture relevant changes of the face due to the facial expression. In particular, each deformation is expected to identify an expression, for example, as suggested by the intuition, the corners of the mouth and the cheeks are mainly deformed for the happiness expression.

With the proposed approach, the feature extraction process starts by computing for each 3D frame in a given video sequence the Dense Scalar Field with respect to the previous one. In this way, we obtain as many fields as the number of frames in the sequence (decreased by one), where each field contains as

many scalar values as the number of points composing the collection of radial curves representing the facial surface. In practice, the size of $DSF(F_{t-1}, F_t)$ is 1×5000, considering 5000 points on the face, similarly to the feature vector used in the first scheme of classification (mean deformation-based). Since the dimensionality of the resulting feature vector is high, also in this case we use LDA to project the scalar values to a 5-dimensional feature space, which is sensitive to the deformations induced by different expressions. The 5-dimensional *feature vector* extracted for the 3D frame at instant t of a sequence is indicated as f^t in the following. Once extracted, the feature vectors are used to train HMMs and to learn deformations due to expressions along a temporal sequence of frames.

In our case, sequences of 3D frames constitute the temporal dynamics to be classified, and each prototypical expression is modeled by an HMM (a total of 6 HMMs λ^{expr} is required, with $expr \in \{an, di, fe, ha, sa, su\}$). Four states per HMM are used to represent the temporal behavior of each expression. This corresponds to the idea that each sequence starts and ends with a neutral expression (state S_1). The frames that belong to the temporal intervals where the face changes from neutral to expressive and back from expressive to neutral are modeled by the *onset* (S_2) and *offset* (S_4) states, respectively. Finally, the frames corresponding to the highest intensity of the expression are captured by the apex state (S_3). This solution has proved its effectiveness in clustering the expressive states of a sequence also in other works (194). Figure 4.20 exemplifies the structure of the HMMs used in our framework.

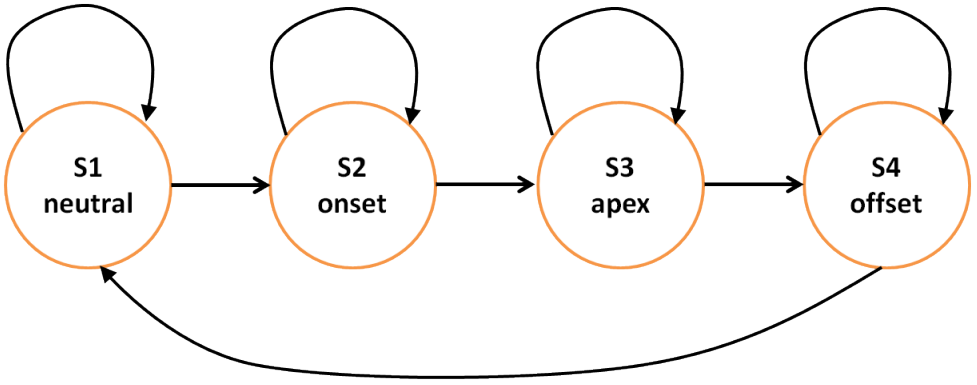


FIGURE 4.20: Structure of the HMMs modeling a 3D facial sequence. The four states model, respectively, the *neutral*, *onset*, *apex* and *offset* frames of the sequence. As shown, from each state it is possible to remain in the state itself or move to the next one (this is known as *Bakis* or left-right HMM).

The training procedure of each HMM is summarized as follows:

- Feature vectors f^t of the training sequences are first clustered to identify a *codebook* of symbols using the standard LBG algorithm (198). This provides the required mapping between multidimensional feature vectors, taking values in a continuous domain, with the alphabet of symbols emitted by the HMM states;
- Expression sequences are considered as observation sequences $O = \{O^1, O^2, \ldots, O^T\}$, where each observation O^t at time t is given by the feature vector f^t;

- Each HMM λ^{expr} is initialized with random values and the *Baum-Welch* algorithm (199) is used to train the model using a set of training sequences. This estimates the model parameters, while maximizing the conditional probability $P(O|\lambda^{expr})$.

Given a 3D sequence to be classified, it is processed as in Sect. 4.4.3, so that each feature vectors f^t corresponds to a *test* observation $O = \{O^1 \equiv f^1, \ldots, O^T \equiv f^T\}$. Then, the test observation O is presented to the six HMMs λ^{expr} that model different expressions, and the *Viterbi* algorithm is used to determine the best *path* $\bar{Q} = \{\bar{q}^1, \ldots, \bar{q}^T\}$, which corresponds to the state sequence giving a maximum of likelihood to the observation sequence O. The likelihood along the best path, $p^{expr}(O, \bar{Q}|\lambda^{expr}) = \bar{p}^{expr}(O|\lambda^{expr})$ is considered as a good approximation of the true likelihood given by the more expensive *forward* procedure (199), where all the possible paths are considered instead of the best one. Finally, the sequence is classified as belonging to the class corresponding to the HMM whose log-likelihood along the best path is the greatest one.

4.4.5 Experimental Results

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

4.4.5.1 BU-4DFE Database: Description and Preprocessing

To investigate the usability and performance of 3D dynamic facial sequences for facial expression recognition, a dynamic 3D facial expression database has been created at *Binghamton University* (175). The Dimensional Imaging's 3D dynamic capturing system (200), has been used to capture a sequence of stereo images and produce the depth map of the face according to a passive stereo-photogrammetry approach. The range maps are then combined to produce a temporally varying sequence of high-resolution 3D images with an RMS accuracy of 0.2mm. At the same time, 2D texture videos of the dynamic 3D models are also recorded. Each participant (subject) was requested to perform the six prototypical expressions (i.e., *angry*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*) separately. Each expression sequence contains neutral expressions in the beginning and the end, so that each expression was performed gradually from neutral appearance, low intensity, high intensity, and back to low intensity and neutral. Each 3D sequence captures one expression at a rate of 25 frames per second and each 3D sequence lasts approximately 4 seconds with about 35,000 vertices per scan (i.e., 3D *frame*). The database consists of 101 subjects (58 female and 43 male, with an age range of 18-45 years old) including 606 3D model sequences with 6 prototypical expressions and a variety of ethnic/racial ancestries (i.e., 28 Asian, 8 African-American, 3 Hispanic/Latino, and 62 Caucasian). More details on the BU-4DFE can be found in (175). An example of a 3D dynamic facial sequence of a subject with "happy" expression is shown in Fig. 4.21, where 2D

frames (not used in our solution) and 3D frames are reported. From left to right, the frames illustrate the intensity of facial expression passing from *neutral* to *onset*, *offset*, *apex* and *neutral* again.

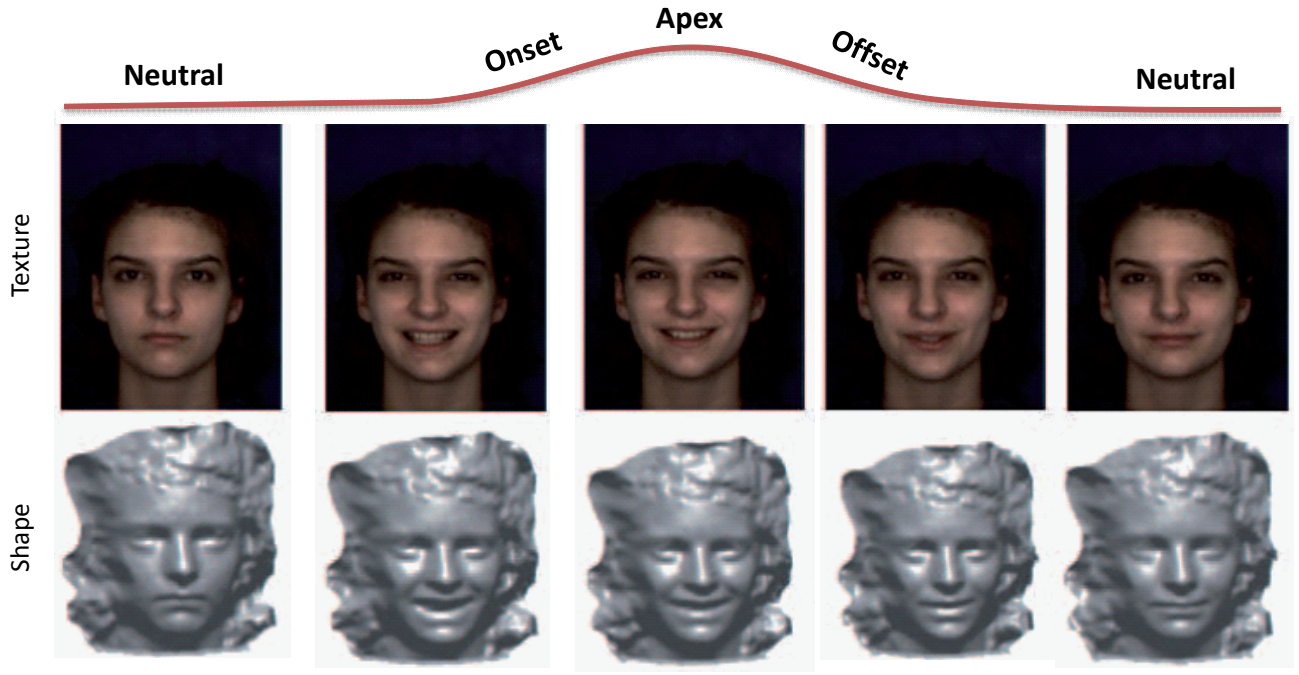


FIGURE 4.21: Examples of 2D and 3D frames extracted from a dynamic 3D video sequence of the BU-4DFE dataset.

It can be observed that the 3D frames present a near-frontal pose with some slight changes occurring mainly in the azimuthal plane. The scans are affected by large outliers, mainly acquired in the hair, neck and shoulders regions (see Fig. 4.21). In order to remove these imperfections from each 3D frame a preprocessing pipeline is performed. The main steps of this pipeline are summarized as follows (see also Fig. 4.22):

- Identify and fill the holes caused, in general, by self-occlusions or open mouth. The holes are identified by locating boundary edges, then triangulating them;
- Detect the nose tip on the face scan in the first frame of the sequence. The nose tip is detected by analyzing the peak point of the face scan in the depth direction. The nose tip is then tracked on all the subsequent frames when the search area is limited to a specific neighborhood around the nose tip detected on the first frame;
- Crop the facial area using a sphere centered on the detected nose tip with a constant radius set to 90mm based on some observations;
- Normalize the pose of a given frame according to its previous frame using the Iterative Closest Point (ICP)-based alignment. We point out that our implementation uses the following parameters to perform the ICP algorithm: (i) Match the nose tips of the faces first; (ii) Number of vertices

considered to find the optimal tranformation=50; and (iii) Number of iterations=5. In addition to permit effective alignment, this set of parameters allows also an attractive computational cost.

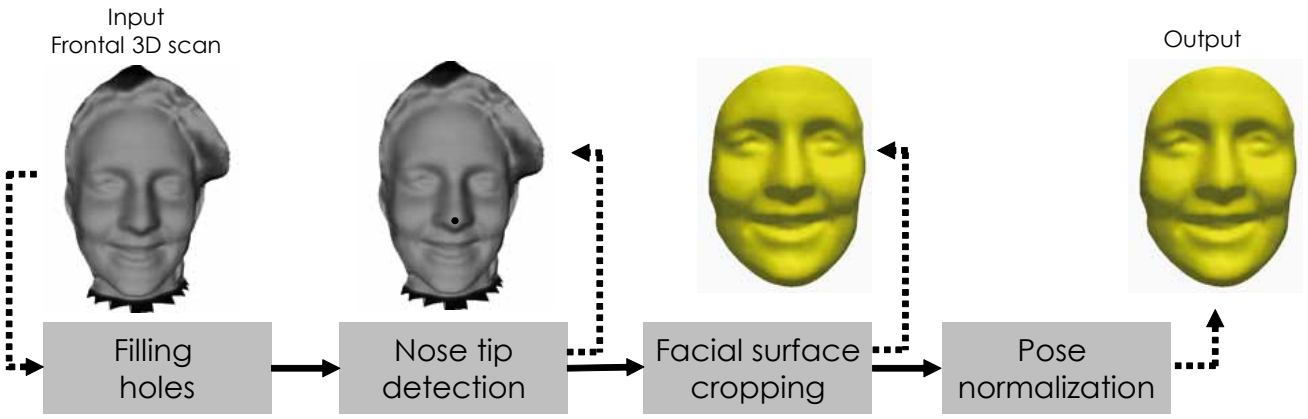


FIGURE 4.22: Outline of the preprocessing steps. A pipeline of four filters is applied to each 3D sequence: (i) Filling holes, if any; (ii) Nose tip detection (for the first frame) and tracking (for remaining frames); (iii) Face cropping using a sphere centered on the nose tip and of radius 90mm; (iv) Pose normalization based on the ICP algorithm.

In a real-world scenario of use, the head can move freely and rotate, whereas in our experiments only near-frontal faces are considered, as the BU-4DFE database does not contain non-frontal acquisitions. To account for the capability of our approach to cope with non-frontal scans, we report in Fig. 4.23 some registration results when applying an artificial rotation to one of the 3D faces to be aligned. It is clear that the registration method is able to handle with moderate pose variations (up to about 30/45 degrees). Instead, the registration method is not able to register a frontal face with a profile face (right side of the figure).

In the proposed framework, after preprocessing and Dense Scalar Fields computation across the 3D sequences, we designed two deformation learning and classification schemes. The first scheme consists on averaging, within a sliding window, the DSF computed on each frame with respect to the first frame of the window. This produces dense deformations across the sliding windows that are learned using a Multi-class Random Forest algorithm (see Sect. 4.4.4.1). The second scheme consists on a dynamic analysis through the 3D sequences using conventional temporal HMMs-modeling. Here, the 3D motion (deformation) is extracted and then learned for each class of expression, as described in Sect. 4.4.4.2. In both the cases, a common experimental set up has been used. In particular, data of 60 subjects have been selected to perform recognition experiments according to the evaluation protocol followed in other works (184, 187, 188). The 60 subjects have been randomly partitioned into 10 sets, each containing 6 subjects, and 10-fold cross validation has been used for training/test, where at each round 9 of the 10 folds (54 subjects) are used for the training, while the remaining fold (6 subjects) is used for the test. In the following, we report experimental evaluation and comparative analysis with respect to previous studies.

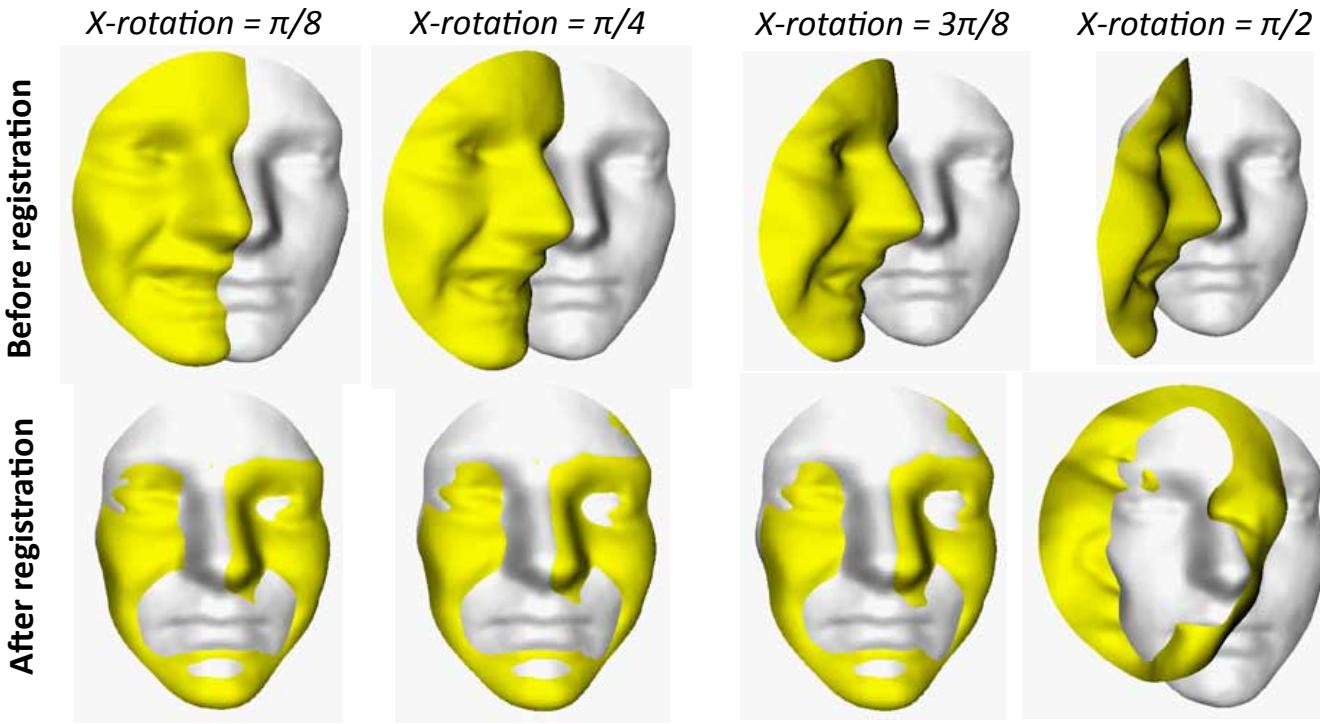


FIGURE 4.23: Registration results using the ICP algorithm when rotating around the X-axis one of the 3D preprocessed faces. The first row shows the initial rotation applied on the yellow model (before the alignment) and the second row shows the alignment output (after alignment).

4.4.5.2 Mean deformation-based Expression Classification

Following the experimental protocol proposed in (184), a large set of subsequences are extracted from the original expression sequences using a sliding window. The subsequences have been defined in (184) with a length of 6 frames with a sliding step of one frame from one subsequence to the following one. For example, with this approach, a sequence of 100 frames originates a set of $6 \times 95 = 570$ subsequences, each subsequence differing for one frame from the previous one. Each sequence is labelled to be one of the six basic expressions, thus extracted subsequences have the same label. This accounts for the fact that, in general, the subjects can enter the system not necessarily starting with a neutral expression, but with an arbitrary expression. The classification of these short sequences is regarded as an indication of the capability of the expression recognition framework to identify individual expressions. According to this, we first compute for each subsequence the Mean Deformation, which is then presented to multi-class Random Forest, as outlined in Sect. 4.4.4.

The performance of Random Forest classifier varies with the number of trees. Thus, we perform the experiments with different numbers of trees; the results of this experimentation is shown in Fig. 4.24. As illustrated in this figure, the average recognition rate raises with the increasing number of trees until 40, when the recognition rate reaches 93.21%, and then becomes quite stable. Thus, in the following we consider 40 trees and we report detailed results (confusion matrix) with this number of trees in Tab. 4.5.

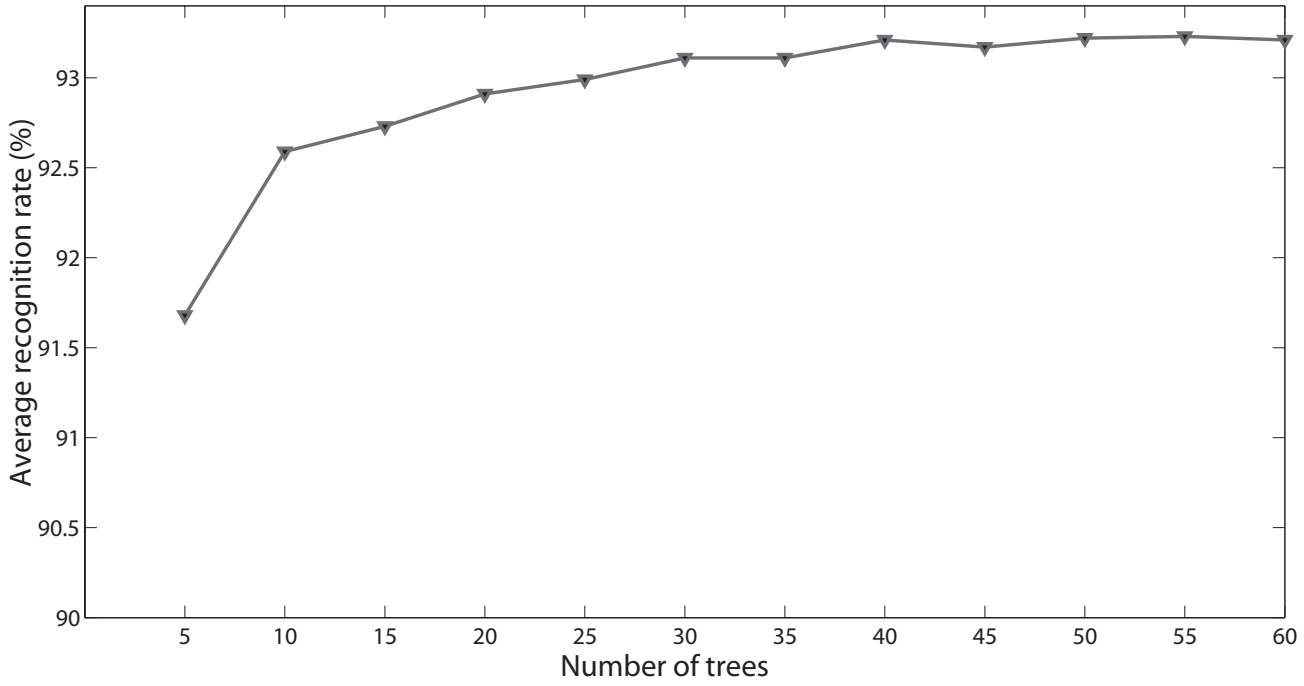


FIGURE 4.24: 4D expression recognition results using a Random Forest classifier when varying the number of trees.

We recall that the rates are obtained by averaging the results of the 10-independent runs (10-fold cross validation). It can be noted that the largest confusions are between the *disgust (Di)* expression and the *angry (An)* and *Fear (Fe)* expressions. Interestingly, these three expressions capture negative emotive states of the subjects, so that similar facial muscles can be activated. The best classified expressions are *happy (Ha)* and *Surprise (Su)* with recognition accuracy of 95.47% and 94.53%, respectively. The standard deviation from the average performance is also reported in the table. The value of this statistical indicator suggests that small variations are observed between different folds.

TABLE 4.5: Confusion matrix for Mean Deformation and Random Forest classifier (for 6-frames window).

%	*An*	*Di*	*Fe*	*Ha*	*Sa*	*Su*
An	**93.11**	2.42	1.71	0.46	1.61	0.66
Di	2.3	**92.46**	2.44	0.92	1.27	0.58
Fe	1.89	1.75	**91.24**	1.5	1.76	1.83
Ha	0.57	0.84	1.71	**95.47**	0.77	0.62
Sa	1.7	1.52	2.01	1.09	**92.46**	1.19
Su	0.71	0.85	1.84	0.72	1.33	**94.53**
Average recognition rate =**93.21±0.81%**						

Effect of the Subsequence Size

We have also conducted additional experiments when varying the temporal size of the sliding window used to define the subsequences. In Fig. 4.25, we report results for a window size equal to 2, 5 and 6, and using the whole length of the sequence (on average this is about 100 frames). From the figure, it clearly emerges that the recognition rate of the six expressions increases when increasing the temporal length of the window. This reveals the importance of the temporal dynamics and shows that the spatio-temporal analysis outperforms a spatial analysis of the frames. By considering the whole sequences for the classification, the result reach 100%. We decided to report detailed results when considering a window length of 6-frames to allow comparisons with previous studies.

