Multimodal emotion recognition from physiological signals and facial expressions

Yujin Wu

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LILLE

Ecole Doctorale MADIS-631
Mathématiques et Sciences du numérique

Spécialité de doctorat : Informatique et Applications

Soutenue publiquement à Lille le 13/04/2023, par :

Yujin Wu

Reconnaissance multimodale des émotions à partir de signaux physiologiques et d'expressions faciales

Multimodal emotion recognition from physiological signals and facial expressions

---

Composition du Jury:

Irene Cheng Professeur, University of Alberta, Canada	**Rapporteure**
Denis Hamad Professeur, Université du Littoral Côte d'Opale, France	**Rapporteur**
Shaun Canavan Professeur Assistant, University of South Florida, United States	**Examinateur**
Stefano Berretti Professeur Associate, University of Florence, Italy	**Examinateur**
Deise Santana Maia Maîtresse de Conférences, Université de Lille, France	**Examinatrice**
Gilles Lebuffe Professeur, Université de Lille, France	**Président du jury**
Mohamed Daoudi Professeur, IMT Nord Europe, France	**Directeur de thèse**
Ali Amad Professeur, Université de Lille, France	**Co-Directeur de thèse**
Juan-Carlos Álvarez Paiva Professeur, Université de Lille, France	**Invité**
Rachid Oulad Haj Thami Professeur, ENSIAS, Université Mohamed V, Maroc	**Invité**

Abstract

Emotion recognition is a subfield of affective computing and a critical research direction for the development of human-centered artificial intelligence, which targets the identification and interpretation of human emotions through machines in an objective and automatic manner. Emotions can be expressed and conveyed through a variety of channels which can be broadly categorized into externally observable behavioural modalities such as facial expressions, body postures, speech and intrinsic physiological modalities such as brain activities, skin conductance, heart rate, etc. Therefore, emotion recognition methods have been developed accordingly based on these modalities. Recent advances in sensor technology and human-computer interaction systems offer the possibility of migrating the deployment of emotion recognition systems from the laboratory to the real world. However, most of the existing research is still directed towards the former. In this context, we are interested in designing emotion recognition algorithms for real-life scenarios from physiological, behavioural and multimodal perspectives and exploring the implications and impact of their interventions on practical applications.

In this thesis, from the physiological perspective, we propose a self-supervised multimodal representation learning method for wearable emotion recognition based on peripheral physiological signals to cope with the overfitting problem posed by limited labelled data and the bias of inaccurate real-world labelling. From the behavioural perspective, we suggest a non-Euclidean metric learning method for 2D facial landmark-based expression recognition to tackle problems such as pose variation and intra-class variation. From a multimodal perspective, we present a deep geometric framework based on a symmetric positive definite matrix representation for multimodal stress and pain detection to address the fusion issue of physiological and behavioural modalities. Additionally, we also realize the deployment of emotion recognition in a real-life scenario, i.e. medical simulation training, where the previously proposed algorithm is integrated into a graphical user interface to test its feasibility and validity on the collected multimodal data, while exploring its pedagogical effects.

Keywords: physiological signals, facial landmarks, emotion recognition, facial expression recognition, stress/pain detection, self-supervised learning, metric learning, manifold.

Résumé

La reconnaissance des émotions est un sous-domaine de l'informatique affective et une direction de recherche critique pour le développement de l'intelligence artificielle centrée sur l'homme, qui vise l'identification et l'interprétation des émotions humaines par des machines de manière objective et automatique. Les émotions peuvent être exprimées et transmises par divers canaux qui peuvent être classés en deux grandes catégories : les modalités comportementales observables de l'extérieur, telles que les expressions faciales, les postures corporelles et la parole, et les modalités physiologiques intrinsèques, telles que les activités cérébrales, la conductivité de la peau, le rythme cardiaque, etc. Des méthodes de reconnaissance des émotions ont donc été développées sur la base de ces modalités. Les récentes avancées en matière de technologie des capteurs et de systèmes d'interaction homme-machine offrent la possibilité de faire passer le déploiement des systèmes de reconnaissance des émotions du laboratoire au monde réel. Cependant, la plupart des recherches existantes sont encore orientées vers le premier. Dans ce contexte, nous nous intéressons à la conception d'algorithmes de reconnaissance des émotions pour des scénarios de la vie réelle d'un point de vue physiologique, comportemental et multimodal et à l'exploration des implications et de l'impact de leurs interventions sur des applications pratiques.

Dans cette thèse, du point de vue physiologique, nous proposons une méthode d'apprentissage de représentation multimodale auto-supervisée pour la reconnaissance d'émotion portable basée sur des signaux physiologiques périphériques pour faire face au problème du surapprentissage posé par des données étiquetées limitées et le biais d'étiquetage inexact dans le monde réel. Du point de vue comportemental, nous suggérons une méthode d'apprentissage métrique non euclidienne pour la reconnaissance d'expressions basées sur des points de repère faciaux en 2D afin de résoudre des problèmes tels que la variation de pose et la variation intra-classe. D'un point de vue multimodal, nous présentons un cadre géométrique profond basé sur une représentation matricielle symétrique définie positive pour la détection multimodale du stress et de la douleur afin de résoudre le problème de la fusion des modalités physiologiques et comportementales. En outre, nous réalisons le déploiement de la reconnaissance des émotions dans un scénario réel, à savoir la formation par simulation médicale, où l'algorithme proposé précédemment est

intégré dans une interface utilisateur graphique pour tester sa faisabilité et sa validité sur les données multimodales collectées, tout en explorant ses effets pédagogiques.

Mots clés: signaux physiologiques, repères faciaux, reconnaissance des émotions, reconnaissance des expressions faciales, détection du stress et de la douleur, apprentissage auto-supervisé, apprentissage métrique, manifold.

Acknowledgments

First and foremost, I would like to thank my supervisor, Mohamed Daoudi, for his invaluable guidance, encouragement, and unwavering support throughout my research. He has demonstrated an exceptional ability to connect with people on a personal level, and his kindness and understanding have created a supportive and positive research environment. Moreover, I have benefited greatly from his approach to balancing life and research, which will be an asset for the rest of my life.

Secondly, I would like to express my sincere gratitude to my co-supervisor, Ali Amad, for his patient listening and unstintingly positive feedback during every meeting. His experience and expertise in the field of cognitive science has been instrumental in enabling me to continue my exploration and research in cross-disciplinary areas.

I would also like to thank Prof. Juan-Carlos Álvarez Paiva from the University of Lille for his invaluable contributions to my research on facial expression recognition. His solid knowledge and expertise in the field of mathematics have been essential in advancing my understanding of complex mathematical concepts and their applications for learning tasks.

I would like to express my special thanks to the members of the PhD committee, Prof. Gilles Lebuffe, Prof. Irene Cheng, Prof. Denis Hamad, Assistant Prof. Shaun Canavan, Associate Prof. Stefano Berretti, and Associate Prof. Deise Santana Maia for their time, insights, and precious feedback of my research.

I extend my heartfelt thanks to the Presage Simulation Center and I-SITE in Lille, France for their financial support of this thesis project with reference ANR-16-IDEX-0004 ULNE.

I would like to thank Dr. Thibaut Denis from the Faculty of Medicine of the University of Lille and all the staff of the Presage Simulation Center for their outstanding efforts and contributions to the acquisition and analysis of real-world data for my research. Through collaboration

and open communication with them, I have been able to bridge the gap between theory and practice.

I am also grateful to all my colleagues from the 3D SAM team, Emery Pierson, Baptiste Chopin, Kévin Feghoul, Estephe Arnaud, Deise Santana Maia, Thomas Besnier who shared the joy and challenges of research with me, supporting and encouraging each other to progress.

Last but not least, I would like to thank my parents, Hongwei and Qunying, and my friend Yiheng. My parents' emotional and financial support has been a constant source of motivation and inspiration to me and has enabled me to overcome challenges and pursue my academic goals with confidence and determination. Their belief in me has been a driving force behind my success, and I am forever grateful for their sacrifice and dedication. My friend Yiheng is always a constant source of support and encouragement throughout my PhD journey. Her companionship and positivity have helped me to stay motivated and focused, even during the most challenging times.

Contents

List of Figures

List of Tables

List of Algorithms

List of acronyms

ACC	3-axis accelerometer
ANS	autonomous nervous system
BVP	blood volume pressure
CNN	convolutional neural networks
CNS	central nervous system
ECG	electrocardiogram
EDA	electrodermal activity
EMG	electromyography
EOG	electrooculography
FCN	fully connected neural networks
FE	facial expressions
GUI	graphical user interface
HR	heart rate
HRV	heart rate variability
IBI	inter beat interval
KNN	k-nearest neighbour
LMNN	large margin nearest neighbors
LSTM	long short-term memory

PNS	peripheral nervous system
PPS	peripheral physiological signals
RESP	respiration
SPD	symmetric positive definite
SSL	self-supervised learning
SVM	Support Vector Machine
TCN	temporal convolutional network
TEMP	skin temperature

1

Introduction

Contents

1.1 Motivation

Emotions are sets of complex physiological, cognitive and behavioral responses that are triggered by internal or external stimuli. They perform a central role in human daily life and significantly influence mechanisms such as attention, decision-making, learning, memorisation and perception. Therefore, the identification of emotional states is crucial for a thorough understanding of human intelligence, behaviour and cognition [10, 11]. In 1995, Picard's pioneering work [12] in affective computing laid the foundation for enabling machines to intelligently recognise, interpret, simulate, and respond to human emotions. It is an interdisciplinary field that integrates computer science, psychology, social and cognitive sciences. One of the main topics that emerged from it is emotion recognition [13] which attempts to empower computers with the ability to automatically infer human emotions and has become increasingly essential for the advancement of human-centred artificial intelligence.

In recent years, rapid advances in sensors, algorithms, and computing resources are driving the deployment of emotion recognition systems from the laboratory to the real world. The corresponding application scenarios include:

- **Healthcare:** Emotion recognition provides diagnosis, prediction, and proactive intervention for psychological disorders such as depression [14, 15];
- **Automotive:** Emotion recognition can assist in driving by identifying whether drivers are fatigued or aggressive in order to ensure their safety [16, 17];
- **Human-computer interaction:** Emotion recognition can interpret emotional cues to enhance the user experience of various interactive devices such as robots, games, smart speakers, etc [18, 19];
- **Marketing and advertising:** Emotion recognition enables precise marketing or the design of specialized advertising by interpreting the emotional responses of the target customers [20];

In addition to the above examples, another compelling application is **Emotion-aware education** [21, 22], which aims to incorporate emotion recognition into education programs by monitoring the students' emotional state to provide better quality tutoring. Previous studies [23–25] have suggested a strong association between emotions and learning procedures. In [26], emotions are considered to be a major impact factor in learning. In addition, the impact of different types of emotions on the learning process varies, with positive emotions influencing aspects such as motivation, concentration and self-regulation of learning, while negative emotions are

detrimental to students' performance and achievement [22]. One of the most representative negative emotions is stress, defined in Hans Selye's pioneering work as "the non-specific response of the body to any demand for change" [27], and commonly referred to as a state of high arousal. According to the Yerkes–Dodson law [1], a person's arousal level tends to lead to a decrease in performance after a certain threshold has been exceeded (i.e. when in a state of stress). An illustration is shown in Fig. 1.1. The evidence reviewed here indicates the demand to introduce automatic recognition of students' emotions, especially their stress states, into the educational sessions, which is indeed the intention of this thesis.

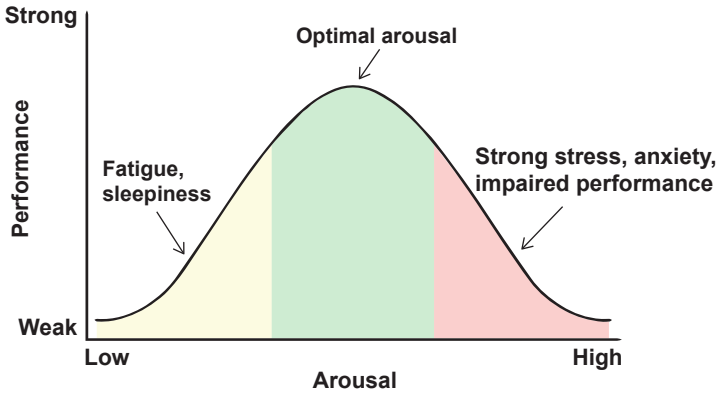


Figure 1.1: The correlation between arousal levels and performance described in Yerkes–Dodson law [1]. Excerpt from [2].

More specifically, this thesis concerns medical simulation training for health professionals at the PRESGAE Health Simulation Centre[1]. During the simulation training, learners may experience stressful states due to factors such as "high fidelity" medical scenarios, complex medical tasks, etc., which can result in decreased performance and learning efficiency. Therefore, the main objective of this thesis is to develop a new tool for identifying the emotions, especially stress, felt by learners during simulations, thereby offering them targeted debriefings to improve their learning. In the end, two sub-goals are concluded from the main objective:

- **Technical objective:** emotions are multi-faceted and involve changes in behavior, physiology, and psychology. Thus, novel automatic emotion recognition algorithms suitable for real-life settings can be designed and evaluated from behavioural, physiological, and multimodal perspectives to achieve a more thorough exploration.

- **Practical objective:** this sub-goal can be further divided into three steps: firstly, emotion data needs to be collected under real-life conditions, secondly, an emotion recognition

[1] https://medecine.univ-lille.fr/presage.

tool (i.e. a graphical user interface) that integrates the designed intelligent algorithms needs to be developed and tested under real-life conditions, and finally, the pedagogical impact of this tool needs to be analyzed and evaluated.

1.2 Challenges

A wide range of human emotional information can be captured, broadly classified into two categories: behavioural modality-based (e.g., facial expressions, body gestures, audio, etc) and physiological modality-based (e.g., electroencephalography (EEG) signals and a series of peripheral physiological signals), representing controllable external manifestations and uncontrollable internal variations under certain emotions, respectively. Considering the necessity of switching from in-lab settings to real-world applications, we focus on peripheral physiological signals and facial landmarks for unimodal emotion recognition, while the multimodal approach is based on the fusion of physiological modalities (i.e., peripheral physiological signals) and two behavioral modalities (i.e., facial expressions and body movements represented by facial landmarks and accelerations, respectively). The reason for selecting these modalities is further explained in Section 2.2. In terms of emotion modality, automatic emotion recognition methods are accordingly differentiated into behaviour-based approach, physiology-based approach and multimodal (behavioural+physiological) approach. In the following, we list the challenges encountered by the different types of approaches for deploying accurate and robust emotion recognition systems in a practical setting.

1.2.1 Physiological chanllenges

-Feature engineering of multivariate signal Physiology-based emotion recognition is a complex time-series analysis involving multichannel heterogeneous signals, which typically require domain knowledge in neuropsychology [28]. To achieve favorable recognition performance, critical features need to be extracted and selected from each variable, which turns out to be computationally expensive and time-consuming, especially for traditional machine learning-based algorithms. Furthermore, the combinations of multimodal features are generally based on early or late fusion which cannot simultaneously capture intra- and inter-modal correlations.

-Inaccurate annotation and noisy data Existing peripheral physiological datasets were mainly collected in a laboratory environment, where the experimental protocols were carefully designed to stimulate the expected emotional responses and obtain accurate labeling (the bound-

ary between emotional and non-emotional data is quite clear based on the protocol). When shifting the focus to real-life scenarios, the continuous nature of emotions makes it challenging to determine the beginning and end of emotional events. In addition, subjective errors introduced by self-evaluation also contribute to the uncertainty of the labels. Besides, unrestricted body movements can lead to signal artefacts which significantly degrade data quality [29].

-Limited labeled data and low subject diversity Labeled data volume and subject diversity are critical to obtaining models with generalization capabilities, particularly for supervised deep learning models, whose automatic representation learning can effectively address the plague of inefficient feature engineering. To date, the majority of emotion-related physiological datasets are based on EEG signals, which are not applicable in real-life scenarios. In addition, current emotion recognition studies typically use physiological data covering no more than 50 subjects [10].

1.2.2 Behaviroual chanllenges

-Pose variations In an unconstrained environment, the head pose may not be frontal to the camera, and this variation in appearance results in facial displacement, scale changes, in-plane and out-of-plane rotations, preventing us from directly using the detected facial feature points for facial expression analysis [30]. From the perspective of landmark configurations, these changes can be summarised as rigid transformations of the face shape in the 3D case and more complex projection transformatios in the 2D case. In addition, pose changes are often accompanied by other factors such as self-occlusion, texture distortion, etc.

-Intra-class variation inter-subject differences due to gender, age, culture, etc. often hinder the analysis of facial points, resulting in large intra-class variations. This is manifested as different individuals expressing the same facial expression in different ways or, in a more extreme case, the same person presenting the same expression in different ways at different moments.

1.2.3 Multimodal chanllenges

Since the manifestation and evolution of emotions involves signals from multiple domains, recent research has turned to behaviroual-physiological fusion with the expectation of leveraging complementary properties between modalities to achieve more robust predictions [31]. In addition, multimodal-based approaches can deal with the problem of missing modalities. For

example, when dealing with complex emotions such as stress, the intensity of facial expressions is commonly weak, whereas physiological information can assist in identifying emotions.

-Lack of multimodal data Currently, the emotion recognition community struggles with a lack of public available datasets which contain both behavioural and physiological data. Therefore, existing efforts tend to concentrate on a single perspective (i.e., either behavioural or physiological).

-Varying data structure Multimodal data also suffers from diversity of data structures. For example, the facial expression videos captured by the camera are a 2D matrix sequences with a high-spatial, low temporal resolution, while the peripheral physiological signals captured by the biosensors are a 1D time series with a lower-spatial, higher-temporal resolution.

-Multimodal fusion A major issue is how to efficiently fuse multimodal features to improve model performance. Simply concatenating feature vectors to generate multimodal representations is the most frequently used technique for integrating behavioral and physiological data. However, the resulting vectors are typically high-dimensional, associated with redundancy, and fail to effectively capture intra-modal correlations. An alternative way in existing research is decision-level fusion, where features from different modalities are extracted independently to provide the corresponding decisions. However, the interactions between the modalities are not well established during this process.

1.3 Thesis contributions

In order to achieve the main objective of this thesis, i.e. to develop novel emotion recognition algorithms and to facilitate their deployment in the real world, the challenges mentioned in the previous section should be addressed. To this end, the contributions achieved in this work can be grouped accordingly into technical and practical levels.

Technical level Emotion recognition algorithms based on physiological, behavioural and multimodal signals are proposed. Thus, the corresponding contributions are three-fold, depending on the type of modality used:

- **From the physiological viewpoint**, we proposed a self-supervised wearable emotion recognition based on peripheral physiological signals. First, in contrast to most machine learning-based physiological recognition methods, a deep neural network based

on residual temporal convolution and transformer was designed to (1) automatically capture more discriminative abstract features, and (2) flexibly fuse multimodal features at different levels, allowing the encoding of intra- and inter-modal correlation information. Secondly, unlike most supervised learning-based approaches, a self-supervised learning (SSL) scheme was adopted to (3) address the problem of de-generalisation due to the limited amount of labeled data and the small number of subjects, (4) get rid of the interference caused by inaccurate labeling and various noises in real-life scenarios, and (5) obtain an effective classification model that can be applied to a range of emotionally relevant downstream tasks.

- **From the behavioural viewpoint**, we propose a non-Euclidean metric learning method that can be applied to facial expression recognition. Different from traditional landmark-based approaches which construct the facial shape representation in Euclidean space, we considered the equivalence classes of 2D landmark configurations, which can be considered as points in the complex projective space to (1) obtain a more robust shape representation that is invariant under the class of affine transformations which contains rotation, scaling, translation, and (2) better measure the similarity between different facial shapes by using non-Euclidean metrics. Moreover, a metric learning algorithm was applied to (3) reduce the intra-class variations while increasing the inter-class distances to achieve a more discriminative feature space.

- **From the multimodal viewpoint**, we presented a novel LSTM-based geometric framework for multimodal stress and pain detection tasks. Unlike commonly used early fusion or late fusion strategies, we computed the symmetric positive definite (SPD) matrix of multimodal data as input to the learning model to (1) cope with structural differences between visual and physiological data and obtain compact, high-order multimodal representations, and (2) simultaneously capture intra- and inter-modal correlations at different instants. To the best of our knowledge, this is the first use of the geometry of SPD matrices to merge physiological and behavioural signals.

Practical level Emotion recognition in unconstrained conditions, i.e. during medical simulation training, was performed and the corresponding contribution was also three-fold:

- **From the experimental viewpoint**, a large-scale multimodal dataset containing motion signals and peripheral physiological signals was collected via a wearable device for emotion analysis, where a series of psychological questionnaires were assigned to subjects to generate ground truth for partial data.

- **From the implementation viewpoint**, an emotion analysis tool, i.e., a graphical user interface was developed to (1) load, synchronize and display video and multimodal signal recordings, (2) execute the proposed emotion recognition algorithms and (3) test its feasibility in a real-life setting.
- **From the pedagogical viewpoint**, the impact of the intervention of the emotion recognition system on pedagogy was analyzed and evaluated.

1.4 Thesis outline

This thesis is organised as follows: In Chapter 2, we conducted a systematic literature review of emotion representations, emotion-related modalities, and emotion recognition approaches based on physiological, behavioural and multimodal modalities. Chapters 3 and 4 correspond to our technical contributions on the physiological and multimodal sides, respectively, where Chapter 3 presents a self-supervised learning approach for emotion recognition based on peripheral physiological signals, and Chapter 4 introduces a deep geometric framework based on SPD representation for multimodal stress and pain detection. Chapter 5 refers to the practical contributions, including the collection of multimodal real-life data, the development and application of the graphical interface, and the pedagogical relevance analysis of emotion recognition. Regarding the technical contributions on the behavioral side, a landmark-based non-Euclidean metric learning approach for facial expression recognition is presented in Chapter 6. Finally, in Chapter 7, we summarise the work presented in this thesis and discuss its corresponding limitations and suggest potential future work.

1.5 Publications

- **Yujin Wu**, Mohamed Daoudi, Ali Amad, "Transformer-based Self-supervised Multimodal Representation Learning for Wearable Emotion Recognition", IEEE Trans. On Affective Computing, Accepted for publication.
- **Yujin Wu**, Mohamed Daoudi, Ali Amad, Laurent Sparrow, Fabien D'Hondt, "Fusion of Physiological and Behavioural Signals on SPD Manifolds with Application to Stress and Pain Detection". IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2022): 2949-2955.
- **Yujin Wu**, Mohamed Daoudi, "Metric Learning on Complex Projective Spaces", International Conference on Smart Multimedia (ICSM 2022):116–127.
- Eman A. Abdel-Ghaffar, **Yujin Wu**, Mohamed Daoudi, "Subject-Dependent Emotion Recognition System Based on Multidimensional Electroencephalographic Signals: A Riemannian Geometry Approach", IEEE Access 10: 14993-15006 (2022)
- **Yujin Wu**, Mohamed Daoudi, Ali Amad, Laurent Sparrow, Fabien D'Hondt, "Unsupervised Learning Method for Exploring Students' Mental Stress in Medical Simulation Training", ACM International Conference on Multimodal Interaction Companion (ICMI 2020): 165-170.

2

Related Work

Contents

In this chapter, we introduce the basic theoretical background on emotions and review recent emotion recognition methods. Firstly, Section 2.1 introduces two ways of describing emotions: the discrete emotion model and the continuous emotion model. Then, in Section 2.2, we present the common physiological and behavioral indicators associated with emotions and explain the reasons for focusing on peripheral physiological signals and facial landmarks in this thesis. Accordingly, emotion recognition methods can be classified in terms of modality: emotion recognition methods based on physiological signals, landmark-based facial expression recognition methods, and multimodal emotion recognition methods based on physiological and behavioral signals, which are presented in Sections 2.3, 2.4, and 2.5, respectively.

2.1 Emotion models

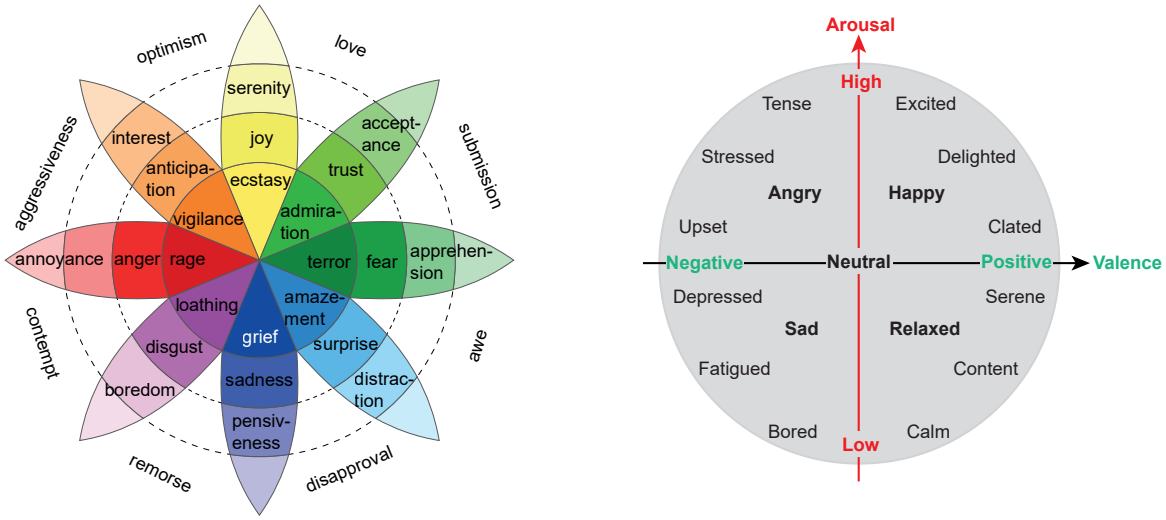
For the purpose of interpreting and quantifying human emotional states, two types of generic emotion models have been established by psychologists: the categorical model and the dimensional model, in which emotions are either denoted by discrete categories or mapped into a multidimensional space, respectively [32]. The emotion recognition tasks involved in this thesis and previous related studies are all based on these two emotion models.

2.1.1 Categorical models

The categorical models define emotions as discrete classes. This type of approach assumes that there exists a set of basic emotions that are innate to humans, whose expression and recognition is generally undifferentiated for individuals from different races or cultures [33]. A series of studies were carried out to build the collection of basic emotions, the most prominent of which are Ekman's basic emotion model [34] and Plutchik's 'wheel of emotions' [35]. Ekman proposed six basic emotions: *joy*, *sadness*, *anger*, *fear*, *disgust*, and *surprise*, based on the following attributes: 1) People are born with emotions rather than acquired; 2) People show the same emotions when they are in the same situations; 3) People exhibit analogous physiological or behaviroual patterns when expressing the same emotions. Plutchik's emotional wheel model [35] contains 8 primary bipolar emotions: *joy* vs *sadness*, *anger* vs *fear*, *trust* vs *disgust*, *surprise* vs *anticipation*. Fig. 2.1(a) shows an illustration of the 'wheel of emotion'. The intensity of the emotion ranges from weak to strong, as it shifts from the periphery to the centre of the wheel. In addition, more complex emotions can be also generated by combining primary emotions, e.g., the combination of *joy* and *trust* is *love*.

2.1.2 Dimensional models

For categorical models, complex emotions are more difficult to process since they cannot cover all emotion. Moreover, the continuity of changes in emotional states might not be well captured by discrete models. To address the above issues, researchers have turned to the dimensional models which describe emotions as points in a space of different dimensions. Russell [36] proposed a two-dimensional model, also known as the circumplex model, which is the most frequently adopted dimensional model for identifying emotions [37]. In this model, each emotional state is mapped as a discrete point in a two-dimensional space, with horizontal and vertical coordinates represented by valence and arousal, respectively. The valence axis represents the emotions ranging from unpleasant (negative) to pleasant (positive), and the arousal axis ranges from passive (low) to active (high), indicating its intensity level [38]. For example, stress can be defined as a specific emotional state of negative valence and active arousal. A graphical representation of the circumplex model is shown in Fig. 2.1(b)



(a) Plutchik's 'wheel of emotions', adapted from [39].

(b) Russell's 2D circumplex model based on valence and arousal.

Figure 2.1: Two types of generic emotion models.

2.2 Emotion related modalities

Automatic emotion recognition can be realized on the basis of behavioural or physiological indicators of the human body. Fig. 2.2 illustrates common behavioural and physiological modalities

applied to automatic emotion recognition, which are described in more detail below.

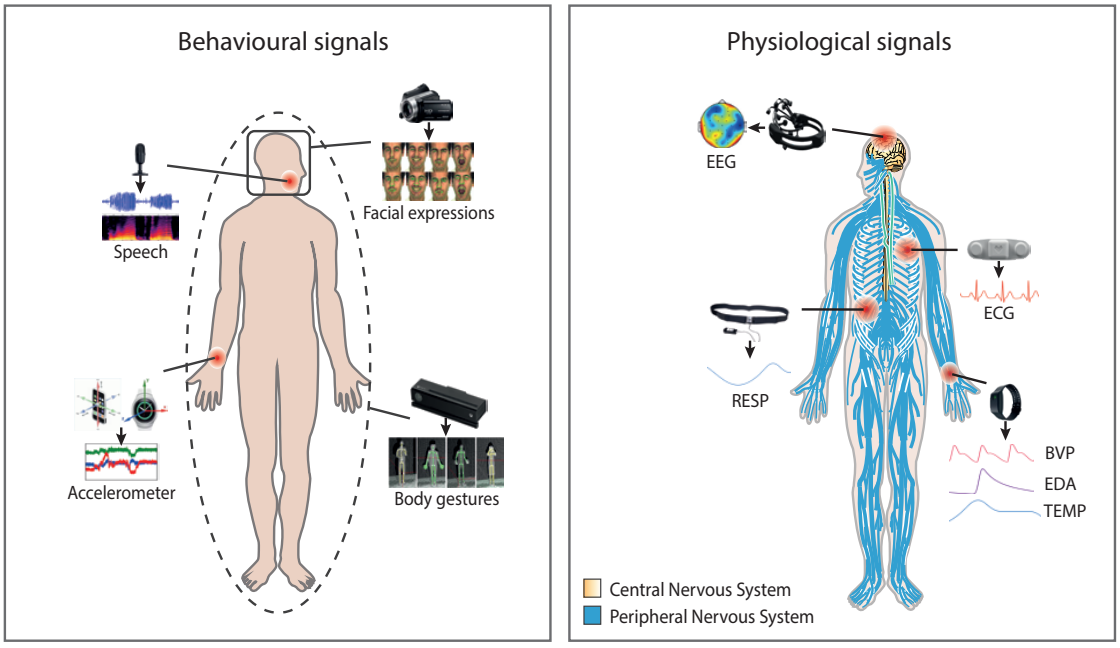


Figure 2.2: Common emotion-related behavioral and physiological modalities and their corresponding data collection devices.

2.2.1 Behaviroual modalities

Behavioural responses to emotional states are actions that can be voluntarily controlled or changed and externally observed [40]. Facial expressions [41], body gestures [42], speech [43], etc, are representative modalities of conveying one's emotional state, with the former two appearing to provide richer relevant information. A body language-based emotion recognition study indicates that non-verbal signals such as facial expressions, body posture, gestures, etc, have a more significant impact on the delivery of information than verbal signals, accounting for 65% of the total.

The facial expression captured by the camera is a typical non-verbal indicator of emotional states projected from human physical appearance. It is generally the result of an action or combination of actions formed by facial muscles and organs. Previous research [44] indicated that facial expressions are the most dominant form in which humans convey their emotions during social communication. According to the pioneering work of Ekman and Friesen [45], there is a high degree of consistency in facial expressions across human groups of different ethnic and cultural backgrounds. In addition, they adopted the facial action coding system (FACS) that systematically defines facial expression classification criteria by describing the

facial muscle movements and concluded six basic emotions: *joy*, *sadness*, *anger*, *fear*, *disgust*, and *surprise*. The strong connections between emotions and facial expressions established by neuroscience and psychology research have inspired researchers in the fields of computer vision and machine learning to explore automatic facial expression recognition algorithms. This also explains the reason for selecting facial expressions as the input modality for behaviour-based and multimodal emotion recognition in this thesis.

Another compelling visual modality that delivers human emotional information is body posture, which represents the movement of body parts and joints, It can normally be captured by cameras, Kinect sensors, or motion capture systems. Though most emotion analysis studies consider facial expressions as the primary source, there is still a body of literature dedicated to uncovering the link between body gestures and emotions. For example, a person's inner emotion changes can be expressed through the placement of hands or legs and the way they perform actions such as walking, sitting or standing [46]. The position of the head is also one of the emotion-related cues, a raised chin, for example, may indicate an arrogant attitude and a sense of superiority [47]. However, different from facial expressions, body posture seems to be greatly influenced by cultural or gender differences. An alternative way of obtaining information on human posture and activity for emotional analysis is to record 3-axis accelerometer (ACC) data from a smartphone or smartwatch. The ACC was initially used in wearable activity recognition systems, recent research in emotion recognition has applied it to the acquisition of contextual information on user actions [37]. In addition, some research [48] has demonstra0ted the feasibility of detecting stress states by estimating the intensity level of activities. In our work, it is applied to the multimodal emotion recognition task.

2.2.2 Physiological modalities

Physiological reactions triggered by emotions form a natural part of the human body functioning which are involuntary and involve internal manifestations that are imperceptible to the naked eye, such as changes in heart rate, respiratory rate, etc. Therefore, the physiological measures can provide more objective decisions for emotion recognition systems [32]. The physiological modalities primarily consist of electroencephalography (EEG) signal captured from Central Nervous System (CNS) and a series of signals measured from the Peripheral Nervous System (PNS) such as electrocardiogram (ECG), respiration (RESP), blood volume pressure (BVP), electrodermal activity (EDA), skin temperature (TEMP), etc.

The central nervous system as indicated by the yellow area in Fig. 2.2, consisting of the brain and spinal cord, is the main part of the human nervous system, where the electrical activity of

brain neurons can be accessed through EEG signals, providing information highly relevant to emotions [29]. However, the acquisition of EEG data requires the support of bulky and invasive head-mounted sensors, which typically consist of a large number of electrodes placed in different areas of the scalp, making their implementation in real-life scenarios quite challenging [37]. In addition, they are highly sensitive to physiological artefacts (e.g. blinking) and electrostatic artefacts introduced by the use of electrodes [49].

The peripheral nervous system refers to all nerves apart from the brain and spinal cord (as shown by the blue area in Fig. 2.2). In response to external or internal emotional stimuli, the autonomous nervous system (ANS), as one of the primary components of the PNS, is activated, thereby triggering a series of changes in peripheral physiological signals. For instance, elevated respiratory rate and increased electrodermal activity appear when subjects are in a high arousal state (e.g., anger, stress, fear) [11]. The peripheral physiological signals that are frequently used for emotion recognition include:

- **Electrocardiography (ECG)**: is a recording of the electrical activity generated during the contraction or relaxation of the cardiac muscles. It is a crucial indicator in emotion recognition research which is commonly applied to estimate parameters related to the cardiac cycle such as the heart rate (HR), heart rate variability (HRV), etc.

- **Blood Volume Pulse (BVP)**: photoplethysmography detects changes in blood volume by the degree to which the artery absorbs the light it emits, thus inferring cardiac cycle information.

- **Respiration (RESP)**: a measurement of exhaled or inhaled air volume and respiratory rate by contraction or relaxation of a chest strap fixed near the chest cavity.

- **Electrodermal Activity (EDA)**: this signal reflects sweat secretion by measuring skin conductance. It can be broken into two basic components: tonic and phasic, providing slowly habituating measures of arousal and moment-by-moment measures of arousal which reflect stimulus-specific responses [50], respectively.

- **Skin Temperature (TEMP)**: is an indicator of the ”fight or flight” response produced by exposure to a stressful situation. A drop in temperature of the extremities can be observed when under this response [37].

In contrast to EEG signal, the measurement of peripheral physiological signals can be accomplished with convenient and less invasive wearable sensors such as wristbands, patches, smart clothes, etc. Thus, their applicability in real life is greatly encouraged. Overall, we chose peripheral signals for physiological-based emotion recognition in this thesis.

2.2.3 Behavioural signal vs Physiological signal

The application of behavioural or physiological modalities in emotion recognition systems has its own benefits and drawbacks. Behavioural signals have received considerable attention due to their non-invasive nature in data acquisition. However, their reliability cannot be verified, as people can autonomously control these physical manifestations to hide their true emotions (i.e., social masking) [51]. Peripheral physiological measures are still slightly more intrusive than non-contact behavioural modalities, however, they allow for a long-term objective assessment of the subject's true internal emotional state. Moreover, considering practical context, they may be more suitable than visual data such as facial expressions, whose performance may be significantly influenced by the location of the camera and the computational cost of image or video analysis is more expensive than time series. Naturally, there is a third option for emotion recognition systems, i.e., the integration of behavioural and physiological modalities, which generally results in more reliable decisions and can cope with missing modalities [13].

2.3 Emotion recognition based on peripheral physiological signals

In this section, emotion recognition approaches based on peripheral physiological signals are presented. Fig. 2.3 illustrates the primary steps included in these methods for predicting emotions. Raw physiological data collected in restricted or unrestricted environments usually need to be pre-processed, which mainly consists of filtering, segmentation, normalization, etc. A brief review of these these operations is given in Section 2.3.1. Once pre-processing is accomplished, clean data segments can be fed into various emotion recognition algorithms, which can be broadly classified into two categories, i.e., fully supervised learning-based algorithms and non-supervised learning-based algorithms. The former can be further subdivided into traditional machine learning-based methods and deep neural network-based methods, which are described in detail in Sections 2.3.2 and 2.3.3, respectively. In the end, Section 2.3.4 describes in detail the methods based on other learning paradigms: unsupervised learning and semi-supervised learning.

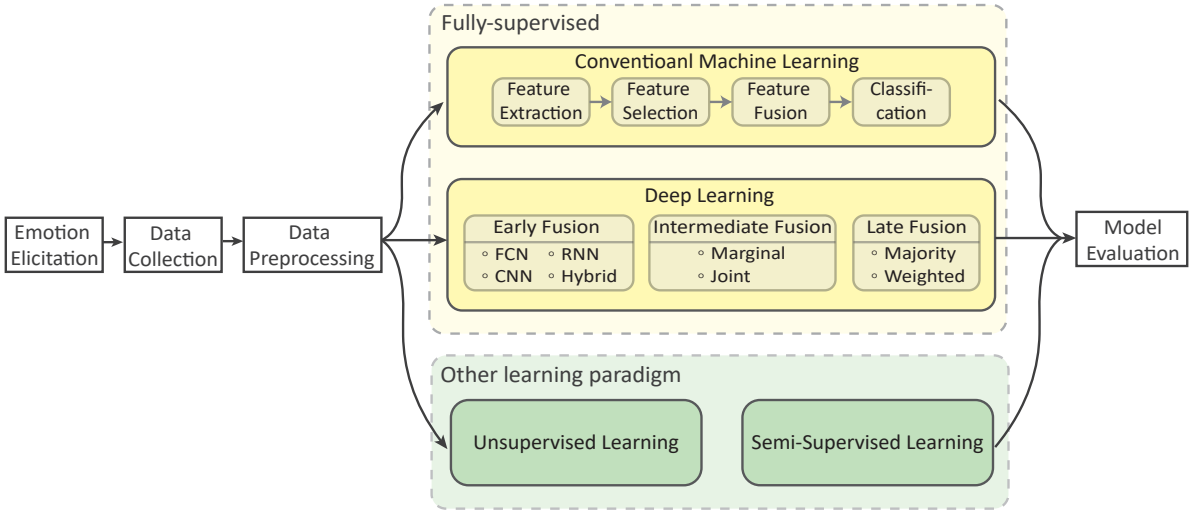


Figure 2.3: Emotion recognition pipelines based on peripheral physiological signals.

2.3.1 Physiological data preprocessing

The primary part of pre-processing is signal filtering, with the purpose of eliminating the impact of noise due to electromagnetic interference, motion artefacts, etc [51]. Since the data acquisition setting in the literature is diverse, resulting in inconsistent frequencies of the obtained data, hence there is not yet a unified standard for filter selection. For example, low-pass filters

with cutoff frequencies of 1 Hz and 2.5 Hz were applied for RESP and EDA signals in [52]. Schmidt et al. [53] employed a 0.1 Hz to 0.35 Hz bandpass filter for RESP signal while a 5 Hz low-pass filter for the raw EDA data. In [54], Butterworth bandpass filters with different orders and cutoff frequencies were utilized for ECG, EMG and BVP signals. After the filtering operation, the clean signal record is generally split into small data segments via a sliding window, from which meaningful patterns are subsequently extracted for identification. The time interval between an emotional stimulus and the physiological response it triggers tends to vary with different factors such as individual, signal modality, etc [11]. Therefore, it is challenging to define an appropriate window size for emotion recognition systems. In the work of Kreibig [55], they observed that the most common average duration of physiological responses was 60 s or 30 s in a survey of 134 publications. Other frequently occurring average intervals are based on 0.5 s, 10 s, 120 s, 180 s, or 300 s. Once the data is segmented, manual or automated feature engineering is performed to extract discriminative representations for emotion classification. For subject-independent recognition algorithms, the data are further normalized to remove inter-individual differences in physiological variables. The common normalization techniques are min-max normalization or z-score normalization [29]. The former scales the original data to the [0, 1] interval through the maximum and minimum values in the subject's data, while the latter converts the mean and standard deviation of the samples to 0 and 1, respectively.

2.3.2 Supervised machine learning-based methods

Traditional machine learning approaches typically require extracting and selecting the most relevant features from pre-processed physiological signals and feeding them into a series of classification algorithms for emotion recognition. This process can be further divided into four stages: feature extraction, feature selection, feature fusion and classification, which will be presented accordingly in the subsequent sections.

1) Feature Extraction

A majority of early research in wearable emotion recognition was based on machine learning, aiming to predict emotional states using crucial features manually extracted from preprocessed peripheral physiological signals in the time-domain, frequency-domain, time-frequency domain and non-linear domain as input to classification algorithms.

- Time domain From the temporal perspective, Hernandez et al. [56] employed a series of statistical featues from EDA time series such as duration, max, min, mean, standard deviation,

signal slope between the first and last points to detect stress in the workplace with a wearable wrist sensor. For ECG and BVP signals reflecting cardiac cycle activity, three important parameters: heart rate (HR), inter beat interval (IBI) and Heart Rate Variability (HRV) can be derived depending on the position of the signal peaks, from which rich emotionally relevant information can be computed. Bong et al. [57] suggested three time-domain features: heart rate (HR), mean R peak amplitude (MRAmp), and mean R-R intervals (MRRI) extracted from the ECG signal for a two-stage emotional stress detection. Kim et al. [58] calculated a series of statistical ECG features: mean, the standard deviation of all normal-to-normal(NN) intervals (SDNN), the standard deviation of the first difference of the HRV, the number of pairs of successive NN intervals differing by more than 50 ms (NN50), etc., for arousal-based emotion recognition. Zhang et al. [59] utilized 7 statistical features: min, max, mean, median, standard deviation, minRatio and maxRatio from EMG and EDA signals.

- Frequency domain An alternative method of extracting effective features is to convert the time domain signal to the frequency domain. Fast Fourier Transform (FFT) [58] was applied on HRV decuced from ECG signal to generate spectral features such as dominant frequency, power of very low frequency band (VLF: 0.003-0.04Hz), low frequency band (LF: 0.04-0.15Hz) and high frequency band (HF: 0.15-0.4Hz), and ratio of power LF/HF. Jerritta et al. [60] performed Hilbert Huang Transform (HHT) on normalized QRS derivative ECG data for calcultaing low and high frequency features.

- Time-Frequency domain Some studies [61, 62] focused on the time-frequency analysis using Wavelet Transform (WT). Guendil et al. [62] employed WT to decompose each signal into six levels and the resulting WT coefficients were treated as emotion-related components in the six frequency bands. Xie et al. [61] calculated WT related features: min, max, mean, and standard deviation in wavelet coefficients after decomposition for ECG, EMG and EDA signals.

- Non-linear domain Apart from the traditional temporal/spectral features, non-linear models have also been suggested for emotion recognition. Rubin et al. [63] derived 19 non-linear ECG features based on complex analysis such as sample entropy, maximal lyapunov exponent, correlation dimension for classifying states of panic and pre-panic. Valenza et al. [52] extracted various non-linear features based on Recurrence Plot (RP), Deterministic Chaos (DC) and Detrended Fluctuation Analysis (DFA) on ECG, EDA, and RESP signals for recognizing multilevel aoursal/valence states. Their experimental results indicated that the non-linear features enhanced the performance in comparison with the common time-frequency domain features.

2) Feature Selection and Reduction

Due to the complexity and individual variability of emotions, recognition methods often incorporate multiple physiological modalities to provide better performance [32], resulting in high-dimensional vectors that cause the curse of dimensionality. Furthermore, some redundant features are not beneficial for identification which may lead to overfitting problems and cause weak generalization of the learning model [51]. To solve these problems, feature selection algorithms are commonly applied to select the most relevent features for emotion recognition tasks. Some filter methods were employed, which are independent of the learning model and filter features by statistical metrics. For example, Yan et al. [54] used Mutual Information (MI) to reduce the feature dimension. Ayata et al. [64] suggested the use of mRMR (minimum redundancy maximum relevance) algorithm to select the most appropriate feature subset for emotion classification. However, the drawback of the filter method is that it completely ignores the impact of the selected features on classification algorithm performance [29]. To address this issue, wrapper methods were proposed which asses the quality of selected features by introducing a classifier. In the work of [58], Sequential backward selection (SBS) was applied, which is a top-down method that iteratively removes a feature from the entire feature set whose elimination improves the classification performance until the number of features is satisfied. Xie et al. [61] selected a similar approach to SBS, Sequential forward selection (SFS), which starts with an empty set. Their experimental results showed that by applying SFS, the performance of the classifier is significantly improved by about 30%. However, the high computational cost associated with exhaustive search on features make wrapper methods impractical for real-life applications [29]. In contrast to these previous time-consuming methods, some approaches adopted dimensionality reduction techniques, such as Principal Component Analysis (PCA) [52] which maps data to a lower dimensional space while retaining the components that contribute most to the variance. To avoid the information loss arising from mapping, a Kernel Principal Component Analysis (KPCA) method was applied in [59], which first projects features into a high-dimensional space using Radial Basis Function (RBF) kernel for non-linear information extraction, and then reduces the dimensions similarly to PCA.