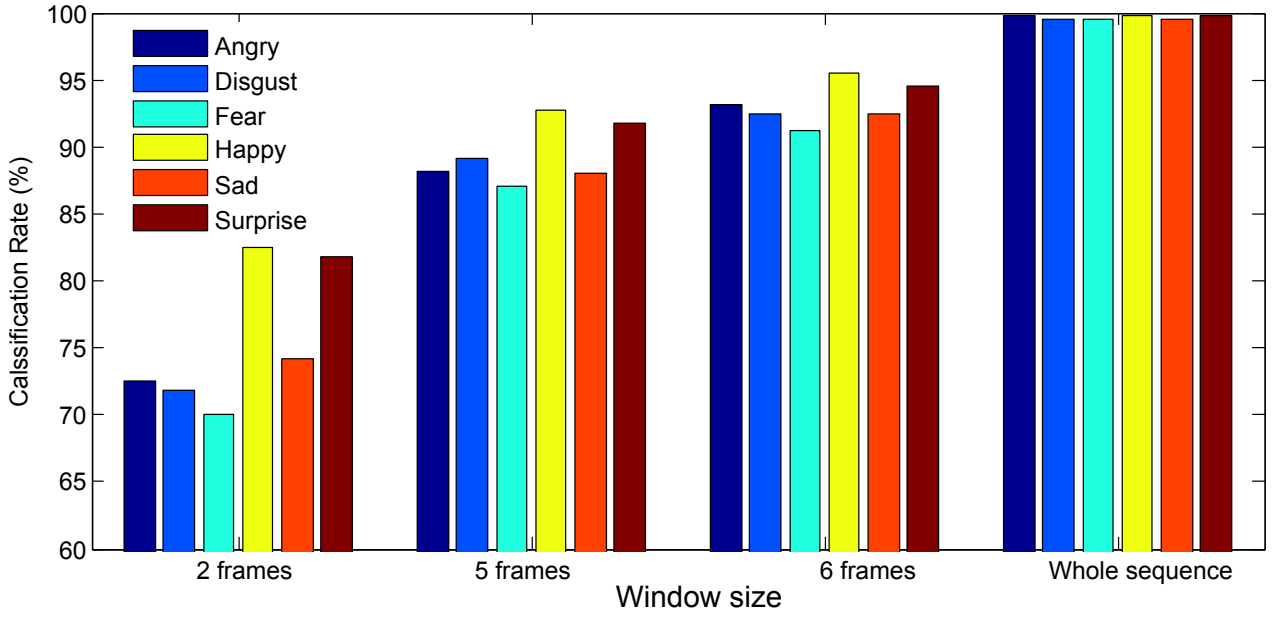


FIGURE 4.25: Effect of the temporal size of the sliding window on the results. The classification rates increase when increasing the length of the temporal window.

Effect of the Spatial Resolution of 3D Faces

In the proposed face representation, the DSF is computed for the points of a set of radial curves originating from the nose tip. Due to this, the density of the scalar field depends on the number of radial curves and the number of points per curve. So, the resolution used for the number of curves and points per curve can affect the final effectiveness of the representation. To investigated this aspect, we have conducted experiments when varying the spatial resolution of the 3D faces (i.e., the number of radial curves and the number of points per curve). Figure 4.26 expresses quantitatively the relationship between the expression classification accuracy (on the BU-4DFE) and the number of radial curves and the number of points per curve. This can give an indication of the expected decrease in the performance in the case the number of radial curves or points per curve is decreased due to the presence of noise and spikes in the data. From

these results, we can also observe that the resolution in terms of number of curves has more importance than the resolution in terms of points per curve.

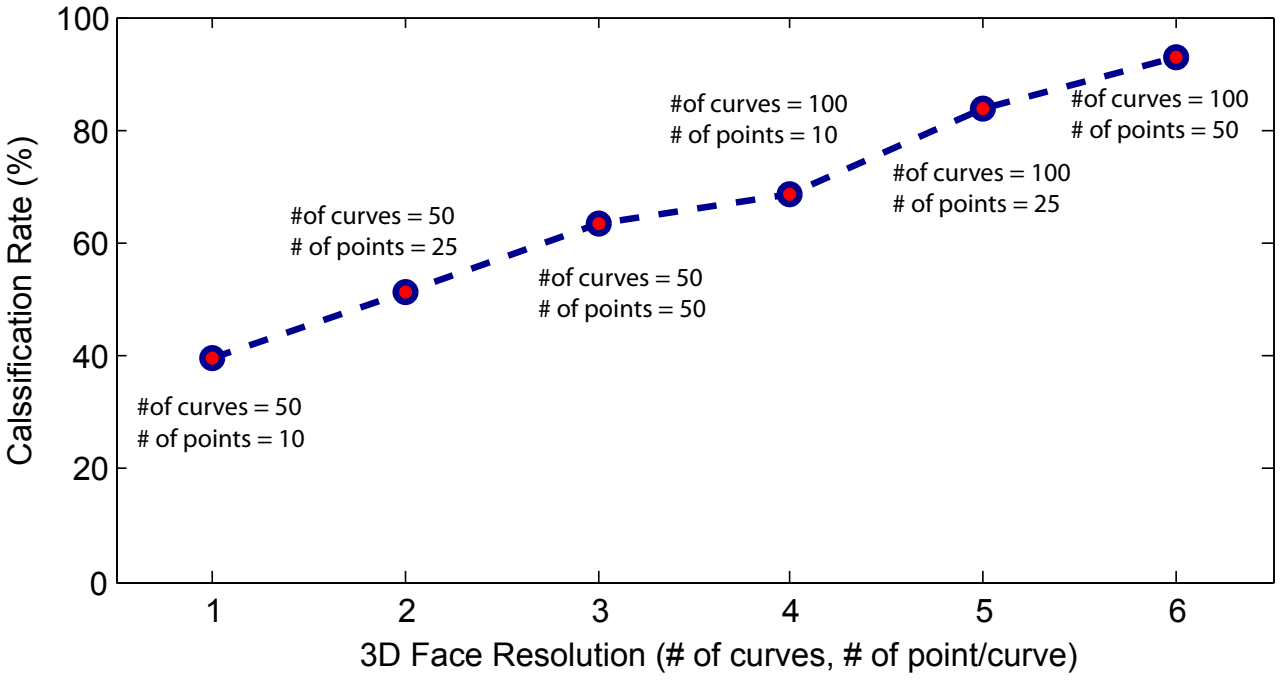


FIGURE 4.26: Effects of varying the 3D face resolution on the classification results.

4.4.5.3 HMM-based Expression Classification

Following the same setup as in previous section (originally defined in (184)), for this experiment we trained the HMMs on 6 frames subsequences constructed as discussed above. The 4-state structure of the HMMs resulted adequate to model the subsequences. Also in this experiment, we performed 10-folds cross validation on the overall set of subsequences derived from the 60×6 sequences (31970 in total). The achieved results by classifying individual subsequences of the expression sequences (*frame-by-frame* experiment) are reported in the confusion matrix of Tab. 4.6. Values in the table have been obtained by using features of 6-frames subsequences as input to the 6 HMMs and using the maximum emission probability criterion as decision rule. It is clear that the proposed approach is capable to accurately classify individual frames by analyzing the corresponding subsequence of previous 5 frames. The average recognition rate is equal to 93.83%, slightly higher than the one displayed by Mean Deformation plus Random Forest classification schema (though the standard deviation among different folds shows a greater value in this case). It can also be noted that, compared to the previous classifier, the same tendency of recognition rates is in general achieved. In fact, correct classification of *angry* is high despite the difficulty of this expression analysis. This learning scheme achieved better recognition than the first one for *angry (An)* expression. Actually, whereas the *angry (An)* expression is known for its subtle motions, our classifier achieved 93.95% of correct classification, which demonstrates the ability of the

proposed DSF to capture subtle deformations across the 3D sequences. These similar good achievements are mainly the effect of the proposed deformation scalar field.

TABLE 4.6: Confusion matrix for Motion extraction and HMM classifiers (for 6-frames window).

%	*An*	*Di*	*Fe*	*Ha*	*Sa*	*Su*
An	**93.95**	1.44	1.79	0.28	2.0	0.54
Di	3.10	**91.54**	3.40	0.54	1.27	0.15
Fe	1.05	1.42	**94.55**	0.69	1.67	0.62
Ha	0.51	0.93	1.65	**94.58**	1.93	0.40
Sa	1.77	0.48	1.99	0.32	**94.84**	0.60
Su	0.57	0.38	3.25	0.38	1.85	**93.57**
Average recognition rate = ***93.83±1.53%***						

Comparion with the FFD feature

The proposed framework can also fit with different deformation fields than the proposed DSF. So, considering alternative features to densely capture the deformation fields on the lattice of points of the radial curves of the face can permit a direct comparison of our DSF feature with different ones. In particular, we considered the Free-Form Deformation (FFD) (196) feature, which is a standard method for non-rigid registration and has been successfully proved in the context of expression recognition (186) (see also Sect. 4.4.3.B). Table 4.7 reports the confusion matrix obtained by posing FFD in our classification framework, using the same setting as in the experiments above (i.e., 100 radial curves, with 50 sampled points each, and LDA reduction of the deformation field from a 5000-dimensional vector to a 5-dimensional subspace). The overall result is that the proposed DSF feature provides a finer discriminative capability with respect to FFD, thus resulting in a better classification accuracy. This can be motivated by the nice invariant properties of the proposed Riemannian framework (as discussed in Sect. 4.4.3).

TABLE 4.7: Confusion matrix for Free-Form Deformation (FFD) and HMM classifiers (for 6-frames window).

%	*An*	*Di*	*Fe*	*Ha*	*Sa*	*Su*
An	**78.45**	4.51	5.72	1.97	6.49	2.86
Di	8.63	**76.1**	6.65	2.82	4.18	1.62
Fe	3.1	5.5	**80.23**	1.99	6.31	2.87
Ha	1.43	2.02	3.77	**86.12**	5.31	1.35
Sa	5.79	1.4	4.99	0.86	**85.83**	1.13
Su	1.73	2.04	6.13	1.55	3.9	**84.65**
Average recognition rate = $81.9 \pm 2.35\%$						

4.4.5.4 Discussion and Comparative Evaluation

To the best of our knowledge, the works reporting results on expression recognition from dynamic sequences of 3D scans are those in (181, 185, 187, 190), and recently (191). These works have been evaluated on the BU-4DFE dataset, but the testing protocols used in the experiments are sometimes different, so that a direct comparison of the results reported in these papers is not immediate. In the following, we discuss these solutions with respect to our proposal, also evidencing the different settings under which the expression recognition results are obtained.

TABLE 4.8: Comparison of this work to earlier studies. Protocol description: #subjects (S), #expressions (E), Win size (Win). T: temporal only/S-T: spatio-temporal. Accuracy on sliding window/whole sequence (or subsequence).

Authors	Method	Features	Classification	Protocol	T/S-T	RR (%)
Sun et al. (184)	MU-3D	12 Motion Units	HMM	60 S, 6 E, Win=6	T	70.31, —
Sun et al. (184)	T-HMM	Tracking model	HMM	60 S, 6 E, Win=6	T	80.04, —
Sun et al. (184)	P2D-HMM	Curvature+Tracking	T-HMM+S-HMM	60 S, 6 E, Win=6	S-T	82.19, —
Sun et al. (184)	R-2DHMM	Curvature+Tracking	2D-HMM	60 S, 6 E, Win=6	S-T	90.44, —
Sandbach et al. (186)	3D Motion-based	FFD+Quad-tree	GentleBoost+HMM	—, 3 E, Win=4	T	73.61, 81.93
Sandbach et al. (181)	3D Motion-based	FFD+Quad-tree	GentleBoost+HMM	—, 6 E, variable Win	T	64.6, —
Le et al. (187)	Level curve-based	segment-wise distances	HMM	60 S, 3 E, —	S-T	—, 92.22
Fang et al. (188, 190)	AFM Fitting	LBP-TOP	SVM-RBF	100 S, 6 E, —	T	—, 74.63
Fang et al. (188, 190)	AFM Fitting	LBP-TOP	SVM-RBF	100 S, 3 E, —	T	—, 96.71
Reale et al. (191)	Spatio-temporal volume	Nebula Feature	SVM-RBF	100 S, 6 E, Win=15	S-T	—, 76.10
This work	Geometric Motion	3D Motion	LDA-HMM	60 S, 6 E, Win=6	T	93.83, —
This work	Geometric Mean	Mean Deformation	LDA-Random Forest	60 S, 6 E, Win=6	T	93.21, —

Table 4.8 summarizes approaches and results reported previously on the BU-4DFE dataset, compared to those obtained in this work. The testing protocols used in the experiments are quite different especially the number of verified expressions, all the six basic expressions in (184, 185, 188, 190), and (191) whereas (186, 187) reported primary results on only three expressions. The number of subjects considered is 60, except in (186) where the number of subjects is not specified. In general, sequences in which the required expressions are acted accurately are selected, whereas in (188) and (190) 507 sequences out of the 606 total are used for all subjects. In our experiments, we conducted tests by following the same setting proposed by the earliest and more complete evaluation described in (184). The training and the testing sets were constructed by generating subsequences of 6-frames from all sequences of 60 selected subjects. The process were repeated by shifting the starting index of the sequence every one frame till the end of the sequence.

We note that the proposed approaches outperforms state-of-the-art solutions following similar experimental settings. The recognition rates reported in (184) and (185) based on temporal analysis only was 80.04% and spatio-temporal analysis was 90.44%. In both studies subsequences of constant window width including 6-frames ($win = 6$) is defined for experiments. We emphasize that their approach is not completely automatic requiring 83 manually annotated landmarks on the first frame of the sequence to allow accurate model tracking.

The method proposed in (186) and (181) is fully automatic with respect to the processing of facial frames in the temporal sequences, but uses *supervised* learning to annotate individual frames of the sequence in order to train a set of HMMs. Though performed off-line, supervised learning requires manual annotation and counting on a consistent number of training sequences that can be a time consuming operation. In addition, a drawback of this solution is the computational cost due to Free-Form Deformations based on B-spline interpolation between a lattice of control points for nonrigid registration and motion capturing between frames. Preliminary tests were reported on three expressions: *(An)*, *(Ha)* and *(Su)*. Authors motivated the choice of the happiness and anger expressions with the fact that they are at either ends of the valence expression spectrum, whereas surprise was also chosen as it is at one extreme of the arousal expression spectrum. However, these experiments were carried out on a subset of subjects accurately selected as acting out the required expression. Verification of the classification system was performed using a 10-fold cross-validation testing. On this subset of expressions and subjects, an average expression recognition rate of 81.93% is reported. In (181), the same authors have reported 64.6% classification rate when in their evaluation they consider all the six basic expressions.

In (187) a fully automatic method is also proposed, that uses an *unsupervised* learning solution to train a set of HMMs (i.e., annotation of individual frames is not required in this case). Expression recognition is performed on 60 subjects from the BU-4DFE database for the expressions of *happiness*, *sadness* and *surprise*. The recognition accuracy averaged on 10 rounds of 10-fold cross-validation show an overall value of 92.22%, with the highest performance of 95% obtained for the happiness expression. However, the authors reported recognition results on whole facial sequences, but this hinders the possibility of the methods to adhere to a real-time protocol. In fact, reported recognition results depends on the preprocessing of whole sequences unlike our approach and the one described in (184), which are able to provide recognition results when processing very few 3D frames.

In (188) and (190), results are presented for expression recognition accuracy on 100 subjects picked out from BU-4DFE database. However, 507 sequences are selected manually according to the following criteria: (1) the 4D sequence should start by neutral expression, and (2) sequences containing corrupted meshes are discarded. In addition, to achieve recognition rate of 75.82%, whole sequences should be analyzed. The authors reported highest recognition rates when only (Ha), (An), and (Su) expressions (96.71%) or (Ha), (Sa) and (Su) (95.75%) are considered.

The protocol used in (191) is quite different from the others. First, the onset frame for each of the six canonical expressions has been marked manually on each sequence of the BU-4DFE database. Then, a fixed size window of 15 frames starting from the onset frame has been extracted from each expression of 100 subjects. So, although sequences from 100 subjects are used by this approach, it also uses a manual intervention to detect the onset frame and just 15 frames from the onset one are used for the classification (and these typically correspond to the most intense expression, including the apex frames).

According to this comparative analysis, the proposed framework compares favorably with state-of-the-art solutions. It consists of two geometric deformation learning schemes with a common feature extraction

module (DSF). This demonstrates the effectiveness of the novel mathematical representation called Dense Scalar Field (DSF), under the two designed schemes.

4.5 Subtle Facial Motions for Effective 4D Expression Recognition

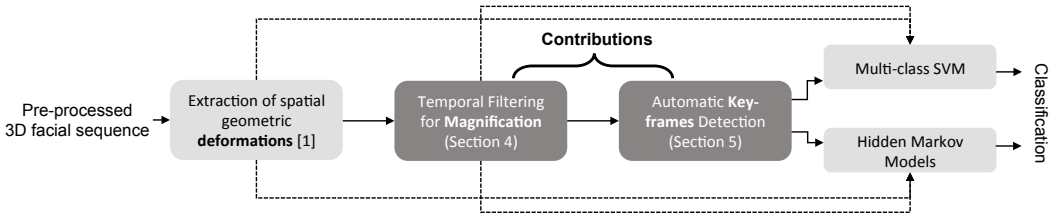


FIGURE 4.27: Overview of the proposed pipeline for effective 4D FER. From left-to-right : Extraction of geometric (deformation) features termed DSFs (201) – **Magnification of facial motions** – **Key-frames (*i.e.* Onset-Apex-Offset) detection** – Classification technique : SVM or HMM. The following schemes are studied later on (i) DSFs (ii) Magnified DSFs and (iii) Magnified DSFs on Key frames.

Even though the performance of *FER* has been substantially boosted by 4D data in recent years, there still exist some unsolved problems. On the one hand, some reputed similar expressions are difficult to distinguish since the facial deformations are sometimes really slight (202). On the other hand, the detection of keyframes which include much expressions information in the 3D face video is not paid sufficient attention. This contribution is not detailed in the manuscript, please refer to (122) for more details.

We present a novel and effective pipeline to automatic 4D *FER*. The main difficulty is to establish a vertex-level dense correspondence between the frames. This point has been solved in the previous step of Dense Scalar Fields (*DSFs*) computation (201) as an accurate registration of neighboring faces is achieved trough an approximation by elastic radial curves as presented in previous section. Then, based on an Eulerian spatio-temporal processing the facial motions, in particular the subtle ones are magnified. Due to such amplification, the deformations of certain expressions with low intensities are amplified to improve the classification accuracy.

We represent 3D facial surfaces by collections of radial curves emanating from the tip of the nose inorder to extract the DSF geometric feature propsoed in section 4.2. In the pre-processing step, the 3D mesh in each frame is first aligned to the first one and then cropped. The facial surfaces are then approximated by indexed collections of radial curves β_α, where the index α denotes the angle formed by the curve with respect to a reference radial curve. These curves are then uniformally resampled.

Let χ reveals the shape difference of two facial surfaces by deforming one mesh into another through an accurate registration step. We propose to adapt the Eulerian spatio-temporal processing (203) to the 3D domain. This method and its application to 3D face videos are presented in the subsequent. The Eulerian spatio-temporal processing was introduced for motion magnification in 2D videos, and has proved its effectiveness in (203). Its basic idea is to amplify the variation of pixel values over time, in

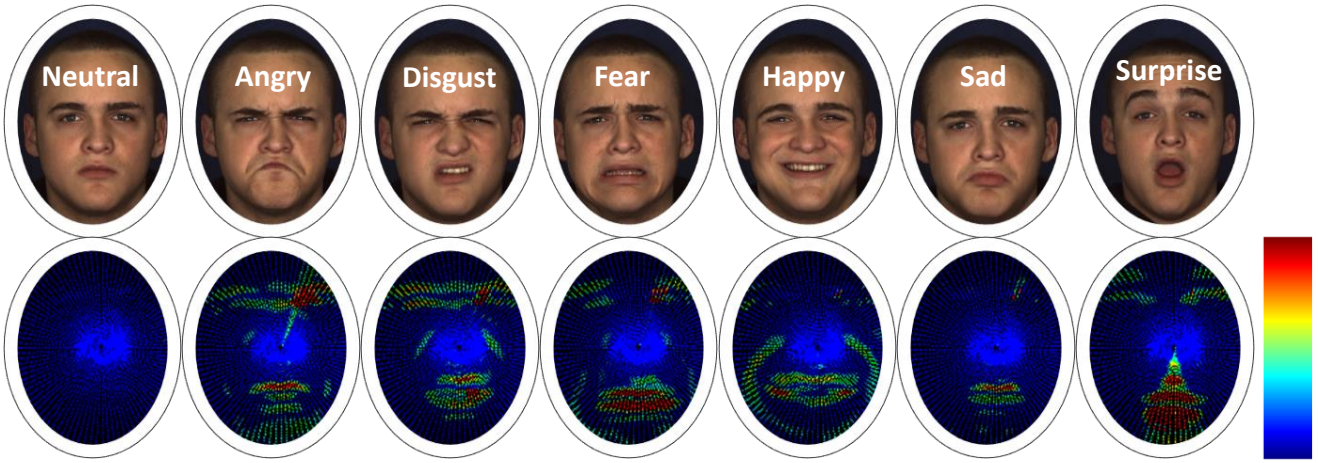
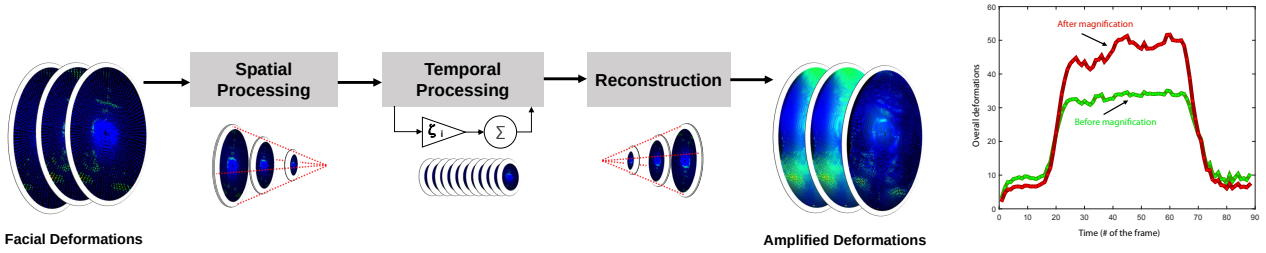


FIGURE 4.28: Top row: facial texture images of an individual with different expressions. Bottom row: facial deformations in Riemannian space. Where, warm colors are associated to the high χ and correspond to facial regions with high deformations, cold colors reflect the most static parts of the 3D face.



(a) (b)

FIGURE 4.29: (a) Overview of 3D video magnification. The original facial deformation features are first decomposed into different spatial frequencies, and the temporal filter is applied to all the frequency bands. The filtered spatial bands are then amplified by a given factor ζ, added back to the original signal, and collapsed to the output sequence. (b) An example of facial expression deformation (norm of the velocity vector) before (green) and after (red) magnification.

a spatially-multiscale manner, without explicitly estimating motion but rather exaggerating motion by amplifying temporal color changes at fixed positions. It relies on a linear approximation related to the brightness constancy assumption that forms the basis of the optical flow algorithm. However, the case is not that straightforward in 3D, because the vertex correspondence across frames cannot be achieved as easy as that in 2D. Fortunately, during the computation of χ, such correspondence is established by surface registration and remeshing. We can thus adapt Eulerian spatio-temporal processing to 3D face video. We take into account the values of the time series χ at any spatial location and highlight the differences in a given temporal frequency band of interest. It thus combines spatial and temporal processing to emphasize subtle changes in a 3D face video.