3) Feature Fusion

Features from different physiological modalities are generally fused in an early fashion or a late fashion to yield superior performance over unimodal recognition methods. Early fusion [52, 58, 59, 62, 65] is a simple and intuitive strategy, where unimodal features are concatenated to form a single high-dimensional vector as input to the classifier. However, this approach is

struggling with missing data and multimodal asynchrony issue [51]. Decision fusion can effectively tackle these problems, where classifiers learn independently on unimodal data and the final decision is a combination of predictions from all classifiers. Majority voting [61, 64, 66] is widely deployed for emotion recognition, where the most frequent prediction is served as the final decision. However, this approach ignores the relationships between different modalities, since each classifier is trained independently. To enhance the robustness of the decision, an adaptive decision fusion strategy [54] was proposed which first constructed sub-classifier weights by computing accuracy matrix, correlation coefficient, instability coefficient and linearly merged the weighted classification results from each classifier. Though this method allow us to capture the cross-modal interaction between different predictions, however, the connections among multimodal features are ignored.

4) Classification

Common machine learning based classifiers in the literature are Support Vector Machine (SVM) [54, 56, 57, 59, 61–63, 65–67], K-Nearest Neighbour (K-NN) [49, 57, 59, 61, 63, 66, 67], Random Forest (RF) [61, 63, 64, 67], Gradient Boosting Decision Tree (GBDT) [59, 61, 67], Linear Discriminant Analysis (LDA) [58, 66], etc. Among these learning models, the support vector machine (SVM) is the most frequently applied in the field of physiological emotion sensing [11]. Some studies [61, 65, 67], have found that the SVM can achieve better performance in emotion recognition than other classifiers. For instance, Cheng et al. [67] evaluated various classifiers using ECG and HRV features from linear, non-linear, time and frequency domains, mong which SVM achieved the highest accuracy of 79.51% on the task of detecting positive and negative emotions. In [61], wavelet transform features of different modalities were extracted for the binary emotion classification, and ultimately the best accuracy of 94.81% was achieved by fusing the SVM classifiers from ECG and EMG signals. Nevertheless, some studies [59, 63] have revealed that the ensemble methods may yield superior performance. In [63], the fusion of time-domain, frequency-domain and non-linear ECG features was fed into a series of classifiers for evaluation, where the RF demonstrated the best performance, achieving an accuracy of 97.2% and 90.7% for panic and pre-panic detection, respectively. Zhang et al. [59] implemented four classifiers: GBDT, SVM, KNN and Gaussian Naive Bayes (GaussianNB) for performance evaluation. With the property of auto-interaction over multiple feature sets, The GBDT outperformed other classifiers in all experimetal settings, achieving an accuracy of 93.42% in 4-class emotion recognition task.

5) Discussion of supervised machine learning-based methods

Table 2.1 briefly summarizes the supervised machine learning-based approaches. While these approaches have demonstrated impressive results, they present certain limitations in the feature engineering process. First, emotion-related features are required to be extracted for each modality. This relies heavily on domain knowledge which results in difficulties of transferring the model to other tasks, especially in the multimodal case. Second, the most reliable features should be selected to cope with the curse of dimensionality introduced by multimodal data while obtaining better recognition performance. However, this process is quite time-consuming and may lead to loss of information. In addition, the above two steps limit the flexibility of fusing multimodal data at different levels since one has to always consider the modalities involved as well as the features extracted and selected.

Table 2.1: Overview of supervised machine-learning based methods for physiological emotion recognition. (EF: early fusion; LF: late fusion.)

Modality	Type	Selection	Fusion	Classification	Paper
ECG	Time	-	-	KNN, SVM	[57]
ECG	Freq	-	-	KNN	[60]
EDA	Time	-	-	SVM	[56]
ECG	Mix	-	-	RF, PA, GB, DT, RC, SVM, KNN, LR	[63]
ECG	Mix	-	-	SVM, RF, DT, KNN, GBDT	[67]
ECG, EMG, EDA, RSPs	Mix	SBS	EF	pLDA	[58]
ECG, EDA, RSP	Non-linear	PCA	EF	QRC	[52]
ECG, EMG, EDA, BVP	Mix	KPCA	EF	GBDT, SVM, KNN, GBN	[59]
ECG, EMG, EDA, RSP	Time-Freq	-	EF	SVM	[62]
EDA, EKG, RSP	Mix	SBL-PCA	EF	SVM, ELM	[65]
ECG, EMG, EDA	Time-Freq	SFS	LF	SVM, RF, NB, DT, KNN, GBDT	[61]
ECG, EDA, TEMP	Mix	-	LF	SVM, KNN, QDA, LDA	[66]
BVP, RSP, TEMP	Time	mRMR	LF	RF	[64]
ECG, EMG, EDA, BVP	Mix	MI	LF	SVM	[54]

2.3.3 Supervised deep learning-based methods

Deep learning-based methods have recently gained extensive attention in the field of emotion recognition and have been shown to outperform machine learning methods in several studies [15, 53, 68–70]. Its major advantage is the automatic mining of complex and abstract features in the raw input, through multi-level non-linear transformations introduced by successive layers, thus drastically simplifying the process of feature extraction and selection compared to machine learning methods [28]. Due to their automatic representation learning properties, deep learning approaches for physiological emotion recognition typically employ multiple modalities to

achieve more accurate and robust performance. This raises a crucial challenge for emotion recognition based on multimodal physiological signals, i.e. how to capture the marginal and joint representation of multiple modalities and enable their combination to be beneficial for model performance? Marginal representation is meaningful pattern discovered from unimodal input and is capable of modelling heterogeneity of multimodal signals (i.e., intra-modal correlations), while joint representation is the encoding of potential complementary or cooperative relationships in multimodal data (i.e., cross-modal correlations). Fig. 2.4 shows an illustration of the deep learning-based fusion strategies for multimodal physiological signals. Three fusion strategies are proposed to attain the above two representations, denoted as early fusion, intermediate fusion and late fusion [71]. In the following sections, each of these three fusion techniques and their corresponding emotion recognition methods are described. Additionally, for the latter two fusion modes, approaches employing EEG signals were also presented as very few studies targeted exclusively peripheral physiological signals.

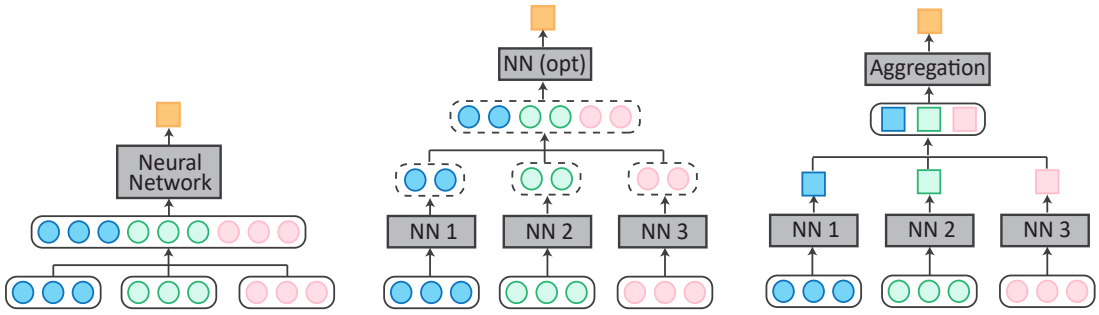


Figure 2.4: Deep learning-based fusion strategies for physiological emotion recognition. a) Early fusion employs concatenated multimodal data as input and extracts efficient inter-modal correlation information through joint learning. b) Intermediate fusion first learns marginal representations for each modality independently, which are then combined and fed directly into the decision layer or continue to learn advanced joint representations. c) Late fusion aggregates decisions derived from marginal representations of unimodal data to generate final decisions. (NN: Neural Network; NN (opt): optinal neural network for joint leanring.)

1) Early Fusion

Early fusion-based approaches typically treat features from different modalities as a unity, from which the joint multimodal representation is learned directly. There exists two approaches for the construction of multimodal inputs, depending on whether or not dependencies in physiological signals are considered. The first is to directly concatenate the raw data or features of individual modalities into a 1D high-dimensional vector, regardless of the time scale. The

second one stacks multimodal data at different instants as a 2D matrix with the dimension determined by the number of modalities (or features) and the number of time steps. Consequently, different types of deep learning architectures can be applied to emotion recognition tasks based on the structure of the input data. Fig. 2.5 illustrates three common deep models in physiological emotion recognition approaches which will be described separately in the following paragraphs.

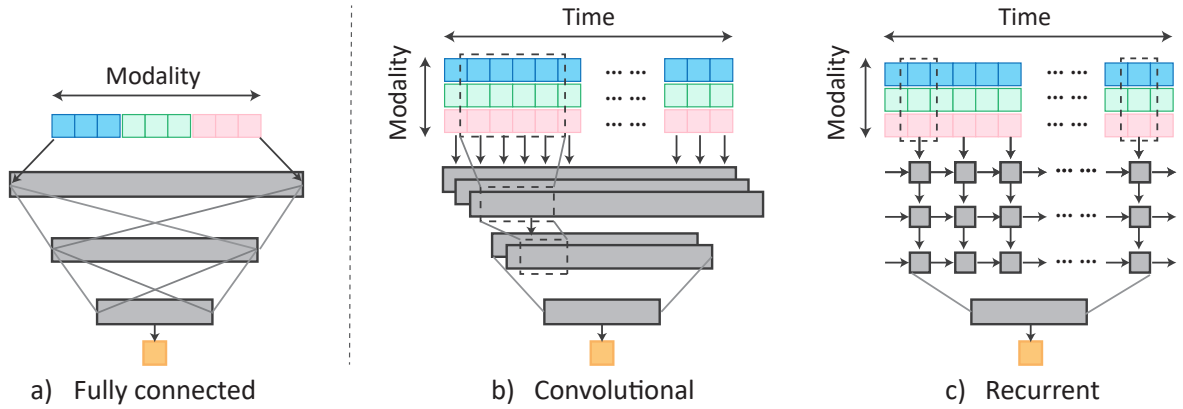


Figure 2.5: Typical early fusion-based emotion recognition architectures. a) Fully connected neural network can be applied directly to multimodal vectors when the temporal dependencies between data samples are ignored. b) Convolutional neural networks can compute temporal dependencies on local scales in 2D multimodal inputs in a parallel manner. c) Recurrent neural networks capture long-term dependencies in physiological sequences in a sequential manner.

- Fully connected The 1D concatenated multimodal vector can be passed directly through fully connected deep networks to learn complex transformations. For example, Saeed et al. [72] proposed a Multi-Task fully connected Neural Network (MT-NN) for personalized driver stress detection. A vector of 16 time-domain features extracted from the EDA and HR signals is first fed into a shared layer to learn a cross-task joint representation, followed by a subject-specific layer to cope with individual differences in stress responses. Ultimately, dense layers with sigmoid cells were used for person-specific stress detection tasks. Experimental results indicated that MT-NN can further improve performance compared to coventional machine learning models such as logistic regression (LR) and support vector machine (SVM). In [73], a similar MT-NN based framework was proposed for subject-dependent pain detection task. Their fully-connected architecture with ECG and EDA features also achieved superior performance over traditional machine learning methods. Despite the promising results of the previous approaches, the fully-connected model did not incorporate the temporal dependencies between physiological signal values.

- Convolutional While convolutional neural networks (CNNs) are extensively employed for tasks involving 2D images, considerable researches have applied them for time series analysis [74]. Such models apply and slide a series of time-invariant filters in a parallel fashion over the input sequences to capture the local dependencies between samples. Multimodal physiological signals are multivariate time series which can be sent into CNN-based models to learn discriminative representations for emotion recognition tasks. In [68], a CNN model consisting primarily of 1-dimensional (1D) convolutional layers and pooling layers was employed to extract abstract representations from ECG and EDA signals. Fully-connected layers were positioned at the end of the model to output prediction probabilities. Finally, the proposed model outperformed classic machine learning algorithms in identification of arousal and valence states. Wang et al. [75] proposed a 1D CNN based architecture for stress detection in real-world driving environment. Change point density was first calculated for each filtered physiological signals, where a change point is defined as a point with a sharp increase or decrease in signal value. The physiological signals were then randomly shifted and combined with the corresponding change point density to form the multichannel input to the 1D CNN. Their experimental results indicated that the proposed model generalized well on stress detection task for different subjects with similar cognitive abilities. Despite the impressive results of CNN-based approaches for emotion recognition tasks, CNNs are not sensitive to temporal order (not beyond the local scale) and are thus ineffective in modelling long-term dependencies [74].

- Recurrent Recurrent neural networks (RNNs) and their gated variants such as the long short-term memory (LSTM) are traditional deep neural networks for processing time-series data, whose decisions at current time steps are influenced by past time steps, thus capturing the temporal dependencies of elements in large scale sequences [74]. Several efforts have been conducted using recurrent models to fully explore the temporal relationships in physiological sequences. For instance, Awais et al. [76] proposed an intelligent remote Internet of Things (IoT) framework for real-time emotion recognition, where wearable devices and IoT technology allowed for wireless data collection and communication, and six peripheral physiological signals were fed into a deep LSTM-based model to classify four categorical emotions, achieving a high performance (F1-score) of 95 %. Zitouni et al. [77] applied a Bidirectional LSTM-based (BiLSTM) architecture for emotion detection in real-life debates using three peripheral signals: EDA, HR and TEMP. In comparison with the baseline models, i.e. Gaussian Naive Bayes (GNB) and eXtreme Gradient Boosting (XGBoost), the proposed model obtained the best performance with an average accuracy of 90.79% for arousal, 90.53% for valence and 86.18% for 4-class classification. Although RNNs have performed well in the above studies, they still

exhibit some limitations. Theoretically, RNNs can model arbitrarily long sequences, However in practice, vanilla RNN, including LSTM face vanishing gradient problem when dealing with extra long series. Futhermore, the sequential processing makes the model training extremely time consuming compared to CNNs.

- Hybrid In order to take advantage of both the parallel computation of CNNs and the long-term dependency modelling of RNNs, some researches have adopted hybrid models on emotion detection tasks. For example, Keren et al. [78] proposed a CNN-LSTM model where convolution and pooling operations first compressed the input sequence to obtain a compact high-level representation, which is subsequently fed into the LSTM to further capture underlying temporal correlations in sequential data. In experiments of identifying arousal and valence, the hybrid model consistently produced better results, compared to methods based on hand-crafted features.

In general, early fusion proved to be the preferred option for most emotion recognition methods based on peripheral physiological signals due to its simplicity of installation, where the joint representation can be extracted directly from the combined multimodal input, emphasizing the importance of inter-modal interactions. However, it neglects the necessity of designing special sub-models to further capture intra-modal properties.

2) Intermediate Fusion

Early fusion treats multimodal data as a whole, where marginal representations are not learned explicitly and hence cannot differentiate which modality the learned features are derived from. Intermediate fusion can effectively solve this problem by first learning marginal representations to capture intra-modal correlations, thereby injecting the priori knowledge into the model. Subsequently, these bottom marginal representations are either fed directly into the decision layer (i.e., marginal intermediate fusion, presented in Fig. 2.6 (a)) or jointly learn a higher-level representation for prediction (i.e., joint intermediate fusion, presented in Fig. 2.6 (b)). Later on, we will discuss emotion recognition methods based on these two intermediate fusion strategies.

- Marginal In marginal intermediate fusion, representations extracted from different modalities can be directly concatenated as input to the classification. Ma et al. [79] proposed a multimodal residual LSTM (MMResLSTM) network for emotion recognition. The EEG signal and other peripheral signals: EOG (electrooculography) and EMG (electromyography) are fed separately into four-layer LSTM based model interspersed with residual connections and layer

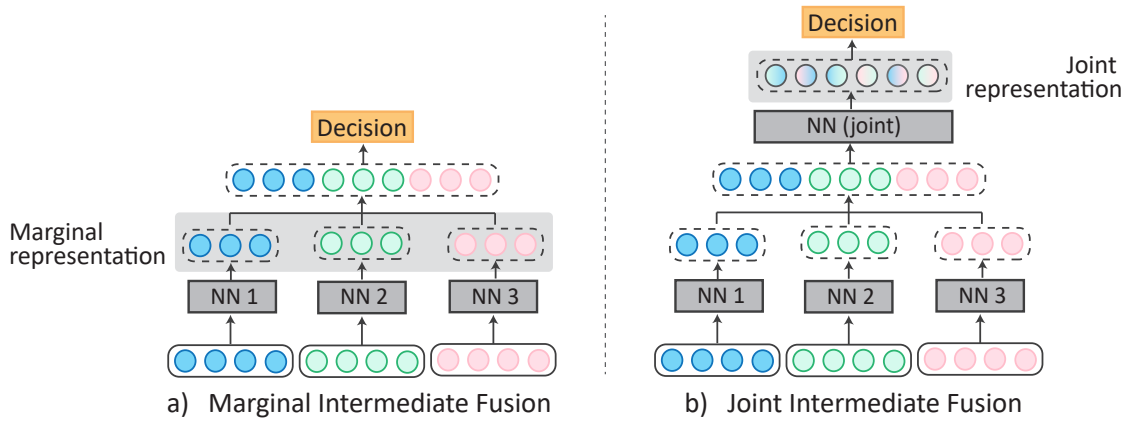


Figure 2.6: Common intermediate fusion strategies for physiological emotion recognition. a) Integrated marginal representation is applied directly to decision making. b) The marginal representations are first combined to jointly learn a high-level multimodal representation for better exploiting inter-modal correlations, which is later fed into the decision layer.

normalization to efficiently learn discriminative temporal features, and eventually the marginal representations of the two branches are concatenated and sent to a fully-connected layer with softmax function for predicting emotional arousal and valence. Zhu et al. [80] designed a emotion recognition framework based on multi-hypergraph neural networks (MHGNN). Three hypergraphs were first built from three physiological modalities (EEG, EOG and EMG), where the vertices and edges of the graph represent the subjects described by the stimuli and the latent correlations between subjects, respectively. Then, two-layer hypergraph neural networks acting on each modality extracted representations encoding both intra-modal and cross-subject correlations. Finally, a dense layer accepted all marginal representations and classified them into different emotion categories. In [81], a hybrid model was proposed for emotion recognition in arousal-valence space, where 2D-CNN was applied to extract spatial features from multichannel EEG signals and temporal features of multiple peripheral signals were captured via LSTM, and ultimately multimodal features were concatenated for fully connected-based classification. The marginal fusion-based approach described above fully leveraged intra-modal correlations, however, cross-modal correlations may not be effectively modelled. Therefore, there exists the potential to further enhance recognition performance by exploring inter-modal cooperation.

- Joint Marginal representations from different branches can also be concatenated and passed through a shared encoder to obtain a joint representation for classification, which exploits the complementary and collaborative nature of the modalities. Zhang et al. [82] proposed an emotion recognition framework based on regularized deep fusion of kernel machine. Low-level unimodal representation obtained from ensemble deep kernel machine optimization (eDKMO)

and bimodal joint representations aggregated by a fully connected layer formed the final representation set, which are fed into the global fusion layer with regularization to generate the final fused representation for classification. This type of method captures cross-modal relationships by including additional integration layers to further enhance the recognition performance. However, this increases the complexity of the model which may result in overfitting, especially on limited physiological data.

Overall, intermediate fusion extracts marginal and joint representations in a hierarchical fashion to simultaneously capture intra- and inter-modal relationships. However, few researches have explored the performance of emotion recognition purely based on peripheral physiological signals.

3) Late Fusion

Considering that different modalities can provide distinct information for emotion recognition, separate sub-models can be trained on unimodal data with their resulting decisions being combined to produce the final decision. According to the fusion techniques of sub-decisions, we divide the existing emotion recognition methods into two categories, namely majority voting-based methods and weighted voting-based methods, both of which will be discussed separately in the following.

- Majority voting The majority voting is considered to be the most intuitive late fusion solution, with the most voted category as the classification result. Dar et al. [83] proposed a hybird framework for multimodal emotion recognition. A 2DCNN-based model and a 1DCNN-LSTM-based model were applied to EEG signals and peripheral physiological signals (i.e., ECG, EDA) respectively to generate modality-specific decisions. The final predictions regarding the quadratic classification of arousal-valence were determined based on the majority voting strategy. There are certain limitations to this approach, e.g. a subset of the classifiers (which are quantitatively dominant) may occasionally give the incorrect classification at the same time [84].

- Weighted voting An alternative approach for decision fusion is to average the outputs of all sub-models, implying that all branches are allocated equal weights. Hssayeni et al. [85] explored the effectiveness of this fusion technique, in which a 1DCNN-based architecture was trained independently on different peripheral physiological signals. The class probabilities generated by the CNNs were averaged as the final output of the model. In experiments for esti-

mating positive and negative emotions, this late fusion approach exhibited better performance against early fusion. Considering that the contribution of different modalities to emotion recognition may be unequally distributed, different weights should be assigned to individual sub-decisions. For example, Li et al. [86] proposed a variable weight fusion strategy for multimodal emotion recognition. EEG, ECG and EDA signals were first converted into spectrograms, an Attention-based Bi-LSTM model was then applied to extract the most relevant time-frequency domain features from each signal. Finally, the predictions from all sub-models are fed into the fully connected layer to automatically learn the corresponding weights for each modality. In the classification experiments of arousal and valence states, the proposed fusion strategy significantly improved the performance of the equal weight scheme.

On the whole, late fusion focuses on marginal representations which can encode intra-modal correlations and this is particularly effective for data with certain modalities missing or extremely heterogeneous data. However, its major drawback is its inability to capture the interaction between different modalities, especially for highly relevant physiological signals in particular emotional states.

4) Discussion of supervised deep learning-based methods

In the previous sections we reviewed multimodal physiological emotion recognition approaches based on the three fusion strategies, a brief summary of which is presented in Table 2.2. Among the three fusions, early fusion and late fusion can only capture inter- and intra-modal correlations, respectively, while intermediate fusion is encouraging for efficient encoding of both intra-modal and inter-modal associations at different levels. However, it can be clearly observed from Table 2.2 that there are no existing intermediate fusion methods that specifically target low-frequency peripheral physiological signals. From the perspective of real-life applications, EEG signals are inappropriate in this context due to their strong invasive nature. Hence, there is a need to develop such fusion methods that rely solely on peripheral physiological signals. Moreover, for supervised deep learning methods, training a sufficiently accurate and generalizable model commonly depends on a large amount of labeled data, which is challenging for physiological data, as the annotation is time-consuming, expensive, and requires the intervention of domain experts. In addition to the quantitative limitation, the quality of the collected labels can also be a problem. This is particularly true in real-life situations where the ground truth of emotions is based on self-report, which can introduce subjective errors.

Table 2.2: Overview of supervised deep learning-based methods for physiological-based emotional recognition.

Fusion	Modality	Real-life?	Taxonomy	Papers
			Architecture	
Early	EDA, HR	Yes	Fully connected	[72]
	ECG, EDA	Yes	Fully connected	[73]
	ECG, EDA	Yes	Convolutional	[68]
	EDA, BR, HR, HRV	Yes	Convolutional	[75]
	BVP, ECG, EDA, EMG, RESP, TEMP	Yes	Recurrent	[76]
	EDA, HR, TEMP	Yes	Recurrent	[77]
	ECG, EDA HR	Yes	Hybrid	[78]
			Representation	
Intermediate	EEG, EMG , EOG	No	Marginal	[79]
	EEG, EMG, EOG	No	Marginal	[80]
	EEG, EMG, EOG, EDA, BVP, RESP, TEMP	No	Marginal	[81]
	EEG, EMG, EOG, ECG, EDA, RESP	No	Joint	[82]
			Aggregation	
Late	EEG, ECG, EDA	No	Majority	[83]
	ECG, EMG, EDA, RESP	Yes	Weighted	[85]
	EEG, ECG, EDA	No	Weighted	[86]

2.3.4 Other learning paradigm based methods

In the previous section, we have demonstrated that deep neural network-based approaches are effective in feature extraction and fusion, thus achieving superior performance over traditional machine learning-based approaches. However, developing fully-supervised deep fusion models for emotion recognition in a mobile environment using only wearable peripheral physiological signals still suffers from the following problems: 1) How to train a generalised model using limited labelled data? and 2) How to deal with the biases introduced by inaccurate and uncertain ground truth collected in real-life settings? Since we always have access to large amounts of unlabeled data, recent researches have shifted towards other learning schemes: unsupervised learning, or semi-supervised learning with the intention of leveraging unlabelled data to discover the true distribution of samples. In general, unsupervised learning uses entirely unlabelled data, while semi-supervised learning uses a mixture of unlabelled and labelled data. In the following, we will present physiological emotion recognition methods based on these two learning strategies.

1) Unsupervised

Unsupervised learning algorithms expect to automatically seek hidden structures in data without relying on supervision from labels [87]. A common technique technique applied in unsuper-

vised learning-based emotion recognition methods is data clustering, which aims to allocate data to different clusters by maximizing intra-cluster similarity while minimizing inter-cluster similarity [88]. For instance, Birjandtalab et al. [89] explored the performance of unsupervised emotion recognition algorithms using multimodal wearable physiological signals consisting of EDA, HR and arterial oxygen level (SPO2). A Gaussian mixture (GMM) model was applied to the clustering of multimodal features, ultimately achieving an accuracy of over 84% in the task of differentiating between four emotional states. Fiorini et al. [90] examined the validity of unsupervised learning models with ECG, HR and brain activity signals. Multimodal features extracted from the time and frequency domains were fed into three unsupervised algorithms: K-Means, K-medoids and Self-organizing maps. Although the above methods can solve the problems associated with data labelling, however they are all based on hand-crafted features.

For the automatic learning of more complex non-linear representations, there is an increasing trend towards the use of deep neural networks for unsupervised learning. A typical model is an autoencoder, which extracts meaningful representations through the compression and reconstruction of the unlabeled data. Several studies have explored the feasibility of this technique for emotion recognition. In [91], stacked convolutional autoencoders (SAE) were applied independently on unlabelled ECG and EDA data to obtain generalized latent representations for arousal classification, achieving better performance than the fully supervised approaches. Though this method effectively modeled the heterogeneity of multimodal signals, i.e. using different models to extract valid unimodal features, however, it neglected the collaborative and complementary nature of multimodality. Different from the previous approach, Zhang et al. [70] presented a correlation-based emotion recognition algorithm (CorrNet), where intra-modal features are first obtained with separate convolutional autoencoders, followed by covariance and cross-covariance computation between each pair of modalities to obtain inter-modal features. However, these unsupervised learning methods based on autoencoders did not introduce supervised signals in pre-training. As a result, the learned representations may contain patterns that are irrelevant to emotions, resulting in unsatisfactory performance.

2) Semi-supervised

Semi-supervised learning typically uses both a large amount of unlabelled data and a small amount of labelled data, assuming that the presence of unlabelled data contributes to a better learning process [92]. Depending on the training strategy, semi-supervised learning methods can be classified into one-stage based or multi-stage based [93]. Most of the semi-supervised emotion recognition methods belong to the latter category, which use unlabelled data for pre-training to provide a good initialisation of the model, followed by fine-tuning on labelled

data. For example, Luo et al. [94] presented a semi-supervised learning algorithm for Valence-Arousal-Dominance emotion recognition based on EEG signals and peripheral physiological signals. Stacked denoising autoencoders (SDAE) was first pre-trained and then fine-tuned in a supervised manner to generate robust representations that were ultimately served as input to KNN and SVM for classification. Finally, the proposed method showed slight improvements over other supervised and semi-supervised methods in the performance comparisons. However, as with [91], it adopted independent autoencoders for reconstruction of different modalities and thus failed to effectively capture the cross-modal correlation of multiple data. Yu et al. [95] proposed a semi-supervised multimodal stress detection framework. A LSTM autoencoder-based (LSTM-AE) model was first trained on labelled samples and the resulting latent representations were clustered with a GMM model, which was subsequently applied to select unlabelled samples with a similar distribution to the labelled data. Later, the selected unlabelled data was applied to pre-train the LSTM-AE model providing the initial parameters of the encoder in the supervised architecture. In the end, the proposed method enhanced the classification performance by 7.7% to 13.8%, compared to the supervised approaches. Though the methods described above have yielded promising results, allowing us to obtain robust and generalized emotion recognition models with limited labeled data. However, their approach regarded the multimodal data as a whole. The autoencoder-based model was designed to compress and reconstruct the concatenated hand-crafted feature sequences, thus ignoring the correlation information within the modalities.

3) Discussion of non-supervised learning based methods

In the above sections, we explore non-supervised learning approaches for emotion recognition, some of which have shown competitive or even better performance compared to fully supervised algorithms, offering a potential to eliminate overfitting effects associated with limited labeled data. Table 2.3 summarizes the corresponding methods demonstrated in the previous sections. As can be observed from the table, first, the number of non-supervised studies, whether based on traditional machine learning algorithms or deep neural networks, remains significantly lower than that of fully supervised ones. Second, for deep unsupervised/semi-supervised learning methods, the autoencoder appears to be a popular model option for learning generalized representations. However, these methods cannot simultaneously model the heterogeneity and collaboration of multimodal physiological signals when pre-trained with unlabeled data; in other words, they typically employ a single model or independent models for representation learning, thus ignoring intra- and inter-modal correlations, respectively. In summary, how to learn meaningful representations from large amounts of unlabeled data while effectively integrating

multimodal physiological signals are crucial factors for accurate and robust emotion recognition in real-life settings.

Table 2.3: Overview of other learning paradigm- based methods for physiological emotion recognition. (PPS: a series of Peripheral Physiological Signals, SAE: stacked autoencoder; SDAE: stacked denoising autoencoders, $C_{intra/inter}$: columns used to indicate whether the deep model takes into account intra- and intermodal correlations in the pre-training phase.)

Learning scheme	Modality	Real-life?	C_{intra}	C_{inter}	Architecture	Papers
Unsupervised	PPS	Yes	-	-	Hand-crafted	[89]
	EEG, PPS	No	-	-	Hand-crafted	[90]
	PPS	Yes	✓	✗	SAE	[91]
	PPS	Yes	✓	✓	CorrNet	[70]
Semi-supervised	EEG, PPS	No	✓	✗	SDAE	[94]
	PPS	Yes	✗	✓	LSTM-AE	[95]

2.4 Emotion recognition based on facial expressions

The process of recognizing human emotions based on facial cues is referred to as facial expression recognition (FER) and its corresponding methods can be broadly categorized into two types, namely appearance-based methods and geometry-based methods. The former encodes the pixel intensity information of a still image or video frame, while the latter analyses the location information of detected facial key points around the main facial components such as eyebrows, eyes, nose and lips in static images or videos. These detected 2D or 3D points are known as facial landmarks and can be considered as a compact representation of a facial shape. Compared with appearance-based methods, facial expression recognition methods based on landmarks have several advantages: first, they are faster and more efficient for long-term emotion recognition in real-world scenes since they only require the tracking of low-dimensional point sets. The number of facial landmarks used in different methods may vary, with the most common being 49 points [96, 97], 66 points [98, 99], or 68 points [100, 101]. An illustration is shown in Fig. 2.7. This is computationally less expensive than analyzing pixel intensity information. Second, they can avoid the privacy issues typically related to appearance-based methods, which require storage or transmission of the original facial images. Thus, landmark-based methods offer a more privacy-friendly option for practical applications. Third, landmark detection and tracking techniques are relatively insensitive to illumination variations in images, and recent methods are even robust to occlusions. Therefore, we only considered the facial landmark-based expression recognition approach in our work.

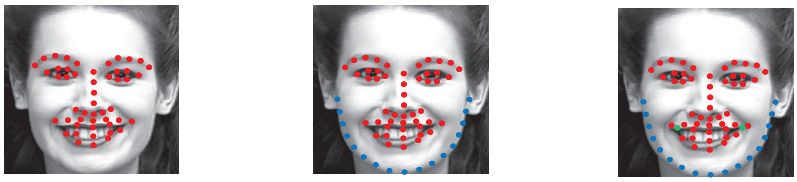


Figure 2.7: Common facial landmark configurations in the literature, where the 66 points-based configuration has 17 more landmarks (i.e. the blue dots) on the face contour than the 49 points-based configuration, and the 68 points-based configuration is formed by adding 2 additional landmarks (i.e. the green dots) on the inner corner of the mouth to the 66 points-based configuration.

Fig. 2.8 demonstrates the common steps included in the landmark-based FER systems. The acquired emotion-related facial images are subjected to a pre-processing consisting of face detection and alignment to remove irrelevant variations of facial expression from the data so as to normalize the visual semantic information [102]. A brief introduction to pre-processing is provided in Section 2.4.1. Subsequently, valid facial shape representations can be captured for a range of emotion-related classification or regression tasks. Depending on the type of feature support, we classified the corresponding recognition methods as point-based methods, local region-based methods, and full face-based methods, which are presented in Sections 2.4.2, 2.4.3 and 2.4.4, respectively.

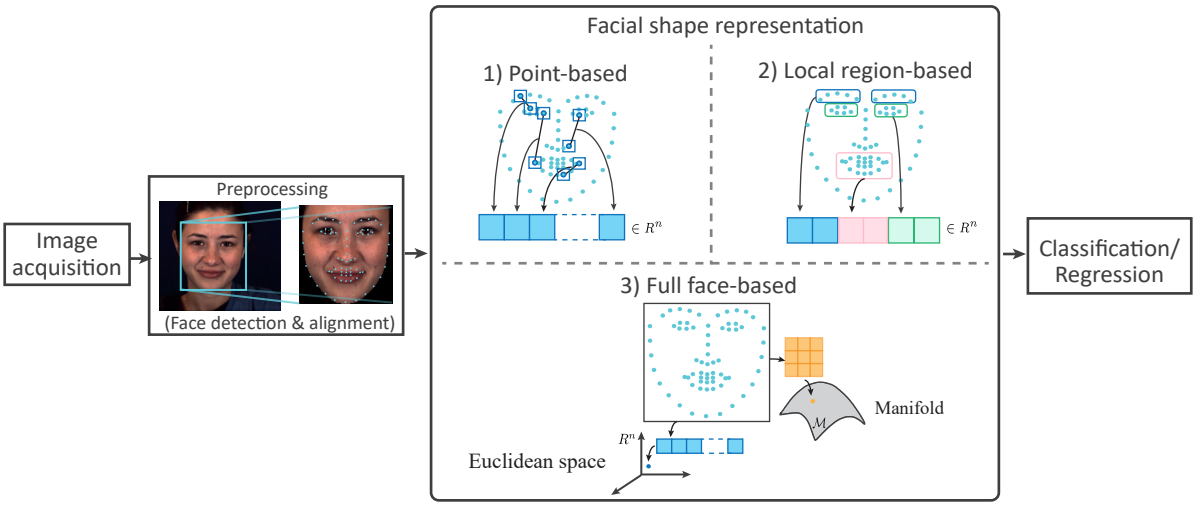


Figure 2.8: Landmark-based facial expression recognition pipeline.

2.4.1 Visual data preprocessing

In this section, we briefly introduce the two basic steps before performing facial expression recognition. Given an image containing the subject, the first step of data preprocessing is to determine the location of the region of interest (ROI), i.e., the face. One of the most widely employed face detection algorithms is the Viola and Jones (V&J) object detection framework, which can achieve accurate results on data collected in a controlled environment [103]. When migrating to real-world applications, this task tends to be challenging, as the human head posture may be non-frontal and obscured by hand motions or glasses, etc. Some robust face detection algorithms [104, 105] based on the Deformable Parts Model framework [106] were developed to cope with this situation. Once face detection is accomplished, facial landmark detection algorithms such as the Active Appearance Model (AAM) and the Supervised Descent

Method (SDM) are then employed to locate points of critical facial areas. The obtained facial points are further applied to compute transformations (e.g., Procrustes transformation) with predefined facial shapes to align face images, thus eliminating variations including translations, uniform scaling, and rotations [102]. At the end of these two operations, the normalized shapes can now be exploited to extract facial representations for recognizing emotions.

2.4.2 Point-based feature support

Approaches using point-based representations generally select specific facial landmarks from local regions (i.e., primary components of the face) which are further exploited to derive effective features. Depending on whether dependencies between facial landmarks were considered, such methods can be further classified into independent feature-based and non-independent feature-based. For example, Asmara et al. [107] tracked 20 facial points located in the eyebrow, eye and mouth regions in the facial expression video, and calculated the shifted Euclidean distance of the coordinate points between consecutive frames. In this case, the motion features of each facial landmark were encoded independently for facial expression classification.

Considering the relationship between facial points when expressing emotions, Munasinghe [108] calculated the distances between the points within the eyebrow and mouth regions as crucial information for identifying emotions. Taking into account the variation in facial sizes, the obtained distance vectors were subsequently normalized by a specific distance in each region. In the end, a random forest (RF) classifier achieved an average accuracy of 90% in predicting four emotions (i.e., anger, happiness, sad, and surprise). Huang et al. [109] proposed a 1D convolution-based neural network to explore the geometric information in facial landmarks. Vectors of normalized distances between facial points and a reference point (i.e., the point on the tip of the nose) were first computed and subsequently fed into 1D convolutional layers and pooling layers to extract discriminative features for pain intensity estimation. Ryumina and Karpov [110] proposed a deep geometric facial expression recognition framework based on Long Short-Term Memory (LSTM). They first selected random samples of facial expressions and calculated the Euclidean distances between facial landmarks. Subsequently, three emsemble classifiers: Random Forest Classifier (RF), Extra Trees Classifier (ET) and AdaBoost Classifier (AB) were adopted to calculate the average importance score of the features for filtering the most relevant distances. An LSTM-based deep neural network was applied to capture the temporal relationships in the selected facial distance sequences and achieved superior expression recognition performance than the appearance-based approaches.

In contrast to the previous studies based on facial landmark pairs, Ghimire et al. [111] pro-

posed a triangle-based representation for a better description of the relationship between points. Three landmarks were randomly chosen as the triangle's vertices, and the derived representation contained four components, i.e., two side lengths and two angle differences between the current frame and the first frame. Ultimately, the most discriminative triangle features were selected by a multi-class AdaBoost with Extreme Learning Machine (ELM) classifier and applied to the SVM-based classification. Palestra et al. [112] employed a hybrid representation containing 32 geometric features for facial expression recognition. The distance between two points, the slope of the line connecting the two points, the area of the irregular polygon, and the ratio of the major and minor axes of the defined ellipse were extracted from the regions of eyebrows, eyes, cheeks, nose, and mouth, respectively, as inputs to a series of classifiers. As a result, the best performance of up to 95.46% was obtained by the RF classifier. Though the above local representation-based approaches exhibited promising performance in facial expression recognition, the geometric features extracted from selected facial landmarks cannot fully describe the complex associations between facial components.

2.4.3 Local region-based feature support

This type of approach typically divides facial landmarks into several groups based on the main components of the face, from which corresponding local features are extracted and combined to form the final representation for the learning task. Liu et al. [113] applied separate LSTM models to process landmarks detected in facial sub-regions such as eyes, eyebrows, and mouth, where landmark coordinate series were fed into the LSTM cells to explore their relative position dependencies. In the end, the geometric features extracted from seven local regions were concatenated and weighted by a convolution-based attention model for expression recognition. Zhang et al. [114] proposed a Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN) for identifying facial expressions. The facial landmarks were first divided into 4 parts based on facial regions: eyes, eyebrows, nose, and mouth. Then, the landmark sequences of each region were processed by independent Bidirectional Recurrent Neural Network (BRNN) models to extract local features. A cascade fusion strategy was introduced, where the upper facial representation was obtained by concatenating the eyebrow-eye-nose features, while the bottom facial representation was a combination of nose-mouth features. The global representation was finally derived from the upper and lower facial representations for efficient facial expression classification.

Apart from the above concatenation-based methods for integrating local features, some methods employ graph models to capture the relationship between facial local areas. Wang

et al. [115] proposed a facial expression recognition framework based on the Interval Temporal Bayesian Network (ITBN), where they defined a facial expression as a set of primitive events that occur continuously or in parallel, with primitive events being the movements of facial points to approximate local muscle activities. The ITBN model established the spatio-temporal dependencies between local facial components by constructing a directed graph with facial primitive events as nodes, resulting in superior recognition performance in comparison with classical appearance- and geometry-based approaches. Although the approaches described above take into account the connections between facial regions, however, the selection of these specific regions is subjective and diverse. There may be an optimal division of regions that may require the intervention of expert knowledge. Furthermore, the division strategy may also vary with specific tasks. Moreover, these approaches may lack robustness to variations in the head pose as it does not regard the context of the whole face.

2.4.4 Full face-based feature support

Unlike the previous approaches based on features extracted from special points or facial local regions, this type of approach treats the entire facial area as a whole, where the shape representations are extracted from the coordinates of all facial landmarks.

1) Shape representation in Euclidean space

Some approaches tend to construct the global shape representation in Euclidean space, which is typically represented by a multidimensional vector. For instance, Jain et al. [116] treated all facial landmarks as a whole, i.e., a high-dimensional vector composed of 2D coordinates. Latent Dynamic Conditional Random Fields (LDCRFs), an undirected graph model was then selected as a classifier, permitting the parallel encoding of motion patterns within a single facial expression and between different expressions. Experimental results showed that LDCRFs obtained more robust performance on landmark-based shape features than on appearance features. Lorincz et al. [117] presented a kernel-based facial expression recognition method. The 3D landmarks captured from facial expression videos were regarded as multidimensional time series, and two kernels: the Dynamic Time Warping (DTW) kernel and the Global Alignment (GA) kernel were applied to capture the similarity information between the sequences. Additionally, they also adopted the alternating projection method on the former to ensure the positive semi-definite nature of the kernel. Ultimately, Gram matrices induced by these two kernels were applied for SVM-based emotion classification. In general, the above approaches encoded the facial shape as a Euclidean vector consisting of 2D or 3D coordinates of the landmarks for ex-

pression recognition. However, this commonly results in a high-dimensional representation. In the work of [117], they employed the PCA technique for the compression of the shape descriptor, which in turn led to information loss. To solve this problem, some approaches adopted deep neural networks for effective representation learning from coordinate vectors. Qiu et Wan [118] first selected specific reference points to normalize the 68 detected facial points, and then the coordinates of the 2D landmark configurations were fed into a fully connected neural network in the form of 1D vectors of size 136 to classify facial expressions. Jung et al. [119] also employed a deep architecture consisting of fully connected layers. Facial movements were encoded by landmark trajectories, which are concatenations of 1D coordinate vectors of all frames in the facial videos and served as input to the model for facial expression recognition. Though the above approaches can provide a more holistic and global understanding of facial expressions, however, they ignore the geometric properties of facial shapes, thus failing to accurately capture the deformations of the face, especially when dealing with complex expressions.

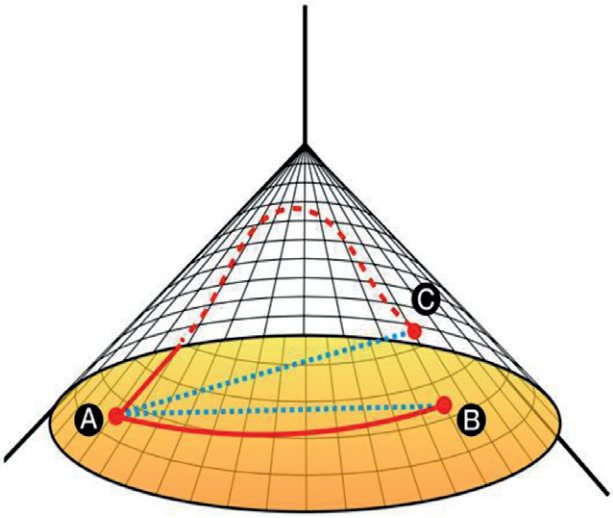


Figure 2.9: The cone of SPD matrices. Points A, B and C represent covariance-based shape representations, where the blue and red lines symbolize distance measurements in Euclidean space and Riemannian manifold, respectively. It can be seen that in the manifold the point A is closer to the point B, while in Euclidean space it has the opposite result. Excerpt from [3].

2) Shape representation on manifold

Different from previous methods that use flattened coordinate vectors in Euclidean space as the face descriptor, this class of methods builds compact, discriminative facial shape representations on non-linear manifolds that are invariant to affine transformations [120]. One successful case is the covariance descriptor [121–123] in the field of computer vision, which efficiently fuses high-dimensional features by computing the corresponding covariance matrix, and this is

what we refer to as compactness. The resulting covariance representations are the symmetric positive definite (SPD) matrices that will induce the manifold structure by defining the appropriate metric [124]. More specifically, the set of SPD matrices forms a cone-shaped Riemannian manifold, on which each covariance-based representation can be considered as a point [3]. Fig. 2.9 shows an illustration of the corresponding Riemannian manifold. Consequently, one can obtain more precise similarity measures between Riemannian representations rather than Euclidean ones, which is related to better discriminative power. In addition, representations based on Riemannian geometry are invariant under affine deformation, thus offering a possibility for robust facial shape analysis. Hence, some studies have turned their attention to constructing shape representations in nonlinear spaces based on the foregoing merits. These methods typically embed the motion of the facial landmarks in the time-parameterized curves according to the Riemannian geometry of manifold data. For example, Taheri et al. [125] proposed a robust facial expression recognition approach based on shape manifold analysis to eliminate the disturbance due to viewpoint variations. The global facial deformation evoked by the emotional expression can be registered by a sequence of facial landmarks, which were further treated as a curve on the Grassmann manifold. The 2D landmark configurations in successive frames were considered as adjacent points on the curve, the connections between which were characterized by the velocity vectors in the tangent space. Consequently, the shape representation derived on the manifold can better discriminate between different facial expressions than that in Euclidean space. Tanfous et al. [126] conducted a view-invariant facial expression analysis using geometry of non-linear manifolds. Continuous 2D landmark configurations detected in facial expression videos were first modeled as trajectories in Kendall's shape space, which were subsequently encoded as sparse time series in Euclidean vector space by Riemannian sparse coding and dictionary learning (SCDL) to suit traditional linear classification algorithms. In the end, the proposed method achieved competitive results in macro- and micro-facial expression recognition tasks. Kacem et al. [127] presented a novel facial expression recognition framework based on the geometry of the space of Gram matrices. Gram matrices computed from a sequence of facial landmarks were considered as points on the Riemannian manifold of positive semidefinite matrices of fixed-rank, where the matrices of adjacent frames were joined by pseudo-geodesics to model the facial movements. The Dynamic Time Warping (DTW) technique was adopted to measure the similarity between trajectories, which was finally fed into a pairwise proximity function SVM (ppfSVM)-based classification. Szczapa et al. [128] proposed a Riemannian framework for pain intensity estimation using facial landmarks. The Gram matrices were first computed from the combination of the face point coordinates and the velocities between consecutive frames. The matrix sequences were then be regarded as curves

on the Riemannian manifold, where a curve-fitting algorithm was applied to cope with missing or incorrect data. The input data for Support Vector Regression (SVR)-based pain intensity prediction was finally provided by Global Alignment Kernel (GAK) through the calculation of the similarity scores between trajectories. Overall, the methods discussed above incorporate the intrinsic geometry of the shape space, therefore yielding facial representations that are robust to affine transformations. They generally employed a fixed metric to measure the similarity between shape representations. However, the selected metric may not always be optimal for a specific task or dataset. In such cases, the underlying relationships between facial shapes may not be captured effectively.