The process is illustrated in Fig. 4.29(a). Specifically, the video sequences are first decomposed into different spatial frequency bands by Gaussian pyramid, and these bands might be magnified differently. We consider that the time series correspond to the values of χ on the mesh surfaces in a frequency band and apply a band pass filter to extract the frequency bands of interest. The temporal processing, $\mathfrak{T}$, is

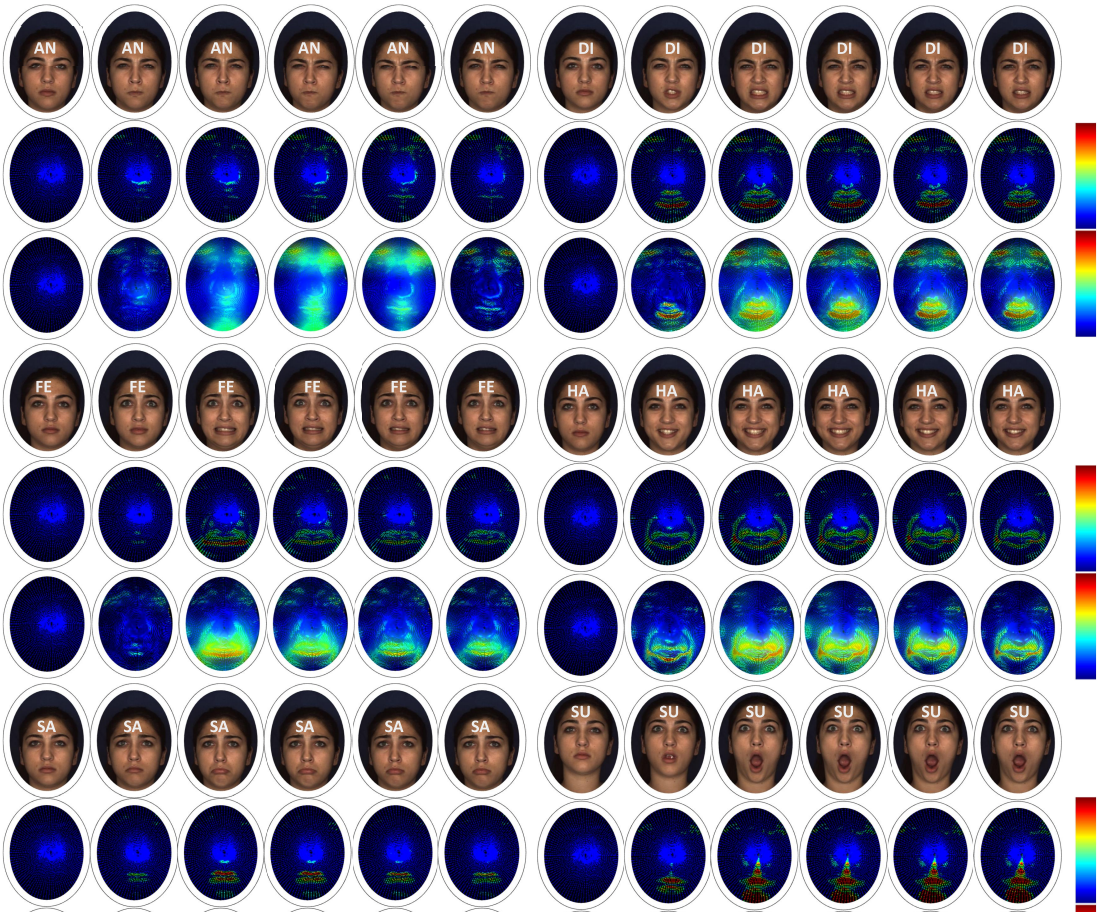


FIGURE 4.30: Illustrations of the deformation magnification on sequences of the same subject performing the three expressions: (a) sad, (b) angry, and (c) fear. One can appreciate the magnification effects on 3D deformations (third row) compared to those of the original facial deformation features (*DSFs*) (second row). Texture images corresponding to each of the 3D face models are also displayed.

uniform for all spatial levels, and for all χ within each level. We then multiply the extracted band passed signal by a magnification factor ζ, and add the magnified signal to the original and collapse the spatial pyramid to obtain the final output.

Qualitative results are shown in Fig. 4.30 where the deformations are amplified compared to the original features. The color-maps obtained after magnification reflect high amplitude of deformation areas which are not visible in the original features color-maps.

4.5.1 Experimental Results

This section describes the experiments conducted to validate the proposed method for 4D *FER*. A brief description of the BU-4DFE and BP4D datasets is first presented followed by the experimental setting and obtained classification accuracy, when including different steps of the pipeline presented in Fig. 4.27. A comparative study with existing methods is then presented.

4.5.1.1 Dataset Description and Experimental Settings

We conduct the results on the BU-4D dataset, this dataset has been presented in the previous section. Our experiments are conducted on the following sub-pipelines – (1) the whole video sequence (denoted by *WV*), (2) the magnified whole video sequence (denoted as *MWV*), (3) the key frames of the video sequence (denoted by *KFV*), and finally (4) the magnified key-frames of the video sequence (*i.e. MKFV*).

As described in (201), $\bar{\chi} = \frac{1}{n}\sum_{t=1}^{n}\chi(t)$ summarizes the overall deformation of the facial surface, where n denotes the video length. Note a major difference with the work presented in (201) is that *DSFs* are computed between successive frames, however, the χ are quantified between the current frame and a reference frame in this approach. In addition, instead of using a Random Forest classifier, a multi-class Support Vector Machine (*SVM*) is considered here where $\bar{\chi}$ is treated as a feature vector to predict the video label in the Subtle Deformation Magnification experiment (using *SVM Classifier*). We also adopt *HMM* to encode the temporal behavior of each expression, and get the expression type. To allow fair comparison with previous studies, we randomly select 60 subjects from the BU-4DFE dataset to perform our experiments under a 10-fold cross-validation protocol. BP4D is another dataset in this domain and is composed of 3D face videos belonging to 41 individuals, who are asked to perform 8 tasks, corresponding to 8 expressions. Besides the 6 universal ones, there are two additional expressions, namely embarrassment and pain. The data in BP4D are very similar to those in BU-4DFE; however, the facial expressions are elicited by a series of activities, including film watching, interviews, experiencing a cold pressor test, etc., and are thus spontaneous. The experiment is carried out in a cross-dataset way, i.e., training on BU-4DFE and testing on BP4D, according to the protocol used in (204). The parameters are the same as the ones in the previous experiments on BU-4DFE.

In literature, a number of studies report 4D *FER* results on the BU-4DFE database, however they differ in their experimental settings. In this section, we compare our results with existing approaches when considering these differences.

The state of the art results reported on the BU-4DFE dataset are demonstrated in Table 4.9. In this table, #E means the number of expressions, #S is the number of subjects, #-CV provides the number of cross-validation used, *Full Seq./Win* means the decision is made based on the analysis of the full sequence or on sub-sequences captured using a sliding window. The studies (205, 206) report one of the highest accuracy when using a sliding window of 6 frames, nevertheless, the approach requires manual annotation of 83 landmarks on the first frame. Moreover, the vertex-level dense tracking scheme is time consuming. In a more recent work from the same group developed by Reale *et al.* (207), the authors propose 4D (*Space-Time*) features termed *Nebula* computed on a fixed-size window of 15 frames. The best accuracy reported is 76.9% using sequences of 100 subjects, when the 3D video segmentation to limit the expressive time interval is performed manually. In (209), Fang *et al.* obtained an accuracy of 74.63% with 507 sequences of 100 subjects. Le *et al.* (210) evaluate their algorithm consisting of level curves and *HMMs* on 60 subjects sequences on three expressions (*HA*, *SA* and *SU*) and display an accuracy of 92.22%. It should be noted

TABLE 4.9: A comparative study of the proposed approach with the state-of-the-art on BU-4DFE.

Method	**Experimental Settings**	**Accuracy**
Sun *et al.* (205)	6E, 60S, 10-CV, Win=6	90.44%
Sun *et al.* (206)	6E, 60S, 10-CV, Win=6	94.37%
Reale *et al.* (207)	6E, 100S, –, Win=15	76.9%
Sandb. *et al.* (208)	6E, 60S, 6-CV, Win	64.6%
Fang *et al.* (209)	6E, 100S, 10-CV, –	74.63%
Le *et al.* (210)	3E, 60S, 10-CV, Full seq.	92.22%
Xue *et al.* (211)	6E, 60S, 10-CV, Full seq.	78.8%
Berretti *et al.* (212)	6E, 60S, 10-CV, Full seq.	79.4%
Berretti *et al.* (212)	6E, 60S, 10-CV, Win=6	72.25%
Ben Amor *et al.* (201)	6E, 60S, 10-CV, Full seq.	93.21%
Ben Amor *et al.* (201)	6E, 60S, 10-CV, Win=6.	93.83%
This work – *SVM* on $\bar{\chi}$	6E, 60S, 10-CV, keyframes	**94.46**%
This work – *HMM* on $\chi(t)$	6E, 60S, 10-CV, keyframes	**95.13**%

that the proposed approach is evaluated when considering full sequences which is a major difference to the works (205–208). It is pointed out in (212) that the problem of the window-based evaluation protocol is to label all sub-sequences from the neutral intervals as one of the six expressions which can bias the final result. For that reason, we are interested in classifying the sequences which are performed during 2-3 seconds as did in (210–212). In comparison with these approaches, our approach does not require any manual or automatic landmarking. The computational complexity of the proposed approach is lower than the *FFD* used in (208) and the pipeline for 4D matching/registration (*i.e. Spin Images*, *MeshHOG*, *RANSAC* and *AFM*) and feature extraction (*LBP-TOP*) proposed in (209). Compared to the results listed in Table 4.9, the proposed approach outperforms existing approaches, where (1) no landmark detection is required; (2) no dimensionality reduction or feature selection techniques are applied; and (3) A vertex-level 4D dense registration and quantification of the deformations are led jointly through a Riemannian approach. The temporal filtering amplifies these deformations and consequently reveals hidden (subtle) 4D facial motions.

In view of the discussion made above, the main contribution which lie in the temporal magnification of the extracted geometric features, allows revealing subtle deformations and thus a more distinguishable power of the six universal expressions. An improvement of more than 10% in the classification rate is achieved. In addition, the introduction of an automatic key-frames detection (onset-apex-offset frames) permits to avoid exhaustive analysis by locating the most relevant frames among the 3D sequence. Results achieved using only the key-frames are comparable to those achieved on full 3D videos. The main limitation of our approach is its dependence to a reference frame which, in fact the computation of *DSFs* are based on this reference frame, taken to be a quasi-neutral 3D face from the sequence.

4.5.1.2 Cross-dataset Evaluation on BP4D

To validate the generalization ability of the proposed method, we launch a cross-dataset experiment on BP4D according to (204). BU-4DFE is employed for training and a subset of BP4D (i.e. Task 1 and Task 8, consisting of happy and disgust faces) for testing. It is actually a ternary classification problem on the samples of happy, disgust and neutral expressions. For computation simplicity, SVM is adopted as the classifier.

TABLE 4.10: Cross-dataset evaluation on the BP4D database.

Method	Training in BU-4DFE	Testing in BP4D	Accuracy (%)
Zhang et al. (204)	Happy, Disgust, Neutral	Task 1 and Task 8	71.0
The proposed work (before magnification)	Happy, Disgust, Neutral	Task 1 and Task 8	75.6
The proposed work (after magnification)	Happy, Disgust, Neutral	Task 1 and Task 8	*81.7*

Table 4.10 demonstrates the results with and without the magnification step of the proposed method on the BP4D database, where similar conclusions can be drawn as in BU-4DFE. Compared with the performance in (204), the score on the original DSF features are better. When the deformations are further enhanced, the accuracy is improved to 81.7%, delivering a 10% result gain. It indicates that our method has the potential to be generalized to spontaneous situations.

4.6 Conclusions

In this chapter we investigated the 3D face analysis using shape analysis of facial curves. Thus, the 3D facial surface is parameterized by a collection of radial curves and a novel Deformation Sector Field (DSF) which accurately describes local deformations between 3D facial surfaces has been proposed. A facial surface parameterization by their radial curves allows the definition of this descriptor on each facial point based on Riemanniann Geometry. This feature has been used to capture the deformation between a given 3D face and a reference one in order to extract 3D facial averageness/symmetry difference that were investigated as high-level features for gender classification and age estimation. By investigating the age estimation separately on Female and Male subsets, we have achieved better age estimation results, which justifies that the general aging effect of face differs considerably with gender. Moreover, the proposed geometric feature (DSF) has been successfully used to describes local deformations across 3D facial sequences for facial expression recognition. The deformations obtained are then amplified by using the temporal filter over the 3D face video. The combination of these two ideas performs accurate vertex-level registration of 4D faces and highlights hidden shape variations in 3D face video sequences. Furthermore, we present an automatic technique to localize relevant frames (onset-apex-offset frames) in the video based on clustering facial shape deformation manifolds. Through comprehensive experiments, we demonstrate the contribution of such joint spatio-temporal analysis in recognizing facial expressions

from 4D data. The proposed approach outperforms the existing ones on the BU-4DFE dataset and shows good potential to be generalized to spontaneous scenarios as in BP4D.

Chapter 5

Towards Statistical Analysis of Surfaces

The main results presented in this chapter have been published in the following journal papers: Computers & Graphics (2015) (213), IEEE PAMI (2016) (214).

5.1 Introduction

In this chapter we seek a framework for analysing shapes of a certain class of 3D objects. Although the general goal in shape analysis is to develop tools for full statistical analysis – statistical averaging, finding principal modes of variations in a population, and shape classification, we restrict to more basic goals of quantifying shape differences and generating deformations. While there have been many efforts in shape analysis of 3D objects, the problem is far from solved and the current solutions face many technical and practical issues. For instance, many general techniques for shape analysis rely on quantifying shape differences by spatially matching geometric features across objects. Therefore, it becomes important to establish a correspondence of parts between objects, i.e. which part in one object corresponds to which part in the other? This was an important bottleneck in a majority of previous efforts on 3D shape analysis where the correspondence (or registration) of objects was either presumed or solved as an independent pre-processing step. More recently, there has been progress in establishing frameworks that formulate the registration and comparison problems jointly. These newer frameworks, using techniques from differential geometry, focus on shape analysis of parameterized surfaces and treat the problem of shape comparison as the problem of computing geodesic paths in shape spaces under a chosen metric. Here shapes are compared using a Riemannian metric on a *pre-shape space* $\mathcal{F}$ consisting of embeddings or immersions of a model manifold (like the sphere, or the disc) into the 3D Euclidean space $\mathbb{R}^3$. Two embeddings correspond to the same shape in $\mathbb{R}^3$ if and only if they differ by an element of a shape-preserving transformation group, such as rigid motion, scaling, and reparameterization. The shape space is therefore the quotient space of the pre-shape space by these shape-preserving groups. If the Riemannian metric on the pre-shape space is preserved by the action of the shape-preserving group

then it induces a Riemannian metric on the quotient space. The construction of geodesics in shape space provide optimal deformations between surfaces and is a very important tool in *statistical* analysis of shapes. Interestingly, the problem of registration is handled using parameterizations of surfaces such that the points denoting the same parameter values on two objects are considered registered.

While these geometric ideas are powerful and comprehensive, there are two important issues that one needs to deal with: (1) the choice of Riemannian metric to define geodesics, geodesic lengths, and the eventual shape metric, and (2) the task of computing geodesic paths between arbitrary shapes. In terms of the first issue, the choice of a metric, an important requirement is that the metric should be invariant to action of the reparameterization group, to enable a well-defined distance on the eventual quotient space or the shape space of surfaces. There is a related requirement for the shape analysis to be invariant to parameterizations of objects since parameterizations are only artificial impositions designed to help navigate along objects. The physical intuition we have is that shape tools, such as the deformation (path or geodesic) from one shape to another, are physical processes that are independent of the way surfaces may be parameterized. These dual requirements rule out the use of commonly-used quantities such as the $\mathbb{L}^2$ norm on the space $\mathcal{F}$ directly. In terms of the second issue, the lack of standard metrics makes it complicated to compute geodesic paths even when the underlying manifold is a vector space, and one needs numerical algorithms for approximating geodesic paths.

In this chapter, we propose two main contributions to handle the discussed challenges. First, we propose a framework to model the shape of the 3D face. The key idea to to represent the 3D face directly as an element of a manifold (and not passing through curve manifold) and perform geodesic and statistical calculus on the quotient space obtained after filtering the scale, rotation, re-parameterization and translation. Thus, we adapt a recent elastic shape analysis framework (22, 23) to the case of hemispherical surfaces, and explore its use in a number of 3D face processing applications including face deformation, template computation, summarization of variability in different expression classes, random generation of 3D faces from a Gaussian-type generative model, and symmetry analysis. The framework in (22, 23) was previously defined for quadrilateral and spherical surfaces, and was later extended to the case of cylindrical surfaces in (23). All of the considered tasks are performed under a unified Riemannian metric allowing principled definition of various tools including registration via surface reparameterization, deformation and symmetry analysis using geodesic paths, intrinsic shape averaging, principal component analysis, and definition of generative shape models.
The second main contribution presented in this chapter is a novel framework proposed for spherical surfaces. The novelty lies in defining the Riemannian metric directly on the quotient (shape) space, rather than inheriting it from pre-shape space (as proposed il all proposed frameworks in this habilitation), and using it to formulate a path energy that measures only the normal components of velocities along the path. In other words, we define and solve for geodesics directly on the shape space and avoid complications resulting from the quotient operation. This comprehensive framework is invariant to arbitrary parameterizations of surfaces along paths, a phenomenon termed as gauge invariance.

5.2 Statistical Analysis of 3D Faces

In this section, we adapt a recent elastic shape analysis framework (22) to the case of hemispherical surfaces, and explore its use in a number of 3D face processing applications. This framework was previously defined for quadrilateral, spherical and cylindrical surfaces. All of the considered tasks are performed under an elastic Riemannian metric allowing principled definition of various tools including registration via surface re-parameterization, deformation and symmetry analysis using geodesic paths, intrinsic shape averaging, principal component analysis, and definition of generative shape models. Thus, the main contributions of this work are:

(1) We extend the framework of Jermyn et al. (22) to allow statistical shape analysis of hemispherical surfaces.

(2) We consider the task of 3D face morphing using a parameterized surface representation and a proper, parameterization-invariant elastic Riemannian metric. This provides the formalism for defining optimal correspondences and deformations between facial surfaces via geodesic paths.

(3) We define a comprehensive statistical framework for modeling of 3D faces. The definition of a proper Riemannian metric allows us to compute intrinsic facial shape averages as well as covariances to study facial shape variability in different expression classes. Using these estimates, one can form a generative 3D face model that can be used for random sampling.

(4) We provide tools for symmetry analysis of 3D faces.

(5) We study expression and identity classification under the defined metric. We compare our performance to the state-of-the-art method in (215). The main idea behind presenting this application is to showcase the benefits of an elastic framework in the recognition task. We leave a more thorough study of classification performance and comparisons to other state-of-the-art methods as future work.

5.2.1 Mathematical Framework

In this section, we describe the main ingredients in defining a comprehensive, elastic shape analysis framework for facial surfaces. We note that these methods have been previously described for the case of quadrilateral, spherical and cylindrical surfaces in (22, 23). We extend these methods to hemispherical surfaces and apply them to statistical shape analysis of 3D faces. Let $\mathcal{F}$ be the space of all smooth embeddings of a closed unit disk in $\mathbb{R}^3$, where each such embedding defines a parameterized surface $f : \bar{\mathbb{D}} \to \mathbb{R}^3$. Let Γ be the set of all boundary-preserving diffeomorphisms of $\bar{\mathbb{D}}$. For a facial surface $f \in \mathcal{F}$, $f \circ \gamma$ represents its re-parameterization. In other words, γ is a warping of the coordinate system on f. As previously shown in (22), it is inappropriate to use the $\mathbb{L}^2$ metric for analyzing shapes of parameterized surfaces, because Γ does not act on $\mathcal{F}$ by isometries. Thus, we utilize the square-root

normal field (SRNF) representation of surfaces and the corresponding Riemannian metric proposed in (22).

Let $s = (u, v) \in \bar{\mathbb{D}}$ define a polar coordinate system on the closed unit disk. The SRNF representation of facial surfaces is then defined using a mapping $Q : \mathcal{F} \to \mathbb{L}^2$ as $Q(f)(s) = \frac{n(s)}{|n(s)|^{1/2}}$. Here, $n(s) = \frac{\partial f}{\partial u}(s) \times \frac{\partial f}{\partial v}(s)$ denotes a normal vector to the surface f at the point $f(s)$. The space of all SRNFs is a subset of $\mathbb{L}^2(\bar{\mathbb{D}}, \mathbb{R}^3)$, henceforth referred to simply as $\mathbb{L}^2$, and it is endowed with the natural $\mathbb{L}^2$ metric. The differential of Q is a smooth mapping between tangent spaces, $Q_{*,f} : T_f(\mathcal{F}) \to T_{Q(f)}(\mathbb{L}^2)$, and is used to define the corresponding Riemannian metric on $\mathcal{F}$ as $\langle\langle w_1, w_2 \rangle\rangle_f = \langle Q_{*,f}(w_1), Q_{*,f}(w_2) \rangle_{\mathbb{L}^2}$, where $w_1,\ w_2 \in T_f(\mathcal{F})$ (23). Using this expression, one can verify that the re-parameterization group Γ acts on $\mathcal{F}$ by isometries, i.e. $\langle\langle w_1 \circ \gamma, w_2 \circ \gamma \rangle\rangle_{f\circ\gamma} = \langle\langle w_1, w_2 \rangle\rangle_f$. Another advantage of this metric is that it has a natural interpretation in terms of the amount of stretching and bending needed to deform one surface into another. For this reason, it has been referred to as the partial elastic metric (22). Furthermore, this metric is automatically invariant to translation. Scaling variability can be removed by rescaling all surfaces to have unit area. We let $\mathcal{C}$ denote the space of all unit area surfaces (pre-shape space).

Rotation and re-parameterization variability is removed from the representation space using equivalence classes. Let $q = Q(f)$ denote the SRNF of a facial surface f. A rotation of f by $O \in SO(3)$, Of, results in a rotation of its SRNF representation, Oq. A re-parameterization of f by $\gamma \in \Gamma$, $f \circ \gamma$, results in the following transformation of its SRNF: $(q, \gamma) = (q \circ \gamma)\sqrt{J_\gamma}$, where J_γ is the determinant of the Jacobian of γ. Now, one can define two types of equivalence classes, $[f] = \{O(f \circ \gamma) | O \in SO(3), \gamma \in \Gamma\}$ in $\mathcal{C}$ endowed with the metric $\langle\langle \cdot, \cdot \rangle\rangle$ or $[q] = \{O(q, \gamma) | O \in SO(3), \gamma \in \Gamma\}$ in $\mathbb{L}^2$ endowed with the $\mathbb{L}^2$ metric; each equivalence class represents a shape uniquely in its respective representation space. This results in two strategies to account for the rotation and re-parameterization variabilities in 3D face data. Given two surfaces $f_1, f_2 \in \mathcal{C}$, the exact solution comes from the following optimization problem: $(O^*, \gamma^*) = \text{arginf}_{(O,\gamma)\in SO(3)\times\Gamma}\, d_{\mathcal{C}}(f_1, O(f_2 \circ \gamma))$. Unfortunately, there is no closed form expression for the geodesic distance $d_{\mathcal{C}}$ because of the complex structure of the Riemannian metric $\langle\langle \cdot, \cdot \rangle\rangle$. There is a numerical approach, termed path-straightening, which can be used to compute this geodesic distance, but it is computationally expensive. Thus, we use an approximate solution to the registration problem in our analysis, which can be computed using the SRNF representation as $(O^*, \gamma^*) = \text{arginf}_{(O,\gamma)\in SO(3)\times\Gamma} \|q_1 - (Oq_2, \gamma)\|$. This problem is much easier to solve and provides a very close approximation to the original problem, because the partial elastic metric on $\mathcal{C}$ is the pullback of the $\mathbb{L}^2$ metric from the SRNF space.