Table 2.4: Overview of behavioral emotion recognition methods.(FCN: Fully Connected Network)

Feature Support	Shape Space	Architecture	Papers
Point-based	Euclidean	Hand-crafted	[107]
	Euclidean	Hand-crafted	[108]
	Euclidean	1DCNN	[109]
	Euclidean	LSTM	[110]
	Euclidean	Hand-crafted	[111]
	Euclidean	Hand-crafted	[112]
Local region-based	Euclidean	Attn-LSTM	[113]
	Euclidean	PHRNN	[114]
	Euclidean	Hand-crafted	[115]
Full face-based	Euclidean	Hand-crafted	[116]
	Euclidean	Hand-crafted	[117]
	Euclidean	FCN	[118]
	Euclidean	FCN	[119]
	Non-linear Manifold	Hand-crafted	[125]
	Non-linear Manifold	Hand-crafted	[126]
	Non-linear Manifold	Hand-crafted	[129]
	Non-linear Manifold	Hand-crafted	[128]

2.4.5 Discussion of landmark-based FER methods

We classified facial landmark-based emotion recognition methods into three types based on the range of facial feature support and introduced them accordingly in the previous sections. Table 2.4 provides a brief summary of these methods. First, we notice that most of the methods still

construct their facial feature representations in Euclidean space, which are commonly disturbed by affine transformations such as translation, scaling, and rotation. Some full face-based methods solved this issue by defining the facial shapes on non-linear manifolds, thus achieving more robust representations to face deformations. However, these methods use predefined metrics to describe the similarity relationships between different facial shapes. Thus they are not flexible enough to adapt to the variable data structures in different tasks or datasets. In addition, all existing methods do not explicitly address the significant intra-class variability presented by facial expression images in the wild. Hence, there exists the possibility of further enhancing the discriminative power of feature space to obtain better recognition performance.

2.5 Emotion recognition based on multimodal signals

To date, most emotion recognition studies have focused on behavioral or physiological responses. Recently, a few studies have focused on fusing explicit behavioral signals such as facial expressions and body motions represented by ACC (3-axis acceleration) signals and implicit physiological signals to leverage the complementary nature of both to provide more robust and enhanced emotion detection. In this section, we classify these multimodal methods into early fusion-based (Section 2.5.1), intermediate fusion-based (Section 2.5.2), late fusion-based (Section 2.5.3), and hybrid fusion-based (Section 2.5.4) according to the integration fashion. The principles of the first three fusions are consistent with those mentioned in the previous section 2.3.3, while the last one combines the early fusion and late fusion. The details of these multimodal approaches will be discussed below.

2.5.1 Early Fusion

Similar to the early fusion mentioned in Section 2.3.3, data from the behavioral and physiological domains are concatenated and treated as a single input for emotion recognition. Li et al. [130] proposed a multimodal fusion framework for emotion recognition based on EEG signals and facial expressions. Energies of different channels are extracted by DWT as EEG features, while a series of geometric features such as distances between facial landmarks, angles of the corner of the lip and slopes of the eyebrows are selected as facial features. Ultimately, the SVM classifier with fused feature vectors as input obtained an improved performance over that using solely EEG data. Werner et al. [131] collected facial distances and gradient-based features from video frames which were combined with the statistical features calculated from biological signals. The resulting multimodal vectors were employed to train a random forest model for pain assessment. Experimental results indicated that the combination of facial and physiological data improved unimodal performance in most cases, especially when predicting high pain intensity. Gil-Martin et al. [132] introduced cube root (CR) and constant Q transform (CQT) in the computation of the spectrum for wearable physiological and motion signals and fed the corresponding spectrum variants as single-channel images into a 2D convolutional and fully connected based neural network for multimodal emotion recognition. Experimental results showed that the proposed model achieved significant performance improvements on multimodal data compared to the unimodal settings. However, early fusion approaches sometimes fail to deliver performance gains. Schmidt et al. [53] collected physiological and motion data using a wrist device and a chest device and extracted various time- and frequency-domain multimodal features for training a range of common machine learning classifiers. However, the results of

experiments conducted on both devices demonstrated that the simple concatenation of features from different modalities did not lead to improved performance.

2.5.2 Intermediate Fusion

Marginal features computed from physiological and behavioral modalities are either applied directly to prediction or used to jointly learn advanced joint representations. For instance, Wu et al. [133] suggested a facial-EEG based emotion recognition system. The spectral energies of the three frequency bands are first extracted from the EEG signals and transformed into image sequences. Then the facial and EEG image sequences are fed separately into independent fully connected neural networks to extract marginal representations. A hierarchical LSTM model with self-attention was applied to fuse and learn more abstract multimodal representations, which achieved the state-of-the-art results on the emotion classification task. Huynh et al. [69] employed an auto-designed deep neural network based on Neural Architecture Search (NAS) for affective states and stress detection. Filter banks and mixed features of physiological signals and the body motion signal (i.e., ACC signal) were extracted to search for the highest scoring architectures from 10,000 randomly generated neural network candidates for training of individual modalities. The obtained unimodal representations were concatenated for final classification. The obtained optimal architecture performed better on multimodal data compared to the setup using only physiological data. Lai et al. [15] proposed a residual temporal convolution-based deep neural network (Res-TCN) to capture the effective features of multimodal data. Two fusion strategies were implemented, the first of which fed sensor data from all modalities as a whole into the Res-TCN model for predicting emotions, while the second combines the marginal representations derived from the independent convolutional models for the final classification. In comparison experiments, the latter exhibited superior performance.

2.5.3 Late Fusion

This type of approach generally integrates decisions derived from physiological and behavioral models. Saffaryazdi et al. [134] conducted multimodal emotion recognition experiments using facial expressions, EEG and peripehral signals. First, the apex frame in the facial expression video was detected and fed into a 3D convolutional network along with its surrounding frames to predict emotions. For physiological signals, various models such as SVM, KNN, RF and LSTM were implemented to classify features of EEG and peripheral signals. Decisions from facial expressions, EEG and peripheral signals were fused with two strategies, i.e., majority

voting and weighted sum. The evaluation results showed that the fused data yielded improved emotion recognition performance compared to unimodal settings.

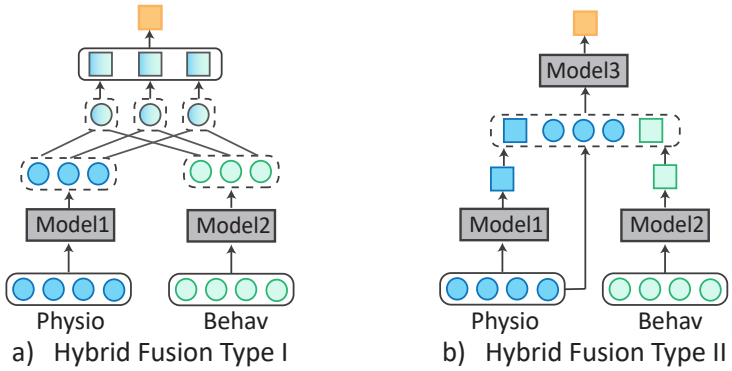


Figure 2.10: Hybrid fusion strategies for multimodal emotion recognition. a) Physiological and behavioural features are first merged within segmented data and their corresponding predictions are aggregated for the final decision [4]. b) Decisions of physiological and behavioural modalities are combined with unimodal features for emotion detection [5].

2.5.4 Hybrid Fusion

Hybrid fusion can be regarded as a mixture of early fusion and late fusion. Two hybrid fusion strategies have been found in the literature for multimodal emotion recognition, and their corresponding illustrations are shown in Fig. 2.10. Regarding the first strategy, Zhong et al. [4] proposed a multimodal emotion recognition framework based on hybrid fusion of facial expressions and peripheral physiological signals. The facial features extracted from different views were first fused with a series of physiological signal features within each data segment. Next, separate weak learners were trained on the fused features and assigned a decision for each segment. The final decision was a combination of all the weak learners' decisions. The proposed fusion strategy was compared with the common early and late fusion strategies and obtained better performance in predicting emotional valence and arousal. Concerning the second strategy, Gjoreski et al. [5] developed a multimodal stress detection method suitable for the real-life setting. First, a stress detector was trained on biological data collected from the laboratory environment. Meanwhile, an action recognizer learned the subject's behavioral information from the ACC (3-axis acceleration) signal. Then, physiological features, the output of the stress detector and action recognizer were fused and fed into a series of machine learning algorithms for continuous pressure prediction. Finally, the experimental results indicated that the incorporation of behavioral information could enhance recognition performance in unconstrained scenes.

2.5.5 Discussion of multimodal emotion recognition methods

We introduced in the previous sections the emotion recognition approaches integrating physiological data and behavioral data (i.e., facial expressions and body motion from ACC signal), which are briefly listed in Table 2.5. From the table, we can first notice that half of the multimodal methods are on the basis of traditional machine learning algorithms. Second, half of the solutions employed either the early fusion or late fusion strategy, neither of which could capture simultaneously intra-modal and inter-modal correlation information. Third, facial expressions are considered a key aspect in recognizing emotions, however, few studies have investigated the usefulness of combining facial expressions with peripheral physiological signals for predicting emotions, especially in practical environments. These studies that have been conducted typically involve EEG signals, which are not suitable for use outside of a controlled laboratory setting. Additionally, the lower focus on integrating physiological signals with facial expressions compared to body motions from the ACC signal is likely due to the technical difficulties in combining visual and physiological data, which have distinct structures and characteristics

Table 2.5: Overview of behaviroual-physiological emotion recognition methods. (ACC: 3-axis Acceleration, FE: Facial Expressions, PPS: Peripheral Physiological Signals, P+F: Physiology+Face, P+M: Physiology+Body Motion.)

Fusion	Modality	Domain	Real-life?	Architecture	Papers
Early	EEG, FE	P+F	No	Hand-crafted	[130]
	PPS, FE	P+F	Yes	Hand-crafted	[131]
	PPS, ACC	P+M	Yes	Hand-crafted	[53]
	PPS, ACC	P+M	Yes	2DCNN	[132]
Intermediate	EEG, FE	P+F	No	Hier-SA-LSTM	[133]
	PPS, ACC	P+M	Yes	NAS	[69]
	PPS, ACC	P+M	Yes	Res-TCN	[15]
Late	EEG, PPS, FE	P+F	No	3DCNN-LSTM	[134]
Hybrid	PPS, FE	P+F	Yes	Hand-crafted	[4]
	PPS, ACC	P+M	Yes	Hand-crafted	[5]

3

Transformer-based Self-supervised Multimodal Representation Learning for Wearable Emotion Recognition

Contents

This chapter shows technical contributions on physiological aspects, where a transformer-based self-supervised learning framework for wearable emotion recognition is proposed, which can solve the overfitting problem due to limited supervised data while performing effective multimodal fusion. In the following, first, we introduce wearable emotion recognition and its challenges in Section 3.1. Next, Section 3.2 shows self-supervised learning-based approaches in the field of emotion recognition and related research on multimodal signal fusion. In Section 3.3 we describe the proposed approach. Sections 3.4 and 3.5 present the datasets used and the corresponding experimental results on emotion recognition, respectively. Finally, Section 3.6 is a summary of our work.

3.1 Introduction

Emotion recognition is an emerging field of research aimed at developing algorithms and technologies to identify human emotions from various data sources. Most current methods of emotion recognition rely on physical or physiological indicators of the human body. In contrast to physical signals such as facial expressions [41] and speech [43], physiological responses under certain emotional states are involuntary and therefore provide more objective decisions for identification systems [32]. The physiological modalities primarily consist of electroencephalography (EEG) signals and a series of peripheral signals. However, the acquisition of EEG data is challenging for implementation in real-life scenarios. With the advance of non-invasive technologies, emotion recognition methods based on multiple peripheral signals captured by smartphones/wearable watches have attracted some attention. Traditional wearable multimodal emotion recognition used hand-crafted features as input of machine learning-based algorithms for classification. Nevertheless, these well-designed features based on comprehensive domain knowledge can be an obstacle for non-domain experts. Most recent researches focus on deep neural networks, which can automatically extract complex patterns from multimodal signals. However, given that most of them are trained in a supervised manner, it is challenging to obtain generalizable models using limited labeled data, especially in daily life settings, where standard protocols for obtaining accurate emotion labels are not yet well defined. Besides, each specific supervised task requires training the deep model from scratch and its knowledge transfer ability on other tasks is not satisfactory [135]. Self-supervised learning (SSL), as an emerging learning paradigm, eliminates the need for extensive manual labeling and has demonstrated comparable or even superior performance to supervised learning methods in areas of computer vision (CV), natural language processing (NLP). Several SSL-based efforts [136–138] have been done for emotion recognition using EEG signal, but they are not suitable for practical scenes. Only one

work [139] targeted low-frequency wearable peripheral signals, but they ignored the correlation between multimodal signals. In this paper, we propose a self-supervised multimodal representation learning approach for wearable emotion recognition based on peripheral physiological signals. The first stage is model pre-training with the pretext task based on signal transformation recognition, where a large amount of unlabeled multimodal data are automatically assigned labels through a series of transformations. Considering the heterogeneity of multimodal signals, temporal convolution-based modality-specific encoders are first employed separately on the transformed unimodal data to extract low-level features, followed by a transformer-based shared encoder deployed to aggregate unimodal features, enabling the modeling of complementary and collaborative properties between multimodal signals. Finally, modality-specific signal transformation recognition is performed to learn effective multimodal representations for downstream tasks that are robust to perturbations in magnitude or temporal domains. The second stage is supervised emotion recognition, where the SSL pre-trained encoder part is retained as a feature extractor to generate more generalized multimodal representations for emotion classification tasks. The overview of the proposed approach is illustrated in Fig. 3.1. To validate the effectiveness of our method and the knowledge transferability across different datasets, we pre-trained the proposed model on a large-scale unsupervised emotion dataset PRESAGE collected in unrestricted real-life scenarios and evaluated its performance on three public emotion recognition datasets. Overall, our contributions can be summarized as follows:

- We proposed a novel self-supervised learning (SSL) framework to learn generalized representations from a large number of unlabeled samples to cope with the overfitting problem on small-scale physiological data.

- We adopt an intermediate fusion strategy based on temporal convolution and transformer, capable of modeling both the heterogeneity and cross-modal correlation of physiological signals to effectively fuse multimodal data.

- We outperformed state-of-the-art supervised or self-supervised learning-based approaches in various emotion-related classification tasks involving mental stress, affective states, arousal, and valence. Moreover, our model was proven to be more accurate and stable on limited labeled data than fully-supervised models. In addition, multiple ablation studies have been performed to investigate the effectiveness of our method.

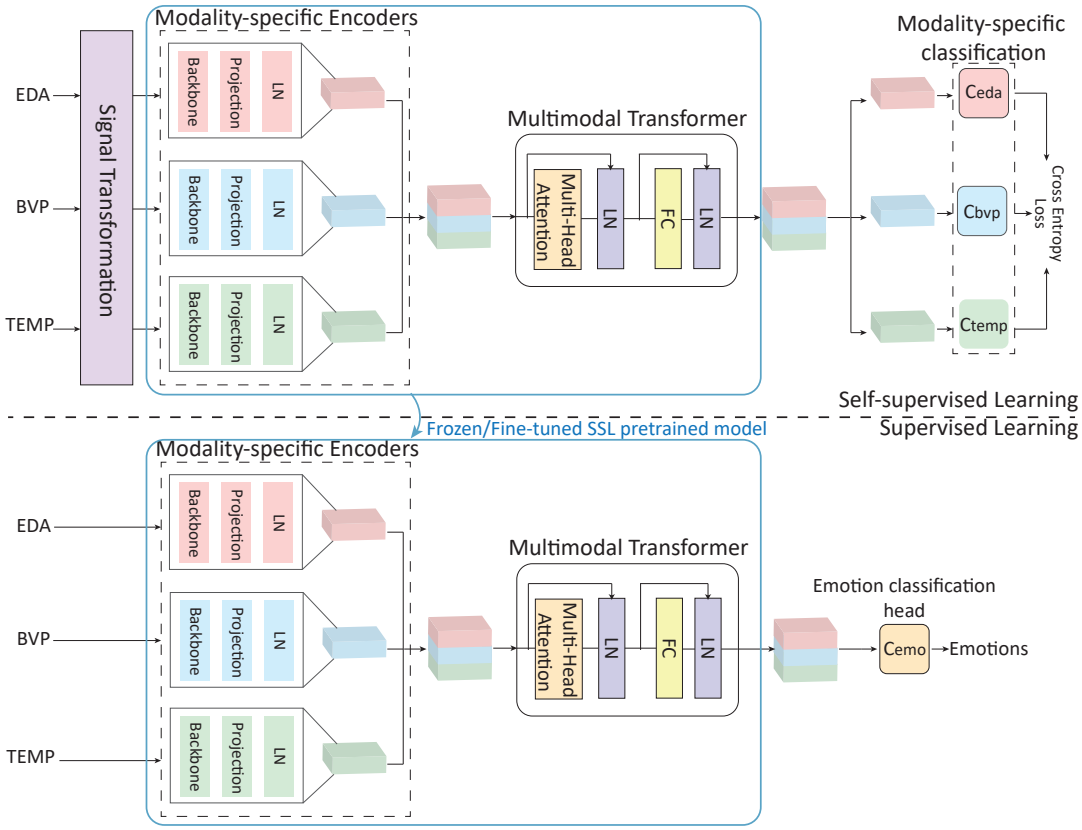


Figure 3.1: Overview of our self-supervised multimodal representation learning framework. The proposed SSL model is first pre-trained with signal transform recognition as the pretext task to learn generalized multimodal representation. The encoder part of the resulting pre-trained model is then served as a feature extractor for downstream tasks which is frozen or fine-tuned on the labeled samples to predict emotion classes.

3.2 Related Work

More recently, researchers have shifted their gaze from the invasive EEG signal to peripheral physiological signals which can be captured by lightweight and more invisible devices. Among numerous peripheral signals, electrocardiogram (ECG), blood volume pressure (BVP), electrodermal activity (EDA), etc, captured from the cardiac system and electrical skin activity are crucial emotion-related modalities and have proven their validity in several studies [15, 53, 69, 70, 140]. However, exploration of the wearable emotion recognition is not yet well established. In Section 2.3, we reviewed methods based on fully-supervised, unsupervised and semi-supervised learning. Among them, fully supervised methods based on deep neural networks struggle with the risk of overfitting and inaccurate emotion labeling, while unsupervised learning methods have unsatisfactory performance due to their weak supervision and there is a lack of exploration for semi-supervised learning algorithms. In addition, the fusion strategies

of most existing approaches are ineffective in capturing both intra- and inter-modal correlations of multimodality. In the following sections we will first detail a novel solution to the problem of overfitting in the low-data regime, namely self-supervised learning, followed by multimodal data fusion methods.

3.2.1 Self-supervised learning (SSL) for limited labelled data

To solve the overfitting problem introduced by the limited available data, one common solution is data augmentation, i.e., applying different transformations on the original samples to obtain more abundant data. However, performing data augmentation on a dataset consisting of a limited number of subjects could not introduce inter-subject variability during training and ultimately yield a generalized model [139]. Recently, a compelling branch in the field of unsupervised representation learning is self-supervised learning (SSL), which can effectively address the de-generalization issue posed by insufficient labeled data. Unlike unsupervised learning which does not involve any labelled data, SSL is designed with a series of pretext tasks that introduce self-supervision to unlabelled data, enabling more effective representation learning for downstream tasks. Each unsupervised sample is automatically labeled through inherent dependencies and associations between the data without human intervention [141]. The SSL model pre-trained on pseudo-labeled data is considered as powerful feature extractor for a variety of downstream tasks. In the domains of computer vision and natural language processing, SSL-based work such as SimCLR [142], Word2Vec [143], and BERT [144] have exhibited competitive and even superior performance on a range of tasks. However, few studies have investigated the performance of SSL models on peripheral physiological signal data. Sarkar et Etemad [135] introduced a self-supervised representation learning framework for ECG-based emotion recognition, where the 1DCNN-based multi-task deep neural network is pre-trained with the objective of identifying the signal transformation types applied to unlabeled data. Their study indicated that the pretext task based on transformation recognition can enable the model to better cope with potential variation factors in the data. However, not all time steps of a signal sequence are associated with the downstream target event (i.e., a specific emotion). Thus, how to filter out irrelevant information during SSL for downstream tasks is an unsolved problem. Exploiting the synchrony of multimodal emotional responses is a potential solution. More specifically, multimodal physiological signals exhibit correlated or consistent temporal changes when emotions are elicited. In this way, modeling the correlation of multimodal signals in SSL can facilitate the capture of emotion-related components in unlabeled data. Regarding multimodal emotion recognition, Dissanayake et al. [139] proposed a self-supervised contrastive learning approach,

which aims to approximate the positive pairs while pushing the negative pairs away from each other. However, their SSL model is obtained by pretraining each modality independently, and thus again ignores the cross-modal correlations. Therefore, more effective multimodal fusion strategies need to be developed for SSL-based wearable emotion recognition.

3.2.2 Multimodal data fusion for emotion recognition

Multimodal data fusion strategies can be generally categorized into: early fusion, intermediate fusion, and late fusion. Most existing approaches for multimodal emotion recognition are based on early fusion, where multimodal data are combined as a whole before performing a learning task. Joint representations can be extracted directly from concatenated vectors with deep models such as 1DCNN [145] and Bi-LSTM [77], which allow for encoding inter-modal correlations. However, since unimodal features are not learned explicitly (i.e., the heterogeneity of the multimodal signal is ignored), this fusion strategy is not effective in capturing intra-modal correlations. Late fusion-based approaches [15, 146] integrate the decisions of multiple independent learning models to predict emotion categories. Thus, in contrast to early fusion, this fusion approach ignores the connections and interactions between modalities.

Different from the previous fusion approaches, intermediate fusion enables both intra- and inter-modal correlation, where independent feature extractors are first applied to different modalities and the obtained unimodal features are then aggregated in an additional fusion module to further learn the joint representation. A variety of options exist for this fusion module. For example, Shu and Wang [147] adopt ed the restricted Boltzmann machine (RBM) model to learn the joint probability distribution of multimodal low-level features to encode cross-modal information exchanges. Zhang et al. [148] modeled the associations between multimodal features by introducing a regularization term to the objective function. More recently, the transformers have also gained popularity in intermediate fusion-based approaches [149–151] for video, audio and text. Regarding studies on emotion recognition, Wu et al. [149] proposed a multimodal Recursive Intermediate Layer Aggregation (RILA) model, which was applied between layers of unimodal deep transformers to capture interactions across modalities through the integration of multimodal intermediate representations. In this work, the transformers were employed to provide valid intermediate features. At the same time, they have also proved to be effective in merging multimodal data [150, 151]. The attention mechanism can capture advanced patterns shared across modalities, thus exhibiting advantages over naive fusion strategies such as concatenation. In terms of practicality, multimodal emotion recognition based on the video, audio and text may not be well suited to real-life scenarios, as it requires considerable com-

putational resources for long-term video stream analysis. In contrast, wearable physiological signals can consistently predict emotions in a low-cost and objective way. However, the validity of transformer-based models has not been well established for wearable emotion recognition. Meanwhile, video, audio and text-based approaches cannot be directly migrated to physiological data due to differences in data structures. In addition, they are susceptible to overfitting problems as they generally have a relatively deep architecture and follow a fully-supervised setup.

3.3 Proposed Method

3.3.1 Overview

Our goal is to employ unlabeled data for capturing generic representations of multimodal physiological signals in order to address the de-generalization problem introduced by a limited number of labeled samples. Hence, we propose a self-supervised learning (SSL) scheme using signal transformation recognition as a pretext objective. An illustration of the proposed approach is shown in Fig. 3.1. In our work, three modalities measured by different sensors are considered: electrodermal activity (EDA), blood volume pressure (BVP) and skin temperature (TEMP). More formally, let $x_m \in \mathbb{R}^{N\times 1}$ represent a 1D time-domain signal from one of the M different modalities (in our work, M = 3), where N is the signal length. Given a set of *n* transform functions $T = \{T_j(\cdot), j \in \{1, \ldots, n\}\}$, the altered multimodal signal dataset can be generated by applying each transformation to individual modality. Based on this, one can easily build a pseudo-labeled dataset $\mathcal{L} = \{(T_j(x_m^i), y^i), y^i \in \{1, \ldots, n\}, m \in \{1, \ldots, M\}, i \in [1, |\mathcal{L}|]\}$ for unlabelled samples through self-supervision enabled by signal transformations. Then, the proposed model consisting of a multimodal encoder E and modality-specific classifiers C is pre-trained to predict the type of transformation applied to samples in $\mathcal{L}$. Ultimately, only the encoder part E of the optimal model obtained after pre-training is retained and is expected to produce generalized multimodal representations in a variety of supervised downstream tasks. Details of the proposed SSL framework are as follows.

3.3.2 Self-supervised learning of multimodal physiological signals

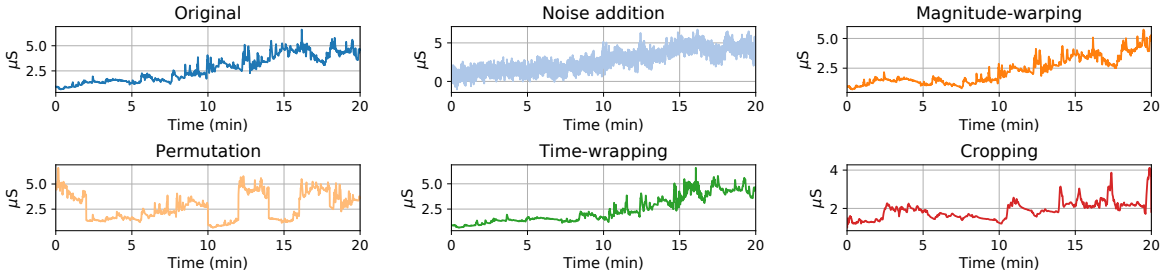


Figure 3.2: The original EDA signal and the disturbed EDA signals after applying five transformations. For each modality, the raw signal data and the transformed signal data are stacked and fed into the proposed SSL model for multimodal representation learning.

3.3.2.A Pretext Task: signal transformation recognition

Signal transformation recognition was adopted as the pretext task in SSL, which proved to be effective in learning generalized representations for downstream tasks such as action recognition [152] and emotion recognition [135]. The random transformations used in the previous SSL methods are one of the common data augmentation techniques for time series, which can generally be classified into two categories: magnitude domain transformations and time domain transformations. The former interferes with the signal values while preserving the time step order, whereas the latter mainly affects the time scale. Previous evaluations of SSL models based on individual transformation recognition [135,152] have indicated that *Noise addition* and *Scaling* ranked highly for magnitude domain transformations, while *Permutation* and *Time-warping* performed outstandingly well among time domain transformations. Meanwhile, according to the review of time series augmentation strategies [153, 154], though most of the suggested transformations have been adopted in previous SSL-based work, two transformations have not been thoroughly evaluated: *Magnitude-warping* and *Cropping*. Ultimately, we selected the five transformations: *Permutation*, *Time-warping*, *Noise addition*, *Magnitude-warping* and *Cropping* for the pretext task. The reason why *Scaling* was omitted is that *Magnitude-warping* can be seen as a special variant of *Scaling*[1].

The above signal transformations are performed on all three modalities and the resulting transformed signal data is fed into the proposed SSL model as input along with the original multimodal signal data. Fig. 3.2 shows the effect of these deformations on a sample of the EDA signal. Details of each transformation are described in subsequent paragraphs. Here, for

[1] *Scaling* multiplies time series values by a random scalar whereas *Magnitude-warping* distorts the signal values by a smooth curve.

simplicity, we write the above-mentioned 1D signal x_m uniformly as $x(t)$, where t represents the time step.

Magnitude domain transformations:

- **Gaussian noise addition:** The original signal $x(t)$ is disturbed by white Gaussian noise $z(t)$, which can be extracted from a zero-mean normal distribution $\mathcal{N}(0, \sigma^2)$. By assigning a preferred signal-to-noise ratio (SNR), the variance σ^2 (i.e., the average power of the noise) of the distribution $\mathcal{N}$ can be derived from the following formula $10^{(P_{sig}-SNR)/10}$, where P_{sig} is the average power of the signal. In the end, the noised signal is calculated as $x(t) + z(t)$.

- **Magnitude-warping:** The magnitudes of the original signal are altered by a random smooth curve formed by cubic spline interpolation function $\phi(\cdot)$. In the end, the transformed signal can be calculated as $x(t) \cdot \phi(x(t))$.

Time domain transformations:

- **Permutation:** The original signal is split into *n* non-overlapping segments $x(t) = \{x_1, x_2, ..., x_n\}$, which are then temporally disrupted and eventually recombined together to form the permuted signal $x(t) = \{x_{p1}, x_{p2}, ..., x_{pn}\}$, where $\{p1, p2, ..., pn\}$ is a shuffled version of the original order.

- **Time-warping:** The original signal is divided into *n* non-overlapping segments $x(t) = \{x_1, x_2, ..., x_n\}$, half of which are randomly selected to be stretched by a linear interpolation function $F(x_i, k)$, where k is the stretch factor, and the remaining half of the segments are squeezed by the function $F(x_i, 1/k)$, where $1/k$ is the squeeze factor. The time-warped signal can be concatenated from the transformed segments and finally resized to the original length.

- **Cropping:** The original signal is divided into *n* non-overlapping segments $x(t) = \{x_1, x_2, ..., x_n\}$, one of which is randomly selected and resampled to the original length.

By identifying the signal transform types, our model is expected to learn a more robust and generalized representation against disturbances in the magnitude or time domains. For example, **Magnitude-warping** and **Gaussian noise addition** can simulate different types of real-world noise, such as measurement errors, signal artefacts caused by the subject's body movements, etc. For time domain transformations, **Permutation** perturbs the order of time steps to prompt the model for capturing temporal dependencies between data points, **Time-warping** simulates duration variations in emotional responses by stretching or squeezing time steps, and **Cropping** allows the model to be more robust to changes in the temporal location of emotional events.

3.3.2.B Self-supervised multimodal representation learning network architecture

The proposed SSL multimodal deep neural network consists of two key elements, namely the encoder E and the modality-specific transformation classifiers C. The encoder E can be further subdivided into temporal convolution-based modality-specific encoders E_p and transformer-based shared encoder E_s, where E_p models the heterogeneity of multimodal signals and E_s activates cross-modal information exchange. Ultimately, the multimodal features obtained from the encoder are used as input to C for identifying transformation types for each modality. The implementation of these key components is described in the following paragraphs.

Modality-specific encoder: Considering the heterogeneity of the multimodal signals, separate encoders are first employed for each modality, with a temporal convolution-based network acting as the backbone to capture low-level intra-modal correlation information. The temporal convolutional network (TCN) [155], in a nutshell, is a combination of dilated causal convolution and residual connections, with parallel computational capability and robust gradients at optimization, thus demonstrating better performance than traditional recurrent networks, such as LSTM and GRU. One basic TCN consists of several residual blocks. The most central components of each block are two dilated causal convolution layers. The causality can be easily achieved when the output at the current moment t depends only on the elements of the past historical moments up to t in the previous layer. Meanwhile, the dilation operation injects holes in the standard convolution map, thereby increasing the reception field. More formally, given the transformed 1D signal of modality m: $x'_m = T_j(x_m) \in \mathbb{R}^{N\times 1}$ with N time steps, and a filter f of size k, the dilated convolution on time step t can be defined as

$$F(t) = \sum_{i=0}^{k-1} f(i) \cdot x'_m(t - d \cdot i) \tag{3.1}$$

where d is the dilation factor. Following each convolutional layer is a weight normalization layer for the convolution filter, a rectified linear unit (ReLU) layer, and a dropout layer for regularization. In the end, a residual connection is created between the input and output of the block, where a 1×1 convolution is introduced to eliminate the mismatch in channel numbers between the input and output. Fig. 3.3 illustrates the detailed structure of the TCN-based backbone. The dilated causal convolution layers in two residual blocks are equipped with 16 filters with a kernel size of 6, where the dilation factors are 1 and 2, respectively. Zero-padding of 5 and 10 are also introduced to ensure that the input and output sequences are of the same length. Subsequently, a modality-specific projection head (i.e., a linear fully connected layer with 128 units) and a layer normalization are then applied to map the low-level features to a higher dimensional

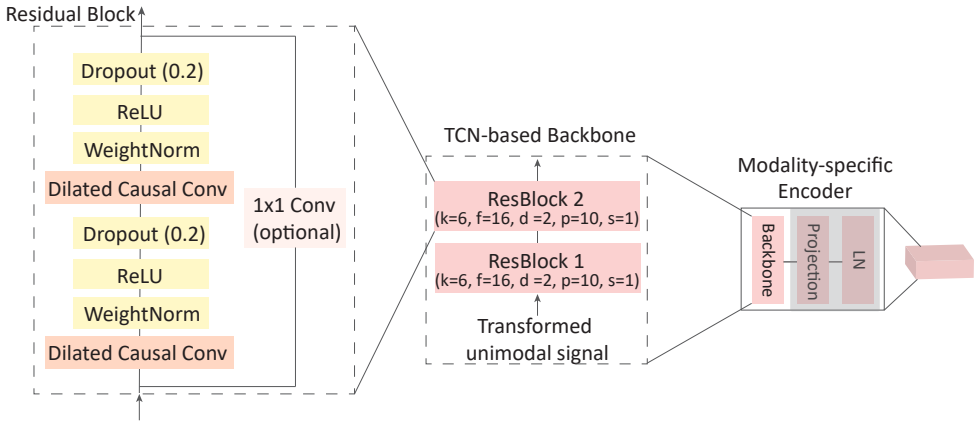


Figure 3.3: Modality-specific backbone based on temporal convolutional network (TCN). Each backbone consists of two residual blocks for capturing low-level features for transformed unimodal signals x'_m. (k: kernel size, f: number of filters, d: dilation factor, p: padding size, s: stride size, weightnorm: weight normalization for convolution filters)

embedding space. Finally, the output of the modality-specific encoder E_p is:

$$z_m = LayerNorm(MLP(TCN(x'_m))) \in \mathbb{R}^{N \times d} \tag{3.2}$$

where d is the embedding dimension.

Shared encoder: As mentioned in Section 3.2.1, encoding of the coordination and interaction between multimodal signals is essential in order to learn generic representations related to the downstream emotion recognition tasks. This can be done through the transformer in which each modality identifies components of other modalities that are highly correlated with itself through the attention mechanism for better signal transformation classification. To achieve this, the low-level features z_m of each modality are first stacked to form a multimodal embedding $z_{multi} = [z_1, \ldots, z_m, \ldots, z_M] \in \mathbb{R}^{MN \times d}$. The scaled dot-product attention proposed in [156] is then applied to calculate the dependencies between different modalities:

$$Attn(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{3.3}$$

where Q, K, V represent queries, keys and values, respectively. More intuitively, the attention layer acts as a weighted sum of values V, where the attention weight associated with each value is generated by the compatibility of the query with its corresponding key. For our shared encoder E_s, queries, keys and values are derived through a linear mapping of multimodal features z_{multi}, and the resulting output of the attention layer is:

$$z^a_{multi} = Attn(z_{multi}W^Q, z_{multi}W^K, z_{multi}W^V) \tag{3.4}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$ are the projection matrices. Fig. 3.4 presents the process of gen-

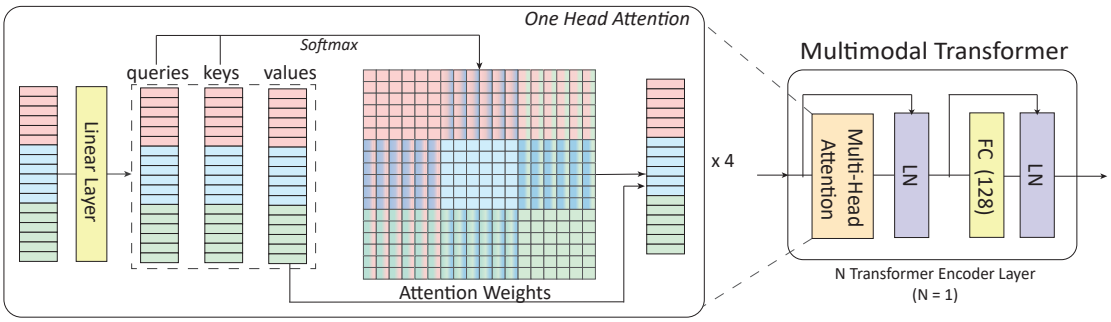


Figure 3.4: Shared encoder based on the multimodal transformer. (FC: fully-connected layer with 128 units, LN: layer normalization)

erating attention weights from multimodal embeddings, where cross-modal communications are activated. For our shared encoder, the one-layer vanilla transformer block proposed in [156] with four-head attention is implemented. The feedforward layer dimension is set to 128. ReLU is selected as the activation function for intermediate layers and a rate of 0.2 is used for Dropout operation. In addition, we did not introduce positional coding information for the stacked multimodal inputs. Since the features of each modality are generated by different encoders, the network performance may not benefit from positional encoding in the context of heterogeneous input. This is further explored in the ablation study (Section 3.5.6.D).

Modality-specific classification head: The multimodal features $h_{multi} \in \mathbb{R}^{MN \times d}$ extracted from the shared encoder E_s are then decomposed to $[h_1, \ldots, h_m, \ldots, h_M]$ for identifying the type of signal transformation applied to each modality. A modality-specific classification head

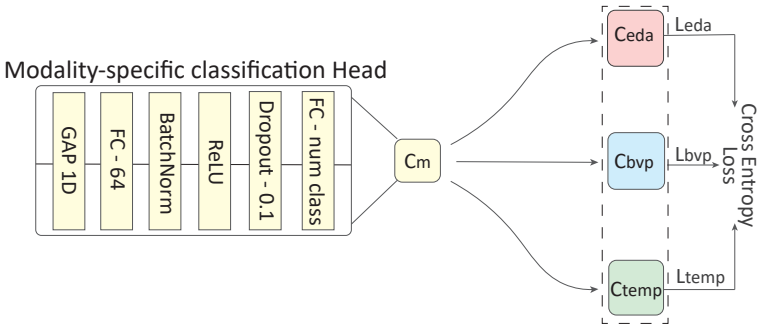


Figure 3.5: Modality-specific classification head C_m for signal transformation recognition task. (GAP: 1D global average pooling, FC: fully-connected layer, BatchNorm: batch normalization, num class: number of signal transformations, i.e., 6 in our work.)

C_m is shown in Fig. 3.5. 1D global average pooling is first applied across all time steps of unimodal features, followed by a fully-connected layer with 64 units. 1D Batch Normalization

is placed before the ReLU layer for more efficient learning and a Dropout layer with a rate of 0.1 is applied to avoid over-fitting. The final fully-connected layer is equipped with a softmax activation function, where the unit number is determined by the number of signal transformations n (i.e., $n = 6$ in our work, 5 transformations plus the original version). In the end, the proposed model is optimized on the pseudo-labeled dataset $\mathcal{L}$ through the total loss L_{total} which is a combination of cross-entropy losses of individual modalities (i.e., EDA, BVP, TEMP in our work):

$$L_m = -\frac{1}{|\mathcal{L}|}\sum_{i=1}^{|\mathcal{L}|} y^i log(C_m(h_m^i)) \tag{3.5}$$

$$L_{total} = \sum_{i=1}^{M} L_m = L_{eda} + L_{bvp} + L_{temp} \tag{3.6}$$

3.3.3 Multimodal emotion recognition based on physiological signals

After pre-training the proposed SSL model with the pretext task on unlabelled data, only the encoder part E is reserved for extracting efficient multimodal representations in a variety of supervised downstream tasks. In this work, we select emotion recognition as our downstream task. A classification head C_{emo} is applied to the output of the encoder E to generate class probabilities for labeled samples $\mathcal{L}_{sup} = \{(x_m^i, y^i), , y^i \in \{1, \ldots, e\}, m \in \{1, \ldots, M\}\}$, where e is the number of the emotion classes. The emotion classification head is constructed in the same way as C_m, except that it accepts multimodal features from encoder E, thus the unit number in the first fully-connected layer changes to 192 (i.e., 64×3), while the unit number in the last fully-connected layer is equal to the emotion class number. The Dropout rate is set to 0.2 to avoid over-fitting. Finally, the proposed model is optimized through the minimization of cross entropy loss L_{sup}.

$$L_{sup} = -\frac{1}{|\mathcal{L}_{sup}|}\sum_{i=1}^{|\mathcal{L}_{sup}|} y^i log(C_{emo}(E(x^i))) \tag{3.7}$$

3.4 Datasets

3.4.1 PRESAGE Dataset

The **PRESAGE dataset** is a large-scale multimodal physiological signal dataset for emotion analysis. The data acquisition is done at the Presage training center[2] in Lille, France, whose mission is to ensure the training of medical students and health professionals through immersion in a recreated hospital environment, where the high-tech mannequins or hired actors, take the place of the patients and students act as doctors. In order to analyze the students' emotional state during the simulation training to optimize the educational program, a large amount of unlabeled multimodal physiological data has been collected from 201 trainees (104 males and 97 females) during five different medical simulation scenarios. Fig. 3.6 (a-e) shows the images of different scenarios captured by the cameras installed in the simulation room. The data collection protocol was approved by the Institutional Review Board of University of Lille with the reference number 2022-626-S108 and all trainees were given a consent form prior to training and were required to fully read the form and provide a signature. To allow students to perform normal medical simulation activities under interference-free conditions, *Empatica E4 Wristband* (Fig. 3.6 (f)), an invasive wearable biometric sensor was adopted to continuously record multimodal physiological signal data of high quality with different frequencies: 3-axis Accelerometer (ACC, 32Hz), Blood Volume Pressure (BVP, 64Hz), Electrodermal Activity (EDA, 4Hz), Skin Temperature (TEMP, 4Hz), Heart Rate (HR, 1Hz), Inter-beat Interval (IBI). In this work, we employed data from three modalities: EDA, BVP and TEMP collected in the five scenarios for self-supervised multimodal representation learning.

3.4.2 WESAD Dataset

The **WESAD dataset** [53] is a multimodal dataset for stress and emotion recognition. Following a study protocol in a restricted laboratory setting, three affective states, namely baseline, stress and amusement, were elicited from 15 subjects during which physiological and motion signals were collected by two separate sensors: RespiBAN (chest-worn device) and Empatica E4 (wrist-worn device). Since we focus on wearable affective computing, only blood volume pressure (BVP, 64 Hz), electrodermal activity (EDA, 4 Hz) and temperature (TEMP, 4 Hz) captured by Empatica E4 were applied to the classification task. According to previous work [15, 53, 69], a stress detection task (*non-stress vs stress*) and a emotion recognition task

[2] https://medecine.univ-lille.fr/presage.

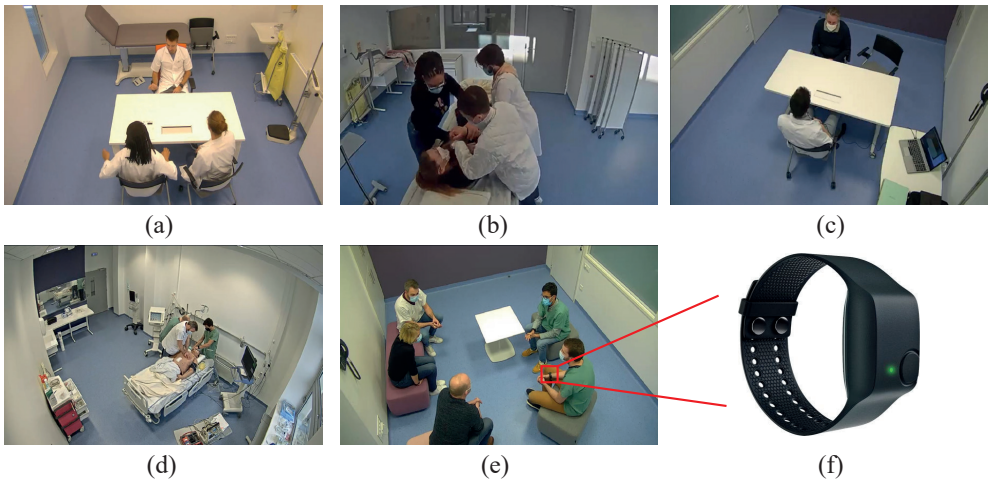


Figure 3.6: Images of different scenarios captured by cameras placed in the simulation training room: (a): Doctor consultation, (b): Prevention of escape for patients in an acute agitated state, (c): Second consultation for patients with suicidal tendencies, (d): Management of cardiac arrest/severe head injury/chest trauma, (e): Diagnostic announcement and (f): the wearable sensor Empatica E4 wristband used for physiological data collection during the simulation training.

(*baseline vs stress vs amusement*) can be performed on the WESAD dataset for supervised learning, where the non-stress class is a combination of the baseline and amusement classes.

3.4.3 CASE Dataset

The **CASE dataset** [157] is a multimodal emotion recognition dataset with continuous annotations. Eight video clips were employed to stimulate four different emotions: amusing, boring, relaxing and scary from 30 subjects. During the experiment, subjects were required to self-assess their own emotional experiences using an annotation interface based on valence-arousal scores, while six physiological signals were recorded at a frequency of 1000 Hz. In our work, we selected blood volume pressure (BVP), electrodermal activity (EDA) and skin temperature (TEMP) signals as in the self-supervised dataset for the classification task. We adopted the same approach as in the literature [70, 139] for the mapping from continuous values of valence and arousal to discrete classes, resulting in a binary (*low vs high valence/arousal*) and a three-class (*low vs medium vs high valence/arousal*) classification problem for supervised learning.

3.4.4 K-EmoCon Dataset

The **K-EmoCon dataset** [158] is a multimodal dataset with multiperspective annotations for emotion recognition in social interactions. 32 subjects were divided into 16 groups for a two-

person debate, during which facial expressions, upper body posture, audio signals, EEG signals and peripheral physiological signals were recorded by different sensors. In our experiments, only blood volume pressure (BVP), electrodermal activity (EDA) and skin temperature (TEMP) signals measured by Empatica E4 were retained for downstream emotion recognition tasks. Tripartite annotations, i.e., self-annotations, partner annotations and external observer annotations were employed to assess subjects' affective states during the debate. Based on the previous work [139], we categorized the arousal- and valence-based annotations into discrete classes, thus forming a binary (low vs high valence/arousal) and a three-class (low vs medium vs high valence/arousal) classification problem for supervised learning.