The optimization problem over $SO(3) \times \Gamma$ is solved iteratively using the general procedure presented in (22, 23). First, one fixes γ and searches for an optimal rotation over $SO(3)$ using Procrustes analysis; this is performed in one step using singular value decomposition. Then, given the computed rotation, one searches for an optimal re-parameterization in Γ using a gradient descent algorithm. This search is initialized using the identity element γ_{id}. One could consider more elaborate initialization schemes but we found that this is not needed in this application. The search over Γ requires the specification of an orthonormal basis for $T_{\gamma_{id}}(\Gamma)$ (space of smooth, boundary-preserving vector fields on $\bar{\mathbb{D}}$). In other

words, we seek a basis of smooth vector fields that map the closed unit disk to itself. In order to define this basis, we make a small simplification. Because all of the initial, facial surface parameterizations were obtained by defining the point $s = (0, 0)$ at the tip of the nose, we treat this point as a landmark, i.e. it is fixed throughout the registration process. Given this simplification, we first construct a basis for $[0, 1]$ as $B_{[0,1]} = \{\sin(2\pi n_1 u), 1 - \cos(2\pi n_1 u), u, 1 - u | n_1 = 1, \ldots, N_1,\ u \in [0, 1]\}$ and a basis for $\mathbb{S}^1$ as $B_{\mathbb{S}^1} = \{\sin(n_2 v), 1 - \cos(n_2 v), v, 2\pi - v | n_2 = 1, \ldots, N_2,\ v \in [0, 2\pi]\}$. We take all products of these two bases while ensuring that the boundary of the unit disk is preserved. Then, we use the Gram-Schmidt procedure to define a finite, orthonormal basis of $T_{\gamma_{id}}(\Gamma)$ denoted by $B_{\bar{\mathbb{D}}} = \{b_1, \ldots, b_N\}$. This basis is used to iteratively deform the initial parameterization of f_2 until it is optimally registered to f_1; three example elements of this basis are shown in Figure 5.1. In the following sections, we let $f_2^* = O^*(f_2 \circ \gamma^*)$, where $O^* \in SO(3)$ is the optimal rotation and $\gamma^* \in \Gamma$ is the optimal re-parameterization. Then, the geodesic distance in the shape space $\mathcal{S} = \mathcal{C}/(SO(3) \times \Gamma)$ is computed using $d([f_1], [f_2]) = \inf_{(O,\gamma) \in SO(3) \times \Gamma} d_{\mathcal{C}}(f_1, O(f_2 \circ \gamma)) \approx d_{\mathcal{C}}(f_1, O^*(f_2 \circ \gamma^*))$. This allows us to compute the geodesic only once, after the two facial surfaces have been optimally registered. In order to improve the search over Γ we use a multi-scale approach with respect to the basis $B_{\bar{\mathbb{D}}}$. We begin with basis elements constructed using $N_1 = 3$ and $N_2 = 3$, which generate large grid deformations. Once the algorithm has converged, we further update the grid using bases with $N_1 = 5,\ N_2 = 5$ and $N_1 = 7,\ N_2 = 7$, which provide more local grid refinements.

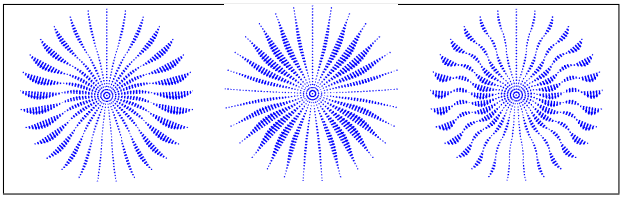


FIGURE 5.1: Three basis elements from $B_{\bar{\mathbb{D}}}$ (smooth vector fields on $\bar{\mathbb{D}}$).

As a next step, we are interested in comparing facial surface shapes using geodesic paths and distances. As mentioned earlier, there is no closed form expression for the geodesic in $\mathcal{C}$, and thus, we utilize a numerical technique termed path-
straightening. In short, this approach first initializes a path between the two given surfaces, and then "straightens" it according to an appropriate path energy gradient until it becomes a geodesic. We refer the reader to (23, 216) for more details. In the following sections, we use $F^{*,pre}$ to denote the geodesic path between two facial surfaces f_1 and f_2 in the pre-shape space (no optimization over $SO(3) \times \Gamma$) and $F^{*,sh}$ to denote the geodesic path in the shape space between f_1 and f_2^*. The length of the geodesic path is given by $L(F^*) = \int_0^1 \sqrt{\langle\langle F_t^*, F_t^* \rangle\rangle_{F^*}} dt$, where $F_t^* = \frac{dF^*}{dt}$. All derivatives and integrals in our framework are computed numerically. The computational cost of the proposed method is similar to that reported in (216).

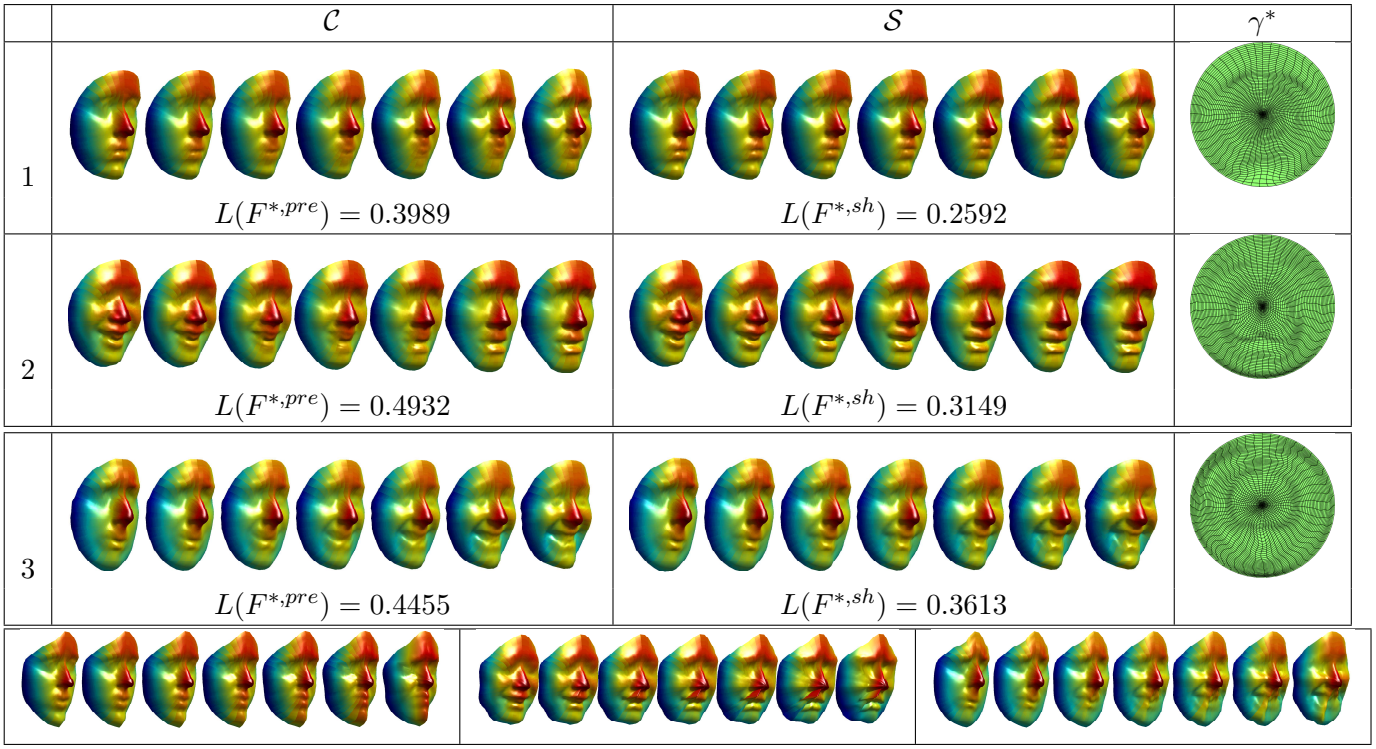


FIGURE 5.2: Top: Comparison of geodesic paths and distances in $\mathcal{C}$ and $\mathcal{S}$ for different persons and expressions ((1) neutral to anger, (2) happiness to disgust, and (3) sadness to happiness) as well as optimal re-parameterizations (allow elastic deformations between 3D faces). Bottom: Geodesics (1)-(3) computed using (215).

5.2.2 Applications

Next, we describe the utility of the presented mathematical tools in various 3D face processing tasks including deformation, template estimation, summarization of variability, random sampling and symmetry analysis. We also present two classification tasks concerned with (1) classifying expressions, and (2) classifying person identities. The 3D faces used in this work are a subset of the BU-3DFE dataset. BU-3DFE is a database of annotated 3D facial expressions, collected by Yin et al. (217) at Binghamton University in Binghamton, NY, USA, which was designed for research on 3D human faces and expressions and to develop a general understanding of human behavior. There are a total of 100 subjects in the database, 56 females and 44 males. A neutral scan was first captured for each subject. Then, each person was asked to perform six expressions reflecting the following emotions: anger, happiness, fear, disgust, sadness and surprise. The expressions varied according to four levels of intensity (low, middle, high and highest). Thus, there were 25 3D facial expression models per subject in the entire database. We use a subset of this data with highest expression intensities (most challenging case) to assess the proposed method.

Each facial surface is represented by an indexed collection of radial curves that are defined and extracted as follows. The reference curve on a facial surface f is chosen to be the vertical curve after the face has been rotated to the upright position. Then, each radial curve β_α is obtained by slicing the facial surface

by a plane P_α that has the nose tip as its origin and makes an angle α with the plane containing the reference curve. We repeat this step to extract radial curves at equally separated angles, resulting in a set of curves that are indexed by the angle α. Thus, the facial surface is represented in a polar (radius-angle) coordinate system. We use 50 radial curves sampled with 50 points in our surface representation (50×50 grid).

Face Deformation: We generate facial shape deformations using geodesic paths. While linear interpolations could also be used here, the geodesic provides the optimal deformation under the defined Riemannian metric. Since we only have to compute the geodesic once per deformation, after the surfaces have been optimally registered, this does not result in a prohibitive computational cost. We compare the results obtained in $\mathcal{C}$ to those in $\mathcal{S}$ in Figure 5.2. We consider three different examples for various persons and expressions. There is a large decrease in the geodesic distance in each case due to the additional optimization over $SO(3) \times \Gamma$. It is clear from this figure that elastic matching of 3D faces is very important when the main goal is to generate natural deformations between them. This is especially evident in the areas of the lips and eyes. Take, for instance, Example 1. In the pre-shape space, the lips are averaged out along the geodesic path and are pretty much non-existent close to the midpoint. But, due to a better matching of geometric features along the geodesic path in the shape space, the lips are clearly defined. The same can be observed in the eye region. As will be seen in the next section, these distortions become even more severe when one considers computing averages and variability within a set of 3D faces. In the right panel of the figure we display the optimal re-parameterizations that achieve the correspondence between these surfaces; these are clearly nonlinear and depict natural transformations. We also generated geodesics for the same examples using the curve-based method in (215) (bottom panel of Figure 5.2). These results suggest that considering the radial curves independently can generate severe distortions in the geodesic paths and produce unnatural deformations between 3D faces.

Face Template: We generate 3D face templates using the notion of the Karcher mean. Tools and results for computing shape statistics for cylindrical surfaces under the SRNF representation have been previously described in (218); we review some of the concepts relevant to current analysis next. Let $\{f_1, \ldots, f_n\} \in \mathcal{C}$ denote a sample of 3D faces. Then, the Karcher mean is defined as $[\bar{f}] = \text{argmin}_{[f] \in \mathcal{S}} \sum_{i=1}^{n} L(F_i^{*,sh})^2$, where $F_i^{*,sh}$ is a geodesic path between $F_i^{*,sh}(0) = f$ and a surface in the given sample $F_i^{*,sh}(1) = f_i^*$ that was optimally registered to f. A gradient-based approach for finding the Karcher mean is given in (218). The Karcher mean is actually an equivalence class of surfaces and we select one element as a representative $\bar{f} \in [\bar{f}]$. As one can see from this formulation, the computation of the Karcher mean requires n geodesic calculations per iteration. This can be very computationally expensive, and thus, we approximate the geodesic using a linear interpolation when computing the facial surface templates. We present all results in Figure 5.3. We compare the facial template computed in $\mathcal{S}$ to a standard sample average computed in $\mathcal{C}$ and the curve-based Karcher mean (215). First, we note from panel (e) that there is a large decrease in energy in each example. The qualitative results also suggest that the 3D face templates computed in $\mathcal{S}$ are much better representatives of the given data than those computed in $\mathcal{C}$ or using the curve-based method. Again, the biggest differences are noticeable around the

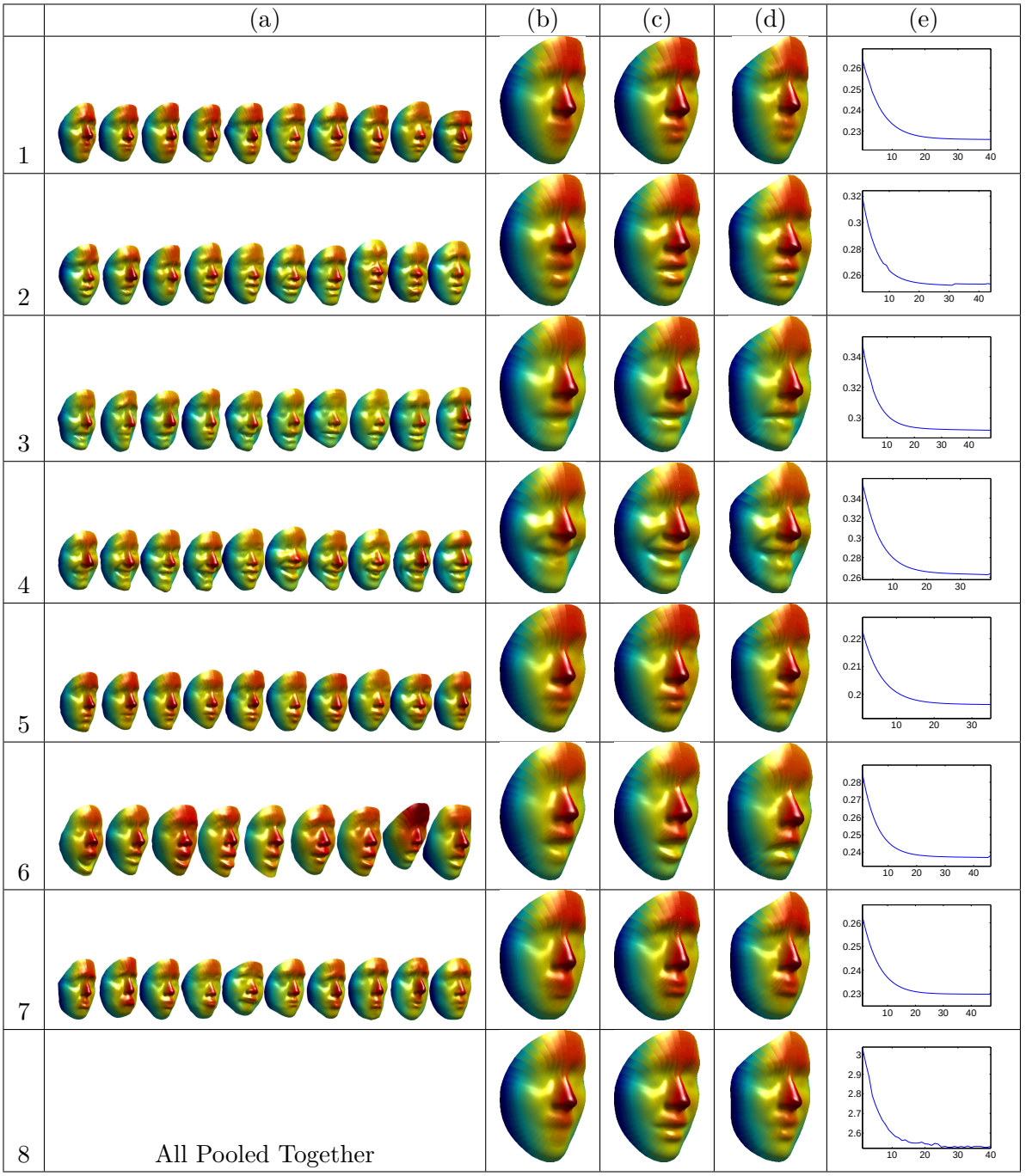


FIGURE 5.3: (a) Sample of surfaces used to compute the face template for each expression: (1) anger, (2) disgust, (3) fear, (4) happiness, (5) neutral, (6) surprise, (7) sadness, and (8) all samples pooled together. (b) Sample average computed in $\mathcal{C}$. (c) Karcher mean computed in $\mathcal{S}$. (d) Karcher mean computed using (215). (e) Optimization energy in $\mathcal{S}$ (sum of squared distances of each shape from the current average) at each iteration.

mouth and eyes. In fact, when looking at panels (b) and (d), it is difficult to recognize the expression; this distinction is much clearer in panel (c).

FIGURE 5.4: The first two principal directions of variation (PD1 and PD2) computed in the pre-shape ($\mathcal{C}$) and shape ($\mathcal{S}$) spaces for expressions (1)-(8) in Figure 5.3.

Summary of Variability and Random Sampling: Once the sample Karcher mean has been computed, the evaluation of the Karcher covariance is performed as follows. First, we optimally register all surfaces in the sample to the Karcher mean $\bar{f}$, resulting in $\{f_1^*, \ldots, f_n^*\}$, and find the shooting vectors $\{\nu_1, \ldots, \nu_n\}$ from the mean to each of the registered surfaces. The covariance matrix K is computed using $\{\nu_i\}$, and principal directions of variation in the given data can be found using standard principal component analysis (PCA). Note that due to computational complexity, we do not use the Riemannian metric $\langle\langle \cdot, \cdot \rangle\rangle$ to perform PCA; thus, we sacrifice some mathematical rigor in order to improve computational efficiency. The principal singular vectors of K can then be mapped to a surface f using the exponential map, which we approximate using a linear path; this approximation is reasonable in a neighborhood of the Karcher mean. The results for all eight samples displayed in Figure 5.3 are presented in Figure 5.4. For each example, we display the two principal directions of variation, in $\mathcal{C}$ and $\mathcal{S}$, sampled at $-2, -1, 0, 1, 2$ standard deviations around the mean. The summary of variability in $\mathcal{S}$ more closely resembles deformations present in the original data. This leads to more parsimonious shape models. In contrast to the principal directions seen in $\mathcal{C}$, the ones in $\mathcal{S}$ contain faces with clear facial features.

Given a principal component basis for the tangent space $T_{[\bar{f}]}(\mathcal{S})$, one can sample random facial shapes from an approximate Gaussian model. A random tangent vector is generated as $v = \sum_{j=1}^{k} z_j \sqrt{S_{jj}} u_j$, where $z_j \overset{iid}{\sim} N(0,1)$, S_{jj} is the variance of the jth principal component, and u_j is the corresponding principal singular vector of K. A sample from the approximate Gaussian is obtained using the exponential

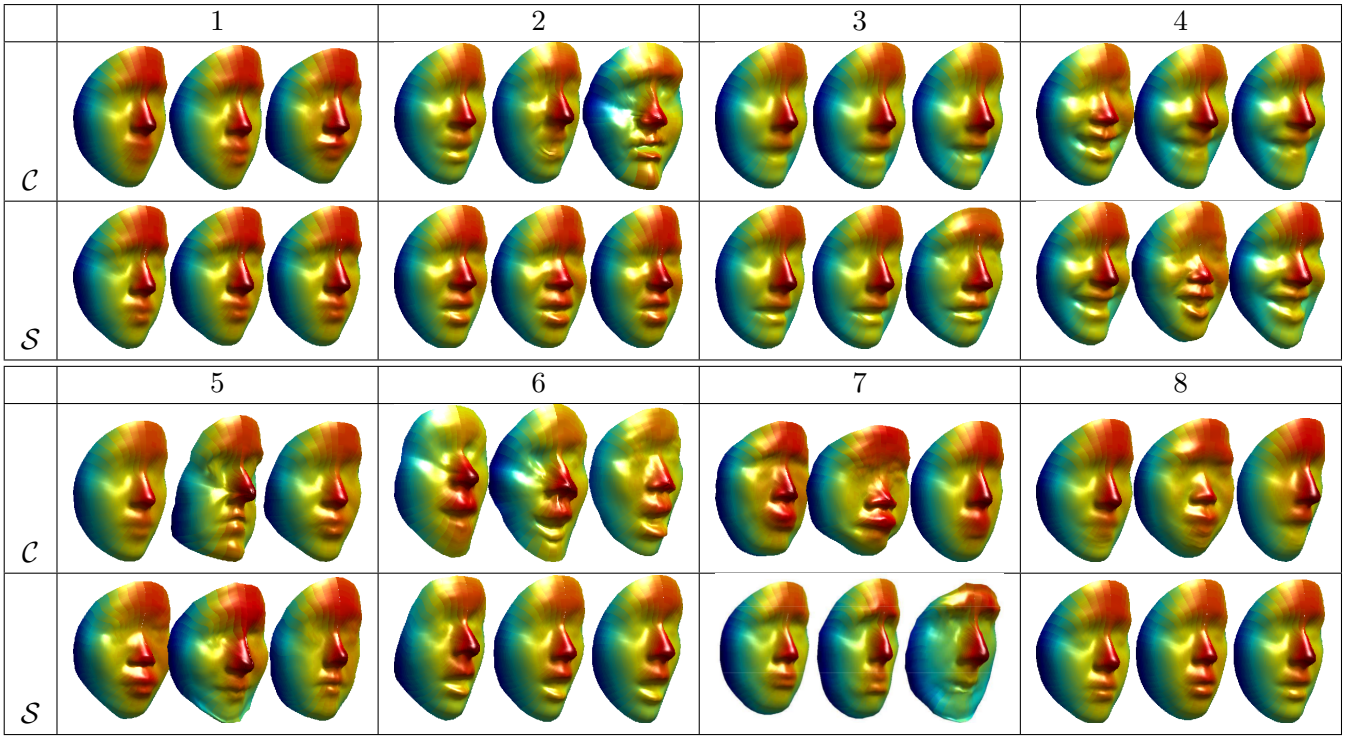


FIGURE 5.5: Random samples generated from the approximate Gaussian distribution in the pre-shape ($\mathcal{C}$) and shape ($\mathcal{S}$) spaces for expressions (1)-(8) in Figure 5.3.

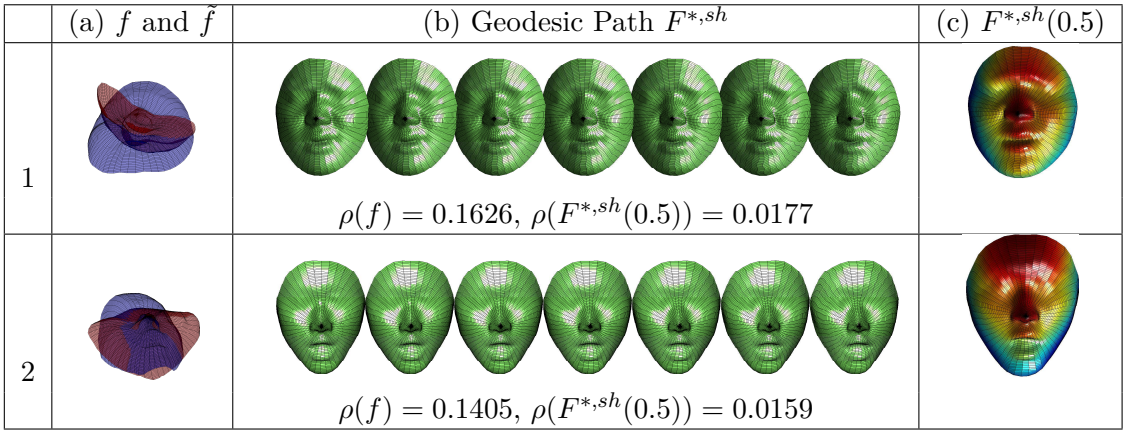


FIGURE 5.6: (a) Facial surface f in blue and its reflection $\tilde{f}$ in red. (b) Geodesic path in $\mathcal{S}$ between f and $\tilde{f}$ and the measure of symmetry $\rho(f)$. We also compute the measure of symmetry for the midpoint of the geodesic $\rho(F^{*,sh}(0.5))$, which is expected to be 0 for perfectly symmetric faces. (c) Midpoint of the geodesic.

map $f_{rand} = \exp_{\bar{f}}(v)$, which again is approximated via a linear path. The results are presented in Figure 5.5. As expected, the 3D faces sampled in the shape space are visually preferred to those sampled in the pre-shape space; this is due to better matching of similar geometric features across 3D faces such as the lips, eyes and cheeks.

Symmetry Analysis: To analyze the level of reflection symmetry of a facial surface f we first obtain

its reflection $\tilde{f} = H(v)f$, where $H(v) = I - 2\frac{vv^T}{v^Tv}$ for a $v \in \mathbb{R}^3$. Let $F^{*,sh}$ be the geodesic path between f and $\tilde{f}^* = O^*(\tilde{f} \circ \gamma^*)$. We define the length of the path $F^{*,sh}$ as a measure of symmetry of f, $\rho(f) = L(F^{*,sh})$. If $\rho(f) = 0$ then f is perfectly symmetric. Furthermore, the halfway point along the geodesic, i.e. $F^{*,sh}(0.5)$, is approximately symmetric (up to numerical errors in the registration and geodesic computation). If the geodesic path is unique, then amongst all symmetric shapes, $F^{*,sh}(0.5)$ is the closest to f in $\mathcal{S}$. Two different examples are presented in Figure 5.6. The measures of symmetry for the geodesic midpoints are 0.0177 and 0.0159, which are very close to 0 (perfect symmetry). In the presented examples, the faces are already fairly symmetric. Nonetheless, the symmetrized faces (right panel) have a natural appearance with clearly defined facial features.

Identity and Expression Classification: In the final application, we explore the use of the proposed framework in two classification tasks. We compare our results to the method presented in (215), which reported state-of-the-art recognition performance in the presence of expressions. We do not compare our performance to any other state-of-the-art methods because many of them are specifically designed for classification experiments (feature based). Our framework is more general as it also allows deformation and statistical modeling of faces. The proposed framework can be tuned to maximize classification performance by extracting relevant elastic features from the computed statistical models, but we believe that this is beyond the scope of the current study.

The first task we consider is concerned with classifying expressions. We selected 66 total surfaces divided into six expression groups (11 persons per group): anger, disgust, fear, happiness, surprise and sadness. We computed the pairwise distance matrices in $\mathcal{C}$, $\mathcal{S}$, and using (215). We calculated the classification performance in a leave-one-out manner by leaving out all six expressions of the test person from the training set. The classification accuracy in $\mathcal{C}$ was 62.12% while that in $\mathcal{S}$ was 74.24%. The classification accuracy of (215) was 68.18%. This result highlights the benefits of elastic shape analysis of hemispherical surfaces applied to this recognition task. It also suggests that considering the radial curves independently, as done in (215), deteriorates recognition performance.

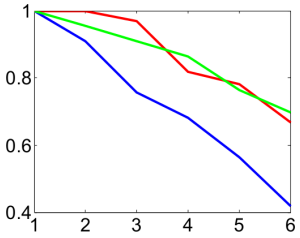


FIGURE 5.7: Identity recognition in $\mathcal{C}$ (blue), $\mathcal{S}$ (red), and using (215) (green).

The second task we consider is identity classification in presence of highest level of facial expressions. Here, we added 11 neutral expression facial surfaces (one per person) to the previously used 66 and computed 11×66 distance matrices in $\mathcal{C}$, $\mathcal{S}$, and using the method in (215). We performed classification by first checking the identity of the nearest neighbor. This resulted in a 100% classification rate for

all methods. Figure 5.7 shows the classification results when accumulating over more and more nearest neighbors (up to six since there are six total expressions for each person). It is clear from this figure that identity classification in the shape space is far superior to that in the pre-shape space. The additional search over Γ allows for the expressed faces to be much better matched to the neutral faces, and in a way provides "invariance" to facial expressions in this classification task. The performance of the proposed method is comparable to (215).

5.3 Gauge Invariant Framework for Shape Analysis of Surfaces

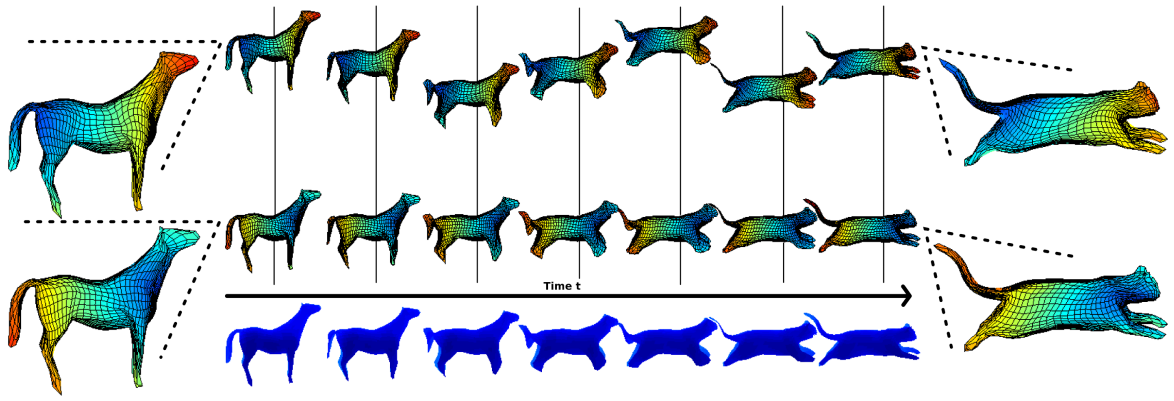


FIGURE 5.8: Two paths in $\mathcal{F}$ with the same sequence of shapes but with different reparameterizations between the corresponding shapes.

Our goal is to develop tools for analyzing shapes of two-dimensional surfaces with certain local constraints (smoothness, no-holes, etc). The main difficulty in comparing shapes of such surfaces is that there is no preferred parameterization that can be used for registering and comparing features across surfaces. Since the shape of a surface is invariant to its parameterization, one would like an approach that yields the same result irrespective of the parameterization.

Furthermore, we are not only interested in the comparison and matching of two shapes, but also in the deformation processes that may transform one shape into another, i.e. metamorphosis. To be physically meaningful, the evolution from one shape to another should be independent of the way surfaces may be parameterized. Our approach to shape analysis presented in this section was therefore initiated by the following question : *What is the natural framework where one can measure deformations of shapes independently of the way shapes are parameterized?* As a motivating example, the sequence of shapes displayed in Fig. 5.8 (bottom) denotes a path where a horse is transformed into a jumping cat. During the transformation process, only the change of shape, drawn in the bottom line as a sequence of blue surfaces, is relevant to us. How the surfaces may be parameterized during the metamorphosis has no importance in our context. To emphasize this idea, two paths of parameterized surfaces corresponding to the same transformation process are displayed in the top two rows. We would like a framework where the physical quantities measured on the path of shapes, such as its length or its energy, are independent of the

parameterizations of surfaces along the transformation process. In particular, in Fig. 5.8, the two paths of parameterized surfaces corresponding to the same transformation process should have the same length. Note that the surfaces along the second path are obtained by applying a *different reparameterization at each time step* to the surfaces along the first path.

Let us emphasize that we are not only interested in how *far* the horse and the jumping cat are from each other, in other words in a quantity like a distance measuring the minimal cost needed to deform the horse into a cat. But, given a metamorphosis between these two shapes, we are also interested in measuring its *length* on one hand, and its *energy* on the other hand, independently of the parameterizations of the transformation process that may have been used to create this metamorphosis. Recall that the length of a path is the integral of the norm velocity function with respect to time and has the dimension of a distance. The energy is the integral of the square of the norm velocity function with respect to time, hence has the dimension of the square of a distance divided by time.

Let us now summarize past work on related subjects. The initial set of papers developed algorithms for geodesic deformations between surfaces while using the given registration of points. They compute geodesics between shapes, under isometric deformations, while assuming the registration (or parameterization) as given. Windheuser et al. (219) proposed to find a geometrically consistent matching of 3D shapes which minimizes an elastic deformation energy but use a linear interpolation between registered pairs of points in $\mathbb{R}^3$ to compute geodesic paths. Another paper by Kilian et al. (220) represents parameterized surfaces by discrete triangulated meshes, assumes a Riemannian metric on the space of such meshes, and computes geodesic paths between given meshes. The main limitation here is that it assumes the correspondence between points across meshes. That is, we need to know beforehand which point on one mesh matches with which point on the second mesh. The same limitation holds for the paper by Heeren et al. (221) also. In contrast, we would like to remove the reparameterization variability so that different surfaces with the same shape but different parameterizations have zero distance between them.

Motivated by progress in shape analysis of curves (222, 223), Kurtek et al. (224), (225) introduced a new representation, termed a *q-map* of surfaces such that the $\mathbb{L}^2$ distance in this representation space is invariant to simultaneous reparameterizations of surfaces. For convenience of the reader, we recall the definition of the q-map but we will not use it in the present work. Let $f : \mathbb{S}^2 \to \mathbb{R}^3$ denote a smooth parameterized surface and $\mathcal{F}$ be the set of such surfaces. Then, this q-map is given by $f \mapsto q$ where $q(s) = \sqrt{r(s)}f(s)$ and $r(s)$ is the area multiplication factor of f at $s \in \mathbb{S}^2$. They defined a Riemannian metric on the space of parameterized surfaces by *pulling back* $\mathbb{L}^2$ metric under the q-map, and used a *path-straightening* algorithm to compute geodesic paths between given surfaces in a pre-shape space. This path-straightening is an iterative algorithm that updates an arbitrary initial path using the gradient of the energy function mentioned above, until the path converges to a geodesic. The energy gradient is approximated numerically using an (approximate) finite basis for $\mathcal{F}$. To remove the effects of original parameterizations, and to obtain geodesics in the shape space, they solve for an optimal reparameterization of one of the surfaces, under the same energy. There are several other papers,

including (226), that focus exclusively on the task of finding optimal correspondence between 3D objects, either using physically-motivated energies or Riemannian metrics. Due to the use of gradient-based searches, these methods and previously mentioned papers do not guarantee a global solution, either for geodesics or for registration. In path-staightening, however, it can be shown that a path that is a local minimum of the path energy is a geodesic path, albeit not the shortest geodesic. To our knowledge, very few methods guarantee a globally-optimal solution to the problem of finding geodesics in shapes spaces of surfaces. Although (224) was the first to provide a geometric framework for joint registration-comparison problem, the Riemannian metric used there has a limitation that it was not translation invariant.

To handle the translation issue mentioned above, Jermyn et al. (227) introduced a comprehensive Riemannian metric that has several improvements, including the fact that it was translation invariant and allows some physical interpretations in its use. This metric, given later in Eqn. (5.2), has terms that can be interpreted as measurements of bending, stretching, and changes in local curvatures of surfaces. It has been termed an *elastic metric* because it is invariant to reparameterizations and the physical interpretations associated with it. Although (227) introduced this metric, it did not use the full metric to compute geodesic paths. Instead, it defined a new map, termed the square-root normal field, given by $q(s) = \sqrt{r(s)}n_f(s)$ where $n_f(s)$ denotes the unit normal to the surface at the point $s \in \mathbb{S}^2$. The square-root normal field has the property that the *last two terms* of the elastic metric transform to the $\mathbb{L}^2$ metric under the map $f \mapsto q$, for some weighting of last two terms in the metric. The first term of the metric is discarded in this analysis. The transformation to $\mathbb{L}^2$ metric is useful since one can apply some common tools from Hilbert space analysis to this problem, including the optimization over the reparameterization group for optimal registration, but this mapping $f \mapsto q$ is not onto and, hence, not invertible. The optimization step is challenging because the reparameterization group is an infinite-dimensional Fréchet Lie group, and the exponential map is not a local diffeomorphism. Since the first term of the elastic metric introduced in (227) is not used by Jermyn et al., it can result in zero shape distance between two surfaces that actually have different shapes. For example, a thin-tall cylinder and a fat-short cylinder, with same surface areas and unit normals, will have zero shape difference under this framework.

Another line of work in shape analysis comes from Michor et al. (228), Bauer et al. (229), (230), (231) and Fuchs et al. (232) (see also (233) for an overview of a lot of mathematical results in this area). Different types of metrics have been studied : Sobolev metrics in (231), curvature weighted metrics in (229), almost local metrics in (230), metrics mesuring the deformations of the interiors of shapes in (232). Let us mention that the first two terms of the metric we use in the present work fit in the general study laid out in (231), and are related to the metrics studied in (234), (235), (236). In this set of papers, the idea is to replace the problem of solving the geodesic equation on shape space by the equivalent problem of solving the equation for horizontal geodesics in the pre-shape space. A geodesic in pre-shape space is horizontal if it is orthogonal to the orbits of the reparameterization group. One task in this strategy is therefore to compute the horizontal space on which the quotient map is an isometry, or equivalently solve a minimization problem for the infinitesimal energy. Depending on the Riemannian metric on the pre-shape space, this task may be computationally trivial or extremely difficult to implement (for metrics

used in (229) and (230) it is just the space of normal vector fields, but for metrics used in (231) and (232) it involves the inversion of a pseudo-differential operator). Another main contribution of these authors is to give sufficient conditions under which the Riemannian metric induced on shape space separates points, i.e. gives a non-zero geodesic distance between pairs of different shapes (a condition that is necessary to make shape comparison). It is worth noting that, in this infinite-dimensional context, vanishing geodesic distance is a common phenomenon (as was first highlighted in (228)). For the metric we use, non-vanishing geodesic distance is guaranteed by the non-vanishing geodesic distance on the space of Riemannian metrics proved in (237) (at least on pairs of shapes inducing different pull-back metrics on the sphere, which is what we are interested in practice).

To summarize, the past approaches involving Riemannian geometry have tended to perform shape analysis in two steps. First, they select a representation space, or a pre-shape space, for objects of interest – curves (222, 223, 238) and surfaces (225, 227, 229–231) – and impose a Riemannian structure on it ensuring that the actions of shape-preserving groups are by isometries. Next, they inherit this metric to the quotient space of the pre-shape space modulo the requisite groups, called the shape space, and seek geodesics between objects in this shape space. The task of inheriting Riemannian metrics to quotient spaces is complicated because reparameterization groups are Fréchet Lie groups and the process of inheriting a metric requires closed orbits, as can be seen in (222, 231, 238, 239), etc. Even though endowing shape space with a Riemannian metric (with positive distance function) seems to be a good approach, inducing this metric by a Riemannian metric on pre-shape space leads to difficulties that one would like to avoid (recall that we are only interested in shapes and not in the way they are parameterized). We will pursue a different strategy where the Riemannian metric is directly imposed on the quotient space, thus avoiding the need to satisfy conditions for inheriting metrics from the pre-shape space or computing an abstract horizontal space. Motivated by an easy implementation of the metrics, we take the point of view where the space of interest is the space of normal vector fields (in contrast with the horizontal space of a Riemannian submersion). Let us emphasize that there is no restriction in doing so : any Riemannian metric on shape space can by expressed as a metric defined on normal vector fields.

Context and Contributions

In this work, we proposed a novel framework for shape analysis of 3D surfaces. This work has been elaborated in collaboration with Dr Barbara Tumpach (Associate professor in Mathematic, Painlevé laboratory, university of Lille) during a CNRS delegation, Pr M. Daoudi and Pr. Anuj Srivastava (Florida State University. The resulting work has been published in the prestigious journal IEEE Transaction on Pattern Analysis and Machine Intelligence PAMI (214). A synthesis of the contribution is presented below, for more details please refer to the paper (214).

5.3.1 Gauge theory in shape analysis

A gauge theory, In physics, is a type of field theory in which the Lagrangian does not change (is invariant) under local transformations from certain Lie groups. The term 'gauge invariance' was proposed by Hermann Weyl in 1918. The term gauge refers to any specific mathematical formalism to regulate redundant degrees of freedom in the Lagrangian. Lagrangian of a dynamic system is a function of dynamic variables which allows the equations of motion of the system to be written concisely. The transformations between possible gauges, called gauge transformations, form a Lie group—referred to as the symmetry group or the gauge group of the theory. Associated with any Lie group is the Lie algebra of group generators. For each group generator there necessarily arises a corresponding field (usually a vector field) called the gauge field. Gauge fields are included in the Lagrangian to ensure its invariance under the local group transformations (called gauge invariance). (source: wikipedia)

The key idea of this work is the application of the gauge theory to the space of functions describing 3D spherical surfaces in order to calculate distances and geodesic paths that are invariant to the re-parameterization of the surface.

Let S define a 3D surface, we shall represent it by an embedding $f : \mathbb{S}^2 \to \mathbb{R}^3$ such that the image $f(\mathbb{S}^2)$ is S. The function f is also called a *parameterization* of the surface S. The space of all such surfaces is defined as $\mathcal{F}$ the set of

$$\mathcal{F} := \{f : \mathbb{S}^2 \to \mathbb{R}^3, f \text{ is an embedding}\}.$$

It is often called the *pre-shape space* since objects with same shape but different orientations or parameterizations may correspond to different points in $\mathcal{F}$. In this space, one can define a path $\Psi : [0,1] \mapsto \mathcal{F}$ between two surfaces (two shapes). This path is composed of a set of shapes and represents a metamorphosis from the initial shape to the final one. The set of such paths is the smooth manifold $\mathcal{P} := \mathcal{C}^\infty([0,1], \mathcal{F})$. Fig. 5.8 shows two elements in $\mathcal{P}$ representing two different paths $\Psi_1 : t \mapsto \Psi(t)$ and $\Psi_2 : t \mapsto \Psi(t)$ from a horse to a cat. For each path, Ψ_i, at each time step $t \in [0,1]$, $\Psi_i(t)$ is a parameterized shape ($i = 1, 2$). $\Psi_1(0) = \Psi_2(0)$ correspond to the initial shape (horse) and $\Psi_1(1) = \Psi_2(1)$ is the final shape (cat). **Our aim is to match the length of any given path** Ψ, let's denote it as $L[\Psi]$, to the length of the path $t \mapsto \Psi(t) \circ \gamma(t)$, where $t \mapsto \gamma(t) \in \Gamma$ is any time-dependent reparameterization of $\mathbb{S}^2$:

$$L[\Psi] = L[\tilde{\Psi}], \quad \text{where } \tilde{\Psi}(t) = \Psi(t) \circ \gamma(t). \tag{5.1}$$

More formally, set $\Gamma = \text{Diff}^+(\mathbb{S}^2)$ and define the group $\mathcal{G} := \mathcal{C}^\infty([0,1], \Gamma)$, of time-dependant reparameterizations that acts on $\mathcal{P}$ according to

$$\begin{array}{ccc} \mathcal{G} \times \mathcal{P} & \longrightarrow & \mathcal{P} \\ (t \mapsto \gamma(t), t \mapsto \Psi(t)) & \longmapsto & (t \mapsto \Psi(t) \circ \gamma(t)). \end{array}$$

The group $\mathcal{G}$ is the *gauge group*, and one says that $\mathcal{G}$ acts by *gauge transformations*. The gauge theory with these ingredients implies a framework where the calculation of the length of any given path is invariant to the re-parameterization. In order to design such a framework, we propose to consider the Γ-invariant Riemannian metric $\langle\langle \cdot, \cdot \rangle\rangle$ on the pre-shape space, and *ignore* the direction tangent to the reparameterization orbit.

We choose the Riemannian metric proposed in (227) and defined as:

$$\begin{aligned} \langle\langle \delta f_1, \delta f_2 \rangle\rangle_f &= \int_{\mathbb{S}^2} ds |g|^{\frac{1}{2}} \Big\{ a \operatorname{Tr}(g^{-1}\delta g_1 \, g^{-1}\delta g_2) \\ &+ \frac{\lambda}{2} \operatorname{Tr}(g^{-1}\delta g_1) \operatorname{Tr}(g^{-1}\delta g_2) + c\delta n_1 \cdot \delta n_2 \Big\}. \end{aligned} \tag{5.2}$$

where $g = f^* \bar{g}$ denotes the pull-back of the Euclidian metric $\bar{g}$ of $\mathbb{R}^3$ and n the unit normal vector field (Gauss map) on $S = f(\mathbb{S}^2)$. δf_1, δf_2 denote two perturbations of a surface f, and $(\delta g_1, \delta n_1) = \Phi_*(\delta f_1)$, $(\delta g_2, \delta n_2) = \Phi_*(\delta f_2)$ denote the corresponding perturbations in (g, n) of f.

More precisely, let $\langle\langle \cdot, \cdot \rangle\rangle$ be a Riemannian metric on pre-shape space $\mathcal{F}$ which is preserved by the action of the group of reparameterizations Γ, that is:

$$\langle\langle \delta f_1 \circ \gamma, \delta f_2 \circ \gamma \rangle\rangle_{f \circ \gamma} = \langle\langle \delta f_1, \delta f_2 \rangle\rangle_f, \tag{5.3}$$

for any $f \in \mathcal{F}$, for any $\delta f_1, \delta f_2 \in T_f\mathcal{F}$ and any $\gamma \in \Gamma$. Given a Γ-invariant sub-bundle H of $T\mathcal{F}$ such that

$$H(f) \oplus Ver(f) = T_f\mathcal{F}, \tag{5.4}$$

denote by $p_H : T_f\mathcal{F} \to H(f)$ the projection onto $H(f)$ with respect to the direct sum decomposition given in Eqn. (5.4). This means that any element $\delta f \in T_f\mathcal{F}$ admits a unique decomposition into the

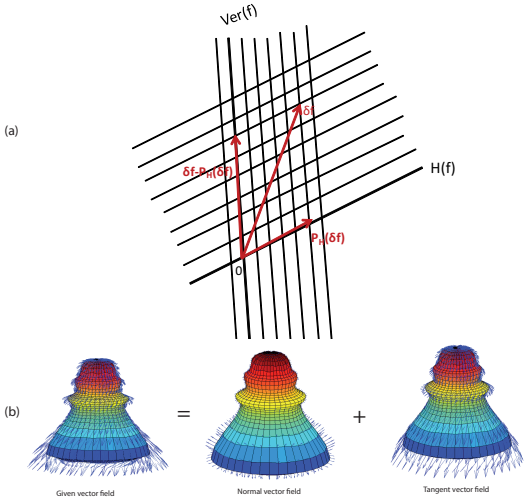


FIGURE 5.9: a. Direct sum decomposition $H(f) \oplus Ver(f) = T_f\mathcal{F}$. b. Vector field decomposition into tangent and normal directions

sum of an element $p_H(\delta f)$ in $H(f)$ and an element in $Ver(f)$. We illustrate this decomposition of vector

spaces in Fig. 5.9.a, while the particular case when H is the space of normal vector fields Nor is shown in Fig. 5.9.b. The non-negative semi-definite inner product on pre-shape space defined by

$$(\!(\delta f_1, \delta f_2)\!)_f := \langle\!\langle p_H(\delta f_1), p_H(\delta f_2)\rangle\!\rangle_f$$

satisfies the gauge-invariance condition given in Eqn. (5.1) and induces a Riemannian metric on quotient space $\mathcal{S}$ such that the quotient map is an isometry between $H(f)$ and the tangent space $T_{[f]}\mathcal{S}$.

5.3.2 Path straightening for geodesic calculation

The path-straightening method is used to find critical points of the energy functional. Starting with an arbitrary path, the method consists of iteratively deforming (or "straightening") the path in the opposite direction of the gradient, until the path converges to a geodesic. The gradient of the path energy is approximated using a basis $\mathcal{B}$ of possible perturbations of a path of surfaces Ψ, as constructed in the previous section. We first compute the directional derivatives $\nabla\mathcal{E}_\Psi(b) = \frac{d}{d\epsilon}(\mathcal{E}(\Psi + \epsilon b))|_{\epsilon=0}$ where b ranges over $\mathcal{B}$. This is done by fixing a small ϵ_1 and approximating the directional derivative by $\nabla\mathcal{E}_\Psi(b) \simeq (\mathcal{E}(\Psi + \epsilon_1 b) - \mathcal{E}(\Psi))\epsilon_1{}^{-1}$. Using a finite orthonormal basis $\mathcal{B}$, we obtain a numerical approximation of the gradient: $\nabla\mathcal{E}_\Psi = \sum_{b\in\mathcal{B}} \nabla\mathcal{E}_\Psi(b)\ b$. In particular, the norm of the gradient is approximately given by $\|\nabla\mathcal{E}_\Psi\|^2 = \sum_{b\in\mathcal{B}} \nabla\mathcal{E}_\Psi(b)^2$. The update of the path is done by replacing Ψ by $\Psi - \epsilon_2\nabla\mathcal{E}_\Psi$, where ϵ_2 is a small parameter that has to be ajusted empirically. The method is detailed in Algorithm 7 below.

Algorithm 7 Path-straightening method.