3.5 Experiments and Results

3.5.1 Data Preprocessing

To eliminate artifacts, we first applied a low-pass Butterworth filter with a cutoff frequency of 0.5 Hz for the EDA and TEMP signals, while the same type of filter with a cutoff frequency of 2 Hz is selected for the BVP signal in PRESAGE, WESAD and K-EmoCon dataset. For the CASE dataset, a low-pass filter with a cutoff frequency of 2 Hz was utilized to clean these three signals. Moreover, we performed z-score normalization as in [159] for each signal recording to reduce the variation in physiological responses between different subjects. Since the four datasets involved in the experiments were collected using sensors with different sampling frequencies, we then uniformly downsampled all signals in the different datasets to the most frequently occurring frequency, i.e., 4 Hz. Subsequently, based on previous work [15, 69], we segmented the signal recordings of all datasets into windows of length 60 s with 99.5% overlap for PRESAGE and WESAD, 99% and 95% overlap for CASE and K-EmoCon, respectively. If the data in a window corresponds to multiple labels, we adopt the same strategy as in the previous work [139], i.e., choosing the one with the majority as the final label. Table 3.1 concludes the learning tasks corresponding to each dataset and the number of samples created after data segmentation. The last column in the table lists the total size of each dataset, where the first dimension represents the total number of samples, while the second and third dimensions represent the signal length at a frequency of 4 Hz in a 60 s window after segmentation (i.e., 240) and the number of modalities (i.e., three modalities: BVP, EDA and TEMP), respectively.

Table 3.1: The learning tasks assigned to each dataset and the corresponding distribution of samples between classes in different datasets. (P: Pretext task, D: Downstream task.)

Dataset	Type	Task	Category (no. of samples)	Total Size
PRESAGE	P	Transformation Recognition	Original version and five transformations (681641)	(4089846, 240, 3)
WESAD	D	Stress-2 Emotion-3	stress (36279), non-stress (85574) baseline (66859), stress (36279), amusement (18715)	(12185, 240, 3)
CASE	D	Arousal-2 Valence-2 Arousal-3 Valence-3	low (33211), high (61919) negative(32017), positive (63113) low (4847), medium (26898), high (63385) negative(9312), neutral (56870), positive (28948)	(95130, 240, 3)
K-EmoCon	D	Arousal-2 Valence-2 Arousal-3 Valence-3	low (3729), high (1488) negative(4050), positive (1167) low (1783), medium (1904), high (1530) negative(1783), neutral (1904), positive (1530)	(5217, 240, 3)

3.5.2 Implementation and model training

The training process of our SSL-based approach consists of two main phases. The first phase is to pre-train the proposed model on the PRESAGE dataset using automatically generated pseudo-labels for signal transformation identification. A set of transformation parameter vectors (15, 0.2, 10, 9, 1.05, 4) was chosen based on the experimental results of the previous study [159] as SNR, magnitude warping variance coefficient, number of permutation segments, number of time warping segments, time-warping stretching coefficient, and number of cropping segments for each modality to generate the five transformations mentioned in Section 3.3.2.A. The pre-training process of the proposed model took approximately 26 hours on an NVIDIA RTX 6000 GPU. The second phase retains only the encoder part of the pre-trained model to extract valid, generalized representations for emotion recognition on WESAD, CASE and K-EmoCon datasets. We did not introduce these three public datasets into pre-training stage in order to verify the knowledge transfer ability of the learned features across different datasets. Ultimately, the proposed model was installed using Pytorch. The optimal models for the pretext and downstream tasks were obtained by the SGD (Stochastic Gradient Descent) optimizer with weight decay parameter of 5e-7 to avoid overfitting. For the first phase (self-supervised pre-training), learning rate, batch size and the number of epochs are set to 5e-3, 32 and 20, respectively. For the second phase (supervised emotion recognition), the learning rate, batch size and number of epochs are set to 1e-4, 128, 20 on WESAD dataset, while for CASE and K-EmoCon datasets, these parameters were set to 1e-3, 64 and 64, respectively.

3.5.3 Evaluation metric and protocol

For a fair comparison, we adopted the same experimental protocol as in [15,53,69,70,139], i.e. Leave-One-Subject-Out cross-validation (LOSO CV), which has the benefit of examining the generalization ability of the model to unpresented subject data. Two metrics, accuracy and F1-score applied in [15,53,69,70,139] were selected to evaluate the performance of the proposed approach on the emotion recognition task. Accuracy represents the proportion of correctly classified samples to the total number of samples. F1-score is considered as a harmonic mean of the precision and recall, which is suggested for evaluating imbalanced datasets.

3.5.4 Baseline Models

Since the exploration of wearable emotion recognition based on peripheral physiological signals has not been well established, a series of baseline models based on fully-supervised learning, unsupervised learning, and self-supervised learning were implemented in addition to available state-of-the-art methods to provide a more comprehensive and reliable performance comparison. The followings are brief descriptions of these models:

Supervised learning-based methods:

- **SimpDCNN** [152]: it is a simple convolutional network consisting of three convolutional blocks with kernel sizes of 24, 16 and 8, each followed by a ReLU activation and a dropout layer.

- **MulT** [160]: it is a transformer-based multimodal fusion method applied to video, audio and text. The unimodal data is first passed through a temporal convolutional network to obtain low-level features, then transformers based on cross-modal attention and self-attention mechanisms are applied successively for effective fusion.

- **ResNet** [161]: it is a 1D convolution-based residual network adapted to physiological signals proposed in [162, 163], which is constructed similarly to ResNet-18, consists mainly of 8 residual blocks with batch normalization (BN) operation and ReLU activation function, where each block contains two convolutional layers. The three modalities: BVP, EDA, TEMP are fed into this network as multi-channel signals.

- **Ours (Supervised)**: it is our proposed multimodal network, trained in a fully-supervised manner.

In addition, three additional supervised methods were applied for the performance comparison on the CASE and K-EmoCon datasets since they lacked baseline results compared to the WESAD dataset.

- **DCNN** [145]: it employs a four-layer 1D convolutional neural network to extract modality-specific features, and a three-layer fully connected network connected at the bottom of the network for classification.
- **Attn-BiLSTM** [164]: it applies a multilayer bidirectional LSTM for capturing valid temporal information for multimodal signals. The attention mechanism was applied to select the most relevant multimodal representation of the emotional state as input for a fully connected layer-based classifier.
- **MMResLSTM** [165]: it uses separate four-layer LSTM-based models for multimodal signals with residual connections. Moreover, the weights of the LSTM layers of both modalities are shared to activate cross-modal communication.

Unsupervised learning-based methods:

- **Autoencoder**: it is an autoencoder with the same encoder part as our proposed model, while the decoder part consists of three transposed convolutional blocks for the reconstruction of the BVP, EDA, TEMP signals. Each unimodal decoder consists of four-layer transposed convolution with the same parameters as the convolutional layers in the encoder.

Self-supervised learning-based methods:

- **SigRep** [139]: it adopts a similar model architecture to SimCLR [142], containing an encoder of four inception-inspired blocks and a projection head consisting of fully connected layers, where each inception block consists of 1D convolutional layers with different kernel sizes and a maximum pooling layer in parallel. The model is applied independently to each signal modality for contrastive representation learning.
- **BENDR** [166]: it is a simpler version of wav2vec 2.0 [167] that was applied to EEG signals. We adapted it for application to peripheral physiological signals at low frequencies. The multi-channel signal consisting of BVP, EDA, TEMP is first passed through a four-layer convolution with kernel sizes of 3, 2, 2, 2, where the GeLU is chosen as the activation function along with GroupNorm and Dropout operations, and the obtained low-level features are randomly masked and fed to the same transformer as our proposed model. The final output features are used to reconstruct the masked features.

For a fair comparison, we followed the parameters provided in these works for the model implementation and applied the same experimental setup. For those models initially designed for non-peripheral physiological signals, the parameters have been slightly adjusted to match the low-frequency wearable data for proper operation.

3.5.5 Experimental Results

3.5.5.A Comparison with state-of-the-art methods

Emotion-related classification tasks were performed on WESAD, CASE, K-EmoCon datasets to evaluate the performance of the proposed SSL model. Tables 3.2, 3.3, 3.4 summarize performance comparisons with state-of-the-art fully-supervised, unsupervised, and self-supervised learning-based methods, which were introduced in Section 3.2 and Section 3.5.4 and hence will not be further described here. For the SSL-based approaches, we report the results under two training modes: **Frozen** (F) and **Fine-Tuned** (T). The first mode refers to freezing the pre-trained encoder part and updating only the parameters of the classification head in the downstream tasks, which is designed to investigate the effectiveness of the learned self-supervised multimodal features. The second mode employs the pre-trained encoder parameters for model initialization and updates all parameters normally to examine the performance gain relative to the **Frozen** mode. From the tables, first, it can be observed that our fully-supervised model obtained better performance than other supervised learning approaches in most emotion recognition tasks, confirming the effectiveness of the proposed architecture. Secondly, regarding our SSL model, the comparison results indicated that, under the **Frozen** mode, our method achieved superior performance over other fully-supervised, unsupervised, and self-supervised based approaches on 6 out of 10 tasks, demonstrating the generalization and high discrimination of the representation learned through the SSL pretext task. In addition, the performance of our model was improved in the **Fine-Tuned** mode, further narrowing the gap with supervised baselines and thus achieving state-of-the-art results in 8 out of 10 tasks. Additionally, it is interesting to note that as the number of supervised samples decreases from WESAD to CASE to K-EmoCon, the higher the performance gain obtained by our SSL-based approach with respect to the supervised approaches. This can be attributed to the fact that supervised learning methods are more prone to overfitting than self-supervised learning methods on low data regimes. Further research on the performance comparison of these two types of methods on limited data is presented in Section 3.5.5.B. Thirdly, in comparison with non-supervised learning methods, we improved the performance of SigRep and BENDR, especially on the CASE and K-EmoCon datasets. The source of this performance gap may be related to the deployed fusion strategies, in addition to

the selected pretext tasks. SigRep [139] learned effective representations for each modality independently through contrastive learning, whereas BENDR [166] regarded multimodal signals as a whole to reconstruct obscured multimodal features. Thus, these two approaches ignored the encoding of inter- and intra-modal correlations, respectively. The impact of different SSL fusion strategies on downstream performance is later investigated in Section 3.5.6.A. Furthermore, the results of the Autoencoder are inferior to other SSL methods. This may be due to the unsupervised nature of its pre-training process which results in more redundant patterns being captured that are irrelevant to the downstream tasks.

Table 3.2: Performance comparison of different emotion recognition tasks with state-of-the-art methods on the WESAD dataset. (Acc: Accuracy, F1: F1-score, SL: supervised learning methods, UL: unsupervised learning methods, SSL: self-supervised learning methods, S: supervised, F: frozen, T: fine-tuned.)

Type	Methods	Stress-2		Emotion-3	
		Acc	F1	Acc	F1
SL	LDA [53]	86.46	83.77	68.85	58.18
	RF [53]	88.33	86.10	76.17	66.33
	SimpDCNN [152]	90.12	88.22	78.30	74.59
	MulT [160]	91.76	91.17	81.09	78.27
	ResNet [161]	91.93	90.97	80.85	79.63
	StressNAS [69]	92.87	-	81.78	-
	Res-TCN [15]	94.16	93.62	83.69	81.61
	Ours (S)	93.83	92.55	84.81	**83.76**
UL	Autoencoder	91.51	90.33	80.39	79.13
SSL	SigRep [139] (F)	92.71	91.99	81.11	78.92
	SigRep [139] (T)	94.91	93.09	84.27	82.35
	BENDR [166] (F)	92.53	91.72	81.98	79.71
	BENDR [166] (T)	93.19	92.55	82.44	80.69
	Ours (F)	94.81	93.69	83.81	82.01
	Ours (T)	**96.29**	**95.11**	**84.94**	82.60

3.5.5.B Self-supervised learning vs Supervised learning on limited labeled data

In the previous section, our self-supervised approach presented state-of-the-art performance on emotion recognition tasks with all labeled data in the dataset. To further investigate the effectiveness of our fine-tuned model on a limited number of labeled samples, we performed a comparison with four supervised learning models: our proposed model with the fully-supervised setting, MulT [160], ResNet [161] and SimpDCNN [152]. MulT and ResNet were selected since they share similar structures to our model and are the best-performing supervised models in addition to ours. Besides, SimpDCNN, as a low-complexity model, is not prone to overfitting

Table 3.3: Performance comparison of different emotion recognition tasks with state-of-the-art methods on the CASE dataset. (Acc: Accuracy, F1: F1-score, SL: supervised learning methods, UL: unsupervised learning methods, SSL: self-supervised learning methods, S: supervised, F: frozen, T: fine-tuned.)

Type	Methods	Valence-2		Valence-3		Arousal-2		Arousal-3	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
SL	SimpDCNN [152]	71.33	68.74	59.20	51.95	67.16	61.60	56.80	53.85
	DCNN [145]	72.35	69.96	59.78	52.80	69.63	63.43	56.09	53.51
	MMResLSTM [165]	73.34	70.96	60.78	53.09	71.12	68.06	57.41	54.69
	Attn-BiLSTM [164]	74.25	71.27	61.97	53.64	70.40	66.52	58.27	54.76
	MulT [160]	74.81	73.17	63.14	62.50	71.28	70.44	62.15	58.48
	ResNet [161]	75.29	74.62	62.89	62.18	72.35	72.19	65.46	59.69
	Ours (S)	76.94	75.06	64.58	63.29	74.15	72.86	**66.32**	**61.78**
UL	Autoencoder	73.23	72.05	60.77	57.32	69.16	67.13	60.08	55.12
	CorrNet [70]	76.37	76.00	60.15	53.00	74.03	72.00	58.22	55.00
SSL	SigRep [139] (F)	71.74	64.78	63.85	54.97	70.79	67.28	63.09	56.99
	SigRep [139] (T)	73.29	69.84	64.63	55.68	72.08	69.45	64.88	58.81
	BENDR [166] (F)	72.94	68.48	61.56	50.86	72.04	67.43	62.37	55.63
	BENDR [166] (T)	72.33	67.62	62.15	53.03	71.51	67.32	63.52	57.01
	Ours (F)	77.49	75.85	65.51	64.07	73.67	70.76	65.09	59.64
	Ours (T)	**78.57**	**77.74**	**66.64**	**64.85**	**74.98**	**73.10**	66.19	60.56

Table 3.4: Performance comparison of different emotion recognition tasks with state-of-the-art methods on the K-EmoCon dataset. (Acc: Accuracy, F1: F1-score, SL: supervised learning methods, UL: unsupervised learning methods, SSL: self-supervised learning methods, S: supervised, F: frozen, T: fine-tuned.)

Type	Methods	Valence-2		Valence-3		Arousal-2		Arousal-3	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
SL	SimpDCNN [152]	77.14	70.06	59.67	48.98	72.48	61.21	46.49	38.34
	DCNN [145]	78.72	72.09	61.97	51.39	73.67	65.53	49.91	39.24
	Attn-BiLSTM [164]	79.76	72.19	62.56	54.35	73.30	66.23	46.95	46.77
	MMResLSTM [165]	78.79	72.76	61.25	51.65	74.31	67.88	44.68	37.19
	MulT [160]	80.13	76.72	63.95	59.07	74.19	72.49	49.25	47.86
	ResNet [161]	80.53	78.04	64.60	62.22	74.35	73.20	50.09	46.77
	Ours (S)	81.51	78.60	64.07	60.83	75.17	73.62	50.42	47.52
UL	Autoencoder	80.58	75.58	63.65	58.32	71.56	69.10	48.83	47.10
SSL	SigRep [139] (F)	78.98	73.15	63.00	54.07	73.36	67.80	47.85	42.58
	SigRep [139] (T)	79.14	73.55	61.74	52.69	73.94	68.71	48.56	43.90
	BENDR [166] (F)	79.83	72.62	61.38	53.20	72.86	66.16	50.68	48.03
	BENDR [166] (T)	78.73	72.15	61.85	54.47	73.82	69.46	52.88	51.24
	Ours (F)	82.95	80.07	66.97	61.28	74.79	73.40	50.76	48.66
	Ours (T)	**84.14**	**81.08**	**68.37**	**63.10**	**76.40**	**74.29**	**54.60**	**52.34**

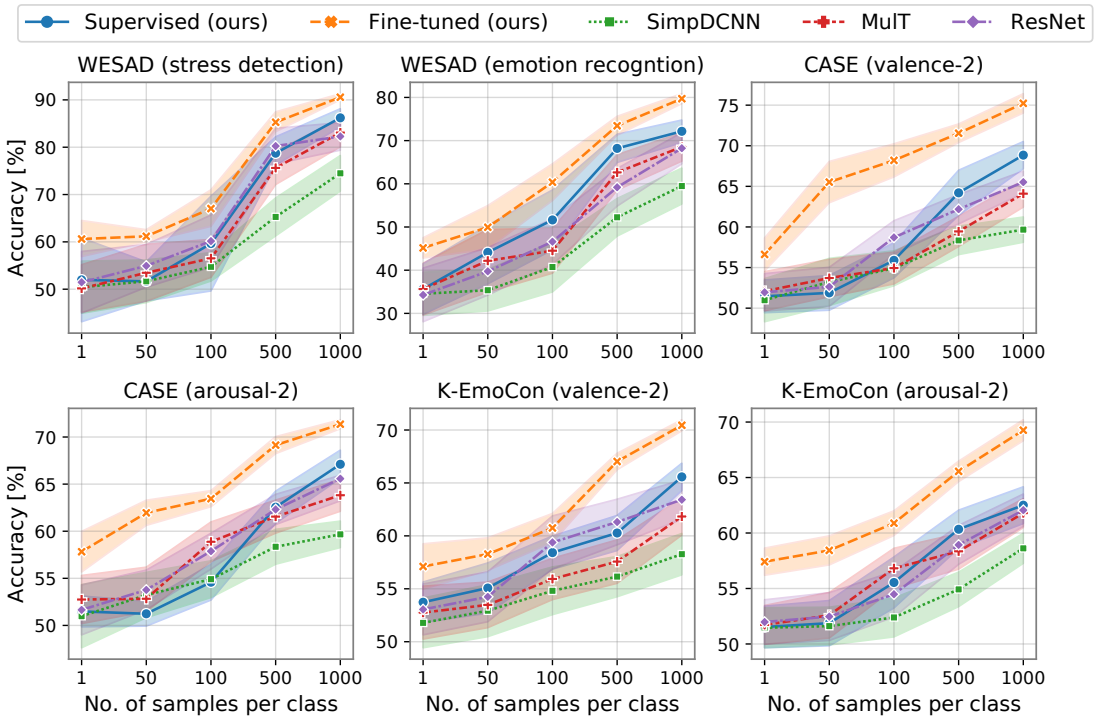


Figure 3.7: Performance comparison with state-of-the-art supervised learning-based methods on limited labeled data sampled from the three emotion recognition datasets. The horizontal axis of each subplot is the number of randomly selected samples from each class, varying from 1 to 1000, while the vertical axis is the corresponding average accuracy.

on limited data, allowing for a more comprehensive performance comparison. We implemented a similar sampling procedure reported in [136, 137], i.e., 1, 50, 100, 500, and 1000 samples were randomly selected for each class in the three datasets for training the classification model. This process was executed 50 times independently for different numbers of samples. The resulting average accuracy and the corresponding standard deviation of all compared models are illustrated in Fig. 3.7. First, our fine-tuned model consistently outperforms other supervised learning-based models for sample sizes varying from 1 to 1000 on the emotion recognition tasks of all three datasets. Among supervised learning-based methods, SimpDCNN exhibited the poorest results, over which our SSL model could achieve significant performance gains of 6.84% - 21.19% for different downstream tasks. Our fully-supervised model yields the highest results compared to other supervised models, whereas the fine-tuned model initialized by self-supervised learning parameters continues to enhance performance by 5.24% - 13.63%. Second, for all downstream tasks, the standard deviation obtained by our fine-tuned model is narrower with respect to the supervised learning-based deep models, demonstrating its superior generalization ability across different samples. The above findings are consistent with those reported

in [168] that the advantage of the self-supervised learning-based method is its better regularisation on low data regimes to avoid overfitting problems compared to fully-supervised methods. As the amount of available labeled data increases, the difference in performance between the two types of models gradually decreases. Overall, the comparison results suggest that the proposed method can produce more meaningful and robust representations for wearable emotion recognition than fully-supervised methods, offering a potential solution to the problem of little labeled data.

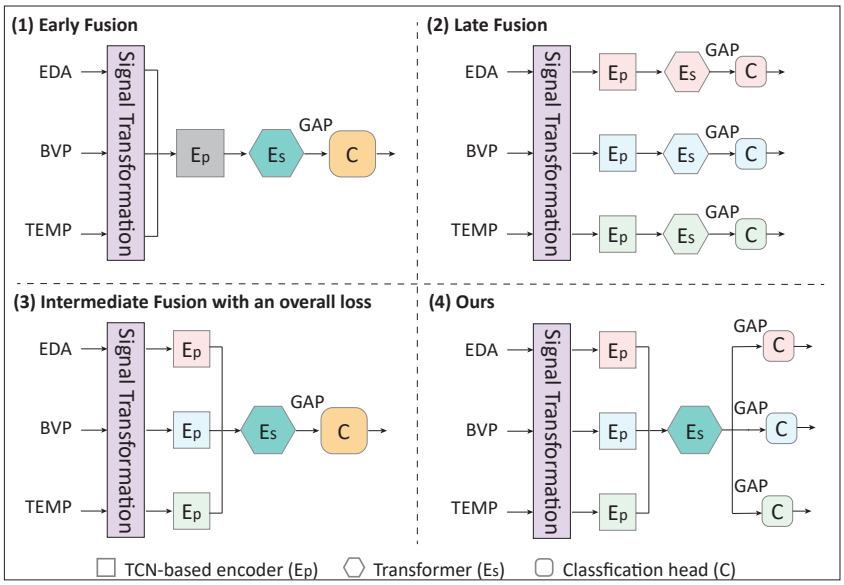


Figure 3.8: Different architectures used in the ablation studies of fusion strategies. (GAP: 1D global average pooling applied before classification.)

3.5.6 Ablation Studies

Different types of ablation experiments were designed and conducted on the WESAD, CASE and K-EmoCon datasets to verify the validity of the proposed method. The encoder part of the models involved was trained in freezing mode and the obtained emotion recognition results are reported in the following sections.

3.5.6.A Ablation study of different fusion strategies

To demonstrate the effectiveness of the selected fusion strategy, we conducted ablation studies on different SSL fusion strategies. Three variants of the proposed model based on early fusion,

late fusion, and intermediate fusion strategies were implemented for comparison. The corresponding model architectures used for comparison are illustrated in Fig 3.8. In all variant models, the TCN-based encoder E_p, transformer E_s and signal transformation classification head C all share the same parameter settings as the proposed model. For the **Early fusion** setup, we treated the multimodal physiological signal as a whole, i.e. a multichannel signal, from which multimodal representations will be learned directly. For the **Late fusion** setup, separate encoders were applied to individual modalities to extract unimodal features for classification. In addition, the third variant model has the same fusion strategy as ours, where unimodal features were first captured and then concatenated to learn more advanced multimodal features. The difference, however, is that this model performs classification by multimodal features. This is to verify the necessity of conducting modality-specific classification in the proposed method, and we refer to this setup as **Intermediate fusion with an overall loss**. Consequently, the corresponding evaluation results are listed in Table 3.5. Our model consistently achieved the best performance on all datasets, demonstrating the effectiveness of the selected fusion strategy, i.e., intermediate fusion. In addition, the intermediate fusion-based models performed better than those based on the other two fusions. This can be attributed to the fact that the intermediate fusion simultaneously models the heterogeneity and coordination of multimodal physiological signals, whereas the other two fusion approaches only consider one of these two properties. Furthermore, the third setting **Intermediate fusion with an overall loss** performs slightly worse than our model, affirming the importance of modality-specific classification. The benefit of applying modality-specific loss functions is that it forces the model to learn, for each modality, generic features that are robust to perturbations in the time or magnitude domain, while the application of an overall loss fails to distinguish each modality's contribution to the learned representation.

Table 3.5: Ablation study of different fusion strategies: average accuracy and F1-score obtained for emotion recognition on WESAD, CASE and K-EmoCon dataset using different variant model. (S-2: Stress-2, E-3: Emotion-3, V-2: Valence-2, A-2: Arousal-2, Inter w/ ol: Intermediate fusion with an overall loss.)

Type	WESAD				CASE				K-EmoCon			
	S-2		E-3		V-2		A-2		V-2		A-2	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Early	91.22	89.94	80.82	78.79	73.01	72.20	69.13	67.06	79.20	74.12	71.57	69.60
Late	93.02	91.73	81.48	80.91	75.58	72.27	71.96	68.52	80.94	76.43	72.81	70.87
Inter w/ ol	93.53	92.77	82.82	81.62	76.69	73.52	72.24	69.11	81.48	77.22	73.06	71.29
Ours	**94.81**	**93.69**	**83.81**	**82.01**	**77.49**	**75.58**	**73.67**	**70.76**	**82.95**	**80.07**	**74.79**	**73.40**

3.5.6.B Ablation study of different modalities

We conducted an ablation study of three modalities: EDA, BVP, TEMP and their combinations to explore their performance on emotion recognition tasks. The resulting average accuracies and F1-scores of our model are summarized in Table 3.6. First, for the unimodal performance, the EDA signal performs outstandingly well among all the modalities, especially when detecting stress and arousal states. This is consistent with previous findings that EDA is one of the most relevant indicators of stress [169] and has even been adopted as ground truth in some studies [170, 171] for the stress analysis of other signals. In addition, it has been proven to correlate linearly with arousal [171]. In the bimodal-based classification, we first observed that the **BVP+EDA** setup performed better on the stress-related tasks (i.e. S-2 and E-3 on the WESAD dataset) than the other setups. This suggests that the BVP signal and the EDA signal are highly coordinated and correlated when the stress state is elicited, making their combination more effective for detection. This finding is quite reasonable. The BVP signal contains information on heart rate (HR) and heart rate variability (HRV) thus providing a strong correlation with stress states. In [40], HRV and EDA were identified as the most relevant physiological indicators for the real-time stress detection task. Secondly, the **EDA+TEMP** setup achieved the best performance on the classification task regarding arousal level. This finding is supported by previous research [172] which indicated that EDA and TEMP had a positive and negative correlation with arousal scores respectively. Lastly, our model achieved performance gains on both bimodal and trimodal data in most cases, confirming again its effectiveness for multimodal fusion.

Table 3.6: Ablation study of different modalities and their combinations: average accuracy and F1-score obtained with different modality combinations in the downstream emotion recognition tasks, where the best performing individual modality and bimodal combinations for each task are underlined. (S-2: Stress-2, E-3: Emotion-3, V-2: Valence-2, A-2: Arousal-2.)

Modality	WESAD				CASE				K-EmoCon			
	S-2		E-3		V-2		A-2		V-2		A-2	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
EDA	92.36	90.58	78.72	75.90	75.21	74.80	72.15	70.13	80.65	74.60	73.09	72.69
BVP	87.82	86.35	75.20	68.80	75.90	75.15	69.23	65.07	80.76	74.13	72.88	70.67
TEMP	78.15	76.91	69.86	65.37	71.64	68.66	68.97	62.16	79.02	72.78	72.52	70.42
EDA + BVP	93.73	92.38	82.32	80.61	76.26	75.13	72.13	70.27	80.87	75.48	73.14	71.78
EDA + TEMP	90.95	89.62	79.74	76.09	76.03	74.97	72.92	70.54	81.70	77.77	74.61	72.93
BVP + TEMP	84.82	80.45	72.88	66.16	72.35	71.31	71.14	68.04	80.12	75.05	72.43	69.86
All	**94.81**	**93.69**	**83.81**	**82.01**	**77.49**	**75.85**	**73.67**	**70.76**	**82.95**	**80.07**	**74.79**	**73.40**

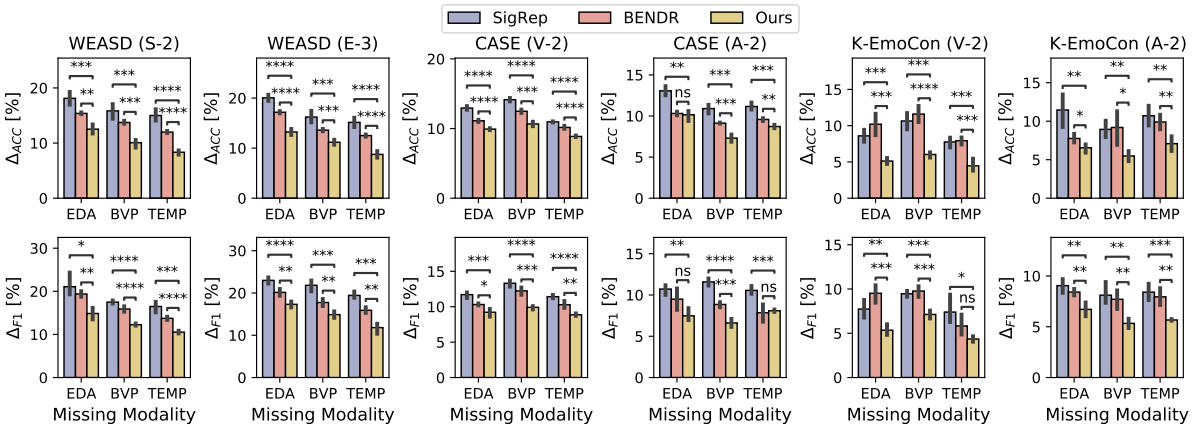


Figure 3.9: Evaluation results of the robustness of the SSL methods in the presence of missing modalities. The horizontal axis of each subplot represents the name of the missing modality, while the vertical axis represents the drops in model performance compared to the case of complete modalities, where the metrics of the vertical axes in the first and second rows are accuracy and F1-score, respectively. (ns: no significant difference; $*$: $p < 0.05$, the more asterisks, the more significant the difference.)

3.5.6.C Ablation study of missing modalities

We investigate the robustness of the proposed SSL model when a signal modality is missing in downstream tasks, which is quite common in real-world scenarios. There exist a variety of experimental setups for incomplete modalities. Based on [173], we selected the most challenging one, i.e., a modality is missing in both training and testing of the downstream task, where 50% of the multimodal samples were first randomly selected and subsequently the data values of a specific modality were set to 0 to simulate its absence. The robustness of the proposed SSL model was measured by calculating its difference in performance in two cases: one with all modalities present and one with missing modalities. The smaller the difference, the more robust the model is considered to be. The above experimental procedure was repeated 10 times. Additionally, we benchmarked our model against the SSL baseline models: SigRep and BENDR. Fig. 3.9 presents the average degradation in accuracy and F1-score of the compared models when a modality is missing in different downstream tasks. A series of t-tests were further conducted on the performance differences for a more systematic robustness comparison. From the evaluation results, we can first observe that the performance drops of our model are significantly lower ($p < 0.05$) than other SSL models on most tasks. This demonstrates the superiority of the proposed method in terms of robustness. Second, we also note that the impact of missing modalities on the robustness of SSL methods is task-dependent. For downstream tasks related to stress and arousal levels, more severe performance declines could be obtained

in the absence of the EDA signal, compared to the other two modalities. This result indicates the importance of the EDA signal for identifying these two emotional states. Similarly, missing the TEMP signal also leads to a considerable reduced performance in arousal-based recognition, whereas, in the valence-based tasks, the loss of the BVP signal has the greatest impact on performance. The above results, consistent with those in Section 3.5.6.B, reconfirm the effect of different modalities on specific emotion recognition.

3.5.6.D Ablation study of different model components

We also investigate the impact of different model components on the performance of downstream classification tasks. To validate the contributions of the modality-specific encoder and the shared encoder, we designed two alternative models: **No TCN** and **No Transformer**. **No TCN** eliminates the temporal convolution network (TCN) where unimodal data is passed directly through the projection layer (i.e. a fully connected layer with 128 units) in the modality-specific encoder shown in Fig. 3.3 and the resulting unimodal low-level features are then concatenated as a whole and fed into the transformer. **No Transformer** removes the multimodal transformer, where unimodal features are first extracted by modality-specific encoders and then averaged along the time dimension by the 1D global average pooling (illustrated in Fig. 3.5) for the final classification tasks. Table 3.7 present the classification results of the above two variant models on the three datasets. Our proposed model enhances both the performance of **No TCN** and **No Transformer** models on all classification tasks across different datasets, highlighting the importance of capturing the heterogeneity and cross-modal correlation of multimodal signals simultaneously. Subsequently, we examined whether the addition of positional encoding could lead to better performance for the transformers with heterogeneous embedding as input. We employed two types of positional encoding (PE): **With fixed PE** and **With learnable PE** in the transformer and compared their performance with our PE-free model. **With fixed PE** added the fixed positional encoding obtained from sine and cosine functions of different frequencies as proposed in [156] to the input embedding of the multimodal transformer while **With learnable PE** adopted the same learnable positional encoding in [174]. Table 3.7 also show the classification results of the proposed model with different PE setting. We observed that temporal context information injected by two types of PE did not contribute to model performance on all classification tasks as expected. This can be attributed to the fact that the multimodal embeddings generated by the separate encoders already own different structures, hence the additional positional information introduces redundancy into the model.

Table 3.7: Ablation study of different model components: average accuracy and F1-score obtained for emotion recognition on WESAD, CASE and K-EmoCon dataset using different variant model. (S-2: Stress-2, E-3: Emotion-3, V-2: Valence-2, A-2: Arousal-2)

Model Variants	WESAD				CASE				K-EmoCon			
	S-2		E-3		V-2		A-2		V-2		A-2	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
No TCN	91.09	90.42	79.86	78.77	64.15	62.20	66.87	57.87	79.39	75.34	71.24	66.59
No Transformer	92.18	91.55	81.08	79.30	74.65	71.53	70.04	68.40	80.67	76.17	72.09	70.66
w/ fixed PE	93.49	91.63	82.38	80.47	76.37	74.45	72.59	70.05	80.32	77.43	73.35	71.20
w/ learnable PE	92.68	91.32	82.42	81.24	76.46	75.35	73.16	70.33	81.59	78.62	74.08	72.09
Our Model	**94.81**	**93.69**	**83.81**	**82.01**	**77.49**	**75.58**	**73.67**	**70.76**	**82.95**	**80.07**	**74.79**	**73.40**

3.5.6.E Ablation study of different signal transformation task

We further explored the impact of using individual transformations and their combinations in the pretext task on downstream emotion recognition performance. As mentioned in Section 3.3.2.A, the five transforms employed can be divided into two classes, i.e., magnitude domain transformations and time domain transformations. Therefore, the types of combinations are arranged accordingly as combinations of transformations within the same domain and combinations of transformations across domains. The evaluation results obtained on different emotion classification tasks are presented in Table 3.8. First, we noticed that **Permutation** and **Time-Warping**, which perturbed the temporal order and duration of events within the window, performed best among the individual signal transformations, which is consistent with the results in [135, 152], demonstrating the necessity to encode the temporal relationships of signals for emotion recognition. Second, the pre-trained models obtained by combining the same domain or cross-domain transformations generally perform better than those based on individual transformations. The performance of these combinations varies depending on the specific task. For the same domain transformation combinations, **P+T+C** performs better for stress-related tasks, whereas **N+M** is more appropriate for arousal and valence-based tasks. For the cross-domain combinations, **N+T** exhibited the best performance on the classification tasks regarding stress, while **N+P** and **M+C** performed best in predicting the arousal and valence states. Finally, we found that models based on cross-domain combinations outperformed those based on the same domain combinations in two-thirds of the downstream tasks. Meanwhile, our pre-trained models using the full set of transformations consistently achieved superior performance in the classification tasks. This can be attributed to the fact that different types of transformations inject diverse prior knowledge for multimodal representation learning, thus contributing to the generalizability of the network.

Table 3.8: Ablation study of individual signal transformations and their combinations: average accuracy and F1score obtained for emotion recognition on WESAD, CASE and K-EmoCon datasets using different transformations in self-supervised pertaining, where the best performing transformations and combinations of transformations in each task are underlined. (N: Noise addition, M: Magnitude-warping, P: Permutation, T: Time-warping, C: Cropping.)

Type	WESAD				CASE				K-EmoCon			
	S-2		E-3		V-2		A-2		V-2		A-2	
Single	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
N	90.18	89.16	78.04	76.24	73.53	70.42	69.16	61.48	79.69	74.30	70.23	67.38
M	89.74	87.86	76.90	73.77	68.75	67.46	68.28	66.94	80.13	75.44	71.11	70.54
P	91.20	89.33	78.38	75.67	73.61	71.17	71.46	69.82	80.22	78.12	71.69	70.82
T	91.34	90.87	81.15	80.44	71.17	69.53	70.81	69.01	80.39	78.31	70.59	67.90
C	89.48	88.06	79.21	77.38	69.35	62.46	69.17	66.60	79.68	74.07	70.87	69.05
Same Domain	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
N+M	89.69	88.43	77.52	75.59	75.87	74.22	71.83	69.62	81.46	78.95	73.14	72.08
P+T+C	93.67	92.88	82.31	80.47	73.71	71.95	70.92	69.05	80.77	78.58	72.66	71.11
Cross Domain	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
N+P	92.15	91.12	80.02	78.93	75.77	73.57	72.85	70.09	80.26	78.54	73.09	71.73
M+P	91.75	90.47	79.24	77.31	74.01	72.92	72.39	69.87	80.34	79.15	71.89	70.76
N+T	92.95	91.69	83.11	81.14	73.51	71.21	70.03	68.91	81.57	79.43	72.04	69.75
M+T	91.51	90.18	79.85	77.52	71.45	70.07	71.61	69.14	80.29	79.04	71.39	69.61
N+C	91.28	90.15	78.81	76.66	73.96	72.60	70.57	68.44	79.52	77.83	73.13	71.52
M+C	90.08	89.62	78.27	76.40	71.90	69.45	69.53	67.89	82.14	79.61	73.89	72.51
All	**94.81**	**93.69**	**83.81**	**82.01**	**77.49**	**75.85**	**73.67**	**70.76**	**82.95**	**80.07**	**74.79**	**73.40**

3.6 Conclusion

In this paper, we have proposed a self-supervised multimodal representation learning framework for emotion recognition based on physiological signals. Signal transformation recognition is defined as a pretext task, where a large amount of unsupervised data is automatically labeled by the imposed signal transformation category for pre-training of the SSL model. Subsequently, the encoder part of the pre-trained model consisting of a temporal convolution network and transformer is maintained to extract effective multimodal representations for the downstream task, i.e. emotion recognition. Eventually, we executed the pre-training on a large-scale unrestricted emotion dataset PRESAGE and verified the validity of the proposed method on three public multimodal emotion recognition datasets. Experimental results indicated that our approach surpassed fully-supervised or self-supervised learning methods, achieving state-of-the-art results in various emotion-related tasks. Additionally, the proposed method performs better than the fully-supervised learning approaches on limited labeled data, demonstrating its superior generalization ability to avoid overfitting problems. A series of ablation studies have also confirmed the efficiency of the designed model architecture. In the next chapter, we will report on the technical contributions on multimodal aspects, i.e., stress and pain detection based on physiological and behavioral signals.

4

Fusion of Physiological and Behavioural Signals on SPD Manifolds with Application to Stress and Pain Detection

Contents

This chapter mainly concerns technical contributions to multimodal aspects, where a deep geometric framework with the symmetric positive definite (SPD) matrix as the joint representation of multimodal signals was proposed to cope with stress and pain detection tasks. In Section 4.1, we shortly introduce stress and pain detection methods and their challenges, Section 4.2 presents relevant multimodal studies and applications of SPD-based representations. A detailed description of the proposed method is given in Section 4.3, followed by Section 4.4 showing details of stress and pain detection experiments and evaluation results of our method, and in Section 4.5, we briefly summarise our work.

4.1 Introduction

From a healthcare point of view, two frequently desired tasks in emotion recognition are stress and pain detection. Stress is a specific emotional state defined by high arousal and negative valence condition according to the dimensional model. When exposed to long-term stressors, a person's mental and physical state can be negatively affected, which can further lead to chronic health problems such as headaches, insomnia or cardiovascular disease [175–177]. On the other hand, pain is characterized as an unpleasant sensory and emotional experience associated with actual or potential tissue damage by the International Association for the Study of Pain (IASP) [178]. Its accurate assessment plays a key role in diagnosing the condition, monitoring post-operative progress and optimizing treatment options [179]. Due to the side effects of these two sensations on health conditions, automatic detection methods are greatly encouraged, especially if the patient is alert or uncooperative. A variety of sensors can be applied to collect stress/pain indicators from different perspectives, whereby the corresponding approaches for stress/pain detection can be mainly divided into three categories: 1) physiological based methods via bio-signals (e.g., electroencephalogram (EEG), electrodermal activity (EDA), etc.); 2) behavioural-based methods via physical signals (e.g., facial expressions, speech, body movements, etc.); 3) multimodal-based methods via a combination of physiological and behavioural signals.

As complementary information between multimodalities contributes more to the robustness and reliability of the system, therefore, stress/pain detection combining physiological and behavioural indicators has become more attractive. However, the application of multimodal data for automatic identification requires knowledge from different domains, constituting an obstacle for most researchers. In addition, how to effectively fuse multimodal data remains an important challenge for such systems. From the fusion strategy perspective, the few existing efforts are mostly based on early or late fusion and therefore unable to incorporate both intra-modal and inter-modal correlations. From the feature engineering perspective, handcrafted feature-based

methods dominate, however, they are typically cumbersome and inefficient for multimodal data. In this study, we address the above problem by introducing a geometric framework that employs symmetric positive definite (SPD) matrices extracted from physiological and behavioural signals as a joint multimodal feature representation on SPD manifold for stress/pain detection. Continuous multimodal data recording can first be converted into SPD matrix sequences. The tangent space mapping method is then applied to locally flatten the manifold and approximate the SPD matrix sequences by tangent space vector sequences, where an LSTM-based deep neural network can be implemented to learn the context correlations of input sequences for classification. The overview of the proposed method is illustrated in Fig. 4.1. In summary, the main contributions of this work are:

- We made the first attempt of applying the SPD matrix to fuse behavioural signals and physiological signals for stress/pain detection.
- We employed the SPD-based representation to efficiently capture correlations within and across modalities at different time steps.
- We adopted tagent space mapping for the linearization of manifold data which facilitates high-level temporal feature extraction by the LSTM-based deep neural network.
- We obtained state-of-the-art results on both stress and pain detection tasks using the proposed method. Furthermore, Our integration strategy is proven to be more effective than early fusion or late fusion.

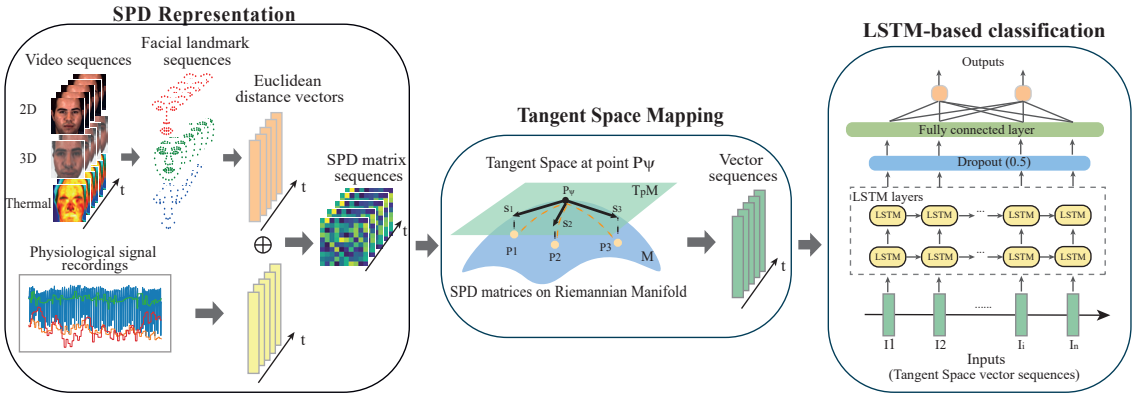


Figure 4.1: Overview of the proposed framework. First, the SPD matrix sequences that incorporate the correlation information between multimodal data (i.e. physiological and behavioural signals) can be extracted from the segmented data records. Subsequently, the tangent space mapping projects the SPD matrix sequences to the vector sequences in the tangent space. Finally, these vectors can be used as input to the LSTM-based classification network for stress/-pain recognition.

4.2 Related Work

The multimodal framework is promising to improve the performance of stress/pain detection. However, few studies can be found using combined data from different fields (i.e. physiological and behavioural).

Multimodal stress/pain detection Some of the multimodal methods presented [5, 15, 53, 69, 132] in Section 2.5 are targeted at stress and pain detection tasks. Among these methods, multimodal stress detection approaches that combine physiological signals and motion signals (i.e., accelerometers) are relatively popular, as they can be captured by wearable sensors and are well suited to the real-life setting. Alternatively, a few stress detection methods also attempt to fuse physiological modalities with behavioural modalities based on visual data. Aigrain et al. [180] captured body video, high-resolution facial images and physiological signals from 25 subjects during a mental arithmetic test for stress detection, where 101 features from behavioural and physiological signals were extracted to train an SVM classifier. For pain detection, a majority of methods are based merely on images or videos of facial expressions, thus the effectiveness of multimodality has not been well investigated. In the work of [181], statistical/time/frequency domain features of the multi-physiological signals (electrocardiogram (ECG), electromyography (EMG), skin conductance level (SCL)) and appearance/geometric features extracted from the facial videos were served as input to a Random Forest (RF) algorithm for continuous estimation of pain intensity. Werner et al. [131] collected facial distances and gradient-based features from video frames which were combined with the statistical features calculated from biological signals. The resulting multimodal vectors were employed to train a Random Forest (RF) model for pain assessment. The multimodal approach described above shows promising results in terms of stress/pain detection. However, regarding the integration of multimodal data, most of them applied the early fusion strategy, i.e. simply concatenating the features of different modalities as a whole for the corresponding learning task. Although inter-modal interactions could be captured from the obtained joint representation, intra-modal correlations were relatively ignored. In addition, most of these methods employed handcrafted features, their extraction and selection proving to be a challenging task in the multimodal case, as not all features contributed to the learning performance.

Symmetric Positive Definite (SPD) matrices Recently, covariance-based representations have gained great popularity in computer vision and machine learning. This success can be explained by three major advantages. Firstly, several characteristics can be fused into a single tensor and deliver higher-order statistical information. Secondly, the covariance matrices are sym-

metric positive definite (SPD) matrices with well-established mathematical theoretical properties [182, 183]. In addition, the SPD matrices have shown impressive accomplishments in many real-world applications such as pedestrian detection [184], facial expression recognition [185], brain-computer interfaces [186], etc. However, all of these applications concentrate solely on behavioural perspectives or physiological perspectives. Liu et al. [187] proposed a multimodal emotion recognition approach using video and audio modalities. Covariance matrix, linear subspace, and Gaussian distribution were built from facial video frames and regarded as points on Riemannian manifolds. Subsequently, the similarity matrix calculated using different Riemann kernels is fed into multiple classifiers (i.e., SVM, partial least squares, and logistic regression). However, they only constructed the SPD matrix for the video modality, the audio features were extracted using an existing toolkit, and fed into the same type of classifier as the video modality. In the end, the final fusion of the two modalities was established on the decision level. Thus, the exploration of inter-modal correlations using SPD matrix is still missing here. In the work of [183], they presented a more general covariance-based SPD representation, containing additional cross-covariance information from different time steps for action recognition. Inspired by [183], we migrate this new type of tensor representation to the scenario of multimodal stress/pain detection and show its effectiveness in this paper. Different from the previous work, we fuse behavioural and physiological information into one single SPD matrix-based representation, which not only incorporates intra-modal correlations, but also allows for exchanges across two modalities. To the best of our knowledge, this is the first use of the geometry of SPD manifold matrices to merge physiological and behavioural signals.