Require: A path Ψ between two parameterized surfaces f_1 and f_2, a basis of perturbation $\mathcal{B}$.
Ensure: The minimal energy needed to deform f_1 into f_2 given by the value of the cost function E, the geodesic path between f_1 and f_2. Set $\|\nabla E\|^2 = 1$.
while $\|\nabla E\|^2 > 10^{-3}$ **do**
2- Compute the energy E of the path Ψ.
3- Set $\Psi_{\text{upd}} = 0$ and $\|\nabla E\|^2 = 0$.

for $i \leftarrow 1, size(\mathcal{B})$ **do**
4- Add a perturbation to the current path Ψ: define $\Psi(i) = \Psi + \epsilon_1\,\mathcal{B}(i)$, where $\mathcal{B}(i)$ is the element of the perturbation basis $\mathcal{B}$ of index i and $\epsilon_1 > 0$ is small.

5- Compute the energy $E(i)$ of the perturbed path $\Psi(i)$.
6- Compute the gradient of energy $\nabla E(i)$ in the direction $\mathcal{B}(i)$ using the approximation $\nabla E(i) \sim \frac{E(i)-E}{\epsilon_1}$.
7- Compute the updating path: $\Psi_{\text{upd}} \leftarrow \Psi_{\text{upd}} + \nabla E(i)\cdot\mathcal{B}(i)$.
8- Compute the squarred norm of the gradient of energy at path Ψ:
$\|\nabla E\|^2 \leftarrow \|\nabla E\|^2 + (\nabla E(i))^2$.
end for
10- Update the path: $\Psi = \Psi - \epsilon_2\Psi_{\text{upd}}$
end while

5.3.3 Discretization of infinite dimensional space

Once the mathematical framework was set up, we encountered difficulties emanating from the transition from the theoretical infinite dimension to the finite dimension used by computers: surface discretization, approximation of curvatures, significant basis of shape deformation (the tangent space to the shape space is of infinite dimension, it was necessary to constitute a sufficiently large base of deformations to treat the forms considered but not taking up too much memory).

Especially, in order to form possible directions for use in path-straightening, we propose to built orthonormal basis of deformatinos. The first basis we used is a variation of the one given in (225). We start with a basis $\mathcal{B}_1 = \{Y_l^m, 1 \leq l \leq N, -l \leq m \leq l\}$ of spherical harmonics of degree less than N, available in Matlab as function SPHARM (see (240) for more information on spherical harmonics). We make three copies of this basis of $\mathbb{R}$-valued functions in order to obtain a basis $\mathcal{B}_2$ of the space $L^2(\mathbb{S}^2, \mathbb{R}^3)$ of $\mathbb{R}^3$-valued functions. Path-straightening method requires perturbations that vanish at $t = 0$ and $t = 1$ so that f_1 and f_2 remain fixed. Therefore, we want a basis of $L^2(\mathbb{S}^2 \times [0, 1], \mathbb{R}^3)$ with elements that have this property. To ensure this, each element of $\mathcal{B}_2$ is multiplied by a basis element of $L^2([0, 1], \mathbb{R})$ of the form $P_j(t) = \frac{1}{4}\sin(\pi j t)$, $1 \leq j \leq J$. Next we orthonormalize the L^2-basis with respect to an H^1-type scalar product (i.e. that measures also the variation of the derivatives).

The number of basis elements, used in path-straightening is crucial and affect the resulting geodesic path. Given two parameterized surfaces f_1 and f_2, we initialize the path with the linear interpolation to a different surface f_3 in the middle of the path. This initial path is shown in the upper row of Fig 5.10. Then, we compute the geodesic path using different number of basis elements. We show the geodesic paths that use 52, 432 and 1728 basis elements, respectively. We can see that the larger the number of basis elements, the better the final result is. We also provide the trade-off between the number of basis elements and the minimum energy value obtained. The trade-off confirms our assertion. At the bottom of the figure, we show the geodesic path obtained when the path-straightening Algorithm is initialized with the linear interpolation between f_1 and f_2. This path is also calculated using the number of basis elements corresponding to the lowest energy. This path can be seen as ground truth to visually interpret the previous geodesics (with more complicated initial conditions and fewer basis elements).

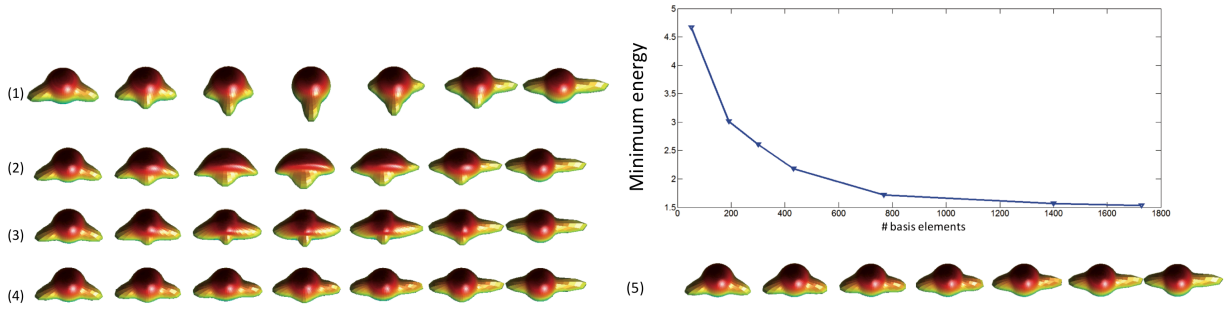


FIGURE 5.10: The effect of the number of basis elements, (1) initial path, (2) geodesic path using 52 basis elements, (3) geodesic path using 432 basis elements, (4) geodesic path using 1728 basis elements, (5) geodesic path using 1728 basis elements after linear interpolation initialization.

5.3.4 Examples of geodesics obtained by path-straightening

The 3D realistic models used in our experiments are part of the TOSCA (241) dataset. Their spherical parameterizations were initially implemented in (242).

First we apply the path-straightening method to the case where the surfaces at the extremes of the initial path have the same shape, but different parameterizations. More precisely, we consider the special case where $\Psi_0(0) = f_1$, $\Psi_0(1) = f_1 \circ \gamma$ for some diffeomorphism γ and where we initialize the path with piecewise linear interpolation to a different surface f_3 in the middle of the path, i.e. $\Psi_0(\frac{t}{2}) = f_3$. This situation is illustrated in Fig. 5.11. The proposed gauge-invariant approach is expected to reach a path with constant shape as a geodesic, despite the different shapes appearing in the initial path and the different parameterization of shapes at the end points of the path (to emphasize the differences in parameterization, zoom-ins of these surfaces are also shown). Once we have the geodesic path Ψ between the given surfaces, the distance in the shape space between f_1 and $f_1 \circ \gamma$, $d_\Psi(f_1, f_2)$, is the length of Ψ. As expected, the resulting geodesic path, shown in Fig. 5.11, is constant with the same shape as the either end, and with $d_\Psi(f_1, f_2) = 0$. Using path-straightening, we obtain a 99.28% decrease in the energy function from the initial path to the final path.

In Fig. 5.12 we consider more challenging shapes. The top-two rows display the case where we have $\Psi_0(0) = f_1$, $\Psi_0(1) = f_1$ (a cat) and where we initialize the path with piecewise linear interpolation to a horse in the middle of the path. The upper row shows the initial path and the second row the geodesic path. We can see that the geodesic path has a constant shape throughout, as expected. We also plot the evolution of the path energy on the right during path-straightening. We can see that the energy decreases until it reaches a relatively small value; the theoretical minimum is, of course, zero for a contant path. In the last two rows of Fig. 5.12, we consider the case of two hands. We initialize the path with linear interpolation (third row in Fig. 5.12), and the resulting path is shown in the last rows of Fig. 5.12. The energy evolution is shown on the right and we can see the energy decreasing until it reaches a constant value; thus, the final path is a geodesic. It can be seen that the deformation along the geodesic path is more natural than the original path.

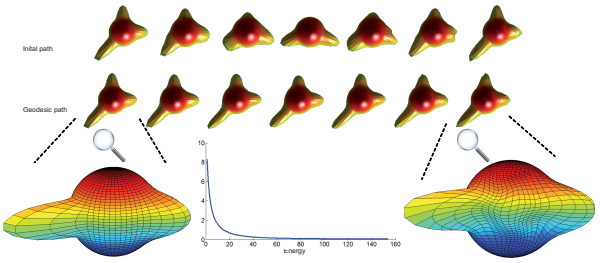


FIGURE 5.11: Illustration of initial path (upper row) and geodesic path in shape space (middle row). The energy is reported in the buttom row. The surfaces at the end points of the path have different parameterizations.

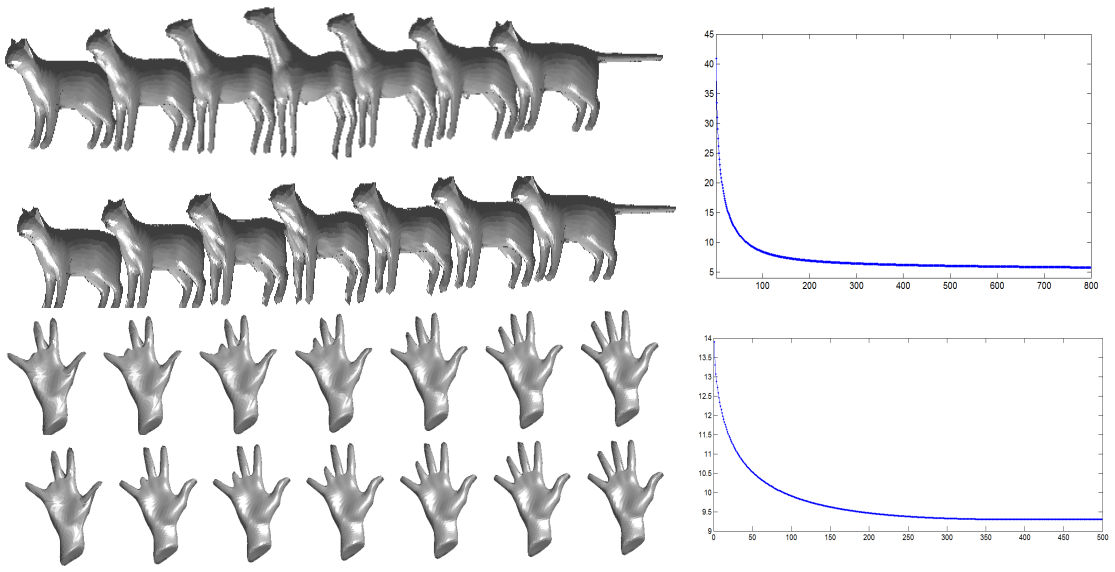


FIGURE 5.12: The top row shows an initial path formed by linear interpolation between a cat to a horse and back to the cat. The second row illustrates the geodesic obtained after 800 iterations of path-straightening. The corresponding evolution of the energy is shown on the right. Similarly, the third row shows a linear path between two hands with bad correspondence and the last row shows the final geodesic, with the corresponding energy is shown on the right.

5.3.5 Classification of 3D shapes

As mentioned earlier, the geodesic paths provide us with tools for comparing, and deforming parameterized surfaces. We suggest a comparison of shapes of 3D objects using geodesic distances between their boundary surfaces in the shape space. This section presents a specific application to illustrate that idea. In this section, we study several shapes belonging to four classes: horses, hands, cats and centaurs.

We begin by computing the pairwise geodesic distances between corresponding 3D surfaces. The distance matrix and the classification dendrogram are shown in Fig. 5.13. In the distance matrix, we can easily distinguish four classes corresponding to four blue boxes. Actually the cold colors in the illustrated matrix correspond to small values of distances versus hot colors that correspond to greater distances. The clustering obtained using the *dendrogram* (command in matlab) can be interpreted by slicing the top of the dendrogram by a horizontal line to split the shapes into the desired number of classes, and then sliding the horizontal line to the bottom in order to refine the classification. The coarsest classification results by slicing the dendrogram into two classes (by a horizontal line close to the top), the shapes 4, 5 and 6 (the hands) forms a first class and the remaining (horses, cats and centaurs) are grouped together as a second class. The next level in classification distinguishes the shapes 1, 2, and 3 (the horses) and 12, 13 (the centaurs) from the shapes 7, 8, 9, 10, 11 (the cats). The finest level separates the horses and the centaurs in different classes and results in four classes. Thus, we argue that the proposed framework provides a powerful tool for shape classification.

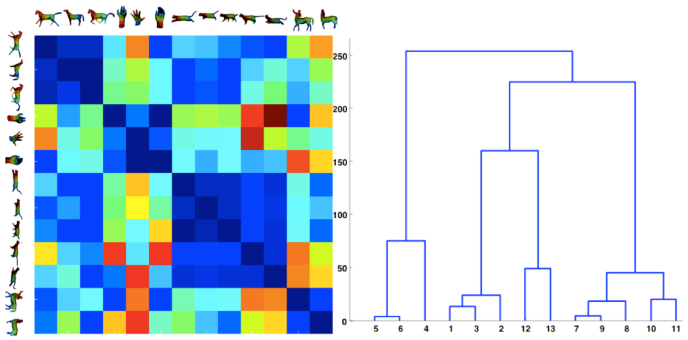


FIGURE 5.13: Classification performance; left: the distance matrix. right: the dendrogram.

5.4 Conclusion

In this chapter, we first defined a Riemannian framework for statistical shape analysis of hemispherical surfaces and applied it to various 3D face modeling tasks including morphing, averaging, exploring variability, defining generative models for random sampling, and symmetry analysis. We considered two classification experiments, one on expressions and one on person identities, to showcase the benefits of elastic shape analysis in this application. The second contribution presented in this chapter is a novel Riemannian framework for computing geodesic paths between shapes of parameterized surfaces. These geodesics are invariant to rigid motion, scaling and most importantly reparameterization of individual surfaces. The novelty lies in defining a Riemannian metric directly on the quotient (shape) space, rather than inheriting it from pre-shape space, and in using it to formulate a path energy that measures only the normal components of velocities along the path. The geodesic computation is based on a path-straightening technique that iteratively corrects paths between surfaces until geodesics are achieved.

Chapter 6

Ongoing Research and Perspectives

I have shown in this *habilitation* the interest of using shape analysis on manifolds for several computer vision applications, especially related to human behaviour understanding. In particular, to filter some undesirable transformations, the shape extracted from the human body and face are represented as elements of a shape space defined as the invariant under the action of groups modeling the undesirable transformations. Due to the non linearity of the underlying spaces, tools from differential geometry are very useful and provide geometric interpretations such as the notion of geodesic and its relevance to find the most efficient way to deform one shape to another. Moreover statistical computation on manifolds, like the definition of mean shape of set of shapes, covariances and explicit statistical models on the tangent space of a sample mean (13) to model the class and the variability within the class are suitable for shapes classification and clustering. The manifolds and the groups acting on them are defined according to the desired application and to the available data. For instance, to handle action recognition problem using skeleton data, the landmarks issued from the skeleton are modeled on Kendall shape space where the comparison is invariant to scale, translation and rotation. Thus, the geodesic distance in Kendall space measures the difference in shape and the intrinsic means computed represent the mean of shapes. The main contribution performed on this manifold is the intrinsic sparse coding of 3D human skeleton (respectively 2D facial landmarks) that consider the geometry of the underlying space and avoid distorsion while projecting on the tangent spaces. The key idea is to code each 3D human skeleton (respectively 2D facial landmarks) on its tangent space once the dictionary elements are calculated on the manifold. The proposed methods are competitive or outperform the state-of-the-art on various and challenging datasets in two recognition tasks: 3D action recognition based on skeleton data, 2D micro and macro facial expression recognition based on facial landmarks.

As 3D face as concerned, the 3D face is parametrized by facial curves that are modeled on an infinite dimension manifold and riemannian geometry tools have been used to 3D face analysis through curve shape analysis. The same undesired transformations have been removed in addition to curve-reparameterization. The features used to classify soft biometric using 3D faces are the shooting vector

along a geodesic path (constant speed curves on the manifold) between shapes as it captures effectively the deformations between them, through accurate registration. This feature has been used as spatial representation of the 3D face to recognize soft-biometric caracterisctics of the 3D face like the gender. Moreover, it has been coupled with a temporal modeling to handle 4D facial expression recognition, the results presented are competitive with state-of-the-art. Finally, a framework for more complicated objects is presented, that surfaces parametrized by a disk or a sphere are modeled on a more complicated manifold and present the algorithms to calculate geodesic paths, distances and intrinsic means. Two strategies have been adopted on this manifold, the first one is to filter the undesirable transformation groups (rotation, translation, scale and surface reparameterization) then perform computation on quotient spaces. This is an extension of the pre-presented frameworks on landmarks and curves manifolds. The second strategy proposed for surface analysis is more complicated theoretically and is based on a gauge framework capable to compute the geodesic paths, intrinsic means of surfaces on shape space without any need to filter the re-parameterization group.

In summary, the main contribution presented in this *habilitation* is a unified computational framework for human behavior analysis through multiple manifolds according to the kind of data provided, with different applications ranging from action recognition to soft-biometrics estimation including facial expression analysis and classification.

Moreover, a comprehensive study of the intrinsic approaches versus extrinsic ones is presented in this *habilitation*. The extrinsic approaches use kernel to handle the non linearity of data. This study has been conducted only for the landmarks data on Kendall shape space. The first paradigm (intrinsic approach) allows the calculus on the manifold tangent spaces. The second paradigm performs calculus in Hilbert space after mapping the manifold-valued data using a kernel embedding which gives a richer representation of the data and helps identifying complex patterns. These two paradigms have been evaluated and compared in the context of action and expression recognition based on 2D and 3D landmarks. The main conclusion of this comparative study leads to the following statement: the extrinsic approach is more efficient to represent 2D trajectories while the intrinsic one is more suitable for 3D trajectories.

In my future research, I will continue this path on the theoretical and application plan while exploring new approaches. It would also be interesting to take advantage of the power of new learning techniques such as deep learning techniques and combine them with geometric approaches. A particular intention will be brought to the analysis of forms of faces and human bodies. Many final applications could be targeted with the developed tools and strategies, in the future. For example motion-controlled gaming to improve kids life by limiting their disables and attract (force) them to communicate with others by sharing space, style motion analysis, etc.

6.1 Geometry-aware Deep Learning

Even deep learning models have impressively surpassed conventional models while assuming an underlying Euclidean structure of the space of the data, this assumption may not always be valid for images and/ or data extracted from images. Actually, the topology of a given space characterizes the proximity between data and plays a vital role in pattern recognition. Pattern analysis takes place in the context of data lying in some inherent geometrical structure. Simply ignoring the geometrical aspect, or naively treating the space as Euclidean, may cause undesired effect. I propose to consider the geometry of the data while designing deep architecture in my future research.

6.1.1 Shape-GAN for motion generation

First, I am interested by the GAN networks due to their emergence in computer vision. For instance, it can serve as a data augmentation tool which remarkably relieves the burden of manual annotations and thereby contributes to the development of various video understanding tasks such as action and activity recognition. On the other hand, human video synthesis allows for many human-centric applications such as avatar animation. Problems related to realistic image generation with GANs has already shown impressive results (243–245). However, the extension from generating images to generating videos turns out to be a highly challenging task with the introduction of the temporal dimension. Considering this fact, a typical generative model needs to learn the plausible physical motion models of objects in addition to learning their appearance models. To this end, some approaches in the literature opted for a disentangled solution by first generating plausible motion then synthesizing coherent appearances. For instance, Yang *et al.* (246) proposed a two stage approach. In the first stage, pose sequences are used to learn a generative model due to their ability to encode motion dynamics. The newly generated pose sequences are then used to guide the generation of video frames while preserving coherent appearances in the input image. Regarding the first stage, they proposed a pose sequence GAN (PSGAN) to generate skeletal sequences. Their generator transforms an input pose into a pose sequence by adopting a encoder-decoder architecture. The output of the decoder is then fed into a LSTM module for temporal pose modeling. However, their resulting sequences may contain corrupted poses which would affect the synthesis of coherent appearance in the second stage. One can follow the same disentangled solution. In the first stage, an encoder-GAN model which generates new samples in the encoder latent space which are then transformed into skeletal sequences can be designed. In contrast to previous works, it guarantees the shape and motion consistencies of the generated pose sequences while having a simpler architecture. In the second stage, given an input image, each pose of a sequence can be transformed to an image, thereby constitutes the final video. The latter procedure can use a recent generative model that learns to transfer a person image to new poses (247).

Given an input video presenting a human action, one can first extract a skeleton from each video frame using a state-of-the-art detector (27) where each skeleton is represented by 18 body landmarks in 2D.

Then, training sequences will be projected to the shape manifold. The resulting trajectories are encoded using SCDL in the shape manifold framework propsoed in chapter 3. The obtained latent samples can the used to train a GAN. This allows to generate new samples which are then transformed to skeletal sequences in the shape manifold using the weighted intrinsic mean algorithm. This reconstruction procedure will be performed with respect to the pre-trained dictionary which insures that the obtained samples lie in the space of skeletons avoiding any noisy or corrupted poses.

6.1.2 Towards Deep learning on Shape spaces

The targeted shape-GAN framework presented in section 6.1.1 aims to generate skeletons on the manifold thanks to the sparse coding framework. A regular GAN will be applied on the linear sparse codes and the geometry of the generated samples can be preserved during the geometric reconstruction of the sparse codes generated using the GAN. Furthermore, it will be interesting to design a deep architecture on the manifold itself. Some previous approaches have proposed deep architectures on non linear spaces such as Lie group, grassmann manifold, SPD (248–250) however no deep architectures are proposed on shape spaces. A first challenging aspect is the extension of the convolution operator on Riemannian manifolds in order to design deep convolutional neural network (CNN)-based architectures. A more complicated issue is to consider the quotient spaces resulting on filtering out the rotations (and re-parameterizations for some manifolds). Finally, it will be interesting have a more-in-depth study and to establish the properties of the learning procedure on the Riemannian manifold of interest including stability, convergence, impact of the Riemannian metric, etc. examples of ground truth correspondences between exemplar shapes.

6.2 Facial emotion recognition in Adverse Conditions

The proposed face analysis frameworks presented in this *habilitation* have been evaluated on datasets that were collected in controlled environments and do not present considerable view-variations. Hence, I would like to extend these approaches on more challenging datasets such as the AFEW database (251) where the data were collected from movies showing close-to-real-world conditions, which simulates the spontaneous expressions in uncontrolled environment. In addition, the view variations for 2D landmarks yield projective transformations which are more complex to filter out, thereby it would be interesting to investigate this problem. An interesting strategy to handle the spontaneous expressions is to design a micro-to-macro expression translation in order capture the subtle deformations. Unsupervised setting has to be considered for that.

6.2.1 Body movement for emotion recognition

The use of facial expression algorithms in real-world applications is difficult, when the user conveys spontaneous emotions involving, in general, the all his body. I propose in this challenging task to explore the contribution of the the dynamics of body parts as proposed in (252–254) to perform emotion classification. For instance, human gait conveys affects similarly to voice or face expressions (255, 256). Several studies provide valuable information, particularly for human psychology understanding and for human-machine interaction applications. Literature on the ability of human to recognize individuals from motion is abundant, in particular, it has been shown that one can recognize a known person or even him/herself accurately from gait data. Most of the gait studies make use of parameters such as the stance phase, the gait cycle frequency, the length of the footsteps that can be easily measured. Nevertheless, gait is radically changed with the affect. Whether one feels fearful, happy, sad, angry, normal or a mixture of these basic emotions, gait is modified (257, 258). Intuitively, one is able to recognize these types of conveyed emotions.

6.2.2 Towards Extrinsic analysis of 3D shape for subtle expression recognition

Recall that instead of performing calculus on tangent spaces, the extrinsic approach tends to embed the manifold-valued data into Hilbert spaces which are higher dimensional vector spaces where linear calculus becomes possible. The main difficulty here arises from the fact that this embedding relies on a kernel function which, according to Mercer's theorem, should be positive definite. **Kernel methods** attempt to compute similarities between data-points (*e.g.* landmark configurations, features extracted from them, etc.) using kernel functions. This enable them to operate in a high-dimensional, implicit feature space. The latter is also called the inner product space since only inner products between data-points are computed, without ever computing the coordinates of the data in the new space. These approaches are known to bring a richer representation of the original data since the inner product space is usually higher-dimensional which helps classification methods to identify complex patterns. In chapter 3 a comprehensive comparison of kernel methods versus intrinsic methods on the Kendall shape space of 2D landmarks for expression recognition is presented. Extrinsic and intrinsic sparse coding have been performed and compared on 2D landmarks data (issued from facial landmarks). It has been demonstrated that the proposed extrinsic approach performs better than the intrinsic on micro-expression. This motivates me to investigate that direction to handle subtle deformation issued from spontaneous expressions acquired in adverse conditions. Therefore, the presented extrinsic approach to code shape trajectories is only suitable for the 2D shape manifold where a positive definite kernel exists in the literature. This is not the case for the 3D shape manifold where our extension of the Procrustes Gaussian kernel is not always valid. Thus, I would like to further study the existence of positive definite kernels in the shape manifold of 3D landmarks. Moreover, the extrinsic approach has been only considered for manifold of sparse representations (landmarks). An interesting direction will be to investigate this direction for manifolds of dense representations of data (curves and surfaces).

6.3 Functional Analysis on manifold

With the advances in algorithmic solutions for skeletal data estimation in video streams – infrared (IR) watching retro-reflective markers, color video streams (259, 260) or RGB-D (Depth) (28) – more and more shape-related data are being recorded in real-time with high temporal resolution. They belong to a second generation of "functional data" (261) of high potential in Computer Vision applications (32). This kind of data is nowadays preferred in a set of problems, in particular *Human Analytics* (e.g. action and activity recognition, emotional state classification, pedestrian behavior understanding, and 3D gait recognition, to cite a few), for different reasons, (1) they are independent of the scene background and are robust to the changes in illumination conditions; (2) they allow robustness to pose variations and allow to handle the inherent camera projection transformations; (3) compared to video streams, they represent compact amount of data suitable in real-time processing. However, they are highly dependent on the sensor's reliability (e.g. in depth estimation) and body joints estimation accuracy. Several space-time representations have been recently proposed for the purpose of classification, clustering, detection and prediction of human behaviors. Among them, time-parameterized trajectories on Riemannian spaces (Lie groups, Kendall shape space, Grassmann manifolds, manifold of SPD matrices of fixed rank) seems to be a good choice. In (37), the authors proposed to represent skeletal motions as trajectories in the Special Euclidean group $SE(3)^n$, and later on $SO(3)^n$ (262), both Lie Groups. These representations are then mapped into the correspondent Lie algebra $\mathfrak{se}(3)^n$, respectively $\mathfrak{so}(3)^n$, the tangent spaces attached to the Lie groups at the identity elements, where they are processed and classified (n is related to the number of body joints). By exploiting the same representation on Lie Groups, Anirudh et al. (263) used the framework of Transported Square-Root Velocity Fields (T-SRVF) (264) to encode motion trajectories in $SO(3)^n$. They extended existing coding methods such as PCA, KSVD, and Label Consistent KSVD to the Riemannian action trajectories. Grounding on the T-SRVF framework also, Ben Amor et al. have proposed another alternative by extending Kendall's shape theory to trajectories (32). Here, trajectories are transported to a reference tangent space attached to the shape space at a reference point, then analyzed under the T-SRVF formulation. In particular, an elastic metric to compare shape trajectories and a geometric toolbox for denoising, smoothing, averaging and resampling trajectories have been proposed. The drawback of these approaches is that the parallel translation to a common tangent space often introduces distortions, in particular when the reference point is far from the data to be analyzed (e.g. Lie Algebra), or inversely. To avoid this problem, an intrinsic sparse coding and dictionary learning (SCDL) formulation were proposed in chapter 3. The intrinsic coding allows to encode more locally human shapes (i.e. around predefined dictionary atoms). So, initial trajectories give rise to sparse code time-series with suitable computational properties, including the sparsity and the linearity. Taking another direction, Kacem et al. have proposed in (102) to map skeletal data into trajectories of Grammian matrices in the cone of positive semidefinite matrices of fixed-rank. A Gram matrix summarizes all pairwise distances between the joints and is by construction rotational invariant. They adopted the pairwise proximity function SVM (ppf-SVM) grounding on a geometry-aware similarity measure defined on the space of interest for trajectory classification.

From the review made above, it is well established that the trajectory representation on Riemannian spaces is suitable in several *Human Analytics* tasks. However, existing approaches omitted the *functional* nature of the trajectories in hands (high-dimensional or infinite-dimensional). The scientific branch of *Functional Data Analysis (FDA)* deals with data of such kind. It often requires that the functions, subject of the analysis, are element of a Hilbert space $\mathbb{L}^2(\mathcal{T})$, i.e. the set of all functions f such that the integral of f^2 over the compact time domain $\mathcal{T}$ is finite. In particular, the functional Principal Component Analysis (fPCA) is an efficient dimension reduction technique which could be applied on realizations of any stochastic process, also assumed to be in a Hilbert space of a compact time domain $\mathcal{T}$. fPCA transforms initial infinite- or high-dimensional sample functions to a small set of uncorrelated variables with respect to a set of modes of variations defined around a mean function (261). While its theory and implementations are well formulated for data lying to Hilbert spaces, only few works started tackling its extension to curved spaces (265–267). The applicability of fPCA on the shape space can be interesting research direction with application on 3D gait analysis and recognition based on skeleton data. One can build, on top of the shape trajectory representation introduced in (32), a comprehensive functional PCA framework on the Kendall's shape space. In literature, the extension of fPCA to curved spaces is still at the beginning (265, 268). From a geometric perspective, the targeted formulation has to accounts for the *Spherical* structure of the pre-shape space and the *Orbifold* geometry of the shape space. To our knowledge, the analysis using fPCA of this kind of data is completely novel in literature. Kendall fPCA will transform submanifolds of shape trajectories to latent spaces, an interesting solution to the dimensionality curse problem. That is, Kendall's fPCA approximates initial high-dimensional shape trajectories to a finite set of uncorrelated variables.

Bibliography

[1] G. V. Kale and V. H. Patil, "A study of vision based human motion recognition and analysis," *IJACI*, vol. 7, no. 2, pp. 75–92, 2016. [Online]. Available: https://doi.org/10.4018/IJACI.2016070104

[2] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1114–1127, 2008.

[3] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, 2005. [Online]. Available: https://doi.org/10.1109/TPAMI.2005.39

[4] M. Ronchetti and M. Avancini, "Using kinect to emulate an interactive whiteboard," *MS in Computer Science. University of Trento.*

[5] A. Sepehri, Y. Yacoob, and L. S. Davis, "Employing the hand as an interface device," *Journal of Multimedia*, vol. 1, no. 7, pp. 18–29, 2006. [Online]. Available: https://doi.org/10.4304/jmm.1.7.18-29

[6] M. Hoffmann, H. Marques, A. Arieta, H. Sumioka, M. Lungarella, and R. Pfeifer, "Body schema in robotics: A review," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 4, pp. 304–324, 2010.

[7] B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C. J. Bula, and P. Robert, "Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 6, pp. 711–723, 2003.

[8] J. F. Lin and D. Kulić, "Online segmentation of human motion for automated rehabilitation exercise analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 168–180, 2014.

[9] T. Watanabe, N. Ohtsuka, S. Shibusawa, M. Kamada, and T. Yonekura, "Design of lower limb chair exercise support system with depth sensor," in *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing*

and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops, 2014, pp. 104–111.

[10] T.-T. Dao, H. Tannous, P. Pouletaut, D. Gamet, D. Istrate, and M. H. B. Tho, "Interactive and connected rehabilitation systems for e-health," *Irbm*, vol. 37, no. 5-6, pp. 289–296, 2016.

[11] G. Kale and V. Patil, "Bharatna yam adavu recognition from depth data," in *2015 Third International Conference on Image Information Processing (ICIIP)*, 2015, pp. 246–251.

[12] D. G. Kendall, "Shape manifolds, Procrustean metrics, and complex projective spaces," *Bulletin of the London Mathematical Society*, vol. 16, no. 2, pp. 81–121, 1984.

[13] I. L. Dryden and K. Mardia, *Statistical shape analysis*, ser. Wiley series in probability and statistics: Probability and statistics. J. Wiley, 1998.

[14] F. L. Bookstein, "Size and shape spaces for landmark data in two dimensions," *Statistical Science*, vol. 1, no. 2, pp. 181–222, 05 1986.

[15] U. Grenander, Y. Chow, and D. Keenan, *Hands: A Pattern Theoretic Study of Biological Shapes*, ser. Research notes in neural computing. Springer-Verlag, 1991. [Online]. Available: http://books.google.com/books?id=6nRtQgAACAAJ

[16] L. Younes, "Computable elastic distance between shapes," *SIAM Journal of Applied Mathematics*, vol. 58, pp. 565–586, 1998.

[17] E. Klassen, A. Srivastava, W. Mio, and S. Joshi, "Analysis of planar shapes using geodesic paths on shape spaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 3, pp. 372–383, March 2004.

[18] S. H. Joshi, E. Klassen, A. Srivastava, and I. Jermyn, "A novel representation for riemannian analysis of elastic curves in r^n," in *CVPR*, 2007.

[19] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1415–1428, 2011.

[20] H. Drira, B. Ben Amor, M. Daoudi, and A. Srivastava, "Pose and expression-invariant 3D face recognition using elastic radial curves," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2010, pp. 1–11, doi:10.5244/C.24.90.

[21] H. Drira, B. A. Boulbaba, S. Anuj, M. Daoudi, and R. Slama, "3D Face Recognition Under Expressions,Occlusions and Pose Variations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, to appear.

[22] I. H. Jermyn, S. Kurtek, E. Klassen, and A. Srivastava, "Elastic shape matching of parameterized surfaces using square root normal fields," in *European Conference on Computer Vision*, 2012, pp. 804–817.

[23] C. Samir, S. Kurtek, A. Srivastava, and M. Canis, "Elastic shape analysis of cylindrical surfaces for 3D/2D registration in endometrial tissue characterization," *IEEE Trans. Medical Imaging*, vol. 33, no. 5, pp. 1035–1043, 2014.

[24] A. Ben Tanfous, H. Drira, and B. Ben Amor, "Sparse coding of shape trajectories for facial expression and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[25] M. Meng, H. Drira, and J. Boonaert, "Distances evolution analysis for online and off-line human object interaction recognition," *Image Vis. Comput.*, vol. 70, pp. 32–45, 2018. [Online]. Available: https://doi.org/10.1016/j.imavis.2017.12.003

[26] A. B. Tanfous, H. Drira, and B. B. Amor, "Coding kendall's shape trajectories for 3d action recognition," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 2840–2849. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Tanfous_Coding_Kendalls_Shape_CVPR_2018_paper.html

[27] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[28] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1297–1304.

[29] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1859–1866.

[30] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[31] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 532–539.

[32] B. Ben Amor, J. Su, and A. Srivastava, "Action recognition using rate-invariant analysis of skeletal shape trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 1–13, 2016.

[33] D. Bryner, E. Klassen, H. Le, and A. Srivastava, "2d affine and projective shape analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 998–1011, 2014.

[34] H. E. Cetingül and R. Vidal, "Sparse riemannian manifold clustering for hardi segmentation," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. IEEE, 2011, pp. 1750–1753.

[35] H. E. Cetingul and R. Vidal, "Intrinsic mean shift for clustering on stiefel and grassmann manifolds," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1896–1902.

[36] J. Ho, Y. Xie, and B. Vemuri, "On a nonlinear generalization of sparse coding and dictionary learning," in *International conference on machine learning*, 2013, pp. 1480–1488.

[37] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[38] M. Harandi, R. Hartley, C. Shen, B. Lovell, and C. Sanderson, "Extrinsic methods for coding and dictionary learning on grassmann manifolds," *International Journal of Computer Vision*, vol. 114, no. 2-3, pp. 113–136, 2015.

[39] M. T. Harandi, C. Sanderson, R. I. Hartley, and B. C. Lovell, "Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach," *CoRR*, vol. abs/1304.4344, 2013. [Online]. Available: http://arxiv.org/abs/1304.4344

[40] S. Jayasumana, M. Salzmann, H. Li, and M. Harandi, "A framework for shape analysis via hilbert space embedding," in *IEEE ICCV*, 2013, pp. 1249–1256.

[41] P. Li, Q. Wang, W. Zuo, and L. Zhang, "Log-euclidean kernels for sparse representation and dictionary learning," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 1601–1608.

[42] R. Anirudh, P. Turaga, J. Su, and A. Srivastava, "Elastic functional coding of human actions: From vector-fields to latent variables," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3147–3155.

[43] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, ser. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press, Dec. 2002.

[44] A. Cherian and S. Sra, "Riemannian dictionary learning and sparse coding for positive definite matrices," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–13, 2017.

[45] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[46] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," MINNESOTA UNIV MINNEAPOLIS INST FOR MATHEMATICS AND ITS APPLICATIONS, Tech. Rep., 2008.

[47] J. Shi, X. Ren, G. Dai, J. Wang, and Z. Zhang, "A non-convex relaxation approach to sparse dictionary learning," in *CVPR 2011*. IEEE, 2011, pp. 1809–1816.

[48] B. Liu, J. Huang, L. Yang, and C. Kulikowsk, "Robust tracking using local sparse appearance model and k-selection," in *CVPR 2011*. IEEE, 2011, pp. 1313–1320.

[49] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2691–2698.

[50] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 5. IEEE, 1999, pp. 2443–2446.

[51] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[52] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[53] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Kernel methods on riemannian manifolds with gaussian rbf kernels," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 12, pp. 2464–2477, 2015.

[54] W. M. Boothby, *An introduction to differentiable manifolds and Riemannian geometry*. Academic press, 1986, vol. 120.

[55] Y. M. Lui, "Advances in matrix manifolds for computer vision," *Image Vision Comput.*, vol. 30, no. 6-7, pp. 380–388, Jun. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2011.08.002

[56] I. Dryden and K. Mardia, *Statistical shape analysis*. Wiley, 1998.

[57] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Comm. Pure Appl. Math.*, vol. 30, no. 5, pp. 509–541, Sep. 1977. [Online]. Available: http://dx.doi.org/10.1002/cpa.3160300502

[58] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2479–2494, June 2013.

[59] M. Harandi and M. Salzmann, "Riemannian coding and dictionary learning: Kernels to the rescue," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3926–3935.

[60] H. E. Çetingül, M. J. Wright, P. M. Thompson, and R. Vidal, "Segmentation of high angular resolution diffusion mri using sparse riemannian manifold clustering," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 301–317, Feb 2014.

[61] C. Yuan, W. Hu, X. Li, S. Maybank, and G. Luo, *Human Action Recognition under Log-Euclidean Riemannian Metric*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 343–353.

[62] M. T. Harandi, R. Hartley, B. Lovell, and C. Sanderson, "Sparse coding on symmetric positive definite manifolds using bregman divergences," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 6, pp. 1294–1306, 2016.

[63] R. Vemulapalli and R. Chellapa, "Rolling rotations for recognizing human actions from 3d skeletal data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4471–4479.

[64] 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*, 2013.

[65] S. Taheri, P. Turaga, and R. Chellappa, "Towards view-invariant expression analysis using analytic shape manifolds," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 306–313.

[66] A. Kacem, M. Daoudi, B. Ben Amor, and J. Carlos Alvarez-Paiva, "A novel space-time representation on the positive semidefinite cone for facial expression recognition," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[67] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[68] Z. Zhang, D. Pati, and A. Srivastava, "Bayesian clustering of shapes of curves," *Journal of Statistical Planning and Inference*, vol. 166, pp. 171 – 186, 2015, special Issue on Bayesian Nonparametrics. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378375815000798

[69] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE transactions on medical imaging*, vol. 23, no. 8, pp. 995–1005, 2004.

[70] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *COLT/EuroCOLT*, 2001.

[71] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *IEEE Inter. Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB)*, 2010, p. 9–14.

[72] G. Garcia-Hernando and T.-K. Kim, "Transition forests: Learning discriminative temporal transitions for action recognition and detection."

[73] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang, "Graph based skeleton motion representation and similarity measurement for action recognition," in *European Conference on Computer Vision*. Springer, 2016.

[74] P. Koniusz, A. Cherian, and F. Porikli, "Tensor representations via kernel linearization for action recognition from 3d skeletons," in *European Conference on Computer Vision*. Springer, 2016, pp. 37–53.

[75] Z. Huang, C. Wan, T. Probst, and L. Van Gool, "Deep learning on lie groups for skeleton-based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6099–6108.

[76] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1110–1118.

[77] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833.

[78] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, "Intraface," 05 2015.

[79] Z. Wang, S. Wang, and Q. Ji, "Capturing complex spatio-temporal relations among facial muscles for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3422–3429.

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

[81] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 2983–2991.

[82] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. PP, 03 2017.

[83] Y.-H. Oh, J. See, A. C. Le Ngo, R. C.-W. Phan, and V. M. Baskaran, "A survey of automatic facial micro-expression analysis: Databases, methods and challenges," *Frontiers in psychology*, vol. 9, p. 1128, 2018.

[84] H. Zheng, X. Geng, and Z. Yang, "A relaxed k-svd algorithm for spontaneous micro-expression recognition," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2016, pp. 692–699.

[85] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[86] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 382–386.

[87] D. Y. Choi, D. H. Kim, and B. C. Song, "Recognizing fine facial micro-expressions using two-dimensional landmark feature," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1962–1966.

[88] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1290–1297.

[89] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 129–136.

[90] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[91] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[92] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[93] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[94] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2117–2126.

[95] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[96] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, 2012.

[97] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013, Portland, OR, USA, June 23-28, 2013*, 2013, pp. 479–485. [Online]. Available: https://doi.org/10.1109/CVPRW.2013.77

[98] L. Xia, C.-C. Chen, and J. 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*. IEEE, 2012, pp. 20–27.

[99] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[100] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part II*, 2012, pp. 872–885.

[101] C. Wang, Y. Wang, and A. L. Yuille, "Mining 3d key-pose-motifs for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2639–2647.

[102] A. Kacem, M. Daoudi, B. B. Amor, S. Berretti, and J. C. Alvarez-Paiva, "A novel geometric framework on gram matrix trajectories for human behavior understanding," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[103] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2017, pp. 148–157.

[104] L. L. Presti and M. L. Cascia, "3d skeleton-based human action classification: A survey," *Pattern Recognition*, vol. 53, no. Supplement C, pp. 130 – 147, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320315004392

[105] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4041–4049.

[106] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[107] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1623–1631.

[108] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[109] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[110] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons, 2016.

[111] S. Elaiwat, M. Bennamoun, and F. Boussaid, "A spatio-temporal rbm-based model for facial expression recognition," *Pattern Recognition*, vol. 49, pp. 152–161, 2016.

[112] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1749–1756.

[113] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 2562–2569.

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

[115] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 2010, p. 65.

[116] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, p. e86041, 2014.

[117] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, 2016.

[118] S.-T. Liong, J. See, R. C.-W. Phan, Y.-H. Oh, A. C. Le Ngo, K. Wong, and S.-W. Tan, "Spontaneous subtle expression detection and recognition based on facial strain," *Signal Processing: Image Communication*, vol. 47, pp. 170–182, 2016.

[119] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[120] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[121] B. B. Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, "4-d facial expression recognition by learning geometric deformations," *IEEE Trans. Cybernetics*, vol. 44, no. 12, pp. 2443–2457, 2014. [Online]. Available: https://doi.org/10.1109/TCYB.2014.2308091

[122] Q. Zhen, D. Huang, H. Drira, B. B. Amor, Y. Wang, and M. Daoudi, "Magnifying subtle facial motions for effective 4d expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 524–536, 2019.

[123] B. Xia, B. B. Amor, H. Drira, M. Daoudi, and L. Ballihi, "Combining face averageness and symmetry for 3d-based gender classification," *Pattern Recognit.*, vol. 48, no. 3, pp. 746–758, 2015. [Online]. Available: https://doi.org/10.1016/j.patcog.2014.09.021

[124] B. Vicki, B. A. Mike, H. Elias, H. Pat, M. Oli, and C. Anne, "Sex discrimination: How do we tell the difference between male and female faces?" in *Perception*, vol. 22(2), 1993, pp. 131–152.

[125] Z. Ziqing, L. Douglas, B. Stacey, R. Raymond, and S. Ronald, "Facial anthropometric differences among gender, ethnicity, and age groups," vol. 54, no. 4, pp. 391–402, 2010.

[126] A. Mehrabian and M. Wiener, "Decoding of inconsistent communications," *Journal of Personality and Social Psychology*, vol. 6, no. 1, pp. 109–114, May 1967.

[127] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Proc. Nebraska Symposium on Motivation*, vol. 19, Lincoln, NE, 1972, pp. 207–283.

[128] P. Ekman and W. V. Friesen, *Manual for the the Facial Action Coding System.* Palo Alto, CA: Consulting Psychologist Press, 1977.

[129] H. Drira, B. Ben Amor, M. Daoudi, A. Srivastava, and S. Berretti, "3d dynamic expression recognition based on a novel deformation vector field and random forest," in *21st International Conference on Pattern Recognition*, 2012.

[130] H. Drira, B. Ben Amor, M. Daoudi, and A. Srivastava, "Pose and expression-invariant 3D face recognition using elastic radial curves," in *Proc. British Machine Vision Conference*, Aberystwyth, UK, August 2010, pp. 1–11.

[131] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1415–1428, 2011.

[132] S. Joshi, E. Klassen, A. Srivastava, and I. Jermyn, "A novel representation for Riemannian analysis of elastic curves in $\mathbb{R}^n$," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, Jun. 2007, pp. 1063–6919.

[133] M. G. Rhodes, "Age estimation of faces: a review," in *Appl. Cognit. Psychol*, vol. 23, 2009, pp. 1–12.

[134] N. Ramanathan, R. Chellappa, and S. Biswas, "Computational methods for modeling facial aging: A survey," in *Journal of Visual Languages & Computing*, vol. 20, no. 3. Elsevier, 2009, pp. 131–144.

[135] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," in *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,*, vol. 34, no. 1, 2004, pp. 621–628.

[136] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, 2002, pp. 442–455.

[137] G. Guo, Y. Fu, T. S. Huang, and C. R. Dyer, "Locally adjusted robust regression for human age estimation," in *IEEE Workshop on Applications of Computer Vision, 2008. WACV 2008*, 2008, pp. 1–6.

[138] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," vol. 17, no. 7, pp. 1178–1188, 2008.

[139] T. Wu, P. Turaga, and R. Chellappa, "Age estimation and face verification across aging using landmarks," vol. 7, no. 6, 2012, pp. 1780–1788.

[140] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning ordinal discriminative features for age estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2570–2577.

[141] G. Guodong, M. Guowang, F. Yun, and H. T. S, "Human age estimation using bio-inspired features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 112–119.

[142] G. Xin, Z. Zhi-Hua, and S.-M. Kate, "Automatic age estimation based on facial aging patterns," vol. 29, no. 12, pp. 2234–2240, 2007.

[143] G. Xin, Z. Zhi-Hua, Z. Yu, L. Gang, and D. Honghua, "Learning from facial aging patterns for automatic age estimation," in *ACM international conference on Multimedia*, 2006, pp. 307–316.

[144] S. Jinli, Z. Song-Chun, S. Shiguang, and C. Xilin, "A compositional and dynamic model for face aging," vol. 32, no. 3, 2010, pp. 385–401.

[145] N. Lakshmiprabha, J. Bhattacharya, and S. Majumder, "Age estimation using gender information," in *Computer Networks and Intelligent Computing*, 2011, pp. 211–216.

[146] U. Kazuya, S. Masashi, and I. Yasuyuki, "Perceived age estimation under lighting condition change by covariate shift adaptation," in *International Conference on Pattern Recognition (ICPR)*, 2010, pp. 3400–3403.

[147] B. Xia, B. B. Amor, M. Daoudi, and H. Drira, "Can 3d shape of the face reveal your age?" in *International Conference on Computer Vision Theory and Applications*, 2014.

[148] F. C. Farkas LG, Katic MJ, "International anthropometric study of facial morphology in various ethnic groups/races," *Journal of Craniofacial Surgery*, vol. 16, no. 4, pp. 615–646, 2005.