4.3 Proposed Method

4.3.1 Symmetric positive definite (SPD) matrix for multimodal signal

Let $\mathbf{x_i} = [v_1, ..., v_D]^T \in \mathbb{R}^D, D \geq 2$ represent a multimodal signal vector comprising the behavioural and physiological signals at the i-th timestamp, where the number of signals is denoted by D. A short-segment can be extracted from the continuous signal recordings of a trial, resulting in a centered matrix $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_i}..., \mathbf{x_N}] \in \mathbb{R}^{D \times N}$, where N is the number of time instants for each segment and the sample mean vector is $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i} = \mathbf{0}$. The outer product operation (denoted by the symbol $\otimes$) is then performed on all signal column vector

pairs $(\mathbf{x_i}, \mathbf{x_j})$ for i and j $= 1, ..., N$ in $\mathbf{X}$ and consequently produces a partition matrix :

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{x_1} \otimes \mathbf{x_1} & \cdots & \mathbf{x_1} \otimes \mathbf{x_N} \\ \mathbf{x_2} \otimes \mathbf{x_1} & \cdots & \mathbf{x_2} \otimes \mathbf{x_N} \\ \vdots & \ddots & \vdots \\ \mathbf{x_N} \otimes \mathbf{x_1} & ... & \mathbf{x_N} \otimes \mathbf{x_N} \end{bmatrix} \in \mathbb{R}^{DN \times DN} \tag{4.1}$$

where the element of $\mathbf{\Omega}$ at position (i, j) is given by:

$$\mathbf{\Omega(i, j)} = \mathbf{x_i} \otimes \mathbf{x_j} = \mathbf{x_i x_j}^T \in \mathbb{R}^{D \times D} \tag{4.2}$$

The sample covariance matrix $\mathbf{S}$ which is a SPD matrix can be derived from the diagonal elements of $\mathbf{\Omega}$:

$$\mathbf{S} = \frac{1}{N-1} \sum_{i=1}^{N} \mathbf{\Omega(i, i)} = \frac{1}{N-1} \sum_{i=1}^{N} \mathbf{x_i x_i}^T \in \mathbb{R}^{D \times D} \tag{4.3}$$

As can be seen from (4.3), the correlation statistics on the pairs of the signal vector $(\mathbf{x_i}, \mathbf{x_j})$ at the two particular time instant i and j are completely ignored. As such, the second SPD matrix defined as cross-covariance can be extracted from the off-diagonal elements of $\mathbf{\Omega}$, which contains the correlation information of the signal vectors at different timestamps and is denoted by $\mathbf{C}$:

$$\begin{aligned} \mathbf{C} &= \frac{1}{N^2 - N} \sum_{i=1, j=1, i \neq j}^{N, N} \mathbf{\Omega(i, j)} \\ &= \frac{1}{N^2 - N} \sum_{i=1, j=1, i \neq j}^{N, N} \mathbf{x_i x_j}^T \in \mathbb{R}^{D \times D} \end{aligned} \tag{4.4}$$

The covariance $\mathbf{S}$ and cross-covariance $\mathbf{C}$, are then combined in a symmetric manner to form a more generalised covariance-based representation which remains a SPD matrix and is denoted by $\mathbf{P}$:

$$\mathbf{P} = \begin{bmatrix} \mathbf{S} & \mathbf{C} & \cdots & \mathbf{C} \\ \mathbf{C} & \mathbf{S} & \cdots & \mathbf{C} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C} & \mathbf{C} & \cdots & \mathbf{S} \end{bmatrix} \in \mathbb{R}^{(m \times D) \times (m \times D)} \tag{4.5}$$

where m is the dimension of the new SPD matrix $\mathbf{P}$, in other words, $\mathbf{P}$ is composed of m blocks of $\mathbf{S}$ and $m(m-1)$ blocks of $\mathbf{C}$. A larger m corresponds to a higher computational cost, while the information ratio of the covariance to the cross-covariance decreases [183].

4.3.2 Riemannian Geometry of SPD Matrices

4.3.2.A Mathematical Preliminaires

A manifold $\mathcal{M}$ is a topological space that is locally homeomorphic to Euclidean space in a neighbourhood of each point. A differentiable manifold is a topological manifold endowed with a smooth structure, allowing differential calculus to be performed on $\mathcal{M}$. Each point of a differentiable manifold is attached to a tangent space which is a real vector space containing all tangent vectors to $\mathcal{M}$ at the point. A differential manifold is referred to as a Riemannian manifold if the tangent space at each point p on the manifold defines a inner product and its value varies smoothly with p. The set of symmetric positive definite (SPD) matrices $\mathcal{P}(n) = \left\{\mathbf{P} \in S(n), \mathbf{u}^T\mathbf{P}\mathbf{u} > 0, \forall \mathbf{u} \in \mathbb{R}^n\right\}$ that we wish to work with, forms a Riemannian manifold, where $S(n) = \left\{\mathbf{S} \in M(n), \mathbf{S}^T = \mathbf{S}\right\}$ is the space of all $n \times n$ symmetric matrices in the space of square real matrices $M(n)$.

4.3.2.B Riemannian metric and geodesic distance

Given a smooth manifold $\mathcal{M}$, a Riemannian metric g on $\mathcal{M}$ is a smooth family of inner products $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ on the tangent spaces $T_p\mathcal{M}$ of $\mathcal{M}$. The smoothness condition for g is satisfied if the function $p \in \mathcal{M} \to g_p(X_p, Y_p) \in \mathbb{R}$ is smooth for each of the local smooth vector fields X, Y in $\mathcal{M}$. With the definition of the inner product g, we can then define the length of the curve joining any two points on $\mathcal{M}$. Let $\gamma : [a, b] \to \mathcal{M}$ be a smooth curve. Then the length $L[\gamma]$ of γ is given by

$$L[\gamma] = \int_a^b g_{\gamma(t)}(\gamma^{'}(t), \gamma^{'}(t))^{1/2} dt \tag{4.6}$$

The curve γ with the shortest length between two points on $\mathcal{M}$ is called the *geodesic*. Then, the distance between two points p_1 and p_2 on $\mathcal{M}$ is defined as

$$d(p_1, p_2) = inf\{L[\gamma] \mid \gamma : [a, b] \to \mathcal{M}, \gamma(a) = p_1, \gamma(b) = p_2\} \tag{4.7}$$

Since the space of SPD matrices $\mathcal{P}(n)$ is a Riemannian manifold, various efforts have been made to measure the distance between SPD matrices with different metrics. In this study, we consider one of the most popular metrics, the Affine-invariant Riemannian metric (AIRM) metric [188] which measures the geodesic distance induced by intrinsic geometry of Riemannian manifold. By applying this metric, the *geodesic* connecting any two SPD matrices $\mathbf{P_i}$ and $\mathbf{P_j}$ can be easily derived:

$$\gamma(t) = \mathbf{P_i}^{\frac{1}{2}} \exp(t \log(\mathbf{P_i}^{-\frac{1}{2}} \mathbf{P_j} \mathbf{P_i}^{-\frac{1}{2}})) \mathbf{P_i}^{\frac{1}{2}} \tag{4.8}$$

and the geodesic distance induced by the Riemannian metric is defined as

$$\delta_R(\mathbf{P_i}, \mathbf{P_j}) = \|\log(\mathbf{P_i}^{-1}\mathbf{P_j})\|_F = \Big[\sum_{i=1}^{n} \log^2\lambda_i\Big]^{\frac{1}{2}} \tag{4.9}$$

where $\|\cdot\|_F$ is the Frobenius norm operator, and $\lambda_i, i = 1 \cdots n$ denote the eigenvalues of $\mathbf{P_i}^{-1}\mathbf{P_j}$.

4.3.2.C Exponential and Logarithm Maps

Given a SPD matrix $\mathbf{P} \in \mathcal{M}$, a tangent space $T_\mathbf{p}\mathcal{M}$ can be defined, composed by all derivatives of curves through $\mathbf{P}$. Then the exponential map $\mathrm{Exp}_\mathbf{P}(\cdot) : T_\mathbf{P}\mathcal{M} \to \mathcal{M}$ and its inverse logarithm map, $\mathrm{Log}_\mathbf{P}(\cdot) : \mathcal{M} \to T_\mathbf{P}\mathcal{M}$ are defined over Riemannian manifolds for exchanging between the manifold and its tangent space at $\mathbf{P}$. More formally, let $\mathbf{S} \in T_\mathbf{p}\mathcal{M}$ be a tangent vector at $\mathbf{P}$ and $\gamma_\mathbf{S}(t)$ the unique geodesic such that $\gamma(0) = \mathbf{P}$ and $\gamma'(0) = \mathbf{S}$, then the exponential map is given as:

$$\mathrm{Exp}_\mathbf{P}(\mathbf{S}) = \gamma_\mathbf{S}(1) \tag{4.10}$$

Intuitively, the exponential map travels from $\mathbf{P}$ at a constant speed $\mathbf{S}$ along the *geodesic* for one unit of time, arriving at another point on $\mathcal{M}$. For the SPD manifold $\mathcal{P}(n)$ equipped with the Riemannian metric (4.9), its exponential map is defined as:

$$\mathrm{Exp}_\mathbf{P}(\mathbf{S}) = \mathbf{P_i} = \mathbf{P}^{\frac{1}{2}}\exp(\mathbf{P}^{-\frac{1}{2}}\mathbf{S}\mathbf{P}^{-\frac{1}{2}})\mathbf{P}^{\frac{1}{2}} \tag{4.11}$$

The logarithm operator, as the inverse of the exponential map, can project the point $\mathbf{P_i}$ on $\mathcal{M}$ to the tangent space $T_\mathbf{p}\mathcal{M}$ [189] and is defined as:

$$\mathrm{Log}_\mathbf{P}(\mathbf{P_i}) = \mathbf{S} = \mathbf{P}^{\frac{1}{2}}\log(\mathbf{P}^{-\frac{1}{2}}\mathbf{P_i}\mathbf{P}^{-\frac{1}{2}})\mathbf{P}^{\frac{1}{2}} \tag{4.12}$$

An illustration of these two operations is shown in Fig 4.2.

4.3.2.D Tangent Space Mapping

In current literature, few approaches have been suggested to tackle the non-linearity of the SPD manifold. A common method of dealing with this problem is to estimate the manifold-value data by mapping them into the tangent space of a specific point on the manifold (e.g., the mean of the data) [190]. To achieve this, we first need to construct the minimal representation on

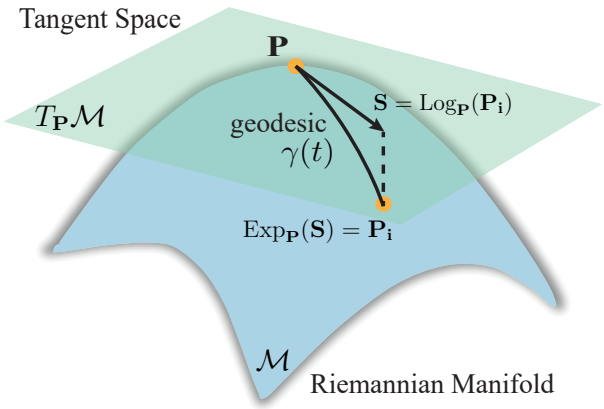


Figure 4.2: Exponential map and logarithm map between the Riemannian Manifold $\mathcal{M}$ and the tangent space $T_{\mathbf{p}}\mathcal{M}$ at $\mathbf{P}$.

the tangent space. Given the expression of the Riemannian metric, we can derive its equivalent definition as

$$\delta_R(\mathbf{P}, \mathbf{P_i}) = \|\log_{\mathbf{P}}(\mathbf{P_i})\|_{\mathbf{P}} = \|\mathbf{S}\|_{\mathbf{P}} = \|\mathbf{P}^{-\frac{1}{2}}\log_{\mathbf{P}}(\mathbf{P}_i)\mathbf{P}^{-\frac{1}{2}}\|_{\mathbf{I}} \tag{4.13}$$

where $\mathbf{I}$ is the identity matrix. Then the minimal representation can be created by the vector operation which maps a $n \times n$ matrix into a $1 \times n^2$ column vector. The coordinates of tangent vector $\mathbf{S} \in T_{\mathbf{p}}\mathcal{M}$ at $\mathbf{P}$ is defined as:

$$\mathrm{vec}_{\mathbf{P}}(\mathbf{S}) = \mathrm{vec}_{\mathbf{I}}(\mathbf{P}^{-\frac{1}{2}}\log_{\mathbf{P}}(\mathbf{P}_i)\mathbf{P}^{-\frac{1}{2}}) \tag{4.14}$$

Since the tangent space is the space of $n \times n$ symmetric matrices, there exist only $\frac{n\times(n+1)}{2}$ independent coefficients (i.e., upper/lower triangular part of the matrix) for the coordinate system. Thus, the vector operation is reformulated by:

$$\mathrm{vec}_{\mathbf{I}}(\mathbf{S}) = [s_{1,1}, \sqrt{2}s_{1,2}, \sqrt{2}s_{1,3}, \cdots, s_{2,2}, \sqrt{2}s_{2,3}, \cdots, s_{n,n}] \in \mathbb{R}^{\frac{n(n+1)}{2}} \tag{4.15}$$

The off-diagonal elements are multiplied by the weight of $\sqrt{2}$ as they are counted twice when calculating the norm at the identity [188]. Now, each point on the SPD manifold is projected into the tangent space through the use of logarithm map $\mathrm{Log}_{\mathbf{P}}(\cdot)$ which preserves the local structure of the manifold, and can be represented as an $n(n+1)/2$-dimensional vector for typical machine learning algorithms. In general, the projection $\mathcal{M} \rightarrow T_{\mathbf{P}}\mathcal{M}$ can occur at any point $\mathbf{P}$ on the manifold. However, in the work of [184], they reported that the best approximation can be

obtained using the mean of the data. Therefore, we employ the geometric mean [191] of a set of SPD matrices $\{\mathbf{P_1}, \mathbf{P_2}, \ldots, \mathbf{P_I}\}, I \geq 1, \mathbf{P_i} \in \mathcal{P}(n)$ to flatten the manifold

$$\mathbf{P}_\psi = \psi(\mathbf{P_1}, \mathbf{P_2}, \ldots, \mathbf{P_I}) = \underset{\mathbf{P} \in P(n)}{\arg\min} \sum_{i=1}^{I} \delta_R(\mathbf{P}, \mathbf{P_i})^2 \tag{4.16}$$

Since the SPD manifold is a manifold with non-positive cross-sectional curvature, the minimum of the problem defined in (4.16) exists and is unique [192]. An effective gradient descent algorithm is adopted for mean computation [193]. In the end, each SPD matrix $\mathbf{P_i}$ is converted into the Euclidean vector s_i for classification:

$$s_i = \mathrm{vec}_{\mathbf{I}}({\mathbf{P}_\psi}^{-\frac{1}{2}} \mathrm{Log}_{\mathbf{P}_\psi}(\mathbf{P_i}) {\mathbf{P}_\psi}^{-\frac{1}{2}}) \in \mathbb{R}^{\frac{n(n+1)}{2}} \tag{4.17}$$

4.3.3 Classification of SPD matrix sequences

Let us consider a set $\mathcal{L} = \bigcup_{q=1}^{Q} \mathcal{L}^q$ consisting of data from Q subjects. The qth subset $\mathcal{L}^q$ is represented by $\mathcal{L}^q = \{([\mathbf{P_i}, ..., \mathbf{P_{i+T}}], y_i), \mathbf{P_i} \in \mathcal{P}(n), i \in [1, |\mathcal{L}^q|]\}$,where $[\mathbf{P_i}, ..., \mathbf{P_{i+T}}]$ is a segmented sequence consisting of T subsequences, $\mathbf{P_i}$ can be considered as a representation of the correlation information for the corresponding subsequence living on the manifold, and y_i is the stress/pain label associated with the entire SPD matrix sequence, such that $y_i = f([\mathbf{P_i}, ..., \mathbf{P_{i+T}}])$. For each subset, its corresponding geometric mean ${\mathbf{P}_\psi}^q$ can be obtained with the equation (4.16). Then each SPD matrix $\mathbf{P_i}$ in the qth subset $\mathcal{L}^q$ is mapped into the tangent space and the derived corresponding subset of vector sequences is denoted by $s^q = \left\{([s_i, ..., s_{i+T}], y_i), s_i \in \mathbb{R}^{\frac{n(n+1)}{2}}, i \in [1, |s^q|]\right\}$ using the equation (4.17). The above process will be repeated for each subject's data. In the end, $\mathcal{L}^* = \bigcup_{q=1}^{Q} s^q$ is considered as the input of the LSTM-based deep neural network in Fig. 4.1. During training, the temporal contextual relationships within the tangent space vector sequence are explored by the 2-layer LSTM network, and the output features are then fed into the fully connected layer, followed by the sigmoid function to obtain the predicted probabilities. Algorithm 4.1 summarizes the process of classifying the SPD matrices.

Algorithm 4.1: Classification of SPD matrix sequences

Input:

A dataset $\mathcal{L}$ consisting of labeled SPD matrix sequence with T data segments collected from Q subjects $\mathcal{L} = \bigcup_{q=1}^{Q} \mathcal{L}^q = \bigcup_{q=1}^{Q} \{([\mathbf{P_i}, ..., \mathbf{P_{i+T}}], y_i), \mathbf{P_i} \in \mathcal{P}(n), i \in [1, |\mathcal{L}^q|]\}$.

Output:

$\hat{y}$ predicted labels of test set $\mathcal{L}^*_{te}$.

`/* Preparation of inputs to the classifier */`

$\mathcal{L}^* \leftarrow \varnothing$.

for each subject data $\mathcal{L}^q$ **do**

  Compute geometric mean of $\mathcal{L}^q$;

  $\mathbf{P}_\psi{}^q = \psi(\mathbf{P_i}, i = 1, \ldots, |\mathcal{L}^q|)$;

  Tangent Space Mapping of SPD matrices in $\mathcal{L}^q$;

  $s_i = vec((\mathbf{P}_\psi{}^q)^{-\frac{1}{2}} \mathrm{Log}_{\mathbf{P}_\psi{}^q}(\mathbf{P_i})(\mathbf{P}_\psi{}^q)^{-\frac{1}{2}})$

  $s^q = \left\{([s_i, ..., s_{i+T}], y_i), s_i \in \mathbb{R}^{\frac{n(n+1)}{2}}, i \in [1, |s^q|]\right\}$

  $\mathcal{L}^* \leftarrow \mathcal{L}^* \cup s^q$

end for

`/* LSTM-based classification */`

Split $\mathcal{L}^*$ into $\mathcal{L}^*_{tr}$ and $\mathcal{L}^*_{te}$ and Train LSTM using $\mathcal{L}^*_{tr}$;

$\hat{y} \leftarrow LSTM(\mathcal{L}^*_{te})$

4.4 Experiments and Results

To evaluate the validity of the proposed method, we conducted multimodal stress detection experiments on WESAD dataset while multimodal pain detection experiments were carried out on BP4D+ dataset.

4.4.1 Datasets

The **WESAD dataset** [53] is one of the most widely used public datasets for stress and affect recognition. In a restricted laboratory setting, multimodal data consisting of motion and physiological signals from 15 subjects were captured by two wearable devices, a wrist sensor and a chest sensor, respectively, and the experimental protocol was designed to stimulate three different emotional states (baseline, stress, amusement) in the participants. Based on previous work [15,53,69], a binary stress detection problem (*stress* vs. *non-stress*) can be formulated on the WESAD dataset by fusing baseline class and amusement class to form the *non-stress* class.

The **BP4D+ dataset** [194] is a large-scale multimodal spontaneous emotion database. 140 subjects were required to complete 10 tasks to elicit 10 different emotions, during which 2D RGB images, 3D model sequences, thermal videos and 8 physiological signal sequences with 1.5 million frames were captured by different sensors. In addition, the metadata are also provided, including 2D/3D/thermal facial landmarks, hand-labelled FACS codes and auto-tracked head poses. In this work, we focus on the identification of pain. As the dataset only provides the most facially-expressive segments for four emotions: happy, embarrassment, fear and pain, we therefore performed a pain detection task (*pain* vs. *non-pain*) by combining happy, embarrassment, fear as the *non-pain* class as proposed in [195]. An example of 2D texture images/3D model sequences/thermal images from the *Pain* class and their corresponding facial landmarks is shown in Fig. 4.3.

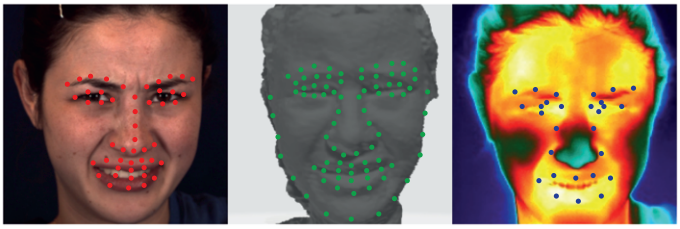


Figure 4.3: An example of 2D texture images/3D model sequences/thermal images from the *Pain* class and their corresponding facial landmarks provided in BP4D+ dataset.

4.4.2 Data Preprocessing and SPD matrix construction

For WESAD dataset, physiological and motion signals captured by the two sensors are filtered and downsampled to the same frequency, followed by a 10-second non-overlapping segmentation as proposed by [159]. Finally, the SPD matrix series can be calculated from multimodal data segments. For BP4D+ dataset, we first calculated the Euclidean distance between each of the two facial landmarks for each video frame using the provided 2D/3D/thermal facial landmarks, and automatically selected the 10 most discriminative distances by feature selection based on Anova F-value. Then the distance vector and the synchronized physiological signal vector were concatenated together to form a new augmented vector. Subsequently, all the augmented vectors are sliced into 1-second non-overlapping data segments. In the end, the obtained SPD matrix sequences can be extracted from prepared data segments. Fig. 4.4 shows an SPD matrix of a *pain* class sample in BP4D+ dataset that generated from the 10 facial distances between 2D landmarks and 8 physiological signals. During the pain task, the subject was asked to immerse hands in ice water, and her mouth was involuntarily opened, hence the distances

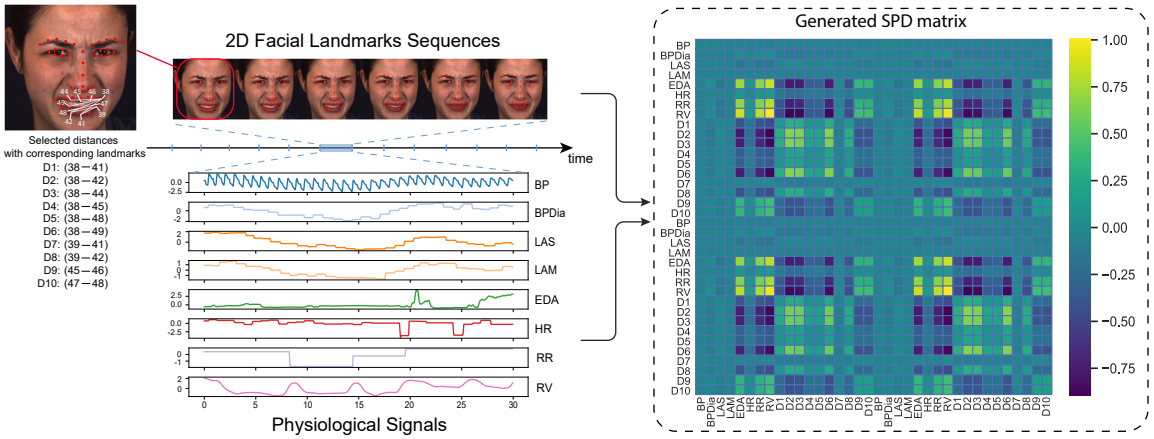


Figure 4.4: The multimodal SPD representation generated by a pain sample in the BP4D+ dataset, where correlations within and across two modalities (i.e., vision and physiology) can be observed. (D1-D10: 10 distances automatically selected based on Anova F-value, BP: raw blood pressure, BPDia: diastolic blood pressure, LAS: systolic blood pressure, LAM: mean blood pressure, EDA: electrodermal activity, HR: heart rate, RR: respiration rate, and RV: respiration volts.)

automatically selected were all based on the landmarks in the lip region, which is consistent with findings in the literature that lip movements such as oblique lip raising [196], horizontal lip stretching [197], etc. are related to pain. Both intra-modal and inter-modal correlations can be observed from the SPD matrix. Among all physiological signals, there was a strong correlation between electrodermal activity (EDA) and respiration signal (respiration rate (RR) and respiration volt (RV)) when the subject was suffering from pain. Among the selected distances, the 2nd, 3rd and 6th distances were more correlated with each other. Furthermore, the association between physiological and facial indicators can be explored, i.e. EDA, RR and RV were also positively correlated with the 9th and 10th distances.

4.4.3 Implementation and evaluation

All the classification models were implemented using Pytorch. To avoid overfitting, dropout operation was employed after the LSTM layers with a hidden state dimension of 128. The Adam optimizer with a learning rate lr=0.001 was selected to minimize the binary cross-entropy loss function during model training of 50 epoch. Decay coefficients of the first and second moment estimation β_1 and β_2 were set to 0.9 and 0.999, respectively. In the end, the proposed framework is evaluated using Leave-One-Subject-Out cross-validation (LOSO-CV) on WESAD dataset followed by [15, 53, 69] and Subject independent 10-fold cross-validation on BP4D+ dataset followed by [195] with two selected metrics: Accuracy and F1-score. Both experimental pro-

tocols assess the model's capacity to generalize across previously unseen subject data.

4.4.4 Stress detection results on WESAD

Binary stress detection experiments were performed using the wrist/chest-based data of all subjects in the WESAD dataset. We first verify the necessity of fusing the sample covariance matrix S and cross-covariance matrix C to form the proposed representation P. The evaluation results of these three representations are shown in Fig. 4.5 (a). It can be observed that the proposed SPD matrix optimizes the detection performance for all modality combinations, compared to those using only matrix S and matrix C. Secondly, to investigate the impact of increasing the dimensionality of the SPD matrix defined in the equation (4.5) on the detection results, we performed experiments using representations with different numbers of blocks of S and C (e.g., m=2 meaning that P consists of 2 blocks of S and 2 blocks of C). We found that using more blocks of S and C to compose the proposed representation P slightly enhanced the classification performance. This can be attributed to the increased proportion of cross-covariance information in the high-dimensional SPD matrix, further demonstrating the benefit of correlations between multiple modalities at different instants for the classification task. Besides, we only test up to m=4, since we noticed that sometimes the m=3 case performs best, following the trade-off between classification performance and computational cost. Finally, we also explored whether fusing data from different modalities, i.e., motion signals and physiological signals, could lead to a performance gain. The corresponding results are reported in Table 4.1. For experiments based on wrist/chest sensor data, combining these two types of data for detection yielded better performance, compared to results based on physiological signals only. When fusing data from both devices (wrist + chest), results using all modalities were not improved, which can be attributed to the redundant information generated by the same motion signals from both devices. When combining all physiological-based modalities, the highest performance (96.88% accuracy and 96.44% F1-score) was obtained with $P(m = 4)$.

Comparison with State-of-the-art To validate the effectiveness of the proposed method for fusing motion and physiological information, Table 4.2 shows the comparison results with 5 state-of-the-art methods using multimodal features. For a fair comparison, only methods that use the same experimental protocol were considered. In the work of Schmidt et al. [53], features from the time and frequency domains are used to train a variety of traditional machine learning classifiers, among which the Random Forest (RF) model achieved the best performance. Samyoun et al. [198] presented GAN/RNN-based deep model to generate gold standard chest sensor

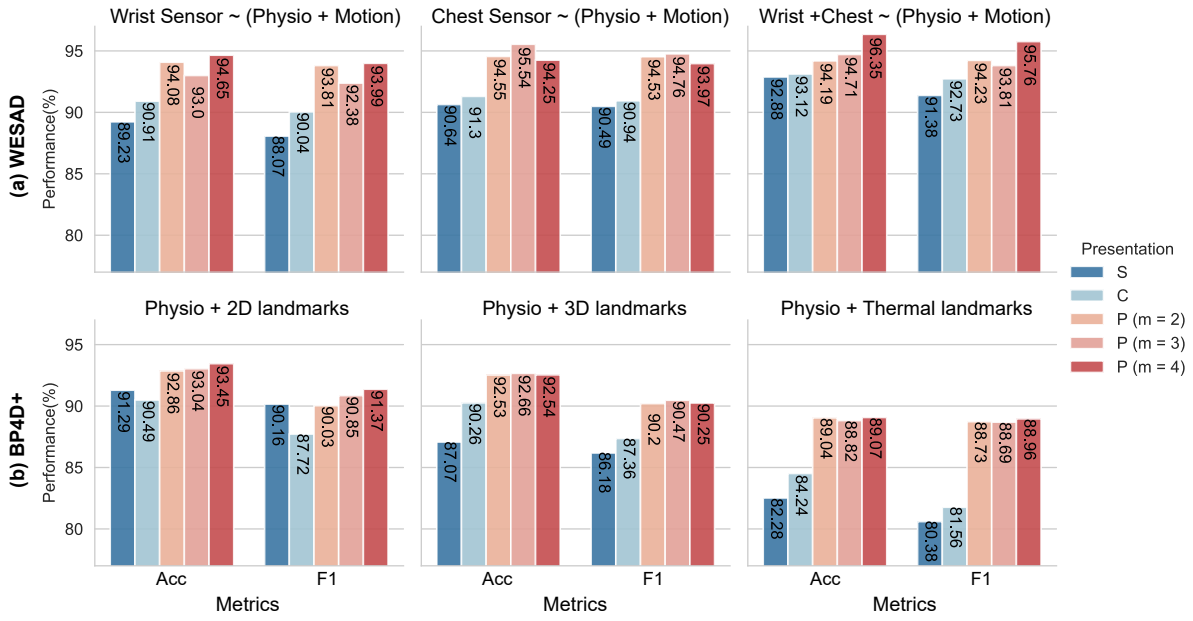


Figure 4.5: (a) Stress detection performance on WESAD dataset and (b) pain detection performance on BP4D+ dataset using the sample covariance matrix S, the cross-covariance matrix C and the proposed SPD representation P. (Acc: Accuracy, F1: F1-score)

Table 4.1: Stress detection performance of uni-modal setting (only physiology) and multi-modal setting (physiology + motion) on WESAD dataset using the proposed SPD representation P. (Acc: Accuracy, F1: F1-score, ↑ (↓): multimodal performance is improved (decreased) compared to the unimodal one.)

Sensor Type	Modalities	$\mathbf{P}(m=2)$		$\mathbf{P}(m=3)$		$\mathbf{P}(m=4)$	
		Acc	F1	Acc	F1	Acc	F1
Wrist	Physio	92.15	92.20	93.00	92.38	93.47	93.25
	Physio + Motion(↑)	**94.08**	**93.81**	**93.10**	**92.78**	**94.65**	**93.99**
Chest	Physio	92.76	92.61	92.96	92.64	92.12	91.75
	Physio + Motion(↑)	**94.55**	**94.33**	**95.54**	**94.76**	**94.25**	**93.97**
Wrist + Chest	**Physio**	**95.33**	**95.29**	**95.09**	**94.15**	**96.88**	**96.44**
	Physio + Motion(↓)	94.23	94.19	94.71	93.81	96.35	95.76

features from wrist data, and classified emulated features with various machine learning algorithms, among which RF (Random Forest) performed the best. Gil-Martin et al. [199] proposed a CNN-based architecture to extract meaningful features from three transforms: the Fourier transform (F), cube root (C), and constant q spectral transform (Q) of signal windows. Huynh et al. [69] suggested a training scheme using neural architecture search (StressNAS). Filter banks extracted from different modalities were utilised as model input, and the optimal model was selected automatically for each modality from 10,000 deep neural networks for training.

Finally, features of all modalities were concatenated for classification. Lai et al. [15] employed residual-temporal convolution network (Res-TCN) to process the filtered multimodal signals and proposed various fusion strategies. The above work for comparison simply spliced features from different modalities for prediction and thus ignored the cross-modality correlations. Overall, our proposed method using the joint SPD representation achieves the state-of-the-art results on wrist sensor data and competitive results on chest/wrist+chest sensor data, respectively, demonstrating its efficiency for integrating multimodal data.

Table 4.2: Comparison with State-of-the-art Methods on WESAD dataset (*Stress* vs. *Non-stress*)

Sensor Type	Methods	Accuracy	F1-score
Wrist	RF [53]	87.12	84.11
	GAN/RNN-RF [198]	92.1	89.7
	FCQ-CNN [199]	92.7	92.55
	StressNAS [69]	93.14	-
	Res-TCN [15]	94.16	93.62
	Ours with $\mathbf{P}(m=2)$	94.08	93.81
	Ours with $\mathbf{P}(m=3)$	93.10	92.78
	Ours with $\mathbf{P}(m=4)$	**94.65**	**93.99**
Chest	GAN/RNN-RF [198]	91.1	90.2
	RF [53]	92.83	91.07
	FCQ-CNN [199]	93.10	93.01
	Res-TCN [15]	**96.69**	**96.61**
	Ours with $\mathbf{P}(m=2)$	94.55	94.33
	Ours with $\mathbf{P}(m=3)$	<u>95.54</u>	<u>94.76</u>
	Ours with $\mathbf{P}(m=4)$	94.25	93.97
Wrist + Chest	RF [53]	92.28	90.74
	GAN/RNN-RF [198]	94.7	93.4
	FCQ-CNN [199]	96.62	96.63
	Res-TCN [15]	**97.75**	**97.74**
	Ours with $\mathbf{P}(m=2)$	94.23	94.19
	Ours with $\mathbf{P}(m=3)$	94.71	93.81
	Ours with $\mathbf{P}(m=4)$	<u>96.35</u>	<u>95.76</u>

4.4.5 Pain detection results on BP4D+

To further assess the validity of the proposed method, we conducted unimodal and multimodal pain detection experiments on the BP4D+ dataset. Similar to the process performed on WESAD, we first consider the results obtained using S or C alone as baseline to explore the im-

portance of combining them in the proposed representation. Based on the results in Fig. 4.5 (b), we reach the same conclusion that the joint SPD representation can improve classification performance. In addition, we can infer that increasing the dimension of P can further boost performance for all modalities. In the end, we also explored the performance of different modality combinations. The evaluation results are summarised in Table 4.3. We noticed that all the multimodal settings exhibit performance gains compared to unimodal detection results. Among four unimodality (i.e. physiological signal, 2D/3D/thermal facial landmarks), the trained model has the best performance using 2D facial landmarks where recognition accuracy and F1-score achieved 91.59%, 89.46% respectively. In the multimodal experiments, the best results with accuracy and F1-score of 93.45% and 91.37% can be observed with **2D + Physio** setting.

Table 4.3: Pain detection performance of uni-modal setting (only physiology) and multi-modal setting (physiology + vision) on BP4D+ dataset using the proposed SPD representation P. (2D/3D/Thermal: 2D/3D/Thermal facial landmarks, Physio: all physiological signals, Acc: Accuracy, F1: F1-score, ↑ (↓): multimodal performance is improved (decreased) compared to the unimodal one.)

Modalities	$\mathbf{P}(m=2)$		$\mathbf{P}(m=3)$		$\mathbf{P}(m=4)$	
	Acc	F1	Acc	F1	Acc	F1
Physio	81.82	81.72	83.24	82.42	82.81	81.42
Thermal Landmarks	82.82	82.70	83.18	82.93	83.53	83.37
3D Landmarks	91.01	88.84	91.13	89.04	91.27	89.30
2D Landmarks	91.44	89.24	91.59	89.46	90.94	88.56
Thermal + Physio(↑)	89.04	88.73	88.82	88.69	89.07	88.96
3D + Physio(↑)	92.53	90.20	92.66	90.47	92.54	90.25
2D + Physio(↑)	**92.86**	**90.03**	**93.04**	**90.85**	**93.45**	**91.37**

Comparison with State-of-the-art Table 4.4 shows the comparison results with 4 state-of-the-art methods using 2D facial landmarks and physiological signals. Due to the diversity of problem formulations and experimental settings, few pain detection efforts can be directly compared to our framework. Hinduja et al. [195] trained a random forest classifier on features extracted from physiological signals and facial action units (AUs) for pain detection. Here we only presented the comparison results based on physiological signals with them, as we did not use AUs for the detection. Our proposed method improves the accuracy by about 5.5%. Moreover, our framework achieved a more balanced pain detection with an F1-score of 82.42%. Since most pain detection datasets contain only vision-related information, very little pain recognition work has been carried out based on data from two different domains, i.e. vision and physiology. Therefore, to validate the effectiveness of our proposed approach on fused multimodal data,

Table 4.4: Comparison with State-of-the-art Methods on BP4D+ dataset (*pain* vs. *non-pain*).

Modality	Methods	Accuracy	F1-score
Physiology	RF [195]	77.7	30.0
	Ours with $\mathbf{P}(m=2)$	81.82	81.72
	Ours with $\mathbf{P}(m=3)$	**83.24**	**82.42**
	Ours with $\mathbf{P}(m=4)$	82.81	81.42
Vision	$\beta_G(t)$ on $\mathcal{S}^+(d,n)$ [200]	81.86	77.34
	Dual-layer 3DCNN [201]	84.03	83.98
	HybNet [202]	86.43	85.71
	Ours with $\mathbf{P}(m=2)$	91.44	89.24
	Ours with $\mathbf{P}(m=3)$	**91.59**	**89.46**
	Ours with $\mathbf{P}(m=4)$	90.94	88.56
Vision + Physiology	$^{(L)}$ $\beta_G(t)$ on $\mathcal{S}^+(d,n)$ [200]	82.77	76.32
	$^{(E)}$ $\beta_G(t)$ on $\mathcal{S}^+(d,n)$ [200]	84.32	78.83
	$^{(E)}$HybNet [202]	87.94	87.16
	$^{(L)}$Dual-layer 3DCNN [201]	89.08	88.68
	$^{(L)}$HybNet [202]	89.36	89.13
	$^{(E)}$Dual-layer 3DCNN [201]	89.80	89.46
	Ours with $\mathbf{P}(m=2)$	92.86	90.03
	Ours with $\mathbf{P}(m=3)$	93.04	90.85
	Ours with $\mathbf{P}(m=4)$	**93.45**	**91.37**

$^{(E)}$: Early fusion, $^{(L)}$: Late fusion.

state-of-the-art pain recognition methods that accept only visual data were implemented and combined with our physiological signal-based model for comparison. For a fair comparison, only the facial landmark-based methods were considered. We used the code provided by the authors, and if the code was not available, we followed the parameters provided in their article. Szczapa et al. [200] represented the facial landmark sequences as trajectories on the Riemannian manifold $\mathcal{S}^+(d,n)$. Each point of the trajectory is a Gram matrix computed from the 2D facial landmarks. Then the Global Alignment Kernel (GAK) was used to calculate the similarity matrix between the trajectories, which was used as feature for SVR-based (Support Vector Regression) pain estimation. To compare the classification performance, we replaced the SVR with an SVM (Support Vector Machine). Huang et al. [202] proposed a hybrid network (HybNet) which combined 1D, 2D and 3D CNN to extract geometric features from facial landmarks and spatio-temporal features for pain recognition. Choo et al. [201] employed a dual-layer 3D CNN for capturing the spatial-temporal features of the 2D facial landmark sequences. When comparing the performance of pain recognition based on solely visual information, our model performs better as shown in Table 4.4. To compare the performance based on multimodal information, we used two fusion techniques that are commonly used in the literature, feature level fusion and

decision level fusion, respectively. We first note that the performance of all the vision-based models used for comparison is improved when combined with our physiology-based model, providing side evidence that our model learns discriminative physiological features. Secondly, our model outperforms other multimodal approaches, both in terms of feature level fusion and decision level fusion, which confirms that the correlation between two modalities is well captured by the proposed method and that inter-modal communication can further contribute to the classification performance. Overall, our method achieves the state-of-the-art results on both unimodal data and multimodal data, validating again its effectiveness.

4.5 Conclusion

In this work, we explore for the first time the feasibility of SPD matrix-based representations for efficiently fusing physiological and behavioural signals, which can capture simultaneously correlation information within and across modalities. Tangent space mapping converts the generated SPD matrix time series into linear vector sequences for its application to the LSTM-based classification. The effectiveness of the proposed method was evaluated on public stress and pain detection datasets. In the end, the proposed framework shows state-of-the-art results on both stress and pain detection tasks. In the next chapter, our practical contribution will be presented, i.e. the real-world application of emotion recognition systems.

5

Emotion recognition in high-fidelity medical simulation training

Contents

This chapter mainly reports on practical contributions, where emotion recognition experiments in real-life scenarios (i.e., medical simulation training) are performed to explore its pedagogical implications. We first introduce what is medical simulation, why emotion recognition systems need to be integrated into teaching programs, and the corresponding challenges in Section 5.1. Subsequently, Section 5.2 details the data acquisition performed in the unconstrained environment. In Section 5.3, we conduct emotion recognition experiments on collected real-life data to verify the effectiveness of the proposed algorithms in Chapters 3 and 4. Finally, we demonstrate an intelligent graphical user interface (GUI) incorporating the proposed emotion recognition algorithms and explore the pedagogical impact of its intervention in Sections 5.4 and 5.5, respectively.

5.1 Introduction

Simulation is the artificial replication of complex real-world processes or events with sufficient fidelity that is designed to enhance the learning process through immersion, practice, reflection and feedback without encountering the risks that may be present in real-life environments [203]. In the medical field, the term medical simulation, as a subset of simulation, refers to modern educational methodologies for training healthcare professionals by recreating clinical situations. Its benefits are twofold: first, students can acquire and maintain clinical skills better in simulation training than in didactic teaching, considerably increasing self-confidence in practical situations, and second, simulation training can reduce the risk of harm to patients caused by inexperienced personnel during treatment [203]. Due to the preceding reasons, simulation has been used extensively as a teaching tool in medical schools. However, immersive learning in sufficiently realistic simulated scenarios tends to induce special emotional experiences in students, with stress being the relatively common emotional state. This stress can arises from a variety of factors or circumstances, for example, complex or high-risk tasks such as cardiac arrest, time constraints in performing tasks, unfamiliar procedures or equipment, or not receiving the desired feedback. Recently, there has been a tendency for teaching programmes to incorporate stress assessment sessions. The benefits are twofold: from the student's perspective, it reduces the negative impact of stressful states on acquisition of new skills, while from the teacher's perspective, stress detection allows training programmes to be adapted to individual needs and promotes teaching effectiveness. The traditional way of assessing emotions such as stress is subjective with the use of a series of psychological scales, however the process of collection and analysis is time-consuming and therefore not suitable for direct application in the pedagogical process.

With the rapid development of Artificial intelligence (AI) technology and sensor devices, intelligent detection algorithms are being developed considerably to provide a solution for the objective identification of students' emotions, especially stress states. However, current stress detection studies are mainly carried out in laboratory settings. In addition, very few researches can be found that have studied the impact of the intervention of emotion recognition tools on medical simulation training. To address these issues, we collected multimodal data from participants in the high-fidelity medical simulation environment and examined the validity of the emotion recognition algorithms proposed in Chapter 3 and 4 on real-life data. Subsequently, a software that integrates the recognition algorithms, i.e. a graphical user interface, was developed and applied during formation to explore the pedagogical impact of the emotion recognition tool. To summarise, the main contributions of this work are as follows:

- We collected a multimodal dataset consisting of peripheral physiological and motion signals via a wearable device in a real-world setting.
- We conducted emotion recognition experiments on real-life data and demonstrated the effectiveness of the previously proposed algorithms.
- We designed and developed a graphical interface that can be equipped with the recognition algorithms, and investigated its feasibility and usefulness as an emotion recognition tool in the simulation training process.

5.2 Data Collection

5.2.1 Subjects

The data involved in this chapter were collected during simulation sessions at the PRESAGE centre over the academic years 2020-2022. The subjects were 28 students in their first year of psychiatry internship or second year of healthcare education. All of these students were enrolled at University of Lille and had never participated in the simulation sessions.

5.2.2 Sensors and Multimodal signals

For emotion-related data collection, a medical-grade wearable device, the *Empatica E4* wristband was applied to monitor students' physiological and motion signal data unobtrusively during the simulation training. The E4 wristband is equipped with sensors designed to gather

high-quality data. The following multimodal signals were collected at different frequencies: 3-axis Accelerometer (ACC, 32Hz), Blood Volume Pressure (BVP, 64Hz), Electrodermal Activity (EDA, 4Hz), Skin Temperature (TEMP, 4Hz), Inter-beat Interval (IBI), Heart Rate (HR). The IBI signal and HR signal are derived from the BVP signal using the E4 wristband's built-in algorithm, which represents the time interval between two consecutive BVP peaks and the number of beats per minute, respectively. This research mainly focuses on ACC, BVP, EDA, and TEMP signals for exploring students' emotional status, particularly stress, during the simulation. We did not use the IBI and HR signals provided by the E4 wristband since its BVP signal analysis algorithm is not robust to the subject's movement, resulting in a large number of missing values in the signal recordings. This is consistent with the findings in [204]. Ultimately, all subjects were required to wear the E4 wristband on their non-dominant hand for data acquisition. Fig. 5.1 illustrates the E4 wristband and its acquired multimodal signals.