[149]

[150] Y. Hu, J. Yan, and P. Shi, "A fusion-based method for 3D facial gender classification," in *Computer and Automation Engineering (ICCAE)*, vol. 5, 2010, pp. 369–372.

[151] X. Han, H. Ugail, and I. Palmer, "Gender classification based on 3D face geometry features using svm," in *CyberWorlds*, 2009, pp. 114–118.

[152] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," in *Pattern Analysis and Machine Intelligence*, vol. 33, 2008, pp. 1962 –1977.

[153] C. Wang, D. Huang, Y. Wang, and G. Zhang, "Facial image-based gender classification using local circular patterns," in *International Conference on Pattern Recognition*, 11 2012.

[154] E. Makinen and R. Raisamo, "An experimental comparison of gender classification methods," in *Pattern Recognition Letters*, vol. 29, 2008, pp. 1544–1556.

[155] Y. Liu and J. Palmer, "A quantified study of facial asymmetry in 3D faces," in *Analysis and Modeling of Faces and Gestures*, 2003, pp. 222–229.

[156] W. Yang, C. Chen, K. Ricanek, and C. Sun, "Gender classification via global-local features fusion," in *Biometric Recognition*, vol. 7098, 2011, pp. 214–220.

[157] C. Shan, "Learning local binary patterns for gender classification on real-world face images," in *Pattern Recognition Letters*, vol. 33, 2012, pp. 431–437.

[158] B. Xia, B. B. Amor, D. Huang, M. Daoudi, Y. Wang, and H. Drira, "Enhancing gender classification by combining 3d and 2d face modalities," in *European Signal Processing Conference (EUSIPCO)*, 2013.

[159] G. Toderici, S. O'Malley, G. Passalis, T. Theoharis, and I. Kakadiaris, "Ethnicity- and gender-based subject retrieval using 3-D face-recognition techniques," in *International Journal of Computer Vision*, vol. 89, 2010, pp. 382–391.

[160]

[161] T. Huynh, R. Min, and J.-L. Dugelay, "An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data," in *ACCV 2012, Workshop on Computer Vision with Local Binary Pattern Variants*, 2012.

[162] J. Wu, W. Smith, and E. Hancock, "Gender classification using shape from shading," in *International Conference on Image Analysis and Recognition*, 2007, pp. 499–508.

[163] B. Xia, B. B. Amor, H. Drira, M. Daoudi, and L. Ballihi, "Gender and 3D facial symmetry: What's the relationship?" in *IEEE Conference on Automatic Face and Gesture Recognition*, 2013.

[164] J. Alphonse, J. Cox, J. Clarke, P. Schluter, and A. McLennan, "The effect of ethnicity on 2d and 3d frontomaxillary facial angle measurement in the first trimester," in *Obstetrics and Gynecology International*, 2013.

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

[166] I. A. Kakadiaris, G. Passalis, G. Toderici, N. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis, "Three-dimensional face recognition in the presence of facial expressions: An annotated deformable approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 640–649, April 2007.

[167] A. S. Mian, M. Bennamoun, and R. Owens, "Keypoint detection and local feature matching for textured 3D face recognition," *Int. Journal of Computer Vision*, vol. 79, no. 1, pp. 1–12, Aug. 2008.

[168] C. Samir, A. Srivastava, M. Daoudi, and E. Klassen, "An intrinsic framework for analysis of facial surfaces," *Int. Journal of Computer Vision*, vol. 82, no. 1, pp. 80–95, April 2009.

[169] S. Berretti, A. Del Bimbo, and P. Pala, "3D face recognition using iso-geodesic stripes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2162–2177, Dec. 2010.

[170] Y. Wang, J. Liu, and X. Tang, "Robust 3D face recognition by local shape difference boosting," *IEEE Transactions on Pattern Analysis and Machine Untelligence*, vol. 32, no. 10, pp. 1858–1870, Oct. 2010.

[171] A. Maalej, B. Ben Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Shape analysis of local facial patches for 3D facial expression recognition," *Pattern Recognition*, vol. 44, no. 8, pp. 1581–1589, Aug. 2011.

[172] S. Berretti, B. B. Amor, M. Daoudi, and A. D. Bimbo, "3d facial expression recognition using sift descriptors of automatically detected keypoints," *The Visual Computer*, vol. 27, no. 11, pp. 1021–1036, 2011.

[173] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Southampton, UK, Apr. 2006, pp. 211–216.

[174] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Proc. First COST 2101 Workshop on Biometrics and Identity Management*, May 2008.

[175] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition (FGR08)*, Amsterdam, The Netherlands, Sep. 2008, pp. 1–6.

[176] B. Matuszewski, W. Quan, and L.-K. Shark, "High-resolution comprehensive 3-D dynamic database for facial articulation analysis," in *Proc. IEEE Int. Conf. on Computer Vision Workshops*, Barcelona, Spain, Nov. 2011, pp. 2128–2135.

[177] B. J. Matuszewski, W. Quan, L.-K. Shark, A. S. McLoughlin, C. E. Lightbody, H. C. Emsley, and C. L. Watkins, "Hi4D-ADSIP 3-D dynamic facial articulation database," *Image and Vision Computing*, vol. 30, no. 10, pp. 713–727, 2012 2012.

[178] D. Cosker, E. Krumhuber, and A. Hilton, "A FACS valid 3d dynamic action unit database with applications to 3D dynamic morphable facial modeling," in *Proc. IEEE Int. Conf, on Computer Vision*, Barcelona, Spain, Nov. 2011, pp. 2296–2303.

[179] S. Berretti, A. Del Bimbo, and P. Pala, "Superfaces: A super-resolution model for 3D faces," in *Proc. of Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment*, Florence, Italy, Oct. 2012, pp. 73–82.

[180] Y. Li, A. Mian, W. Lu, and A. Krishna, "Using kinect for face recognition under varying poses, expressions, illumination and disguise," in *Proc. IEEE Workshop on Applications of Computer Vision*, Tampa, FL, Jan. 2013, pp. 186–192.

[181] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.

[182] L. Benedikt, V. Kajic, D. Cosker, P. Rosin, and D. Marshall, "Facial dynamics in biometric identification," in *Proc. British Machine Vision Conf.*, Leeds, UK, Sep. 2008, pp. 1–10.

[183] L. Benedikt, D. Cosker, P. L. Rosin, and D. Marshall, "Assessing the uniqueness and permanence of facial actions for use in biometric applications," *IEEE Transactions on Systems, Man and Cybernetics - Part A*, vol. 40, no. 3, pp. 449–460, May 2010.

[184] Y. Sun and L. Yin, "Facial expression recognition based on 3D dynamic range model sequences," in *Proc. Eur. Conf. on Computer Vision*, Marseille, France, Oct. 2008, pp. 58–71.

[185] Y. Sun, X. Chen, M. J. Rosato, and L. Yin, "Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis," *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, vol. 40, no. 3, pp. 461–474, 2010.

[186] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "A dynamic approach to the recognition of 3D facial expressions and their temporal models," in *Proc. IEEE Conf. on Automatic Face and Gesture Recognition*, Santa Barbara, CA, Mar. 2011, pp. 406–413.

[187] V. Le, H. Tang, and T. S. Huang, "Expression recognition from 3D dynamic faces using robust spatio-temporal shape features," in *IEEE Conference on Automatic Face and Gesture Recognition*, Santa Barbara, CA, Mar. 2011, pp. 414–421.

[188] T. Fang, X. Zhao, S. Shah, and I. Kakadiaris, "4D facial expression recognition," in *Proc. IEEE Int. Conf. on Computer Vision Workshop*, Barcelona, Spain, Nov. 2011, pp. 1594–1601.

[189] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[190] "3d/4d facial expression analysis: An advanced annotated face model approach," *Image and Vision Computing*, vol. 30, no. 10, pp. 738 – 749, 2012.

[191] M. Reale, X. Zhang, and L. Yin, "Nebula feature: A space-time feature for posed and spontaneous 4d facial behavior analysis," *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 0, pp. 1–8, 2013.

[192] X. Zhang, L. Yin, J. F. Cohn, S. J. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *FG*, 2013, pp. 1–6.

[193] H. Drira, B. Ben Amor, M. Daoudi, A. Srivastava, and S. Berretti, "3D dynamic expression recognition based on a novel deformation vector field and random forest," in *Proc. Int. Conf. on Pattern Recognition*, Tsukuba, Japan, Nov. 2012, pp. 1104–1107.

[194] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3D facial expression dynamics," *Image and Vision Computing*, vol. 30, no. 10, pp. 762–773, 2012.

[195] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes, "Nonrigid registration using free-form deformations:application to breast mr images." *IEEE Transactions on medical imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[196] ——, "Nonrigid registration using free-form deformations: application to breast mr images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, p. 712–721, 1999.

[197] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[198] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–94, Jan. 1980.

[199] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[200] Di3D, "http://www.di3d.com," 2006.

[201] B. Ben Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, "4-d facial expression recognition by learning geometric deformations," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2443–2457, 2014.

[202] X. Yang, D. Huang, Y. Wang, and L. Chen, "Automatic 3d facial expression recognition using geometric scattering representation," in *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, vol. 1, 2015, pp. 1–6.

[203] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–8, 2012.

[204] X. Zhang, L. Yin, J. F. Cohn, S. J. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, 2014.

[205] Y. Sun and L. Yin, "Facial expression recognition based on 3d dynamic range model sequences," in *Europeon Conference on Computer Vision*, 2008, pp. 58–71.

[206] Y. Sun, X. Chen, M. Rosato, and L. Yin, "Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 40, no. 3, pp. 461–474, 2010.

[207] M. Reale, X. Zhang, and L. Yin, "Nebula feature: A space-time feature for posed and spontaneous 4D facial behavior analysis." in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[208] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3d facial expression dynamics," *Image and Vision Computing*, vol. 30, no. 10, pp. 762–773, 2012.

[209] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3d/4d facial expression analysis: An advanced annotated face model approach," *Image and Vision Computing*, vol. 30, no. 10, pp. 738–749, 2012.

[210] V. Le, H. Tang, and T. Huang, "Expression recognition from 3d dynamic faces using robust spatio-temporal shape features," in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, 2011, pp. 414–421.

[211] M. Xue, A. Mian, W. Liu, and L. Li, "Automatic 4d facial expression recognition using dct features," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 199–206.

[212] S. Berretti, A. Del Bimbo, and P. Pala, "Automatic facial expression recognition in real-time from dynamic sequences of 3d face scans," *The Visual Computer*, vol. 29, no. 12, pp. 1333–1350, 2013.

[213] S. Kurtek and H. Drira, "A comprehensive statistical framework for elastic shape analysis of 3d faces," *Comput. Graph.*, vol. 51, pp. 52–59, 2015. [Online]. Available: https://doi.org/10.1016/j.cag.2015.05.027

[214] A. B. Tumpach, H. Drira, M. Daoudi, and A. Srivastava, "Gauge invariant framework for shape analysis of surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 46–59, 2016. [Online]. Available: https://doi.org/10.1109/TPAMI.2015.2430319

[215] H. Drira, B. Ben Amor, A. Srivastava, D. Daoudi, and R. Slama, "3D face recognition under expressions, occlusions, and pose variations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2270–2283, 2013.

[216] S. Kurtek, E. Klassen, J. Gore, Z. Ding, and A. Srivastava, "Elastic geodesic paths in shape space of parameterized surfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1717–1730, 2012.

[217] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Automatic Face and Gesture Recognition*, 2006, pp. 211–216.

[218] S. Kurtek, C. Samir, and L. Ouchchane, "Statistical shape model for simulation of realistic endometrial tissue," in *International Conference on Pattern Recognition Applications and Methods*, 2014.

[219] T. Windheuser, U. Schlickewei, F. R. Schmidt, and D. Cremers, "Geometrically consistent elastic matching of 3d shapes: A linear programming solution," in *Proceedings of the 2011 International Conference on Computer Vision*, ser. ICCV '11, 2011, pp. 2134–2141.

[220] M. Kilian, N. J. Mitra, and H. Pottmann, "Geometric modeling in shape space," *ACM Trans. Graphics*, vol. 26, no. 3, 2007, Proc. SIGGRAPH.

[221] B. Heeren, M. Rumpf, M. Wardetzky, and B. Wirth, "Time-discrete geodesics in the space of shells," *Comp. Graph. Forum*, vol. 31, no. 5, pp. 1755–1764, Aug. 2012.

[222] L. Younes, "Computable elastic distances between shapes," *SIAM J. of Applied Math*, pp. 565–586, 1998.

[223] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1415–1428, 2011.

[224] S. Kurtek, E. Klassen, Z. Ding, and A. Srivastava, "A novel Riemannian framework for shape analysis of 3d objects," in *CVPR*, 2010, pp. 1625–1632.

[225] S. Kurtek, E. Klassen, J. C. Gore, Z. Ding, and A. Srivastava, "Elastic geodesic paths in shape space of parameterized surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1717–1730, 2012.

[226] N. Litke, M. Droske, M. Rumpf, and P. Schröder, "An image processing approach to surface matching," in *Proceedings of the Third Eurographics Symposium on Geometry Processing*, ser. SGP '05, 2005.

[227] I. H. Jermyn, S. Kurtek, E. Klassen, and A. Srivastava, "Elastic shape matching of parameterized surfaces using square root normal fields," in *ECCV (5)*, 2012, pp. 804–817.

[228] P. W. Michor and D. Mumford, "Vanishing geodesic distance on spaces of submanifolds and diffeomorphisms," *Doc. Math.*, vol. 10, pp. 217–245, 2005.

[229] M. Bauer, P. Harms, and P. W. Michor, "Curvature weighted metrics on shape space of hypersurfaces in n-space," *Differential Geom. Appl.*, vol. 30, pp. 33–41, 2012.

[230] ——, "Almost local metrics on shape space of hypersurfaces in n-space," *SIAM J. Img. Sci.*, vol. 5, no. 1, pp. 244–310, Mar. 2012.

[231] ——, "Sobolev metrics on shape space of surfaces," *Journal of Geometric Mechanics*, vol. 3, no. 4, pp. 389–438, 2011.

[232] M. Fuchs, B. Jüttler, O. Scherzer, and H. Yang, "Shape metrics based on elastic deformations," *J. Math. Imaging Vis.*, vol. 35, no. 1, pp. 86–102, 2009.

[233] M. Bauer, M. Bruveris, and P. W. Michor, "Overview of the geometries of shape spaces and diffeomorphism groups," *Journal of Mathematical Imaging and Vision*, vol. 50, no. 1-2, pp. 60–97, 2014.

[234] D. G. Ebin, "The manifold of Riemannian metrics," *Global Analysis (Proc. Sympsos. Pure Math., Vol. XV, Berkeley, Calif.,1968)*, pp. 11–40, 1970.

[235] D. S. Freed and D. Groisser, "The basic geometry of the manifold of Riemannian metrics and of its quotient by the diffeomorphism group," *Michigan Math. J.*, 1989.

[236] O. Gil-Medrano and P. W. Michor, "The Riemannian manifold of all Riemannian metrics," *Quaterly J. Math. Oxford*, vol. 2, 1991.

[237] B. Clarke, "The metric geometry of the manifold of Riemannian metrics over a closed manifold," *Calculus of Variations and Partial Differential Equations*, vol. 39:533–545, 2010.

[238] M. Bauer, M. Bruveris, and P. W. Michor, "Constructing reparametrization invariant metrics on spaces of plane curves," *Differential Geometry and its Applications*, vol. 34, pp. 139–165, 2014.

[239] S. Lahiri, D. Robinson, and E. Klassen, "Precise matching of PL curves in $\mathbb{R}^N$ in the square root velocity framework," 2015.

[240] R. Courant and D. Hilbert, *Methods of Mathematical Physics*. Wiley-Interscience, 1962, vol. Volume I.

[241] A. Bronstein, M. Bronstein, and R. Kimmel, *Numerical Geometry of Non-Rigid Shapes*. Springer, 2008.

[242] S. Kurtek, A. Srivastava, E. Klassen, and H. Laga, "Landmark-guided elastic shape analysis of spherically-parameterized surfaces," *Eurographics*, vol. 32, no. 2, 2013.

[243] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[244] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network." in *CVPR*, vol. 2, 2017, p. 4.

[245] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," *arXiv preprint arXiv:1610.09585*, 2016.

[246] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin, "Pose guided human video generation," in *European Conference on Computer Vision*. Springer, 2018, pp. 204–219.

[247] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2347–2356.

[248] Z. Huang, C. Wan, T. Probst, and L. V. Gool, "Deep learning on lie groups for skeleton-based action recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 1243–1252. [Online]. Available: https://doi.org/10.1109/CVPR.2017.137

[249] Z. Huang and L. V. Gool, "A riemannian network for SPD matrix learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 2017, pp. 2036–2042. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14633

[250] Z. Huang, J. Wu, and L. V. Gool, "Building deep networks on grassmann manifolds," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th*

AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, 2018, pp. 3279–3286. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16846

[251] A. Dhall, R. Goecke, S. Lucey, T. Gedeon *et al.*, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[252] H. Halim, K. Hideki, G. Julie, and B. Alain, "Modeling, simulation and optimization of bipedal walking." Springer Berlin Heidelberg, 2013, ch. The combined role of motion-related cues and upper body posture for the expression of emotions during human walking, pp. 71–85.

[253] G. Venture, H. Kadone, T. Zhang, J. Grèzes, A. Berthoz, and H. Hicheur, "Recognizing emotions conveyed by human gait," *I. J. Social Robotics*, vol. 6, no. 4, pp. 621–632, 2014. [Online]. Available: https://doi.org/10.1007/s12369-014-0243-1

[254] A. Kacem, M. Daoudi, B. B. Amor, S. Berretti, and J. C. Á. Paiva, "A novel geometric framework on gram matrix trajectories for human behavior understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 1–14, 2020. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2872564

[255] A. P. Atkinson, M. L. Tunstall, and W. H. Dittrich, "Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures," *Cognition*, vol. 104, no. 1, pp. 59–72, 2007.

[256] A. P. Atkinson, W. H. Dittrich, A. J. Gemmell, and A. W. Young, "Emotion perception from dynamic and static body expressions in point-light and full-light displays," *Perception*, vol. 33, no. 6, pp. 717–746, 2004.

[257] B. De Gelder, "Towards the neurobiology of emotional body language," *Nature Reviews Neuroscience*, vol. 7, no. 3, pp. 242–249, 2006.

[258] C. L. Roether, L. Omlor, A. Christensen, and M. A. Giese, "Critical features for the perception of emotion from gait," *Journal of vision*, vol. 9, no. 6, pp. 15–15, 2009.

[259] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[260] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1812.08008, 2018.

[261] J.-L. Wang, J.-M. Chiou, and H.-G. Müller, "Functional data analysis," *Annual Review of Statistics and Its Application*, vol. 3, no. 1, pp. 257–295, 2016.

[262] R. Vemulapalli and R. Chellapa, "Rolling rotations for recognizing human actions from 3d skeletal data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4471–4479.

[263] R. Anirudh, P. Turaga, J. Su, and A. Srivastava, "Elastic functional coding of riemannian trajectories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 922–936, 2017.

[264] J. Su, A. Srivastava, F. D. M. de Souza, and S. Sarkar, "Rate-invariant analysis of trajectories on riemannian manifolds with application in visual speech recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 620–627.

[265] X. Dai and H.-G. Müller, "Principal component analysis for functional data on riemannian manifolds and spheres," *arXiv preprint arXiv:1705.06226*, 2017.

[266] Z. Zhang, E. Klassen, and A. Srivastava, "Phase-amplitude separation and modeling of spherical trajectories," *Journal of Computational and Graphical Statistics*, vol. 27, no. 1, pp. 85–97, 2018.

[267] N. Hosni, H. Drira, F. Chaieb, and B. Ben Amor, "3d gait recognition based on functional pca on kendall's shape space," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2130–2135.

[268] Z. Lin and F. Yao, "Intrinsic riemannian functional data analysis," *arXiv preprint arXiv:1812.01831*, 2018.

Title Shape Analysis for Human Behavior Understanding

Abstract As one of the most active research areas in computer vision, visual analysis of human motion attempts to detect, track and identify people, and more generally, to interpret human behaviors, from image sequences involving humans. The main concern of this dissertation is the issue of shape analysis of imaging data with application to human behavior analysis. In particular, to filter some undesirable transformations, the shape extracted from the human body and face are represented as elements of a shape space defined as the invariant under the action of groups modeling the undesirable transformations. The main contribution presented in this dissertation is a unified framework for human behavior analysis through multiple manifolds representing different data, with different applications ranging from action recognition to soft-biometrics estimation including facial expression analysis and classification. First, the landmarks issued from the skeleton or facial landmarks were modeled on Kendall shape space where the comparison is invariant to scale, translation and rotation. An intrinsic sparse coding and dictionary learning SCDL on the Kendall Shape Space were performed with application to action and expression recognition using dynamic landmarks. A comparative study to an extrinsic sparse coding is also presented to understand the benefit of each methodology. Second, the facial curves were viewed as points on an infinite-dimensional, dfferentiable manifold and shooting vector along a geodesic representing the deformations between 3D faces has been proposed with application to soft-biometric recognition from 3D faces and expression recognition from 3D dynamic faces. Finally, a framework for 3D parametrized surfaces is presented. We present the algorithms to calculate geodesic paths, distances and intrinsic means. A novel idea based on gauge theory capable to compute the geodesic paths on shape space without any need to filter the re-parameterization group is proposed. Experiments conducted on the main benchmarks of action, facial expression and soft-biometric recognition demonstrate the efficiency of the proposed framework on the task of human behavior understanding.

Keywords shape analysis ; riemannien geometry ; action recognition ; facial expression recognition ; dynamic faces analysis

Titre Analyse de formes pour la compréhension du comportement humain

Résumé L'analyse visuelle des mouvements humains est l'un des domaines de recherche les plus actifs en vision par ordinateur. Elle vise à détecter, suivre et identifier les personnes, et plus généralement, d'interpréter les comportements humains, à partir de séquences d'images impliquant des humains. Cette Habilitation a pour thème principal l'analyse de forme des données d'imagerie avec application à l'analyse du comportement humain. En particulier, pour filtrer certaines transformations indésirables, les formes extraites du corps et du visage humain sont représentées comme des éléments d'un espace de formes défini comme invariant sous l'action de groupes modélisant les transformations indésirables. La principale contribution présentée dans cette habilitation est un cadre unifié pour l'analyse du comportement humain à travers de multiples variétés représentant différentes données, avec différentes applications allant de la reconnaissance d'action à l'estimation de la biométrie douce, y compris l'analyse et la classification des expressions faciales. Premièrement, les landmarks issus des skeletons humains ou du visage sont modélisés sur l'espace de forme de Kendall où la comparaison est invariante à l'échelle, à la translation et à la rotation. Un codage parcimonieux intrinsèque sur l'espace de forme de Kendall a été effectué avec une application à la reconnaissance d'action et d'expression à partir de landmarks dynamiques. Une étude comparative à un codage extrinsèque parcimonieux est également présentée pour comprendre les avantages de chaque méthodologie. Deuxièmement, les courbes faciales ont été vues comme des points sur une variété de dimension infinie et un vecteur de vélocité le long d'une géodésique représentant les déformations faciales entre les visages 3D a été proposé avec une application à la reconnaissance des biométries douces à partir de visages 3D et à la reconnaissance d'expressions faciales à partir de visages 3D dynamiques. Enfin, un cadre pour les surfaces 3D paramétrées est présenté. Nous présentons les algorithmes pour calculer les géodésiques, les distances et les moyens intrinsèques. Une nouvelle idée basée sur la théorie de jauge capable de calculer les chemins géodésiques sur l'espace de forme sans avoir besoin de filtrer le groupe de re-paramétrisation est proposée. Les expériences menées sur les principaux benchmarks de reconnaissance d'action, d'expression faciales et de reconnaissance de biométries douces démontrent l'efficacité du cadre proposé pour l'analyse et la compréhension du comportement humain.

Mots-clés analyse de formes ; géométrie riemanniennne ; reconnaissance d'action ; reconnaissance des expressions faciales ; analyse de visages dynamiques