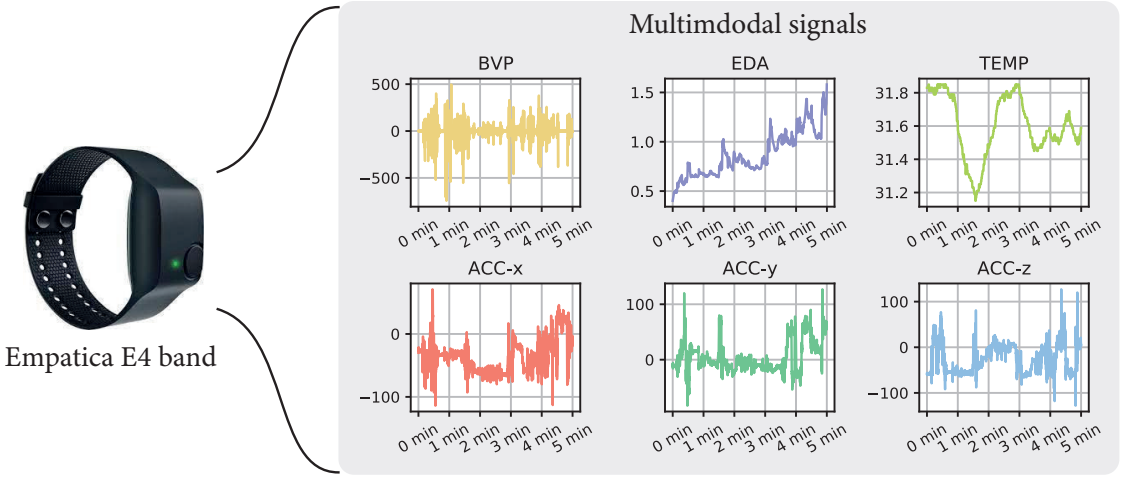


Figure 5.1: The wearable device used in this study: E4 wristband and collected raw multimodal signals.

5.2.3 Experimental protocol

The medical simulation training along with the emotional data collection was conducted at the PRESAGE Center, where each simulation session consisted of three phases: pre-simulation phase, simulation training phase, and educational debriefing phase. Fig. 5.2 demonstrates the entire experimental procedure. Details of each phase will be disclosed below.

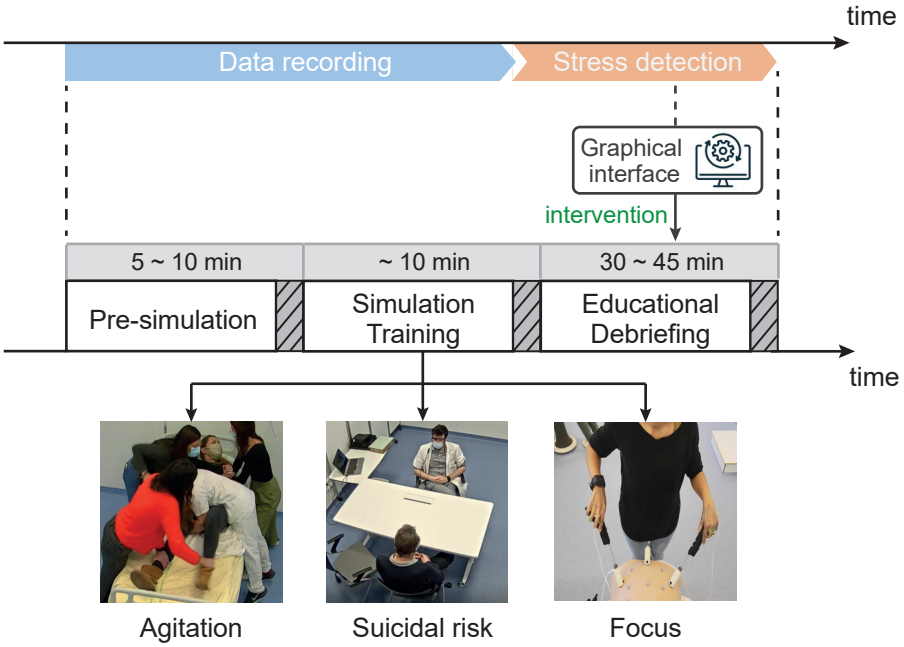


Figure 5.2: Experimental protocol. The grey part with grid lines represents the subject's self-evaluation, which was further used as ground truth for emotion recognition task.

Pre-simulation phase

Before entering the simulation room, the trainer first introduced the simulation scenario and the teaching objectives to the trainees, i.e. the participants of the experiment. Then, participants received an information letter, an image rights form, a consent form describing the entire experiment, and a set of psychometric scales. Meanwhile, a staff assisted participants in wearing the E4 bracelet and perform a short device test to ensure the quality of the data collected. Upon completion and signature of all relevant documents, participants were requested to press the tag button on their E4 wristbands to mark the start time of the simulation training and then enter the simulation room to perform the appropriate medical tasks. The above process lasted about 5 to 10 minutes

Simulation training phase

Multimodal data was collected from three different simulation scenarios: *Agitation*, *Suicidal risk*, *Focus* for the analysis of the trainee's emotional state. In each scene, highly realistic mannequins or hired actors, took on the role of patients, while trainees acted as doctors. The following are short descriptions of these scenarios:

- ***Agitation***: the scenario simulated a patient who was apparently in an acute agitated state,

showing psychotic symptoms and trying to break out of supervision. trainees were assigned the task of acting reasonably to prevent the patient's escape.

- ***Suicidal risk***: the scenario simulated consultation with a patient who was motivated by suicidal ideation, in which a trainee was required to make appropriate decisions based on the assessed level of urgency and risk of suicide.

- ***Focus***: the scenario required students to pass various tests on a surgical simulator. To provoke stress states, they were told that the obtained score on the tests would be recorded in the final grade of the semester.

Once the simulation was complete, the trainees were also required to press the tag button on the E4 wristband to mark the end time. The same psychometric scales as in the first phase were collected from the trainees to annotate their emotional experiences during the simulation training. Each simulation session lasted about 10 minutes.

Educational debriefing phase

Finally, all participants were invited to a collective debriefing with the actors who played patients to review the simulation training process and discuss the trainees' experience, the difficulties they encountered, and potential solutions for improvement. The debriefing session was led by a senior physician with teaching experience. In this session, the physician will apply the emotion detection tool developed (i.e., a graphical interface with embedded AI algorithms) to optimize the teaching process. At the end of the debriefing, trainees were invited to only fill in questionnaires concerning their perceived level of competence and the difficulty of the task, which were then used to study the intervention effects of the emotion detection tool.

5.2.4 Ground truth collection and evaluation

To establish the ground truth of multimodal data for emotion recognition, we collected trainees' subjective emotional experiences before/during/after the simulation training through three types of questionnaires: Self-Assessment Manikins (SAM), State-Trait Anxiety Inventory (STAI) and Likert scale of perceived competence and task difficulty. The collection and analysis of the subjects' self-reports were carried out in collaboration with Thibaut Denis, a graduate student from the Lille Neuroscience and Cognition Research Center. In the following, we describe these questionnaires in detail and present the statistical analysis of the collected ground truth data.

Psychometric scales

- ***Self-Assessment Manikins***: Self-Assessment Manikins (SAM) [6] is a picture-oriented emotion assessment tool, allowing subjects to rate the level of four dimensions: arousal, valence, liking, and dominance associated with emotional responses to diverse stimuli. Each scale is rated from 1 to 9, with 1 indicating low values for each dimension and 9 indicating high values for each dimension. In our study, we concentrated on the arousal and valence dimensions. The illustration of SAM is shown in Fig. 5.3. We only collected this scale after the first and second phases.

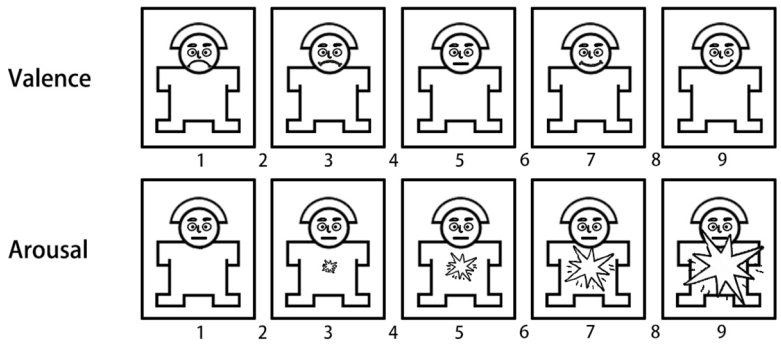
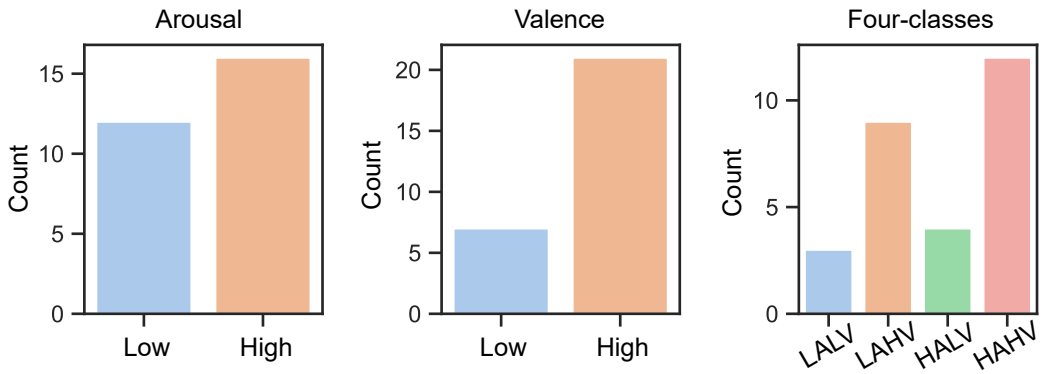


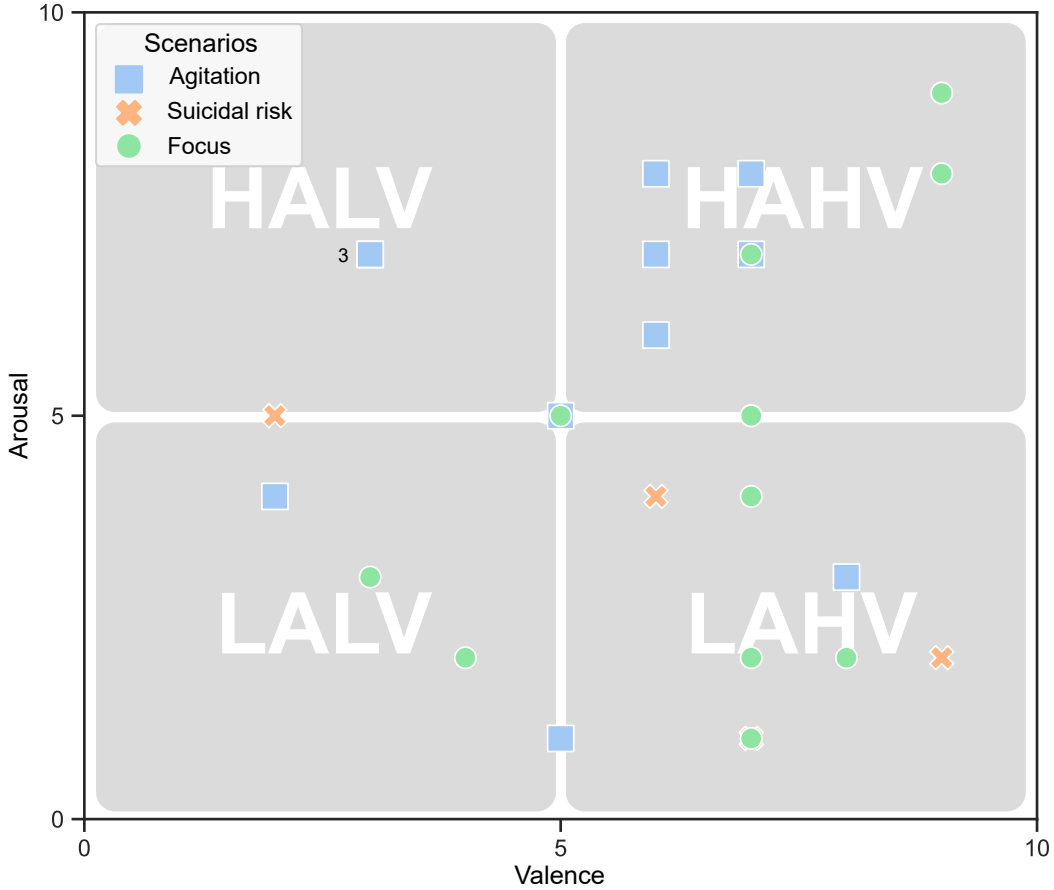
Figure 5.3: Self-Assessment Manikins. Excerpt from [6].

- ***State-Trait Anxiety Inventory***: The State-Trait Anxiety Inventory (STAI) [205] is a simple test to measure anxiety levels which provides insight into feelings of apprehension, tension, nervousness and worry. It consists of a series of four-point Likert scales (1. not at all, 2. somewhat, 3. moderately so, 4. very much so), the corresponding results of which can be considered as an approximation of a person's acute stress state. We only collected this scale after the first and second phases.

- ***Likert scale on the level of perceived competence and difficulty of the task***: stress can be considered as the result of an interaction between the learner and his or her cognitive appraisal, in accordance with Lazarus and Folkman's model [206]. Thus, each trainee in the simulation was repeatedly presented with a measure of his or her perceived level of competence in managing the simulation (primary assessment) and a measure of their perception of the difficulty of the task (secondary evaluation). Based on this, a likert scale of nine items was proposed. In the end, we collected this scale after all the phases.



a) Distribution of subjective ratings on A-V dimensions for all scenarios



b) Distribution of subjective ratings on A-V space for different scenarios

Figure 5.4: Analysis of Self-Assessment Manikin (SAM) data.

Figure 5.5: Analysis of State-Trait Anxiety Inventory (STAI) data before and after simulation training. (The green triangle symbol represents the mean value.)

Analysis of ground truth data

We analyzed the emotional experience of the 28 trainees during the simulation training based on the obtained subjective assessment data. First, we presented descriptive statistics on arousal and valence dimensions according to SAM scores in Fig. 5.4. The categorical representations can be deduced from the continuous scores of arousal and valence ranging from 1 to 9, i.e. scores less than 5 are classified as *Low* arousal/valence, while those greater than or equal to 5 belong to *High* arousal/valence. When combining the arousal and valence dimensions, it implies a 4-class representation: *LALV*, *LAHV*, *HALV*, *HAHV*. Fig. 5.4.a shows the categorical distribution of the SAM scoring for all scenarios. In general, the *Low* and *High* samples did not differ significantly in arousal dimensions, while for valence, the vast majority of trainees had positive emotional experiences. For the binary dimensions, *HAHV* and *LAHV* occupy the top two positions, representing excited and calm states, respectively, while a small portion of the sample corresponds to the *HALV* and *LALV* categories, implying stressful and boring emotional experiences. We further investigated the effect of different simulation scenarios on the emotional state of the trainees (Fig. 5.4.b). We observed that the *agitation* scenario differed from the other two scenarios in that the samples were mainly concentrated in the high arousal area and a considerable portion of the samples were distributed in subareas with negative valence. This may be related to the difficulty of the tasks performed by the subjects in this scenario, which required them to cope with the unexpected behavior of the agitated patients and was often accompanied by relatively violent physical conflicts in this case. In contrast, in the *focus* and *suicidal risk* scenarios, subjects were exposed to a more calm environment, especially the *focus* scenario involved only human-machine interaction, thus most of the samples presented with low arousal and positive emotions. Second, to examine the impact of simulation training on trainees' perceived stress from a holistic perspective, we studied samples of STAI scales before

and after the simulation. Fig. 5.5 demonstrates the data distribution of STAI scores collected in different scenarios. For the *agitation* and *suicidal risk* scenarios, the average stress level of the trainees showed modest increases at the end of the simulation. The Wilcoxon signed-rank tests showed a significant difference between pre- and post-simulation scores in the *agitation* scenario (*p-value*= 0.029), whereas there was no significant difference in the *suicidal risk* (*p-value*=0.375) and *focus* scenario (*p-value*=0.141), this finding is consistent with the analysis in arousal-valence space.

5.3 Emotion recognition experiments with real-life data

In Chapter 3 and 4, we validate the proposed emotion recognition methods based on self-supervised learning and supervised deep learning on data collected in a laboratory setting, respectively. We will continue to explore their performance in unrestricted real-life scenarios in this section.

5.3.1 Data preprocessing

First, the same low-pass filter as in Chapter 3 was used to remove the EDA, BVP and TEMP signal artefacts. In addition, a Butterworth low-pass filter with a cut-off frequency of 5 Hz was used to denoise the ACC signal. Then, z-score normalization was applied to eliminate cross-subject variation in multimodal signals. Finally, all signal modalities were sampled uniformly to 4Hz and segmented into windows of 60 seconds with 90% overlap for emotion classification. For arousal and valence scores varying from 1 to 9, a threshold of 5 was chosen to generate the labels of the samples for the emotion recognition task, where scores less than 5 were classified as *Low* and vice versa as *High*, resulting in two binary classification problems and a quadratic classification problem (i.e., *LALV*, *LAHV*, *HALV*, *HAHV*) on the arousal and valence dimensions. In addition, stress recognition is a task of particular interest, where *HALV* states can be categorised as *Stress* and other categories in the quadratic classification can be regarded as *Non-stress*. An overview of the pre-processed real-life dataset is shown in Table 5.1.

5.3.2 Methods for comparision

The following methods were implemented for performance comparison:

- ***Handcrafted-SVM***: This is a traditional supervised machine learning-based approach where a series of handcrafted features are extracted from the multimodal data as input

Table 5.1: A summary of pre-processed real-life data.

Task	Category (number of samples)	Total
Arousal-2	low (3417), high (4857)	8274
Valence-2	low(2591), high (5683)	
Stress-2	non-stress (6687), stress (1587)	
Four-class	LALV(1004), LAHV(2413) HALV(1587), HAHV(3270)	

Table 5.2: List of multimodal features for machine learning-based emotion recognition.

Signal	Features
ACC	mean, std of each axis/summation of all axis absolute integral of each axis/summation of all axis peak frequency of each axis
BVP	std of intervals between heartbeats (SDNN) std of successive differences between adjacent R-R intervals (SDSD) rms of successive differences between adjacent R-R intervals (RMSSD) proportion of differences greater than 50ms / 20ms (PNN50/20) frequency power in 0.15 - 0.40Hz (HF) and the 0.04 - 0.15Hz (LF)
EDA	Tonic: mean, std, 20th percentile, 80th percentile, quartile deviation Phasic: peaks prominence, peaks width, peaks per 100 s, strong peaks (peaks that are more than 1 μSiemens) per 100s
TEMP	mean, std, min, max, slope

to the SVM classifier. For ACC and TEMP signals, multiple statistical features presented in [53] were calculated directly from the pre-processed signal data. The EDA signal was first decomposed into two basic components: tonic and phasic using the cvxEDA method proposed by Greco et al. [207], from which the corresponding time-domain features were then extracted separately. For the BVP signal, we first detected the peaks to recover the IBI signal and then calculated a series of HRV features for the recognition task. The features employed are listed in Table 5.2.

- ***SPD-LSTM***: The multimodal fusion approach based on Riemannian geometry of the SPD matrices proposed in Chapter 4. An LSTM-based model was used to extract and classify high-level features from the linearised manifold data. For the comparison experiments, we set the dimension of the SPD representation m to 3 and the other parameters to the same settings as described in Chapter 4.

- ***TCN-TRANS***: The self-supervised learning approach based on 1D temporal convolution (TCN) and multimodal transformer (TRANS) proposed in Chapter 3. The experimental

parameters were configured as described in Chapter 3. For the comparison experiments, the frozen setting was applied, i.e. the encoder part of the pre-trained model was frozen in the downstream emotion recognition tasks.

5.3.3 Evaluation metric and protocol

To validate the generalizability of the model on different subject data, we adopted Leave-One-Subject-Out cross-validation protocol that is frequently used in real-life scenarios. Two common metrics: Accuracy and F1-score were selected for the performance evaluation.

5.3.4 Experimental Results

We performed the four emotion-related classification tasks mentioned in Section 6.3.1 on the collected data and investigated the performance of three modality combinations:

- Behaviour: all behavioral modalities, i.e., 3-axis accelerometer.
- Physiology: all physiological modalities, i.e., EDA, BVP and TEMP.
- Behaviour+Physiology: all modalities. i.e., ACC, EDA, BVP and TEMP.

The corresponding results are summarised in Table 5.3. The three methods used for comparison: *Handcrafted-SVM*, *SPD-LSTM*, *TCN-TRANS* in the table correspond to supervised machine learning, supervised deep learning and self-supervised learning, respectively. From the classification results, we can first conclude, in line with the work presented in Chapter 3 and Chapter 4, that both of the latter two methods can significantly enhance the performance of the algorithms based on handcrafted features. Furthermore, SSL learning-based methods exhibit the best performance, which can be attributed to the generalized representation learned from a large amount of unlabelled data. With regard to the different modality combinations, from a unimodal perspective, physiological modalities show superior performance to motion modalities in a majority of learning tasks. From a multimodal perspective, the two proposed methods both demonstrate the effectiveness of their fusion strategy on multimodal data, whereas the *Handcrafted-SVM* models with concatenated multimodal features as input fail to consistently achieve improved performance compared to unimodal ones.

Table 5.3: Performance comparison of different emotion recognition tasks with state-of-the-art methods on the Presage dataset. (Acc: Accuracy, F1: F1-score; B:Behaviour, P:Physiology, B+P:Behaviour+Physiology.)

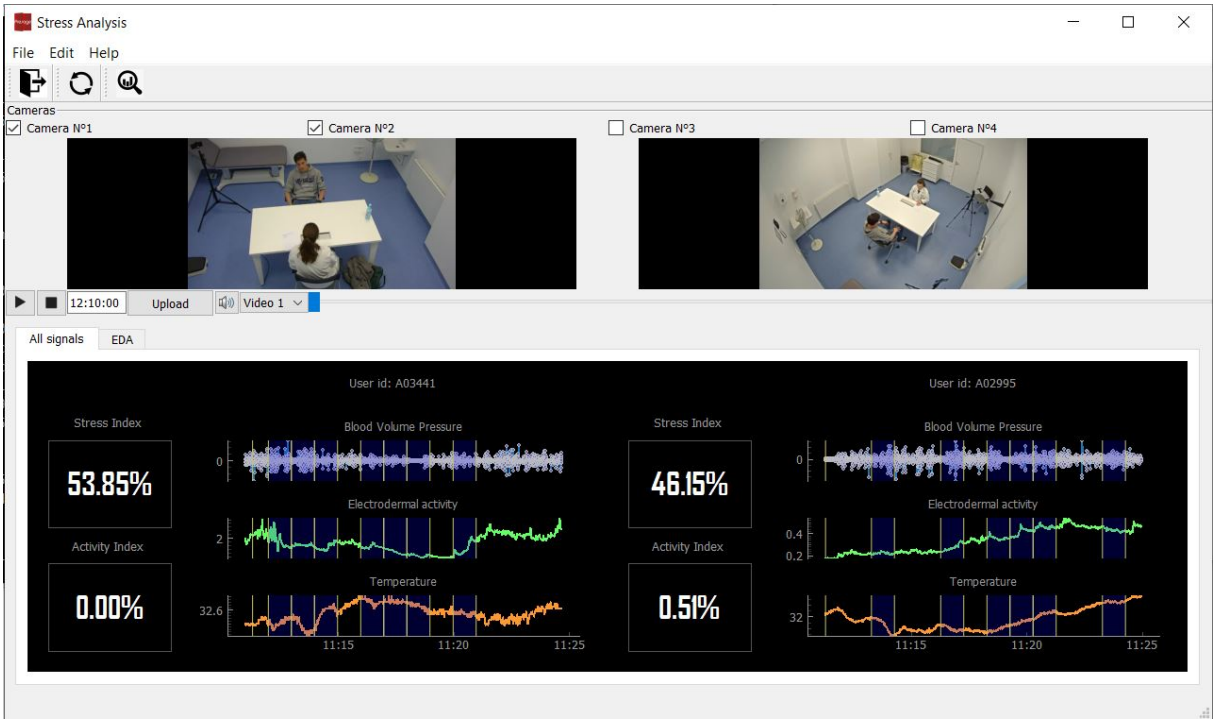
Type	Methods	Arousal-2		Valence-2		Stress-2		4-class	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
B	Handcrafted-SVM	57.39	52.09	63.07	59.62	64.41	61.83	44.55	41.56
	SPD-LSTM	72.15	70.45	75.26	71.07	76.57	72.55	61.39	54.06
	TCN-TRANS	**74.54**	**71.79**	**77.92**	**74.43**	**79.31**	**75.47**	**63.47**	**58.75**
P	Handcrafted-SVM	54.35	49.80	65.82	61.13	66.15	62.44	48.27	43.93
	SPD-LSTM	71.44	69.82	77.27	73.57	80.42	75.17	66.08	61.57
	TCN-TRANS	**73.62**	**71.64**	**80.27**	**75.61**	**83.59**	**79.42**	**67.29**	**63.11**
B+P	Handcrafted-SVM (↓)	55.68	50.76	66.87	65.25	61.94	59.93	46.86	41.96
	SPD-LSTM (↑)	73.69	71.81	78.52	75.72	82.66	78.70	67.21	62.06
	TCN-TRANS (↑)	**76.96**	**74.24**	**82.14**	**78.06**	**86.81**	**83.58**	**70.98**	**65.05**

5.4 Pratical application of stress analysis tool

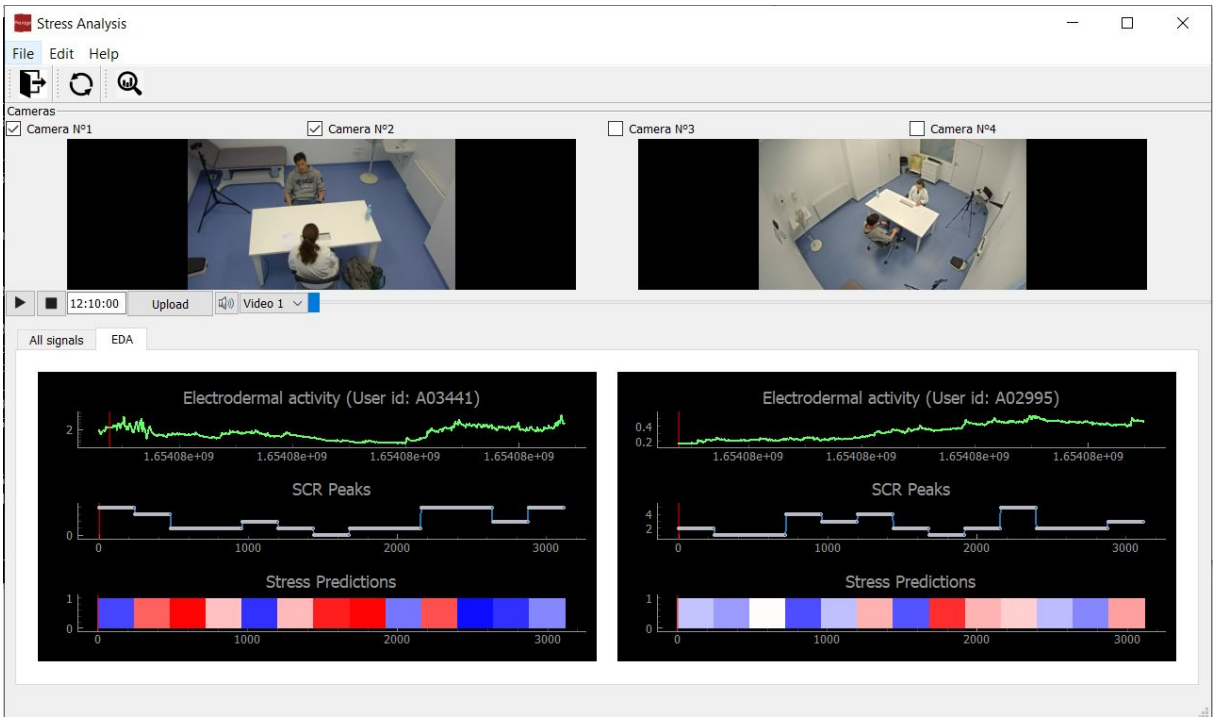
In the previous sections, we described the process of collecting real-life data on which we validated the effectiveness of the emotion recognition methods proposed in Chapter 3 and Chapter 4. Next, in order to achieve the main objective of the project, i.e. to verify whether the intervention of emotion detection algorithms, and in particular the stress detection algorithm, could improve the debriefing phase thus facilitating the learning of the trainees, a graphical interface was developed and applied as a stress analysis tool to the debriefing process. In the following, we will show the details of the tool and the evaluation results of its application in realistic scenarios in Sections 5.4.1 and 5.4.1 respectively.

5.4.1 Development of the stress analysis tool

Fig. 5.6 (a) shows a graphical interface developed based on PyQt, allowing to load, synchronize and visualize the trainees' multimodal data collected in the simulation rooms. The graphical interface consists of two parts, the upper part of which allows the display of video recordings from the different cameras in the simulation room, while the lower part shows the physiological signals of the trainees during the entire simulation training, and the parts of the signal curves framed in blue are the 60 s stress segments detected by the SSL emotion recognition algorithm proposed in Chapter 3. Besides, the areas circled by rectangles show the percentage of stress periods during simulation training and the physical activity index (AI) calculated proposed in [208], which represents the intensity of the motion obtained from the 3-axis accelerometry every 60 s. During the development of the interface, we designed a simplified version based



(a) The initial graphical interface



(b) Simplified graphical interface

Figure 5.6: The graphical interface with the built-in SSL-based stress detection algorithm applied in debriefing phase after simulation training.

on user feedback as shown in Fig. 5.6 (b). At the debriefing phase, the teacher expects to quickly locate and review the most stressful segments of the trainee and provide instruction, thus the probabilities of stress in every 60 s represented by a diverging colormap with values ranging from [0,1] is introduced, where blue and red are biased towards the *non-stress* and *stress* states, white corresponds to a neutral state (i.e., a probability of 0.5). In addition, all the physiological signals initially proposed were reduced to the EDA signal for simplification since it is considered to be one of the best real-time relevant indicators of stress [40].

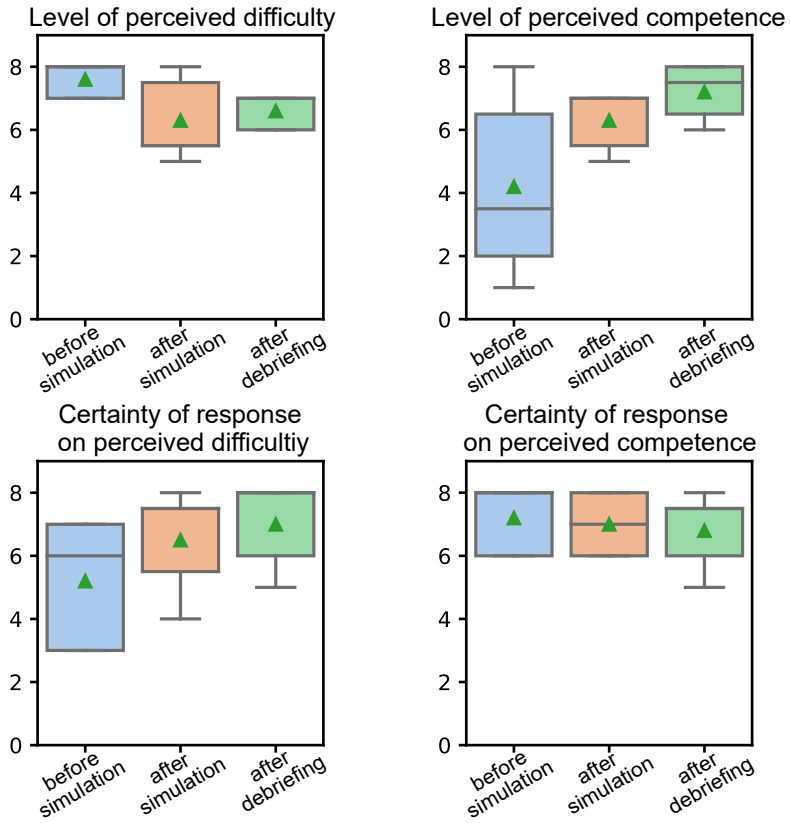


Figure 5.7: Visualization of the subjective assessment of the task difficulty and the level of competence at three moments: pre-simulation, post-simulation and post-debriefing.

5.4.2 Evaluation of the stress analysis tool

The pedagogical relevance and feasibility of the stress detection method were validated in a simulation training based on *suicidal risk* scenario in June 2022. To explore the impact of the application of the stress analysis tool on the debriefing phase, particularly in terms of trainees' competence development, subjective variables were first monitored in different phases of the

simulation training. As mentioned in Section 5.2.4, a 9-item Likert scale was distributed to the trainees at three moments: pre-simulation, post-simulation and post-debriefing, to measure their perception of task difficulty, as well as their level of ability to manage the task. In addition, a metacognitive dimension of certainty about the answers was added. Fig. 5.7 presents a visualization of the subjective assessment. We can observe a correspondence between a potential decrease in perceived difficulty and an increase in perceived competence, while their corresponding levels of certainty of answers remain either increasing or basically invariant. This is in line with expectations, as the trainers can target the stressful moments of the trainees during the simulation with the aid of the developed tool, which allows them to help the students to better solve their problems and thus promote their personal competence. Secondly, we collected feedback from the trainees on their usage of the stress analysis tool through a 7-item Likert scale. The feedback from trainees generally gave positive feedback on the role of the interface in debriefing and its ability to identify key moments (Fig. 5.8). Moreover, in addition to the the discussion of typical open questions during the debriefing process, the use of the simplified version (i.e. the colormap of stress detection probabilities) made it possible for trainees to rapidly initiate a certain degree of reflective processing based on long-duration video material. Therefore, the incorporation of the trainee's stress detection in the debriefing session can also enhance their participation and engagement. Furthermore, when reviewing the detected stressful moments with the trainees, we noticed that these segments were usually accompanied by emotional verbal expressions. This further affirms the effectiveness of our proposed algorithm and also inspires the possibility of incorporating audio data in future work to provide more accurate detection.

5.5 Conclusion

In this chapter, we investigated the application of an emotion recognition system in a real-life situation, i.e., medical simulation training. Students' multimodal signals were recorded during the simulation via a wearable sensor, while a series of psychometric scales were applied to establish the ground truth of emotion. Subsequently, emotion recognition experiments involving arousal, valence and stress were conducted on the collected data, proving again the validity of the recognition algorithms presented in Chapters 3 and 4. In order to explore the feasibility and usefulness of the intervention of emotion recognition in the pedagogical process, a graphical interface incorporating the emotion recognition algorithm was developed and used in the post-simulation debriefing session. Based on feedback from debriefing participants, the positive effects of the emotion recognition tool in educational practice were initially confirmed,

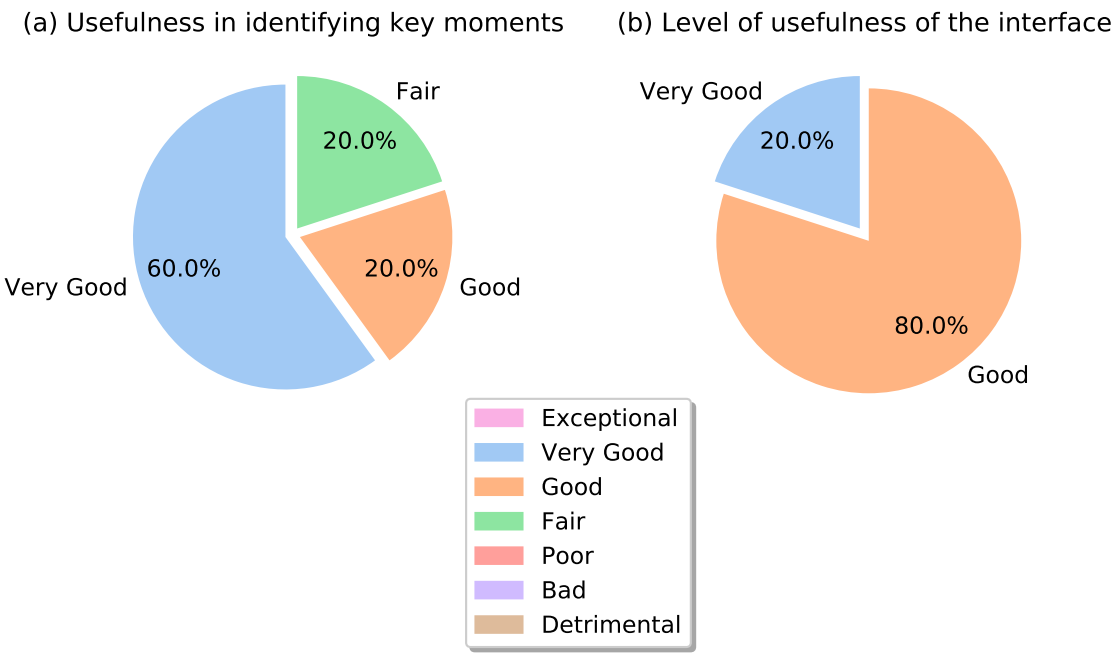


Figure 5.8: Visualisation of trainees' feedback regarding the application of the stress analysis tool.

including the facilitation of skill acquisition and increased engagement. In the next chapter we will show the technical contribution on the behavioural aspect, i.e., the landmark-based facial expression recognition approach.

6

Metric Learning on Complex Projective Spaces

Contents

The work presented in this chapter corresponds to the technical contributions on the behavioral aspect, where a non-euclidean metric learning method with application to facial expression recognition is proposed. In Section 6.1 we briefly introduce landmark-based shape analysis and its limitations, followed by Section 6.2 in which we present metric learning methods applied to facial expression recognition. In Section 6.3 we describe in detail the shape representation in the complex projective space, the Fubini-Study metric and its optimization algorithm, in Section 6.4 we focus on facial expression recognition experiments and explore the effectiveness of the proposed method, and finally Section 6.5 presents a summary of the work.

6.1 Introduction

Recently landmark detection and tracking methods for human faces and bodies became reliable and accurate. This has greatly encouraged landmark-based shape analysis methods, aiming at shape description and comparison through the location of a pre-defined set of points and the association between them. Since a set of landmarks detected in a video frame or still image is a natural choice for modelling the facial shape or body shape, such methods are widely deployed in the field of computer vision and multimedia such as face verification [209], person re-identification [210], facial expression recognition [211–213] and action recognition [214]. Depending on the type of shape space, landmark-based methods can be divided into Euclidean methods and Riemannian methods. The latter are based on the geometry of Riemannian manifolds and are more robust to variation factors in the data, such as affine transformations, thus gaining broader interest compared to the Euclidean methods. In the work of [127], the face and body landmark sequences were parametrized as trajectories on the Riemannian manifold of symmetric positive definite (SPD) matrices of fixed-rank. Daoudi et al. [215] used the Gram matrix of 2D landmarks as the representation of the body shape on a non-linear manifold to assess the depression severity of a patient. Szczapa et al. [216] encoded the facial movement as trajectories on the manifold of symmetric positive-semidefinite (PSD) matrices for pain intensity estimation. The above approaches always involve metrics defined on the manifold to provide similarity measures for the learning task at hand. However, these Riemannian methods typically employ pre-defined metrics, which may not be the most appropriate for distance measures on a given dataset. In response to this problem, a large number of metric learning algorithms have been proposed in order to obtain more discriminative metrics to further facilitate learning performance. However, most conventional metric learning methods for these vision tasks are designed for the optimization of metrics in the Euclidean space and are therefore not applicable to manifolds.

Considering the above issues, we propose a metric learning method suited to Riemannian geometry. First, we treated the equivalence class of complex landmarks as points in the complex projective space whose shape representation is invariant under a class of affine transformations that consists of translations, rotations, and scaling. Based on the seminal work on shape analysis of Kendall [217], a family of metrics - *Fubini-Study metrics* in this space was selected for distance measurement, where metric learning algorithm based on Large Margin Nearest Neighbors (LMNN) is exploited to improve the discriminative power of the proposed metric. Finally, the learned metric was employed in a similarity-based classification task and its validity was verified in the facial expression recognition scenario. Fig. 6.1 demonstrates the overview of the proposed method. Overall, the main contribution of the paper are:

- We proposed a landmark-based affine-invariant shape representation in complex projective space on the basis of Riemannian geometry.
- We adapted the LMNN algorithm to learn the optimal metric, and the resulting similarity features have smaller intra-class differences and larger inter-class differences.
- We performed a comparison between the Fubini-Study metric and the Euclidean metric on the facial expression recognition task, validating the effectiveness of the proposed metric for shape analysis. Furthermore, our approach also showed competitiveness in comparison with state-of-the-art solutions.

6.2 Related work

As presented in Section 2.4.4, most Riemann methods for facial expression recognition tasks rely on full-face feature support, i.e. treating full-face landmark points as a whole and deriving robust shape representations from them. However, such methods ignore the possibility of optimizing the used metric to enhance learning performance. Due to the need for suitable metrics to measure the similarity between a set of landmarks, a series of metric learning-based approaches have been proposed for facial expression recognition tasks. The purpose of metric learning is to find an adequate metric to better capture the potential similarity relationship hidden in the data. The optimal metric obtained from metric learning can bring samples from the same class as close as possible while keeping the differently labeled samples far from each other. In the following, we focus on reporting on previous work that adopted this learning model for facial shape analysis.

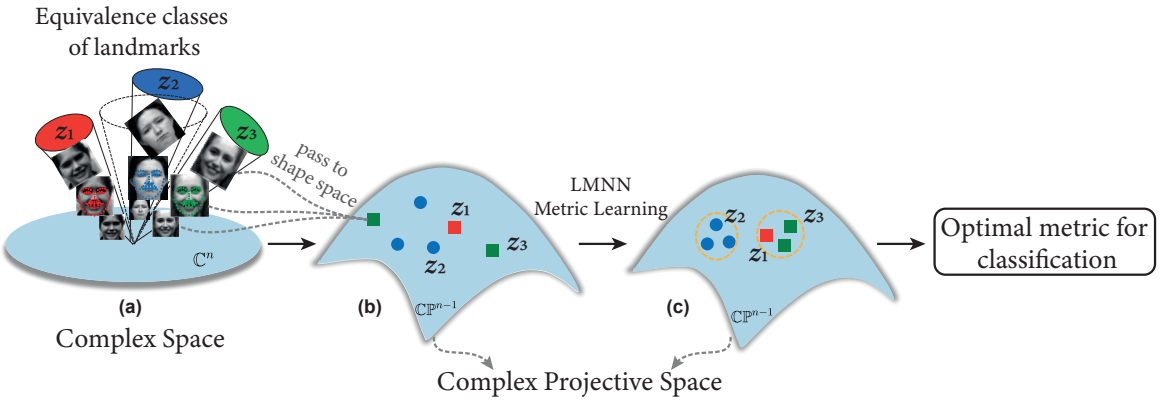


Figure 6.1: Overview of the proposed approach. (a) 2D landmarks detected in images of facial expressions are first converted to vectors in complex space; (b) The variants obtained by affine transformation constitute the equivalence class of landmarks, which can be considered as the same point in the complex projective space; (c) The application of metric learning enables the reduction of intra-class differences while enlarging inter-class distances. Finally, the learned metric is used for similarity-based classification.

6.2.1 Euclidean metric learning

In the work of Wan et al. [213], the facial shape representation was defined as the combination of the shape vector represented by the facial points and the appearance feature vector. Mahalanobis distance-based Euclidean metric learning algorithms was employed to obtain the optimal metric, where similar shapes are relatively closer in the learned feature space. Experimental results demonstrated that the KNN-based classifier with the learned metric achieved superior spontaneous facial expression recognition performance than the methods using the original metric and exhibited good generalization ability on the spontaneous expressions. Kacem et al. [212] exploited a chart of barycentric coordinates to map affine equivalence classes of facial landmarks to the Euclidean space and then applied metric learning to the Mahalanobis distance for the facial expression recognition task. The DTW technique combined with the optimal metric was used to align facial shape sequences and constructed similarity-based inputs for the pairwise proximity function SVM (ppfSVM) classifier. In the end, the optimal metric yields better recognition accuracy, confirming the necessity of metric learning. However, their method relies on the stability of a reference triangle, which is used to obtain an affine-invariant shape representation for metric learning.

6.2.2 Non-Euclidean metric learning

Some non-linear methods were also proposed to learn a more discriminant metric, which can encode the intrinsic geometry of the manifold. Daoudi et al. [211] propose a metric learning over a family of metrics on the space of oriented ellipses centered at the origin in Euclidean n-space and the double cover of the manifold of n $\times$ n positive semi-definite (PSD) matrices of rank two. They first adopted the Gram matrices as the facial shape representations. The Average Neighborhood Margin Maximization (ANMM) algorithm was then generalized to learn the optimal quotient Riemannian metric on the manifold data. The proximity measures calculated between facial shapes using the obtained metric were fed into the SVM classifier for static facial expression recognition, obtaining improved performance over that before metric learning.

6.3 Proposed Method

6.3.1 Complex projective space

A 2D landmark configuration $\mathbf{z}$ consists of n ordered points (x_1, y_1), ..., (x_n, y_n) on the plane. Expressing the points as complex numbers (i.e., writing $z_j := x_j + iy_j$ instead of (x_j, y_j)), we identify $\mathbf{z}$ as a complex vector $(z_1, \ldots, z_n) \in \mathbb{C}^n$. We will consider two landmark configurations $\mathbf{z}$ and $\mathbf{w}$ to be *equivalent* if the points w_j $(1 \leq j \leq n)$ in the second configuration are obtained from the points z_j in the first by means of a common translation, rotation, and scaling. This equivalence is established by the fact that the three transformations only alter the position, orientation or size of the shapes accordingly, without destroying their semantic information. Therefore, the set of equivalence classes of 2D landmark configurations is the *shape space* in which we wish to work. This shape space can be considered as a generalization of the original space of landmark configurations, enabling a more realistic measure of the similarity/dissimilarity between shapes, as differences caused by transformations are excluded. Since the landmark points are represented in complex form, the equivalence of two landmark configurations $\mathbf{z}$ and $\mathbf{w}$ translates into the existence of a nonzero complex number a and a complex number v so that $w_j = az_j + v$, $\forall j \in \{1, \ldots, n\}$, where $a = \lambda e^{i\theta}$ controls the variation factors of scaling and rotation while v regulates translation. If we consider only *centered* configurations where $z_1 + \cdots + z_n = 0$, then two centered configurations $\mathbf{z}$ and $\mathbf{w}$ are equivalent if and only if there exists a nonzero complex number a for which $\mathbf{w} = a\mathbf{z}$. This is precisely the definition of *complex projective space* (i.e.,the desired shape space).

Definition 1. *Complex projective space of (complex) dimension n, $\mathbb{CP}^n$, is the set of equivalence*

classes of nonzero vectors in $\mathbb{C}^{n+1}$ *with the equivalence relation*

$$(z_1, \ldots, z_n + 1) \sim (az_1, \ldots, az_n + 1)$$

for any nonzero complex number a.

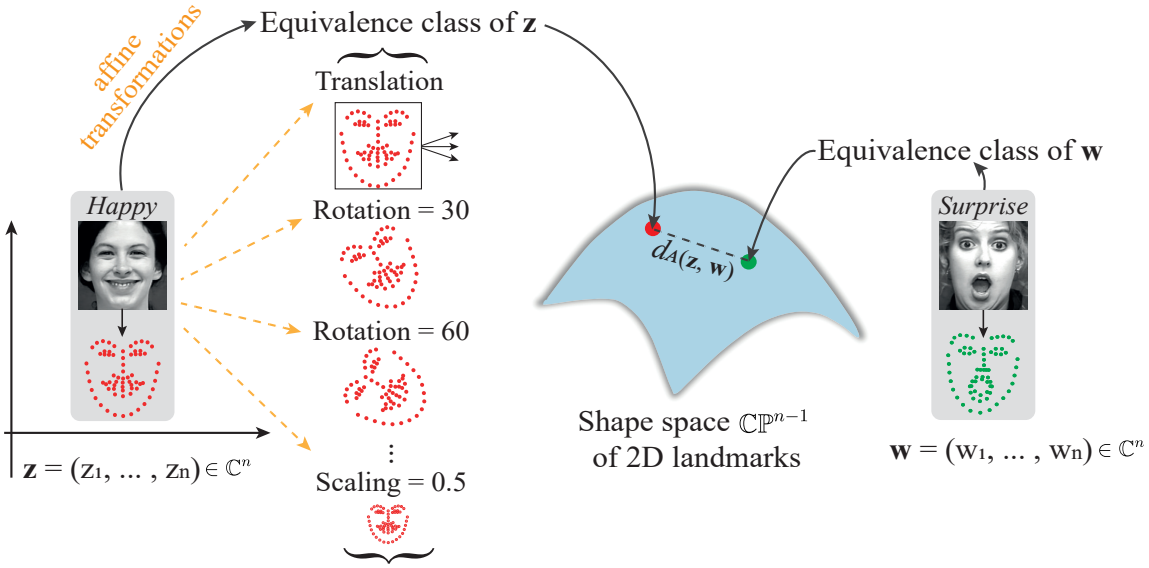


Figure 6.2: An illustration of 2D facial landmark configurations and its corresponding shape space in the facial expression recognition scenario. Matching and comparison of facial shapes in the complex projective space is not disturbed by traditional affine transformations: translation, rotation and scaling.

A centered landmark $\mathbf{z}$ with coordinates $(z_1, \ldots, z_n)$ gives rise to an equivalence class $[\mathbf{z}]$ which can also be described by its *homogeneous coordinates* $[z_1 : z_2 : \ldots : z_n]$. The equivalence class $[\mathbf{z}]$ is a point in $\mathbb{CP}^{n-1}$, which, moreover, lies in the projective hyperplane given by the homogeneous equation $z_1 + \cdots + z_n = 0$. Fig. 6.2 illustrates the complex projective space constructed in a facial expression recognition scene, where changes in facial landmarks due to translation, rotation or scaling do not modify the emotional state conveyed by the facial expressions. Hence, the altered landmark configurations can be all regarded as the same point in the complex projective space, providing a better measure of the similarity between different facial expressions.

6.3.2 Fubini-Study Metrics

A useful feature of complex projective space is that it carries a family of simply-defined and well-studied metrics: let A be a positive-definite Hermitian matrix, the *Fubini-Study distance*, d_A, associated to A between points $\mathbf{z} = [z_1 : \ldots : z_n]$ and $\mathbf{w} = [w_1 : \ldots : w_n]$ in $\mathbb{CP}^{n-1}$ is defined by

$$\frac{|\langle A\mathbf{z}, \mathbf{w}\rangle|^2}{\langle A\mathbf{z}, \mathbf{z}\rangle\langle A\mathbf{w}, \mathbf{w}\rangle} = \cos^2(d_A(\mathbf{z}, \mathbf{w})) \tag{6.1}$$

We recall that the expression $\langle \mathbf{z}, \mathbf{w}\rangle = z_1\bar{w}_1 + \cdots z_n\bar{w}_n$ is the standard complex-valued inner product on $\mathbb{C}^n$.

6.3.3 Euclidean Metrics

Given a positive-definite Hermitian matrix, its associated Fubini-Study metric allows us to define the distance between two equivalence classes of landmark configurations. However, there exists also a simple alternative. Assume that for all of our landmark configurations, the i-th landmark is different from the origin. For each landmark configuration $\mathbf{z}$, instead of considering the point in $\mathbb{CP}^{n-1}$ given by $\mathbf{z} = [z_1 : \ldots : z_n]$, we can consider the vector

$$\mathbf{z}_i := \left(\frac{z_1}{z_i}, \ldots, \frac{z_{i-1}}{z_i}, \frac{z_{i+1}}{z_i}, \ldots, \frac{z_n}{z_i}\right) \in \mathbb{C}^{n-1}.$$

Note that any two landmark configurations $\mathbf{z}$ and $\mathbf{z}'$ are equivalent if and only if $\mathbf{z}_i = \mathbf{z}'_i$. In this way, our shape space, or rather the piece of it that interests us, becomes the linear space $\mathbb{C}^{n-1} = \mathbb{R}^{2n-2}$. This enables us to use different Euclidean metrics to measure the distance between equivalence classes of landmark configurations.

Recall that each Euclidean metric in $\mathbb{R}^{2n-2}$ is determined by a positive-definite $(2n-2) \times (2n-2)$ real matrix M and that the associated distance is given by the formula

$$d_M(\mathbf{x}_i, \mathbf{y}_i) = \sqrt{(\mathbf{x}_i - \mathbf{y}_i)^T M(\mathbf{x}_i - \mathbf{y}_i)} \tag{6.2}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ are coordinates in $\mathbb{C}^{n-1} = \mathbb{R}^{2n-2}$ associated to the equivalence classes $[\mathbf{x}]$ and $[\mathbf{y}]$ in $\mathbb{CP}^{n-1}$.

6.3.4 Distance function for Fubini-Study metrics

Recall that given a positive-definite Hermitian matrix A, the *Fubini-Study distance* associated to A between points $\mathbf{z} = [z_1 : \ldots : z_n]$ and $\mathbf{w} = [w_1 : \ldots : w_n]$ in $\mathbb{CP}^{n-1}$ is defined by

$$\frac{|\langle A\mathbf{z}, \mathbf{w}\rangle|^2}{\langle A\mathbf{z}, \mathbf{z}\rangle\langle A\mathbf{w}, \mathbf{w}\rangle} = \cos^2(d_A(\mathbf{z}, \mathbf{w})).$$

Before performing metric learning using this family of distances, we must consider a cost function that is a real-valued function on the space of positive-definite $n \times n$ Hermitian matrices:

$$C : \mathcal{H}_n \longrightarrow \mathbb{R}.$$

Usually, the cost function takes the following form: there exist points $\mathbf{z}_1, \ldots, \mathbf{z}_k$ in $\mathbb{CP}^{n-1}$ and a real-valued smooth function:

$$c : \mathbb{R}_+^{\binom{k}{2}} \longrightarrow \mathbb{R}$$

such that for any positive-definite $n \times n$ Hermitian matrices, the value of $C(A)$ only depends on the mutual distances between the points $\mathbf{z}_1, \ldots, \mathbf{z}_k$:

$$C(A) = c(d_A(\mathbf{z}_1, \mathbf{z}_2), \ldots, d_A(\mathbf{z}_{k-1}, \mathbf{z}_k)).$$

Sometimes it is easier not to explicitly describe the dependence of the cost function C in terms of the distances $d_A(\mathbf{z}_i, \mathbf{z}_j)$, but in terms of the square or other function f of these distances:

$$C(A) = c(f(d_A(\mathbf{z}_1, \mathbf{z}_2)), \ldots, f(d_A(\mathbf{z}_{k-1}, \mathbf{z}_k))).$$

Since we need to find an optimal metric with the aid of the cost function C, i.e., to find the best positive-definite Hermitian matrix A that assigns small distances to similar points, while assigning relatively large distances to dissimilar points, the function $f : \mathbb{R}_+ \to \mathbb{R}_+$ should be smooth, strictly increasing and satisfy $f(0) = 0$. Given the formula for the Fubini-Study distance, it will be convenient to work with the function $f(t) = 1 - \cos^2(t)$ as part of the cost function for metric learning. This function is smooth, satisfies $f(0) = 0$, and it is strictly increasing in the interval $[0, \frac{\pi}{2}]$. Notice that $\frac{\pi}{2}$ is precisely the diameter of $\mathbb{CP}^{n-1}$ with any of the Fubini-Study metrics, so that whatever happens to f for $t > \frac{\pi}{2}$ is irrelevant. As we are particularly interested on the dependency of the positive-definite Hermitian matrix A, the

distance function is represented by the term $F(A)$

$$F(A) = 1 - \cos^2(d_A(\mathbf{z}_i, \mathbf{z}_j)) = 1 - \frac{|\langle A\mathbf{z}_i, \mathbf{z}_j\rangle|^2}{\langle A\mathbf{z}_i, \mathbf{z}_i\rangle\langle A\mathbf{z}_j, \mathbf{z}_j\rangle} \tag{6.3}$$

for a number of chosen landmark configurations $\mathbf{z}_1, \ldots, \mathbf{z}_k$ in $\mathbb{CP}^{n-1}$. To summarize, we will take the cost function C(A) to be a function of $F(A)$. In order to obtain the optimal metric through the cost function, the gradient of $F(A)$ with respect to the variable A can be first given by the formula

$$\begin{aligned}\nabla_A F(A) = \frac{-1}{\left(\langle A\mathbf{z}_i, \mathbf{z}_i\rangle\langle A\mathbf{z}_j, \mathbf{z}_j\rangle\right)^2} \times \big[&\langle A\mathbf{z}_i, \mathbf{z}_i\rangle\langle A\mathbf{z}_j, \mathbf{z}_j\rangle(\langle A\mathbf{z}_j, \mathbf{z}_i\rangle\mathbf{z}_i\mathbf{z}_j^* + \langle A\mathbf{z}_i, \mathbf{z}_j\rangle\mathbf{z}_j\mathbf{z}_i^*) \\ &- |\langle A\mathbf{z}_i, \mathbf{z}_j\rangle|^2 \left(\langle A\mathbf{z}_j, \mathbf{z}_j\rangle\mathbf{z}_i\mathbf{z}_i^* + \langle A\mathbf{z}_i, \mathbf{z}_i\rangle\mathbf{z}_j\mathbf{z}_j^*\right)\big]\end{aligned} \tag{6.4}$$

Proof. Recall that the computation of a gradient of a function F on an inner product space is a two-step process. We first compute the directional derivative of $F(A)$ at point A in the direction of X:

$$dF(A)(X) := \frac{dF(A + tX)}{dt}|_{t=0}, \tag{6.5}$$

where X is any positive-definite $n \times n$ Hermitian matrices. The resulting derivative is given as

$$\begin{aligned} dF(A)(X) &= -\frac{d}{dt}\frac{|\langle (A+tX)\mathbf{z}_i, \mathbf{z}_j\rangle|^2}{\langle (A+tX)\mathbf{z}_i, \mathbf{z}_i\rangle\langle (A+tX)\mathbf{z}_j, \mathbf{z}_j\rangle}|_{t=0} \\ &= -\frac{d}{dt}\frac{\langle (A+tX)\mathbf{z}_i, \mathbf{z}_j\rangle\overline{\langle (A+tX)\mathbf{z}_i, \mathbf{z}_j\rangle}}{\langle (A+tX)\mathbf{z}_i, \mathbf{z}_i\rangle\langle (A+tX)\mathbf{z}_j, \mathbf{z}_j\rangle}|_{t=0} \\ &= -\frac{d}{dt}\frac{\langle (A+tX)\mathbf{z}_i, \mathbf{z}_j\rangle\langle (A+tX)\mathbf{z}_j, \mathbf{z}_i\rangle}{\langle (A+tX)\mathbf{z}_i, \mathbf{z}_i\rangle\langle (A+tX)\mathbf{z}_j, \mathbf{z}_j\rangle}|_{t=0} \end{aligned}$$

For convenience, we can re-write the above equation in the following form:

$$\begin{aligned} dF(A)(X) &= -\frac{d}{dt}\frac{f(t)}{g(t)}|_{t=0} \\ &= -\frac{f'(t)g(t) - f(t)g'(t)}{g(t)^2}|_{t=0} \end{aligned} \tag{6.6}$$

We can first calculate the derivative $f'(t)$ as follows:

$$\begin{aligned} f'(t) &= \frac{d}{dt}\langle (A+tX)\mathbf{z}_i, \mathbf{z}_j\rangle\langle (A+tX)\mathbf{z}_j, \mathbf{z}_i\rangle|_{t=0} \\ &= \frac{d}{dt}\langle A\mathbf{z}_i + tX\mathbf{z}_i, \mathbf{z}_j\rangle\langle A\mathbf{z}_j + tX\mathbf{z}_j, \mathbf{z}_i\rangle|_{t=0} \\ &= \langle X\mathbf{z}_i, \mathbf{z}_j\rangle\langle A\mathbf{z}_j, \mathbf{z}_i\rangle + \langle X\mathbf{z}_j, \mathbf{z}_i\rangle\langle A\mathbf{z}_i, \mathbf{z}_j\rangle \end{aligned}$$

We can say that $(\mathbf{z}+\overline{\mathbf{z}}) = 2\Re(\mathbf{z})$, thus $f'(t)$ can eventually be written as

$$f'(t) = 2\Re(\langle X\mathbf{z}_i, \mathbf{z}_j\rangle\langle A\mathbf{z}_j, \mathbf{z}_i\rangle) \tag{6.7}$$

Next, the derivative $g'(t)$ can be computed as follows:

$$\begin{aligned} g'(t) &= \frac{d}{dt}\langle (A+tX)\mathbf{z}_i, \mathbf{z}_i\rangle\langle (A+tX)\mathbf{z}_j, \mathbf{z}_j\rangle|_{t=0} \\ &= \frac{d}{dt}\langle A\mathbf{z}_i + tX\mathbf{z}_i, \mathbf{z}_i\rangle\langle A\mathbf{z}_j + tX\mathbf{z}_j, \mathbf{z}_j\rangle|_{t=0} \\ &= \langle X\mathbf{z}_i, \mathbf{z}_i\rangle\langle A\mathbf{z}_j, \mathbf{z}_j\rangle + \langle X\mathbf{z}_j, \mathbf{z}_j\rangle\langle A\mathbf{z}_i, \mathbf{z}_i\rangle \end{aligned} \tag{6.8}$$

Finally, we can bring equations (6.7) and (6.8) back into (6.6) to obtain the final form of $dF(A)(X)$:

$$\begin{aligned} dF(A)(X) &= -\frac{f'(t)g(t) - f(t)g'(t)}{g(t)^2}|_{t=0} \\ &= \frac{-1}{\left(\langle A\mathbf{z}_i, \mathbf{z}_i\rangle\langle A\mathbf{z}_j, \mathbf{z}_j\rangle\right)^2} \times \big[2\Re(\langle X\mathbf{z}_i, \mathbf{z}_j\rangle\langle A\mathbf{z}_j, \mathbf{z}_i\rangle)\,\langle A\mathbf{z}_i, \mathbf{z}_i\rangle\langle A\mathbf{z}_j, \mathbf{z}_j\rangle \\ &\quad - |\langle A\mathbf{z}_i, \mathbf{z}_j\rangle|^2\,(\langle X\mathbf{z}_i, \mathbf{z}_i\rangle\langle A\mathbf{z}_j, \mathbf{z}_j\rangle + \langle X\mathbf{z}_j, \mathbf{z}_j\rangle\langle A\mathbf{z}_i, \mathbf{z}_i\rangle)\big] \end{aligned}$$

With the directional derivative $dF(A)(X)$ defined in the above equation, we now need to find the unique vector field $\nabla F(A)$ (i.e., the gradient on the Riemannian manifold of Hermitian matrices) that satisfies:

$$\langle \nabla F(A), X\rangle = dF(A)(X) \tag{6.9}$$

Since the inner product of two Hermitian matrices is the trace of their product, the equation

(6.9) can be written in the following form:

$$\begin{aligned}\mathrm{tr}(\nabla F(A)X) = \frac{-1}{\left(\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle\right)^2} \times \big[&2\Re(\langle X\mathbf{z}_i,\mathbf{z}_j\rangle\langle A\mathbf{z}_j,\mathbf{z}_i\rangle)\,\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle \\ &- |\langle A\mathbf{z}_i,\mathbf{z}_j\rangle|^2\left(\langle X\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle + \langle X\mathbf{z}_j,\mathbf{z}_j\rangle\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\right)\big]\end{aligned}$$

With the trick $\langle X\mathbf{z}_i,\mathbf{z}_j\rangle = \mathrm{tr}(X\,\mathbf{z}_i\mathbf{z}_j^*)$ where $\mathbf{z}_i$ and $\mathbf{z}_j$ are considered as $n\times 1$ matrices, we can rewrite the equation as:

$$\begin{aligned}\mathrm{tr}(\nabla F(A)X) &= \frac{-1}{\left(\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle\right)^2} \times \big[2\Re\left(\mathrm{tr}(X\mathbf{z}_i\,\mathbf{z}_j^*)\langle A\mathbf{z}_j,\mathbf{z}_i\rangle\right)\,\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle \\ &\quad - |\langle A\mathbf{z}_i,\mathbf{z}_j\rangle|^2\left(\mathrm{tr}(X\mathbf{z}_i\,\mathbf{z}_i^*)\langle A\mathbf{z}_j,\mathbf{z}_j\rangle + \mathrm{tr}(X\mathbf{z}_j\,\mathbf{z}_j^*)\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\right)\big] \\ &= \frac{-1}{\left(\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle\right)^2} \times \big[\left(\mathrm{tr}(X\mathbf{z}_i\,\mathbf{z}_j^*)\langle A\mathbf{z}_j,\mathbf{z}_i\rangle + \mathrm{tr}(X\mathbf{z}_j\,\mathbf{z}_i^*)\langle A\mathbf{z}_i,\mathbf{z}_j\rangle\right) \\ &\quad \langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle - |\langle A\mathbf{z}_i,\mathbf{z}_j\rangle|^2\left(\mathrm{tr}(X\mathbf{z}_i\,\mathbf{z}_i^*)\langle A\mathbf{z}_j,\mathbf{z}_j\rangle + \mathrm{tr}(X\mathbf{z}_j\,\mathbf{z}_j^*)\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\right)\big] \\ &= \frac{-1}{\left(\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle\right)^2} \times \big[\left(\mathrm{tr}(X\mathbf{z}_i\,\mathbf{z}_j^*\langle A\mathbf{z}_j,\mathbf{z}_i\rangle) + \mathrm{tr}(X\mathbf{z}_j\,\mathbf{z}_i^*\langle A\mathbf{z}_i,\mathbf{z}_j\rangle)\right) \\ &\quad \langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle - |\langle A\mathbf{z}_i,\mathbf{z}_j\rangle|^2\left(\mathrm{tr}(X\mathbf{z}_i\,\mathbf{z}_i^*\langle A\mathbf{z}_j,\mathbf{z}_j\rangle) + \mathrm{tr}(X\mathbf{z}_j\,\mathbf{z}_j^*\langle A\mathbf{z}_i,\mathbf{z}_i\rangle)\right)\big] \\ &= \frac{-1}{\left(\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle\right)^2} \times \big[\left(\mathrm{tr}(X(\mathbf{z}_i\,\mathbf{z}_j^*\langle A\mathbf{z}_j,\mathbf{z}_i\rangle + \mathbf{z}_j\,\mathbf{z}_i^*\langle A\mathbf{z}_i,\mathbf{z}_j\rangle)\right) \\ &\quad \langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle - |\langle A\mathbf{z}_i,\mathbf{z}_j\rangle|^2\left(\mathrm{tr}(X(\mathbf{z}_i\,\mathbf{z}_i^*\langle A\mathbf{z}_j,\mathbf{z}_j\rangle + \mathbf{z}_j\,\mathbf{z}_j^*\langle A\mathbf{z}_i,\mathbf{z}_i\rangle)\right)\big] \\ &= \mathrm{tr}(X((\frac{-1}{\left(\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle\right)^2} \times \big[\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle(\langle A\mathbf{z}_j,\mathbf{z}_i\rangle\mathbf{z}_i\mathbf{z}_j^* + \langle A\mathbf{z}_i,\mathbf{z}_j\rangle\mathbf{z}_j\mathbf{z}_i^*) \\ &\quad - |\langle A\mathbf{z}_i,\mathbf{z}_j\rangle|^2\left(\langle A\mathbf{z}_j,\mathbf{z}_j\rangle\mathbf{z}_i\mathbf{z}_i^* + \langle A\mathbf{z}_i,\mathbf{z}_i\rangle\mathbf{z}_j\mathbf{z}_j^*\right)\big]))\end{aligned}$$

In the end, the gradient $\nabla F(A)$ is given by:

$$\begin{aligned}\nabla_A F(A) = \frac{-1}{\left(\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle\right)^2} \times \big[&\langle A\mathbf{z}_i,\mathbf{z}_i\rangle\langle A\mathbf{z}_j,\mathbf{z}_j\rangle(\langle A\mathbf{z}_j,\mathbf{z}_i\rangle\mathbf{z}_i\mathbf{z}_j^* + \langle A\mathbf{z}_i,\mathbf{z}_j\rangle\mathbf{z}_j\mathbf{z}_i^*) \\ &- |\langle A\mathbf{z}_i,\mathbf{z}_j\rangle|^2\left(\langle A\mathbf{z}_j,\mathbf{z}_j\rangle\mathbf{z}_i\mathbf{z}_i^* + \langle A\mathbf{z}_i,\mathbf{z}_i\rangle\mathbf{z}_j\mathbf{z}_j^*\right)\big]\end{aligned} \tag{6.10}$$

which will be used for the optmization of the Fubini-Study metrics.

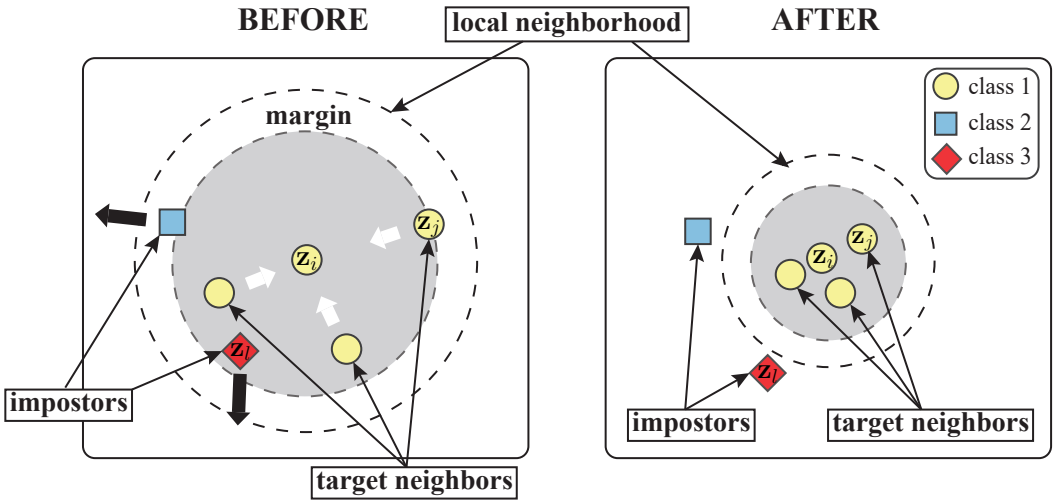


Figure 6.3: An illustration of the LMNN algorithm modified from [7]. The proposed Fubini-study metric will be optimised such that the distances between each sample $\mathbf{z}_i$ and its $k = 3$ target neighbours are reduced, while the distances to the impostors are increased.

6.3.5 Metric Learning with Fubini-Study metrics

Large Margin Nearest Neighbors (LMNN) [7] was selected as the metric learning algorithm. The purpose of this algorithm is to reduce the distance between each sample and its *target neighbors*, which are the k pre-selected nearest neighbor samples of the same class, while trying to keep it away from its *imposters*, which are differently labeled samples that invade the margin established by those target neighbor. An illustration of the LMNN algorithm is shown in Fig. 6.3. Assuming that the target neighbor set has been selected (the nearest neighbors of each sample are calculated by the Fubini-Study distance), the loss function of LMNN consists of two terms. The first term is the target neighbors pulling term, given by

$$\varepsilon_{pull}(A) = \sum_{j \rightsquigarrow i} F_{ij}(A) = 1 - cos^2(d_A(\mathbf{z}_i, \mathbf{z}_j)) \tag{6.11}$$

where $F_{ij}(A)$ is the distance function in Eq.(6.3) between complex-valued samples $\mathbf{z}_i$ and $\mathbf{z}_j$ corresponding to Hermitian matrix A. $j \rightsquigarrow i$ iff jth sample is a target neighbor of ith sample. The second term is the impostors pushing term, given by

$$\varepsilon_{push}(A) = \sum_{i} \sum_{j \rightsquigarrow i} \sum_{l} (1 - y_{il})[1 + F_{ij}(A) - F_{il}(A)]_+ \tag{6.12}$$

where $y_{il} = 0$ if ith sample and lth sample are differently labeled and 1 otherwise, the term $[\cdot]_+$ is defined as $[z]_+ = max\{z, 0\}$. Finally, the loss function is given by,

$$\varepsilon(A) = (1-\mu)\varepsilon_{pull}(A) + \mu\varepsilon_{push}(A), \mu \in [0,1] \tag{6.13}$$

where μ is the weighting parameter for balancing push and pull effects. The gradient of the loss function can be formulated as

$$\nabla_A(\varepsilon(A)) = (1-\mu)\nabla_A(\varepsilon_{pull}(A)) + \mu\nabla_A(\varepsilon_{push}(A)), \mu \in [0,1] \tag{6.14}$$

where,

$$\nabla_A(\varepsilon_{pull}(A)) = \sum_{j \rightsquigarrow i} \nabla_A F_{ij}(A) \tag{6.15}$$

and,

$$\nabla_A(\varepsilon_{push}(A)) = \sum_i \sum_{j \rightsquigarrow i} \sum_l (1-y_{il})[\nabla_A F_{ij}(A) - \nabla_A F_{il}(A)] \tag{6.16}$$

The Riemannian Steepest Descent algorithm in the toolbox Pymanopt [218] was implemented using the cost function and gradient in Eq.(6.13) and Eq.(6.14), respectively, to seek the optimal solution for positive-definite Hermitian matrix A. The main procedure for our metric learning on complex projective space is given in Algorithm 6.1.

Algorithm 6.1: Metric learning in Complex Projective Space $\mathbb{CP}^{n-1}$

Data: N training samples $\mathcal{Z} = \{(Z^i, y^i)\}_1^N$ with their associated labels, k is the number of *target neighbors*.

Result: The optimal positive-definite Hermitian matrix $A^\star$

$A \leftarrow I(n)$, identity complex matrix of dimension n. ;

for $i = 1 \ldots N$ **do**

  Define k *target neighbors* $\mathcal{N}_s^A(i)$ and their corresponding *impostors* $\mathcal{N}_o^A(i)$ for each sample Z^i using the Fubini-Study metric in Eq.(6.1) ;

end

$Cost \leftarrow \varepsilon(A, \mathcal{N}_s^A, \mathcal{N}_o^A)$, $\varepsilon(.)$ is given by Eq.(6.13);

$Grad \leftarrow \nabla_A(\varepsilon(A)), \nabla_A(.)$ is given by Eq.(6.14);

$A^\star \leftarrow SteepestDescent(A, Cost, Grad)$, $SteepestDescent(.)$ is the Riemannian Steepest Descent optimization algorithm in $\mathcal{H}_n$;

6.4 Experiments

To investigate the effectiveness of the proposed metric in the complex domain, we conducted several experiments for the facial expression recognition task. First, we introduced the CK+ and Oulu-CASIA datasets used in the experiments. All details of the experimental setup are then reported in Section 6.4.2. In the end, classification results using the proposed metric and comparative results with the Euclidean metric and other state-of-the-art approaches in facial expression recognition are presented in Section 6.4.3.

6.4.1 Datasets

The proposed non-Euclidean metric learning algorithm was evaluated on two public facial expression datatsets. The **Cohn-Kanade Extended (CK+) dataset** [8] consists of 327 annotated frontal video sequences performed by 118 subjects, wherein each subject is required to exhibit seven facial expressions during the experiment, namely – *anger*, *contempt*, *disgust*, *fear*, *happy*, *sad* and *surprise*. One sequence contains images from neutral expression (first frame) to peak expression (last frame). For the classification experiments, we exploit only the last frame, where the intensity of facial expression attains its peak. The **Oulu-CASIA dataset** [9] consists of frontal facial videos from 80 subjects captured under three illumination conditions: dark, normal and weak normal, during which the subjects were required to imitate six classic facial expressions: *anger*, *disgust*, *fear*, *happy*, *sad* and *surprise*. Similar to the CK+ dataset, the last frame of each sequence has the highest expression intensity. Therefore, we use the last frame of the 480 video sequences under normal illumination for the classification task. Randomly selected images of facial expressions from the two datasets are presented in Fig. 6.4.

6.4.2 Experimental setting

For each landmark configuration consisting of 68 points in CK+ and Oulu-CASIA datasets, we first excluded the 17 points of face contour. The 2D centered landmarks are then written in complex form to represent the facial shape $\mathbf{z} \in \mathbb{CP}^{51}$ which is invariant to translation, scaling and rotation. Inspired by previous work [211, 212], proximity data were constructed to serve as input to classifiers by using the optimal Fubini-study metric obtained from the LMNN algorithm, where each sample $\mathbf{z}_i$ is represented by its similarity with all samples $[d_A(\mathbf{z}_1, \mathbf{z}_i), \ldots, d_A(\mathbf{z}_k, \mathbf{z}_i)]^T$ in the dataset. In the end, K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) were selected as classifiers which are most frequently used in metric learning-based methods [175]. The hyperparameters of KNN: number of neighbors n

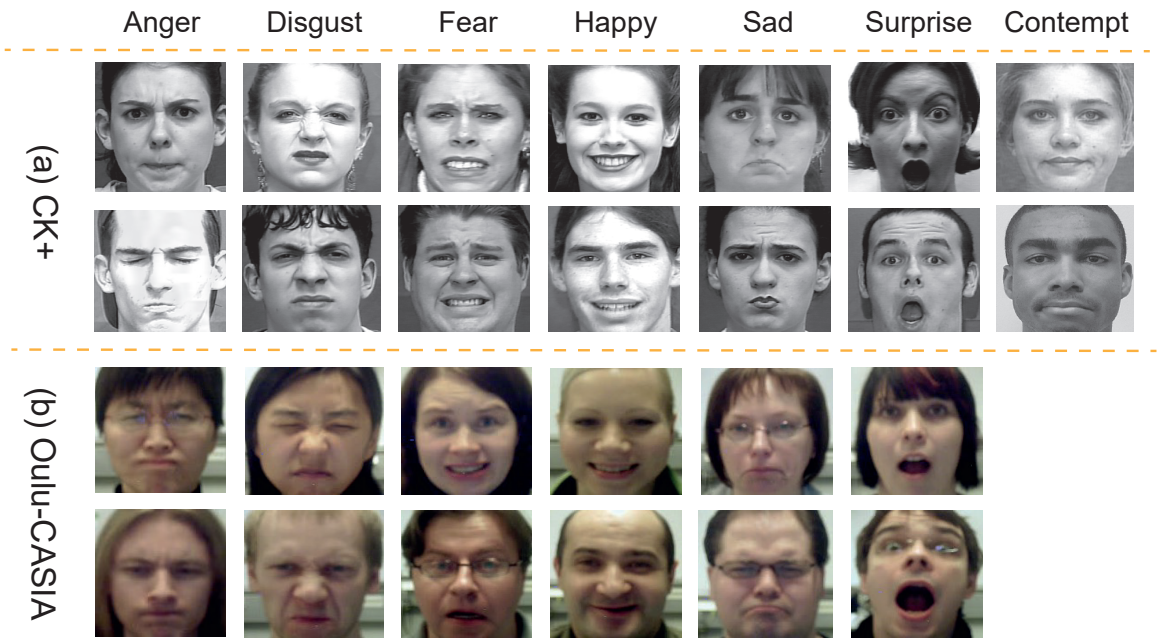


Figure 6.4: Sample images of facial expressions extracted from the CK+ [8] and Oulu-CASIA [9] datasets.

and hyperparameters of SVM with Gaussian kernel: regularization parameter C, kernel coefficient γ were determined through grid search method. The search field was configured as follows: the number of neighbors $n \in$ [1, 10], regularization $C \in$ [1e-3, 1e+3] with step 10, kernel coefficient $\gamma \in$ [1e-5, 10]. For evaluation, we performed the same experimental protocol in the literature [219], [127], [220], [211], namely 10-fold subject-independent cross-validation on CK+ dataset and Oulu-CASIA dataset with two metrics: accuracy and confusion matrix. All subjects were divided into 10 groups in ascending order of ID, 9 groups were used for training, and the remaining group was used for testing. In the end, the average classification performance is reported in the following sections.

6.4.3 Results and Discussion

First, we investigated the effect of target neighbors parameter k in the LMNN algorithm on the classification performance. The evolutions of accuracy when varying k are shown in Fig. 6.5. The optimal performance was achieved on CK+ and Oulu-CASIA datasets when $k = 8$ and $k = 5$, respectively, for the KNN classifier, while $k = 7$ and $k = 5$ give the best results for the SVM classifier. The above settings of parameter k were retained for subsequent experiments. We then verified the effectiveness of the LMNN algorithm based on the proposed metric. Table 6.1 presents the average classification accuracy of the Fubini-Study metric before and after LMNN-

based distance metric learning for facial expression recognition task on CK+ and Oulu-CASIA datasets, respectively. From the obtained results, we can observe that the optimal Fubini-Study metric $d_A{}^\star$ obtained maximum gains of 4.84% and 9.89% in classification accuracy on CK+ and Oulu-CASIA datasets compared to the original ones, which demonstrates that the use of a suitable metric can enhance facial expression recognition performance. In addition, the SVM classifier generated better classification performance than KNN, both in terms of the original and the optimal metric.

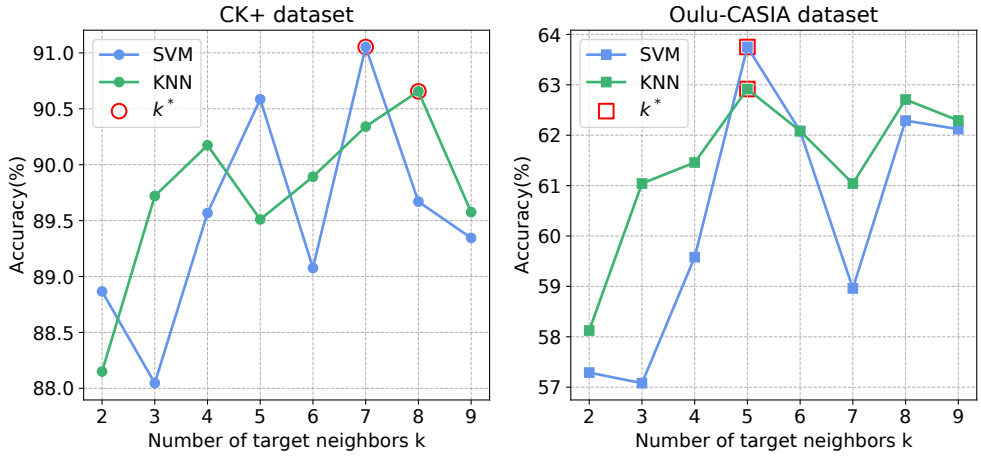


Figure 6.5: Average classification accuracy on the CK+ and Oulu-CASIA datasets using the Fubini-Study metric when varying the target neighbor parameter k.

Table 6.1: Overall accuracy (Acc %) of of two types of metrics, i.e., the Euclidean metrics and the Fubini-Study metrics, before and after LMNN-based metric learning on the CK+ and Oulu-CASIA datasets. ($[\cdot]^I$: identity matrix; $[\cdot]^\star$: optimal matrix.)

Dataset	Type	Metric	Classifier	
			KNN	SVM
CK+	Euclidean	Original Metric $d_M{}^I$	79.84	79.97
		Optimal Metric $d_M{}^\star$	86.41	89.15
	Fubini-Study	Original Metric $d_A{}^I$	85.81	87.53
		Optimal Metric $d_A{}^\star$	<u>90.65</u>	**91.05**
Oulu-CASIA	Euclidean	Original Metric $d_M{}^I$	51.75	50.49
		Optimal Metric $d_M{}^\star$	61.66	61.84
	Fubini-Study	Original Metric $d_A{}^I$	53.01	55.62
		Optimal Metric $d_A{}^\star$	<u>62.91</u>	**63.75**

6.4.3.A Fubini-Study metric vs Euclidean metric

A comparison with the Euclidean metric in Eq.(6.2) in measuring the similarity between different facial expressions is also executed on CK+ and Oulu-CASIA datasets to validate the

Table 6.2: Confusion matrix using two optimal metrics, i.e., Fubini-Study metric and Euclidean metric obtained after metric learning for the CK+ dataset. (*Angry* (An); *Contempt* (Co); *Disgust* (Di); *Fear* (Fe); *Happy* (Ha); *Sad* (Sa); *Surprise* (Su). Underlining indicates superior performance of the proposed metric.)

(a) Fubini-Study metric

	An	Co	Di	Fe	Ha	Sa	Su
An	**84.4**	2.2	4.4	0.0	0.0	8.9	0.0
Co	5.5	**61.1**	5.6	11.1	5.6	11.1	0.0
Di	0.0	0.0	**98.3**	0.0	0.0	1.7	0.0
Fe	0.0	4.0	0.0	**80.0**	4.0	4.0	8.0
Ha	0.0	0.0	0.0	1.4	**98.6**	0.0	0.0
Sa	14.3	3.6	0.0	0.0	0.0	**82.1**	0.0
Su	0.0	1.2	0.0	0.0	0.0	0.0	**98.8**

(b) Euclidean metric

	An	Co	Di	Fe	Ha	Sa	Su
An	**84.4**	2.2	4.4	0.0	0.0	8.9	0.0
Co	11.1	**50.0**	0.0	5.6	0.0	27.8	5.6
Di	0.0	0.0	**100.0**	0.0	0.0	0.0	0.0
Fe	0.0	0.0	0.0	**72.0**	16.0	8.0	4.0
Ha	0.0	0.0	0.0	0.0	**100.0**	0.0	0.0
Sa	14.3	0.0	3.6	0.0	0.0	**78.6**	3.6
Su	1.2	1.2	0.0	1.2	0.0	0.0	**96.4**

Table 6.3: Confusion matrix using two optimal metrics, i.e., Fubini-Study metric and Euclidean metric obtained after metric learning for the Oulu-CASIA dataset. (*Angry* (An); *Disgust* (Di); *Fear* (Fe); *Happy* (Ha); *Sad* (Sa); *Surprise* (Su). Underlining indicates superior performance of the proposed metric.)

(a) Fubini-Study metric

	An	Di	Fe	Ha	Su	Sa
An	**58.8**	12.5	3.8	2.5	1.2	21.2
Di	20.0	**50.0**	5.0	3.8	3.7	17.5
Fe	5.0	5.0	**65.0**	8.7	10.0	6.3
Ha	1.2	1.3	11.2	**85.0**	0.0	1.3
Su	2.5	5.0	15.0	0.0	**76.3**	1.2
Sa	26.2	10.0	12.5	1.3	2.5	**47.5**

(b) Euclidean metric

	An	Di	Fe	Ha	Su	Sa
An	**51.2**	21.3	6.3	0.0	0.0	21.2
Di	25.0	**38.8**	7.5	2.5	3.7	22.5
Fe	5.0	5.0	**60.0**	13.8	6.2	10.0
Ha	5.0	0.0	11.3	**83.7**	0.0	0.0
Su	1.2	5.0	11.3	0.0	**82.5**	0.0
Sa	21.3	13.7	11.2	0.0	1.3	**52.5**

effectiveness of the proposed metric. From the results of the Euclidean metric in Table 6.1, we can observe that the performance of KNN and SVM is improved by the proposed metric, regardless of whether metric learning is employed or not. These results demonstrate that facial deformations can be better modeled in complex projective space rather than in Euclidean space. To further investigate the validity of the proposed metric for identifying each facial expression, we provide also the confusion matrix obtained using the two learned metrics, i.e., Fubini-Study metric and Euclidean metric obtained after metric learning, for the CK+ and Oulu-CASIA datasets in Table 6.2 and Table 6.3, respectively. For CK+ dataset, the *Disgust*, *Happy*, and *Surprise* expressions can be well recognized using the Fubini-study metric with an accuracy rate of over 95%, while the main confusions happened in the *Contempt* expression. This can be explained by its slight facial changes and its smaller sample size compared to other expressions. When compared with the Euclidean metric, the recognition performance of the *Con-*

tempt, *Fear*, *Sad*, *Surprise* expressions is enhanced by up to 11.1%. For Oulu-CASIA dataset, the two metrics can both better identify *Happy* and *Surprise* than other expressions. Among all expressions, *Angry*, *Disgust*, *Fear*, *Happy* can be better classified by the Fubini-Study metric with a maximum gain of 11.2%. Fig. 6.6, Fig. 6.7 show the 2D visualization of the CK+ and Oulu-CASIA datasets using the t-SNE [221] with the two metrics. For CK+ dataset, the 2D visualization obtained by the original Euclidean metric in Fig. 6.6 (c) shows an approximately linear relationship, and the points of different expressions are mixed together, while the original Fubini-study metric (Fig. 6.6 (a)) has been able to well distinguish the two expressions: *Happy* and *Surprise*. After metric learning, the Fubini-study metric can lead to a more uniform feature distribution for different expressions compared to the Euclidean metric. We reached the same conclusion on the Oulu-CASIA dataset that our proposed metric always yields a more discriminative feature representation both before and after metric learning, with respect to the Euclidean metric.

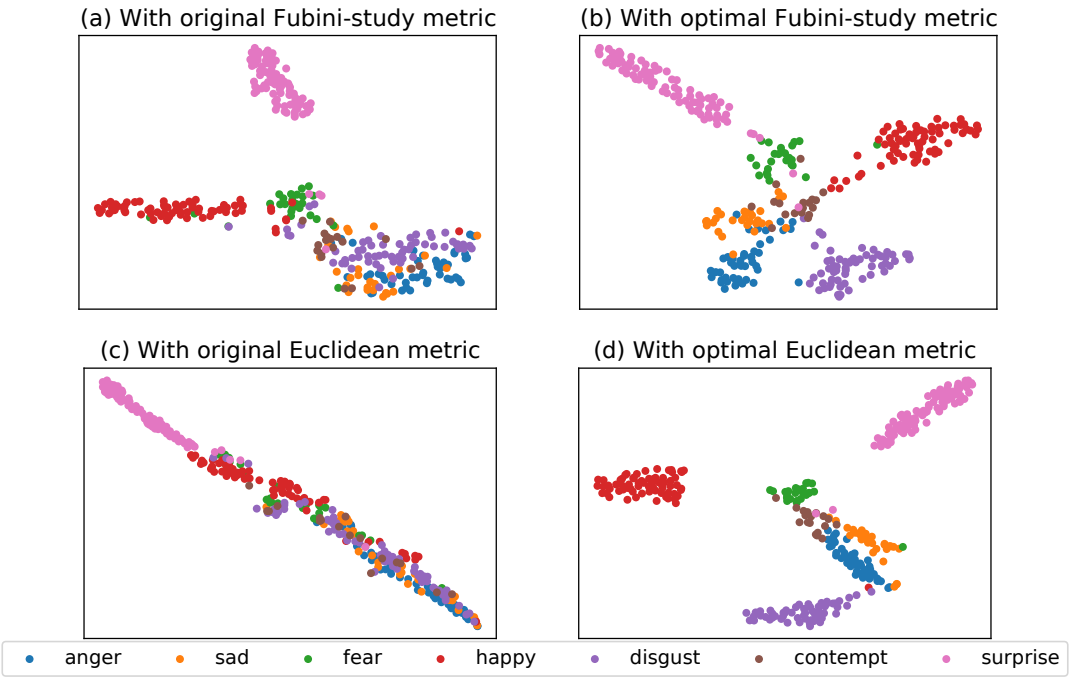


Figure 6.6: 2D visualization of CK+ dataset using t-SNE method with the proposed metrics. (a-b) show the visualization results before and after metric learning with Fubini-Study metric. (c-d) show the visualization results before and after metric learning with Euclidean metric.

6.4.3.B Comparison with the State-of-the-Art

We first compared the proposed method with state-of-the-art approaches for CK+ dataset, with the details reported in Table 6.4. For a fair comparison, only the geometry-based methods

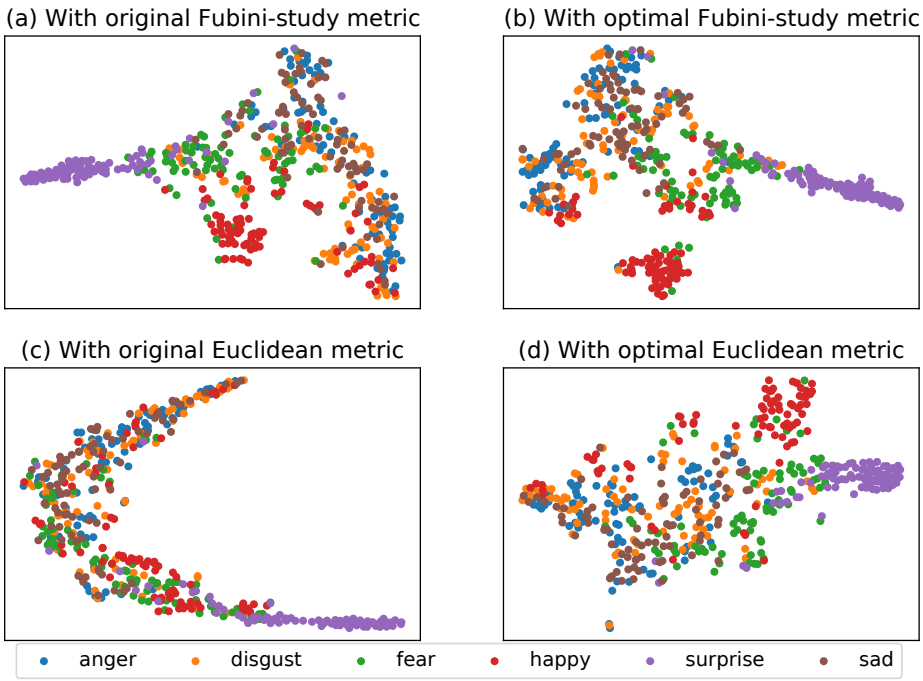


Figure 6.7: 2D visualization of Oulu-CASIA dataset using t-SNE method with the proposed metrics. (a-b) show the visualization results before and after metric learning with Fubini-Study metric. (c-d) show the visualization results before and after metric learning with Euclidean metric.

that use facial landmarks were considered in the comparative evaluation. We divided these methods into two categories, depending on whether the input is a static image or a dynamic video sequence. The only comparable static method [211] used the Gram matrix computed from landmarks as the shape representation and conducted a non-Euclidean metric learning on the manifold of oriented ellipses centered at the origin in Euclidean n-space. The comparison results show that our method performs better, both in terms of the original metric and the optimal metric. For dynamic approaches, the ranked-first approach is the work of [127], where the facial landmark sequences were modeled as parametrized trajectories on the Riemannian manifold of positive semi-definite matrices of fixed-rank. However, their metric applied to calculate the similarity between trajectories contains a weight parameter, which needs to be carefully selected for each dataset. According to the performance ranking, the following method [220] encoded time-varying facial shapes and used deep neural networks for classification. Compared with these work, our method using solely the last frame achieves competitive results on facial expression recognition. For Oulu-CASIA dataset, few geometry-based methods are available for comparison. Based on this, we implemented the static method [211] using the code provided by the author and the comparison results are shown in Table 6.5. Consistent with the results on CK+, our approach always yields higher accuracy rates, regardless of whether metric learning is used, proving its effectiveness once again for facial expression recognition.

Table 6.4: Comparison with state-of-the-art geometric methods on the CK+ dataset.

Input	Methods	Accuracy
Video sequence	Shape velocity on $\mathcal{G}$ [222]	82.80
Video sequence	LDCRFs [223]	85.84
Video sequence	ITBN [224]	86.30
Video sequence	Intrinsic SCDL - SVM [220]	89.43
Video sequence	Extrinsic SCDL - Bi-LSTM [220]	95.73
Video sequence	**Shape trajectories on** $\mathcal{S}^+(d,n)$ [127]	**96.87**
Static image	original metric on $\mathcal{S}_c^+(2,n)$ [211]	85.83
Static image	**original metric on** $\mathbb{CP}^{n-1}$ **(ours)**	**87.53**
Static image	optimal metric on $\mathcal{S}_c^+(2,n)$ [211]	90.53
Static image	**optimal metric on** $\mathbb{CP}^{n-1}$ **(ours)**	**91.05**

Table 6.5: Comparison with state-of-the-art geometric methods on the Oulu-CASIA dataset.

Input	Methods	Accuracy
Static image	original metric on $\mathcal{S}_c^+(2,n)$ [211]	53.33
Static image	**original metric on** $\mathbb{CP}^{n-1}$ **(ours)**	**55.62**
Static image	optimal metric on $\mathcal{S}_c^+(2,n)$ [211]	62.08
Static image	**optimal metric on** $\mathbb{CP}^{n-1}$ **(ours)**	**63.75**

6.5 Conclusion

In this chapter, a non-Euclidean metric learning method was proposed and applied to facial expression recognition, where the facial shapes were directly encoded by the equivalent class of the landmarks in the complex projective space. The similarity between them can be measured by Fubini-Study metrics. A metric learning algorithm based on Large Margin Nearest Neighbors (LMNN) was adapted to perform metric optimization on the Riemannian manifold, i.e. the space of positive-definite Hermitian matrices. Facial expression recognition experiments were conducted on two public datasets to verify the validy of the proposed method. The experimental results showed that, compared to the metric defined in the Euclidean space, the Fubini-Study metrics are more effective and discriminative for identification of facial expressions. The latter one also achieves comparable performance to the state-of-the-art approaches. In the next chapter, we will discuss the limitations of the methods involved in this thesis and the perspectives for future research on emotion recognition.

7

Conclusion

Contents

Emotion recognition, an important branch of affective computing, has received considerable attention for enabling machines to automatically recognise human emotions. With advances in sensor technologies and human-computer interaction systems, there is a growing trend to migrate from laboratory environments to unrestricted scenarios such as medical simulation training. However, there is still a lack of relevant research on the latter. In order to build emotion recognition systems in real-life settings, from a technical point of view, appropriate intelligent algorithms for emotion recognition should be designed at the behavioural, physiological, and even multimodal levels. From a practical point of view, recognition algorithms need to be integrated into the user interface and the effects of the emotional feedback obtained should be investigated. The work we present in this thesis concentrates on exactly these two aspects. At the end of the thesis, we review the main contributions and their corresponding limitations in Sections 7.1 and 7.2, and we discuss the perspectives for future work in Section 7.3.

7.1 Contributions

7.1.1 Technical level

From the physiological perspective, we proposed a novel emotion recognition method based on peripheral physiological signals. A deep neural network based on the intermediate fusion strategy was designed to optimize feature extraction and fusion of multivariate sequences, where a modality-specific encoder based on residual temporal convolution was first employed to extract low-level unimodal features, and then a transformer-based shared encoder with stacked multimodal features as input was applied to capture both inter-modal correlation information. To obtain a generalized model, we adopted a self-supervised learning (SSL) scheme. The proposed model was first pre-trained on a large-scale unlabelled physiological dataset with the pretext objective of signal transformation classification. A series of downstream emotion recognition tasks involving mental stress, emotional state, arousal and valence were then performed to validate the effectiveness of the pre-trained model. Finally, our approach achieved state-of-the-art results in comparison with fully-supervised, unsupervised, and self-supervised-based methods.

From the behavioural perspective, We proposed a non-Euclidean metric learning method for the facial expression recognition task. We encoded facial shapes as equivalence classes in the complex projective space using the complex form of detected 2D landmarks, resulting in an affine-invariant shape representation. A metric learning algorithm based on Large Margin

Nearest Neighbors (LMNN) was applied to optimize the metrics in the non-linear space, thereby yielding a feature space with smaller intra-class differences and larger inter-class differences. The proposed method finally obtained superior performance to Euclidean metric learning methods and yielded competitive results in comparison with state-of-the-art methods.

From the multimodal perspective, we presented a novel deep geometric framework for multimodal stress and pain detection tasks. To fully explore the potential interactions between different modalities, we fused behavioural and physiological information into a single symmetric positive definite (SPD) matrix-based representation that incorporates both intra- and inter-modal correlations. To cope with the non-linearity of the manifold induced by the SPD matrices, we employed a tangent space mapping technique to map the resulting SPD matrix sequences into vector sequences in linear space, enabling the use of an LSTM-based deep neural network for classification. To the best of our knowledge, this is the first use of the geometry of the SPD matrix for fusing physiological and behavioural signals. This novel fusion strategy allows the proposed method to achieve improved performance in comparison with feature-level or decision-level fusion-based methods.

7.1.2 Practical level

We conducted emotion recognition experiments in a real-life setting, i.e., high-fidelity medical simulation, where multimodal data consisting of peripheral physiological and action signals and a series of psychological scales were collected from students. The evaluation results of the experiments confirmed the validity of the proposed algorithms in Chapters 3 and 4, which were then integrated into a graphical user interface to explore its impact on the pedagogical debriefing sessions. Ultimately, the users' feedback showed that emotion-aware teaching can facilitate students' acquisition of competencies and increase their engagement.

7.2 Limitations

7.2.1 Technical level

From the physiological perspective, first, we stacked the embedding sequences from the output of modality-specific encoders as inputs to the transformer, expecting to learn inter-modal dependencies through the self-attention mechanism. However, since the complexity of the transformer is $O(N^2)$ at an input sequence length of N, this natural concatenation and dense

cross-modal attention operation entail high computational costs, especially for large-scale learning [225]. Secondly, regarding the pretext task of self-supervised learning, we adopted a fixed number of signal transformations to learn generalised representations for downstream tasks. However, there may be an optimal combination strategy for signal transformations and an optimal set of parameter settings for these transformations. Furthermore, the optimal transformation policy probably varies from one modality to another. Third, the self-supervised pretraining process suffers from an early de-generalisation problem, where the model possibly converges too fast on the discriminative pretext task thereby losing its generalisation in downstream tasks [226].

From the behavioural perspective, our proposed method has two major limitations. First, it is based on static facial landmarks and therefore only takes into account spatial information. However, the execution of facial expressions is a dynamic process. Thus, the method neglects to encode the temporal changes of facial point positions. Secondly, recent methods have attempted to apply 3D facial landmarks for the identification of facial expressions, which can better tackle the distortions introduced by head pose variations and obtain more meaningful distance measurements. However, the proposed method cannot be extended to the 3D case as it is not possible to express the coordinates of 3D points in complex form.

From the multimodal perspective, there are two main limitations in our approach. First, the multimodal signals are temporally inconsistent, which is manifested by dense continuous physiological signals and relatively sparse video images. We adopted the most intuitive approach, a downsampling techinique to match multimodal data sizes for the computation of the proposed SPD-based representation. However, this may distort the correspondence between instances of multimodality [227]. Second, to address the non-linearity of the SPD manifold, we adopted an intrinsic solution of projecting the manifold data onto a tangent space at some reference point via a logarithmic map. However, this may introduce approximation errors, especially when the projected point lies further away from the reference point.

7.2.2 Practical level

First, in terms of ground truth collection, a series of self-report questionnaires were applied to measure emotional states. However, they inevitably have some limitations. They only record emotions at a particular moment in time, rather than providing a continuous measure, and there exist recall biases in the post hoc collection of subjective experiences. In addition, there may be a delay between the real emotional experience and the emotion perceived by the individual.

All of these factors affect the quality of the labeling. Second, from the perspective of the recognition algorithm, our approach employed a binary classification policy for the target emotion, i.e. stress. However, given the complexity and non-linearity of human emotional responses, differentiating only between stressed and unstressed appears to be ineffective in capturing the subtle evolution of emotional states.

7.3 Future work

7.3.1 Technical level

From the physiological perspective, for transformers with stacked multimodal features as input, one of the future research lines is how to reduce their computational complexity. One possible solution is to adopt a sparse attention mechanism, which measures similarity only for certain key positions in a sequence, and the resulting sparse attention matrix allows for reduced computational time and memory requirements. Furthermore, in order to maximise the generalisation of the representations learned by the self-supervised model, a scheme that can automatically learn the optimal set of signal transformations for each modality needs to be designed for its integration into the original framework. Moreover, the same idea could be applied to the design of upstream pretext tasks, as sometimes learned features on the basis of the adopted pretext task show poor transferability on downstream data.

From the behavioural perspective, the proposed method can be extended to dynamic scenarios. Sequences of landmarks will be explored to model the temporal relationship, where successive landmark configurations are connected by geodesic to create a parameterized curve in the complex projective space and the final classification will be executed based on the comparison of similarity between the curves. In addition, we will also explore the application of other types of landmarks, such as body landmarks, for emotion recognition. Body movements involve larger muscle groups with distinct changes, whereas facial motions tend to be more subtle and may be difficult to discern. In such cases, the use of body landmarks as an emotion recognition modality may provide more information than facial landmarks. Moreover, body landmarks are frequently involved in another interesting direction of research, namely action recognition.

From the multimodal perspective, in order to better address the inconsistency on the time scale of multimodal data, an implicit alignment module can be introduced into our proposed model to obtain better performance. Different backbones can first be applied to the physiologi-

cal and behavioural modalities respectively to extract low-level features, followed by the calculation of the similarity between the sub-components to align the features from different modalities before proceeding to fusion. Furthermore, future work could develop deep neural networks suitable for non-linear SPD matrices in order to avoid distortion of the manifold data due to tangent space mapping. Recent studies [228, 229] have proposed end-to-end deep Riemannian architectures that can be directly applied to SPD matrices. An example is SPDNet [228], which is inspired by classical convolutional neural networks, where a bilinear mapping layer and an eigenvalue rectification layer were built to simulate convolutional and linear rectification operations.

7.3.2 Practical level

First, in order to enhance the reliability of the acquired emotion's ground truth, assessments from a third party such as a trainer or a patient can be incorporated in the data collection procedure. Second, in order to optimise the quality of pedagogy, a multi-level emotion classification system should be developped. A more optimal solution would be performing regression tasks to predict arousal and value scores, allowing to obtain a more comprehensive, continuous emotional evolution of the trainee during the medical simulation. In addition, person-specific emotion recognition models can be also deployed to facilitate individualised teaching programmes for students. Third, during the tests of the developed graphical interface, we revealed correlations between the recognized stress fragments and verbal expressions, which guided us to incorporate speech signals in future recognition frameworks to achieve more accurate and robust performance.

Bibliography

[1] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, vol. 18, no. 5, pp. 459–482, Nov. 1908. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/cne.920180503

[2] A. Saeed, S. Trajanovski, M. Van Keulen, and J. Van Erp, "Deep physiological arousal detection in a driving simulator using wearable sensors," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2017, pp. 486–493.

[3] K. You and H.-J. Park, "Re-visiting riemannian geometry of symmetric positive definite matrices for the analysis of functional connectivity," *NeuroImage*, vol. 225, p. 117464, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920309496

[4] B. Zhong, Z. Qin, S. Yang, J. Chen, N. Mudrick, M. Taub, R. Azevedo, and E. Lobaton, "Emotion recognition with facial expressions and physiological signals," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–8.

[5] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *Journal of Biomedical Informatics*, vol. 73, pp. 159–170, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046417301855

[6] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0005791694900639

[7] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, p. 207–244, Jun. 2009.

[8] 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*, 2010, pp. 94–101.

[9] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[10] P. J. Bota, C. Wang, A. L. N. Fred, and H. Plácido Da Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140 990–141 020, 2019.

[11] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Wearable affect and stress recognition: A review," 2018. [Online]. Available: https://arxiv.org/abs/1811.08854

[12] R. W. Picard, *Affective Computing*. MIT Press, Jul. 2000, google-Books-ID: GaVncRTcb1gC.

[13] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517300738

[14] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems," *Journal of Medical and Biological Engineering*, vol. 40, no. 2, pp. 149–157, Apr. 2020.

[15] K. Lai, S. N. Yanushkevich, and V. P. Shmerko, "Intelligent stress monitoring assistant for first responders," *IEEE Access*, vol. 9, pp. 25 314–25 329, 2021.

[16] H. Silva, A. Lourenço, and A. Fred, "In-vehicle driver recognition based on hand ecg signals," in *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, ser. IUI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 25–28.

[17] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[18] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, and J. Mao, "A facial expression emotion recognition based human-robot interaction system," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 668–676, 2017.

[19] M. Granato, D. Gadia, D. Maggiorini, and L. A. Ripamonti, "Feature extraction and selection for real-time emotion recognition in video games players," in *2018 14th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, 2018, pp. 717–724.

[20] Y. Liu, O. Sourina, and M. R. Hafiyyandi, "Eeg-based emotion-adaptive advertising," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 843–848.

[21] G. Li and Y. Wang, "Research on leamer's emotion recognition for intelligent education system," in *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2018, pp. 754–758.

[22] O. El Hammoumi, F. Benmarrakchi, N. Ouherrou, J. El Kafi, and A. El Hore, "Emotion recognition in e-learning systems," in *2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, 2018, pp. 1–6.

[23] G. Bower, "Mood and memory," *American Psychologist*, vol. 36, pp. 129–148, 03 1981.

[24] J. Harley, S. Lajoie, C. Frasson, and N. Hall, "An integrated emotion-aware framework for intelligent tutoring systems," vol. 9112, 06 2015.

[25] M. Feidakis, T. Daradoumis, and S. Caballe, "Endowing e-learning systems with emotion awareness," in *2011 Third International Conference on Intelligent Networking and Collaborative Systems*, 2011, pp. 68–75.

[26] M. H. Immordino-Yang and A. Damasio, "We feel, therefore we learn: The relevance of affective and social neuroscience to education," *Mind, Brain, and Education*, vol. 1, no. 1, pp. 3–10, 2007. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-228X.2007.00004.x

[27] H. Selye, *The stress of life*. New York: McGraw-Hil Edition, 1956.

[28] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519302532

[29] P. J. Bota, C. Wang, A. L. N. Fred, and H. Plácido Da Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140 990–141 020, 2019.

[30] C. Ding and D. Tao, "Pose-invariant face recognition with homography-based normalization," *Pattern Recognition*, vol. 66, pp. 144–152, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320316303831

[31] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[32] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, and W. Zhang, "A systematic review on affective computing: emotion models, databases, and recent advances," *Information Fusion*, vol. 83-84, pp. 19–52, 2022.

[33] G. Colombetti, "From affect programs to dynamical discrete emotions," *Philosophical Psychology*, vol. 22, no. 4, pp. 407–425, 2009. [Online]. Available: https://doi.org/10.1080/09515080903153600

[34] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992. [Online]. Available: https://doi.org/10.1080/02699939208411068

[35] R. Plutchik, "A psychoevolutionary theory of emotions," *Social Science Information*, vol. 21, no. 4-5, pp. 529–553, 1982. [Online]. Available: https://doi.org/10.1177/053901882021004003

[36] J. A. Russell, "Affective space is bipolar," *Journal of Personality and Social Psychology*, vol. 37, pp. 345–356, 1979, place: US Publisher: American Psychological Association.

[37] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Wearable affect and stress recognition: A review," *arXiv:1811.08854 [cs]*, Nov. 2018.

[38] P. J. Bota, C. Wang, A. L. N. Fred, and H. Placido Da Silva, "A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals," *IEEE Access*, vol. 7, pp. 140 990–141 020, 2019.

[39] R. Plutchik, *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*. American Psychological Association, 2003. [Online]. Available: https://books.google.fr/books?id=a0p4QgAACAAJ

[40] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *Journal of Biomedical Informatics*, vol. 59, pp. 49–75, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046415002750

[41] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognition Letters*, vol. 120, pp. 69 – 74, 2019.

[42] S. Piana, A. Staglianò, F. Odone, A. Verri, and A. Camurri, "Real-time automatic emotion recognition from body gestures," *ArXiv*, vol. abs/1402.5047, 2014.

[43] K. Huang, C. Wu, Q. Hong, M. Su, and Y. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5866–5870.

[44] A. Mehrabian, "Communication without words," 1968.

[45] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, pp. 124–129, 1971, place: US Publisher: American Psychological Association.

[46] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 505–523, 2021.

[47] M.-Z. Balaban (Ghitan), "The Definitive Book of Body Language by Allen & Barbara Pease, Orion Books Ltd, London 2005," *Lingua. Language and Culture*, vol. XVI, no. 2, pp. 153–156, 2017, publisher: Universitatea »Babes Bolyai« Cluj - Facultatea de St. Economice si Gestiunea Afacerilor. [Online]. Available: https://www.ceeol.com/search/article-detail?id=620945

[48] J. Ramos Rojas, J.-H. Hong, and A. Dey, "Stress recognition: A step outside the lab," 01 2014.

[49] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion Recognition: a review," in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. Penang, Malaysia: IEEE, Mar. 2011, pp. 410–415.

[50] D. Alexander, C. Trengove, P. Johnston, T. Cooper, J. August, and E. Gordon, "Separating individual skin conductance responses in a short interstimulus-interval paradigm," *Journal of Neuroscience Methods*, vol. 146, no. 1, pp. 116–123, Jul. 2005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0165027005000464

[51] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A Review of Emotion Recognition Using Physiological Signals," *Sensors*, vol. 18, no. 7, p. 2074, Jun. 2018.

[52] G. Valenza, A. Lanata, and E. P. Scilingo, "The role of nonlinear dynamics in affective valence and arousal recognition," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 237–249, 2012.

[53] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18*, Boulder, CO, USA, 2018, pp. 400–408.

[54] M. Yan, Z. Deng, B. He, C. Zou, J. Wu, and Z. Zhu, "Emotion classification with multichannel physiological signals using hybrid feature and adaptive decision fusion," *Biomedical Signal Processing and Control*, vol. 71, p. 103235, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421008326

[55] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological Psychology*, vol. 84, no. 3, pp. 394–421, 2010, the biopsychology of emotion: Current theoretical and empirical perspectives. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301051110000827

[56] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," in *Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 125–134.

[57] S. Z. Bong, M. Murugappan, and S. Yaacob, "Analysis of electrocardiogram (ecg) signals for human emotional stress classification," in *Trends in Intelligent Robotics, Automation, and Manufacturing*, S. G. Ponnambalam, J. Parkkinen, and K. C. Ramanathan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 198–205.

[58] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[59] X. Zhang, C. Xu, W. Xue, J. Hu, Y. He, and M. Gao, "Emotion recognition based on multichannel physiological signals with comprehensive nonlinear processing," *Sensors*, vol. 18, no. 11, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/11/3886

[60] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob, "Emotion recognition from electrocardiogram signals using hilbert huang transform," in *2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT)*, 2012, pp. 82–86.

[61] J. Xie, X. Xu, and L. Shu, "Wt feature based emotion recognition from multi-channel physiological signals with decision fusion," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 2018, pp. 1–6.

[62] Z. Guendil, Z. Lachiri, C. Maaoui, and A. Pruski, "Emotion recognition from physiological signals using fusion of wavelet based features," in *2015 7th International Conference on Modelling, Identification and Control (ICMIC)*, 2015, pp. 1–6.

[63] J. Rubin, R. Abreu, S. Ahern, H. Eldardiry, and D. G. Bobrow, "Time, frequency complexity analysis for recognizing panic states from physiologic time-series," in *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*, ser. PervasiveHealth '16. Brussels, BEL: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2016, p. 81–88.

[64] "Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems," vol. 40. [Online]. Available: http://link.springer.com/10.1007/s40846-019-00505-7

[65] L. lan Chen, Y. Zhao, P. fei Ye, J. Zhang, and J. zhong Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Expert Systems with Applications*, vol. 85, pp. 279–291, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417417300477

[66] A. S. Anusha, J. Jose, S. P. Preejith, J. Jayaraj, and S. Mohanasankar, "Physiological signal based work stress detection using unobtrusive sensors," *Biomedical Physics Engineering Express*, vol. 4, no. 6, p. 065001, sep 2018. [Online]. Available: https://dx.doi.org/10.1088/2057-1976/aadbd4

[67] Z. Cheng, L. Shu, J. Xie, and C. L. P. Chen, "A novel ecg-based real-time detection method of negative emotions in wearable applications," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, 2017, pp. 296–301.

[68] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, vol. 7, pp. 57–67, 2019.

[69] L. Huynh, T. Nguyen, T. Nguyen, S. Pirttikangas, and P. Siirtola, *StressNAS: Affect State and Stress Detection Using Neural Architecture Search*. New York, NY, USA: Association for Computing Machinery, 2021, p. 121–125.

[70] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "Cornet: Fine-grained emotion recognition for video watching using wearable physiological sensors," *Sensors*, vol. 21, no. 1, 2021.

[71] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Briefings in Bioinformatics*, vol. 23, no. 2, 01 2022, bbab569. [Online]. Available: https://doi.org/10.1093/bib/bbab569

[72] A. Saeed and S. Trajanovski, "Personalized driver stress detection with multi-task neural networks using physiological signals," 2017. [Online]. Available: https://arxiv.org/abs/1711.06116

[73] D. Lopez-Martinez and R. Picard, "Multi-task neural networks for personalized pain recognition from physiological signals," 2017. [Online]. Available: https://arxiv.org/abs/1708.08755

[74] J. Bradbury, S. Merity, C. Xiong, and R. Socher, "Quasi-recurrent neural networks," 2016. [Online]. Available: https://arxiv.org/abs/1611.01576

[75] K. Wang, Y. L. Murphey, Y. Zhou, X. Hu, and X. Zhang, "Detection of driver stress in real-world driving environment using physiological signals," in *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, vol. 1, 2019, pp. 1807–1814.

[76] M. Awais, M. Raza, N. Singh, K. Bashir, U. Manzoor, S. U. Islam, and J. J. P. C. Rodrigues, "Lstm-based emotion detection using physiological signals: Iot framework for healthcare and distance learning in covid-19," *IEEE Internet of Things Journal*, vol. 8, no. 23, pp. 16 863–16 871, 2021.

[77] M. S. Zitouni, C. Y. Park, U. Lee, L. Hadjileontiadis, and A. Khandoker, "Arousal-valence classification from peripheral physiological signals using long short-term memory networks," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2021, pp. 686–689.

[78] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 985–990.

[79] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual lstm network," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 176–183. [Online]. Available: https://doi.org/10.1145/3343031.3350871

[80] J. Zhu, X. Zhao, H. Hu, and Y. Gao, "Emotion recognition from physiological signals using multi-hypergraph neural networks," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 610–615.

[81] J. Liao, Q. Zhong, Y. Zhu, and D. Cai, "Multimodal physiological signal emotion recognition based on convolutional recurrent neural network," *IOP Conference Series: Materials Science and Engineering*, vol. 782, no. 3, p. 032005, mar 2020. [Online]. Available: https://dx.doi.org/10.1088/1757-899X/782/3/032005

[82] X. Zhang, J. Liu, J. Shen, S. Li, K. Hou, B. Hu, J. Gao, T. Zhang, and B. Hu, "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine," *IEEE Transactions on Cybernetics*, vol. 51, no. 9, pp. 4386–4399, 2021.

[83] M. N. Dar, M. U. Akram, S. G. Khawaja, and A. N. Pujari, "Cnn and lstm-based emotion charting using physiological signals," *Sensors*, vol. 20, no. 16, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/16/4551

[84] R. K. Shahzad and N. Lavesson, "Comparative analysis of voting schemes for ensemble-based malware detection," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 4, no. 1, pp. 98–117, 2013, open Access Journal.

[85] M. D. Hssayeni and B. Ghoraani, "Multi-modal physiological data fusion for affect estimation using deep learning," *IEEE Access*, vol. 9, pp. 21 642–21 652, 2021.

[86] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition,"

Information Processing Management, vol. 57, no. 3, p. 102185, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457319307204

[87] H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu, "Unsupervised learning based on artificial neural network: A review," in *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 2018, pp. 322–327.

[88] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey," *A Review of Machine Learning Techniques for Processing Multimedia Content*, 09 2005.

[89] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, "A non-eeg biosignals dataset for assessment and visualization of neurological status," in *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, 2016, pp. 110–114.

[90] L. Fiorini, G. Mancioppi, F. Semeraro, H. Fujita, and F. Cavallo, "Unsupervised emotional state classification through physiological parameters for social robotics applications," *Knowledge-Based Systems*, vol. 190, p. 105217, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119305465

[91] K. Ross, P. Hungler, and A. Etemad, "Unsupervised multi-modal representation learning for affective computing with multi-corpus wearable data," *Journal of Ambient Intelligence and Humanized Computing*, Oct. 2021. [Online]. Available: https://doi.org/10.1007/s12652-021-03462-9

[92] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, Feb. 2020. [Online]. Available: http://link.springer.com/10.1007/s10994-019-05855-6

[93] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, "A survey on semi-, self- and unsupervised learning for image classification," *IEEE Access*, vol. 9, pp. 82 146–82 168, 2021.

[94] J. Luo, Y. Tian, H. Yu, Y. Chen, and M. Wu, "Semi-supervised cross-subject emotion recognition based on stacked denoising autoencoder architecture using a fusion of multi-modal physiological signals," *Entropy*, vol. 24, no. 5, 2022. [Online]. Available: https://www.mdpi.com/1099-4300/24/5/577

[95] H. Yu and A. Sano, "Semi-supervised learning and data augmentation in wearable-based momentary stress detection in the wild," 2022. [Online]. Available: https://arxiv.org/abs/2202.12935

[96] 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 (CVPR)*, June 2013.

[97] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2610–2617.

[98] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[99] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 1944–1951.

[100] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast aam fitting in-the-wild," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 593–600.

[101] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[102] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022.

[103] B. Martinez and M. F. Valstar, *Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition*. Cham: Springer International Publishing, 2016, pp. 63–100. [Online]. Available: https://doi.org/10.1007/978-3-319-25958-1_4

[104] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790–799, 2014, best of Automatic Face and Gesture Recognition 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885613001765

[105] J. Orozco, B. Martinez, and M. Pantic, "Empirical analysis of cascade deformable models for multi-view face detection," *Image and Vision Computing*, vol. 42, pp. 47–61, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885615000967

[106] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[107] R. A. Asmara, P. Choirina, C. Rahmad, A. Setiawan, F. Rahutomo, R. D. R. Yusron, and U. D. Rosiani, "Study of drmf and asm facial landmark point for micro expression recognition using klt tracking point feature," *Journal of Physics: Conference Series*, vol. 1402, no. 7, p. 077039, dec 2019. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1402/7/077039

[108] M. I. N. P. Munasinghe, "Facial expression recognition using facial landmarks and random forest classifier," in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, 2018, pp. 423–427.

[109] Y. Huang, L. Qing, S. Xu, L. Wang, and Y. Peng, "Hybnet: A hybrid network structure for pain intensity estimation," *Vis. Comput.*, vol. 38, no. 3, p. 871–882, mar 2022. [Online]. Available: https://doi.org/10.1007/s00371-021-02056-y

[110] E. Ryumina and A. Karpov, "Facial expression recognition using distance importance scores between facial landmarks," *Proceedings of the 30th International Conference on Computer Graphics and Machine Vision (GraphiCon 2020). Part 2*, pp. paper32–1, 12 2020.

[111] D. Ghimire, J. Lee, Z.-N. Li, S. Jeong, S. Park, and H. Choi, "Recognition of facial expressions based on tracking and selection of discriminative geometric features," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, pp. 35–44, 03 2015.

[112] G. Palestra, A. Pettinicchio, M. Del Coco, P. Carcagnì, M. Leo, and C. Distante, "Improved performance in facial expression recognition using 32 geometric features," vol. 9280, 09 2015.

[113] C. Liu, K. Hirota, J. Ma, Z. Jia, and Y. Dai, "Facial expression recognition using hybrid features of pixel and geometry," *IEEE Access*, vol. 9, pp. 18 876–18 889, 2021.

[114] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[115] Z. Wang, S. Wang, and Q. Ji, "Capturing complex spatio-temporal relations among facial muscles for facial expression recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3422–3429.

[116] 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)*, 2011, pp. 1642–1649.

[117] A. Lorincz, L. A. Jeni, Z. Szabó, J. F. Cohn, and T. Kanade, "Emotional expression classification using time-series kernels," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 889–895.

[118] Y. Qiu and Y. Wan, "Facial expression recognition based on landmarks," in *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, vol. 1, 2019, pp. 1356–1360.

[119] 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)*, 2015, pp. 2983–2991.

[120] A. Cherian and S. Sra, *Positive Definite Matrices:Symmetric positive definite (SPD) matrices Data RepresentationData representationand Applications to Computer Vision*. Cham: Springer International Publishing, 2016, pp. 93–114. [Online]. Available: https://doi.org/10.1007/978-3-319-45026-1_4

[121] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 589–600.

[122] ——, "Pedestrian detection via classification on riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.

[123] J. Cavazza, A. Zunino, M. S. Biagio, and V. Murino, "Kernelized covariance for action recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 408–413.

[124] K. Zhao, S. Yang, A. Wiliem, and B. C. Lovell, "Landmark manifold: Revisiting the riemannian manifold approach for facial emotion recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 1095–1100.

[125] S. Taheri, P. Turaga, and R. Chellappa, "Towards view-invariant expression analysis using analytic shape manifolds," in *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2011, pp. 306–313.

[126] A. B. Tanfous, H. Drira, and B. B. Amor, "Sparse coding of shape trajectories for facial expression and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2594–2607, 2020.

[127] 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," *arXiv:1807.00676 [cs]*, Jun. 2018, arXiv: 1807.00676. [Online]. Available: http://arxiv.org/abs/1807.00676

[128] B. Szczapa, M. Daoudi, S. Berretti, P. Pala, A. D. Bimbo, and Z. Hammal, "Automatic estimation of self-reported pain by interpretable representations of motion dynamics," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 2544–2550.

[129] 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*, vol. 42, no. 1, pp. 1–14, 2020.

[130] D. Li, Z. Wang, Q. Gao, Y. Song, X. Yu, and C. Wang, "Facial expression recognition based on electroencephalogram and facial landmark localization," *Technology and Health Care*, vol. 27, pp. 1–15, 01 2019.

[131] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain recognition from video and biomedical signals," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 4582–4587.

[132] M. Gil-Martin, R. San-Segundo, A. Mateos, and J. Ferreiros-Lopez, "Human stress detection with wearable sensors using convolutional neural networks," *IEEE Aerospace and Electronic Systems Magazine*, vol. 37, no. 1, pp. 60–70, 2022.

[133] D. Wu, J. Zhang, and Q. Zhao, "Multimodal fused emotion recognition about expression-eeg interaction and collaboration using deep learning," *IEEE Access*, vol. 8, pp. 133 180–133 189, 2020.

[134] N. Saffaryazdi, S. T. Wasim, K. Dileep, A. F. Nia, S. Nanayakkara, E. Broadbent, and M. Billinghurst, "Using facial micro-expressions in combination with eeg and physiological signals for emotion recognition," *Frontiers in Psychology*, vol. 13, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.864047

[135] P. Sarkar and A. Etemad, "Self-supervised ecg representation learning for emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[136] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, "Uncovering the structure of clinical EEG signals with self-supervised learning," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046020, mar 2021.

[137] X. Jiang, J. Zhao, B. Du, and Z. Yuan, "Self-supervised contrastive learning for eeg-based sleep staging," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[138] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," 2020.

[139] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara, "Sigrep: Toward robust wearable emotion recognition with contrastive representation learning," *IEEE Access*, vol. 10, pp. 18 105–18 120, 2022.

[140] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob, "Emotion recognition from electrocardiogram signals using hilbert huang transform," in *2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT)*, 2012, pp. 82–86.

[141] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[142] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020.

[143] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," 2019.

[144] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[145] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, vol. 7, pp. 57–67, 2019.

[146] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems," *Journal of Medical and Biological Engineering*, vol. 40, no. 2, pp. 149–157, Apr. 2020.

[147] Y. Shu and S. Wang, "Emotion recognition through integrating eeg and peripheral signals," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2871–2875.

[148] X. Zhang, J. Liu, J. Shen, S. Li, K. Hou, B. Hu, J. Gao, T. Zhang, and B. Hu, "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine," *IEEE Transactions on Cybernetics*, vol. 51, no. 9, pp. 4386–4399, 2021.

[149] Y. Wu, Z. Zhang, P. Peng, Y. Zhao, and B. Qin, "Leveraging multi-modal interactions among the intermediate representations of deep transformers for emotion recognition," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, ser. MuSe' 22. New York, NY, USA: Association for Computing Machinery, 2022, p. 101–109. [Online]. Available: https://doi.org/10.1145/3551876.3554813

[150] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," *arXiv:1906.00295 [cs]*, Jun. 2019, arXiv: 1906.00295.

[151] Y. Ding, A. Rich, M. Wang, N. Stier, M. Turk, P. Sen, and T. Höllerer, "Sparse fusion for multimodal transformers," 2021.

[152] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 2, jun 2019.

[153] T. T. Um, F. M. J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić, "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 216–220.

[154] B. K. Iwana and S. Uchida, "An empirical survey of data augmentation for time series classification with neural networks," *PLOS ONE*, vol. 16, no. 7, pp. 1–32, 07 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0254841

[155] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *ArXiv*, vol. abs/1803.01271, 2018.

[156] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[157] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Scientific Data*, vol. 6, no. 1, p. 196, Oct. 2019, number: 1 Publisher: Nature Publishing Group.

[158] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, no. 1, p. 293, Sep. 2020, number: 1 Publisher: Nature Publishing Group.

[159] P. Sarkar and A. Etemad, "Self-supervised ecg representation learning for emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[160] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," 2019. [Online]. Available: https://arxiv.org/abs/1906.00295

[161] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[162] X. Jiang, J. Zhao, B. Du, and Z. Yuan, "Self-supervised contrastive learning for eeg-based sleep staging," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[163] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network," *Nature medicine*, vol. 25, no. 1, pp. 65–69, Jan. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6784839/

[164] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition," *Information Processing & Management*, vol. 57, no. 3, p. 102185, 2020.

[165] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual lstm network," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 176–183.

[166] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," *Frontiers in Human Neuroscience*, vol. 15, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2021.653659

[167] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[168] A. Newell and J. Deng, "How useful is self-supervised pretraining for visual tasks?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[169] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[170] J. Wijsman, B. Grundlehner, H. Liu, J. Penders, and H. Hermens, "Wearable physiological sensors reflect mental stress state in office-like situations," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 600–605.

[171] J. Hernandez, P. Paredes, A. Roseway, and M. Czerwinski, "Under pressure: Sensing stress of computer users," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 51–60.

[172] W. Sato, T. Kochiyama, and S. Yoshikawa, "Physiological correlates of subjective emotional valence and arousal dynamics while viewing films," *Biological Psychology*, vol. 157, p. 107974, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301051120301344

[173] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, pp. 2302–2310, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16330

[174] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2114–2124.

[175] J. L. Suárez, S. García, and F. Herrera, "A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges," *Neurocomputing*, vol. 425, pp. 300–322, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220312777

[176] B. S. McEwen and E. Stellar, "Stress and the individual. mechanisms leading to disease." *Archives of internal medicine*, vol. 153 18, pp. 2093–101, 1993.

[177] G. P. Chrousos and P. W. Gold, "The concepts of stress and stress system disorders. overview of physical and behavioral homeostasis." *JAMA*, vol. 267 9, pp. 1244–52, 1992.

[178] S. N. Raja, D. B. Carr, M. Cohen, N. B. Finnerup, H. Flor, S. Gibson, F. Keefe, J. S. Mogil, M. Ringkamp, K. A. Sluka, X.-J. Song, B. Stevens, M. Sullivan, P. Tutelman, T. Ushida, and K. Vader, "The Revised IASP definition of pain: concepts, challenges, and compromises," *Pain*, vol. 161, no. 9, pp. 1976–1982, Sep. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7680716/

[179] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. W. Picard, "Automatic recognition methods supporting pain assessment: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 530–552, 2022.

[180] J. Aigrain, M. Spodenkiewicz, S. Dubuisson, M. Detyniecki, D. Cohen, and M. Chetouani, "Multimodal stress detection from multiple assessments," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 491–506, 2018.

[181] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm, "Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity," in *Engineering Applications of Neural Networks*, ser. Communications in Computer and Information Science, L. Iliadis and C. Jayne, Eds. Cham: Springer International Publishing, 2015, pp. 275–285.

[182] R. Bhatia, *Positive Definite Matrices*. Prinecton University press, 2009.

[183] J. Zhang, Z. Feng, Y. Su, and M. Xing, "Cross-covariance matrix: Time-shifted correlations for 3D action recognition," *Signal Processing*, vol. 171, p. 107499, Jun. 2020.

[184] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.

[185] N. Otberdout, A. Kacem, M. Daoudi, L. Ballihi, and S. Berretti, "Automatic analysis of facial expressions based on deep covariance trajectories," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 10, pp. 3892–3905, 2020.

[186] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Classification of covariance matrices using a Riemannian-based kernel for BCI applications," *Neurocomputing*, vol. 112, pp. 172–178, Jul. 2013.

[187] M. Liu, R. Wang, S. Li, Z. Huang, S. Shan, and X. Chen, "Video modeling and learning on Riemannian manifold for emotion recognition in the wild," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 113–124, Jun. 2016.

[188] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian Framework for Tensor Computing," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, Jan. 2006. [Online]. Available: http://link.springer.com/10.1007/s11263-005-3222-z

[189] M. Faraki, M. Palhang, and C. Sanderson, "Log-euclidean bag of words for human action recognition," *IET Computer Vision*, vol. 9, no. 3, pp. 331–339, jun 2015.

[190] S. Jayasumana, R. I. Hartley, M. Salzmann, H. Li, and M. T. Harandi, "Kernel methods on riemannian manifolds with gaussian RBF kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2464–2477, 2015.

[191] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass Brain–Computer Interface Classification by Riemannian Geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, Apr. 2012.

[192] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Communications on Pure and Applied Mathematics*, vol. 30, no. 5, pp. 509–541, 1977. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160300502

[193] P. T. Fletcher and S. Joshi, "Principal Geodesic Analysis on Symmetric Spaces: Statistics of Diffusion Tensors," in *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis*, ser. Lecture Notes in Computer Science, M. Sonka, I. A. Kakadiaris, and J. Kybic, Eds. Berlin, Heidelberg: Springer, 2004, pp. 87–98.

[194] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 3438–3446.

[195] S. Hinduja, S. J. Canavan, and G. Kaur, "Multimodal fusion of physiological signals and facial action units for pain recognition," in *15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020, Buenos Aires, Argentina, November 16-20, 2020*. IEEE, 2020, pp. 577–581.

[196] C. J. Patrick, K. D. Craig, and K. M. Prkachin, "Observer judgments of acute pain: facial action determinants," *Journal of Personality and Social Psychology*, vol. 50, no. 6, pp. 1291–1298, Jun. 1986.

[197] L. LeResche, "Facial expression in pain: A study of candid photographs," *Journal of Nonverbal Behavior*, vol. 7, no. 1, pp. 46–56, Sep. 1982.

[198] S. Samyoun, A. Sayeed Mondol, and J. A. Stankovic, "Stress Detection via Sensor Translation," in *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. Marina del Rey, CA, USA: IEEE, May 2020, pp. 19–26.

[199] M. Gil-Martin, R. San-Segundo, A. Mateos, and J. Ferreiros-Lopez, "Human stress detection with wearable sensors using convolutional neural networks," *IEEE Aerospace and Electronic Systems Magazine*, vol. 37, no. 1, pp. 60–70, 2022.

[200] B. Szczapa, M. Daoudi, S. Berretti, P. Pala, A. D. Bimbo, and Z. Hammal, "Automatic estimation of self-reported pain by interpretable representations of motion dynamics," in *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*. IEEE, 2020, pp. 2544–2550.

[201] K. W. Choo and T. Du, "Pain detection from facial landmarks using spatial-temporal deep neural network," in *Thirteenth International Conference on Digital Image Processing (ICDIP 2021)*, X. Jiang and H. Fujita, Eds., vol. 11878, International Society for Optics and Photonics. SPIE, 2021, pp. 593 – 597.

[202] Y. Huang, L. Qing, S. Xu, L. Wang, and Y. Peng, "Hybnet: a hybrid network structure for pain intensity estimation," *Vis. Comput.*, vol. 38, no. 3, pp. 871–882, 2022.

[203] D. D. G. Krishnan, D. A. V. Keloth, and D. S. Ubedulla, "Pros and cons of simulation in medical education: A review," *International Journal of Medical and Health Research*, vol. 3, no. 6, pp. 84–87, Jun. 2017. [Online]. Available: http://www.medicalsciencejournal.com/archives/2017/vol3/issue6/3-6-15

[204] S. Ollander, C. Godin, A. Campagne, and S. Charbonnier, "A comparison of wearable and stationary sensors for stress detection," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Budapest, Hungary: IEEE, Oct. 2016, pp. 4362–4366. [Online]. Available: http://ieeexplore.ieee.org/document/7844917/

[205] "The State-Trait Anxiety Inventory | Revista Interamericana de Psicología/Interamerican Journal of Psychology," Apr. 2020. [Online]. Available: https://journal.sipsych.org/index.php/IJP/article/view/620

[206] R. S. Lazarus, *Fifty Years of the Research and theory of R.s. Lazarus: An Analysis of Historical and Perennial Issues*. Psychology Press, Jun. 2013.

[207] A. Greco, G. Valenza, A. Lanata, E. Scilingo, and L. Citi, "cvxEDA: a Convex Optimization Approach to Electrodermal Activity Processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 1–1, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7229284/

[208] J. Bai, C. Di, L. Xiao, K. R. Evenson, A. Z. LaCroix, C. M. Crainiceanu, and D. M. Buchner, "An activity index for raw accelerometry data and its comparison with other activity metrics," *PLOS ONE*, vol. 11, no. 8, pp. 1–14, 08 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0160644

[209] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[210] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[211] M. Daoudi, N. Otberdout, and J.-C. Á. Paiva, "Metric learning on the manifold of oriented ellipses: Application to facial expression recognition," in *Pattern Recognition. ICPR International Workshops and Challenges*, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds. Cham: Springer International Publishing, 2021, pp. 196–206.

[212] A. Kacem, M. Daoudi, and J.-C. Alvarez-Paiva, "Barycentric representation and metric learning for facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 443–447.

[213] S. Wan and J. Aggarwal, "Spontaneous facial expression recognition: A robust metric learning approach," *Pattern Recognition*, vol. 47, no. 5, pp. 1859–1868, 2014.

[214] L. Shao, L. Liu, and M. Yu, "Kernelized Multiview Projection for Robust Action Recognition," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 115–129, Jun. 2016.

[215] M. Daoudi, Z. Hammal, A. Kacem, and J. F. Cohn, "Gram matrices formulation of body shape motion: An application for depression severity assessment," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2019, pp. 258–263.

[216] B. Szczapa, M. Daoudi, S. Berretti, P. Pala, A. D. Bimbo, and Z. Hammal, "Automatic estimation of self-reported pain by interpretable representations of motion dynamics," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 2544–2550.

[217] 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.

[218] J. Townsend, N. Koep, and S. Weichwald, "Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation," *Journal of Machine Learning Research*, vol. 17, no. 137, pp. 1–5, 2016.

[219] 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)*, 2015, pp. 2983–2991.

[220] A. B. Tanfous, H. Drira, and B. B. Amor, "Sparse Coding of Shape Trajectories for Facial Expression and Action Recognition," *arXiv:1908.03231 [cs]*, Aug. 2019, arXiv: 1908.03231. [Online]. Available: http://arxiv.org/abs/1908.03231

[221] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[222] S. Taheri, P. Turaga, and R. Chellappa, "Towards view-invariant expression analysis using analytic shape manifolds," in *Face and Gesture 2011*. Santa Barbara, CA, USA: IEEE, Mar. 2011, pp. 306–313. [Online]. Available: http://ieeexplore.ieee.org/document/5771415/

[223] S. Jain, Changbo Hu, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. Barcelona, Spain: IEEE, Nov. 2011, pp. 1642–1649. [Online]. Available: http://ieeexplore.ieee.org/document/6130446/

[224] Z. Wang, S. Wang, and Q. Ji, "Capturing Complex Spatio-temporal Relations among Facial Muscles for Facial Expression Recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, Jun. 2013, pp. 3422–3429.

[225] Y. Ding, A. Rich, M. Wang, N. Stier, M. Turk, P. Sen, and T. Höllerer, "Sparse fusion for multimodal transformers," 2021. [Online]. Available: https://arxiv.org/abs/2111.11992

[226] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021. [Online]. Available: https://doi.org/10.1109%2Ftkde.2021.3090866

[227] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," 2017. [Online]. Available: https://arxiv.org/abs/1705.09406

[228] Z. Huang and L. Van Gool, "A riemannian network for spd matrix learning," 2016. [Online]. Available: https://arxiv.org/abs/1608.04233

[229] R. Wang, X.-J. Wu, Z. Chen, T. Xu, and J. Kittler, "Dreamnet: A deep riemannian network based on spd manifold learning for visual classification," 2022. [Online]. Available: https://arxiv.org/abs/2206.07967